RadThinking: A Dataset for Longitudinal Clinical Reasoning in Radiology

Wenxuan Li[1] **Pedro R. A. S. Bassi**[1] **Xinze Zhou**[1] **Jakob Wasserthal**[2]
Alan L. Yuille[1] **Zongwei Zhou**[1,3*]

[1]Department of Computer Science, Johns Hopkins University
[2]Clinic of Radiology and Nuclear Medicine, University Hospital Basel
[3]Department of Oncology, Johns Hopkins School of Medicine

Code, Models & Data: https://huggingface.co/datasets/wenxuanchelsea/RadThinking

Abstract

Cancer screening is evidence-based reasoning. A radiologist collects imaging evidence on the scan, compares it to prior scans, integrates clinical context, and reaches a diagnostic conclusion confirmed by pathology. The same evidence cues then drive the next step: searching for similar patients and choosing the management that helped them. Public CT datasets release the pixel and the segmentation, but not the evidence trail. We present RadThinking, a Visual Question Answering (VQA) dataset that makes this evidence trail explicit and trainable. RadThinking releases VQA pairs at three difficulty tiers. *Foundation VQAs* are atomic evidence cues read from the image. *Single-step reasoning VQAs* apply one clinical rule to one cue. *Compositional VQAs* combine cues across rules to reach a guideline category such as LI-RADS-5. For every compositional VQA, we release the chain of foundation VQAs that supports it. The chain follows the rules of the governing clinical reporting standard. The dataset spans 20,362 CT scans from 9,131 patients across 43 cancer groups, plus 2,077 verified healthy controls with $\geq$1-year follow-up. To our knowledge, RadThinking is the first cancer-screening VQA corpus that stratifies imaging evidence by reasoning depth, grounds compositions in clinical reporting standards, and anchors conclusions to pathology. The foundation tier supplies atomic evidence supervision. The compositional tier supplies chain-of-thought data and verifiable rewards for reinforcement-learning recipes such as DeepSeek-R1 and OpenAI o1. Findings and diagnosis are released as separate verifiable layers; a correct diagnosis with no supporting findings is a black box, and RadThinking makes the supporting cues explicit so it cannot stay one. RadThinking enables systematic training and evaluation of whether AI systems can *reason from evidence* about cancer, and supplies the structured evidence substrate needed to search at population scale across patients no single radiologist could ever review.

1 Introduction

A radiologist screening a CT scan for cancer does not simply look for bright or dark spots. They reason from evidence. They collect imaging evidence on the scan, compare it to prior scans, integrate clinical context, and reach a diagnosis confirmed by pathology. The same evidence then drives the next step: searching for similar patients in the literature or in institutional memory, and choosing management based on what those patients showed. Cancer screening is therefore an evidence-collection task wrapped around a patient-similarity search. AI systems can excel at the search side

*Correspondence to: Zongwei Zhou (ZZHOU82@JH.EDU)

Preprint.

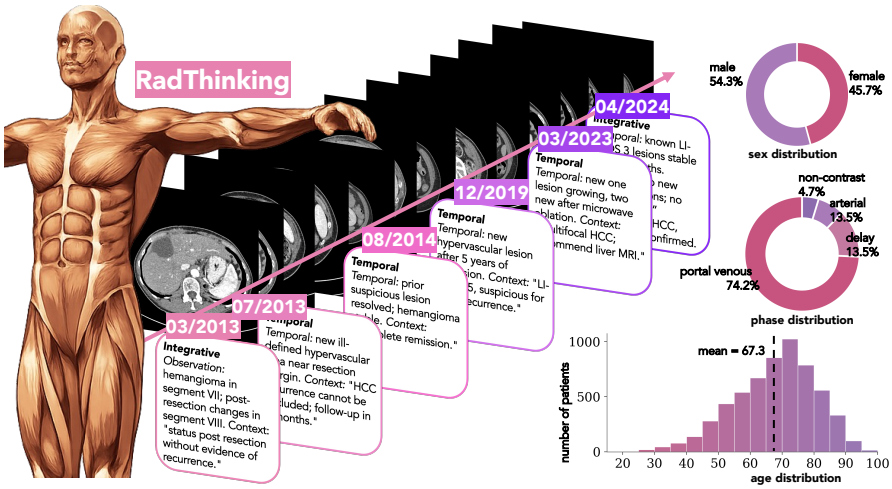

Figure 1: **Overview of RadThinking.** *Left:* the reasoning trajectory of an illustrative hepatocellular carcinoma (HCC) patient. We monitored this patient across 26 CT scans over 11 years (2013–2024). Six selected timepoints show how reasoning complexity evolves: post-resection baseline (Scan 1), a new suspicious lesion (Scan 2), resolution confirming a benign finding (Scan 5), recurrence after 5 years of remission (Scan 17), progression after ablation (Scan 24), and stability that downgrades concern (Scan 26). Each scan carries a four-step reasoning chain: observations, temporal comparison, clinical context, and a pathology-confirmed conclusion. *Right:* dataset characteristics. RadThinking has 20,362 CT scans from 9,131 patients across 19 organ screening targets (43 cancer groups). The distributions cover patient age, sex, and contrast phase.

because they can be trained on far more patient trajectories than any single radiologist will ever read. The bottleneck is the evidence side, where the cues are hard to elicit, hard to standardize, and hard to verify.

Public CT datasets reduce this process to perception. They provide a scan and a segmentation mask [19, 49, 70, 83, 8]. Evaluation asks one question: did the model find the tumor? They lack longitudinal trajectories. They lack radiology reports. They lack clinical variables. Models trained on them are optimized for perception, not for the evidence chain [13, 75, 16, 113]. On the text side, clinical retrieval tools such as OpenEvidence [88] place published literature evidence at a clinician's fingertips. They cannot read a CT scan. They cannot generate the per-patient imaging cues that decide the differential. The hardest cancer-screening cases need exactly these cues. Early-stage tumors are ambiguous on a single scan. Temporal comparison and clinical context turn ambiguous pixels into decisive evidence [71, 64].

We introduce RadThinking (Fig. 1), a VQA dataset for multicancer screening at three difficulty tiers. Foundation VQAs expose the atomic evidence cues a radiologist reads from the image. Compositional VQAs combine these cues, under the rules of a clinical reporting standard, to reach a guideline category. The two tiers decouple *finding* from *diagnosis*: findings are descriptive and interpretable, a diagnosis without findings is a black box, and the right findings can be reused to support a better diagnosis. The radiology lexicon RadLex [95, 40] already encodes this decoupling at the vocabulary level by separating imaging observations from disorders; RadThinking extends it to a full chain of trainable VQA pairs. This mirrors how compositional reasoning in general AI is built from atomic primitives [93, 121, 60, 44, 103]. RadThinking trains vision-language models (VLMs) along two paths. The foundation tier supplies atomic visual skills [52]. The compositional tier supplies chain-of-thought data [109, 114, 99] and verifiable rewards for RL recipes such as DeepSeek-R1 [43] and OpenAI o1 [87]. The two tiers form a curriculum [17, 34]. The same structured evidence is the natural index for retrieving similar patients at population scale, a use case where AI can leverage many more cases than any one radiologist remembers. We release RadThinking under CC BY-NC-SA 4.0.

Related work. Table 1 positions RadThinking against representative datasets across seven properties. We summarize each group below and identify the gap that RadThinking fills.

Table 1: **RadThinking in context.** Comparison against representative cancer-screening, medical-VQA, and clinical-reasoning resources across seven properties: 3D imaging, voxel-wise tumor masks, paired radiology reports, longitudinal scans, VQA pair release, multi-tier difficulty stratification, and pathology-confirmed ground truth. ✓ present, ✗ absent.

Dataset	Modality	3D	Voxel	Report	Long.	VQA	Tiers	Path.
Cancer segmentation datasets (perception only).								
KiTS/LiTS/MSD/PanTS [49, 19, 8, 70]	CT	✓	✓	✗	✗	✗	✗	✓
AbdomenAtlas 2.0 [23]	CT	✓	✓	✗	✗	✗	✗	✗
ULS23 [33]	CT	✓	✓	✗	✗	✗	✗	✗
Imaging paired with text.								
CT-RATE [46]	Chest CT	✓	✗	✓	✗	✗	✗	✗
MIMIC-CXR [57]	X-ray	✗	✗	✓	✗	✗	✗	✗
RadGPT [15]	Abd. CT	✓	✓	✓	✗	✗	✗	✗
Medical visual question answering.								
VQA-RAD/SLAKE/PathVQA [63, 73, 47]	2D mixed	✗	✗	✗	✗	✓	✗	✗
OmniMedVQA [53]	2D mixed	✗	✗	✗	✗	✓	✗	✗
M3D-VQA [10]	3D Med.	✓	✓	✗	✗	✓	✗	✗
3D-RAD [41]	3D CT	✓	✗	✗	✓	✓	✗	✗
DeepTumorVQA [24]	CT	✓	✗	✗	✗	✓	✗	✗
Kvasir-VQA-x1 [42]	Endoscopy	✗	✗	✗	✗	✓	✓	✗
Reasoning chains for medicine.								
MedReason / HuatuoGPT-o1 [111, 21]	Text	✗	✗	✗	✗	✓	✗	✗
PhysicianBench [76]	EHR text	✗	✗	✗	✓	✗	✗	✗
CheXthought [100]	X-ray	✗	✗	✓	✗	✗	✗	✗
RadThinking (ours)	**CT**	✓	✓	✓	✓	✓	✓	✓

Cancer segmentation datasets. KiTS, LiTS, PanTS, BraTS, MSD [49, 19, 70, 83, 8] pair scans with voxel masks. Multi-organ atlases [67, 68, 23, 74, 12, 33] scale up. None pair scans with text or reasoning.

Imaging plus text. CT-RATE, MIMIC-CXR, CT2Rep, RadGPT [46, 57, 45, 15] pair imaging with reports. They cover one body region and lack reasoning structure.

Medical VQA. 2D resources [63, 73, 47, 119, 53] pair images with short factual answers. 3D extensions [10, 41, 24, 108] reach volumes but do not stratify questions by reasoning depth. Kvasir-VQA-x1 [42] stratifies questions for endoscopy. Compositional-VQA work in general AI [93, 121, 60, 44, 103, 6, 98, 58, 55] establishes the decomposition primitive. It has not been operationalized for cancer screening.

Reasoning chains in medicine. MedReason, HuatuoGPT-o1, Med-PRM [111, 21, 116] are text-only. PhysicianBench [76] decomposes EHR tasks into 670 checkpoints, demonstrating that visible chains expose model failures, but operates without imaging. CheXthought [100] releases free-form CoT for chest X-ray, not structured VQA. Medical VLMs Merlin, RadFM, Med-Gemini [20, 110, 96] establish broader radiology VLM training.

Radiology lexicons and structured reporting. Free-form radiology reports are a major bottleneck for both clinical interoperability and AI training: the same finding can be written a dozen ways. RadLex [95, 62, 40], the RSNA's controlled vocabulary of over 75,000 terms, was built to standardize this language. Its ontology explicitly separates *imaging observations* from *disorders*, formalizing the principle that findings and diagnoses are different categories of evidence. RadLex is also increasingly used to remind radiologists which features to search for on a given scan. RadThinking inherits both functions. Its Step 1 features are designed to be compatible with the RadLex finding vocabulary, and its foundation VQAs serve as a per-organ evidence checklist that mirrors how RadLex prompts a reader. The four-step chain extends the finding-versus-disorder decoupling end-to-end: every finding is a foundation VQA, every diagnosis is a compositional VQA conclusion, and the chain that connects them is auditable. RadLex provides the vocabulary; RadThinking provides the trained reasoning data on top of it.

Text-only clinical evidence retrieval. A complementary line of work surfaces evidence from the published literature rather than from the patient's own imaging. OpenEvidence [88] is the most widely used example. It is an AI copilot for clinicians that runs retrieval-augmented generation over peer-reviewed journals (e.g., NEJM, JAMA, Lancet, Cochrane) and returns cited summaries to free-form clinical questions. As of 2026 it is used by more than 40% of practicing physicians in the

U.S. and partners with NEJM. OpenEvidence treats evidence as something to be *retrieved* from prior publications. RadThinking treats evidence as something to be *extracted* from the patient's own CT scan and verified against pathology. The two are complementary: literature evidence tells a clinician what similar patients tended to show; per-patient imaging evidence is what determines whether *this* patient is similar in the first place. RadThinking supplies the missing imaging side at scale.

Innovation. RadThinking is the first resource that releases cancer-screening VQA pairs at multiple reasoning-depth tiers, decomposes each compositional question into foundation VQAs organized by clinical reporting standards, and anchors conclusions to pathology with longitudinal voxel-grounded imaging. The released foundation VQAs constitute a per-patient evidence vocabulary that is grounded in pixels, aligned with RadLex, scored against guidelines, and verified by tissue diagnosis. Findings and diagnoses are released as separate auditable layers, not collapsed into a single label. This is the substrate the field needs to move from population-level literature retrieval (e.g., OpenEvidence) to patient-level reasoning and patient-similarity search.

2 The RadThinking Dataset

2.1 What RadThinking Contains

The primary released artifact is a corpus of (CT scan, question, answer) triples at three difficulty tiers (§3). Compositional triples additionally carry a chain of foundation triples that solves them. Supporting artifacts are released alongside: standardized NIfTI CT volumes, voxel-wise tumor masks across 19 organ screening targets (twelve organs with no prior public CT tumor annotations), paired de-identified radiology reports and clinical variables, pathology labels for cancer-positive patients, and a confirmation of $>$1-year cancer-free follow-up for verified healthy patients.

Table 2 summarizes the per-patient JSON. It records the four-step reasoning chain that underlies every compositional VQA. Each trace stores the four steps (§4) plus metadata, parsed report, and risk category. The full schema appears in Appendix A. The JSON is the source from which the VQA pairs are generated (§6).

2.2 Cohort, Annotation, and Multimodal Data

RadThinking contains 9,131 patients and 20,362 pelvic, abdominal, and thoracic CT scans from 10 European institutions, acquired 2012 to 2025 under IRB and ethics approval. The median follow-up is 1.17 years per patient. The cancer-positive cohort has confirmed malignancies across 43 cancer groups spanning 19 organ screening targets, with all prior scans retained per patient. The verified-healthy cohort has $>$1-year follow-up after the last CT. We split strictly at the patient level, stratified by cancer type, organ target, and institution: $N_{\text{train}} = 7{,}305$ patients (16,290 scans) and $N_{\text{test}} = 1{,}826$ patients (4,072 scans).

CT volumes are released in standardized NIfTI format with harmonized orientation and voxel spacing. Voxel-wise tumor masks cover all 19 organ screening targets; twelve of these targets have no prior public CT tumor annotation[2]. Annotation used a three-stage protocol: 28 radiologist residents produced initial masks via MONAI Label [36], two of eight board-certified radiologists independently reviewed each case, and a separate radiologist adjudicated discrepancies. The protocol builds on prior scalable CT annotation [94, 69, 118, 28, 14, 25]. Mean inter-reviewer Dice on a 200-patient validation cohort is 62.2% (per-organ in Appendix B). Cases with Dice < 0.30 receive an *ambiguity flag* that feeds the complexity stratification (§B.6).

Each scan is paired with the de-identified radiology report and the clinical variables available at imaging time. A parsing pipeline (Appendix B) extracts findings, impression, and recommendation. Pathology serves as the ground-truth conclusion (§B.4) for cancer-positive patients; absence of cancer over $>$1-year follow-up is the negative ground truth for healthy patients. All chain inputs reflect only information available at or before imaging time, which prevents future-information leakage.

[2]The seven organs with existing public CT tumor masks are liver [19], kidney [49], and pancreas, colon, lung, spleen, prostate [8]. The twelve targets without prior public masks are thyroid, breast, esophagus, gallbladder, stomach, duodenum, adrenal, bladder, uterus, ovary, lymph node, and bone.

Table 2: **The per-patient JSON file in RadThinking.** Every patient is a single JSON record. Patient-level fields summarize the patient. The `reasoning_traces` list contains one structured reasoning chain per CT scan. Step 1 to Step 4 mirror the chain definition in Eq. 1.

field	description
Patient-level fields	
`patient_id`	anonymized patient identifier; one folder of CT volumes per id
`primary_cancer`	resolved cancer type, confidence, source, all candidate scores, metastasis flag
`clinical_history`	list of prior diagnoses, procedures, and oncological status
`num_scans`, `date_range`	length and time span of the longitudinal sequence
`reasoning_traces`	list of one reasoning chain per CT scan (fields below)
Per-scan trace fields	
`metadata`	scan id, accession, date, scan index, age, sex, contrast phase, malignancy/metastasis flags
`step1_observations`	list of findings; each finding stores organ, location, size, attenuation, tumor type, certainty, malignancy/metastasis flags, governing clinical standard (§B.1)
`step2_temporal`	per-lesion change labels (NEW, GROWING, STABLE, SHRINKING, RESOLVED); interval to prior scan; counts of new, matched, and resolved findings (§B.2)
`step3_clinical_context`	parsed report (findings, impression, recommendation), RECIST assessment, organ-specific risk category, structured clinical variables, raw report text (§B.3)
`step4_conclusion`	primary cancer with confidence, organ-level diagnosis, ICD-10 code, metastatic disease flag (§B.4)
`reasoning_complexity`	one of PERCEPTUAL, TEMPORAL, INTEGRATIVE, AMBIGUOUS (§B.6)

3 VQA Tiers and Compositional Structure

RadThinking organizes its VQA pairs into three difficulty tiers. *Foundation VQAs* (§3.1) are atomic evidence cues whose answers come from a single annotated field. *Single-step reasoning VQAs* (§3.2) apply one explicit clinical rule to one foundation cue. *Compositional VQAs* (§3.3) require multiple foundation cues to be composed via the rules of a clinical reporting standard. The three tiers form a curriculum from atomic perception to multi-step clinical reasoning. Box 1 makes the chain visible. It shows a complete reasoning trace for one compositional VQA, decomposed into the foundation VQAs that supply the evidence at each step. This mirrors how recent agentic medical benchmarks expose the reasoning chain through structured checkpoints [76, 100], but with vision in the loop and with each checkpoint formulated as a verifiable, evidence-grounded VQA pair.

The chain trace in Box 1 is the central artifact RadThinking releases. Every compositional VQA in the dataset comes with its trace recorded in this form. The atomic Q_i at each step is itself a foundation VQA from §3.1. The composition rule is the published rule of the governing standard. The final answer is verified by tissue diagnosis or by structured follow-up. We now define each tier formally.

3.1 Foundation VQA: Atomic Evidence Cues

Foundation VQAs ask questions whose answers are read directly from one field of the JSON record. They cover the atomic evidence cues a radiologist reads from the image and train atomic visual skills [52, 99]. Examples grouped by source field follow.

Modality and acquisition. "What modality is this image?" (CT). "What is the contrast phase?" (portal venous). "What body region is shown?" (abdomen). Source: scan metadata.

Anatomical presence. "Is the liver visible in this scan?" (yes/no). "Are the kidneys included?". Source: canonical organ list and body-region tag.

Lesion presence and basic geometry. "Are any lesions present?". "How many lesions in the liver?". "What is the size of the largest hepatic lesion?". "Where is the lesion located?". Source: `step1_observations`.

Lesion attenuation and morphology. "Is the lesion hypodense in the portal-venous phase?". "Does the lesion have a calcified rim?". Source: `step1_observations`.

Patient demographics and history. "What is the patient's age?". "Does the patient have known cirrhosis?". Source: `clinical_variables` and `clinical_history`.

Temporal change atoms. "Is the lesion present in both this scan and the prior?". "Did the longest diameter increase?". Source: `step2_temporal`.

Box 1: Reasoning Chain Trace for a Compositional VQA (Patient HCC, Scan 17, 2019-12)

Compositional VQA. *What is the LI-RADS category of the new hepatic lesion in segments VI/VII?*
Patient context. HCC post-resection (2013); chronic at-risk for HCC; six prior scans.

① Step 1: Imaging Observations (Foundation VQAs)

Q_1: Are any new hepatic lesions present? **A:** Yes, in segments VI/VII.
Q_2: What is the lesion size? **A:** 1.4 cm.
Q_3: Is arterial-phase hyperenhancement present? **A:** Yes.
Q_4: Is portal-venous washout present? **A:** Yes.
Q_5: Is an enhancing capsule visible? **A:** Yes.
Q_6: Is tumor in vein present? **A:** No.

② Step 2: Temporal Comparison (Foundation VQAs)

Q_7: Was this lesion present in the prior scan? **A:** No (NEW).
Q_8: How long since the prior imaging? **A:** 5 months.

③ Step 3: Clinical Context (Foundation VQAs)

Q_9: Cancer history? **A:** Hepatocellular carcinoma (post-resection 2013).
Q_{10}: Is the patient at risk for HCC? **A:** Yes.
Q_{11}: Is the LI-RADS pathway eligible? **A:** Yes.

④ Step 4: Compose under LI-RADS v2018 [26]

Rule input. At-risk (yes) $\wedge$ size $\geq$1 cm (1.4 cm) $\wedge$ arterial-phase hyperenhancement (yes) $\wedge$ washout (yes) $\wedge$ enhancing capsule (yes).
Rule output. LR-5 (definite HCC).
Verification. Pathology-confirmed HCC recurrence after 5-year remission.

3.2 Single-Step Reasoning VQA

Single-step reasoning VQAs apply one explicit clinical rule to one foundation observation. They require one inference step. Examples follow.

Threshold rules. "Is the renal lesion at or above 1 cm by Bosniak criteria?" [101]. "Is the liver lesion 2 cm or larger by LI-RADS threshold?" [26]. The atom is the size; the rule is the standard's threshold.

Single-feature classification. "Does the lesion show arterial-phase hyperenhancement?". "Does the lesion show portal-venous washout?". The atom is the attenuation; the rule is the LI-RADS feature definition.

Single-change rules. "Is the lesion growing per RECIST 1.1?" [39]. The atom is the volume ratio; the rule is the 20% threshold.

Single-context rules. "Is the patient eligible for the LI-RADS pathway?". The atoms are cirrhosis status and chronic hepatitis B status; the rule is the LI-RADS at-risk definition.

3.3 Compositional VQA: Multi-Step Chain-of-Thought

Compositional VQAs require multiple foundation answers to be combined under the rules of a clinical reporting standard. The combination rule is given by the standard, not learned. The hard VQAs are the clinical decisions that radiologists actually make.

LI-RADS category. "What is the LI-RADS category of the largest hepatic lesion?". Foundation chain: (i) Is the patient at risk for HCC? (ii) What is the lesion size? (iii) Does it show arterial-phase hyperenhancement? (iv) Does it show washout? (v) Does it show an enhancing capsule? (vi) Does it meet threshold growth? The composition rule maps these to LR-1 through LR-5.

RECIST response. "What is the RECIST 1.1 response category?". Foundation chain: sum of target-lesion diameters at baseline and now, presence of new lesions, percent change. Rule: complete response, partial response, stable disease, or progressive disease.

TNM staging. "What is the T-stage of this gastric tumor?" [2]. Foundation chain: wall thickness, serosal invasion, perigastric fat involvement, adjacent-organ invasion. Rule: TNM T1 to T4.

Differential under ambiguity. "Given imaging, history, and prior scans, what is the most likely diagnosis?". Foundation chain: lesion characteristics, temporal change, clinical history, risk factors. Rule: the radiologist's standard differential workflow.

3.4 Clinical Reporting Standards as Compositional Grammars

For every organ screening target, professional societies have codified a compositional grammar. LI-RADS [26], PI-RADS [106], BI-RADS [38], Bosniak [101], and the others listed in Table 3 each define (a) the atomic features to evaluate and (b) a deterministic rule that maps these features to the final risk or diagnostic category. RadThinking uses these grammars directly. The atomic features become foundation VQAs. The deterministic rule becomes the program for the compositional VQA. The four-step scaffold introduced next (§4) is the canonical decomposition path: every compositional VQA is built from foundation VQAs that fall into one of four steps, namely imaging observations, temporal comparison, clinical context, and the diagnostic conclusion.

4 Constructing Structured Clinical Reasoning Chains

The four-step chain is the scaffold that decomposes compositional VQAs (§3.3) into foundation VQAs (§3.1). Steps 1 to 3 collect three complementary kinds of evidence (imaging, temporal, contextual). Step 4 anchors the resulting decision to pathology. For each scan I_t with tumor mask M_t, radiology report R_t, clinical variables C_t, pathology P, and governing standard $\mathcal{S}$ (Table 3), we construct

$$\mathcal{T}_t = \langle\, \mathcal{O}_t,\ \Delta_t,\ \mathcal{C}_t,\ \mathcal{D}_t \,\rangle, \tag{1}$$

where $\mathcal{O}_t$ are imaging observations grounded in M_t and $\mathcal{S}$, Δ_t is the temporal change relative to prior scans, $\mathcal{C}_t$ is the clinical context parsed from R_t and C_t, and $\mathcal{D}_t$ is the pathology-confirmed conclusion. We do not invent a reasoning vocabulary. We extract from each report the features that the governing standard prescribes. The construction pipeline (Algorithm 1 in Appendix B) processes preprocessing, per-step extraction, and quality control. Per-step formal definitions and validation metrics are in Appendix B.

Validation. The 200-patient cohort used for annotation QC (§2.2) is reused for every pipeline step. Eight board-certified radiologists assess each step independently. The headline numbers are 62.2% inter-annotator Dice for spatial annotations, 94.6% feature-extraction accuracy (Step 1), 95.9% temporal-label agreement (Step 2), 97.1% report-parsing accuracy (Step 3), and Fleiss' $\kappa = 0.947$ for complexity stratification. Per-step breakdowns are in Appendix B.

Complexity stratification. Each chain is labeled PERCEPTUAL, TEMPORAL, INTEGRATIVE, or AMBIGUOUS, reflecting the information depth required to reach the conclusion. The decision rules and validation are in Appendix B.

Table 3 maps each of the 19 organ screening targets to its governing standard. The features become foundation VQAs and the standard's rule becomes the program for compositional VQA.

5 Dataset Statistics and Analysis

RadThinking contains 20,362 structured reasoning chains, one per CT scan, from 9,131 patients.

Cancer type distribution. After merging clinically equivalent subtypes (Appendix E), the dataset spans 43 cancer groups (Fig. 2) with a pronounced long-tail distribution. The five most frequent groups are hepatocellular carcinoma, pancreatic IPMN, renal cell carcinoma, breast carcinoma, and colorectal carcinoma. Together they account for over 50% of cancer-positive patients. Of the 43 groups, 31 map to the 19 organ screening targets (Table 3). The remaining 12 (507 patients) are cancers detected outside the screened organs via metastatic deposits or incidental findings (Appendix F). Healthy controls account for 2,077 patients (22.7%). Among cancer-positive patients, 2,563 (36.3%) have longitudinal imaging with 2 to 26 scans per patient.

Reasoning complexity distribution. Across all chains (Appendix B.6), *integrative* reasoning has the largest share at 39.2%. *Ambiguous* follows at 36.4%, then *perceptual* at 12.9%, and *temporal* at

Table 3: **Clinical reporting standards governing each cancer type in RadThinking.** Each standard defines the imaging features, risk categories, and management recommendations used in clinical practice. Our reasoning chains extract report content and align it with these organ-specific feature vocabularies.

target organ	standard	key imaging features	risk stratification
thyroid	ACR IF [51]	nodule size on CT, density, calcification, extrathyroidal extension	$<$1 cm (ignore) to $>$2.5 cm (further imaging)
lung	Lung-RADS v2022 [30]	nodule size, density (solid/GGO/part-solid), growth rate, spiculation	1 (negative) to 4X (suspicious)
breast	BI-RADS / ACR IF [38, 4]	mass density, margins, enhancement, calcification on CT	benign (ignore) to suspicious (tissue sampling)
esophagus	NCCN + TNM [3]	wall thickness, luminal narrowing, fat plane invasion	T/N staging criteria
liver	LI-RADS [26]	arterial hyperenhancement, washout, enhancing capsule, threshold growth	LR-1 (benign) to LR-5 (definite HCC)
gallbladder	ACR IF [97]	wall thickness, mucosal enhancement, polyp size	thin-wall (benign) to thick/enhancing (surgery)
stomach	NCCN + Borrmann [2]	wall thickness, enhancement pattern, serosal invasion, morphology	Borrmann I–IV + T staging
pancreas	ACR IF + Fukuoka [81, 104]	duct dilation, mural nodules, solid component, cyst size	low/high-risk stigmata
spleen	ACR IF [48]	lesion size, homogeneity, multiplicity, enhancement	$<$1 cm (benign) to heterogeneous/growing (workup)
duodenum	NCCN [18]	mass size, obstruction, vascular encasement	resectability criteria
colon	C-RADS [92, 117]	polyp size, morphology, location, number	C0–C4 categories
kidney	Bosniak v2019 [101]	septa, wall thickness, enhancement, calcification	I/II (benign) to IV (surgical)
adrenal	ACR IF [80, 50]	size, HU on unenhanced CT, washout characteristics	$\leq$1 cm (ignore) to $>$4 cm (surgery)
bladder	VI-RADS [90]	muscularis integrity, stalk morphology, signal on DWI	VI-RADS 1–5
prostate	PI-RADS v2.1 [106]	T2 signal, DWI restriction, DCE, size, location by zone	PI-RADS 1–5
uterus	FIGO [5]	endometrial thickness, myometrial invasion depth, cervical extension	FIGO stage I–IV
ovary	O-RADS / ACR IF [7, 9]	cyst size, wall/septa thickness, solid component, enhancement	O-RADS 1 (normal) to 5 (high risk)
lymph node	Lugano [27]	short-axis diameter, morphology, enhancement, FDG avidity (PET)	measurable ($>$1.5 cm) vs. non-measurable
bone	WHO / RECIST [85, 39]	lytic/sclerotic morphology, cortical destruction, soft-tissue component	benign features to aggressive (biopsy)

11.5%. Integrative and ambiguous cases together cover 75.6% of the dataset. This confirms that the majority of cancer screening requires reasoning beyond single-scan perception.

Quality control summary. Automated QC (Appendix B) flagged 905 issues: timeline oscillations (304, concentrated in patients with $\geq$12 scans), uncertain primary cancer assignment (422), and malignancy flag inconsistencies (179).

6 Training Vision-Language Models with RadThinking

RadThinking supplies the supervised and RL stages of the modern VLM training stack. Modern open-source VLMs follow a four-stage recipe [107, 11, 37, 102, 32, 72, 78, 112, 65, 122, 1, 86]: unimodal pretraining, vision-language alignment, supervised fine-tuning (SFT), and preference or reinforcement-learning post-training. A second post-training axis is rule-based RL with verifiable rewards [43, 87, 31, 84, 54, 82, 91, 35, 77], mostly limited so far to math and perception [89, 61, 56, 21, 111]. RadThinking targets the SFT and RL stages with grounded multimodal CoT and pathology-confirmed rewards. We do not claim experimental results.

Curriculum from foundation to compositional VQA. The three tiers (§3) form a natural SFT curriculum [17, 34]. Foundation samples are (image, atomic evidence question, short answer) triples that train visual skills [52, 99]. Compositional samples are (image, hard question, four-step evidence chain, answer) tuples that train chain-of-thought in the spirit of LLaVA-CoT [114] and Visual-CoT [99]. Box 2 shows examples for one patient. Two properties distinguish these chains from text-distilled CoT: they are grounded in voxel masks and clinical reporting standards, and their

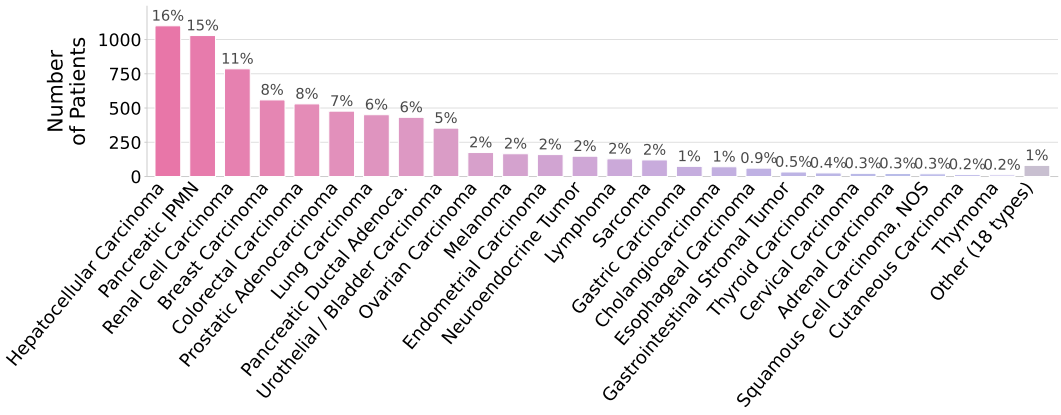

Figure 2: **Primary cancer type distribution across cancer-positive patients in RadThinking.** Clinically equivalent subtypes are grouped together. For example, colon, rectal, and colorectal NOS are grouped into Colorectal Carcinoma. Bladder and urothelial cancers are grouped into a single category. The result is 43 distinct cancer groups with a pronounced long-tail distribution. Healthy controls (2,077) are omitted for clarity.

conclusions are verified by tissue diagnosis. The 20,362 scans yield several hundred thousand SFT pairs across the three tiers.

Verifiable rewards for RL. The chain exposes four rule-based reward axes for GRPO-style RL [43, 89, 61]: pathology match against $\mathcal{D}_t$, organ-level malignancy and metastasis flags, organ-specific risk category $\kappa_{\mathcal{S}}$ with within-± 1 partial credit, and temporal change labels. A format reward enforces the four-step output structure. Details are in Appendix C.1.

Evaluation. We recommend reporting accuracy stratified by VQA tier (foundation, single-step, compositional) and orthogonally by case complexity (§B.6). Together they show *where* a model fails. Per-protocol details are in Appendix C.2.

Position against the open VLM ecosystem. RadThinking fills four gaps that frontier-lab reports leave open. Only Google publishes a medical VLM recipe (Med-Gemini [96], Med-PaLM M [105]) and neither releases data. Most open VLMs use 2D backbones; RadThinking provides 3D CT with voxel grounding. Reasoning-RL works repeatedly flag scarcity of multimodal CoT outside math [54, 82, 91, 35, 77, 84]; RadThinking supplies grounded medical CoT. No other corpus supplies time-ordered scans with grounded change labels.

7 Discussion and Conclusion

RadThinking reframes cancer-screening AI as evidence-based visual question answering. The three tiers form a curriculum from atomic evidence cues to multi-step compositional reasoning. Every compositional VQA carries the chain of foundation VQAs that supplies its evidence. The chains are grounded in voxel-wise tumor masks and in organ-specific clinical reporting standards. The dominance of integrative and ambiguous cases (75.6% of chains) confirms that most cancer screening lies beyond single-scan perception. The chain's quality is anchored in an eight-radiologist validation cohort: 62.2% inter-annotator Dice, 94.6% feature-extraction accuracy, 95.9% temporal-label agreement, 97.1% report-parsing accuracy, and Fleiss' $\kappa = 0.947$ for complexity stratification.

Open question: structured evidence as the substrate for patient-similarity search. A radiologist's workflow does not stop at reading the image. After collecting the imaging evidence, she searches for similar patients in the literature or in institutional memory, and lets their trajectories inform the management of the patient on the table. Literature-side tools such as OpenEvidence [88] now scale this search across millions of published articles, but they cannot read the scan and cannot supply the per-patient cues that determine which patients count as similar. RadThinking supplies those cues. Every foundation VQA is a structured, verifiable evidence field; every compositional chain ends in a pathology-confirmed outcome. This makes RadThinking the natural index for patient-similarity search at population scale, where AI can leverage many more patient trajectories than any one

Box 2: VQA Examples Across the Three Tiers (Patient P000001, Scan 2)

Patient context. Known metastatic breast carcinoma, post-mastectomy. Prior scan: 2017-07-24. Six lesions tracked across organs.

① Foundation VQA (atomic evidence cues, no reasoning)

Q_1: What modality is this image? **A:** CT.
Q_2: Is the liver visible in this scan? **A:** Yes.
Q_3: What is the size of the renal lesion in the left inferior pars? **A:** 12 mm.
Q_4: Was the right gluteal soft-tissue lesion present in the prior scan? **A:** Yes.

② Single-Step Reasoning VQA (one rule)

Q_5: Is the renal lesion at or above 1 cm by Bosniak criteria?
A: Yes (12 mm $\geq$ 10 mm threshold).

Q_6: Has the right gluteal soft-tissue lesion resolved relative to the prior scan?
A: Yes (RESOLVED; not present at the current time point).

③ Compositional VQA (multi-step chain-of-thought)

Q_7: What is the RECIST 1.1 response category for this scan?
A.
[Observations] Five lesions present at both timepoints; renal lesion stable at 12 mm.
[Temporal] One lesion RESOLVED; no new lesions; no growth meeting RECIST progressive disease threshold.
[Context] Post-mastectomy follow-up; baseline 2018-01-24.
[Conclusion] Stable Disease per RECIST 1.1.

Q_8: Synthesize imaging and clinical context to give the most likely diagnosis.
A. Metastatic breast carcinoma with stable disease. Pathology-confirmed.

radiologist remembers. We do not study retrieval in this paper. We provide the substrate the field needs to study it.

Limitations. *(i) Information structure, not narrative.* Chains encode what was observed and how it changed. They do not encode free-form inferential narrative. *(ii) Constructed from existing documentation.* The pipeline parses clinical reports rather than eliciting think-aloud protocols. *(iii) CT only.* The four-step framework generalizes to mammography (BI-RADS) and MRI (PI-RADS) where standards exist. *(iv) Incomplete multimodal data.* Some patients lack prior scans or structured reports; this mirrors clinical reality and provides an evaluation axis for reasoning under missing information. *(v) Limited follow-up window.* The healthy cohort requires only >1-year cancer-free follow-up. *(vi) Point-estimate validation.* Quality metrics are reported without significance testing; downstream auditing [79] is encouraged.

Conclusion. RadThinking supplies the medical VQA training data missing from current open VLM stacks: foundation-tier SFT for atomic visual skills, compositional-tier CoT SFT, and verifiable rewards for RL recipes such as DeepSeek-R1 [43, 89, 61]. It complements parallel work on tumor synthesis [22, 115], continual learning for medical data [29, 120], and partial-label assembly [59].

Acknowledgments and Disclosure of Funding

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the National Institutes of Health (NIH) under Award Number R01EB037669. We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC. We thank Yucheng Tang, Ho Hin Lee, Sucheng Ren, Junfei Xiao, Yuyin Zhou, and Jieneng Chen for their constructive suggestions at several stages of the project. We thank Jaimie Patterson for writing a news article about this project. Paper content is covered by patents pending.

References

[1] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. De Monicault, S. Garg, T. Gervet, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024.

[2] J. A. Ajani, T. A. D'Amico, K. Almhanna, D. J. Bentrem, J. Chao, P. Das, C. S. Denlinger, P. Fanta, F. Farjah, C. S. Fuchs, et al. Gastric cancer, version 3.2016, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 14(10): 1286–1312, 2016.

[3] J. A. Ajani, T. A. D'Amico, D. J. Bentrem, J. Chao, C. Corvera, P. Das, C. S. Denlinger, P. C. Enzinger, P. Fanta, F. Farjah, et al. Esophageal and esophagogastric junction cancers, version 2.2019, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 17(7):855–883, 2019.

[4] S. Al-Katib, G. Gupta, A. Brudvik, S. Ries, J. Krauss, and M. Farah. A practical guide to managing ct findings in the breast. *Clinical imaging*, 60(2):274–282, 2020.

[5] F. Amant, M. R. Mirza, M. Koskas, and C. L. Creutzberg. Cancer of the corpus uteri. *International Journal of Gynecology & Obstetrics*, 143:37–50, 2018.

[6] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] R. F. Andreotti, D. Timmerman, L. M. Strachowski, W. Froyman, B. R. Benacerraf, G. L. Bennett, T. Bourne, D. L. Brown, B. G. Coleman, M. C. Frates, et al. O-rads us risk stratification and management system: a consensus guideline from the acr ovarian-adnexal reporting and data system committee. *Radiology*, 294(1):168–185, 2020.

[8] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

[9] M. Atri, A. Alabousi, C. Reinhold, E. A. Akin, C. B. Benson, P. R. Bhosale, S. K. Kang, Y. Lakhman, R. Nicola, P. V. Pandharipande, et al. Acr appropriateness criteria® clinically suspected adnexal mass, no acute symptoms. *Journal of the American College of Radiology*, 16(5):S77–S93, 2019.

[10] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, and B. Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[11] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[12] P. R. Bassi, W. Li, Y. Tang, F. Isensee, Z. Wang, J. Chen, Y.-C. Chou, Y. Kirchhoff, M. Rokuss, Z. Huang, J. Ye, J. He, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. H. Maier-Hein, P. Jaeger, Y. Ye, Y. Xie, J. Zhang, Z. Chen, Y. Xia, Z. Xing, L. Zhu, Y. Sadegheih, A. Bozorgpour, P. Kumari, R. Azad, D. Merhof, P. Shi, T. Ma, Y. Du, F. Bai, T. Huang, B. Zhao, H. Wang, X. Li, H. Gu, H. Dong, J. Yang, M. A. Mazurowski, S. Gupta, L. Wu, J. Zhuang, H. Chen, H. Roth, D. Xu, M. B. Blaschko, S. Decherchi, A. Cavalli, A. L. Yuille, and Z. Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 37:15184–15201, 2024. URL https://github.com/MrGiovanni/Touchstone.

[13] P. R. Bassi, W. Li, J. Chen, Z. Zhu, T. Lin, S. Decherchi, A. Cavalli, K. Wang, Y. Yang, A. L. Yuille, and Z. Zhou. Learning segmentation from radiology reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–315. Springer, 2025. URL https://github.com/MrGiovanni/R-Super.

[14] P. R. Bassi, Q. Wu, W. Li, S. Decherchi, A. Cavalli, A. Yuille, and Z. Zhou. Label critic: Design data before models. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025. URL https://github.com/PedroRASB/LabelCritic.

[15] P. R. Bassi, M. C. Yavuz, I. E. Hamamci, S. Er, X. Chen, W. Li, B. Menze, S. Decherchi, A. Cavalli, K. Wang, Y. Yang, A. Yuille, and Z. Zhou. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23720–23730, 2025. URL https://github.com/MrGiovanni/RadGPT.

[16] P. R. Bassi, X. Zhou, W. Li, S. Płotka, J. Chen, Q. Chen, Z. Zhu, J. Prządo, I. E. Hamamci, S. Er, X. Chen, M. C. Yavuz, Y.-C. Chou, T. Lin, K. Wang, Y. Tang, J. B. Cwikla, S. Decherchi, A. Cavalli, Y. Yang, A. L. Yuille, and Z. Zhou. Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks. *arXiv preprint arXiv:2510.14803*, 2025. URL https://github.com/MrGiovanni/R-Super.

[17] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, 2009.

[18] A. B. Benson, A. P. Venook, M. M. Al-Hawary, M. A. Arain, Y.-J. Chen, K. K. Ciombor, S. A. Cohen, H. S. Cooper, D. A. Deming, I. Garrido-Laguna, et al. Small bowel adenocarcinoma, version 1.2020, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 17(9):1109–1133, 2019.

[19] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

[20] L. Blankemeier, J. P. Cohen, A. Kumar, D. Van Veen, S. J. S. Gardezi, M. Paschali, Z. Chen, J.-B. Delbrouck, E. Reis, C. Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pages rs–3, 2024.

[21] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

[22] Q. Chen, X. Chen, H. Song, Z. Xiong, A. Yuille, C. Wei, and Z. Zhou. Towards generalizable tumor synthesis. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11147–11158, 2024. URL https://github.com/MrGiovanni/DiffTumor.

[23] Q. Chen, X. Zhou, C. Liu, H. Chen, W. Li, Z. Jiang, Z. Huang, Y. Zhao, D. Yu, J. He, Y. Zheng, L. Shao, A. Yuille, and Z. Zhou. Scaling tumor segmentation: Best lessons from real and synthetic data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 24001–24013, 2025. URL https://github.com/BodyMaps/AbdomenAtlas2.0.

[24] Y. Chen, W. Xiao, P. R. Bassi, X. Zhou, S. Er, I. E. Hamamci, Z. Zhou, and A. 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. URL https://github.com/Schuture/DeepTumorVQA.

[25] Y. Chen, Z. Zhou, W. Li, and A. Yuille. Large-scale label quality assessment for medical segmentation via a vision-language judge and synthetic data. *arXiv preprint arXiv:2601.14406*, 2026.

[26] V. Chernyak, K. J. Fowler, A. Kamaya, A. Z. Kielar, K. M. Elsayes, M. R. Bashir, Y. Kono, R. K. Do, D. G. Mitchell, A. G. Singal, et al. Liver imaging reporting and data system (li-rads) version 2018: imaging of hepatocellular carcinoma in at-risk patients. *Radiology*, 289(3): 816–830, 2018.

[27] B. D. Cheson, R. I. Fisher, S. F. Barrington, F. Cavalli, L. H. Schwartz, E. Zucca, and T. A. Lister. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: the lugano classification. *Journal of clinical oncology*, 32(27): 3059–3067, 2014.

[28] Y.-C. Chou, B. Li, D.-P. Fan, A. Yuille, and Z. Zhou. Acquiring weak annotations for tumor localization in temporal and volumetric data. *Machine Intelligence Research*, pages 1–13, 2024. URL https://github.com/johnson111788/Drag-Drop.

[29] Y.-C. Chou, Z. Zhou, and A. Yuille. Embracing massive medical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–35. Springer, 2024. URL https://github.com/MrGiovanni/OnlineLearning.

[30] J. Christensen, A. E. Prosper, C. C. Wu, J. Chung, E. Lee, B. Elicker, A. R. Hunsaker, M. Petranovic, K. L. Sandler, B. Stiles, et al. Acr lung-rads v2022: assessment categories and management recommendations. *Journal of the American College of Radiology*, 21(3): 473–488, 2024.

[31] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[32] W. Dai, N. Lee, B. Wang, Z. Yang, Z. Liu, J. Barker, T. Rintamaki, M. Shoeybi, B. Catanzaro, and W. Ping. NVLM: Open frontier-class multimodal LLMs. *arXiv preprint arXiv:2409.11402*, 2024.

[33] M. de Grauw, E. T. Scholten, E. J. Smit, M. J. Rutten, M. Prokop, B. van Ginneken, and A. Hering. The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. *Medical image analysis*, 102:103525, 2025.

[34] H. Deng, H. Zhang, M. Ou, Z. Li, J. Liu, H. Wang, and T.-Y. Lin. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.

[35] Y. Deng, H. Bansal, F. Yin, N. Peng, W. Wang, and K.-W. Chang. OpenVLThinker: Complex vision-language reasoning via iterative SFT-RL cycles. *arXiv preprint arXiv:2503.17352*, 2025.

[36] A. Diaz-Pinto, S. Alle, V. Nath, Y. Tang, A. Ihsani, M. Asad, F. Pérez-García, P. Mehta, W. Li, M. Flores, et al. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. *Medical Image Analysis*, 95:103207, 2024.

[37] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[38] C. D'Orsi, L. Bassett, and S. Feig. Breast imaging reporting and data system (bi-rads). *Oxford University Press, New York*, 2018.

[39] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. 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.

[40] R. W. Filice and C. E. Kahn. Ontologies in the new computational age of radiology: RadLex for semantics and interoperability in imaging workflows. *RadioGraphics*, 43(1), 2022.

[41] X. Gai, J. Liu, Y. Li, Z. Meng, J. Wu, and Z. Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *arXiv preprint arXiv:2506.11147*, 2025.

[42] H. L. Gamage, L. Wijerathne, Y. Wickramasinghe, M. Riegler, and P. Halvorsen. Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy. *arXiv preprint arXiv:2506.09958*, 2025.

[43] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[44] T. Gupta and A. Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[45] I. E. Hamamci, S. Er, and B. 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.

[46] I. E. Hamamci, S. Er, C. Wang, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, O. F. Durugol, B. Hou, S. Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19, 2026.

[47] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[48] M. T. Heller, M. Harisinghani, J. D. Neitlich, P. Yeghiayan, and L. L. Berland. Managing incidental findings on abdominal and pelvic ct and mri, part 3: white paper of the acr incidental findings committee ii on splenic and nodal findings. *Journal of the American College of Radiology*, 10(11):833–839, 2013.

[49] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[50] B. R. Herts, S. G. Silverman, N. M. Hindman, R. G. Uzzo, R. P. Hartman, G. M. Israel, D. A. Baumgarten, L. L. Berland, and P. V. Pandharipande. Management of the incidental renal mass on ct: a white paper of the acr incidental findings committee. *Journal of the American College of Radiology*, 15(2):264–273, 2018.

[51] J. K. Hoang, J. E. Langer, W. D. Middleton, C. C. Wu, L. W. Hammers, J. J. Cronan, F. N. Tessler, E. G. Grant, and L. L. Berland. Managing incidental thyroid nodules detected on imaging: white paper of the acr incidental thyroid findings committee. *Journal of the American College of Radiology*, 12(2):143–150, 2015.

[52] H. Hong, H. Kim, H. Lee, S. Choi, et al. Decomposing complex visual comprehension into atomic visual skills for vision language models. *arXiv preprint arXiv:2505.20021*, 2025.

[53] Y. Hu, T. Li, Q. Lu, W. Shao, J. He, Y. Qiao, and P. Luo. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[54] W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Y. Hu, and S. Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[55] D. A. Hudson and C. D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[56] J. Jeong, S. Yun, H. Lim, and J. Kang. Med-PRM: Medical reasoning models with stepwise, guideline-verified process rewards. *arXiv preprint arXiv:2506.11474*, 2025.

[57] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[58] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[59] M. Kang, B. Li, Z. Zhu, Y. Lu, E. K. Fishman, A. Yuille, and Z. Zhou. Label-assemble: Leveraging multiple datasets with partial labels. In *IEEE International Symposium on Biomedical Imaging*, pages 1–5. IEEE, 2023. URL https://github.com/MrGiovanni/LabelAssemble.

[60] T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, and A. Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *International Conference on Learning Representations (ICLR)*, 2023.

[61] Y. Lai, J. Zhong, M. Li, S. Zhao, and X. Yang. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

[62] C. P. Langlotz. RadLex: A new method for indexing online educational materials. *RadioGraphics*, 26(6):1595–1597, 2006.

[63] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

[64] B. Li, Y.-C. Chou, S. Sun, H. Qiao, A. Yuille, and Z. Zhou. Early detection and localization of pancreatic cancer by label-free tumor synthesis. *MICCAI Workshop on Big Task Small Data, 1001-AI*, 2023. URL https://github.com/MrGiovanni/SyntheticTumors.

[65] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[66] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[67] W. Li, C. Qu, X. Chen, P. R. Bassi, Y. Shi, Y. Lai, Q. Yu, H. Xue, Y. Chen, X. Lin, Y. Tang, Y. Cao, H. Han, Z. Zhang, J. Liu, T. Zhang, Y. Ma, J. Wang, G. Zhang, A. Yuille, and Z. Zhou. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, page 103285, 2024. URL https://github.com/MrGiovanni/AbdomenAtlas.

[68] W. Li, A. Yuille, and Z. Zhou. How well do supervised models transfer to 3d image segmentation? In *International Conference on Learning Representations*, 2024. URL https://github.com/MrGiovanni/SuPreM.

[69] W. Li, P. R. Bassi, T. Lin, Y.-C. Chou, X. Zhou, Y. Tang, F. Isensee, K. Wang, Q. Chen, X. Xu, J. Ye, Z. Zhu, S. Decherchi, A. Cavalli, A. L. Yuille, and Z. Zhou. Scalemai: Accelerating the development of trusted datasets and ai models. *arXiv preprint arXiv:2501.03410*, 2025. URL https://github.com/MrGiovanni/ScaleMAI.

[70] W. Li, X. Zhou, Q. Chen, T. Lin, P. R. Bassi, X. Chen, C. Ye, Z. Zhu, K. Ding, H. Li, K. Wang, Y. Yang, Y. Tang, D. Xu, A. L. Yuille, and Z. Zhou. Pants: The pancreatic tumor segmentation dataset. In *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. URL https://github.com/MrGiovanni/PanTS.

[71] W. Li, P. R. A. S. Bassi, L. Wu, X. Zhou, Y. Zhao, Q. Chen, S. Plotka, T. Lin, Z. Zhu, M. Martin, J. Caskey, S. Jiang, X. Chen, J. B. Ćwikła, A. Sankowski, Y. Wu, S. Decherchi, A. Cavalli, C. Lall, C. Tomasetti, Y. Guo, X. Yu, Y. Cai, H. Qiao, J. Bao, C. Hu, X. Wang, A. Sitek, K. Ding, H. Li, M. Wang, D. Yu, G. Zhang, Y. Yang, K. Wang, A. L. Yuille, and Z. Zhou. Early and prediagnostic detection of pancreatic cancer from computed tomography. *arXiv preprint arXiv:2601.22134*, 2026. URL https://github.com/BodyMaps/ePAI.

[72] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han. VILA: On pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[73] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2021.

[74] J. Liu, A. Yuille, Y. Tang, and Z. Zhou. Clip-driven universal model for partially labeled organ and pan-cancer segmentation. In *MICCAI 2023 FLARE Challenge*, 2023. URL https://github.com/ljwztc/CLIP-Driven-Universal-Model.

[75] J. Liu, Y. Zhang, K. Wang, M. C. Yavuz, X. Chen, Y. Yuan, H. Li, Y. Yang, A. Yuille, Y. Tang, and Z. Zhou. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. *Medical Image Analysis*, page 103226, 2024. URL https://github.com/ljwztc/CLIP-Driven-Universal-Model.

[76] R. Liu, I. Q. Mohiuddin, A. J. Schoeffler, K. Renduchintala, A. Nayak, P. L. Vemu, S. C. Vedak, K. C. Black, J. L. Havlik, I. Ogunmola, S. P. Ma, R. Dhatt, and J. H. Chen. PhysicianBench: Evaluating LLM agents in real-world EHR environments. *arXiv preprint arXiv:2605.02240*, 2026.

[77] Z. Liu, Z. Sun, Y. Zang, X. Dong, Y. Cao, H. Duan, D. Lin, and J. Wang. Visual-RFT: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[78] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, et al. DeepSeek-VL: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[79] A. Lubonja, P. R. Bassi, W. Li, H. Qiao, R. Burns, A. L. Yuille, and Z. Zhou. Auditing significance, metric choice, and demographic fairness in medical ai challenges. *arXiv preprint arXiv:2512.19091*, 2025. URL https://github.com/ariellubonja/RankInsight.

[80] W. W. Mayo-Smith, J. H. Song, G. L. Boland, I. R. Francis, G. M. Israel, P. J. Mazzaglia, L. L. Berland, and P. V. Pandharipande. Management of incidental adrenal masses: a white paper of the acr incidental findings committee. *Journal of the American College of Radiology*, 14(8): 1038–1044, 2017.

[81] A. J. Megibow, M. E. Baker, D. E. Morgan, I. R. Kamel, D. V. Sahani, E. Newman, W. R. Brugge, L. L. Berland, and P. V. Pandharipande. Management of incidental pancreatic cysts: a white paper of the acr incidental findings committee. *Journal of the American College of Radiology*, 14(7):911–923, 2017.

[82] F. Meng, L. Du, Z. Liu, Z. Zhou, Q. Lu, D. Fu, B. Shi, W. Wang, J. He, K. Zhang, et al. MM-Eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[83] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993, 2015.

[84] Microsoft Research. Phi-4-Reasoning technical report. Technical report, Microsoft, 2025. https://www.microsoft.com/en-us/research/publication/phi-4-reasoning/.

[85] W. C. of Tumours Editorial Board et al. *Soft tissue and bone tumours*, volume 3. World Health Organization, 2020.

[86] OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[87] OpenAI. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[88] OpenEvidence. OpenEvidence: An AI copilot for evidence-based clinical decision making at the point of care. https://www.openevidence.com, 2026. Retrieval-augmented generation over peer-reviewed medical literature (e.g., NEJM, JAMA, Cochrane). Accessed May 2026.

[89] J. Pan, C. Liu, J. Wu, F. Liu, J. Zhu, H. B. Li, C. Chen, C. Ouyang, and D. Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

[90] V. Panebianco, Y. Narumi, E. Altun, B. H. Bochner, J. A. Efstathiou, S. Hafeez, R. Huddart, S. Kennish, S. Lerner, R. Montironi, et al. Multiparametric magnetic resonance imaging for bladder cancer: development of vi-rads (vesical imaging-reporting and data system). *European urology*, 74(3):294–306, 2018.

[91] Y. Peng, G. Zhang, M. Zhang, Z. You, J. Liu, Q. Zhu, K. Yang, X. Xu, X. Geng, and X. Yang. LMM-R1: Empowering 3B LMMs with strong reasoning abilities through two-stage rule-based RL. *arXiv preprint arXiv:2503.07536*, 2025.

[92] P. J. Pickhardt, J. R. Choi, I. Hwang, J. A. Butler, M. L. Puckett, H. A. Hildebrandt, R. K. Wong, P. A. Nugent, P. A. Mysliwiec, and W. R. Schindler. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. *New England Journal of Medicine*, 349(23):2191–2200, 2003.

[93] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

[94] C. Qu, T. Zhang, H. Qiao, J. Liu, Y. Tang, A. Yuille, and Z. Zhou. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 21, 2023. URL https://github.com/MrGiovanni/AbdomenAtlas.

[95] Radiological Society of North America (RSNA). RadLex radiology lexicon. https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon. Controlled radiology terminology with over 75,000 terms; distinguishes imaging observations and findings from disorders.

[96] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[97] S. Sebastian, C. Araujo, J. D. Neitlich, and L. L. Berland. Managing incidental findings on abdominal and pelvic ct and mri, part 4: white paper of the acr incidental findings committee ii on gallbladder and biliary findings. *Journal of the American College of Radiology*, 10(12): 953–956, 2013.

[98] R. R. Selvaraju, P. Tendulkar, D. Parikh, E. Horvitz, M. T. Ribeiro, B. Nushi, and E. Kamar. SQuINTing at VQA models: Introspecting VQA models with sub-questions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[99] H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 2024.

[100] S. Sharma, J. Long, G. Shih, S. Eid, C. Bluethgen, F. L. Jacobson, E. B. Tsai, A. M. Alaa, C. P. Langlotz, and Global Radiology Consortium. CheXthought: A global multimodal dataset of clinical chain-of-thought reasoning and visual attention for chest X-ray interpretation. *arXiv preprint arXiv:2604.26288*, 2026.

[101] S. G. Silverman, I. Pedrosa, J. H. Ellis, N. M. Hindman, N. Schieda, A. D. Smith, E. M. Remer, A. B. Shinagare, N. E. Curci, S. S. Raman, et al. Bosniak classification of cystic renal masses, version 2019: an update proposal and needs assessment. *Radiology*, 292(2):475–488, 2019.

[102] A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, A. Gritsenko, M. Minderer, A. Sherbondy, S. Long, et al. PaliGemma 2: A family of versatile VLMs for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

[103] D. Surís, S. Menon, and C. Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[104] M. Tanaka, C. Fernández-del Castillo, T. Kamisawa, J. Y. Jang, P. Levy, T. Ohtsuka, R. Salvia, Y. Shimizu, M. Tada, and C. L. Wolfgang. Revisions of international consensus fukuoka guidelines for the management of ipmn of the pancreas. *Pancreatology*, 17(5):738–753, 2017.

[105] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al. Towards generalist biomedical AI. *New England Journal of Medicine AI*, 2024. arXiv:2307.14334.

[106] B. Turkbey, A. B. Rosenkrantz, M. A. Haider, A. R. Padhani, G. Villeirs, K. J. Macura, C. M. Tempany, P. L. Choyke, F. Cornud, D. J. Margolis, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 76(3):340–351, 2019.

[107] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[108] S. Wang, Y. Liu, J. Yang, Y. Wang, H. Lin, B. Wang, Y. Liu, and Y. Wang. MedFrameQA: A multi-image medical VQA benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025.

[109] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[110] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 2025.

[111] J. Wu, W. Deng, X. Li, S. Liu, T. Mi, Y. Peng, Z. Xu, Y. Liu, H. Cho, C.-I. Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

[112] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[113] Y. Xia, Q. Yu, L. Chu, S. Kawamoto, S. Park, F. Liu, J. Chen, Z. Zhu, B. Li, Z. Zhou, A. L. Yuille, E. K. Fishman, and R. H. Hruban. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022.

[114] G. Xu, P. Jin, H. Li, Y. Song, L. Sun, and L. Yuan. LLaVA-CoT: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

[115] Y. Yang, Z.-Y. Wang, Q. Liu, S. Sun, K. Wang, R. Chellappa, Z. Zhou, A. Yuille, L. Zhu, Y.-D. Zhang, and J. Chen. Medical world model: Generative simulation of tumor evolution for treatment planning. *arXiv preprint arXiv:2506.02327*, 2025. URL https://github.com/scott-yjyang/MeWM.

[116] J. Yun, J. Sohn, J. Park, H. Kim, X. Tang, D. Shao, Y. H. Koo, K. Minhyeok, Q. Chen, M. Gerstein, et al. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16565–16582, 2025.

[117] M. E. Zalis, M. A. Barish, J. R. Choi, A. H. Dachman, H. M. Fenlon, J. T. Ferrucci, S. N. Glick, A. Laghi, M. Macari, E. G. McFarland, et al. Ct colonography reporting and data system: a consensus proposal. *Radiology*, 236(1):3–9, 2005.

[118] T. Zhang, X. Chen, C. Qu, A. Yuille, and Z. Zhou. Leveraging ai predicted and expert revised annotations in interactive segmentation: Continual tuning or full training? In *IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024. URL https://github.com/MrGiovanni/ContinualLearning.

[119] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[120] Y. Zhang, X. Li, H. Chen, A. L. Yuille, Y. Liu, and Z. Zhou. Continual learning for abdominal multi-organ and tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2023. URL https://github.com/MrGiovanni/ContinualLearning.

[121] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[122] J. Zhu, Z. Chen, W. Wang, H. Tian, L. Lu, B. Li, Y. Cui, Z. Cai, E. Zhao, S. Wang, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A JSON Schema of the Released Reasoning Chains

This appendix documents the complete schema of the per-patient JSON file. Each patient is one JSON record. Each record has patient-level fields and a list of per-scan reasoning traces. We list every field with its type and value range. We then show a worked example for one scan.

A.1 Patient-Level Fields

field	type	description / value range
`patient_id`	string	anonymized identifier (e.g., `P000001`)
`primary_cancer`	object	resolved primary cancer record (sub-fields below)
`primary_cancer`	string	cancer type (e.g., `breast cancer`, `hepatocellular carcinoma`)
`confidence`	string	`high`, `medium`, or `low`; defined in App. B
`source`	list[string]	resolver sources that voted (`icd10`, `tumor_keyword`, `tumor_flag`, `report_nlp`, `report_nlp_definitive`)
`all_candidates`	object	mapping from candidate name to weighted vote score
`metastasis_sites`	list[string]	organs flagged as metastatic spread sites
`has_metastatic_disease`	bool	true if any metastasis site is flagged
`clinical_history`	list[string]	curated short statements about prior diagnoses, surgeries, oncological status
`num_scans`	int	number of CT scans in the longitudinal sequence
`date_range`	object	`first` and `last` ISO dates of the sequence
`reasoning_traces`	list[object]	one reasoning chain per scan (§A.2)

A.2 Per-Scan Trace Fields

Each entry in `reasoning_traces` has the following five top-level fields plus the complexity label.

`metadata`. `patient_id`, `scan_id`, `accession` (string), `scan_date` (ISO), `scan_index` and `total_scans` (int), `sex` and `age` (string, may be empty), `malignancy` and `metastasis` (string flags: `yes`, `no`, `u`).

`step1_observations`. A list. Each finding has: `finding_id` (string); `raw_organ` and `canonical_organ` (string, one of the 19 targets or an ancillary category); `location` and `standardized_location` (string); `type` (free text describing tumor type); `type_certainty` (`certain`, `high`, `low`); `size_mm` (numeric, or `multiple`, or `U` for unknown); `attenuation` and `standardized_attenuation` (string); `malignancy` and `metastasis` (`yes`, `no`, `u`); `clinical_standard` (object with `name` and `reference`).

`step2_temporal`. `status` is `no_prior_available` or `temporal_comparison_available`. When prior is available: `prior_scan_date` (ISO), `interval_days` and `interval_months` (numeric), `n_matched`, `n_new`, `n_resolved` (int), and a list `changes`. Each change entry has: `organ`, `location`, `tumor_type`, `matched_with_prior` (bool), and `change` (one of `NEW`, `GROWING`, `STABLE`, `SHRINKING`, `RESOLVED`, `PRESENT_BOTH`). When sizes are available, `size_current_mm`, `size_prior_mm`, and `volume_ratio` are populated; otherwise a `size_note` explains the absence.

`step3_clinical_context`. `report_parsed` is an object with `findings` (list of strings), `impression` (string), `recommendation` (string), and `parse_method` (`rule_based` or `llm`). `recist_assessment` is one of `Stable Disease`, `Partial Response`, `Complete Response`, `Progressive Disease`, or null. `risk_category` is the organ-specific category from κ_S (e.g., `LR-5`, `Bosniak IIF`, `PI-RADS 4`, `RECIST 1.1: Stable Disease`, or `not explicitly stated`). `clinical_variables` stores `age`, `sex`, `contrast`, and `clinical_history`. `raw_report` is the de-identified original text.

`step4_conclusion`. `primary_cancer`, `primary_cancer_confidence`, `primary_cancer_source` (mirrors patient-level resolution at scan time). `has_metastatic_disease` (bool) and `metastasis_sites` (list). `overall_malignancy` and `overall_metastasis` (`yes`, `no`, `u`). `icd10_code` and `icd10_organ` (string, may be empty). `organ_level_diagnosis` is an object mapping each organ that appears in this scan to a sub-object with `malignancy`, `metastasis`, and `primary_tumor` fields.

`reasoning_complexity`. One of `PERCEPTUAL`, `TEMPORAL`, `INTEGRATIVE`, `AMBIGUOUS` (§B.6).

A.3 Worked Example

The following abbreviated JSON shows Patient P000001, Scan 2 (the same scan used in Box 2). Long lists are truncated for readability; the released file contains all findings.

```
{
  "patient_id": "P000001",
  "primary_cancer": {
    "primary_cancer": "breast cancer",
    "confidence": "high",
    "source": ["report_nlp", "report_nlp_definitive"],
    "has_metastatic_disease": true },
  "clinical_history": ["Known breast carcinoma", "Status post mastectomy", ...],
  "num_scans": 5,
  "date_range": {"first": "2017-07-24", "last": "2019-04-23"},
  "reasoning_traces": [
    {
      "metadata": {"scan_id": "S000002", "scan_date": "2017-11-28",
                   "scan_index": 2, ...},
      "step1_observations": [
        {"finding_id": "tumor 1", "canonical_organ": "chest_wall",
          "type": "metastasis", "malignancy": "yes",
          "clinical_standard": {"name": "RECIST 1.1 (general)"}}, ... ],
      "step2_temporal": {
        "status": "temporal_comparison_available",
        "interval_months": 4.2,
        "n_matched": 6, "n_new": 0, "n_resolved": 1,
        "changes": [
          {"organ": "kidney", "change": "STABLE", "volume_ratio": 1.0}, ...]},
      "step3_clinical_context": {
        "report_parsed": {
          "findings": [...],
          "impression": "Stable disease per RECIST 1.1...", ...},
        "recist_assessment": "Stable Disease",
        "risk_category": "RECIST 1.1: Stable Disease"},
      "step4_conclusion": {
        "primary_cancer": "breast cancer",
        "has_metastatic_disease": true,
        "organ_level_diagnosis": {
          "liver": {"malignancy": "yes", "metastasis": "yes"}, ...}},
      "reasoning_complexity": "INTEGRATIVE"
    }, ... ]
}
```

The full released file preserves all findings, complete report text, and every clinical variable available at imaging time.

B Reasoning Chain Construction Pipeline

This appendix documents the formal definition and validation of each chain step (§B.1–§B.4), the construction algorithm (Algorithm 1, §B.5), the complexity stratification rules (§B.6), and the data cleaning and quality control pipeline (§B.7–§B.9).

B.1 Step 1: Imaging Observations

From the voxel-wise tumor mask M_t, the CT volume I_t, and the governing clinical standard $\mathcal{S}$, we extract a structured observation for each finding. When $M_t \neq \emptyset$, the observation is a tuple

$$\mathcal{O}_t = \big(\,\mathrm{loc}(M_t),\;\; \mathrm{size}(M_t),\;\; \mathcal{F}_{\mathcal{S}}(R_t),\;\; \mathrm{morph}(I_t, M_t)\,\big). \tag{2}$$

$\mathrm{loc}(\cdot)$ maps the mask centroid to an anatomical region via an organ atlas (e.g., "liver, segment VI"). $\mathrm{size}(\cdot)$ returns bounding-box axes and volume in cm^3 and flags standard-relevant thresholds. $\mathcal{F}_{\mathcal{S}}(R_t)$ are standard-specific descriptive features extracted from the findings section of R_t via LLM-based parsing aligned to $\mathcal{S}$'s vocabulary. $\mathrm{morph}(\cdot)$ captures HU statistics, sphericity, surface irregularity, and calcification. When $M_t = \emptyset$, $\mathcal{O}_t$ records "no suspicious findings."

Validation. On the shared validation cohort (n=200), two independent reviewers per case yield 62.2% inter-annotator Dice. For $\mathcal{F}_{\mathcal{S}}$, eight board-certified radiologists verified each extracted feature against the source report. Feature-level accuracy was 94.6%. Errors concentrate in ambiguous modifiers (e.g., "mildly heterogeneous" vs. "heterogeneous").

B.2 Step 2: Temporal Comparison

For longitudinal scans, we co-register matched lesion pairs $(M_{t_{k-1}},\, M_{t_k})$ and compute the volume ratio $r = \mathrm{vol}(M_{t_k}) \,/\, \mathrm{vol}(M_{t_{k-1}})$ to assign a temporal label:

$$\Delta_{t_k} = \begin{cases} \textsc{new} & M_{t_{k-1}} = \emptyset,\; M_{t_k} \neq \emptyset \\ \textsc{growing} & r > 1.2 \\ \textsc{stable} & 0.8 \leq r \leq 1.2 \\ \textsc{shrinking} & r < 0.8 \\ \textsc{resolved} & M_{t_{k-1}} \neq \emptyset,\; M_{t_k} = \emptyset \end{cases} \tag{3}$$

The 20% threshold follows RECIST-inspired volumetric criteria [39]. Three standards add their own temporal criteria. LI-RADS threshold growth is $\geq$50% diameter increase in $\leq$6 months [26]. Lung-RADS growth rate is a new solid component or $\geq$1.5 mm mean diameter increase [30]. RECIST 1.1 progressive disease is $\geq$20% sum-of-diameters increase [39]. When triggered, the chain carries both labels. For cancer-positive patients, the lead time $\ell_t = t_{\mathrm{index}} - t$ is the interval between each pre-diagnosis scan and pathological confirmation. Patients without prior scans receive $\Delta_t = \textsc{no\_prior}$.

Validation. On the validation cohort, eight radiologists reviewed all matched lesion pairs and their temporal labels. Agreement with automatic labels was 95.9%. Lesion matching was correct in 193 of 200 cases (96.5%). Errors involved small lesions ($<$1 cm) in adjacent segments where co-registration ambiguity was unavoidable.

B.3 Step 3: Clinical Context Integration

Step 3 extracts the radiologist's interpretive synthesis (impression, risk stratification, recommendation) plus non-imaging clinical variables. The report R_t and clinical variables C_t are parsed into

$$\mathcal{C}_t = \big(\,F_t,\;\; \mathcal{I}_t,\;\; \mathcal{R}_t,\;\; \kappa_{\mathcal{S}}(\mathcal{I}_t),\;\; C_t\,\big), \tag{4}$$

where $F_t, \mathcal{I}_t, \mathcal{R}_t$ are the extracted findings, impression, and recommendation, and $\kappa_{\mathcal{S}}(\cdot)$ maps the impression to the risk category defined by $\mathcal{S}$. Parsing uses rule-based section detection plus LLM extraction; non-standard reports ($\sim$12%) use few-shot prompting. Risk categories are extracted in two paths. Regex-based extraction of explicitly stated categories (e.g., "LR-4," "Bosniak IIF") succeeds for 17.7% of cancer scans. Rule-based derivation from observation features following each standard's published criteria fills the rest. Together these assign a risk category to 80.3% of cancer scans. The remaining 19.7% are governed by staging systems (FIGO, TNM) where risk-category scoring does not apply.

Validation. On the validation cohort, section-level parsing accuracy was 97.1% (findings 98.9%, impression 96.3%, recommendation 98.2%). Clinical variable linkage to scan timepoints passed 100% temporal-integrity checks (no future-information leakage). Across all cancer scans, text-extracted and feature-derived risk categories agree exactly in 76.8% of cases and within ± 1 category in 88.2%.

B.4 Step 4: Diagnostic Conclusion

The final step anchors the chain to a definitive ground truth:

$$\mathcal{D}_t = \begin{cases} (c,\ h) \text{ from pathology } P & \text{if cancer-positive} \\ \varnothing\ (>\text{1-year follow-up}) & \text{if healthy} \end{cases} \tag{5}$$

where c is the cancer type and h the histological subtype. Step 4 is a ground-truth anchor, not a reasoning step. It closes the chain so each becomes a self-contained evaluation unit.

B.5 Algorithm: Reasoning Chain Construction

Algorithm 1: Reasoning Chain Construction for a Single Scan

Input: CT scan I_t; tumor mask M_t ($\emptyset$ if healthy); prior scans $\{(I_{t_k}, M_{t_k})\}_{k<t}$; radiology report R_t; clinical variables C_t; pathology P; clinical standard $\mathcal{S}$

Output: Structured reasoning chain $\mathcal{T}_t = \langle\, \mathcal{O}_t,\ \Delta_t,\ \mathcal{C}_t,\ \mathcal{D}_t \,\rangle$

① Step 1: Imaging Observations

if $M_t \neq \emptyset$:
Location $\leftarrow$ organ atlas on M_t; Size $\leftarrow$ bbox axes + vol. (cm^3); Morphology $\leftarrow$ sphericity, surface irregularity
Standard features $\mathcal{F}_{\mathcal{S}} \leftarrow$ extract from R_t aligned to $\mathcal{S}$
$\mathcal{O}_t \leftarrow$ StructuredObservation(location, size, morphology, $\mathcal{F}_{\mathcal{S}}$)
else: $\mathcal{O}_t \leftarrow$ "No suspicious findings"

② Step 2: Temporal Comparison

if prior (I_{t_k}, M_{t_k}) exists:
Co-register $M_{t_k} \rightarrow M_t$; volume ratio $r = \text{vol}(M_t)/\text{vol}(M_{t_k})$
$\Delta_t \leftarrow$ NEW / GROWING ($r>1.2$) / STABLE / SHRINKING ($r<0.8$) / RESOLVED
else: $\Delta_t \leftarrow$ "No prior available"

③ Step 3: Clinical Context Integration

Parse $R_t \rightarrow (F_t,\ \mathcal{I}_t,\ \mathcal{R}_t)$ via rule-based detection + LLM
Extract risk category from $\mathcal{I}_t$ using $\mathcal{S}$ (LI-RADS, Bosniak, PI-RADS, ...); clinical context $\leftarrow$ age, sex, history from C_t
$\mathcal{C}_t \leftarrow$ ClinicalContext($F_t,\ \mathcal{I}_t,\ \mathcal{R}_t,\ C_t$, risk category)

④ Step 4: Diagnostic Conclusion

if cancer-positive: $\mathcal{D}_t \leftarrow$ cancer type and subtype from pathology P
else: $\mathcal{D}_t \leftarrow$ "No malignancy, >1-year follow-up"

B.6 Reasoning Complexity Stratification

We stratify each chain into one of four complexity levels using rule-based decisions over dataset metadata.

Perceptual cases satisfy all four conditions. (a) Tumor longest axis >3 cm. (b) HU difference >20 between lesion and parenchyma. (c) No ambiguity flag from annotation adjudication. (d) High-risk category from imaging alone (LR-5, Bosniak IV, PI-RADS 5). A single scan suffices.

Temporal cases have a decisive temporal change label (NEW, GROWING, RESOLVED) or meet a standard-specific temporal criterion such as LI-RADS threshold growth, but do not meet the Perceptual criteria. The finding becomes apparent only through longitudinal comparison.

Integrative cases require synthesis of imaging with clinical context. The standard assigns an intermediate risk category (LR-3, Bosniak IIF, PI-RADS 3). Reaching the conclusion demands integration of report impression, history, or demographics. There is no inter-radiologist discordance.

Ambiguous cases carry the annotation protocol's ambiguity flag. Radiologists reached different clinical conclusions even after applying the standard. Three discordance types are tagged: *boundary* (same diagnosis, different extent), *classification* (different risk category), and *detection* (presence vs. absence). Pathology is the definitive tiebreaker.

Validation. On the validation cohort, eight radiologists independently classified each patient. Inter-rater reliability was Fleiss' $\kappa = 0.947$. Agreement with automatic labels was 93% (186/200). Of 14 disagreements, 9 were Integrative↔Ambiguous, 3 Perceptual↔Temporal, and 2 other.

B.7 Data Cleaning

Constructing chains at scale requires four cleaning operations. (1) *Organ name normalization* maps 191 raw surface forms to 19 canonical targets. (2) *Primary cancer resolution* combines ICD-10 codes, tumor-type keywords, malignancy/metastasis flags, and report NLP through weighted majority voting. (3) *Lesion-level temporal tracking* matches findings across scans by organ, location, and type, with fuzzy organ-family grouping. (4) *Clinical validation* ensures that metastasis sites are never confused with primary cancers. Pipeline architecture and quality control flag statistics follow.

B.8 Pipeline Architecture

The pipeline takes two data sources as input. (1) Per-scan metadata. This includes radiology reports, ICD-10 codes, clinical variables, and RECIST assessments (20,362 rows). (2) Per-tumor metadata. This includes organ labels, tumor types, locations, sizes, and malignancy flags (45,641 rows). The pipeline produces structured 4-step reasoning chains following Algorithm 1.

Organ name normalization. Raw organ labels in the source data contain 191 unique surface forms. Examples include "chest wall," "thoracic wall," and "pectoral region." We must map these to the 19 canonical cancer screening targets plus categorized ancillary organs. We built a two-pass normalization table. The first pass uses exact string matching. The second pass uses substring matching. Together they consolidate all 191 forms into canonical names. We also define organ *families* (e.g., breast $\approx$ chest wall $\approx$ soft tissue) for fuzzy temporal matching.

Multi-source primary cancer resolver. Determining each patient's primary cancer is essential for grounding the clinical context step. It is non-trivial. ICD-10 codes are available for only 27% of scans. Tumor type labels are heterogeneous (722 unique strings). Radiology reports describe findings without always stating the diagnosis explicitly. Our resolver combines four sources with weighted voting. (1) *ICD-10 primary tumor codes* (weight 4) use only the C00–C76 range for primary neoplasms. We record metastasis codes (C77–C79) solely as spread sites. This prevents clinically incorrect labels such as "bone metastasis" being assigned as a primary cancer. (2) *Tumor type keywords* (weight 3) provide a curated mapping of 180+ tumor type strings to canonical primary cancer names. (3) *Tumor-level flags* (weight 2) infer the primary cancer from the organ of origin for each tumor with `malignancy=yes` and `metastasis=no`. (4) *Two-tier report NLP* treats definitive radiologist statements such as "metastatic breast cancer" or "known melanoma" with weight 5. It treats 50+ standard regex patterns for cancer names, surgical procedures, and diagnostic signs with weight 1 to 3. Votes are aggregated at the patient level across all scans. The candidate with the highest weighted score is selected. Confidence is assigned as *high*, *medium*, or *low*. *High* requires the dominant candidate to score $\geq 2\times$ runner-up with total $\geq$3. *Medium* requires the dominant candidate to exceed the runner-up. *Low* covers the cases with no candidate or a tie.

Lesion-level temporal tracking. For patients with longitudinal imaging, we match lesions across consecutive scans using a two-pass algorithm. (1) Exact match on (canonical organ, location, tumor type). (2) Fuzzy match using organ family grouping for unmatched lesions. Matched lesions receive temporal change labels based on size ratios. The labels are growing ($>1.2\times$), shrinking ($<0.8\times$), or stable. Unmatched prior lesions are labeled *resolved*. Unmatched current lesions are labeled *new*.

Two-stage report parsing. Radiology reports are parsed in two stages. (1) Rule-based section detection identifies headers and dash-delimited lists. It extracts structured findings and RECIST assessments. (2) A sentence-splitting fallback handles unstructured narrative text. Recommendations and clinical impressions are separated from imaging findings.

B.9 Quality Control Flags

Automated QC checks are applied to every patient. They flag three categories of issues. *Timeline oscillations* (304 flags) occur when a lesion appears resolved and then reappears in a later scan. These flags concentrate in patients with $\geq$12 scans. They reflect radiologist reporting variability rather than true biological change. *Primary cancer uncertain* flags (422) arise when primary cancer confidence remains low despite all four resolver sources being consulted. They typically appear in patients whose reports describe non-specific findings without clearly indicating a primary malignancy. *Malignancy flag inconsistencies* (179 flags) occur when the malignancy label for a lesion changes across scans without an intervening treatment or biopsy event.

C Training-Path Details

This appendix expands on the SFT, RL, and evaluation paths summarized in Section 6.

C.1 Verifiable Rewards for Reinforcement Learning

The four-step structure exposes four reward signals that are deterministic and require no further human annotation. They fit GRPO-style RL recipes used by DeepSeek-R1 [43] and its medical variants [89, 61].

(i) Pathology match. The model's predicted cancer type is compared to $\mathcal{D}_t$. Exact match yields reward 1; mismatch yields 0. Available for every cancer-positive patient.

(ii) Organ-level malignancy and metastasis. For each organ in `organ_level_diagnosis`, the predicted (malignancy, metastasis) flags are checked against ground truth. The reward decomposes into a sum across organs and is multi-label.

(iii) Risk category. The predicted LI-RADS, PI-RADS, Bosniak, or analogous category is compared to $\kappa_{\mathcal{S}}(\mathcal{I}_t)$. Exact match is rewarded; within-± 1 category counts as partial credit.

(iv) Temporal change. For multi-scan inputs, the predicted change label per lesion is checked against Δ_{t_k}.

A format reward enforces the four-step output structure. To our knowledge, RadThinking is the first public resource that provides all four signals for cancer screening at scale.

C.2 Recommended Evaluation Protocols

We recommend reporting accuracy at each VQA tier, with case complexity (§B.6) as an orthogonal axis.

Foundation accuracy. Closed-form questions covering the atomic evidence cues (modality, contrast phase, organ presence, lesion presence, lesion size, attenuation). Matches the closed-question protocol of VQA-RAD [63], SLAKE [73], and OmniMedVQA [53].

Single-step reasoning accuracy. One clinical rule applied to one foundation observation. Threshold checks, single-feature classification, and single-change rules. Reports whether the model knows the rule but fails the perception, or vice versa.

Compositional accuracy. Multi-step questions ending in a clinical-guideline category (LI-RADS, PI-RADS, Bosniak, RECIST, TNM). Pathology serves as the terminal ground truth where applicable. Matches the medical VQA setup of LLaVA-Med [66] and Med-R1 [61], but with the chain of foundation answers exposed for step-level diagnosis.

Longitudinal compositional accuracy. Multi-scan input. The model must integrate the temporal trajectory before answering. The longitudinal cohort in RadThinking (2,563 cancer patients with 2 to 26 scans) is sized for this protocol.

C.3 Integration Recipe

A team integrating RadThinking into an existing VLM stack mixes the foundation and compositional SFT pairs into Stage 3 (a 5 to 15% mix ratio avoids oncology overfitting), uses the four verifiable rewards plus a four-step format reward in a Stage 4 GRPO pass, and reports results stratified by VQA tier and case complexity. Aggregate accuracy hides where the gains come from. Stratification keeps temporal and integrative gains visible.

D Illustrative Patient: Scan-by-Scan Reasoning Annotations

The following provides the full reasoning chain annotations for the six selected timepoints of the illustrative HCC patient shown in Figure 1. This patient was monitored over 26 CT scans across 11 years (2013–2024).

Scan 1 (2013-07): *Observation:* hemangioma in liver segment VII; post-resection changes in segment VIII. *Temporal:* no prior. *Context:* report impression: "status post resection of segment VIII without evidence of recurrence." *Complexity:* INTEGRATIVE. The hemangioma must be distinguished from recurrent HCC using clinical context, not imaging alone.

Scan 2 (2013-10): *Temporal:* **new** ill-defined hypervascular area near resection margin. *Context:* "likely perfusion alteration, but HCC recurrence cannot be excluded; follow-up in 3 months." *Complexity:* TEMPORAL. The new lesion is the decisive finding. Its nature is still ambiguous.

Scan 5 (2014-08): *Temporal:* prior suspicious lesion **resolved**; hemangioma stable. *Context:* "complete remission; stable postoperative cyst." *Complexity:* TEMPORAL. The resolution event is the key reasoning step. It confirms that the prior finding was benign.

Scan 17 (2019-12): *Observation:* **new** hypervascular lesion in segments VI/VII. *Context:* "LI-RADS 5, suspicious for HCC recurrence;" also a LI-RADS 3 lesion (indeterminate). *Complexity:* TEMPORAL. A new lesion after 5 years of remission changes the clinical trajectory.

Scan 24 (2023-03): *Observation:* three lesions (13 mm, 6 mm, 3 mm) in segments V/VII after microwave ablation. *Temporal:* one **growing**, two **new**. *Context:* "multifocal HCC; LI-RADS 3 lesions suspicious for recurrence; recommend liver MRI." *Complexity:* TEMPORAL. The pattern is recurrence after ablation.

Scan 26 (2024-04): *Observation:* two LI-RADS 3 lesions, one with arterial enhancement. *Context:* "no new hepatic lesions; known LI-RADS 3 lesions without progression." *Complexity:* INTEGRATIVE. Stability over 8 months, combined with LI-RADS criteria, downgrades concern. *Conclusion:* HCC, pathology-confirmed.

E Data Normalization Vocabulary

The raw dataset contains heterogeneous labels from pathology records, radiology reports, and ICD-10 codes. This appendix documents the complete normalization vocabulary used to construct structured reasoning chains. All original labels are preserved in the released data alongside canonical forms.

Organ name normalization. Table 4 maps all 191 raw organ surface forms to the 19 canonical screening targets plus ancillary organ categories. The normalization table was built by two-pass matching: exact string matching followed by substring matching.

Table 4: **Organ name normalization vocabulary for the 19 organ screening targets.** Each canonical organ is shown with its raw surface forms from the source data (observation count in parentheses). The 45 non-target organ categories (64 additional surface forms) follow RECIST 1.1 [39] general criteria.

canonical	*N*	**raw surface forms (observation count)**
thyroid	2	thyroid (255), thyroid gland (3)
lung	2	lung (5,343), lungs (6)
breast	1	breast (319)
esophagus	2	esophagus (98), esophagus/stomach (1)
liver	3	liver (13,034), lung/liver (1), liver/kidney (1)
gallbladder	1	gallbladder (77)
stomach	2	stomach (185), stomach/small intestine (1)
pancreas	2	pancreas (3,035), duodenum/pancreas (1)
spleen	2	spleen (748), liver/spleen (1)
duodenum	1	duodenum (80)
colon	8	colon (684), appendix (16), rectum (12), anal canal (1), anal region (1), small intestines/colon (1), stomach/colon (1), colon/duodenum (1)
kidney	5	kidney (3,991), left kidney (78), right kidney (48), kidneys (3), renal fossa (1)
adrenal	6	adrenal gland (1,501), adrenal (18), adrenal glands (7), adrenal gland/kidney (1), right adrenal gland (1), adrenal gland/bone (1)
bladder	2	bladder (296), urinary bladder (6)
prostate	1	prostate (170)
uterus	3	uterus (277), cervix (6), uterus/vagina (1)
ovary	4	ovary (288), ovaries (2), right ovary (1), left ovary (1)
lymph_node	14	lymph nodes (1,002), lymph node (782), axilla (19), supraclavicular (3), axillary (3), inguinal canal (2), cervical (2), cervical/supraclavicular (1), mediastinal lymph nodes (1), mediastinal (1), retroperitoneal lymph nodes (1), lymphatic system (1), interaortocaval (1), infracarinal (1)
bone	16	bone (3,957), spine (115), axial skeleton (8), rib (7), skeleton (6), sacrum (4), femur (4), skeletal system (3), skull (2), vertebra (2), skeletal (2), ilium (1), sternum (1), rib cage (1), scapula (1), paravertebral (1)
+ 45 non-target		peritoneum (5 forms), pelvis, soft_tissue (3), muscle (7), skin (3), chest_wall (4), vascular (18), mediastinum, pleura, small_intestine (5), retroperitoneum (2), abdomen, abdominal_wall, neck (2), thorax (2), . . .

Clinical deduplication of cancer types. Table 5 documents all seven merges of clinically equivalent cancer subtypes into unified groups, reducing the label set from 55 to 43 clinically distinct cancer groups.

Table 5: **Clinical deduplication of cancer type labels.** Seven groups of clinically equivalent subtypes are merged. All original fine-grained labels are preserved in the released data.

unified group	original labels merged	clinical justification
colorectal carcinoma	colon cancer (347), rectal cancer (100), colorectal cancer NOS (80), rectosigmoid cancer (3)	anatomically contiguous segments of the large bowel; shared TNM staging, NCCN guidelines, and screening protocols
lung carcinoma	lung cancer (430), small cell lung cancer (14), non-small cell lung cancer (7)	all primary lung malignancies; subtypes share the same organ screening target and imaging vocabulary
pancreatic ductal adenocarcinoma	pancreatic cancer NOS (280), pancreatic ductal adenocarcinoma (152)	PDAC accounts for >85% of pancreatic cancers; NOS labels typically reflect unspecified histology rather than a distinct entity
urothelial / bladder ca.	bladder cancer (216), urothelial carcinoma (137)	urothelial carcinoma constitutes >90% of bladder malignancies; both labels refer to the same disease
endometrial carcinoma	endometrial cancer (121), uterine cancer (39)	endometrial carcinoma is the dominant uterine malignancy (>90%); "uterine cancer" without qualifier denotes endometrial origin
neuroendocrine tumor	neuroendocrine tumor (133), pancreatic NET (12), neuroendocrine carcinoma (3)	shared neuroendocrine lineage; grouped for statistical power while acknowledging grade heterogeneity (NET G1/G2 vs. NEC G3)
cholangiocarcinoma	cholangiocarcinoma NOS (55), intrahepatic CCA (10), extrahepatic CCA (6)	subtypes of the same biliary epithelial malignancy; anatomic subsite preserved in released chain metadata

F Cancer Group to Organ Screening Target Mapping

Each reasoning chain is governed by the clinical reporting standard of the organ in which a finding is detected (Table 3). The patient's *primary cancer* may differ from the screened organ. For example, a breast cancer patient may present with liver metastases evaluated under LI-RADS. Table 6 maps all 43 cancer groups to the 19 organ screening targets. Groups marked *extra-organ* represent primary cancers outside the screened organs, detected via metastatic deposits or incidental findings on CT.

Table 6: **Mapping of 43 cancer groups to the 19 organ screening targets.** 31 groups map to a screened organ. 12 are "extra-organ" primary cancers detected via metastatic deposits or incidental findings on CT.

organ target	cancer groups	patients
thyroid	thyroid carcinoma	26
lung	lung carcinoma	451
breast	breast carcinoma	559
esophagus	esophageal carcinoma	61
liver	hepatocellular carcinoma, cholangiocarcinoma, hepatoblastoma	1,172
gallbladder	gallbladder carcinoma	12
stomach	gastric carcinoma, gastrointestinal stromal tumor	107
pancreas	pancreatic IPMN, pancreatic ductal adenoca., pancreatobiliary ca., pancreatic neoplasm (benign)	1,466
spleen	splenic neoplasm	9
duodenum	duodenal carcinoma, ampullary carcinoma	13
colon	colorectal carcinoma, anal carcinoma	532
kidney	renal cell carcinoma, clear cell carcinoma	787
adrenal	adrenal carcinoma	22
bladder	urothelial / bladder carcinoma	353
prostate	prostatic adenocarcinoma	477
uterus	endometrial carcinoma, cervical carcinoma, vulvar carcinoma	184
ovary	ovarian carcinoma, ovarian borderline tumor	177
lymph node	lymphoma	129
bone	multiple myeloma	10
subtotal: organ-mapped (31 groups)		*6,547*
extra-organ (detected via metastasis or incidental finding)	melanoma, neuroendocrine tumor, sarcoma, squamous cell carcinoma (NOS), cutaneous carcinoma, thymoma, mesothelioma, testicular carcinoma, laryngeal ca., pharyngeal ca., oropharyngeal ca., small intestine ca.	507
subtotal: extra-organ (12 groups)		*507*
total cancer-positive patients		**7,054**