OncoTriad-QA: A Patient-Level Radiology–Pathology–Genomics Benchmark for Pan-Cancer Reasoning

Ahnaf Munir[*1] **Dannong Wang**[*1] **Michael W. McDonald**[2]
Mubarak Shah[1] **Pegah Khosravi**[1,3] **Yu Tian**[†1]

[1]Institute of Artificial Intelligence, University of Central Florida
[2]AdventHealth Medical Group Urology at Celebration, FL
[3]Department of Clinical Sciences, College of Medicine, University of Central Florida
{ahnaf.munir, da304044, shah, pegah.khosravi, yu.tian2}@ucf.edu
michael.mcdonald.md@adventhealth.com

Abstract

Cancer diagnosis and characterization require integrating complementary evidence from radiology, pathology, genomics, and clinical metadata. However, most medical large language model (LLM) and vision-language model (VLM) benchmarks focus on isolated modalities or narrow image-text tasks, leaving patient-level oncology assessment across multiple evidence streams largely untested. We introduce OncoTriad-QA, a patient-level radiology–pathology–genomics benchmark for pan-cancer question answering. OncoTriad-QA contains 86.1k semantic questions across 9,281 TCGA patient cases from 32 cancer cohorts, aligning CT/MRI radiology, whole-slide histopathology, somatic mutations, copy-number alterations, DNA methylation, bulk RNA-seq, and clinical metadata. Case-specific annotations are constructed through a source-grounded LLM-assisted pipeline using curated labels, diagnostic reports, molecular profiles, and modality-derived evidence as primary sources of truth, with automated consistency checks and clinician review. We also introduce OncoVLM, a reference multimodal model that maps modality-native radiology, pathology, DNA methylation, and RNA-seq evidence into an LLM interface through learned projectors. Experiments show that existing general-purpose and medical LLMs remain limited on comprehensive pan-cancer QA, especially when questions require integrating imaging findings, tumor morphology, and molecular evidence. After fine-tuning on OncoTriad-QA, OncoVLM exceeds MedGemma-4B by an average of 10.7 points when using MCQ accuracy and BERTScore-F1, with consistent gains across multiple-choice and open-ended questions under radiology-only, pathology-only, and all-available settings. These results demonstrate the benchmark's value for training and evaluating models for integrated cancer question answering.

Code: `OncoTriad-QA` **Dataset**: `ai-mind-lab/OncoTriad-QA`

1 Introduction

Cancer exhibits substantial heterogeneity across organ systems, histologic subtypes, and molecular states, causing patients with superficially similar tumors to follow markedly different clinical

[*]Equal contribution.
[†]Corresponding author. Senior authorship shared with Pegah Khosravi and Mubarak Shah.

Preprint.

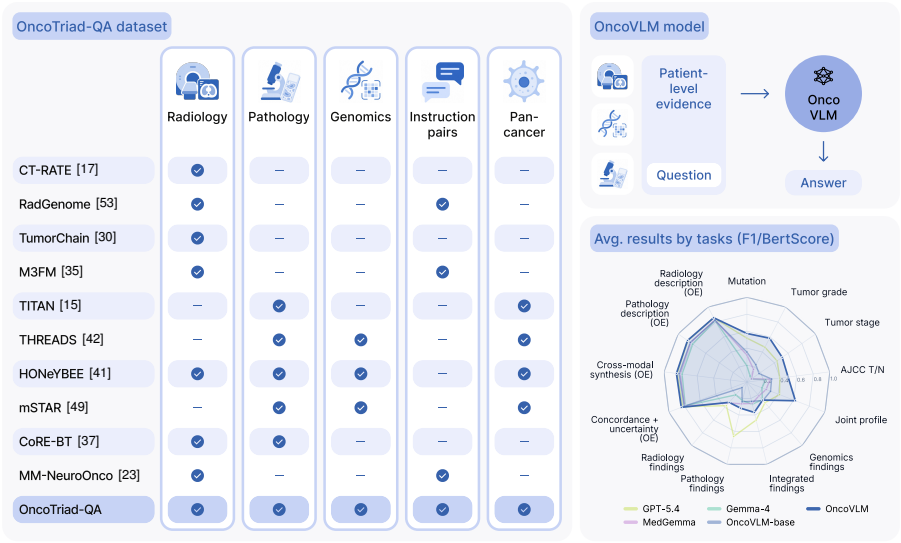

Figure 1: Existing oncology datasets and multimodal medical resources typically cover only subsets of the evidence used in cancer characterization. OncoTriad-QA supports patient-level integration of radiology, pathology, genomics, clinical metadata, and instruction-style QA across pan-cancer cohorts, enabling evaluation of modality-specific, cross-modal, and missing-modality performance.

trajectories [33, 19]. Clinical cancer characterization therefore relies on evidence across multiple scales: macroscopic anatomy from computed tomography (CT) / magnetic resonance imaging (MRI), microscopic architecture from whole-slide histopathology, and molecular alterations from genomic profiling [5, 31]. These streams are complementary rather than interchangeable: radiology captures lesion extent and anatomic context, histopathology captures cellular morphology and tissue organization, and molecular profiling captures subtype-defining alterations and treatment relevance. A benchmark limited to one stream can measure modality-specific recognition, but not whether a model can integrate heterogeneous evidence into a unified cancer interpretation.

Most medical large language model (LLM) and vision-language model (VLM) benchmarks still evaluate isolated or narrow modality combinations. Recent medical VLMs, including Med-Flamingo [34], LLaVA-Med [29], and MedGemma [38], and modality-specific foundation models such as RadFM [45], TITAN [15], and MUSK [46], have shown strong transfer within radiology, pathology, or biomedical image-text tasks. However, performance on pathology reports or captions, radiology, or single-modality image understanding does not ensure patient-level oncology understanding: a model may recognize findings in isolation but still fail to determine whether imaging appearance, tumor morphology, molecular alterations, and clinical context are concordant within the same patient. As illustrated in Figure 1, existing oncology resources usually cover only partial evidence streams: radiology datasets often lack pathology and molecular data, pathology–genomics resources often lack radiology, and general medical instruction datasets rarely provide aligned patient-level evidence across all three modalities plus clinical metadata. This creates an evaluation gap between modality-specific recognition and the integrative interpretation required in oncology, motivating a unified case-level benchmark for modality-specific, cross-modal, and missing-modality QA.

To address this gap, we introduce **OncoTriad-QA**, the first patient-level radiology, pathology, and genomics benchmark for multi-cancer-type (pan-cancer) analysis. Across approximately 9,000 TCGA patient cases from 32 cancer cohorts, OncoTriad-QA links CT/MRI radiology from TCIA [12], whole-slide histopathology from TCGA [44], and GDC [21] clinical and molecular profiles, including somatic mutations, copy-number alterations, DNA methylation, and bulk RNA-seq. The benchmark contains 86.1k semantic questions, which are then expanded into all-available, radiology-only, and pathology-only versions to support evaluation under different evidence settings. Rather than treating each evidence stream as a separate task, OncoTriad-QA frames cancer characterization as patient-level QA, with MCQs and open-ended questions probing modality-specific recognition, molecular

grounding, subtype reasoning, staging and grading, cross-modal concordance, and missing-modality robustness.

Because OncoTriad-QA uses LLM-assisted summarization and QA construction at scale, we make reliability and auditability central to the benchmark design. Each case summary and QA item is generated from structured modality-specific evidence, followed by automatic format and consistency checks and medical expert review on a subset of cases. Specifically, the benchmark design rests on the following safeguards: (1) fixed patient-level data splits, so that all evidence and questions for a given patient fall entirely within either the training or the test partition, preventing case-level leakage across splits; (2) evaluation protocols stratified by cancer cohort, so that performance can be reported and compared per cancer type rather than masked by an aggregate score; and (3) evaluation protocols stratified by modality availability, so that models are assessed separately under all-available, radiology-only, and pathology-only conditions rather than assuming complete evidence is always present. Together, these checks reduce unsupported claims and help assess whether models use the available patient evidence rather than single-modality shortcuts.

To demonstrate how the benchmark supports both training and evaluation, we further introduce **OncoVLM**, a reference multimodal language model trained on OncoTriad-QA. Consistent with the fixed patient-level splits described above, OncoTriad-QA is partitioned into a training set, used to instruction-tune OncoVLM, and a held-out test set, used exclusively for evaluation, with no patient case appearing in both. OncoVLM connects frozen radiology, pathology, methylation, and transcriptomic encoders to an LLM backbone through learned projectors and instruction tuning, allowing heterogeneous patient evidence to be represented in a unified language-model interface. We benchmark OncoVLM against general medical VLMs, and open- and closed-source multimodal LLM baselines. Results show that existing models remain limited on pan-cancer QA, while fine-tuned OncoVLM exceeds MedGemma-4B by approximately 10.7 points when using MCQ accuracy and BERTScore-F1 across radiology-only, pathology-only, and all-available settings.

The main contributions of our work are:

- **OncoTriad-QA, a clinician-guided patient-level benchmark for multimodal pan-cancer assessment.** We construct a multimodal oncology benchmark that aligns imaging, histologic, molecular, and clinical evidence across approximately 9,000 patient cases from 32 cancer cohorts, with medical-expert guided prompt/question design, and case-summary audit for clinical validity. The benchmark includes MCQs and open-ended questions targeting staging and grading, modality-specific recognition, molecular grounding, and cross-modal concordance.
- **OncoVLM, a novel multimodal model for oncology QA.** Unlike many existing multimodal medical VLMs restricted to low-resolution PNG proxies, OncoVLM ingests modality-native clinical data, such as gigapixel whole-slide pathology, multi-slice radiology, and genomic sequences, through modality-specific encoders and learned projectors, and can work with multiple modality combinations and handle missing modalities.
- **A universal evaluation framework of multimodal oncology.** Designed to mimic real clinical settings, our protocol evaluates multimodal cancer QA across arbitrary modality combinations. It remains agnostic to VLM input design by providing fallbacks whenever a model cannot consume original full-scale imaging or molecular profiles. Using this protocol, we evaluate OncoVLM with state-of-the-art proprietary multimodal and open-weight baselines and medical VLMs.

2 Related Work

Medical vision-language models. General-purpose medical VLMs have achieved strong transfer across clinical imaging tasks, with systems ranging from contrastive pretraining on biomedical figure-caption pairs [51] to few-shot multimodal learning [34] and GPT-supervised instruction tuning [29, 6, 38]. Most of these systems, however, are evaluated on single-image VQA benchmarks such as VQA-RAD, SLAKE, and PathVQA [52], which pair one radiology or pathology image with a short question and provide no molecular or cross-modal context. Despite strong in-domain transfer, such systems are trained on single modalities or narrow modality pairs and cannot jointly reason over radiology, histopathology, and molecular data within a unified framework.

Modality-specific foundation models. Within individual modalities, large-scale pretraining has produced capable specialists. In radiology, RadFM [45], CheXagent [10], and BioViL-T [4] achieve strong performance on report generation and image interpretation. In pathology, UNI [9], CONCH [32], MUSK [46], TITAN [15], GigaPath [48], and THREADS [42] encode rich slide-level representations, with recent models incorporating molecular supervision. In genomics, Geneformer [40], scGPT [13], and BulkFormer [28] capture transcriptomic and epigenomic structure at scale. Yet none can jointly process imaging and molecular evidence at inference time, nor support open-ended interpretation across all three modalities.

Patient-level oncology benchmarks. A separate line of work builds multimodal datasets directly from TCGA/TCIA cohorts rather than from image-caption pairs. CT-RATE [17] and RadGenome [53] pair chest CT with reports and grounded findings but stay radiology-only; TumorChain [30] introduces interleaved chain-of-thought traces for tumor analysis but is similarly confined to imaging. On the pathology side, TITAN [15], THREADS [42], HONeYBEE [41], and mSTAR [49] couple whole-slide images with limited genomic or textual supervision for representation learning rather than open-ended QA. Disease-specific multimodal benchmarks such as CoRe-BT [37] and MM-NeuroOnco [23] integrate radiology, pathology, and text for brain tumor typing, but are restricted to a single cancer type. M3FM [35] targets chest CT screening with instruction pairs, again within one modality. None of these resources align radiology, pathology, and a comprehensive molecular profile (mutations, copy-number alterations, methylation, and transcriptomics) at the patient level across a pan-cancer cohort, which is the gap OncoTriad-QA is designed to close.

Multimodal fusion for oncology. Prior work has combined histopathology and molecular data for tasks such as survival prediction and subtype classification [8, 7]. While these methods demonstrate the value of cross-modal integration, they produce scalar task-specific outputs rather than general-purpose language-grounded responses. Like prior medical VLMs, they too remain confined to a subset of modalities, with none jointly incorporating radiology, pathology, and genomics.

Instruction tuning in medical AI. Instruction tuning has proven effective for medical VLMs, as shown by MedTrinity-25M [47] and similar efforts [15]. The complementary framework of learning with privileged information (LUPI) [39], where auxiliary signals available only at training time guide representation learning, has been applied in computational pathology via TriDeNT [18]. Our work extends both paradigms to multimodal oncology. Here, the LLM observes full multimodal inputs and structured clinical ground truth during dataset construction, while OncoVLM learns to replicate this integrated interpretation from encoder-derived representations alone at inference time.

3 Dataset construction

OncoTriad-QA is a patient-level multimodal benchmark spanning 9,281 cases drawn from 32 TCGA cancer cohorts. For every case we assemble all available evidence from three complementary sources and align it under a single patient identifier: CT and MR radiology from The Cancer Imaging Archive (TCIA) [12], whole-slide histopathology from TCGA [44], and molecular and clinical profiles from the Genomic Data Commons (GDC) [21], the latter comprising somatic mutations, copy-number alterations, DNA methylation, and bulk RNA-seq. Because the set of modalities available for any given patient varies widely in routine clinical practice [43, 50], we do not assume complete data. Instead, we record per-case modality availability and materialize each case under three evidence settings—*all-available*, *radiology-only*, and *pathology-only*—so that the benchmark supports both complete multimodal evaluation and realistic missing-modality scenarios.

Our central design goal is reliability at scale. Rather than asking a language model to answer clinical questions from raw images, we use GPT-5.4 [36] only as a *summarizer* and *question writer* conditioned on curated, structured evidence, so that every generated item is traceable to an explicit source field. In total, OncoTriad-QA contains approximately 86.1k distinct semantic questions—43.2k universal MCQs, 21.2k case-specific MCQs, and 21.7k case-specific open-ended questions—which, after modality-condition expansion, yield 182.6k materialized question rows.

The remainder of this section describes the construction pipeline (Figure 2) in the order it is applied: how per-patient evidence is aligned into a single *detailed case summary*, how the three question streams—*universal MCQs*, *case-specific MCQs*, and *case-specific open-ended questions*—are generated from it, and finally the modality-condition expansion and quality-control steps that produce the released benchmark.

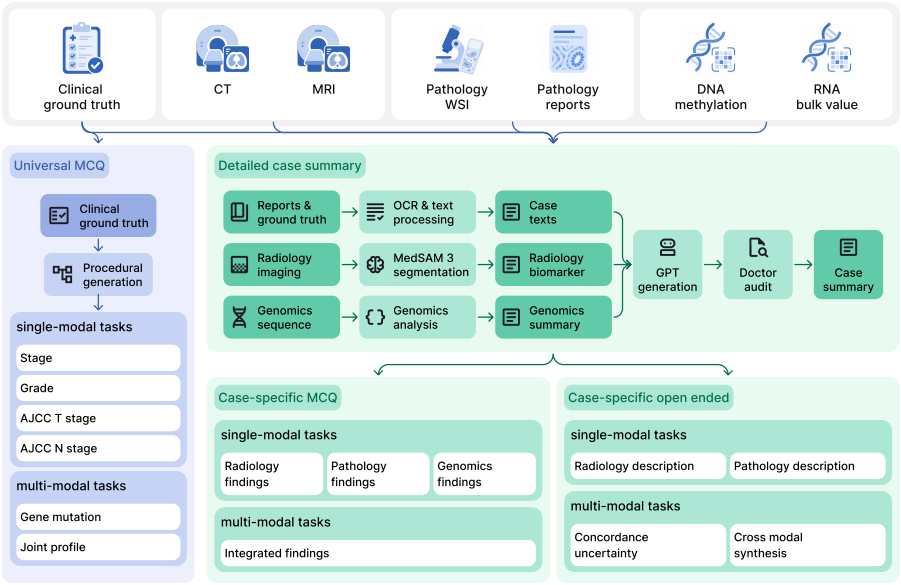

Figure 2: Overview of OncoTriad-QA construction. Clinical metadata, radiology, pathology, DNA methylation, and bulk RNA-seq are converted into structured patient-level evidence. Clinical variables generate universal MCQs, while modality-specific processing supports GPT-based case summaries and QA generation. The final benchmark includes case-specific MCQs and open-ended QAs across radiology, pathology, genomics, integrated findings, cross-modal synthesis, and uncertainty, with physician audit for reliability.

Detailed case summary. We treat the patient case as the unit of the benchmark and convert each patient's heterogeneous evidence into a single structured case summary that serves as the canonical context for all case-specific tasks. For each patient we resolve one canonical record per modality (preferring tumor over normal samples for RNA-seq, masked ensemble MAFs for mutations, and standardized copy-number outputs), and retain a case when it carries curated clinical ground truth plus at least one processable modality. The summary then interleaves three deterministic, modality-specific text blocks with this ground truth.

(i) Radiology. CT and MR tumor regions are segmented with Medical SAM3 [27] using a text prompts applied uniformly across CT, MR, and all 32 cohorts. For each case, imaging series are ranked by segmentation consistency and mean tumor area; we retain the top three (k=3) and two representative slices per series (the middle and largest-tumor-area slices). Geometric, intensity, shape, and texture biomarkers are computed per slice and aggregated to a single case-level vector by a quality-weighted mean, rendered into a fixed-schema text block; masks are also converted to bounding boxes for language-model annotation.

(ii) Pathology. Pathology reports and curated ground-truth labels provide the primary evidence for tumor morphology, normalized into a text block via OCR and text processing. Because closed-source LLMs cannot process gigapixel whole-slide images (WSIs), we additionally supply representative ROI tiles from TCGA-UniformTumor-8K [15] rather than full WSIs, surfacing visual patterns the reports may not capture.

(iii) Genomics. Raw GDC molecular data are converted into a structured genomics text block by a deterministic, generation-free pipeline summarizing RNA expression, DNA methylation, somatic mutations, copy-number alterations, pathway activity, immune and stromal signatures, molecular-subtype indicators, driver-gene status, TMB, MSI status, and HRD score.

Universal MCQs. We procedurally generate universal MCQs from curated structured labels using standardized templates instantiated whenever the required patient-level variable is available, covering tumor stage, tumor grade, AJCC T/N category, mutation status, and joint stage–mutation profiles. Because answers come directly from ground-truth fields, this stream (43.2k questions) is independent of LLM question construction and gives a controlled evaluation setting with a fixed answer schema.

Case-specific MCQs. Case-specific MCQs are generated from the detailed case summary and probe whether a model can match patient-specific evidence rather than rely on cancer-type priors or answer style. To avoid the language-pattern shortcuts that arise when a model writes its own distractors [16, 3], we never prompt GPT for distractors directly. Instead, we extract a *contrastive fingerprint* for each case by showing GPT the target alongside three other cases from the same cohort: the target's fingerprint becomes the correct option and the others serve as distractors. Because all options are drawn from real same-cohort patients, the task depends on discriminating patient-specific evidence rather than stylistic cues. This stream (21.2k questions) spans single-modal radiology, pathology, and genomics findings plus a multi-modal integrated-findings task.

Case-specific open-ended questions. For open-ended QA, GPT uses the case summary to produce concise reference answers across four task categories: radiology description and pathology description (single-modal), and cross-modal synthesis and concordance and uncertainty (multi-modal). Generation follows guidelines emphasizing evidence-grounded wording, explicit uncertainty when information is incomplete, and avoidance of unsupported clinical conclusions. This stream contributes 21.7k questions.

Modality-condition expansion. Each semantic question is finally materialized under every evidence setting whose required modalities are available for its case—the *all-available* setting and any *single-modality* setting that still contains all the evidence the question depends on. This preserves a single semantic intent while exposing models to it under different evidence configurations, yields 182.6k materialized rows, and lets the evaluation protocol be stratified jointly by cohort and modality availability.

Quality control. Before release, every summary and QA item passes automatic checks including schema and format validation, agreement of MCQ answer keys with curated ground-truth fields, distractor well-formedness, and rejection of answers referencing evidence absent from the summary. To further assess quality, we obtained physician feedback on a subset of captions and MCQs. The reviewer flagged cases with insufficient imaging context or incomplete stage, grade, lymph-node, or genomic evidence, and we used this feedback to refine the generation guidelines toward evidence-grounded wording, explicit uncertainty, and avoidance of unsupported conclusions.

4 OncoVLM Architecture

OncoTriad-QA supports both evaluation and supervised training of patient-level multimodal oncology models. Alongside the benchmark, we introduce OncoVLM, a model designed to use modality-native oncology evidence rather than only text summaries or low-resolution image proxies. OncoVLM encodes multi-slice radiology series, gigapixel whole-slide pathology patch sets, DNA methylation beta-value profiles, and bulk RNA-seq expression profiles, then maps these heterogeneous inputs into a shared language-model interface.

Modality encoding and projection. For each patient case, the available evidence is processed through frozen modality-specific encoders. Radiology inputs are encoded with MedSigLIP-448 [38]; Pathology WSIs are encoded with pathology foundation model UNI [9]; RNA-seq expression and DNA methylation beta values are encoded with BulkFormer [28] and CpGPT [14], respectively. Keeping these encoders frozen preserves modality-specific representations learned from large-scale pretraining and reduces overfitting across heterogeneous cancer cohorts [29]. Each encoder is paired with a learned projector that maps its outputs into the language model embedding space. Formally, for a patient case with available modality set $\mathcal{M} \subseteq \{\text{RAD}, \text{PATH}, \text{RNA}, \text{MET}\}$, each modality $m \in \mathcal{M}$ has a raw input x_m, a frozen encoder $\mathcal{E}_{\phi_m}$, and a trainable projector $\mathcal{P}_{\theta_m}$ producing a block of modality tokens

$$\mathbf{z}_m = \mathcal{P}_{\theta_m}(\mathcal{E}_{\phi_m}(x_m)) \in \mathbb{R}^{n_m \times d}, \tag{1}$$

where d is the LLM embedding dimension and n_m the number of tokens contributed by modality m. Since pathology slides produce variable-length and long sets of patch embeddings, the pathology branch uses a Perceiver-style resampler [26, 1] to compress slide-level evidence into a fixed number of tokens $n_{\text{PATH}} = K$, rather than a simple MLP. Radiology, DNA methylation, and RNA-seq use modality-specific MLP projectors for slice-level or case-level embeddings. The projected tokens $\{\mathbf{z}_m\}_{m \in \mathcal{M}}$ are inserted into the instruction prompt with modality tags. Missing modalities are handled by omitting the corresponding token blocks (i.e., restricting $\mathcal{M}$), allowing the model to

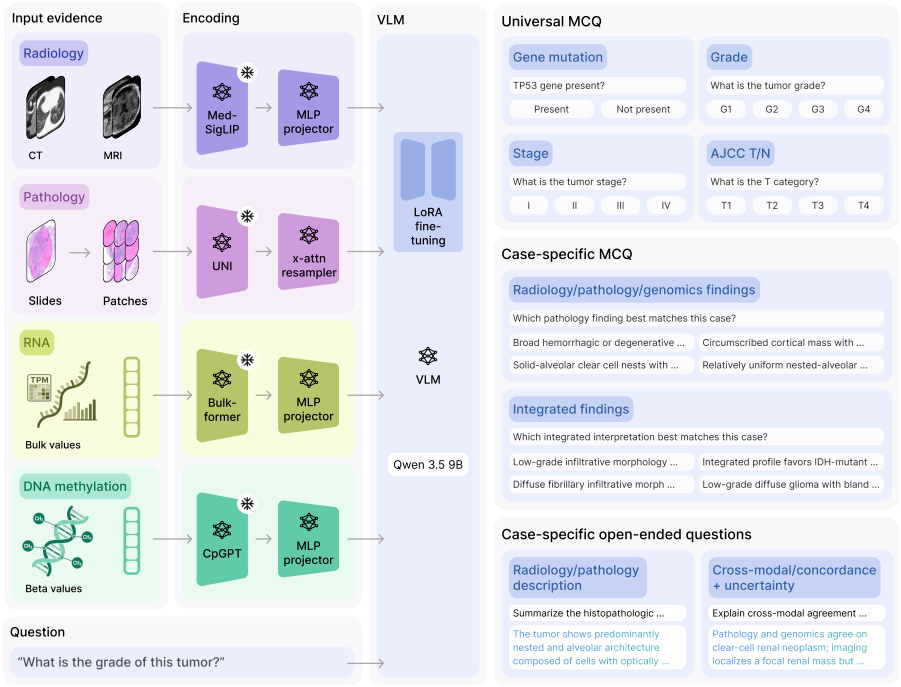

Figure 3: Overview of the proposed multimodal architecture (OncoVLM). Modality-specific encoders map radiology, pathology, and genomic inputs into a shared representation space, which is processed by a LoRA-tuned vision-language model to perform classification and multimodal tasks.

operate under the same evidence configurations used in OncoTriad-QA, including all-available, pathology-only, and radiology-only settings.

Training objective. Both training stages share a single masked autoregressive next-token cross-entropy objective and differ only in the supervision target and the subset of trainable parameters. Let $s = (s_1, \ldots, s_T)$ denote the assembled token sequence (modality tokens, instruction prompt, and target text) and let $\mathbf{1}[t \in \mathcal{T}]$ be a label mask that is 1 for supervised target tokens $\mathcal{T}$ and 0 for conditioning tokens. The objective is

$$\mathcal{L} = -\frac{1}{|\mathcal{T}|} \sum_{t=1}^{T} \mathbf{1}[t \in \mathcal{T}] \, \log p(s_t \mid s_{<t}) \, . \tag{2}$$

Modality token blocks and prompt tokens are always masked out, so they act purely as conditioning context and never contribute gradients to the language-model output.

Two-stage training. In the *modality-alignment stage*, each projector is trained independently on modality-specific captions while the corresponding encoder and the language model remain frozen. Concretely, for modality m with a paired caption $c^{(m)} = (c_1, \ldots, c_{L_m})$, only the projector parameters θ_m are optimized under Eq. (2) with the caption tokens as targets:

$$\mathcal{L}_{\text{align}}(\theta_m) = -\,\mathbb{E}_{(x_m, c^{(m)})} \sum_{t=1}^{L_m} \log p_\Theta(c_t \mid c_{<t},\, \mathbf{z}_m) \, , \qquad \mathbf{z}_m = \mathcal{P}_{\theta_m}(\mathcal{E}_{\phi_m}(x_m)) \, , \tag{3}$$

where the LLM parameters Θ and encoder parameters ϕ_m are held fixed. This maps each frozen biomedical representation into the LLM embedding space before any joint VQA training, giving each modality a stable language interface and reducing early cross-modal interference [25].

In the *supervised fine-tuning (SFT) stage*, the encoders and aligned projectors are frozen, and the language model is adapted with low-rank adaptation (LoRA) [24] on the OncoTriad-QA training split. Given an instruction prompt $\mathbf{q}$, the available modality token blocks $\{\mathbf{z}_m\}_{m \in \mathcal{M}}$, and a target answer $y = (y_1, \ldots, y_{L_y})$, only the LoRA parameters $\Delta\Theta$ are updated while Θ, $\{\phi_m\}$, and $\{\theta_m\}$

remain frozen:

$$\mathcal{L}_{\text{SFT}}(\Delta\Theta) = -\,\mathbb{E}_{(\mathbf{q},\{\mathbf{z}_m\},y)} \sum_{t=1}^{L_y} \log p_{\,\Theta\oplus\Delta\Theta}(y_t \mid y_{<t},\, \mathbf{q},\, \{\mathbf{z}_m\}_{m\in\mathcal{M}})\,. \tag{4}$$

This is the standard autoregressive next-token cross-entropy loss over target answer tokens, with the instruction prompt and modality tokens used as conditioning context. Training over varying modality subsets $\mathcal{M}$ (all-available, pathology-only, radiology-only) lets a single set of LoRA weights serve every evidence configuration without architectural changes.

5 Experiments

We use Qwen3.5-9B as the OncoVLM language model backbone and evaluate the following: (i) **GPT-5.4** [36], a proprietary general-purpose model used as a strong reference rather than a directly comparable trainable system; (ii) **MedGemma-4B** [38], a medical vision-language foundation model; (iii) **Gemma-4-E4B** [20], a recent frontier open-weight model; (iv) **OncoVLM-base**, OncoVLM with trained projectors but without the fine-tuning stage; (v) **OncoVLM-finetuned**, the OncoVLM-base model fine-tuned on OncoTriad-QA. Baselines receive representative pathology ROI images, representative radiology slices, and structured molecular text summaries when available, whereas OncoVLM receives encoder-derived modality representations. We evaluate on the held-out OncoTriad-QA test split. MCQs use accuracy and macro-F1, while open-ended QAs use BERTScore-F1, ROUGE-L, and an LLM-as-judge evaluation on BRCA. Claude Sonnet 4.6 [2] is used as the judge to compare OncoVLM against GPT-5.4 pairwise. We report OncoVLM win rate, counting ties as half wins. Table 1 results are averaged over three independent runs with standard deviations reported.

5.1 Main Results

Instruction tuning improves structured clinical prediction. Table 1 shows that fine-tuned OncoVLM achieves a 21.8-point accuracy gain over the projector-only base model on Universal MCQs. It also consistently outperforms open medical VLM baselines and surpasses the GPT-5.4 reference on the average Universal MCQ score. This pattern is expected because Universal MCQs are generated from curated clinical and molecular labels with a fixed answer schema. The largest gains occur on structured and joint-profile tasks, suggesting that instruction tuning helps the model map patient-level multimodal evidence to standardized oncology labels.

Case-specific MCQs require flexible evidence matching. Case-specific MCQs show a different pattern from Universal MCQs. Although fine-tuned OncoVLM improves over the projector-only model, GPT-5.4 remains strongest overall, and open-weight VLMs are closer to OncoVLM than in the universal tasks. This suggests that the difficulty is not simply a failure to learn fixed oncology labels, but the low-structure nature of the case-specific MCQ task itself. Each option is a contrastive fingerprint extracted by comparing the target case with same-cohort cases, so the discriminative cue can vary across examples. Such heterogeneity offers fewer reusable templates for SFT, making task-level adaptation difficult [11, 22].

LLM-as-judge reveals stronger evidence grounding in open-ended QA. For open-ended QA, BERTScore-F1 is tightly clustered across models, while ROUGE-L and LLM-as-judge results show clearer gains for OncoVLM. As shown in Table 4, the LLM judge prefers OncoVLM over GPT-5.4 on BRCA open-ended questions with an overall win rate of 0.702, with the strongest gains on concordance and uncertainty, cross-modal synthesis, and radiology description. Further analysis revealed that the judge often favors OncoVLM because GPT-5.4 introduces findings absent from the case summary, whereas OncoVLM stays closer to the available patient evidence. This is consistent with Table 3, where OncoVLM better preserves discriminative case-specific findings.

Cohort-level gains vary with supervision scale and heterogeneity. Table 2 shows that fine-tuned OncoVLM improves MCQ performance across the reported cancer groups, but the gains are not uniform. Larger cohorts show clearer improvements, while smaller cohorts show smaller margins over GPT-5.4 and sometimes higher variation. This indicates that cohort-level performance is shaped not only by modality availability, but also by the amount and consistency of paired supervision. For example, Brain cancer remains competitive rather than clearly dominated, suggesting that additional cohort-specific supervision may be as important as adding more modalities.

Table 1: OncoTriad-QA benchmark results by task category on the all-available modality setting. Values are mean and standard deviation over three separate run, and using different seeds on non-GPT model.

Name	GPT-5.4	MedGemma	Gemma-4	OncoVLM	OncoVLM
	–	4B	E4B	Qwen-3.5-9B	Qwen-3.5-9B
Finetuned	—	—	—	—	✓
Use projector	—	—	—	✓	✓
Universal MCQ (Accuracy% / F1)					
Mutation	51.5 / 0.515 ± 1.2 / ± 0.01	46.0 / 0.317 ± 0.0 / ± 0.00	11.5 / 0.190 ± 0.2 / ± 0.00	44.2 / 0.344 ± 0.3 / ± 0.01	**58.2 / 0.577** ± 1.3 / ± 0.01
Tumor grade	52.2 / 0.384 ± 2.7 / ± 0.03	40.3 / 0.162 ± 0.5 / ± 0.02	10.3 / 0.073 ± 1.3 / ± 0.01	37.5 / 0.178 ± 0.8 / ± 0.00	**64.7 / 0.615** ± 0.5 / ± 0.05
Tumor stage	44.5 / 0.464 ± 1.0 / ± 0.01	7.7 / 0.074 ± 0.3 / ± 0.00	5.7 / 0.053 ± 0.8 / ± 0.01	18.8 / 0.180 ± 1.3 / ± 0.02	**51.8 / 0.480** ± 2.5 / ± 0.03
AJCC T/N	42.3 / 0.400 ± 1.5 / ± 0.02	36.4 / 0.309 ± 0.2 / ± 0.00	39.2 / 0.213 ± 0.1 / ± 0.00	39.6 / 0.310 ± 1.4 / ± 0.01	**60.0 / 0.520** ± 1.1 / ± 0.00
Joint profile	38.9 / 0.388 ± 2.3 / ± 0.02	28.9 / 0.246 ± 0.3 / ± 0.00	18.2 / 0.197 ± 1.0 / ± 0.01	31.9 / 0.314 ± 0.6 / ± 0.01	**62.4 / 0.623** ± 1.2 / ± 0.01
Mean	46.0 / 0.449 ± 1.6 / ± 0.02	36.0 / 0.267 ± 0.2 / ± 0.00	17.9 / 0.177 ± 0.5 / ± 0.01	37.6 / 0.306 ± 0.7 / ± 0.01	**59.4 / 0.571** ± 1.3 / ± 0.01
Case-specific MCQ (Accuracy% / F1)					
Radiology findings	**42.3 / 0.419** ± 8.1 / ± 0.08	34.9 / 0.332 ± 3.2 / ± 0.03	23.8 / 0.202 ± 1.6 / ± 0.01	5.8 / 0.082 ± 0.9 / ± 0.01	31.7 / 0.320 ± 4.2 / ± 0.04
Pathology findings	**63.2 / 0.630** ± 1.3 / ± 0.01	24.3 / 0.224 ± 0.1 / ± 0.00	24.0 / 0.235 ± 1.0 / ± 0.01	24.2 / 0.222 ± 0.9 / ± 0.01	32.1 / 0.319 ± 1.0 / ± 0.01
Genomics findings	**30.5 / 0.305** ± 1.1 / ± 0.01	23.6 / 0.231 ± 0.5 / ± 0.00	27.4 / 0.263 ± 0.4 / ± 0.01	26.1 / 0.268 ± 0.6 / ± 0.01	29.4 / 0.293 ± 2.5 / ± 0.02
Integrated findings	**47.8 / 0.478** ± 0.2 / ± 0.00	25.7 / 0.260 ± 0.2 / ± 0.00	27.9 / 0.272 ± 0.4 / ± 0.00	24.4 / 0.242 ± 0.8 / ± 0.01	36.0 / 0.359 ± 2.0 / ± 0.02
Mean	**47.0 / 0.470** ± 1.2 / ± 0.01	25.0 / 0.243 ± 0.4 / ± 0.00	26.3 / 0.254 ± 0.6 / ± 0.01	24.0 / 0.236 ± 0.8 / ± 0.01	32.5 / 0.323 ± 1.9 / ± 0.02
Case specific open-ended (BERT-F1 / ROUGE-L)					
Concordance + uncertainty	0.823 / 0.150 ± 0.00 / ± 0.00	0.795 / 0.120 ± 0.00 / ± 0.00	0.788 / 0.103 ± 0.00 / ± 0.00	0.802 / 0.127 ± 0.00 / ± 0.00	**0.840 / 0.218** ± 0.00 / ± 0.00
Cross-modal synthesis	0.839 / 0.195 ± 0.00 / ± 0.00	0.807 / 0.144 ± 0.00 / ± 0.00	0.782 / 0.109 ± 0.00 / ± 0.00	0.815 / 0.153 ± 0.00 / ± 0.00	**0.850 / 0.241** ± 0.00 / ± 0.00
Pathology description	0.861 / 0.266 ± 0.00 / ± 0.00	0.819 / 0.165 ± 0.00 / ± 0.00	0.788 / 0.106 ± 0.00 / ± 0.00	0.821 / 0.164 ± 0.00 / ± 0.00	**0.863 / 0.314** ± 0.00 / ± 0.00
Radiology description	0.853 / 0.239 ± 0.00 / ± 0.00	0.828 / 0.212 ± 0.00 / ± 0.00	0.820 / 0.164 ± 0.00 / ± 0.00	0.821 / 0.169 ± 0.00 / ± 0.00	**0.857 / 0.289** ± 0.00 / ± 0.01
Mean	0.841 / 0.204 ± 0.00 / ± 0.00	0.808 / 0.147 ± 0.00 / ± 0.00	0.789 / 0.110 ± 0.00 / ± 0.00	0.813 / 0.149 ± 0.00 / ± 0.00	**0.851 / 0.258** ± 0.00 / ± 0.00

Modality ablations show complementary evidence. Table 5 shows that different modalities support different task types. The all-available setting performs best on integrative MCQ categories, supporting the value of combining radiology, pathology, and molecular evidence. Pathology-only inputs preserve much of the performance on grade and pathology-description tasks, while radiology-only inputs are more competitive on radiology findings and remain useful for integrated findings. In contrast, open-ended BERTScore-F1 changes only slightly when modalities are removed, again suggesting that surface-level semantic metrics are less sensitive to missing-evidence errors than structured decisions.

6 Limitations

Interpretation of our results is constrained by the public retrospective cohorts underlying OncoTriad-QA. These cohorts skew toward surgically resectable, treatment-naïve disease and incompletely

Table 2: VQA benchmark results by cancer type on the all-available modality setting. Each cancer is split into MCQ and open-ended rows. MCQ cells report Accuracy% and macro-F1; open-ended cells report BERTScore-F1 and ROUGE-L F1. Values are mean and standard deviation over inference repeats when available.

Name		GPT-5.4	MedGemma	Gemma-4	OncoVLM	OncoVLM
		–	4B	E4B	Qwen-3.5-9B	Qwen-3.5-9B
Finetuned		—	—	—	—	✓
Use projector		—	—	—	✓	✓
Cancer	**Group**	**Performance by model**				
Breast (BRCA)	MCQ	38.3 / 0.345 ± 1.5 / ± 0.01	26.4 / 0.217 ± 0.4 / ± 0.00	24.2 / 0.278 ± 1.5 / ± 0.02	32.2 / 0.279 ± 1.4 / ± 0.02	**55.6 / 0.543** ± 1.4 / ± 0.02
	Open-ended	0.842 / 0.205 ± 0.00 / ± 0.00	0.811 / 0.150 ± 0.00 / ± 0.00	0.791 / 0.111 ± 0.00 / ± 0.00	0.816 / 0.151 ± 0.00 / ± 0.00	**0.855 / 0.274** ± 0.00 / ± 0.00
Kidney (KIRC)	MCQ	40.2 / 0.389 ± 2.7 / ± 0.03	36.7 / 0.339 ± 0.8 / ± 0.01	26.3 / 0.270 ± 1.1 / ± 0.01	24.4 / 0.268 ± 0.8 / ± 0.01	**53.9 / 0.541** ± 2.7 / ± 0.02
	Open-ended	0.848 / 0.228 ± 0.00 / ± 0.00	0.812 / 0.158 ± 0.00 / ± 0.00	0.794 / 0.115 ± 0.00 / ± 0.00	0.804 / 0.149 ± 0.00 / ± 0.00	**0.854 / 0.272** ± 0.00 / ± 0.00
Brain (LGG)	MCQ	49.4 / 0.490 ± 3.1 / ± 0.04	32.5 / 0.325 ± 0.2 / ± 0.00	24.5 / 0.246 ± 1.1 / ± 0.01	37.4 / 0.338 ± 1.6 / ± 0.03	**55.3 / 0.550** ± 2.1 / ± 0.02
	Open-ended	0.841 / 0.206 ± 0.00 / ± 0.00	0.803 / 0.142 ± 0.00 / ± 0.00	0.789 / 0.110 ± 0.00 / ± 0.00	0.820 / 0.153 ± 0.00 / ± 0.01	**0.852 / 0.267** ± 0.00 / ± 0.00
Thyroid (THCA)	MCQ	42.4 / 0.411 ± 1.6 / ± 0.02	33.7 / 0.320 ± 0.6 / ± 0.01	17.1 / 0.168 ± 1.2 / ± 0.01	38.2 / 0.360 ± 2.4 / ± 0.02	**50.3 / 0.485** ± 3.0 / ± 0.03
	Open-ended	0.838 / 0.192 ± 0.00 / ± 0.00	0.804 / 0.145 ± 0.00 / ± 0.00	0.784 / 0.111 ± 0.00 / ± 0.00	0.820 / 0.165 ± 0.00 / ± 0.01	**0.849 / 0.244** ± 0.00 / ± 0.00

Table 3: Open-ended examples comparing GPT-5.4 and fine-tuned OncoVLM. Red spans mark incorrect or weak GPT evidence, green spans mark OncoVLM evidence aligned with the case, and underlined spans mark the corresponding ground-truth evidence.

GPT-5.4	**OncoVLM**	**Ground truth**
Question: Synthesize pathology and genomics.		
...spindle-cell neoplasm with a whorled to storiform architecture, bland cytology, and conspicuous collagenized stroma, without obvious...	...aggressive diffuse glial malignancy, showing necrosis, microvascular proliferation, and marked cellular atypia in a high-grade astrocytic...	...high-grade diffuse glioma with necrosis and microvascular proliferation, a morphologic pattern concordant with glioblastoma. The genomic...
Question: Describe dominant histology.		
...overall low-power appearance is that of a high-grade poorly differentiated urothelial carcinoma, with relatively scant ...	...papillary urothelial neoplasm with complex branching fronds, fibrovascular cores, and crowded epithelial proliferation...	...papillary to nested urothelial neoplasm with broad fused epithelial fronds, complex [...] hyalinized fibrovascular stroma...

Table 4: LLM-as-judge on BRCA open-ended. Win rate is the OncoVLM preference rate over GPT-5.4, ties counted as half win.

Open-ended task	**Win rate**
Radiology description	**0.714**
Pathology description	**0.554**
Cross-modal synthesis	**0.736**
Concordance + uncertainty	**0.805**
Overall	**0.702**

represent racial and ethnic minorities, older adults, and patients from low- and middle-income countries. Inconsistent clinical annotation and uneven modality coverage further reduce per-cancer statistical power and leave room for cohort-level confounding. Radiology introduces additional heterogeneity through variation in scanner protocols, contrast use, resolution, artifacts, DICOM metadata quality, and MRI sequence labels. Because the benchmark uses selected 2D slices rather than harmonized 3D volumes, imaging biomarkers should be viewed as descriptive grounding signals rather than IBSI-standardized radiomic measurements.

Table 5: VQA modality ablation for OncoVLM on cases with pathology and radiology features. MCQ cells report Accuracy% and macro-F1; open-ended cells report BERTScore-F1 and ROUGE-L F1.

	Universal MCQ Accuracy% / F1		**Case-specific MCQ** Accuracy% / F1			**Case specific open-ended** BERT-F1 / ROUGE-L		
Input	**Mutation**	**Grade**	**Path. findings**	**Rad. findings**	**Integrated findings**	**Path. desc.**	**Rad. desc.**	**X-modal synth.**
All	**66.2** **0.660**	**64.0** 0.569	**47.1** **0.460**	32.0 **0.340**	**47.1** **0.463**	**0.861** **0.303**	**0.858** **0.286**	**0.851** **0.257**
Path. only	50.7 0.505	**64.0** **0.617**	39.2 0.371	24.0 0.237	37.3 0.366	0.860 0.295	0.855 0.283	0.848 0.242
Rad. only	55.9 0.559	43.2 0.231	29.4 0.285	**32.9** 0.328	42.5 0.424	0.848 0.259	0.854 0.275	0.842 0.224

7 Conclusion

We introduce OncoTriad-QA, a patient-level radiology-pathology-genomics benchmark for integrated oncology QA, together with OncoVLM, a reference multimodal VLM. Experiments show that OncoVLM improves over general medical and base multimodal baselines, especially on structured clinical tasks, while ablations show different modalities provide complementary signals. Future work will extend the benchmark with volumetric and longitudinal radiology, stronger cross-modal alignment, additional missing-modality settings, and physician-audited adversarial cases.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf`.

[2] Anthropic. Claude Sonnet 4.6. `https://www.anthropic.com/claude/sonnet`, 2025. Accessed: May 2025.

[3] Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.169. URL `https://aclanthology.org/2025.acl-long.169/`.

[4] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, et al. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15016–15027, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01442. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01442`.

[5] Kevin M Boehm, Pegah Khosravi, Rami Vanguri, Jianjiong Gao, and Sohrab P Shah. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer*, 22(2):114–126, October 2021.

[6] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024. URL https://arxiv.org/abs/2406.19280.

[7] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2022. doi: 10.1109/TMI.2020.3021387.

[8] Richard J. Chen, Ming Y. Lu, Drew F.K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, and Faisal Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878.e6, 2022. ISSN 1535-6108. doi: https://doi.org/10.1016/j.ccell.2022.07.004. URL https://www.sciencedirect.com/science/article/pii/S1535610822003178.

[9] Richard J. Chen, Tong Ding, Ming Y. Lu, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, Mar 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://doi.org/10.1038/s41591-024-02857-3.

[10] Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, Sergios Gatidis, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A vision-language foundation model to enhance efficiency of chest x-ray interpretation, 2024. URL https://arxiv.org/abs/2401.12208.

[11] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025. URL https://arxiv.org/abs/2501.17161.

[12] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, December 2013. doi: 10.1007/s10278-013-9622-7.

[13] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, Aug 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

[14] Lucas Paulo et al. de Lima Camillo. Cpgpt: A foundation model for dna methylation. *bioRxiv*, 2024. doi: 10.1101/2024.10.24.619766. URL https://www.biorxiv.org/content/10.1101/2024.10.24.619766v1.

[15] Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Harry Robertson, Bowen Chen, Cristina Almagro-Pérez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Christina S. Chen, Daisuke Komura, Akihiro Kawabe, Mieko Ochi, Shinya Sato, Tomoyuki Yokose, Yohei Miyagi, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11):3749–3761, 2025. ISSN 1546-170X. doi: 10.1038/s41591-025-03982-3. URL https://doi.org/10.1038/s41591-025-03982-3.

[16] Wenjian Ding, Yao Zhang, Jun Wang, and Zhenglu Yang. Rethinking distractor quality in multimodal multiple-choice questions: Automated evaluation and hard benchmark construction. *Computers*, 15(2), 2026. ISSN 2073-431X. doi: 10.3390/computers15020130. URL https://www.mdpi.com/2073-431X/15/2/130.

[17] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar,

Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Mehmet K. Ozdemir, and Bjoern Menze. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arxiv:2403.17834*, 2024.

[18] Lucas Farndale, Robert Insall, and Ke Yuan. Trident : Triple deep network training for privileged knowledge distillation in histopathology. *Medical Image Analysis*, 102:103479, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103479. URL https://www.sciencedirect.com/science/article/pii/S1361841525000271.

[19] Marco Gerlinger, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q. McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R. Santos, Mahrokh Nohadani, Aron C. Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P. Andrew Futreal, and Charles Swanton. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10): 883–892, 2012. doi: 10.1056/NEJMoa1113205. URL https://www.nejm.org/doi/full/10.1056/NEJMoa1113205.

[20] Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4.

[21] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016. doi: 10.1056/NEJMp1607591. URL https://www.nejm.org/doi/full/10.1056/NEJMp1607591.

[22] Arnav Gudibande, Eric Wallace, Charlie Victor Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Kz3yckpCN5.

[23] Feng Guo, Jiaxiang Liu, Yang Li, Qianqian Shi, and Mingkun Xu. Mm-neuroonco: A multimodal benchmark and instruction dataset for mri-based brain tumor diagnosis, 2026. URL https://arxiv.org/abs/2602.22955.

[24] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[25] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9226–9259. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

[26] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention, 2021.

[27] Chongcong Jiang, Tianxingjian Ding, Chuhan Song, Jiachen Tu, Ziyang Yan, Yihua Shao, Zhenyi Wang, Yuzhang Shang, Tianyu Han, and Yu Tian. Medical sam3: A foundation model for universal prompt-driven medical image segmentation. *arXiv preprint arXiv:2601.10880*, 2026. URL https://arxiv.org/abs/2601.10880.

[28] Boming Kang, Rui Fan, Meizheng Yi, Chunmei Cui, and Qinghua Cui. A large-scale foundation model for bulk transcriptomes. *bioRxiv*, 2025. doi: 10.1101/2025.06.11.659222. URL https://www.biorxiv.org/content/early/2025/06/17/2025.06.11.659222.

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

[30] Sijing Li, Zhongwei Qiu, Jiang Liu, Wenqiao Zhang, Tianwei Lin, Yihan Xie, Jianxiang An, Boxiang Yun, Chenglin Yang, Jun Xiao, Guangyu Guo, Jiawen Yao, Wei Liu, Yuan Gao, Ke Yan, Weiwei Cao, Zhilin Zheng, Tony C. W. Mok, Kai Cao, Yu Shi, Jiuyu Zhang, Jian Zhou, Beng Chin Ooi, Yingda Xia, and Ling Zhang. Tumorchain: Interleaved multimodal chain-of-thought reasoning for traceable clinical tumor analysis, 2026. URL https://arxiv.org/abs/2603.05867.

[31] Jana Lipkova, Richard J Chen, Bowen Chen, Ming Y Lu, Matteo Barbieri, Daniel Shao, Anurag J Vaidya, Chengkuan Chen, Luoting Zhuang, Drew F K Williamson, Muhammad Shaban, Tiffany Y Chen, and Faisal Mahmood. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*, 40(10):1095–1110, October 2022.

[32] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, Mar 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02856-4. URL https://doi.org/10.1038/s41591-024-02856-4.

[33] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, May 2012. ISSN 1474-1768. doi: 10.1038/nrc3261. URL https://doi.org/10.1038/nrc3261.

[34] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR, 10 Dec 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

[35] Chuang Niu, Qing Lyu, Christopher D. Carothers, Parisa Kaviani, Josh Tan, Pingkun Yan, Mannudeep K. Kalra, Christopher T. Whitlow, and Ge Wang. Specialty-oriented generalist medical ai for chest ct screening, 2024.

[36] OpenAI. GPT-5.4 (Large Language Model). https://platform.openai.com/, 2026. Accessed: 2026-04-25.

[37] Juampablo E. Heras Rivera, Daniel K. Low, Xavier Xiong, Jacob J. Ruzevick, Daniel D. Child, Wen wai Yim, Mehmet Kurt, and Asma Ben Abacha. Core-bt: A multimodal radiology-pathology-text benchmark for robust brain tumor typing, 2026. URL https://arxiv.org/abs/2603.03618.

[38] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, and other. Medgemma technical report, 2026. URL https://arxiv.org/abs/2507.05201.

[39] Han-Tai Shiao and Vladimir Cherkassky. Learning using privileged information (lupi) for modeling survival data. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 1042–1049, 2014. doi: 10.1109/IJCNN.2014.6889517.

[40] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, Jun 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

[41] Aakash Tripathi, Asim Waqas, Matthew B. Schabath, Yasin Yilmaz, and Ghulam Rasool. Honeybee: enabling scalable multimodal ai in oncology through foundation model-driven embeddings. *npj Digital Medicine*, 8(1):622, Oct 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-02003-4. URL https://doi.org/10.1038/s41746-025-02003-4.

[42] Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H. Song, Tong Ding, Sophia J. Wagner, Ming Y. Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, Richard J. Chen, Dina ElHarouni, Georges Ayoub, Connor Bossi, Keith L. Ligon, Georg Gerber, Long Phi Le, and Faisal Mahmood. Molecular-driven foundation model for oncologic pathology, 2025. URL https://arxiv.org/abs/2501.16652.

[43] Andrew Wang, Ellie Pavlick, and Ritambhara Singh. Handling and interpreting missing modalities in patient clinical trajectories via autoregressive sequence modeling, 2026. URL https://arxiv.org/abs/2604.18753.

[44] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013. doi: 10.1038/ng.2764.

[45] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, Aug 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-62385-7. URL https://doi.org/10.1038/s41467-025-62385-7.

[46] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. A vision–language foundation model for precision oncology. *Nature*, 638(8051): 769–778, Feb 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08378-w. URL https://doi.org/10.1038/s41586-024-08378-w.

[47] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IwgmgidYPS.

[48] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

[49] Yingxue Xu, Yihui Wang, Fengtao Zhou, Jiabo Ma, Cheng Jin, Shu Yang, Jinbang Li, Zhengyu Zhang, Chenglong Zhao, Huajun Zhou, Zhenhui Li, Huangjing Lin, Xin Wang, Jiguang Wang, Anjia Han, Ronald Cheong Kin Chan, Li Liang, Xiuming Zhang, and Hao Chen. A multimodal knowledge-enhanced whole-slide pathology foundation model. *Nature Communications*, 16 (1):11406, Dec 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-66220-x. URL https://doi.org/10.1038/s41467-025-66220-x.

[50] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2418–2428. ACM, August 2022. doi: 10.1145/3534678.3539388. URL http://dx.doi.org/10.1145/3534678.3539388.

[51] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1): AIoa2400640, 2025. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640.

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

[53] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Weiwei Tian, Ya Zhang, Weidi Xie, and Yanfeng Wang. Development of a large-scale grounded vision language dataset for chest ct analysis. *Scientific Data*, 12(1):1636, Oct 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05922-9. URL https://doi.org/10.1038/s41597-025-05922-9.