MEDBENCHAGENT: TOWARDS SYSTEMATIC AUTOMATION OF MEDICAL VLM BENCHMARK CONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale construction of medical vision-language model (VLM) benchmarks is increasingly feasible with richly annotated imaging datasets and large language models (LLMs), yet existing automation largely focuses on generating evaluation items within predefined benchmark specifications. We study the broader problem of automatically deriving the specification itself: *what* to evaluate, *which* annotations support each task, and *how* to translate this evidence into reliable evaluation items. We formulate benchmark construction as *constrained compilation*, in which the benchmark specification is progressively derived from evaluation requirements, heterogeneous annotations, and medical knowledge. Based on this formulation, we introduce **MedBenchAgent**, a multi-agent framework with a **Benchmark Intermediate Representation (BIR)** that encodes task definitions, evidence mappings, evaluation protocols, and item specifications across construction stages. MedBenchAgent separates *planning*, which derives and verifies the specification, from *instantiation*, which constructs and audits items under the locked specification. MedBenchAgent achieves a Task-Space F1 of 90.9%, outperforming direct task induction (79.2–80.0%) and prior-guided induction (85.1%); 994 of 1,000 sampled items from correctly identified tasks pass human audit. We further demonstrate portability to a specialized medical domain and evaluate twelve VLMs, revealing task- and setting-specific variation obscured by aggregate scores. These results establish constrained compilation as a scalable and auditable framework for medical VLM benchmark construction beyond question generation.

1 INTRODUCTION

Large-scale construction of medical vision-language model (VLM) benchmarks is increasingly feasible as richly annotated imaging datasets provide reusable supervision and large language models (LLMs) enable scalable generation of evaluation items (Zhang et al., 2023; Hu et al., 2024; Chen et al., 2024). Yet a benchmark is more than a collection of images and questions: it must specify *what* capabilities are evaluated, *which* source annotations validly support them, *how* annotation evidence is transformed into reference answers, *how* evaluation items are instantiated, and *how* model responses are elicited and assessed. Collectively, we refer to these decisions as the *benchmark specification*, which in turn determines the claims that the resulting evaluation can support. In medical imaging, establishing such a specification requires interpreting heterogeneous annotation schemas, units, terminology, and label semantics, making manual benchmark design time-consuming and expertise-intensive. Automating the specification itself, rather than only large-scale item generation, is therefore an important problem for scalable and reliable medical VLM evaluation.

Existing work approaches automated benchmark construction from two directions. Medical systems construct benchmarks from domain data, but operate within predefined clinical scopes, reporting structures, or question schemas (Xie et al., 2026; Liu et al., 2026a). General-purpose systems offer greater flexibility by using declarative or agentic workflows to translate evaluation requirements into benchmark construction (Li et al., 2025; Xiong et al., 2026; Butt et al., 2025). However, they do not explicitly constrain medical benchmark specification by annotation-supported task validity, reference provenance, and medical knowledge. The remaining challenge, then, is to combine flexible benchmark specification with the semantic and evidential constraints imposed by evaluation requirements and heterogeneous medical data.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We address this challenge with **MedBenchAgent**, a multi-agent framework for systematic medical VLM benchmark construction. We cast benchmark construction as *constrained compilation*, in which the benchmark specification is progressively established from an intended evaluation scope, heterogeneous source annotations, and medical knowledge, with each confirmed decision narrowing the space of subsequent construction. To make these constraints explicit and preserve them across stages, we introduce a **Benchmark Intermediate Representation (BIR)** that records the evaluation scope, task hierarchy, annotation evidence mappings, evaluation protocols, sample allocations, and item-construction specifications. Together, constrained compilation and the BIR provide an explicit mechanism for carrying semantic and evidential constraints from benchmark specification into downstream construction. MedBenchAgent operationalizes this formulation through a staged multi-agent workflow for medical benchmark construction: 1) planning resolves and verifies the specification against dataset annotations and medical references before locking the BIR; and 2) instantiation constructs and audits evaluation items under this fixed specification. In this way, MedBenchAgent extends automation from predefined medical scopes or question structures to benchmark specification itself, while preserving the annotation and medical constraints required for reliable construction.

This formulation enables empirical evaluation of both the benchmark-construction process and the benchmark produced by MedBenchAgent. In construction analysis, MedBenchAgent achieves a Task-Space F1 of 90.9%, compared with 79.2–80.0% for direct task induction and 85.1% when the same task prior is provided to the direct model. For correctly identified tasks, 994 of 1,000 sampled items pass human audit. Together, these results suggest that open-ended decisions about *what should be evaluated* are substantially more challenging than evidence mapping and item construction once benchmark semantics are constrained. We further instantiate the workflow in a specialized medical domain under a different annotation schema and evaluate twelve VLMs on the resulting benchmark. Model evaluation reveals substantial variation across tasks, imaging settings, and evaluation metrics that aggregate scores can obscure, reinforcing fine-grained benchmark specification as an important component of medical VLM evaluation.

Overall, this work makes three contributions: (1) **Conceptually**, we model the construction of medical VLM benchmarks and formulate it as constrained compilation; (2) **Methodologically**, we introduce a multi-agent MedBenchAgent framework and its Benchmark Intermediate Representation (BIR); and (3) **Empirically**, we derive insights into both the benchmark-construction process and medical VLM evaluation, while providing reusable benchmark artifacts.

2 Related Work

Medical VLM benchmarks. Medical VLM benchmarks have progressed from manually authored question–answer datasets such as VQA-RAD and SLAKE (Lau et al., 2018; Liu et al., 2021), to scalable reuse of image–caption and classification supervision in PMC-VQA and OmniMedVQA (Zhang et al., 2023; Hu et al., 2024), and more recently to task-structured evaluation over heterogeneous datasets in GMAI-MMBench and MMBU (Chen et al., 2024; D'Cunha et al., 2026).

Automated benchmark construction. General-purpose benchmark automation has increasingly adopted declarative or agentic construction, whereas medical systems have favored domain-specific, more deterministic workflows. AutoBencher constructs benchmarks from specified evaluation desiderata, BenchmarkAgent translates high-level evaluation requirements into construction procedures, and BenchAgents coordinates multiple agents across benchmark-creation stages (Li et al., 2025; Xiong et al., 2026; Butt et al., 2025). In the medical domain, EHRBench uses a task-specific template pipeline to instantiate clinical QA from EHR data (Xie et al., 2026), while Liu et al. (2026a) packages schema- and report-grounded oncology VQA generation as a reusable agent skill.

3 What Makes a Good Medical VLM Benchmark?

3.1 Benchmark Formulation

A medical VLM benchmark must specify both what is evaluated and how model performance is assessed. Building on medical benchmarking frameworks that distinguish the evaluation objective, challenge data, reference annotations, and assessment procedures (Karargyris et al., 2023; Maier-Hein

et al., 2020), and following recent task-oriented medical VLM benchmarks, we additionally specify the task definitions and queries presented to the model for vision–language evaluation. We represent a benchmark by its intended evaluation scope $\mathcal{U}$, task space $\mathcal{T}$, evaluation-item set $\mathcal{I}$, and evaluation protocol $\mathcal{P}$. A medical VLM benchmark is therefore defined as

$$\mathcal{B} = (\mathcal{U}, \mathcal{T}, \mathcal{I}, \mathcal{P}), \qquad \mathcal{I} = \{\iota_i = (x_i, t_i, q_i, r_i)\}_{i=1}^{N}, \tag{1}$$

where x_i is a medical imaging case, $t_i \in \mathcal{T}$ the evaluated task, q_i the model-facing query, and r_i the benchmark reference; $\mathcal{P}$ specifies each task's response convention and scoring rule.

Because each r_i is a benchmark-level reference, not a raw annotation copied directly from a source dataset, its provenance and semantics must be recorded. The CLAIM [1] guideline requires medical imaging studies to document the source, definition, and acquisition procedure of the reference standard (Tejani et al., 2024). We represent the source datasets as

$$\mathcal{D} = \{\mathcal{D}_m\}_{m=1}^{M}, \qquad \mathcal{D}_m = (\mathcal{X}_m, \mathcal{A}_m, \Sigma_m), \tag{2}$$

where $\mathcal{X}_m$ is the set of imaging cases in dataset m, $\mathcal{A}_m$ is its annotation collection, and Σ_m specifies the annotation schema, annotation unit, granularity, label space, and field semantics. An annotation record $a \in \mathcal{A}_m$ represents evidence encoded under the dataset-specific specification Σ_m; it becomes a benchmark reference r_i only after its semantics have been matched to a task t_i and mapped into the common benchmark representation.

Benchmark construction must therefore determine not only whether the available annotations support a proposed task, but also how the relevant annotation evidence should be interpreted as its reference answer. We therefore abstract benchmark construction as

$$F : (\mathcal{D}, \mathcal{K}, \mathcal{U}) \mapsto \mathcal{B}, \tag{3}$$

where $\mathcal{K}$ denotes the medical knowledge required to interpret clinical concepts and evaluation semantics. Once constructed, the benchmark can be applied to a model set $\mathcal{M}$ to produce empirical evidence $\mathcal{E}$: $(\mathcal{B}, \mathcal{M}) \longrightarrow \mathcal{E}$. This separates benchmark construction, which defines the evaluation object, from benchmark application, which produces evidence about model behavior under the specified scope and protocol. We use *benchmark specification* to denote the pre-instantiation decisions that determine the evaluation scope, dataset-specific task definitions and granularity, supporting evidence and reference construction, evaluation protocol, sample allocation, and item-construction rules. The concrete item set $\mathcal{I}$ is produced by instantiating this specification over the source cases.

3.2 Desiderata for Benchmark Construction

The formulation above motivates three desiderata for medical VLM benchmark construction.

D1: Evidence Groundedness. Because benchmark construction maps source annotations to evaluation tasks and reference answers, annotation availability alone is insufficient to justify either output. Evidence groundedness requires that each task be supported by the schema and semantics of its source annotations and that each item-level reference r_i be entailed by the annotation evidence used to construct it. SLAKE uses physician-provided annotations, while EHRXQA uses templates and executable queries (Liu et al., 2021; Bae et al., 2023).

D2: Medical Alignment. Annotation support establishes that a task is constructible, but does not determine how it should be expressed under the intended medical evaluation scope. Medical alignment requires task definitions, response spaces, and queries presented to the model to be consistent with established medical concepts and terminology, and not merely to verbalize labels to one dataset. GMAI-MMBench reflects this desideratum by organizing evaluation around clinically defined tasks (Chen et al., 2024).

D3: Evaluative Utility. The application $(\mathcal{B}, \mathcal{M}) \rightarrow \mathcal{E}$ further implies that evidence-grounded and medically aligned items do not necessarily yield informative evidence about model behavior. Evaluative utility requires the benchmark to support separate assessment of model capabilities and preserve interpretable differences across tasks and imaging settings. OmniMedVQA, MultiMedEval, and GMAI-MMBench exemplify this desideratum through evaluation across multiple datasets, tasks, modalities, and model families (Hu et al., 2024; Royer et al., 2024; Chen et al., 2024).

[1] Checklist for Artificial Intelligence in Medical Imaging

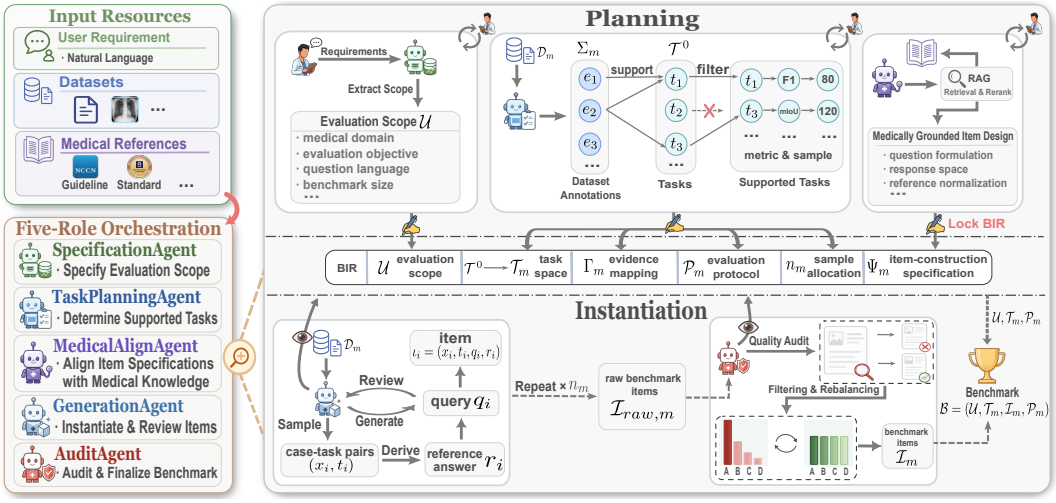


Figure 1: **Overview of MedBenchAgent.** MedBenchAgent coordinates five agents through a shared Benchmark Intermediate Representation (BIR). Planning specifies tasks and their constraints within the BIR, whereas instantiation constructs evaluation items under the locked specifications.

3.3 Benchmark Construction as Constrained Compilation

We therefore formulate benchmark construction as constrained compilation: planning progressively establishes the benchmark specification, and each confirmed decision restricts the admissible choices of subsequent stages within the mapping F. The evaluation scope constrains the relevant task categories; annotation semantics determine which concrete prediction targets can be instantiated within them and which evidence can support their references; medical knowledge and evaluation requirements further constrain task expression and assessment. This process requires an explicit intermediate state that preserves confirmed decisions across stages. Section 4 instantiates this state as the BIR and locks it before item construction.

4 MedBenchAgent

Figure 1 summarizes the five-agent workflow and the separation between planning and instantiation, while Algorithm 1 formalizes the corresponding BIR updates and stage-wise construction procedure.

4.1 Inputs and Benchmark Intermediate Representation

MedBenchAgent takes three benchmark-specific resources as input: a natural-language **user requirement**, **medical imaging datasets**, and **medical reference materials**. Each dataset-specific BIR contains a task-category prior $\mathcal{T}_0$, represented as an unexpanded task tree. The user requirement u defines the intended evaluation scope $\mathcal{U}$ and specifies additional user constraints on benchmark construction. The source collection $\mathcal{D} = \{\mathcal{D}_m\}_{m=1}^{M}$ consists of medical imaging cases and their dataset-specific annotations. The medical knowledge resource $\mathcal{K}$ consists of clinical guidelines, reporting standards, and other domain references relevant to the target setting. The initial task definitions in $\mathcal{T}_0$ are distilled from the task-design principles of existing medical VLM benchmarks and specify candidate tasks and their annotation-support conditions.

Section 3 formulates benchmark construction as $F : (\mathcal{D}, \mathcal{K}, \mathcal{U}) \mapsto \mathcal{B}$. MedBenchAgent implements this transformation through two coupled stages:

$$\mathcal{Z} = F_{\text{plan}}(\mathcal{D}, \mathcal{K}, \mathcal{U}; \mathcal{T}_0), \qquad \mathcal{I} = F_{\text{inst}}(\mathcal{D}, \mathcal{Z}), \tag{4}$$

where $\mathcal{Z}$ denotes the benchmark design produced during planning.

For each source dataset $\mathcal{D}_m$, planning maintains a dataset-specific **Benchmark Intermediate Representation (BIR)**. After the evaluation scope is confirmed, its task-space component is initialized

as the unexpanded task tree $\mathcal{T}_0$. Candidate nodes are retained, removed, or expanded according to their support conditions, the evaluation scope, and available annotation evidence. Planning completes their evidence mappings, evaluation protocols, and sample allocations, while task design supplies the item-construction specifications. The resulting state is then confirmed and locked:

$$\mathrm{BIR}_m^{(0)} = (\mathcal{U}, \mathcal{T}_0, 0, 0, 0, 0)\,, \qquad \mathrm{BIR}_m^{\mathrm{lock}} = (\mathcal{U}, \mathcal{T}_m, \Gamma_m, \mathcal{P}_m, n_m, \Psi_m)\,. \tag{5}$$

Here, $\mathcal{T}_m$ is the dataset-specific task tree obtained by retaining supported nodes and expanding decomposable tasks into distinct prediction targets. Ordinary tasks remain leaf nodes, whereas decomposable tasks have question types as their children. Each child must remain within its parent's task semantics and have annotation support for its own prediction target. The middle four components jointly describe this tree and its node specifications: $\Gamma_m(t)$ records supporting annotation evidence, $\mathcal{P}_m(t)$ specifies the evaluation protocol, and $n_m(t)$ denotes the sample allocation associated with task node t. These specifications are completed for the retained atomic tasks during planning; the zero entries in the initial BIR denote components not yet populated for the dataset. $\Psi_m(t)$ contains the corresponding medically grounded item-construction specification. We write $\mathrm{Leaves}(\mathcal{T}_m)$ for the retained atomic evaluation tasks, so a decomposable parent is not counted in addition to its children.

The planning output is therefore $\mathcal{Z} = \{\mathrm{BIR}_m^{\mathrm{lock}}\}_{m=1}^{M}$. Each $\mathrm{BIR}_m^{\mathrm{lock}}$ encodes the confirmed dataset-level benchmark specification, while instantiation produces the corresponding audited item set $\mathcal{I}_m$. Dataset-specific leaf tasks, item sets, and protocols are then consolidated across datasets, with $\mathcal{T} = \bigcup_{m=1}^{M} \mathrm{Leaves}(\mathcal{T}_m)$, to form the final benchmark $\mathcal{B} = (\mathcal{U}, \mathcal{T}, \mathcal{I}, \mathcal{P})$.

4.2 Planning: Compiling and Locking the BIR

Planning implements F_{plan} via SpecificationAgent, TaskPlanningAgent, and MedicalAlignAgent, which structure requirements, determine supported tasks, and design task items, respectively. The *CKPT* markers in Algorithm 1 denote critical verification checkpoints at consequential planning transitions, validating the proposed BIR state before commitment through deterministic checks, human review, model-based supervision, or hybrid mechanisms.

SpecificationAgent first converts the open-ended natural-language requirement u into a structured evaluation scope $\mathcal{U}$. The specification resolves ambiguous dimensions of the requested benchmark before constraining subsequent planning.

Algorithm 1 Medical VLM benchmark construction

Require: User requirement u, source datasets $\mathcal{D}$, medical references $\mathcal{K}$
Ensure: Benchmark $\mathcal{B} = (\mathcal{U}, \mathcal{T}, \mathcal{I}, \mathcal{P})$

$\mathcal{U} \xleftarrow{\text{CKPT}} \textsc{Specify}(u)$; $\{\mathrm{BIR}_m^{(0)} = (\mathcal{U}, \mathcal{T}_0, 0, 0, 0, 0)\}_{m=1}^{M}$; $\mathcal{I}, \mathcal{I}_{\mathrm{raw},m} \leftarrow \emptyset$, $\forall m$
for all $\mathcal{D}_m \in \mathcal{D}$ **do**
  $(\mathcal{T}_m, \Gamma_m, \mathcal{P}_m, n_m) \xleftarrow{\text{CKPT}} \textsc{AssessTaskSupport}(\mathcal{D}_m, \mathrm{BIR}_m^{(0)})$
  $\mathrm{BIR}_m \leftarrow (\mathcal{U}, \mathcal{T}_m, \Gamma_m, \mathcal{P}_m, n_m, 0)$
  $\Psi_m \xleftarrow{\text{CKPT}} \textsc{DesignTaskItems}(\mathcal{K}, \mathcal{U}, \mathcal{T}_m, \Gamma_m, \mathcal{P}_m)$
  $\mathrm{BIR}_m^{\mathrm{lock}} \leftarrow (\mathcal{U}, \mathcal{T}_m, \Gamma_m, \mathcal{P}_m, n_m, \Psi_m)$
  for all eligible case–task pairs (x_i, t) with $t \in \mathrm{Leaves}(\mathcal{T}_m)$ **do**
    $e_i \leftarrow \textsc{ExtractEvidence}(x_i, \Gamma_m(t))$
    $r_i \leftarrow \textsc{DeriveReference}(e_i, \Psi_m(t))$
    $q_i \leftarrow \textsc{GenerateQuery}(x_i, t, e_i, \Psi_m(t))$
    $\iota_i \leftarrow (x_i, t, q_i, r_i)$
    if $\textsc{ReviewItem}(\iota_i, e_i, \mathrm{BIR}_m^{\mathrm{lock}})$ **then** $\mathcal{I}_{\mathrm{raw},m} \leftarrow \mathcal{I}_{\mathrm{raw},m} \cup \{\iota_i\}$
  end for
  $\mathcal{I}_m \leftarrow \textsc{Audit}(\mathcal{I}_{\mathrm{raw},m}, \mathrm{BIR}_m^{\mathrm{lock}}, \mathcal{D}_m)$
  $\mathcal{I} \leftarrow \mathcal{I} \cup \mathcal{I}_m$
end for
$\mathcal{T} \leftarrow \bigcup_{m=1}^{M} \mathrm{Leaves}(\mathcal{T}_m)$; $\mathcal{P} \leftarrow \textsc{MergeProtocols}(\{\mathcal{P}_m\}_{m=1}^{M})$
return $\mathcal{B} = (\mathcal{U}, \mathcal{T}, \mathcal{I}, \mathcal{P})$

TaskPlanningAgent then extends the task-category prior into the dataset-specific task tree $\mathcal{T}_m$ using the evaluation scope and annotation evidence in $\mathcal{D}_m$. A category is retained only when its support conditions are satisfied, and Γ_m records the eligible annotation fields to construct references for its tasks. When annotations expose multiple distinct prediction targets within a supported task, the agent expands it into child tasks, each representing an annotation-supported answer axis within parent semantics. Semantic duplicates are removed across retained atomic tasks, establishing the task support required by Desideratum D1. TaskPlanningAgent selects a compatible primary evaluation metric for each retained atomic task. The resulting task plan passes through CKPT before commitment to the BIR, after which deterministic allocation assigns n_m. Preserving prediction targets provides the basis for separate task-level analysis under Desideratum D3.

MedicalAlignAgent subsequently converts each retained task into a medically meaningful and executable item-construction specification Ψ_m. It adopts retrieval-augmented generation (RAG) (Lewis et al., 2020) over the relevant medical references in $\mathcal{K}$ to ground the design in medical knowledge.

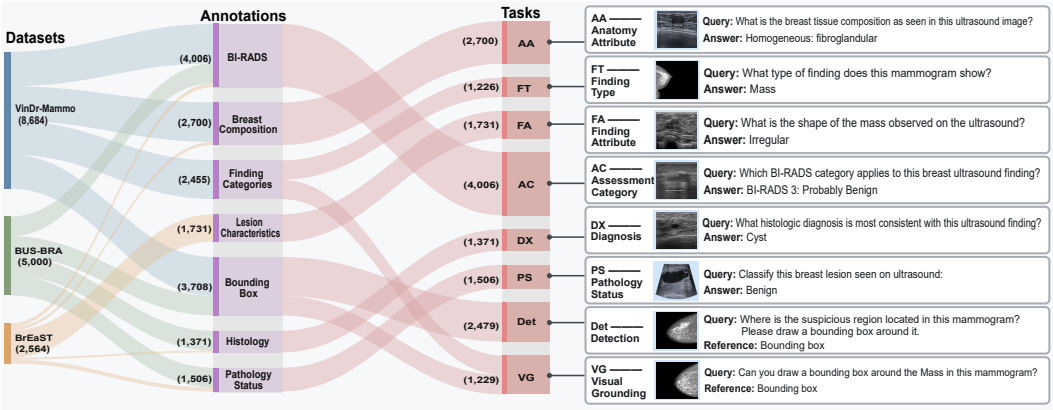


Figure 2: **Benchmark construction and representative items.** Three source datasets are mapped through heterogeneous annotations to the retained evaluation tasks, with one representative item shown for each task. Numbers in parentheses denote item counts.

Grounded on these references and constrained by the retained task semantics and authorized annotation evidence, it specifies the task's question semantics, response space, and reusable question formulations, together with the normalization to benchmark-level reference answers. This step connects an evidence-supported task to its medical evaluation semantics, thereby supporting Desideratum D2. After the complete set of task designs passes through CKPT, the resulting $\mathrm{BIR}_m^{\mathrm{lock}}$ is fixed for downstream instantiation. Downstream stages may read the confirmed task definitions, supporting evidence, protocols, sample allocations, and construction specifications, but cannot revise them.

4.3 Instantiation: Constructing Items under the Locked BIR

Instantiation realizes F_{inst} in Eq. 4 by applying $\mathrm{BIR}_m^{\mathrm{lock}}$ to the source annotations without reopening the confirmed benchmark design. As summarized in Algorithm 1, GenerationAgent operates at the case–task level to instantiate evaluation items, whereas AuditAgent operates over the resulting item set to verify benchmark-level consistency.

For each eligible case–task pair with $t_i \in \mathrm{Leaves}(\mathcal{T}_m)$, **GenerationAgent** constructs an evaluation item $\iota_i = (x_i, t_i, q_i, r_i)$ under the corresponding task specification $\Psi_m(t_i)$. The model generates only the query q_i, varying its wording within the locked task semantics and authorized annotation evidence. Deterministic code derives the reference answer r_i from the supporting evidence specified by $\Gamma_m(t_i)$ and, where applicable, organizes the response options according to the confirmed construction specification. A subsequent model review call screens the generated item for task drift, ambiguity, unsupported additions, and answer leakage without modifying the fixed reference. This separation allows linguistic variation in the query while preserving annotation-grounded reference answers and the item-level entailment requirement of Desideratum D1.

AuditAgent subsequently checks the generated item set for consistency with the locked task-node specifications, source annotations, and benchmark-level requirements. Deterministic checks verify annotation support, response and reference consistency, and exact duplication, while accepted multiple-choice items are balanced with respect to option positions. AuditAgent additionally performs task-stratified consistency screening to identify issues such as task drift, answer leakage, ambiguity, and semantic duplication. The items passing the deterministic audit constitute the finalized evaluation-item set $\mathcal{I}$ and complete the resulting benchmark $\mathcal{B} = (\mathcal{U}, \mathcal{T}, \mathcal{I}, \mathcal{P})$.

5 Experiments: Benchmark Construction and Evaluation

5.1 Experimental Setup

Construction. Figure 2 summarizes the resulting benchmark construction. We instantiate MedBenchAgent using three medical imaging datasets: VinDr-Mammo (Nguyen et al., 2023), BrEaST (Pawłowska et al., 2024), and BUS-BRA (Gómez-Flores et al., 2024). The medical knowledge

Table 1: **Task-space construction and evidence-mapping results.** Task-Space Precision (P), Recall (R), and F1 measure agreement with researcher review; EMA measures agreement of supporting annotation fields (%).

Model	Task-Space Review						Evidence
	TP	FP	FN	P↑	R↑	F1↑	EMA↑
DeepSeek-V4-Flash	19	8	2	70.4	90.5	79.2	100.0
Qwen3.5-9B	19	8	2	70.4	90.5	79.2	100.0
MedGemma-27B	20	9	1	69.0	95.2	80.0	100.0
DeepSeek-V4-Flash + $\mathcal{T}_0$	20	6	1	76.9	95.2	85.1	100.0
MedBenchAgent	**20**	**3**	**1**	**87.0**	**95.2**	**90.9**	**100.0**

Table 2: **Human consistency audit of finalized items.** Pass rates are computed over 1,000 items sampled uniformly from the finalized benchmark.

Criterion	Pass	Fail	Pass Rate
Reference groundedness	1,000	0	100.0%
Task–query alignment	996	4	99.6%
Unambiguity	1,000	0	100.0%
Evaluation integrity	994	6	99.4%
All criteria	**994**	**6**	**99.4%**

resource is grounded in the relevant sections of the ACR BI-RADS v2025 Manual (Destounis et al., 2025; Leung et al., 2025). For the reported runs, we instantiate the planning checkpoints with researcher review.

Evaluation. We evaluate twelve VLMs spanning general-purpose and medical-domain model families, multiple parameter scales, and different inference regimes. The general-purpose models include InternVL3-8B (Zhu et al., 2025), Qwen2.5-VL (7B, 32B) (Bai et al., 2025), and Qwen3-VL (8B, 32B), including an 8B thinking variant (Qwen Team, 2025). The medical models include MedGemma-4B (Google Research & Google DeepMind, 2025), Lingshu (7B, 32B) (LASA Team et al., 2025), Hulu-Med-7B (Jiang et al., 2025), QoQ-Med-7B (Dai et al., 2025), and BreastGPT-8B (Liu et al., 2026b). Multiple-choice tasks are evaluated with Accuracy (Acc.) and Macro-F1 (F1), while spatial tasks use mean intersection over union (mIoU) and $\mathrm{Acc@IoU}_{0.5}$. When a task contains multiple fine-grained tasks, metrics are computed separately and macro-averaged with equal weight.

Implementation. We host DeepSeek-V4-Flash (DeepSeek-AI, 2026) as the language-model backend with vLLM (Kwon et al., 2023), and Qwen3-Embedding-0.6B and Qwen3-Reranker-0.6B (Zhang et al., 2025) for retrieval and reranking. Model evaluation uses deterministic decoding and FlashAttention-2 (Dao, 2024), and all experiments are run on NVIDIA B200 GPUs.

5.2 Benchmark Construction Analysis

Task-space construction. MedBenchAgent uses BIR-mediated task planning to constrain task-space construction with $\mathcal{U}$ and Σ_m, rather than inducing the task space in a single open-ended step. We evaluate this strategy against direct baselines using the same inputs in a single model call, and additionally provide DeepSeek-V4-Flash with the initial task definitions $\mathcal{T}_0$. The direct baselines use MedGemma-27B (Google Research & Google DeepMind, 2025), Qwen3.5-9B (Qwen Team, 2026), and DeepSeek-V4-Flash. We evaluate against a reference task space $\mathcal{T}^\star$, defined before MedBenchAgent and finalized prior to these experiments by dataset-familiar researchers with informal input from a clinical collaborator. Proposals are matched to $\mathcal{T}^\star$ by prediction target and annotation semantics rather than surface wording. We report Task-Space Precision, Recall, and F1 from TP, FP, and FN counts; for matched tasks, Evidence-Mapping Agreement (EMA) measures agreement with researcher-confirmed supporting fields. As shown in Table 1, BIR-mediated construction achieves 90.9% Task-Space F1, compared with 79.2–80.0% for direct induction. Providing $\mathcal{T}_0$ improves direct DeepSeek-V4-Flash induction to 85.1% F1, while MedBenchAgent further increases precision from 76.9% to 87.0% at the same 95.2% recall, reducing false positives from six to three. All methods achieve 100% EMA on matched tasks. These results show that the BIR-mediated design primarily improves the selection of valid prediction targets, while supporting-field identification is reliable once the task is correctly identified.

Human audit of finalized items. MedBenchAgent promotes consistent item construction by locking item-construction requirements in the confirmed specification before instantiation. To assess whether these constraints are preserved in the finalized benchmark, we sample 1,000 items without replacement and audit each for reference groundedness, task–query alignment, unambiguity, and evaluation integrity against its annotation evidence and specification. An item passes only when all four criteria are satisfied. Table 2 shows that 994 items pass, yielding a 99.4% consistency rate. Six failures include four VinDr-Mammo finding-type questions that ask for BI-RADS assessment categories and two BrEaST diagnosis questions with unsupported additions. These cases localize remaining errors

Table 3: **Evaluation performance on tasks with choices.** Results are reported as **Accuracy** / **Macro-F1** (%). † denotes thinking variants. Task abbreviations (AA, FT, etc.) are defined in Figure 2.

Model	Anatomy	Finding		Assessment & Diagnosis			Overall
	AA	FT	FA	AC	DX	PS	
General-purpose VLMs							
InternVL3-8B	36.59 / 17.30	19.09 / 8.65	42.28 / 30.86	16.12 / 12.82	35.35 / 22.86	48.95 / 40.25	33.06 / 22.12
Qwen2.5-VL-7B	48.86 / 23.08	29.77 / 10.17	41.45 / 29.28	21.35 / 11.93	33.76 / 22.67	51.03 / 41.86	37.70 / 23.16
Qwen3-VL-8B-I	34.89 / 18.56	16.31 / 9.03	29.84 / 22.57	39.11 / 19.74	39.98 / 29.65	60.55 / 47.85	36.78 / 24.57
Qwen3-VL-8B-T†	25.52 / 14.18	15.09 / 7.29	56.93 / 35.13	38.58 / 20.03	35.13 / 24.15	53.05 / 42.52	37.38 / 23.88
Qwen2.5-VL-32B	44.68 / 21.59	54.00 / 12.67	46.20 / 29.55	26.83 / 16.79	43.26 / 26.69	52.92 / 42.41	44.65 / 24.95
Qwen3-VL-32B	37.31 / 20.47	**54.65** / 13.25	37.08 / 28.42	34.22 / 20.59	39.37 / 29.83	59.37 / 54.28	43.67 / 27.81
Medical VLMs							
MedGemma-4B	22.39 / 15.69	23.16 / 10.84	47.55 / 33.94	17.98 / 15.38	30.20 / 23.75	55.78 / 44.32	32.84 / 23.99
Lingshu-7B	44.27 / 26.13	32.22 / 12.28	56.84 / 38.66	25.28 / 20.85	52.18 / 29.29	71.21 / 52.53	47.00 / 29.96
Hulu-Med-7B	31.62 / 16.63	17.05 / 9.19	51.14 / 35.49	22.75 / 18.81	49.64 / 34.70	64.14 / 51.59	39.39 / 27.73
QoQ-Med-7B†	48.56 / 22.72	53.83 / 12.03	66.09 / 37.85	20.43 / 10.24	41.45 / 24.07	42.14 / 30.44	45.42 / 22.89
BreastGPT-8B	36.03 / **26.41**	35.73 / **16.96**	**75.22** / **52.50**	**40.72** / **31.54**	**65.09** / **47.06**	**78.58** / **71.30**	**55.23** / **40.96**
Lingshu-32B	**50.89** / 24.70	42.66 / 14.17	57.54 / 43.01	34.19 / 21.65	49.04 / 24.54	70.08 / 55.25	50.73 / 30.55
Family-level Summary							
General-purpose Avg.	37.98 / 19.19	31.49 / 10.18	42.30 / 29.30	29.37 / 16.98	37.81 / 25.97	54.31 / 44.86	38.87 / 24.41
Medical Avg.	38.96 / 22.04	34.11 / 12.58	59.06 / 40.24	26.89 / 19.74	47.93 / 30.57	63.65 / 50.90	45.10 / 29.35

to template semantics and evidence control during generation, distinguishing item-expression errors from the task-specification errors identified in the task-space analysis. The high consistency rate supports locked specifications for large-scale item construction while identifying targets for semantic verification. Detailed setups and additional analyses are provided in Appendix C.

5.3 Model Evaluation

Task-oriented specification preserves non-redundant evaluation evidence. MedBenchAgent retains annotation-supported prediction targets as separate tasks and records their evaluation protocols in the BIR. We test whether these choices affect model assessment beyond increasing benchmark granularity. As shown in Table 3, medical VLMs outperform general-purpose VLMs by 6.23 points in Overall Acc. but trail by 2.48 points on AC, indicating that the aggregate advantage is not uniform across tasks. To quantify such changes, we compare all 66 model pairs and find that 12–24 pairs (18.2–36.4%) reverse ordering between Overall Acc. and individual-task Acc.; within tasks, 4–15 pairs (6.1–22.7%) reverse between Acc. and F1.

Table 4: **Spatial localization performance.** Results are reported as mIoU / Acc@IoU$_{0.5}$ (%).

Model	VinDr-Mammo		BUS-BRA
	Det	VG	Det
General-purpose VLMs			
InternVL3-8B	3.30 / 0.24	3.15 / 0.33	26.00 / 11.20
Qwen2.5-VL-7B	7.60 / 3.91	8.77 / 6.51	28.94 / 22.24
Qwen3-VL-8B-I	10.72 / 7.00	13.68 / 10.98	51.11 / 56.40
Qwen3-VL-8B-T†	10.94 / 7.32	10.42 / 7.00	37.50 / 32.88
Qwen2.5-VL-32B	9.70 / 6.10	10.83 / 7.32	40.63 / 37.52
Qwen3-VL-32B	**15.40** / **12.86**	**18.01** / **16.84**	55.36 / 62.08
Medical VLMs			
MedGemma-4B	1.54 / 0.00	1.57 / 0.00	15.73 / 3.68
Lingshu-7B	4.41 / 0.81	3.65 / 0.65	4.71 / 0.88
Hulu-Med-7B	0.01 / 0.00	0.00 / 0.00	6.97 / 2.96
QoQ-Med-7B†	6.97 / 1.06	6.97 / 1.30	45.63 / 43.52
BreastGPT-8B	5.34 / 0.90	5.44 / 1.14	**72.62** / **82.80**
Lingshu-32B	4.70 / 1.46	5.62 / 1.95	12.46 / 4.40

MedBenchAgent also preserves localization annotations as spatial prediction targets rather than converting them into choice-based questions. For example, Lingshu-7B has higher Overall Acc. than Qwen2.5-VL-32B (47.00% vs. 44.65%), while Qwen2.5-VL-32B performs better on all six spatial measures in Table 4. Across these measures, 28–34 of 66 model pairs (42.4–51.5%) reverse ordering relative to Overall Acc. Thus, rankings vary substantially across evaluation forms. These results support Desideratum D3: preserving annotation-supported tasks across capability dimensions and assigning task-appropriate metrics materially changes the evidence provided by the benchmark, motivating decomposition of aggregate results by prediction target to reveal capability differences obscured by Overall summaries. Additional results are provided in Appendix E.

Annotation grounding provides a model-independent reference layer. MedBenchAgent treats the annotations associated with each source image as the reference standard. This design places

Table 5: **Multiple-choice performance under the blurred-image intervention.** Results are reported as **Acc.** / **F1** (%). Δ Overall denotes the original-to-blurred score drop (pp)

Model	Anatomy	Finding		Assessment & Diagnosis			Overall	Δ Overall
	AA	FT	FA	AC	DX	PS	Acc/F1	Acc/F1
Random Guess	25.00 / 20.64	12.50 / 8.02	36.90 / 30.95	19.76 / 16.84	25.00 / 22.39	41.67 / 37.83	26.81 / 22.78	–
MedGemma-4B	11.56 / 11.58	8.24 / 5.88	39.06 / 27.99	1.83 / 3.38	21.06 / 16.73	40.44 / 36.59	20.36 / 17.02	12.48 / 6.97
Lingshu-7B	26.78 / 19.60	16.48 / 7.03	55.03 / 27.96	32.85 / 16.08	35.98 / 18.74	61.92 / 32.28	38.17 / 20.28	8.83 / 9.68
Hulu-Med-7B	19.06 / 16.62	5.71 / 4.69	54.83 / 28.07	31.07 / 15.89	38.27 / 22.01	52.23 / 41.12	33.53 / 21.40	5.86 / 6.33
BreastGPT-8B	36.66 / 18.00	42.01 / 11.48	58.54 / 28.81	20.12 / 11.37	30.70 / 18.49	35.55 / 21.97	37.26 / 18.35	17.97 / 22.61

trust in clinician annotations rather than external model judgments, avoiding an additional source of uncertainty. To test this strategy, we replace every evaluation image with a Gaussian-blurred image while keeping all non-image benchmark components unchanged. Across four medical VLMs, Acc. decreases by 5.86–17.97 percentage points (pp) and F1 by 6.33–22.61 pp (Table 5). The consistent degradation indicates that benchmark performance remains sensitive to visual evidence from the corresponding image, despite references being derived directly from source annotations rather than model-based visual judgments.

5.4 Cross-Domain Case Study

To assess the cross-domain portability of MedBenchAgent, we apply the complete construction workflow under a specialized medical setting using ISIC 2018 (Codella et al., 2019; Tschandl et al., 2018), with a distinct dataset and annotation schema.

The resulting task space and representative VLM results are summarized in Figure 3. Representative VLMs show different relative strengths across prediction targets rather than a uniform performance pattern. This indicates that the same workflow can instantiate a domain-appropriate task structure under a distinct annotation schema These results provide cross-domain evidence for the portability of MedBenchAgent across annotation semantics and task definitions. Additional details are provided in Appendix D.

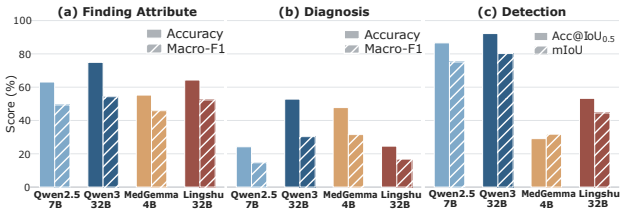


Figure 3: **Cross-domain instantiation on dermoscopy.**

6 Conclusion

We introduced MedBenchAgent, a multi-agent framework that treats medical VLM benchmark construction as constrained compilation and makes benchmark specification an explicit, auditable object of construction. Through the Benchmark Intermediate Representation (BIR), MedBenchAgent separates specification planning from downstream item instantiation, allowing construction decisions to be evaluated stage by stage rather than only through the final benchmark. Our analyses identify open-ended task-space specification as the primary source of construction uncertainty, while annotation-evidence mapping and item instantiation become substantially more reliable once benchmark semantics are fixed. Evaluation of the resulting benchmark further reveals task-, metric-, and imaging-dependent variation across VLMs that aggregate scores can obscure. Together with the cross-domain instantiation under a specialized medical setting, these results show that constrained, multi-agent benchmark construction can support both scalable medical VLM evaluation and systematic study of the benchmark-construction process itself.

Limitations and Future Directions. While MedBenchAgent moves medical benchmark construction towards greater automation, limited human involvement remains at planning checkpoints; we will further reduce this reliance by introducing supervisory agents to verify intermediate specifications. More broadly, MedBenchAgent provides a basis for comparing controlled benchmark compositions and for constructing larger collections of disease-, anatomy-, and modality-specific benchmarks.

AI USE STATEMENT

Generative AI was used as part of the research methodology for benchmark construction, including task planning, evaluation-item generation and review and data preprocessing. Generative AI also assisted with code development, language editing, and polishing; the authors reviewed and validated all AI-assisted code and manuscript changes and take responsibility for the final text, claims, code, analyses, and artifacts produced with generative-AI assistance.

ETHICS STATEMENT

This work constructs research benchmarks from publicly released medical-imaging datasets and their existing annotations, without collecting new patient data. The benchmarks are intended for research evaluation. Their reference standards, population coverage, and potential biases follow the source datasets and annotation protocols. Use and redistribution are subject to the licenses, privacy requirements, and usage conditions of the original datasets and medical reference materials.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the benchmark-construction and evaluation procedures throughout the main text, appendices, and released code. Section 5.1 specifies the source datasets, evaluated model families, metrics, and compute environment. Appendix A documents dataset normalization and profiling, planning checkpoints, LLM, retrieval, and generation settings, a concrete BIR compilation example, and resource usage; Tables 6–8 report the effective configurations and trace-derived resource statistics. The released code contains the complete initialized BIR template and task prior, decision-critical prompts, serialization and validation logic, and the normalized annotation files used in the reported breast-imaging runs. Appendix C provides the direct-baseline setup, task-space matching and aggregation rules, human-audit sampling and decision criteria, and residual-error analysis; the complete baseline prompts and raw task-planning outputs are provided with the released code. Appendix D documents the ISIC cross-domain task allocation and consistency-check procedure, and Appendix E reports the complete model-evaluation results.

REFERENCES

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3867–3880. Curran Associates, Inc., 2023. doi: 10.52202/075280-0170.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923.

Natasha Butt, Varun Chandrasekaran, Neel Joshi, Besmira Nushi, and Vidhisha Balachandran. Benchagents: Automated benchmark creation with agent interaction. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427, 2024.

Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1902.03368, 2019. URL http://arxiv.org/abs/1902.03368.

David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 37406–37453. Curran Associates, Inc., 2025. doi: 10.52202/085713-1256.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024.

Ryan D'Cunha, Alejandro Lozano, Xiaoxiao Sun, Daniel Vela Jarquin, Min Woo Sun, Josiah Aklilu, James Burgess, Yuhui Zhang, Ryan Nayebi, Paola Avila, Robayo, Jin Ye, Ming Hu, Zhongying Deng, Junjun He, Xin Chen, Yue Yao, Robert Tibshirani, Jeffrey J. Nirschl, and Serena Yeung-Levy. MMBU: A massive multi-modal biomedical understanding benchmark to probe the perception capabilities of vision-language models. *CoRR*, abs/2606.06696, 2026. doi: 10.48550/ARXIV.2606.06696.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348, 2026. doi: 10.48550/ARXIV.2606.19348.

S. V. Destounis, S. M. Friedewald, L. J. Grimm, S. P. Poplack, and J. S. Sung. Mammography. In *ACR BI-RADS v2025 Manual*. American College of Radiology, Reston, VA, 2025.

Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical physics*, 51(4):3110–3123, 2024.

Google Research and Google DeepMind. Medgemma technical report. *CoRR*, abs/2507.05201, 2025. doi: 10.48550/ARXIV.2507.05201.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *CoRR*, abs/2510.08668, 2025. doi: 10.48550/ARXIV.2510.08668.

Alexandros Karargyris, Renato Umeton, Micah J Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, et al. Federated benchmarking of medical artificial intelligence with medperf. *Nature machine intelligence*, 5(7):799–810, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *CoRR*, abs/2506.07044, 2025. doi: 10.48550/ARXIV.2506.07044.

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

J. W. T. Leung, J. A. Baker, R. Hooley, V. A. Loving, and J. A. Rapelyea. Ultrasound. In *ACR BI-RADS v2025 Manual*. American College of Radiology, Reston, VA, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Xiang Li, Farzaan Kaiyom, Evan Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. Autobencher: Towards declarative benchmark construction. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94357–94386, 2025.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

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

Yang Liu, Jiajin Zhang, Danyang Tu, Yaojun Hu, Jiao Qu, Jiuyu Zhang, Yu Shi, Wei Fang, Shi Gu, Ling Zhang, and Yingda Xia. Breastgpt: A multimodal large language model for the full spectrum of breast cancer clinical routine. *CoRR*, abs/2606.04911, 2026b. doi: 10.48550/ARXIV.2606.04911.

Lena Maier-Hein, Annika Reinke, Michal Kozubek, Anne L Martel, Tal Arbel, Matthias Eisenmann, Allan Hanbury, Pierre Jannin, Henning Müller, Sinan Onogur, et al. Bias: Transparent reporting of biomedical image analysis challenges. *Medical image analysis*, 66:101796, 2020.

Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023.

Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1):148, 2024.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. In Ninon Burgos, Caroline Petitjean, Maria Vakalopoulou, Stergios Christodoulidis, Pierrick Coupe, Hervé Delingette, Carole Lartizien, and Diana Mateus (eds.), *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 1310–1327. PMLR, 03–05 Jul 2024. URL https://proceedings.mlr.press/v250/royer24a.html.

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

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5:180161, 2018. doi: 10.1038/sdata.2018.161.

Yuzhang Xie, Keqi Han, Yunpeng Xiao, Hejie Cui, Guanchen Wu, Ziyang Zhang, Kai Shu, Jiaying Lu, Xiao Hu, and Carl Yang. Ehrbench: An automated and reliable ehr-based benchmark for clinical decision making with llms. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 10068–10079, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817571.

Shiyun Xiong, Dongming Wu, Peiwen Sun, Yuang Ai, Bokang Yang, Wencheng Han, Xiao-Hui Li, and Xiangyu Yue. Benchmark everything everywhere all at once, 2026. URL https://arxiv.org/abs/2606.06462.

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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025. doi: 10.48550/ARXIV.2504.10479.

A Reproducibility and Construction Details

A.1 Data Normalization, Allocation, and Construction Settings

Table 6 summarizes the effective configuration used in the reported benchmark-construction runs. Here we provide only implementation details that are not specified in the main text but are important for interpreting or reproducing the construction procedure.

TaskPlanningAgent consumes a compact dataset profile containing field types, coverage, cardinality, and representative values. Profiling limits are listed in Table 6. Representative-value summaries exclude identifiers, file paths, and spatial coordinates.

Sample allocation. The requested benchmark size is allocated after the supported task set has been determined. The category weights in Table 6 define relative allocation preferences rather than fixed quotas. If only a subset of the six task categories is represented by the retained tasks, the corresponding weights are renormalized over the active categories. Integer sample counts are then obtained using largest-remainder allocation and distributed within each category across the retained task families and their fine-grained tasks. The resulting counts sum exactly to the user-specified target sample count. They may also be inspected or adjusted at the corresponding construction checkpoint before generation.

Planning checkpoints. MedBenchAgent exposes interruptible verification checkpoints after consequential planning outputs, including requirement specification, dataset and task selection, sample allocation, reference-material selection, and medical task design. At each checkpoint, the current BIR state can be inspected, accepted, or revised before execution resumes. The checkpoint interface is verifier-agnostic: depending on the decision being checked, verification may be implemented through deterministic checks, human review, model-based supervision, or combinations thereof. These checkpoints are intended to verify that the evolving benchmark specification remains medically coherent, annotation-supported, and sufficiently broad in its coverage of meaningful evaluation capabilities before the BIR is locked.

In the reported experiments, we instantiate these checkpoints with researcher review toward this objective. Researchers inspect consequential specification decisions and residual planning errors to assess the medical coherence, annotation support, and breadth of the evolving benchmark specification. This review is not intended to exhaustively enumerate every annotation-supported atomic prediction target or to enforce leaf-by-leaf agreement with an external reference task space. Once the BIR is locked, downstream instantiation follows the resulting specification without requiring manual review of every generated item.

Table 6: Effective configuration used for benchmark construction.

Parameter	Value
LLM inference	
Model	`deepseek-v4-flash`
Serving backend	vLLM `0.22.0`
API interface	OpenAI-compatible
Request timeout	120 s
SDK retries	0
Planning / review temperature	0.0
Question-generation temperature	0.7
Semantic-audit temperature	0.0
Top-p	1.0
Maximum generation length	up to $65{,}536 - N_{\text{prompt}}$ tokens
Numerical dtype	BF16 compute / FP8 weights and KV cache
Generation	
Pipeline seed	0
Maximum tasks per image	20
Question templates per task	3
Record fetch batch size	128
Retrieval and reranking	
Embedding model	`Qwen3-Embedding-0.6B`
Reranker	`Qwen3-Reranker-0.6B`
Chunk size	1,200 characters
Chunk overlap	200 characters
Retrieval candidates	12
Maximum retained passages	5
Minimum reranker score	0.5
Reranker context limit	8,192 tokens
Dataset profiling	
Frequent values per field	3
Maximum displayed value length	200 characters
Task allocation	
Acquisition/basic weight	0.10
Anatomy/global weight	0.20
Finding weight	0.20
Clinical weight	0.20
Report weight	0.10
Spatial weight	0.20
Semantic audit	
Maximum sampled items per task	10

Retrieval, generation, and audit settings. Retrieval is used during medical task design rather than separately for every benchmark item; the exact retrieval, reranking, profiling, and generation configuration is summarized in Table 6. During instantiation, deterministic application-level operations use pipeline seed 0, while reference answers and supporting evidence are derived deterministically from the locked BIR and normalized annotations. Only the question wording is generated stochastically by the LLM. The deterministic hard audit is the final acceptance gate, checking item contracts, provenance, annotation support, reference reconstruction, spatial validity, and exact duplication. A task-stratified semantic sample additionally records task drift, answer leakage, ambiguity, and semantic duplication for diagnostic analysis. All LLM calls use the OpenAI-compatible interface to the locally served `deepseek-v4-flash` model. Planning, review, and semantic-audit calls use temperature 0.0, whereas question generation uses temperature 0.7. The request timeout is 120 s with no SDK-level automatic retries. Parameters not explicitly controlled by MedBenchAgent, including top-p, maximum generation length, numerical dtype, and tensor parallelism, follow the vLLM serving configuration.

Dataset normalization and profiling. Before benchmark construction, MedBenchAgent applies a standardized normalization workflow to convert heterogeneous raw data sources, including tables, JSON files, and dataset metadata, into its canonical image-centric annotation format. This workflow restructures source information into the predefined common `annotation.jsonl` schema while preserving the supplied medical labels. The released artifacts include the normalized annotation files and the common schema used for the three breast-imaging datasets. TaskPlanningAgent interprets the resulting annotation profile under the task definitions and support conditions in the BIR; task support is therefore assessed only against the normalized fields available to planning.

Resource usage. Tables 7 and 8 characterize where computational cost is incurred during benchmark construction. Across the three breast-imaging runs, the three planning agents require 103 calls,

Table 7: Trace-derived LLM usage aggregated across BrEaST, BUS-BRA, and VinDr-Mammo.

Agent	Calls	Input tok.	Output tok.	Latency (s)
Planning				
SpecificationAgent	6	1,355	235	8.7
TaskPlanningAgent	43	96,998	8,776	44.6
MedicalAlignAgent	54	136,804	9,910	53.4
Instantiation and audit				
GenerationAgent	32,614	20,809,793	1,121,145	15,292.5
AuditAgent	20	27,735	7,405	30.9
Total	**32,737**	**21,072,685**	**1,147,471**	**15,430.1**

Table 8: Construction usage by dataset. Per-item quantities use the final audited population. Wall time is observed end-to-end elapsed time and includes both LLM and deterministic pipeline operations.

Dataset	Audited items	Calls	Input tok.	Output tok.	Tok./item	Calls/item	Wall time (h)
BrEaST	2,564	5,230	3,815,429	197,439	1,565.1	2.04	1.00
BUS-BRA	5,000	10,025	5,974,782	353,968	1,265.8	2.00	2.45
VinDr-Mammo	8,684	17,482	11,282,474	596,064	1,367.9	2.01	2.90
Total	**16,248**	**32,737**	**21,072,685**	**1,147,471**	**1,367.6**	**2.01**	**6.35**

254,078 tokens, and 106.7 s of accumulated request latency. The tables report the computational cost of planning and item construction separately.

Across the three datasets, MedBenchAgent produced 16,248 finalized items in 6.35 h of observed end-to-end wall time, averaging 2.01 LLM calls and 1,367.6 processed tokens per finalized item. We report calls, token usage, accumulated request latency, and wall time to characterize computational footprint and construction throughput. The trace therefore shows that the additional specification stages introduced by MedBenchAgent account for only a small fraction of total LLM usage; the dominant computational cost arises from item-level generation and review during instantiation.

A.2 BIR Task Prior and Dataset-Specific Compilation

The BIR is the explicit construction state used throughout benchmark planning. Its initial form contains benchmark-level metadata together with a reusable collection of candidate task objects. TaskPlanningAgent compiles this state against the confirmed evaluation scope and each dataset's normalized annotation schema, retaining supported tasks and expanding decomposable task families into atomic prediction targets.

The complete initialized BIR template and machine-readable task prior used in our experiments are provided with the released code. Here, we retain a concrete compilation example to illustrate how a generic candidate task is specialized into a dataset-specific task specification and subsequently used for deterministic reference construction.

The distinction between BIR initialization and dataset-specific compilation is visible in a concrete example. The initial BIR contains the generic `finding_attribute` task family, whose support condition requires an explicitly annotated image-observable property of a target finding. For BrEaST, evidence-constrained expansion produces child prediction targets for shape, margin, echogenicity, posterior features, calcifications, halo sign, and skin thickening. The following schematic record illustrates the resulting mass-margin construction specification and its item-level reference derivation; the explanatory hierarchy is shown for readability and does not introduce additional fields into the initialized BIR template.

```
initial_bir_task:
  task_id: finding_attribute
  task_type: close_vqa
  response_format: letter
  category: finding_recognition_and_characterization
  sample_count: null
dataset_conditioned_task:
  dataset: BrEaST
```

```
    atomic_task_id: finding_attribute.margin
    supporting_fields: [image.annotation.margin]
    require: image.annotation.margin is present and mappable
    metric: macro_f1
    options: [Circumscribed, Not circumscribed]
    answer_field: image.annotation.margin
    answer_value_mapping:
      ''circumscribed'' → ''Circumscribed''
      ''not circumscribed - indistinct'' → ''Not
  circumscribed''
    question_templates:
      - ''On this breast ultrasound, how would you describe
        the margin of the mass?''
  example_instance:
    image_id: case001.png
    annotation:
      path: image.annotation.margin
      raw_value: ''not circumscribed - indistinct''
    normalization: ''Not circumscribed''
    semantic_reference: ''Not circumscribed''
    reference_source: deterministic label mapping
    provenance: dataset id, image id, field path, raw value,
      BIR hash
```

This example shows two roles of the BIR task representation. The initial task object defines a valid candidate capability. Compilation for each dataset then checks whether the available annotations support that capability and, when needed, splits it into atomic prediction targets. Downstream item construction follows the confirmed specification without redefining the task.

A.3 Prompt and Implementation Artifacts

The complete decision-critical prompts for SpecificationAgent, TaskPlanningAgent, MedicalAlignAgent, GenerationAgent, and AuditAgent are provided with the released code, together with their serialization and validation logic. Section 4 describes the corresponding agent responsibilities and BIR state transitions. Deterministic dataset profiling, sample allocation, reference derivation, and hard auditing do not rely on LLM prompts.

B Constructed Breast-Imaging Benchmark

B.1 Raw Task Proposals and Evidence Mappings

Table 9 transcribes every task that TaskPlanningAgent marked `retain` in the initial `task_plans/v1.json` output, before researcher confirmation. It shows the model's original task decomposition and its exact mapping from each proposed task to authorized annotation evidence. The count-affecting review decisions are reported separately in Appendix C.2.

B.2 Audited Representative Items

The 20 entries in Figures 4 and 5 give one audited item for every atomic dataset–task pair instantiated in the realized breast-imaging benchmark, spanning BrEaST, BUS-BRA, and VinDr-Mammo. For each entry, we show the source dataset, task definition, representative image, generated query, and corresponding benchmark reference answer.

C Construction Analysis and Benchmark Validation

C.1 Reference Task Space and Task-Space Evaluation Protocol

Baselines. We compare MedBenchAgent with three direct task-planning baselines to evaluate BIR-mediated task-space compilation against direct task induction. All methods receive the same

Table 9: TaskPlanningAgent's initial retained task proposals and supporting fields before task-space review.

Dataset	Raw proposed task	Raw supporting field(s)
BrEaST: 11 retained proposals		
BrEaST	anatomy_attribute.tissue_composition	study.annotation.tissue_composition
BrEaST	finding_attribute.shape	image.annotation.shape
BrEaST	finding_attribute.margin	image.annotation.margin
BrEaST	finding_attribute.echogenicity	image.annotation.echogenicity
BrEaST	finding_attribute.posterior_features	image.annotation.posterior_features
BrEaST	finding_attribute.calcifications	image.annotation.calcifications
BrEaST	finding_attribute.halo	image.annotation.halo
BrEaST	finding_attribute.skin_thickening	image.annotation.skin_thickening
BrEaST	assessment_category	image.annotation.birads
BrEaST	diagnosis	image.annotation.diagnosis
BrEaST	pathology_status	image.annotation.classification
BUS-BRA: 6 retained proposals		
BUS-BRA	finding_type	image.annotation.histology
BUS-BRA	assessment_category	image.annotation.birads
BUS-BRA	diagnosis	image.annotation.histology
BUS-BRA	pathology_status	image.annotation.pathology
BUS-BRA	detection	image.annotation.bbox[*]
BUS-BRA	visual_grounding	image.annotation.bbox[*]; image.annotation.histology
VinDr-Mammo: 6 retained proposals		
VinDr	view_identification	image.annotation.view_position
VinDr	anatomy_attribute.breast_density	image.annotation.breast_density
VinDr	finding_type	image.annotation.findings[*].finding_categories[*]
VinDr	assessment_category	image.annotation.breast_birads
VinDr	detection	image.annotation.findings[*].bbox[*]
VinDr	visual_grounding	image.annotation.findings[*].bbox[*]; image.annotation.findings[*].finding_categories[*]

confirmed evaluation scope $\mathcal{U}$ and the same dataset-specific profile derived from the normalized annotation schema Σ_m. MedBenchAgent compiles the initial task tree $\mathcal{T}_0$ within the BIR against the available annotation evidence through task retention and expansion. The direct baselines instead construct a complete set of prediction targets directly from $\mathcal{U}$ and Σ_m in a single model call per dataset. This comparison therefore evaluates catalog-initialized BIR compilation against direct task induction under the same evaluation scope and dataset profile. Direct $+\mathcal{T}_0$ augments the DeepSeek-V4-Flash direct baseline with the initial BIR task definitions as prompt reference.

For this comparison, both MedBenchAgent and the direct baselines are evaluated before human task-space confirmation. The MedBenchAgent result is the initial task-space proposal produced by TaskPlanningAgent, while each direct baseline is evaluated from its raw model output. All LLM baselines return a single JSON object through an OpenAI-compatible endpoint and are hosted locally on 1–2 NVIDIA B200 GPUs.

Reference task-space design. We designed the breast-imaging reference task space $\mathcal{T}^\star$ from the perspective of medical VLM evaluation, with the goal of covering the major stages of imaging diagnosis across the breast cancer diagnostic workflow. Informed by discussions within the research team and with a clinical collaborator, we abstracted this workflow into eight evaluation phases: image recognition and validity, global tissue characterization, lesion detection and localization, finding characterization, diagnostic classification, standardized assessment, management and integrated reasoning, and report generation. These phases provide a clinically structured progression from basic image interpretation to increasingly integrated diagnostic decisions, while remaining expressed as model-evaluable visual prediction capabilities.

We then applied this staged framework using the task definitions and reference annotations available in the public breast-imaging datasets considered in this study. For each phase, we defined a prediction target only when the corresponding annotations provided a clear reference standard, and we represented distinct prediction targets as separate atomic tasks. The framework defines the intended clinical coverage, while each dataset determines which stages have annotation-supported targets.

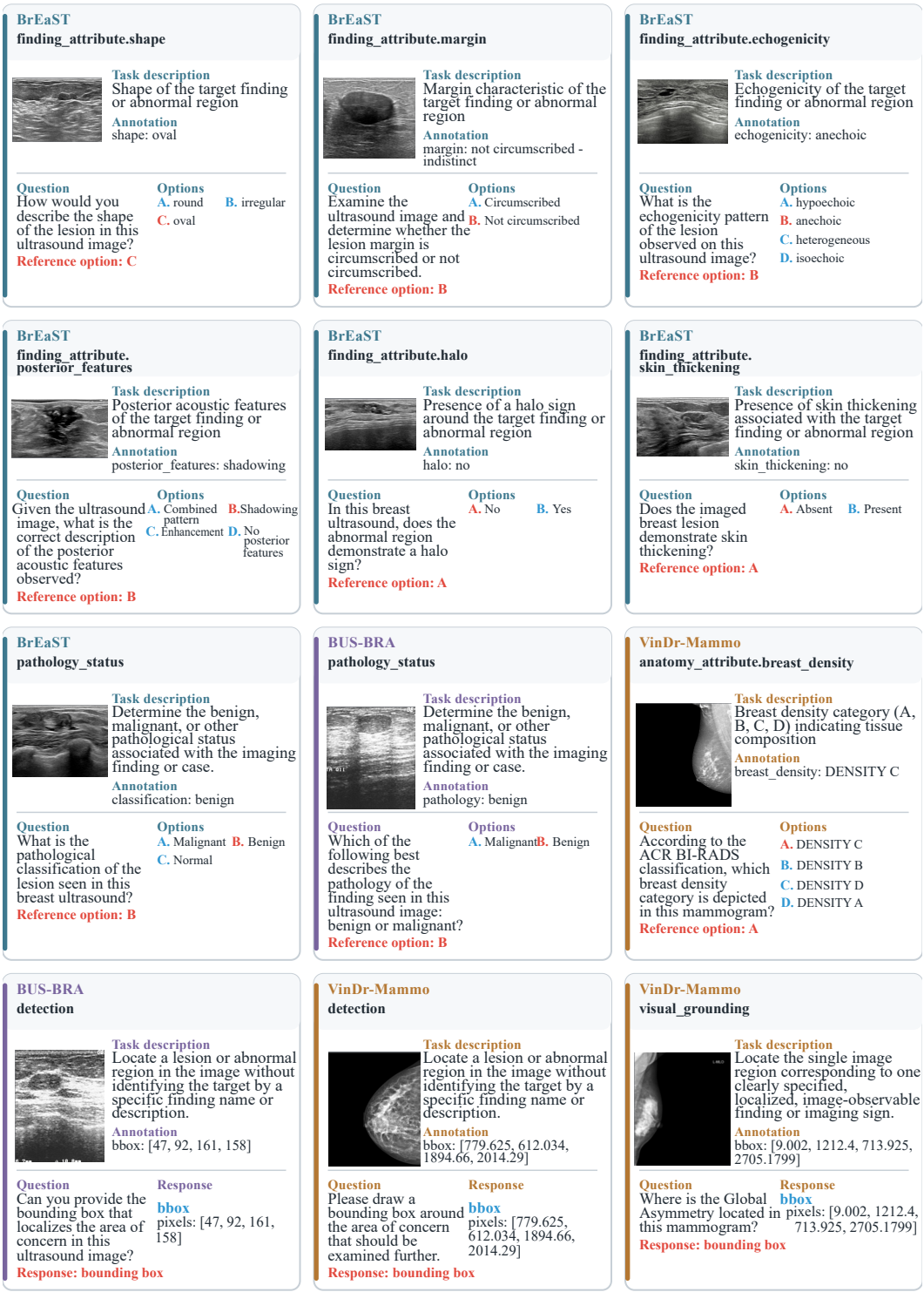


Figure 4: Representative audited items for 12 dataset–task mappings in the breast-imaging benchmark. Each card shows the source dataset and task, source annotation, representative image, generated query, and benchmark reference or spatial response.

Stages without an eligible target remain in the framework, showing both the covered capabilities and the parts of the broader imaging workflow not supported by the available public annotations.

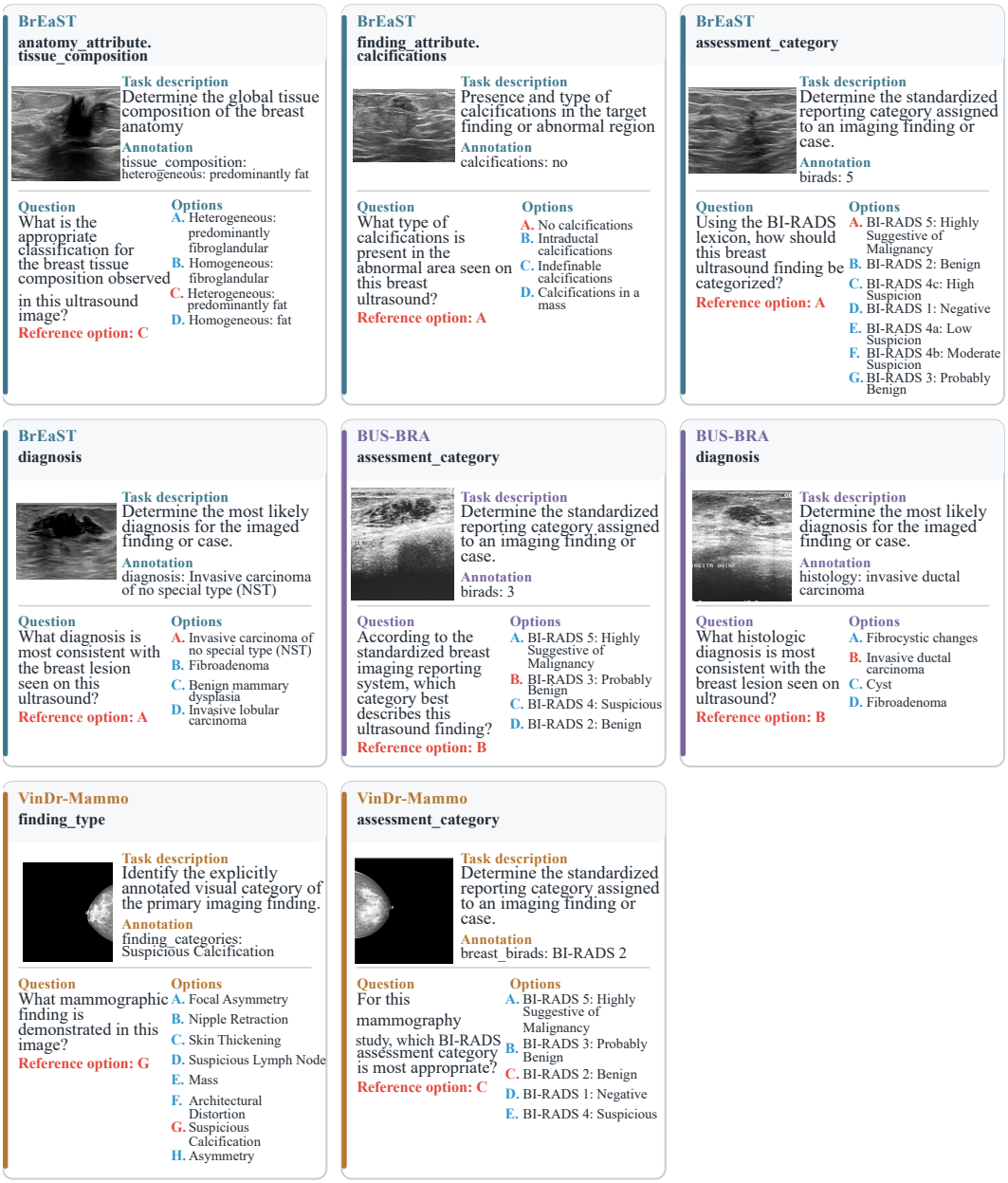


Figure 5: Continued from Figure 4. Representative samples from different dataset–task mappings, completing all 20 pairs.

Figure 6 reports the resulting dataset-conditioned reference task space. Under the atomic matching protocol below, the table contains 21 reference prediction targets in total. A phase may contain multiple atomic targets when the available annotations support distinct prediction targets, whereas a categorical prediction target such as VinDr finding-type classification remains one task even when its answer space contains multiple finding categories.

Independence from automated task-space construction. The staged framework and the resulting $\mathcal{T}^\star$ were developed before the automated task-space experiments and were fixed prior to evaluating MedBenchAgent or any direct-construction baseline. $\mathcal{T}^\star$ was not provided to any evaluated construction method, and the reference task space was not added to, removed from, split, or merged in response to method outputs. It therefore serves only as the external reference for measuring agreement between automatically constructed task spaces and the independently designed breast-imaging evaluation space.

Human involvement in task-space evaluation. Development of the reference framework and adjudication of method outputs were separated from the automated construction procedures. The reference-space design involved benchmark-design discussions among dataset-familiar researchers with clinical input. For the Task-Space F1 analysis reported in this work, matching and adjudication between method proposals and the fixed reference task space were performed by a researcher using the predefined rules below.

Prediction-target matching. The unit of evaluation was a dataset-conditioned prediction target. An ordinary task counted once; a decomposable parent category did not count separately, whereas each distinct prediction target under that parent was evaluated as one task. Matching was based on whether the proposal and reference represented the same prediction target, regardless of task identifier, surface wording, evaluation metric, response format, or JSON nesting. A raw proposal was counted as a true positive (TP) when it matched one reference prediction target in $\mathcal{T}^\star$. A proposal without a corresponding reference task was counted as a false positive (FP), including unsupported, out-of-scope, shortcut-prone, or semantically redundant proposals. A reference task in $\mathcal{T}^\star$ without a semantically equivalent proposal was counted as a false negative (FN).

Matching was one-to-one. If multiple proposals represented the same reference prediction target, at most one was matched as a TP and each additional duplicate was counted as an FP. If a single proposal conflated multiple distinct prediction targets, it could match at most one reference task, with each remaining unmatched target counted as an FN. Differences that did not change the prediction target—including the evaluation metric, response format, bounding-box coordinate convention, surface wording, or exact annotation-field spelling—did not affect task-level matching.

Evidence-Mapping Agreement. For task proposals matched to $\mathcal{T}^\star$, we additionally evaluated whether the method identified the annotation fields required to support the corresponding reference construction. Evidence-Mapping Agreement (EMA) was computed only over matched TPs. For each matched task, agreement was assigned a value of one when the method-proposed supporting-field set was identical to the supporting-field set confirmed during review against the dataset annotation schema Σ_m, and zero otherwise. FPs and FNs were excluded from the EMA denominator.

Metrics and aggregation. For each method–dataset pair, we computed

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2TP}{2TP + FP + FN}.$$

We additionally report the number of task-space discrepancies between a raw proposal and $\mathcal{T}^\star$, defined as $FP + FN$. Overall results micro-aggregate TP, FP, and FN across the three datasets before applying these equations. Thus, datasets contribute according to their numbers of reference and proposed prediction targets rather than receiving equal dataset weight.

Direct task-induction baseline. The direct baselines receive the same confirmed benchmark requirement and dataset profile as MedBenchAgent, but do not receive the predefined task catalogue T_0. They induce the complete task space in a single model call under the same output schema and evidence-admissibility rules, including annotation-supported prediction targets, atomic answer axes, semantic-duplicate removal, and explicit spatial-evidence requirements. The complete baseline prompts and serialization instructions are provided with the released code.

C.2 Task-Space Results and Discrepancy Analysis

Tables 10 and 11 detail task-space discrepancies for MedBenchAgent and the three direct baselines. Their matched tasks have supporting-field mappings consistent with the reference, as reflected by the EMA results in Table 1.

The discrepancy patterns also differ qualitatively across construction strategies. MedBenchAgent's residual errors are concentrated in semantic boundary cases, including a redundant BUS-BRA finding-type proposal, a visual-grounding proposal that did not define a capability distinct from generic detection, a shortcut-prone VinDr view-identification task, and the omitted finding-level assessment target. In contrast, the direct baselines more frequently introduce nonvisual metadata, acquisition or marker shortcuts, invalid spatial evidence, or omit otherwise supported visual prediction targets. This distinction helps explain why Evidence-Mapping Agreement is perfect for matched tasks even when task-space precision differs: once the intended prediction target is identified, annotation-field

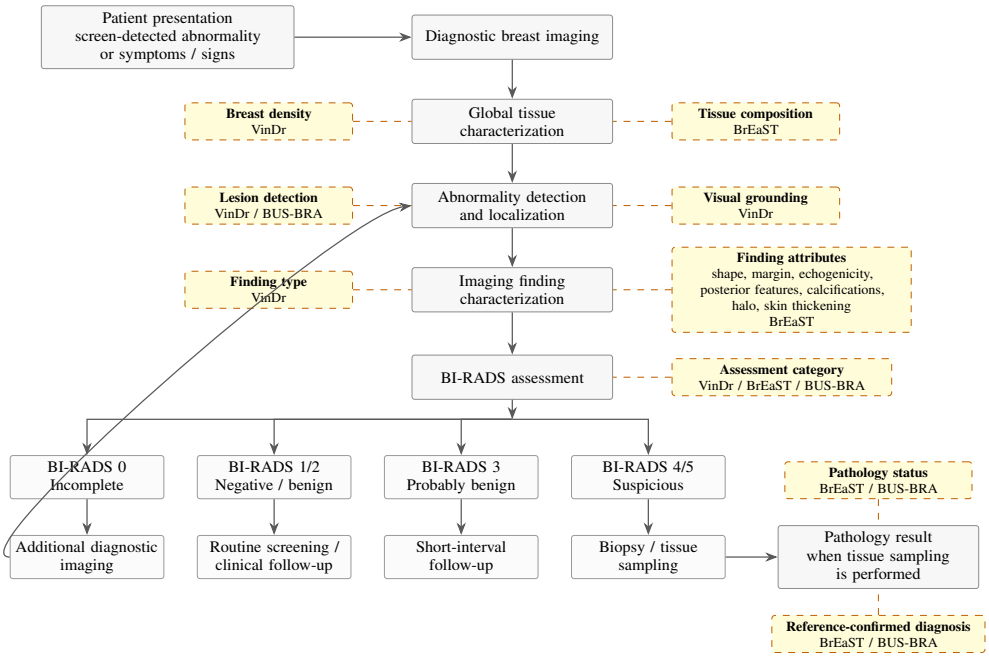


Figure 6: Clinical breast-imaging workflow and its annotation-supported reference task space. Tasks are attached to the corresponding stages of the imaging workflow.

Table 10: Task-space discrepancies affecting TP, FP, or FN counts for MedBenchAgent and Direct DeepSeek. Retained true positives are omitted.

Method	Dataset	Proposal or confirmed task	Label	Review rationale
MedBenchAgent	BUS-BRA	finding type	FP	Duplicated diagnosis on the histology answer axis.
MedBenchAgent	BUS-BRA	visual grounding	FP	Did not define a capability distinct from generic lesion detection.
MedBenchAgent	VinDr	view identification	FP	Burned-in view markers exposed the target as a shortcut.
MedBenchAgent	VinDr	finding-level assessment	FN	Finding-level BI-RADS supported a distinct, initially omitted answer axis.
Direct DeepSeek	BrEaST	clinical signs	FP	Nonvisual clinical metadata.
Direct DeepSeek	BrEaST	patient symptoms	FP	Nonvisual patient information.
Direct DeepSeek	BrEaST	verification method	FP	Nonvisual examination/pathology metadata.
Direct DeepSeek	BrEaST	tumor localization	FP	Treated a mask as bounding-box evidence.
Direct DeepSeek	BrEaST	diagnosis	FN	The explicit diagnosis field was not proposed as an independent task.
Direct DeepSeek	BUS-BRA	side identification	FP	Laterality metadata was outside the intended visual scope.
Direct DeepSeek	BUS-BRA	device identification	FP	Acquisition-device metadata was outside the clinical visual scope.
Direct DeepSeek	VinDr	laterality classification	FP	Metadata/marker shortcut rather than the intended image semantics.
Direct DeepSeek	VinDr	view-position classification	FP	Burned-in view marker leaked the target.
Direct DeepSeek	VinDr	visual grounding	FN	Finding category was not paired with its box.

selection is comparatively stable, whereas deciding which targets should constitute the benchmark remains more difficult.

C.3 Human Item Audit and Residual Errors

Sampling and review setting. The audit population comprised all 16,248 finalized items accepted by the automatic screening and audit procedures in the completed BrEaST, BUS-BRA, and VinDr-Mammo runs. We globally shuffled this combined population and sampled 1,000 items uniformly without replacement.

We reviewed each sampled item using the predefined criteria below. The review materials comprised the dataset and task identifiers, response format, question, options, benchmark reference, and complete

Table 11: Task-space discrepancies affecting TP, FP, or FN counts for Direct Qwen and Direct MedGemma. Retained true positives are omitted.

Method	Dataset	Proposal or confirmed task	Label	Review rationale
Direct Qwen	BrEaST	palpability assessment	FP	Clinical palpation is not image-answerable.
Direct Qwen	BrEaST	interpretation classification	FP	A source conclusion is not an independent visual capability.
Direct Qwen	BrEaST	verification method	FP	Nonvisual confirmation metadata.
Direct Qwen	BrEaST	patient symptom status	FP	Nonvisual patient information.
Direct Qwen	BUS-BRA	lateralization	FP	Laterality metadata was outside the intended scope.
Direct Qwen	BUS-BRA	device identification	FP	Acquisition-device metadata was outside the intended scope.
Direct Qwen	VinDr	laterality classification	FP	Metadata/marker shortcut.
Direct Qwen	VinDr	view-position classification	FP	Burned-in view marker leaked the target.
Direct Qwen	VinDr	breast density	FN	A varying breast-density field supported an omitted global answer axis.
Direct Qwen	VinDr	visual grounding	FN	Finding category was not paired with its box.
Direct MedGemma	BrEaST	anatomy classification	FP	Anatomy was constant and could not form a valid classification task.
Direct MedGemma	BrEaST	modality classification	FP	Modality was constant.
Direct MedGemma	BrEaST	interpretation classification	FP	Source interpretation was not an independent visual capability.
Direct MedGemma	BrEaST	age classification	FP	Patient age was nonvisual metadata.
Direct MedGemma	BrEaST	patient symptoms	FP	Patient symptoms were nonvisual.
Direct MedGemma	BrEaST	signs classification	FP	Clinical-sign metadata was outside the target visual scope.
Direct MedGemma	BrEaST	verification classification	FP	Nonvisual confirmation metadata.
Direct MedGemma	VinDr	laterality classification	FP	Metadata/marker shortcut.
Direct MedGemma	VinDr	view-position classification	FP	Burned-in view marker leaked the target.
Direct MedGemma	VinDr	visual grounding	FN	Finding category was not paired with its box.

normalized annotation. The audit evaluated construction consistency against the source annotations and locked task specification.

Audit criteria. Four binary criteria were applied. *Reference groundedness* required the benchmark reference to follow directly from the authorized annotation evidence. *Task–query alignment* required the question to express the locked prediction target. *Unambiguity* required the question to admit one intended answer under the supplied evidence. *Evaluation integrity* required the absence of answer leakage, overlapping response options, question–reference mismatch, response-format mismatch, and unsupported additions. An item was considered valid only when all four criteria passed:

$$\mathrm{ValidItem}_i = \mathbf{1}\{G_i \wedge A_i \wedge U_i \wedge I_i\},$$

where G_i, A_i, U_i, and I_i denote reference groundedness, task–query alignment, unambiguity, and evaluation integrity, respectively.

Residual errors. Six of the 1,000 sampled items failed at least one audit criterion. Four VinDr-Mammo finding-type items exhibited the same template-level task–query mismatch: the generated question asked for a BI-RADS assessment category while the locked answer space contained finding categories. Two BrEaST diagnosis items introduced unsupported contextual details and therefore failed evaluation integrity. Importantly, the fixed annotation-derived reference answers remained supported by the authorized evidence in all six cases. Across the complete audit, there were four task–query-alignment failures and six evaluation-integrity failures, with no reference-groundedness or unambiguity failures.

D Cross-Domain Portability Study: Dermoscopy

D.1 Dataset, Task Space, and Realized Allocation

The cross-domain run used the ISIC 2018 challenge data from its three official task sources: lesion segmentation (Task 1), dermoscopic-attribute detection (Task 2), and disease classification (Task 3). Tasks 1 and 2 share 1,000 held-out dermoscopic images, while Task 3 contributes 1,512 held-out diagnosis images. After deduplicating the shared Task 1/2 identifiers, the normalized corpus contains 2,512 unique images. Diagnosis labels are available for the Task 3 subset, whereas lesion masks and five dermoscopic-attribute masks are available for the Task 1/2 subset. Because these annotations are defined on different image subsets, no single task is eligible for all 2,512 images. We apply the

Table 12: ISIC 2018 tasks, supporting annotation fields, and realized item counts.

Task	Supporting field	Resp.	Metric	Planned	Realized
Diagnosis	`image.annotation.diagnosis`	letter	Macro-F1	1,667	1,512
Detection	`image.annotation.findings[*].bbox[*]`	bbox	mIoU	1,666	1,000
Globules presence	`image.annotation.attribute_presence.globules`	letter	Accuracy	334	334
Milia-like cyst presence	`image.annotation.attribute_presence.milia_like_cyst`	letter	Macro-F1	334	334
Negative network presence	`image.annotation.attribute_presence.negative_network`	letter	Macro-F1	333	333
Pigment network presence	`image.annotation.attribute_presence.pigment_network`	letter	Accuracy	333	333
Streaks presence	`image.annotation.attribute_presence.streaks`	letter	Macro-F1	333	333
			Total	5,000	4,179

workflow to the ISIC annotation schema and evaluate the resulting task-to-evidence mappings, item consistency, and model performance.

Task space and realized allocation. The requested 5,000-item allocation yielded 4,179 realized items because of source-data eligibility limits. All eligible diagnosis (1,512) and lesion-localization (1,000) cases were retained, with the remaining 1,667 items allocated across the five dermoscopic-attribute tasks. We checked the initial task plan against the documented ISIC annotation schema and label semantics. All seven proposed tasks and their supporting-field mappings were retained without task-space revision.

D.2 Item Consistency and Representative Examples

Item consistency. We additionally performed a consistency check on 200 finalized ISIC items using the same annotation- and contract-based criteria as the breast-imaging audit. All sampled items passed checks of reference–evidence consistency, task–query consistency, and evaluation-format integrity against the normalized annotations and locked task contracts.

Figure 7 provides one audited representative item for each of the seven dermoscopy tasks using the same presentation format as the breast-imaging examples.

D.3 Model Evaluation

Table 13 reports performance across diagnosis, dermoscopic attributes, and localization. Different checkpoints lead on different task metrics, providing complementary evaluation profiles. Together with the task-space review and item audit, these results demonstrate application of MedBenchAgent to the ISIC annotation schema.

Table 13 shows that the cross-domain benchmark preserves substantial variation across task families and model checkpoints. No single checkpoint exhibits uniformly strong performance across diagnosis, dermoscopic attributes, and lesion localization, and the relative behavior of general-purpose and medical VLMs varies across these task types. In particular, categorical performance does not reliably predict spatial-localization performance. These results are consistent with the breast-imaging evaluation in showing that aggregate model quality cannot be inferred from one task family alone, while providing evidence that the same construction workflow can be instantiated under a different annotation schema and clinical vocabulary.

E Additional Model-Evaluation Results

E.1 Evaluation Protocol

For reporting, fine-grained question types in Table 3 are first scored separately and macro-averaged within their parent task where applicable, while dataset-specific results for the same task family are

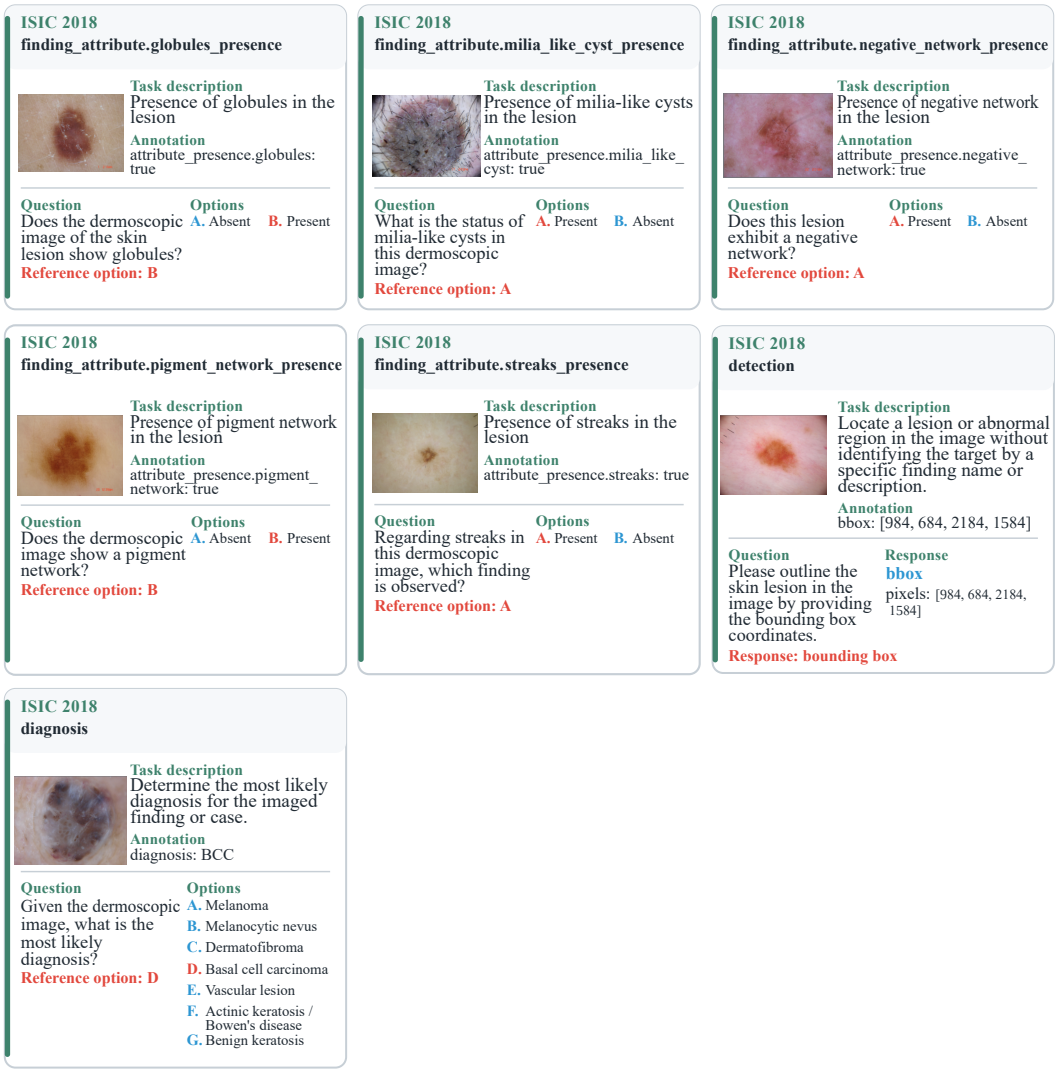


Figure 7: Representative audited ISIC 2018 items for all seven dermoscopy tasks.

then macro-averaged with equal weight, and the reported Overall score macro-averages the resulting task-family scores.

In this section, we report the complete per-dataset model-evaluation results and detailed evaluation protocol. Tables 14, 16, and 17 use the same two-level template to report dataset-level task results for BrEaST, BUS-BRA, and VinDr-Mammo. The abbreviations are AA (anatomy attribute), FA (finding attribute), FT (finding type), AC (assessment category), DX (diagnosis), and PS (pathology status). Table 15 separately expands BrEaST FA into its seven question types. Table 18 reports mIoU and $\mathrm{Acc@IoU}_{0.5}$ for each dataset–task pair. Scores are computed over all evaluated items, assigning zero IoU when no valid bounding box is returned.

Inputs and response parsing. Each model received the image, question, and lettered options when applicable, without the reference answer. Model-specific chat templates were applied with deterministic decoding. Multiple-choice predictions were parsed as a single valid option letter and evaluated using Accuracy and Macro-F1 over semantic answer classes. Invalid responses remained incorrect. Spatial predictions were parsed as a single four-number bounding box, accepting the supported `bbox` or `bbox_2d` fields, bare lists or tuples, and JSON-fenced outputs. Missing,

Table 13: Evaluation performance on ISIC 2018. Choice tasks report Accuracy / Macro-F1 (%), while detection reports mIoU / Acc@IoU0.5 (%). Best results for each metric are shown in bold.

Model	Diagnosis	Finding Attribute					Detection
		Globules	Milia-like	Neg. net.	Pigment net.	Streaks	
General-purpose VLMs							
InternVL3-8B	20.7 / 15.3	31.4 / 31.1	70.7 / 56.6	45.3 / 35.0	72.4 / 63.4	39.3 / 35.7	50.5 / 52.2
Qwen2.5-VL-7B	24.2 / 14.6	57.8 / 53.7	80.8 / 55.5	83.8 / 47.4	59.5 / 59.3	33.6 / 30.7	75.1 / 86.6
Qwen3-VL-8B-I	53.2 / 22.8	28.4 / 27.5	80.2 / **61.8**	80.2 / 47.3	59.5 / 58.4	58.0 / 44.7	**80.3** / 91.4
Qwen3-VL-8B-T†	30.9 / 18.6	60.5 / 52.9	83.2 / 61.1	85.0 / 48.2	57.1 / 56.0	82.9 / 47.0	71.2 / 83.6
Qwen2.5-VL-32B	29.8 / 20.8	25.1 / 23.6	46.7 / 40.2	28.5 / 24.3	72.4 / 64.5	15.9 / 15.9	76.8 / 87.6
Qwen3-VL-32B	52.8 / 30.3	70.1 / 58.2	84.4 / 55.1	74.2 / 45.8	58.0 / 57.3	**88.0** / **55.1**	80.2 / **92.2**
Medical VLMs							
MedGemma-4B	47.8 / 31.5	25.7 / 24.0	66.8 / 52.5	45.3 / 33.5	**79.0** / **71.7**	59.5 / 47.7	31.8 / 29.1
Lingshu-7B	10.6 / 7.0	68.3 / 57.2	82.9 / 53.9	46.8 / 36.2	76.9 / 71.4	46.5 / 38.4	11.7 / 6.8
Hulu-Med-7B	**56.3** / **34.4**	**75.4** / 47.4	83.8 / 59.0	**95.2** / **48.8**	40.8 / 38.6	76.6 / 48.9	0.7 / 0.2
QoQ-Med-7B†	49.5 / 21.4	69.2 / 48.2	**84.7** / 55.6	90.7 / 47.6	41.1 / 39.2	75.7 / 46.8	25.8 / 11.3
Lingshu-32B	24.6 / 16.8	71.9 / **60.3**	75.7 / 53.5	44.1 / 33.9	72.4 / 69.2	57.4 / 44.8	44.5 / 53.3

malformed, nonfinite, degenerate, out-of-range, or multiple distinct boxes were invalid and received zero IoU.

Coordinate conversion and scoring. We strictly follow the protocols of all models that claim to have spatial evaluation capabilities. Predictions were converted from the configured model-specific coordinate representation to absolute original-image coordinates before scoring. Qwen3-VL, InternVL3, and BreastGPT used a fixed $[0, 1000]$ coordinate space; Qwen2.5-VL, Lingshu, and QoQ-Med used coordinates in the resized model input; and Hulu-Med used original-image coordinates. MedGemma used normalized $[0, 1]$ coordinates in $[y_1, x_1, y_2, x_2]$ order, which were reordered and rescaled. The effective coordinate size was recorded with each prediction. Spatial metrics were computed over all evaluated items, including invalid outputs. Accordingly, mIoU and $\mathrm{Acc@IoU}_{0.5}$ measure end-to-end bounding-box prediction under the specified response protocol.

E.2 Pairwise Model-Ordering Analysis

The summary statistics of model-ordering reversals are reported in Section 5.3. Here, we provide the detailed pairwise model-ordering analysis and describe how these statistics are computed. For two models i and j, an ordering reversal between scores s and s' occurs when

$$(s_i - s_j)(s'_i - s'_j) < 0.$$

With 12 evaluated VLMs, there are $\binom{12}{2} = 66$ unordered model pairs. Ties are not counted as reversals.

Overall versus task-level evaluation. We compare the ordering induced by Overall Acc. in Table 3 with the Acc. ordering for each task category. The numbers of reversed pairs are 19 (AA), 22 (FT), 18 (FA), 24 (AC), 12 (DX), and 22 (PS), corresponding to 18.2–36.4% of all model pairs.

Acc. versus F1. Within each task category, we compare model ordering under Acc. and F1. The numbers of reversed pairs are 13 (AA), 11 (FT), 5 (FA), 15 (AC), 15 (DX), and 4 (PS), corresponding to 6.1–22.7%.

Choice-based versus spatial evaluation. We compare Overall Acc. with each of the six spatial measures in Table 4. The numbers of reversed model pairs are 33 and 32 for VinDr-Mammo detection, 34 and 32 for VinDr-Mammo visual grounding, and 29 and 28 for BUS-BRA detection, for mIoU and $\mathrm{Acc@IoU}_{0.5}$, respectively. Thus, 28–34 of the 66 model pairs (42.4–51.5%) reverse their ordering across the spatial measures.

Table 14: **Multiple-choice evaluation on BrEaST by task type.** Results are reported as **Accuracy / Macro-F1** (%). AA contains 200 tissue-composition items; FA equally averages its seven question types; AC and PS each contain 256 items; and DX contains 121 items. † denotes thinking variants.

Model	Anatomy	Finding	Assessment & Diagnosis		
	AA	FA	AC	DX	PS
General-purpose VLMs					
InternVL3-8B	27.5 / 10.8	42.3 / 30.9	18.8 / 10.1	28.9 / 16.3	50.8 / 33.4
Qwen2.5-VL-7B	33.0 / 19.5	41.5 / 29.3	16.4 / 10.5	34.7 / 21.3	54.7 / 36.4
Qwen3-VL-8B-I	27.5 / 10.9	29.8 / 22.6	20.7 / 16.7	38.8 / 26.9	60.9 / 40.3
Qwen3-VL-8B-T†	28.0 / 13.4	56.9 / 35.1	21.1 / 19.3	38.0 / 24.5	49.2 / 32.7
Qwen2.5-VL-32B	25.0 / 15.2	46.2 / 29.6	19.1 / 13.2	48.8 / 27.3	47.3 / 31.6
Qwen3-VL-32B	17.5 / 12.9	37.1 / 28.4	20.7 / 20.8	35.5 / 27.7	62.9 / 53.0
Medical VLMs					
MedGemma-4B	26.5 / 16.0	47.5 / 33.9	19.5 / 13.4	44.6 / 32.4	51.6 / 36.9
Lingshu-7B	**33.5** / **24.0**	56.8 / 38.7	24.6 / 25.0	46.3 / 18.8	**69.1** / 40.4
Hulu-Med-7B	27.0 / 11.3	51.1 / 35.5	20.7 / 16.0	44.6 / 24.4	61.7 / 41.3
QoQ-Med-7B†	30.0 / 20.7	66.1 / 37.8	20.7 / 13.7	43.0 / 23.5	48.8 / 31.5
BreastGPT-8B	21.5 / 19.9	**75.2** / **52.5**	**36.3** / **36.1**	**60.3** / **39.7**	68.4 / **55.5**
Lingshu-32B	31.5 / 22.2	57.5 / 43.0	23.4 / 18.7	47.1 / 20.3	67.2 / 42.3

Table 15: **Multiple-choice evaluation on BrEaST: finding-attribute question types.** Results are reported as **Accuracy / Macro-F1** (%). Best results for each metric are shown in bold.

Model	Finding Attribute						
	Shape	Margin	Echo.	Posterior	Calc.	Halo	Skin
General-purpose VLMs							
InternVL3-8B	54.5 / 30.9	53.2 / 53.0	46.3 / 33.1	21.8 / 15.6	45.2 / 23.0	26.0 / 24.4	48.8 / 35.9
Qwen2.5-VL-7B	55.4 / 26.8	58.3 / 47.4	40.3 / 28.3	19.4 / 8.1	27.8 / 14.7	47.6 / 44.6	41.4 / 35.1
Qwen3-VL-8B-I	34.3 / 28.6	47.2 / 42.2	25.5 / 16.2	37.3 / 24.8	50.4 / 24.0	11.4 / 16.7	2.7 / 5.6
Qwen3-VL-8B-T†	56.6 / 34.4	58.3 / 45.8	61.5 / 28.4	22.2 / 12.4	59.9 / 30.4	63.4 / 46.6	76.6 / 47.8
Qwen2.5-VL-32B	57.4 / 32.3	55.2 / 45.3	61.0 / 35.1	22.2 / 12.4	71.0 / 28.6	47.2 / 43.8	9.4 / 9.4
Qwen3-VL-32B	58.7 / 39.9	32.9 / 40.3	45.0 / 34.7	23.8 / 17.9	52.0 / 24.7	15.9 / 17.3	31.2 / 24.2
Medical VLMs							
MedGemma-4B	52.9 / 40.8	54.8 / 52.8	39.8 / 23.3	52.8 / 24.6	29.0 / 14.2	42.3 / 39.6	61.3 / 42.3
Lingshu-7B	58.3 / 45.6	63.9 / 60.2	63.6 / 32.6	13.9 / 9.4	44.0 / 18.8	65.4 / 48.5	88.7 / 55.5
Hulu-Med-7B	58.3 / 34.1	59.5 / 50.0	58.9 / 38.6	25.8 / 19.1	17.9 / 13.3	64.6 / 45.0	73.0 / 48.4
QoQ-Med-7B†	52.1 / 30.6	65.1 / 62.2	51.9 / 30.5	**57.9** / 25.8	75.4 / 27.2	**72.0** / 41.8	88.3 / 46.9
BreastGPT-8B	**73.6** / **56.1**	**78.6** / **78.0**	**70.6** / **45.3**	54.8 / **33.2**	**89.3** / **35.0**	70.3 / **63.7**	**89.5** / **56.3**
Lingshu-32B	59.9 / 47.5	67.9 / 64.4	62.8 / 44.7	23.0 / 21.0	45.2 / 22.9	65.4 / 44.6	78.5 / 56.1

E.3 Additional Analysis of Model Behavior

The results allow us to examine how aggregate model performance is composed and where model comparisons change with the evaluation target, metric, or imaging setting. We first analyze variation within task families, then examine changes across model scales and thinking variants, and finally compare categorical and spatial performance across imaging settings. Throughout, comparisons refer to the evaluated checkpoints under the reported protocols.

E.3.1 Task Granularity and Metric-Dependent Rankings

Strong task-family performance can coexist with substantial variation among its constituent prediction targets. BreastGPT-8B achieves the highest aggregate finding-attribute scores, yet its BrEaST Macro-F1 ranges from 78.0 on margin and 63.7 on halo to 33.2 on posterior features and 35.0 on calcifications (Table 15). The aggregate FA score therefore combines markedly different attribute-level results. Retaining these prediction targets separately identifies which attributes contribute to a model's overall advantage and which remain comparatively weak under their respective evaluation settings.

The choice of metric also changes comparisons within an individual attribute. On posterior features, QoQ-Med-7B achieves higher Accuracy than BreastGPT-8B (57.9 versus 54.8), whereas BreastGPT-

Table 16: **Multiple-choice evaluation on BUS-BRA by task type.** Results are reported as **Accuracy / Macro-F1** (%). † denotes thinking variants. Best results for each metric are shown in bold.

Model	Assessment & Diagnosis		
	AC	DX	PS
General-purpose VLMs			
InternVL3-8B	25.7 / 24.1	41.8 / 29.4	47.1 / 47.1
Qwen2.5-VL-7B	37.0 / 15.6	32.8 / 24.1	47.4 / 47.3
Qwen3-VL-8B-I	35.3 / 20.5	41.1 / 32.4	60.2 / 55.4
Qwen3-VL-8B-T†	31.2 / 20.9	32.2 / 23.8	56.9 / 52.3
Qwen2.5-VL-32B	34.1 / 21.2	37.8 / 26.0	58.6 / 53.2
Qwen3-VL-32B	36.2 / 20.8	43.2 / 31.9	55.8 / 55.6
Medical VLMs			
MedGemma-4B	22.6 / 22.4	15.8 / 15.1	60.0 / 51.8
Lingshu-7B	38.6 / 26.3	58.1 / 39.8	73.3 / 64.6
Hulu-Med-7B	26.3 / 25.4	54.6 / 45.0	66.6 / 61.8
QoQ-Med-7B†	36.9 / 14.5	39.9 / 24.7	35.4 / 29.4
BreastGPT-8B	**45.0** / **36.5**	**69.8** / **54.4**	**88.8** / **87.1**
Lingshu-32B	33.3 / 28.0	51.0 / 28.7	73.0 / 68.1

Table 17: **Multiple-choice evaluation on VinDr-Mammo by task type.** Results are reported as **Accuracy / Macro-F1** (%). Best results for each metric are shown in bold.

Model	Anatomy	Finding	Assessment & Diagnosis
	AA	FT	AC
General-purpose VLMs			
InternVL3-8B	45.7 / 23.8	19.1 / 8.7	3.9 / 4.2
Qwen2.5-VL-7B	64.7 / 26.6	29.8 / 10.2	10.7 / 9.6
Qwen3-VL-8B-I	42.3 / 26.2	16.3 / 9.0	61.4 / 22.0
Qwen3-VL-8B-T†	23.0 / 14.9	15.1 / 7.3	**63.4** / 19.9
Qwen2.5-VL-32B	64.4 / 28.0	54.0 / 12.7	27.3 / 15.9
Qwen3-VL-32B	57.1 / 28.0	**54.6** / 13.3	45.8 / 20.1
Medical VLMs			
MedGemma-4B	18.3 / 15.4	23.2 / 10.8	11.8 / 10.3
Lingshu-7B	55.0 / 28.3	32.2 / 12.3	12.7 / 11.3
Hulu-Med-7B	36.2 / 22.0	17.0 / 9.2	21.2 / 15.1
QoQ-Med-7B†	67.1 / 24.7	53.8 / 12.0	3.7 / 2.5
BreastGPT-8B	50.6 / **32.9**	35.7 / **17.0**	40.8 / **22.1**
Lingshu-32B	**70.3** / 27.2	42.7 / 14.2	45.8 / 18.3

Table 18: **Spatial evaluation on BUS-BRA and VinDr-Mammo.** Results report **mIoU** and **Acc@IoU**$_{0.5}$ (%). Best results for each metric are shown in bold.

Model	BUS-BRA Detection		VinDr Detection		VinDr Visual Grounding	
	mIoU	Acc@.5	mIoU	Acc@.5	mIoU	Acc@.5
General-purpose VLMs						
InternVL3-8B	26.0	11.2	3.3	0.2	3.2	0.3
Qwen2.5-VL-7B	28.9	22.2	7.6	3.9	8.8	6.5
Qwen3-VL-8B-I	51.1	56.4	10.7	7.0	13.7	11.0
Qwen3-VL-8B-T†	37.5	32.9	10.9	7.3	10.4	7.0
Qwen2.5-VL-32B	40.6	37.5	9.7	6.1	10.8	7.3
Qwen3-VL-32B	55.4	62.1	**15.4**	**12.9**	**18.0**	**16.8**
Medical VLMs						
MedGemma-4B	15.7	3.7	1.5	0.0	1.6	0.0
Lingshu-7B	4.7	0.9	4.4	0.8	3.6	0.7
Hulu-Med-7B	7.0	3.0	0.0	0.0	0.0	0.0
QoQ-Med-7B†	45.6	43.5	7.0	1.1	7.0	1.3
BreastGPT-8B	**72.6**	**82.8**	5.3	0.9	5.4	1.1
Lingshu-32B	12.5	4.4	4.7	1.5	5.6	2.0

8B achieves higher Macro-F1 (33.2 versus 25.8). The same ordering reversal occurs on halo: QoQ-Med-7B leads in Accuracy (72.0 versus 70.3), but BreastGPT-8B leads in Macro-F1 (63.7 versus 41.8). These reversals extend the metric-dependent rankings observed in the main results to atomic prediction targets. Accuracy weights individual items equally, while Macro-F1 averages class-specific F1 scores; consequently, the two metrics capture different aspects of performance over the answer space. Reporting both preserves distinctions that either metric alone would omit.

E.3.2 Changes across Model Scales and Thinking Variants

Similar aggregate scores can conceal large changes in the underlying task profile. Qwen3-VL-8B-T exceeds Qwen3-VL-8B-I by only 0.60 Accuracy points overall, but the task-level changes are much larger (Table 3). FA improves by 27.09 points, while AA, FT, AC, DX, and PS decrease by 9.37, 1.22, 0.53, 4.85, and 7.50 points, respectively. Under the reported equal weighting of task families, the FA improvement contributes approximately 4.52 points to Overall Accuracy, while the other five tasks jointly contribute approximately -3.91 points. The small net change thus results from opposing task-level changes rather than uniformly similar performance.

The FA improvement is itself concentrated in particular attributes. On BrEaST, halo Accuracy increases from 11.4 to 63.4 and skin-thickening Accuracy from 2.7 to 76.6, whereas posterior-feature Accuracy decreases from 37.3 to 22.2 (Table 15). With seven equally weighted attributes, the halo and skin-thickening increases together contribute approximately 18.0 points to the roughly 27.1-point FA improvement, based on the rounded attribute scores. Decomposing the comparison at both levels therefore reveals a pattern that the Overall score obscures: a substantial gain within one task family is driven largely by a subset of its prediction targets and offset by declines elsewhere.

Changes across model scales also vary by task. Within Lingshu, the 32B checkpoint improves Overall Accuracy from 47.00 to 50.73 and FA Macro-F1 from 38.66 to 43.01 relative to the 7B checkpoint, while DX Macro-F1 decreases from 29.29 to 24.54. In contrast, Qwen3-VL-32B improves over Qwen3-VL-8B-I on aggregate multiple-choice performance and all three reported breast-imaging spatial settings (Tables 3 and 4). These checkpoint comparisons show that aggregate improvements can have different task-level compositions. They support examining the resulting performance profile rather than using model scale or the thinking designation as a substitute for task-specific evaluation.

E.3.3 Performance across Imaging Settings and Spatial Tasks

Relative performance on categorical tasks does not consistently carry over to bounding-box prediction. On ISIC, Hulu-Med-7B achieves the highest diagnosis Accuracy and Macro-F1 among the evaluated checkpoints (56.3 and 34.4), but obtains 0.7 mIoU and 0.2 $\mathrm{Acc@IoU}_{0.5}$ on detection. Qwen3-VL-8B-I achieves lower diagnosis scores (53.2 and 22.8), yet substantially higher detection scores (80.3 and 91.4; Table 13). These results describe complementary task profiles within the same dataset family. Because diagnosis and detection use different ISIC annotation subsets, the comparison concerns dataset-level performance rather than paired success and failure on individual images. The spatial scores measure end-to-end box prediction, including response validity under the evaluation protocol.

Spatial rankings also change across imaging settings. BreastGPT-8B leads Qwen3-VL-32B on BUS-BRA detection, with $\mathrm{Acc@IoU}_{0.5}$ of 82.80 versus 62.08, whereas Qwen3-VL-32B leads on VinDr-Mammo detection, with 12.86 versus 0.90 (Table 4). This ranking reversal shows that a model's advantage within one localization setting need not persist in another. Reporting each dataset–task pair separately preserves this distinction and prevents a pooled spatial score from obscuring where the observed advantage applies.