Solve the Missing First Step: Can VLMs Standardize Raw Heterogeneous Medical Data?

Anonymous ACL submission

Abstract

As vision-language models (VLMs) are increasingly applied to medical AI, existing benchmarks mainly focus on evaluating their diagnosis ability over given medical images and texts, implicitly assuming that standardized medical images, texts or question–answer pairs are already prepared. However, this assumption does not hold when we apply VLMs in real clinical practice, where medical data is often raw, heterogeneous, and fragmented across different sources. In this paper, we study this missing step, *i.e.*, raw medical data standardization. Specifically, models are given raw dataset folders and evaluated on their ability to identify source formats, convert raw medical images into VLM-compatible visual inputs, extract relevant textual information, and organize the results into structured image-text pairs. To construct this Medical Data Standardization Benchmark (MDS-Bench), we manually annotate 1,939 raw medical data standardization tasks covering diverse clinical practice, radiology modalities, annotation format, and directory layouts. Extensive experiments show that even the best performing VLMs, *i.e.*, Gemini 3 Flash, achieve only 48.6% end-to-end success rate. Our research highlights raw medical data standardization as a critical bottleneck for medical AI diagnosis in real practice.

1 Introduction

With the rapid progress of VLMs (Li et al., 2023), building intelligent systems that can directly process real-world medical data is becoming increasingly feasible. However, existing medical benchmarks (Chen et al., 2024) usually start from curated images, reports, or question-answer pairs. This design is useful for evaluating perception and reasoning after the input has been cleaned, but it leaves this upstream standardization problem largely unexplored. In practical medical workflows, data are often stored as heterogeneous files, annotations, metadata records, masks, volumes, and dataset specific folders (Hosseinzadeh et al., 2025; Pezoulas et al., 2019; Barret et al., 2025; Zhang et al., 2022; Liu and Ye, 2025). Before a VLM can reason over these resources, it first identify the relevant sources, align them, convert them, and organize them into a usable representation (Wilkinson et al., 2016). This missing step limits our understanding of how well VLMs can operate on real medical data. Figure 1 illustrates this gap: real-world medical resources first be identified, aligned, and standardized before they can serve as VLM-ready image-text inputs.

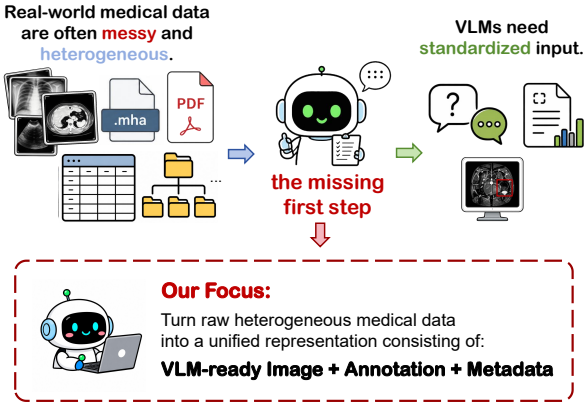


Figure 1: Motivation of raw medical data standardization. When medical AI systems are deployed in the real world, their inputs are not always standardized or VLM-ready. Instead of receiving well-prepared images, reports, or question–answer pairs, VLMs may need to process raw clinical data collected from different sources, where relevant files, annotations, metadata, and directory layouts are not explicitly aligned. Therefore, an upstream standardization step is needed convert raw medical inputs into representations that can be directly used by medical VLMs.

Raw medical data standardization is challenging because medical datasets rarely follow a single organization rule. A single sample may depend on multiple resources, such as a raw image, a segmentation mask (Wasserthal et al., 2023), a label table, a metadata record, or a slice from a volumetric

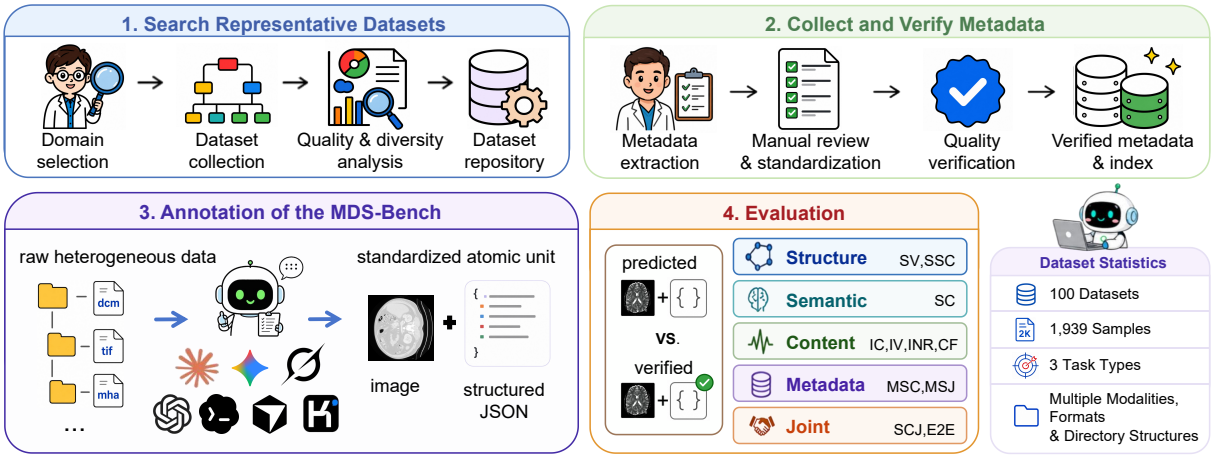


Figure 2: Overview of the MDS-Bench construction pipeline. We collect representative raw medical imaging datasets, trace target samples to their original visual sources, annotations, and metadata, and verify standardized image–JSON outputs against the raw resources. The resulting benchmark covers diverse task types, modalities, raw file formats, annotation styles, and directory structures.

scan. The links among these resources are often implicit in file names, directory structures, annotation conventions, or dataset documentation. Thus, standardization is not merely a format-conversion problem: a model must recover sample structure, select the correct sources, align annotations, and preserve task-relevant medical semantics. If this step fails, downstream inputs may appear complete while still being grounded in the wrong source, annotation, or clinical meaning.

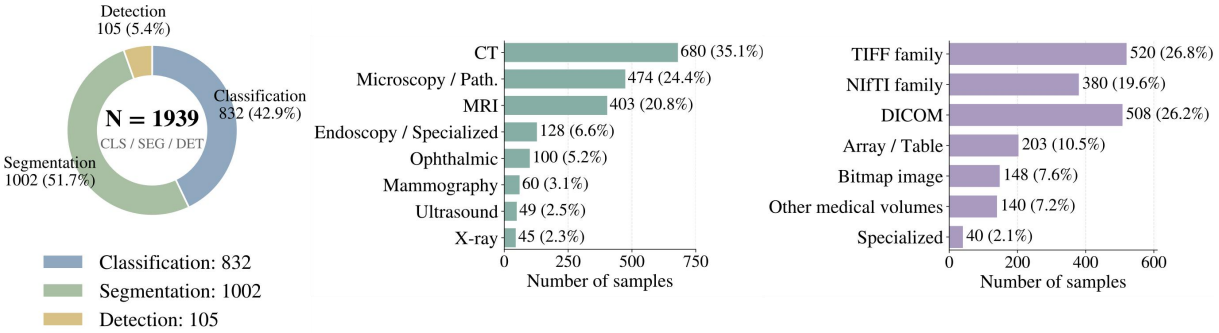
We study this missing step as raw medical data standardization in an agentic VLM setting (Yao et al., 2022; Akella and Narayanam, 2025; Li et al., 2024), where a VLM is embedded in a coding-agent environment and can interact with files, tools, and executable scripts. Specifically, in this task, a model receives a raw dataset folder. We expect our VLMs can identify source files, convert raw medical images into VLM-compatible visual inputs, align annotations and metadata, and generate a unified image and JSON output. The image provides a common visual representation for downstream models, while the JSON records information that cannot be captured by the image alone, such as source references, modality attributes, task type, labels, masks, boxes, and dataset metadata. This task differs from medical image classification, visual question answering, and report generation, which usually assume that the relevant input has already been selected. Our task starts earlier and asks whether VLMs can make raw medical resources usable while preserving their original meaning.

We build MDS-Bench to make raw medical data standardization measurable. As shown in Figure 2, the benchmark is constructed by collecting raw medical imaging datasets, tracing target samples to their original sources, and verifying standardized outputs against source evidence. It contains 1,939 samples across classification, segmentation, and detection, covering diverse modalities, file formats, annotation styles, and directory layouts. Since a model may produce valid JSON while choosing the wrong source or misaligning annotations, we further introduce an eleven-metric evaluation protocol that assesses image–JSON structure, sample-level semantics and fidelity, dataset-level metadata quality, and strict full-pipeline success.

Extensive experiments show that current VLMs still fail to standardize raw medical data reliably. For example, even the best performing VLMs, *i.e.*, Gemini 3 Flash, achieve only 48.6% end-to-end success rate, indicating that this upstream task remains far from solved The main difficulty is not schema generation alone, but preserving source based semantics, content fidelity, metadata consistency, and full pipeline correctness across heterogeneous raw resources. We provide a systematic empirical study showing that raw data standardization is a major bottleneck for applying agentic multimodal models to real medical data.

2 Related Work

Existing Medical AI Benchmarks. Existing medical AI benchmarks mainly evaluate perception, question answering, diagnosis, and reasoning after inputs have already been curated. Representa-



(a) Task type distribution (b) Modality distribution (c) Raw source format distribution

Figure 3: Dataset composition of MDS-Bench. The benchmark spans classification, segmentation, and detection tasks, and includes diverse medical imaging modalities and heterogeneous raw source formats.

tive benchmarks, including MultiMedQA (Singhal et al., 2023), BioASQ (Nentidis et al., 2025), PMC-VQA (Zhang et al., 2024), GMAI-MMBench (Chen et al., 2024), and MMMU (Yue et al., 2024), measure medical knowledge, biomedical question answering, multimodal understanding, or expert level reasoning. These benchmarks have advanced medical AI evaluation, but they usually start from selected images, organized question-answer pairs, or predefined annotations. Recent workflow-oriented systems, such as MedAgents (Tang et al., 2024), MedPerf (Karargyris et al., 2023), and MedAgentBench (Jiang et al., 2025), move closer to real deployment by studying agent collaboration and practical evaluation pipelines. However, they view data preparation as a prerequisite instead of an evaluative capability. Our work benchmarks this missing upstream step: whether VLMs can recover usable samples from raw heterogeneous medical resources, align visual content with annotations and metadata, and produce source-grounded image-text outputs.

Structured Output and Multimodal Structure Extraction. Structured Output and Multimodal Structure Extraction. Structured output generation and multimodal structure extraction study whether VLMs can produce outputs with explicit schemas. In text settings, StructEval (Yang et al., 2025) and JSONSchemaBench (Geng et al., 2025) evaluate schema following for language models. In multimodal settings, Image2Struct (Roberts et al., 2024) and SO-Bench (Feng et al., 2025) evaluate structured generation from visual inputs. These works show that schema compliance is important for VLM reliability, but their inputs are usually already prepared, such as clean images, text passages, schemas, or explicit instructions. They do not require VLMs to recover the raw data organization before producing structured outputs. In contrast, our benchmark starts from raw heterogeneous medical resources and evaluates source identification, visual conversion, annotation alignment, metadata organization, and schema-constrained generation in one standardization task.

3 Task and Benchmark Design

3.1 Task Definition

Raw medical data standardization evaluates whether a VLM can transform raw medical dataset resources into unified, source-grounded representations for downstream use. Given a raw dataset folder and a target schema, the model must identify relevant source files, convert processable visual evidence, align annotations and metadata, and map the recovered information into the required output format. Unlike medical image classification, detection, segmentation (Litjens et al., 2017), or visual question answering (Lin et al., 2023), this task does not assume that the image, label, annotation, and metadata have already been selected and aligned. It therefore targets the missing upstream step before conventional medical VLM evaluation.

The required output contains three components:

- **Standardized image:** converted or rendered from the selected raw source(s) and saved in a common visual format, so downstream models can access heterogeneous medical resources through a unified interface.
- **Per-image json annotation:** records sample-level semantics and provenance, including dataset identity, modality, anatomical context, task type, spatial attributes, source mapping, and task-specific annotations.
- **Dataset-level json annotation:** summarizes global information shared across samples,

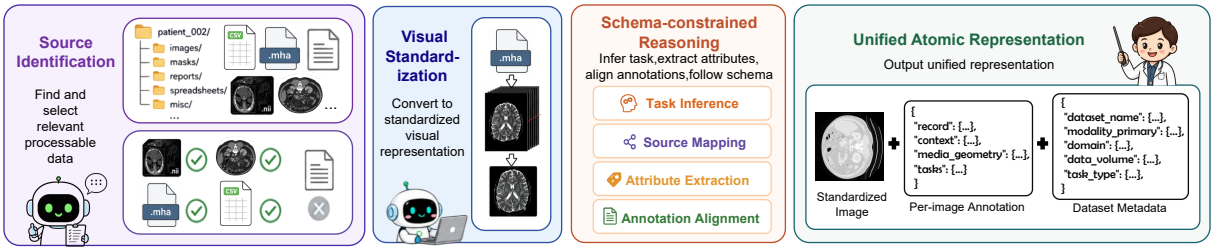


Figure 4: Staged raw medical data standardization workflow, where the agent identifies source evidence, standardizes visual inputs, aligns annotations and metadata, and outputs unified image-text atomic units.

such as dataset name, modality, data volume, task type, label space, and class information.

All annotations must conform to the target schema and remain grounded in evidence from the original raw resources. Figure 2 summarizes the benchmark pipeline from dataset collection and metadata verification to raw input processing, image–JSON output generation, and eleven-metric evaluation.

3.2 Benchmark Construction

Our MDS-bench is designed to evaluate standardization under raw data conditions rather than on previously curated image and annotation pairs. It contains 1,939 target samples across classification, segmentation, and detection. Each sample requires the VLMs to locate source files or source records, render visual evidence when needed, and map aligned information into the target schema. The resulting benchmark covers common medical vision tasks while retaining the upstream challenges of source selection, visual conversion, annotation alignment, and schema mapping that are often abstracted away in downstream VLM evaluation.

We select datasets to cover diversity in task type, modality, raw format, annotation style, and directory layout. As summarized in Figure 3 , the benchmark includes **51** segmentation datasets, **43** classification datasets, and **6** detection datasets. The complete dataset catalog is provided in Appendix A. It covers major modalities such as CT, MRI, microscopy, pathology, X-ray, ultrasound, ophthalmic imaging, endoscopy and supports various raw source formats, including DICOM, TIFF, NIfTI, bitmap images, array files, medical volumes, videos, and specialized formats. This diversity allows us to evaluate if VLMs can standardize raw medical resources across different storage formats, annotation standards, and directory structures, rather than just adhering to predefined schema.

Ground truth is constructed by tracing each target sample back to its original resources manually. For each listed instance, we identify the source files or source records, associated labels, masks, bounding boxes, tables, metadata entries, and visual conversion results. Model assisted extraction is used to draft the standardized target, but human verification is used to check the source mapping and all task relevant fields against the original dataset. If a drafted value is unsupported or conflicts with the raw resources, it is corrected or left empty. This process ensures that the ground truth is based on evidence from the dataset rather than on image appearance or prior medical assumptions.

3.3 Staged Reasoning Design

Raw medical data standardization cannot be handled as direct schema filling, because the required evidence is scattered across images, volumes, masks, label files, metadata, and directory structures. We therefore design the prompt as a staged reasoning workflow, as shown in Figure 4. The workflow first guides the agent to locate source evidence, then convert visual content, then reason over annotations and metadata, and finally check whether the generated outputs remain source-grounded and mutually consistent. This staged design fits the agentic VLM setting, where the model can inspect files, use tools, and run scripts before producing standardized outputs.

The first stage addresses source grounding. Before generating annotation, the agent must identify which raw files or records support the target sample and separate processable visual sources from auxiliary resources such as masks, label tables, and metadata files. Once the relevant source evidence is found, the second stage converts visual content into a common image representation. This conversion is not intended to create a visually plausible image in isolation; instead, it preserves the medical content needed for later annotation reasoning.

The third stage converts grounded evidence into

structured annotations. The agent infers task type, modality, dimension, anatomical information, labels, masks, boxes, and provenance based on support from raw files, annotations, metadata, file names, directory structure, or dataset documentation. The final stage checks consistency across the standardized image, the per-image annotation, and the dataset annotation. Together, these stages connect source identification, visual conversion, annotation alignment, and schema-constrained generation into one reasoning pipeline.

Raw medical data standardization is a coupled pipeline rather than a single schema generation task. A VLM is required to select the correct source, produce a valid structured output, recover the correct medical semantics, preserve task annotations, and keep dataset information consistent with the sample. If any step fails, the final standardized data may look usable but become unreliable. Therefore, our evaluation protocol measures both component quality and full pipeline success.

3.4 Evaluation Protocol

We organize the eleven metrics into five capability groups according to the main failure points in this chain, *i.e.*, **Structure metrics** include Schema Validity (SV) and Schema Semantic Composite (SSC). They test whether the generated JSON can serve as a valid standardized record and whether the chosen schema matches the actual sample semantics. This group is needed because a syntactically valid JSON may still describe the wrong modality, task, or annotation type.

Semantic metrics include Semantic Correctness (SC). SC evaluates the key fields that determine how the sample should be interpreted, such as dataset identity, modality, task type, data dimension, class information, and annotation type. Once the interpretation is correct, the next question is whether the sample annotation contains useful and reliable content. **Content metrics** include Information Completeness (IC), Information Validity (IV), Information Non Redundancy (INR), and Content Fidelity (CF). They measure whether required fields are filled, valid, concise, and faithful to the verified source evidence.

Standardized data also need correct dataset context, not only correct sample annotations. **Metadata metrics** include Metadata Semantic Correctness (MSC) and Meta Sample Joint Score (MSJ). They evaluate dataset information such as modality, task type, data volume, label availability, and class statistics, and further check whether this dataset information agrees with the generated sample content. Finally, **Joint metrics** include Source Content Joint Score (SCJ) and End-to-End Strict Pass (E2E). They test whether source selection, structure, semantics, and content fidelity succeed together, which gives the strictest view of real standardization success.

Source matching is used as a gate for field evaluation. For a dataset with N target samples, let s_i denote the source matching score of sample i. Metrics that compare generated fields with verified sample ground truth, including SC, IC, IV, INR, and CF, are averaged only over samples with $s_i > 0$. This avoids giving field credit to an output that is based on an unrelated source. In contrast, SV, SSC, SCJ, and E2E are computed over all N samples, so missing outputs and wrong source outputs are penalized directly.

We define SCJ as

$$\mathrm{SCJ} = \frac{1}{N}\sum_{i=1}^{N} s_i f_i, \tag{1}$$

where f_i is the content fidelity score. For E2E, let p_i indicate whether the model produces a valid image and JSON pair, g_i denote schema validity, and q_i denote semantic correctness. The strict pass indicator is

$$\begin{aligned} e_i = \mathbb{I}\big[p_i \wedge s_i = 1 \wedge g_i \geq 0.85 \\ \wedge\, q_i \geq 0.5 \wedge f_i \geq 0.5\big]. \end{aligned} \tag{2}$$

The final E2E score is

$$\mathrm{E2E} = \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{3}$$

We use $g_i \geq 0.85$ rather than perfect schema validity because a minor schema defect should not erase credit from an otherwise correct sample.

All metrics are computed within each dataset and then averaged across the 100 datasets with equal weight. This prevents large datasets from dominating the benchmark and better reflects robustness across modalities, source formats, annotation styles, and directory layouts. Full metric definitions and details are provided in Appendix C.

4 Experiment

4.1 Experimental Setup

We evaluate frontier agentic VLMs on MDS-Bench. Each model receives the raw dataset directory and

Model	Structure		Semantic	Content				Metadata		Joint	
	SV	SSC	SC	IC	IV	INR	CF	MSC	MSJ	SCJ	E2E
Claude Haiku 4.5 (Anthropic, 2025)	82.2	49.8	49.1	47.5	41.5	42.5	40.3	56.8	34.1	38.0	24.9
Grok 4.20 (xAI, 2026)	80.0	48.8	55.7	49.0	45.5	46.7	41.7	57.8	35.6	40.0	21.5
Kimi K2.5 (Kimi Team et al., 2026)	83.6	51.7	55.1	54.5	45.5	47.2	46.7	63.4	40.4	44.8	29.4
GPT-5.2 (OpenAI, 2025b)	81.6	52.8	60.2	50.0	37.5	47.3	47.7	74.6	43.4	45.4	23.6
GPT-5.2-Codex (OpenAI, 2025a)	83.7	53.3	57.0	55.5	49.0	48.1	47.1	65.6	40.6	44.8	28.1
Composer 1.5 (Cursor, 2026)	81.9	54.0	59.9	50.5	41.3	46.8	48.5	72.6	43.4	46.7	30.8
Claude Sonnet 4.6 (Anthropic, 2026b)	85.2	56.0	62.5	60.0	56.3	50.5	49.9	69.0	42.7	48.6	33.8
Claude Opus 4.6 (Anthropic, 2026a)	83.9	56.2	**65.2**	59.5	51.8	51.7	53.7	74.4	48.0	52.3	30.5
Gemini 3 Flash (Google DeepMind, 2025)	**88.2**	**62.0**	63.9	**67.5**	**62.5**	**58.7**	**58.7**	**78.0**	**53.7**	**56.3**	**48.6**

Table 1: Benchmark performance on the MDS-bench for nine VLMs with eleven evaluation metrics. All values are percentages (%) averaged across 1,939 samples. The best result in each column is highlighted in green, and the second-best result is highlighted in light green.

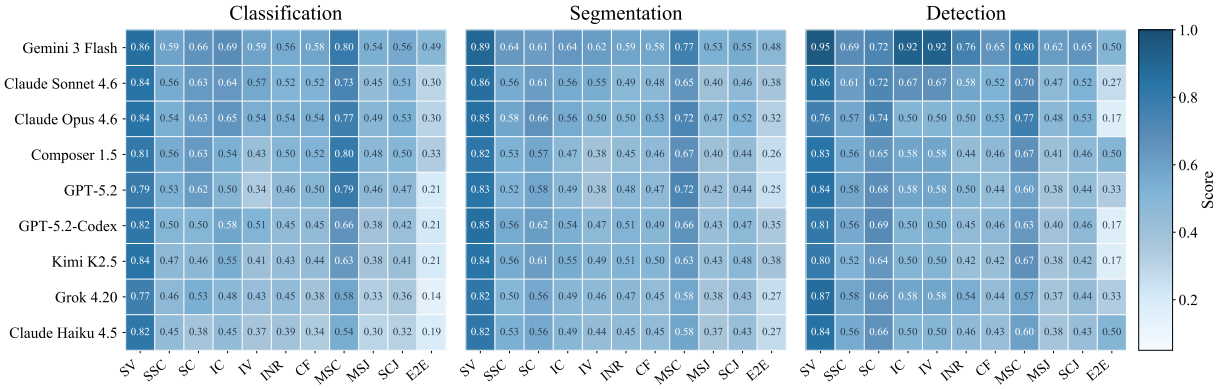


Figure 5: Per-task performance of all evaluated models on the proposed benchmark. Heatmaps report scores on the eleven evaluation metrics for classification, segmentation, and detection subsets, respectively. Rows correspond to models, columns correspond to metrics, and darker colors indicate higher scores.

the target schema, and is required to identify relevant source files or records, convert processable medical data into standardized images, and generate both per-image annotations and a dataset-level annotation. All models are evaluated with the same output format and scoring protocol, and no outputs are manually corrected after generation. We compare Claude Haiku 4.5 (Anthropic, 2025), Claude Sonnet 4.6 (Anthropic, 2026b), Claude Opus 4.6 (Anthropic, 2026a), GPT-5.2 (OpenAI, 2025b), GPT-5.2-Codex (OpenAI, 2025a), Gemini 3 Flash (Google DeepMind, 2025), Grok 4.20 (xAI, 2026), Kimi K2.5 (Kimi Team et al., 2026), and Composer 1.5 (Cursor, 2026). For each model, all metrics are first computed within each dataset and then averaged across datasets with equal weight.

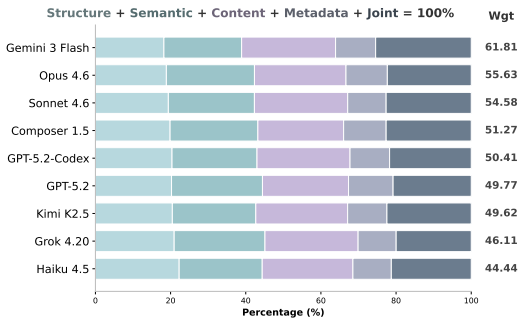


Figure 6: Weighted capability composition on MDS-Bench. Each stacked bar decomposes a model's normalized performance into Structure, Semantic, Content, Metadata, and Joint categories using weights 0.15, 0.20, 0.25, 0.10, and 0.30. The rightmost value reports the overall weighted score (Wgt).

4.2 Overall Results

Models can often satisfy the output schema, but schema validity does not imply correct standardization. As shown in Table 1, most models obtain relatively high SV, ranging from 80.0% to 88.2%. However, SSC is much lower than expected, ranging from 48.8% to 62.0%. This gap shows that models can generate JSON files with the required fields, while still making mistakes in task branch, modality, source reference, or annotation semantics. Thus, schema following is only a basic requirement for raw medical data standardization, not a reliable

Model	Structure		Semantic		Content		Metadata		Joint	
	Score (%)	# Fail	Score (%)	# Fail	Score (%)	# Fail	Score (%)	# Fail	Score (%)	# Fail
Claude Haiku 4.5 (Anthropic, 2025)	66.0	10	49.1	41	42.9	58	45.4	57	31.5	75
Grok 4.20 (xAI, 2026)	64.4	13	55.7	31	45.7	64	46.7	63	30.7	78
Kimi K2.5 (Kimi Team et al., 2026)	67.7	9	55.1	35	48.5	54	51.9	52	37.1	69
GPT-5.2 (OpenAI, 2025b)	67.2	4	60.2	25	45.6	67	59.0	34	34.5	76
GPT-5.2-Codex (OpenAI, 2025a)	68.5	5	57.0	27	49.9	49	53.1	44	36.5	72
Composer 1.5 (Cursor, 2026)	67.9	6	59.9	25	46.8	53	58.0	36	38.7	70
Claude Sonnet 4.6 (Anthropic, 2026b)	70.6	**1**	62.5	23	54.2	45	55.8	43	41.2	65
Claude Opus 4.6 (Anthropic, 2026a)	70.0	3	**65.2**	**20**	54.2	45	61.2	34	41.4	65
Gemini 3 Flash (Google DeepMind, 2025)	**75.1**	4	63.9	23	**61.8**	**39**	**65.9**	**29**	**52.4**	**51**

Table 2: Error attribution across five capability categories. For each category, Score reports the mean category score, and # Fail reports the number of datasets whose category score is below 50%. Higher Score and lower # Fail indicate better performance.

indicator of task success.

Joint metrics show that current models still fail to complete the full standardization pipeline reliably. SCJ ranges from 38.0% to 56.3%, and E2E ranges from 21.5% to 48.6%. These scores are much lower than structure metrics because they require several conditions to hold together: a valid image-text pair, correct source selection, sufficient schema quality, semantic correctness, and content fidelity. The low joint scores indicate that models can often solve individual parts of the task, but errors accumulate when the complete raw-data-to-standardized-output pipeline is evaluated.

Gemini 3 Flash is the strongest overall model, but even the best model remains far from reliable end-to-end standardization. Gemini 3 Flash obtains the best score on most metrics, including SV, SSC, IC, IV, INR, CF, MSC, MSJ, SCJ, and E2E. Claude Opus 4.6 achieves the highest SC, suggesting stronger recovery of key task semantics. Nevertheless, Gemini 3 Flash reaches only 48.6% E2E, meaning that more than half of the target samples still fail at least one required condition. This result highlights the gap between strong component-level performance and reliable full-pipeline standardization.

The same structure-joint gap appears across task subsets, while model rankings vary under different data conditions. Figure 5 shows that schema-level metrics remain higher than joint metrics across classification, segmentation, and detection, with E2E remaining difficult in all three task types. Additional cross-group results in Appendix B show that Gemini 3 Flash is the most robust model overall, while Claude Opus 4.6 performs best on Microscopy / Pathology datasets.

Weighted capability composition further shows that overall performance is driven by content and joint reliability. Figure 6 summarizes model performance through a capability-weighted decomposition. Since Joint and Content receive the largest weights, models with stronger full-pipeline consistency and content fidelity obtain higher overall weighted scores. Gemini 3 Flash ranks first, indicating that its advantage comes not only from schema following but also from stronger content recovery and joint reliability. In contrast, models with reasonable Structure or Semantic contributions can still rank lower when their Joint component is weak. This pattern reinforces our main finding that raw medical data standardization is primarily bottlenecked by source-grounded content preservation and end-to-end consistency.

4.3 Error Attribution Analysis

Dataset-level failure rates show that errors are concentrated in Content and Joint, while Structure failures are rare. Table 2 groups the eleven metrics into five categories and reports both the mean category score and the percentage of datasets whose category score falls below 0.5. Structure has the lowest failure rate for all models, ranging from 1% to 13%, which means that most datasets do not fail because the model breaks the output format. In contrast, Content and Joint have much higher failure rates. For example, Content fails on 67% of datasets for GPT-5.2, 64% for Grok 4.20, and 58% for Claude Haiku 4.5, while Joint fails on more than half of the datasets for every model. This pattern indicates that the primary failures arise when models are required to maintain labels, masks, boxes, source records, and task-specific annotation payloads, as opposed to simply adhering to the schema.

The category breakdown also shows that model strengths do not transfer uniformly across levels of standardization. Gemini 3 Flash has the strongest overall error profile, but other models show different strengths and weaknesses.

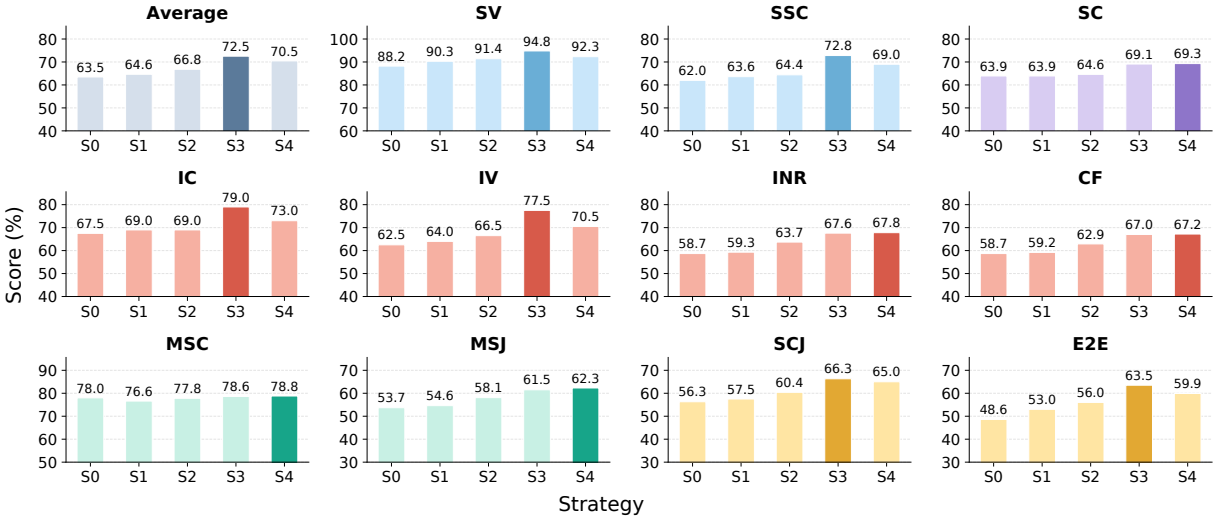


Figure 7: Inference strategy ablation on MDS-Bench using Gemini 3 Flash as the base model. Strategies S0-S4 compare direct generation with increasingly validation-guided refinement and selection procedures. The first panel reports the average score across all eleven metrics, while the remaining panels show metric-specific changes in structure, semantic, content, metadata, and joint performance.

Claude Opus 4.6 achieves the best Semantic score, yet its Joint failure rate remains 65%. GPT-5.2 and Composer 1.5 obtain relatively strong Metadata scores, but both remain weak in Content and Joint. This suggests that models may recover dataset-level context or key task semantics while still failing to preserve reliable sample-level content. The key takeaway from this analysis is that enhancing Structure or Metadata alone is inadequate; effective standardization of raw medical data necessitates improved Content recovery and greater Joint reliability across datasets.

4.4 Inference Strategy Ablation

Validation based inference improves overall standardization quality. Figure 7 compares five strategies using Gemini 3 Flash. S0 directly generates the standardized output, S1 adds self refinement, S2 adds verification guided refinement, S3 selects the best complete candidate with a validator, and S4 applies validator guided field wise selection. Compared with S0, S1 and S2 steadily improve E2E from 48.6% to 53.0% and 56.0%, while SCJ increases from 56.3% to 57.5% and 60.4%. This demonstrates that additional verification can minimize certain source and content-related errors without altering the base model.

S3 is most effective because complete-candidate selection yields the largest gains on joint metrics. S3 achieves the best overall results, with 66.3% SCJ and 63.5% E2E, improving E2E by 14.9 points over S0. It also obtains the highest scores on SV, SSC, IC, and IV. These gains suggest that selecting the best full output is especially useful for raw data standardization, where source correctness, schema quality, semantic correctness, and content fidelity must hold at the same time.

Field wise selection improves local quality but does not produce the best end to end reliability. S4 obtains the best scores on SC, INR, CF, MSC, and MSJ, indicating stronger local semantic, content, and metadata fields. However, its SCJ and E2E are lower than S3, at 65.0% and 59.9%. This suggests that combining locally better fields does not necessarily produce the most coherent standardized sample. Overall, the ablation results indicate that selecting a comprehensive candidate is more effective than independently patching fields.

5 Conclusion

We introduced raw medical data standardization as an upstream task for medical AI and built MDS-Bench to evaluate VLM agents on heterogeneous raw medical resources. Experiments show that current models can often satisfy schema-level requirements, but still struggle with Content and Joint, leading to low E2E performance even for Gemini 3 Flash. Validation based inference improves results, especially through complete-candidate selection, but does not eliminate the core bottleneck. These findings show that reliable medical AI requires not only structured output generation, but also faithful content recovery and source-aligned standardization from raw data.

Limitations

This work has several limitations. The benchmark is built from publicly available medical imaging datasets and may not fully reflect real clinical data environments with private PACS systems, electronic health records, access-control constraints, or incomplete metadata. It focuses on classification, detection, and segmentation, leaving other medical AI settings such as report generation, longitudinal modeling, and treatment-oriented decision support for future work. The evaluation protocol uses schema-based and source-grounded checks, which may miss subtle medical interpretation errors or penalize alternative valid outputs. Ground truth construction relies on large-model-assisted extraction followed by human verification, so residual annotation errors may remain. The benchmark is intended for research evaluation only, and model-generated standardized medical data should not be used for clinical decision making without expert verification.

References

Ashlesha Akella and Krishnasuri Narayanam. 2025. Data wrangling task automation using code-generating language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29616–29618.

Anthropic. 2025. Claude Haiku 4.5 system card.

Anthropic. 2026a. Claude Opus 4.6 system card.

Anthropic. 2026b. Claude Sonnet 4.6 system card.

Nelly Barret, Anna Bernasconi, Boris Bikbov, and Pietro Pinoli. 2025. I-etl: an interoperability-aware health (meta) data pipeline to enable federated analyses. *BMC Medical Informatics and Decision Making*, 25(1):375.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, and 1 others. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427.

Cursor. 2026. Introducing Composer 1.5.

Di Feng, Kaixin Ma, Feng Nan, Haofeng Chen, Bohan Zhai, David Griffiths, Mingfei Gao, Zhe Gan, Eshan Verma, Yinfei Yang, and 1 others. 2025. So-bench: A structural output evaluation of multimodal llms. *arXiv preprint arXiv:2511.21750*.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. Jsonschemabench: A rigorous benchmark of structured outputs for language models. *arXiv preprint arXiv:2501.10868*.

Google DeepMind. 2025. Gemini 3 Flash model card.

Elham Hosseinzadeh, Marziyeh Afkanpour, Mehri Momeni, and Hamed Tabesh. 2025. Data quality assessment in healthcare, dimensions, methods and tools: a systematic review. *BMC Medical Informatics and Decision Making*, 25(1):296.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. 2025. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9):AIdbp2500144.

Alexandros Karargyris, Renato Umeton, Micah J Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, and 1 others. 2023. Federated benchmarking of medical artificial intelligence with medperf. *Nature machine intelligence*, 5(7):799–810.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

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

Lan Li, Liri Fang, and Vetle I Torvik. 2024. Autodcworkflow: Llm-based data cleaning workflow auto-generation and benchmark. *arXiv preprint arXiv:2412.06724*.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Congcong Liu and Fangfang Ye. 2025. A review of multimodal medical data fusion techniques for personalized medicine. In *Proceedings of the 4th International Conference on Biomedical and Intelligent Systems*, pages 338–347.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodríguez-Ortega, Eduard Rodriguez-López, Natalia Loukachevitch, Andrey Sakhovskiy, Elena Tutubalina, Dimitris Dimitriadis, and 1 others. 2025. Overview of bioasq 2025: The thirteenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 173–198. Springer.

OpenAI. 2025a. Addendum to GPT-5.2 system card: GPT-5.2-Codex.

OpenAI. 2025b. Update to GPT-5 system card: GPT-5.2.

Vasileios C Pezoulas, Konstantina D Kourou, Fanis Kalatzis, Themis P Exarchos, Aliki Venetsanopoulou, Evi Zampeli, Saviana Gandolfo, Fotini Skopouli, Salvatore De Vita, Athanasios G Tzioufas, and 1 others. 2019. Medical data quality assessment: On the development of an automated framework for medical data curation. *Computers in biology and medicine*, 107:270–283.

Josselin S Roberts, Tony Lee, Chi H Wong, Michihiro Yasunaga, Yifan Mai, and Percy Liang. 2024. Image2struct: Benchmarking structure extraction for vision-language models. *Advances in Neural Information Processing Systems*, 37:115058–115097.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, and 1 others. 2023. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, and 1 others. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

xAI. 2026. Grok 4.20 model documentation.

Jialin Yang, Dongfu Jiang, Lipeng He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, and 1 others. 2025. Structeval: Benchmarking llms' capabilities to generate structural outputs. *arXiv preprint arXiv:2505.20139*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2024. *URL https://arxiv.org/abs/2305.10415*, 40.

Yong Zhang, Ming Sheng, Xingyue Liu, Ruoyu Wang, Weihang Lin, Peng Ren, Xia Wang, Enlai Zhao, and Wenchao Song. 2022. A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. *Health Information Science and Systems*, 10(1):22.

A Full Dataset Catalog

Table 3–5 list the complete set of datasets used in our benchmark. Some Zenodo-derived datasets are listed using their record-ID-based raw folder names, such as entries ending with `_folder`. These names are retained to preserve direct traceability to the original downloaded raw folders.

B Cross-Group Generalization Analysis

We provide additional results by task type and medical modality. This analysis examines whether model rankings remain stable across different subsets of the benchmark. For each model m and dataset d, we first compute the simple average over the eleven metrics:

$$\mathrm{Overall}(m, d) = \frac{1}{11} \sum_{k=1}^{11} v_k(m, d), \qquad (4)$$

where $v_k(m, d)$ denotes the score of metric k. For a group g with dataset set $\mathcal{D}_g$, the group score is:

$$\mathrm{Group}(m, g) = \frac{1}{|\mathcal{D}_g|} \sum_{d \in \mathcal{D}_g} \mathrm{Overall}(m, d). \qquad (5)$$

All values are reported as percentages. Groups with only three datasets are included for completeness and should be interpreted as descriptive evidence.

B.1 Results by Medical Modality

Table 6 reports model performance across medical modality groups. Gemini3 achieves the highest average score overall and ranks first in CT, MRI, Ultrasound, Ophthalmic Imaging, and X-ray. Opus4.6 performs best on Microscopy / Pathology, while Sonnet4.6 performs best on Endoscopy and Composer1.5 performs best on Mammography. These results show that the strongest model overall is not always the best model for every medical modality. Table 7 further summarizes cross-modality rank stability, showing that Gemini 3 Flash achieves the most stable overall ranking across modalities, although several models remain competitive on specific modality groups.

B.2 Results by Task Type

Table 8 reports results by task type. Gemini3 ranks first on classification, segmentation, and detection, showing the most stable task-level generalization. Opus4.6 and Sonnet4.6 remain strong on classification and segmentation, while the detection group shows larger ranking changes. Because detection contains only six datasets, these results should be treated as indicative rather than definitive. Table 9 further reports task-level rank stability, confirming that Gemini 3 Flash is the only model ranked first across all three task types, while other models show larger rank variation, especially on detection.

B.3 Summary

The cross-group analysis shows that Gemini3 is the most robust model across both task types and medical modalities. It ranks first on all three task types and has the best average modality rank. However, model performance is not fully invariant across data conditions. Opus4.6 performs best on Microscopy / Pathology, Sonnet4.6 performs best on Endoscopy, and Composer1.5 performs best on Mammography. These results suggest that raw medical data standardization difficulty depends not only on the target task, but also on the medical modality and the conventions of the underlying data source.

C Evaluation Details

This appendix provides the full definition of the eleven-metric evaluation protocol used in the main experiments. The protocol is designed to evaluate whether a model can transform raw heterogeneous medical data into source-grounded standardized images, per-image annotations, and dataset annotations. It therefore checks not only output format, but also source selection, semantic alignment, information quality, content fidelity, and dataset-level metadata quality.

C.1 Notation

For each dataset, let N denote the number of target samples specified by the sample manifest. For sample i, the evaluator records the following quantities:

- $p_i \in \{0, 1\}$: whether the sample forms a valid image-text pair. This requires a readable image, a parseable JSON annotation, and a JSON image path that matches the actual output image.
- $s_i \in [0, 1]$: source matching score between the predicted source information and the verified raw source file or source record.
- $g_i \in [0, 1]$: schema-validity score, computed as the fraction of schema checks passed.
- $q_i \in [0, 1]$: semantic correctness score for sample-level fields.

Dataset name	Modality	Task type	Format
Micro_Ultrasound_Prostate_Segmentation_Dataset	Micro-Ultrasound	segmentation	nii
BONBID-HIE	MRI	segmentation	mha
4939348_Elastographic_Map_Simulation	Ultrasound	classification	mat
4975777_folder	Microscopy	classification	czi
RetinalAVM_CaseStudy_5002837	Ophthalmic	classification	tif
PBC_Brain_MRI_5005236	MRI	classification	nii.gz
Organoid_MRI_Longitudinal	MRI	classification	dcm
Placenta_HSI	Hyperspectral	segmentation	tif
15680730_folder	CT	classification	mha
Granger2018_AOSLO_RPE_PR_ROIs	Adaptive optics	segmentation	tif
SLIMfast4C_HeLa_meGFP_MBP_TMD_Zenodo5712332	Microscopy	classification	tif
Zenodo_5903190_Multiplex_Tissue_Imaging	Microscopy	segmentation	tiff
6355622_folder	NA	segmentation	mat
NYU_POAG_RetinalOCT_Glaucoma	OCT	classification	mha
ICIAR2018_BACH_BreastHistology	Microscopy	classification	tif
BACH_ICAR_2018	Microscopy	classification	tif
Brain_MRI_segmentation	MRI	segmentation	tif
breast-cancer-microwave	Microwave imaging	classification	npy
Brain_Tumor_Dataset	MRI	classification	tif
capsule_endoscopy_dataset_kauhc2025	Capsule endoscopy	classification	bmp
Cervical_Cancer_largest_dataset	Microscopy	classification	bmp
chest-xray-unidatapro	X-ray	detection	dcm
colorectal_histology_mnist2019	Histopathology	classification	tif
ct-of-the-spine-scoliosis	CT	classification	dcm
edd2020_endoscopy_detection_and_segmentation2025	Endoscopy	segmentation	tif
expansion-microscopy	Microscopy	segmentation	tif
fazekas-mri	MRI	classification	dcm
fomo-mri	MRI	classification	nii.gz
fundus-image-segmentation	Fundus	segmentation	tif
gastrointestinal_bleeding_images_dataset2024	Endoscopy	classification	bmp
human-protein-atlas-cells-dataset	Microscopy	segmentation	parquet
isic-skin-cancer-train-npz	Dermoscopy	segmentation	npz
kits23-kidney-cancer	CT	segmentation	nii.gz
leukemia-cancer-dataset	Microscopy	classification	bmp

Table 3: Complete dataset catalog used in the proposed benchmark (Part 1 of 3).

- $c_i \in [0, 1]$: information completeness score.
- $d_i \in [0, 1]$: information validity score.
- $r_i \in [0, 1]$: information non-redundancy score.
- $f_i \in [0, 1]$: content fidelity score.

We define the set of source-matched samples as

$$S_{\text{src}} = \{i \mid s_i > 0\}, \qquad N_{\text{src}} = |S_{\text{src}}|. \tag{6}$$

Field-level metrics that require a meaningful comparison with the target ground truth are computed over S_{src}. Joint metrics are computed over all N target samples, so wrong-source and missing-output cases are penalized directly.

C.2 Output Validity and Source Matching

The evaluator first verifies whether each sample has a valid paired output. A pair is considered valid only when the standardized image exists, can be opened as an image file, the JSON annotation exists and can be parsed, and the image path recorded in the JSON points to the actual standardized image. This check produces p_i.

Source matching evaluates whether the predicted source fields can be traced to the verified raw source. The evaluator normalizes source identifiers by handling common path, filename, extension, and stem variations, and compares them against the ground-truth source stem or equivalent source identifiers. The score s_i is set to 1 when the predicted source is equivalent to the verified source and trace evidence is present; otherwise it is set to 0. This check is used as a gate for field-level semantic and content evaluation, because evaluating annotations against the target sample is not meaningful when the model has selected an unrelated source.

C.3 Structure Metrics

Schema Validity (SV). Schema validity measures whether the sample-level JSON follows the required schema. The evaluator checks required top-level blocks, record fields, image-path references, context fields, media fields, task-specific

Dataset name	Modality	Task type	Format
medical-7-2020	MRI	classification	mat
melanoma-train	Dermoscopy	segmentation	bmp
microscopy-demo	Microscopy	classification	tif
mri-rd	MRI	classification	dcm
npz-histopathology-dataset	Microscopy	classification	npz
pancreas-mri	MRI	classification	dcm
tcga-lgg-mribrats	MRI	segmentation	tif
tumorsimulations	MRI	segmentation	nii.gz
adenocarcima	CT	classification	npy
brain-organoids-segmentation-dataset	Microscopy	segmentation	tif
BraTS2020_TrainingValidation	MRI	segmentation	nii
BraTS2020_training_data	MRI	segmentation	h5
breast-cancer-cell-segmentation	Microscopy	segmentation	tif
SIPaKMeD	Microscopy	classification	bmp
computed-tomography-ct-of-the-brain	CT	classification	dcm
DMID-Breast-Cancer-Mammography	Mammography	segmentation	tif
DRIVE	Fundus	segmentation	tif
echonet-pediatric	Ultrasound	classification	avi
electron-microscopy-dataset	Microscopy	segmentation	tif
finding-lungs-in-ct-data	CT	segmentation	tif
fluorescence-microscopy-image-denoising-dataset	Microscopy	segmentation	h5
fuse-my-cells-part-01	Microscopy	classification	tif
fuse-my-cells-part-02	Microscopy	classification	tif
kidney-nori-annotated-images	Microscopy	segmentation	tif
C-NMC_Leukemia	Microscopy	classification	bmp
malignant-lymphoma-classification	Microscopy	classification	tif
MIAS-Mammography	Mammography	classification	pgm
mosmed-covid19-ct-scans	CT	classification	nii
TheChineseMammographyDatabase	Mammography	classification	dcm
PROSTATE_MRI	MRI	classification	dcm
BTCV-TCIA-MultiOrgan-Abdominal	CT	segmentation	nii.gz
Comprehensive-3D-Medical-Dataset	CT	segmentation	nii.gz
Lung-PET-CT-Dx	CT	detection	dcm
Tibial-Plateau-Fracture	CT	classification	mat

Table 4: Complete dataset catalog used in the proposed benchmark (Part 2 of 3).

branches, spatial branches, and branch exclusivity. Let g_i be the fraction of schema checks passed for sample i. Then

$$\mathrm{SV} = \frac{1}{N} \sum_{i=1}^{N} g_i. \tag{7}$$

Schema-Semantic Composite (SSC). SV alone can be high when a model generates a well-formed but semantically incorrect JSON. SSC therefore couples schema validity with semantic correctness:

$$\mathrm{SSC} = \frac{1}{N} \sum_{i=1}^{N} g_i q_i. \tag{8}$$

This metric penalizes outputs that satisfy the surface schema but assign incorrect task, modality, dimension, dataset identity, or annotation semantics.

C.4 Semantic Metric

Semantic Correctness (SC). SC evaluates whether key sample-level fields match the verified ground truth. These fields include dataset identity, task type, modality, data dimension, class information, and annotation type. Since these fields are meaningful only after a related source is identified, SC is computed over the source-matched set:

$$\mathrm{SC} = \frac{1}{N_{\mathrm{src}}} \sum_{i \in S_{\mathrm{src}}} q_i. \tag{9}$$

If no sample passes the source gate, the score is set to 0.

C.5 Information and Fidelity Metrics

Information Completeness (IC). IC measures whether required task-relevant fields are populated. It checks whether expected fields in the sample context, auxiliary fields, and task branch are present rather than missing:

$$\mathrm{IC} = \frac{1}{N_{\mathrm{src}}} \sum_{i \in S_{\mathrm{src}}} c_i. \tag{10}$$

Information Validity (IV). IV measures whether populated fields contain valid and evidence-

Dataset name	Modality	Task type	Format
Brain-MP2RAGE-MPRAGE	MRI	segmentation	nii.gz
CHAOS-Abdominal	CT	segmentation	dcm
LNDb-Lung-Nodule	CT	segmentation	mhd
Totalsegmentator	CT	segmentation	nii.gz
L1-L5-Spine-FineSeg	CT	segmentation	nii.gz
HaN-Seg	CT	segmentation	nrrd
MoNuSAC-Cell-Detection	Microscopy	detection	tif
FLARE22	CT	segmentation	nii.gz
Abdominal-CT-Scans	CT	segmentation	nii
DICOM-Brain-MRI	MRI	classification	dcm
RSNA-Heart-Detection	X-ray	detection	dcm
Hippocampus	MRI	segmentation	nii
KiTS23	CT	segmentation	nii.gz
LiTS-Liver-Tumor-Segmentation	CT	segmentation	nii
LUNA16	CT	detection	mhd
M3D-RefSeg	CT	segmentation	nii.gz
Prostate-MSD	MRI	segmentation	nii
SIIM-Medical-Images	CT	classification	dcm
Breast-MRI-NACT-Pilot	MRI	segmentation	dcm
CPTAC-CCRCC	CT+MRI	segmentation	dcm
CPTAC-PDA	CT+MRI	segmentation	dcm
CPTAC-UCEC	CT+MRI+PET	segmentation	dcm
LIDC-IDRI	CT	segmentation	dcm
Lumbar_Spine_Segmentation	MRI	segmentation	mha
MIDRC-RICORD-1A	CT	classification	dcm
NSCLC-Radiomics	CT	segmentation	dcm
QIN-BREAST	CT+MRI+PET	classification	dcm
Spine-Mets-CT-SEG	CT	segmentation	dcm
rsna-pneumonia-detection-challenge	X-ray	detection	dcm
soft_tissue_sarcoma	CT+MRI+PET	segmentation	dcm
tcga-luad	CT+PET	classification	dcm
LCTSC	CT	segmentation	dcm

Table 5: Complete dataset catalog used in the proposed benchmark (Part 3 of 3).

supported values. It penalizes placeholders, unsupported guesses, empty tokens, and invalid default values:

$$\mathrm{IV} = \frac{1}{N_{\mathrm{src}}} \sum_{i \in S_{\mathrm{src}}} d_i. \tag{11}$$

Information Non-Redundancy (INR). INR measures whether the output avoids repetitive or redundant field filling. It rewards outputs that provide compact, non-duplicative information while still preserving required evidence:

$$\mathrm{INR} = \frac{1}{N_{\mathrm{src}}} \sum_{i \in S_{\mathrm{src}}} r_i. \tag{12}$$

Content Fidelity (CF). CF measures whether recovered task-specific content matches the verified ground truth. This includes labels, class names, masks, boxes, source mappings, expected auxiliary fields, and task-specific annotation payloads:

$$\mathrm{CF} = \frac{1}{N_{\mathrm{src}}} \sum_{i \in S_{\mathrm{src}}} f_i. \tag{13}$$

Together, IC, IV, INR, and CF distinguish evidence-supported content recovery from merely producing a syntactically valid JSON file.

C.6 Metadata Metrics

Each dataset also contains a dataset-level metadata record. The evaluator checks whether this record correctly summarizes global dataset information, including dataset name, primary modality, task type, data volume, valid-image count, and class information.

Metadata Semantic Correctness (MSC). MSC measures whether dataset-level semantic fields match the verified dataset-level ground truth:

$$\mathrm{MSC} = m_{\mathrm{sc}}, \tag{14}$$

where m_{sc} denotes the metadata semantic correctness score.

Meta-Sample Joint Score (MSJ). Dataset-level metadata may be correct even when sample-level outputs are unfaithful. We therefore define a joint metadata score that couples metadata reliability with sample-level content fidelity:

$$\mathrm{MSJ} = (m_{\mathrm{sv}} m_{\mathrm{sc}} m_{\mathrm{cf}})^{1/3} \cdot \mathrm{CF}, \tag{15}$$

Model	CT (%)	Microscopy / Path. (%)	MRI (%)	Ultrasound (%)	Ophthalmic (%)	Endoscopy (%)	X-ray (%)	Mammo. (%)	Mean (%)
Gemini 3 Flash (Google DeepMind, 2025)	**65.5**	59.5	**64.3**	**72.6**	**67.2**	59.8	**66.4**	55.8	**63.9**
Claude Opus 4.6 (Anthropic, 2026a)	55.2	**63.4**	51.2	61.2	61.8	59.7	48.5	57.7	57.3
Claude Sonnet 4.6 (Anthropic, 2026b)	52.9	59.4	55.3	48.5	63.2	**62.1**	65.3	35.8	55.3
Composer 1.5 (Cursor, 2026)	53.8	51.8	48.0	43.9	62.4	52.1	54.3	**62.9**	53.6
GPT-5.2-Codex (OpenAI, 2025a)	51.0	53.1	51.5	52.9	64.4	53.5	46.2	41.5	51.8
GPT-5.2 (OpenAI, 2025b)	50.4	56.7	48.3	50.6	53.3	46.8	55.2	41.7	50.4
Kimi K2.5 (Kimi Team et al., 2026)	55.6	55.8	40.4	32.5	57.0	56.0	48.3	37.8	47.9
Claude Haiku 4.5 (Anthropic, 2025)	49.9	46.1	41.9	40.7	50.2	37.1	42.8	53.8	45.3
Grok 4.20 (xAI, 2026)	55.4	46.8	42.5	29.6	48.8	35.4	50.1	39.3	43.5

Table 6: Cross-modality results measured by the simple average of the eleven metrics. Bold numbers indicate the best model in each group.

Model	Avg. Rank	Std.	Stability	CT	Microscopy / Path.	MRI	Ultrasound	Ophthalmic	Endoscopy	X-ray	Mammo.
Gemini 3 Flash (Google DeepMind, 2025)	**1.5**	**0.76**	Stable	1	2	1	1	1	2	1	3
Claude Opus 4.6 (Anthropic, 2026a)	3.4	1.69	Moderate	4	1	4	2	5	3	6	2
Claude Sonnet 4.6 (Anthropic, 2026b)	3.9	2.64	Unstable	6	3	2	5	3	1	2	9
Composer 1.5 (Cursor, 2026)	4.9	1.89	Moderate	5	7	6	6	4	6	4	1
GPT-5.2-Codex (OpenAI, 2025a)	5.0	2.14	Moderate	7	6	3	3	2	5	8	6
GPT-5.2 (OpenAI, 2025b)	5.4	1.77	Moderate	8	4	5	4	7	7	3	5
Kimi K2.5 (Kimi Team et al., 2026)	6.1	2.36	Moderate	2	5	9	8	6	4	7	8
Grok 4.20 (xAI, 2026)	7.1	2.17	Moderate	3	8	7	9	9	9	5	7
Claude Haiku 4.5 (Anthropic, 2025)	7.8	1.67	Moderate	9	9	8	7	8	8	9	4

Table 7: Cross-modality rank stability. Lower average rank and lower standard deviation indicate more stable performance across modalities.

Model	Seg. (%)	Cls. (%)	Det. (%)	Mean (%)
Gemini 3 Flash (Google DeepMind, 2025)	**62.6**	**63.0**	**74.3**	**66.6**
Claude Sonnet 4.6 (Anthropic, 2026b)	54.8	56.6	59.9	57.1
Claude Opus 4.6 (Anthropic, 2026a)	56.7	57.7	54.9	56.4
Composer 1.5 (Cursor, 2026)	49.5	55.4	56.0	53.6
GPT-5.2 (OpenAI, 2025b)	50.5	51.8	54.3	52.2
GPT-5.2-Codex (OpenAI, 2025a)	54.1	49.8	51.2	51.7
Kimi K2.5 (Kimi Team et al., 2026)	54.7	47.1	49.4	50.4
Grok 4.20 (xAI, 2026)	49.0	44.7	54.3	49.3
Claude Haiku 4.5 (Anthropic, 2025)	49.3	41.2	53.3	47.9

Table 8: Cross-task results measured by the simple average of the eleven metrics.

Model	Avg. Rank	Std.	Stability	Seg.	Cls.	Det.
Gemini 3 Flash (Google DeepMind, 2025)	1.0	**0.00**	Stable	1	1	1
Claude Opus 4.6 (Anthropic, 2026a)	2.7	1.15	Stable	2	2	4
Claude Sonnet 4.6 (Anthropic, 2026b)	2.7	0.58	Stable	3	3	2
Composer 1.5 (Cursor, 2026)	4.7	2.08	Moderate	7	4	3
GPT-5.2 (OpenAI, 2025b)	5.7	0.58	Stable	6	5	6
GPT-5.2-Codex (OpenAI, 2025a)	6.3	1.53	Moderate	5	6	8
Kimi K2.5 (Kimi Team et al., 2026)	6.7	2.52	Unstable	4	7	9
Grok 4.20 (xAI, 2026)	7.3	2.08	Moderate	9	8	5
Claude Haiku 4.5 (Anthropic, 2025)	7.7	1.53	Moderate	8	9	7

Table 9: Cross-task rank stability. Gemini 3 Flash is the only model ranked first across all three task types.

where m_{sv}, m_{sc}, and m_{cf} denote metadata schema validity, metadata semantic correctness, and metadata content fidelity, respectively. The geometric mean penalizes metadata records with any weak component, and multiplication by CF further requires consistency between dataset-level metadata and sample-level standardized content.

C.7 Joint Metrics

Source Content Joint Score (SCJ). SCJ evaluates whether source selection and content preservation succeed simultaneously:

$$\mathrm{SCJ} = \frac{1}{N}\sum_{i=1}^{N} s_i f_i. \tag{16}$$

Unlike CF, which is computed only after source matching, SCJ is averaged over all target samples. Thus, a wrong-source output receives no credit even if its generated content appears plausible.

End-to-End Strict Pass (E2E). E2E measures whether a sample passes the sample-level standardization pipeline. We first define the strict pass indicator:

$$\begin{aligned} e_i = \mathbb{I}\big[&p_i \wedge s_i = 1 \wedge g_i \geq 0.85 \\ &\wedge q_i \geq 0.5 \wedge f_i \geq 0.5\big]. \end{aligned} \tag{17}$$

The E2E score is then

$$\mathrm{E2E} = \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{18}$$

A sample passes E2E only when it has a valid image-text pair, selects the exact source, reaches the schema-validity threshold, and satisfies the minimum semantic and content-fidelity thresholds. We use $g_i \geq 0.85$ rather than perfect schema validity to avoid collapsing otherwise valid samples because of a minor schema defect.

C.8 Dataset-Level Aggregation

All eleven metrics are first computed independently for each dataset. We then average each metric

across datasets with equal weight:

$$\mathrm{Score}(k) = \frac{1}{|\mathcal{D}|} \sum_{D \in \mathcal{D}} \mathrm{Score}_D(k), \qquad (19)$$

where $\mathcal{D}$ is the set of 100 benchmark datasets and k denotes one of the eleven metrics. This prevents large datasets from dominating the final results and makes the reported scores reflect performance across heterogeneous medical domains, modalities, source formats, annotation styles, and directory organizations.

C.9 Metric Groups Used in Analysis

For analysis, we organize the eleven metrics into five capability groups:

- **Structure**: SV and SSC.
- **Semantic**: SC.
- **Content**: IC, IV, INR, and CF.
- **Metadata**: MSC and MSJ.
- **Joint**: SCJ and E2E.

These groups are used in the main result table and in error attribution analysis. They separate outputs that are merely well-formed from outputs that are source-grounded, semantically correct, content-faithful, and usable as standardized medical data.