MedRepBench: Benchmarking Structured Understanding of Medical Report Images

Fangxin Shang[1], Yuan Xia[1✉], Dalu Yang[1], Yahui Wang[1], and Binglin Yang[1]

Baidu Inc., China

Abstract. Medical report understanding from real-world document images is essential for generating patient-facing explanations and enabling structured information exchange in clinical systems.
Existing VLMs and LLMs have shown strong performance on document understanding, but structured understanding of medical reports remains insufficiently benchmarked. Therefore, we introduce MedRepBench, a benchmark with 1,925 de-identified Chinese medical report images spanning diverse departments, patient demographics, and acquisition formats. In MedRepBench, we mainly focus on report-grounded interpretation rather than evaluating diagnostic reasoning, treatment recommendation, or the integration of patient history. The *interpretation* is defined as structured extraction of report fields (e.g., *item*, *value*, *unit*, *reference range*, *abnormal flag*) plus a patient-facing explanation grounded strictly in the report content.
The benchmark primarily evaluates end-to-end VLMs, and also includes a controlled text-only setting (high-quality OCR + LLM) to approximate an upper bound when character recognition errors are minimized. Our evaluation framework provides two complementary protocols: (1) an objective protocol measuring field-level recall of structured items, and (2) an automated subjective protocol that uses an LLM-based judge to score factuality, interpretability, and reasoning quality under a fixed prompt.
Using the objective metric as a reward signal, we also provide a lightweight GRPO-based alignment baseline for a mid-sized VLM, which improves field-level recall by up to 6%.
Finally, we analyze practical limitations of OCR+LLM pipelines, including layout-related errors and additional system latency, showing the need for robust end-to-end vision-based medical report understanding. The dataset and evaluation resources are publicly available on HuggingFace.

1 Introduction

Medical report interpretation is an essential capability in modern healthcare, serving both patients and healthcare professionals. It is also a key component of AI-powered medical consultation system [3, 23, 24]. For patients, it transforms complex clinical documents into user-friendly explanations that support informed decision-making and promote healthier lifestyles outside clinical settings. For physicians and healthcare institutions, it enables efficient electronic

✉ Corresponding author. Email: babyxiayuan@gmail.com

information exchange across disparate systems, improving referral workflows and multi-institutional collaboration. As healthcare systems increasingly adopt electronic documentation, the demand for automated and scalable medical report interpretation solutions continues to grow.

Fig. 1: Overview of the MedRepBench dataset. (a) Joint distribution of document types (Exam./Lab.) and acquisition methods (Photo, Screenshot, e-Doc.); (b) Report counts for the top-5 clinical departments; (c) Patient gender distribution (with *Unspecified* denoting missing metadata); (d) Patient age distribution across the corpus.

Unlike medical report generation [9], which focuses on producing narrative-style descriptions, medical report interpretation aims to extract structured clinical findings [19] and generate patient-facing explanations grounded in factual content. This task sits at the intersection of information extraction and layperson-targeted summarization, supporting applications such as at-home health monitoring, physician-patient communication, and intelligent referral systems.

Although both vision–language models (VLMs) and large language models (LLMs) demonstrate strong capabilities in document understanding, prior work [16] has not adequately evaluated their performance on structured interpretation of medical reports. We therefore focus on assessing the end-to-end capability of VLMs to extract and explain structured clinical content directly from raw report images, and systematically compare them with OCR-assisted pipelines (OCR+VLM and OCR+LLM). From an application standpoint, OCR-based pipelines remain the most accurate and cost-effective option whenever high-quality text can be reliably extracted from medical reports. However, they introduce an additional subsystem that must be engineered and maintained for each domain and document template, and they strip away layout and visual cues from the downstream LLM, leading to architectural overhead and error propagation.

In parallel, emerging vision language models with text-token-free visual encoders [5,22], which encode written content directly as visual tokens, challenge the traditional *OCR-then-LLM* decomposition. These models call for benchmarks that stress end-to-end visual understanding and structured prediction

directly from images, beyond purely text-based extraction. MedRepBench is therefore designed to compare OCR-assisted and no-OCR paradigms under a unified protocol and to serve as a concrete reference point and diagnostic signal for future text-token-free VLMs. Our benchmark reveals that VLMs, although still lagging behind strong text-only models under OCR-assisted settings, remain a promising direction for unified, layout-aware medical document interpretation.

The medical report interpretation task presents unique complexities compared to general document understanding: (1) **diverse data acquisition methods**, such as handheld device photography, PDF documents, and mobile web screenshots; (2) **high variability in image quality**, including occlusions, folds, partial captures, and illumination issues; and (3) **heterogeneous layout styles**, as medical reports lack universally enforced formatting standards. These factors collectively demand robust model performance across diverse and noisy input scenarios.

To address these challenges, we introduce **MedRepBench**, a benchmark designed to evaluate modern models in medical report interpretation. MedRepBench consists of 1,925 de-identified real-world medical reports spanning multiple clinical departments and patient demographics. Each report includes metadata such as document type (e.g., *Exam. Report*, *Lab Report*), department (e.g., *Gynecology*, *Pediatrics*), patient gender, and age.

As shown in Figure 1, our proposed MedRepBench includes two primary document types—exam reports and lab reports—captured through diverse acquisition methods such as photos, screenshots, and electronic documents. These reports originate from a wide range of medical institutions, covering substantial cross-institution layout variations, thus reflecting real-world heterogeneity in medical documentation. Such heterogeneity, combined with broad departmental coverage and diverse patient demographics, is crucial for evaluating the robustness and generalizability of medical report interpretation models.

Our main contributions are as follows:

- We introduce **MedRepBench**, a benchmark for *vision-grounded structured understanding* of real-world Chinese medical report images, featuring 1,925 de-identified laboratory and examination reports. It supports a dual evaluation protocol: (1) *objective* field-level recall over five attributes, and (2) *automated subjective* evaluation of factuality, interpretability, and reasoning quality via an LLM judge.
- We benchmark representative open-source VLMs and LLMs under both end-to-end (image-only) and OCR-assisted settings, revealing current strengths, failure modes, and the remaining gap in the no-OCR configuration.
- We provide a lightweight GRPO-based alignment baseline that uses field-level recall as a reward, showing up to **+6%** absolute recall improvement over SFT for a mid-scale VLM.

The dataset, benchmark protocol, and evaluation resources are released on HuggingFace, including de-identified report images, structured metadata/labels, documentation, prompt templates, and evaluation scripts for reproducible evaluation. The dataset is released under the CC BY-NC 4.0 license.

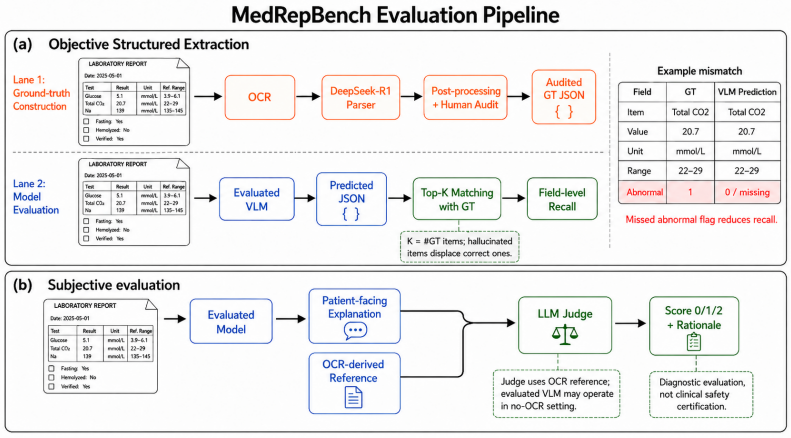

Fig. 2: Evaluation pipeline and qualitative comparison in MedRepBench. (a) Objective evaluation separates OCR-assisted ground-truth construction from VLM prediction and Top-K field-level recall; the inset shows a VLM–GT abnormal-flag mismatch. (b) Subjective evaluation scores patient-facing explanations with an LLM judge using OCR-derived reference content.

2 The Proposed Benchmark

2.1 Dataset Construction

Figure 2 provides an overview of the evaluation pipeline, covering objective structured extraction and subjective scoring of patient-facing explanations. We collect medical report images from a private, real-world online medical service source spanning multiple clinical departments and document types, including handheld photographs, PDFs, and mobile screenshots. The reports were uploaded in practical service scenarios, and their secondary use for benchmark construction is governed by the service's user agreement and institutional compliance requirements. We randomly sample 8,000 reports and apply a multi-stage curation pipeline covering de-identification, quality filtering, deduplication, and content validation, resulting in 1,925 high-quality images.

To obtain structured labels, we first extract text with a proprietary OCR system optimized for medical documents. The OCR text is then parsed into structured JSON by DeepSeek-R1 [8] under a constrained schema, producing item-level fields such as *name*, *value*, *unit*, *reference range*, and *abnormality flag*, as well as document-level metadata including *document type*, *department*, *gender*, and *age*.

Filtering and Quality Control. We remove low-quality or incomplete captures using heuristic quality filters (e.g., short-side resolution below 800 px, severe motion blur, and truncated pages). We then manually inspect a random subset of 500 samples to verify the filtering results. We then deduplicate reports by combining perceptual image hashing with OCR text similarity, discarding

12.3% near-duplicates. Next, we perform content validation to ensure each retained report contains at least one structured laboratory panel with valid reference ranges. To control bootstrapping noise, we enforce schema validation and rule-based normalization (e.g., unit standardization, numeric formatting, and reference-range pattern checks), and further audit a random subset of 200 reports with human annotators. We observe over 97% field-level agreement; the main residual errors involve nested-table field misalignment and OCR errors on handwritten annotations, and all errors found during the audit are corrected before release.

De-identification and Release. Before release, we apply OCR-based personally identifiable information (PII) detection for names, addresses, ID numbers, and other direct identifiers, followed by region-of-interest pixel blackout. All 1,925 images are then manually inspected to ensure that no individual can be traced from the released data. The public HuggingFace repository contains the de-identified images under a censored image directory and a metadata/label CSV file with document-level attributes and structured item annotations.

Finally, we emphasize that our main benchmark conclusions rely on *field-level recall* against the released structured labels, rather than any LLM-judge component used elsewhere in the evaluation framework.

2.2 Structured Interpretation Evaluation

Previous benchmarks for report evaluation—such as BLEU, ROUGE, or CIDEr—focus on surface-level text similarity, which is insufficient for assessing structured information extraction. MedRepBench instead introduces a field-level evaluation protocol based on recall, defined for each field f as:

$$\text{Recall}_f = \frac{\#\text{ correctly extracted field } f}{\#\text{ ground-truth field } f}. \tag{1}$$

We evaluate recall over five core fields as mentioned above. The primary objective metric is the average of field-wise recall across all matched items:

$$\text{AverageRecall} = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \text{Recall}_f, \tag{2}$$

where $\mathcal{F}$ denotes the set of target fields and $|\mathcal{F}| = 5$ in MedRepBench.

To avoid rewarding unconstrained over-generation, we truncate each model output to Top-K items, where K is the number of ground-truth items in the corresponding report. Under this protocol, hallucinated or irrelevant items compete for the same Top-K slots and can displace correct items, directly reducing field-level recall. Thus, although the primary metric is recall, the fixed-budget matching protocol implicitly penalizes spurious generations.

2.3 Automated Subjective evaluation via LLM

In addition to field-level recall, we use an LLM-based evaluator to assess the interpretability and factuality of model outputs. The evaluator is prompted with

the model-generated explanation and OCR-derived reference content, returning a scalar score (e.g., 0–2) and a brief rationale. This subjective score complements the objective metric by evaluating whether the output forms a report-grounded and internally consistent patient-facing explanation.

In our setting, *factuality* measures whether the generated explanation is strictly supported by evidence in the report, i.e., it is consistent with the extracted structured fields and does not introduce non-existent test items, values, units, or reference ranges. *Interpretability* reflects whether the explanation is understandable to lay users by organizing results clearly (e.g., highlighting key abnormal items), using accessible language, and avoiding unsupported diagnoses or treatment recommendations. *Reasoning quality* assesses whether abnormality and key takeaways are derived from value–range comparisons or discrete outcomes (e.g., positive/negative), without internal contradictions.

While DeepSeek-R1 is used for both initial annotation and as the LLM judge, we decouple these roles. Model performance conclusions are based on objective field-level recall, independent of DeepSeek-R1's judgment. DeepSeek-R1 annotations are post-processed and verified by human reviewers to minimize systematic errors. The LLM-as-judge scores are primarily for ranking and qualitative comparison, further validated through a pilot human study in Section 4.2.

2.4 Reinforcement Learning Optimization

To improve structured item extraction, we adopt **Group Relative Policy Optimization (GRPO)** [13, 20, 21] to align a vision–language model with our benchmark objective, using field-level recall as a non-differentiable reward signal. This component is included as a practical instantiation of recent RL-based alignment techniques in the context of medical report understanding, and we do *not* claim algorithmic novelty.

Reward Design and Optimization. We treat report interpretation as sequence generation: given an image or OCR-assisted prompt, the model outputs a JSON list of five-field items (*name*, *value*, *unit*, *range*, *abnormality flag*). The GRPO reward is the average field-level recall over target fields $\mathcal{F}$, where $|\mathcal{F}| = 5$:

$$R = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \text{Recall}_f. \tag{3}$$

Recall is computed between one-to-one matched prediction–reference pairs aligned by normalized item names. Following GRPO, candidate outputs in each group are scored by R, transformed into group-relative advantages, and optimized with a PPO-style clipped objective with KL regularization to the SFT reference policy.

Implementation Details. We apply GRPO on top of an InternVL3-8B model [30] that has been previously fine-tuned via supervised learning (SFT) on a separate dataset for the subjective interpretation task. During GRPO training, the vision

encoder and image-to-text projector are frozen to preserve stable multimodal alignment. We insert LoRA [11] modules (rank=128) into all attention layers of the LLM (query, key, value projections), and only LoRA weights are updated. Training is performed for 2 epochs on a curated dataset of 800 structured report samples following the MedRepBench format.

This lightweight optimization yields a +6% gain in average field-level recall over the SFT baseline, along with improved subjective interpretability (see Section 4.3 for details).

3 Evaluation

3.1 Models and Experimental Setup

To ensure reproducibility and privacy compliance in clinical applications, we evaluate exclusively on publicly available vision-language models (VLMs) and large language models (LLMs). Proprietary frontier APIs (e.g., GPT, Gemini, and Claude) are excluded.

VLMs: We include a broad selection of recent vision-language models capable of document-level understanding, including InternVL2.5 [4], InternVL3 [30], Qwen2.5-VL [2], and LLaMA-4 [18]. These models are evaluated using their image input interface under consistent inference settings.

We intentionally focus on general-domain VLMs to evaluate their transferability from generic document and vision-language understanding tasks to structured medical report extraction. Extending the benchmark to medical-specialized VLMs is a natural direction for future leaderboard development.

LLMs: We additionally benchmark several text-only LLMs, including Qwen3 series [25] and the DeepSeek-R1-Distill [8]. For these models, OCR-extracted text is provided as part of the input prompt. DeepSeek-R1 is also employed as an automatic judge for subjective evaluation.

For objective evaluation, we consider two settings: (1) end-to-end VLM interpretation from raw images, and (2) an auxiliary OCR-assisted setup where recognized text is provided to both VLMs and LLMs. In subjective evaluation, only the evaluator (DeepSeek-R1) receives OCR input, while models generate interpretations from their native inputs (images for VLMs, OCR for LLMs). This design ensures interpretability assessment is decoupled from access to ground-truth text, while remaining anchored to a consistent reference.

Although MedRepBench supports both visual and text-based pipelines, its primary goal is to evaluate the end-to-end capabilities of VLMs. OCR-assisted variants serve only to approximate upper bounds and diagnose non-OCR-related limitations.

3.2 Objective and Subjective Evaluation Protocols

We design two complementary evaluation protocols: objective evaluation, focusing on structured field extraction accuracy, and subjective evaluation, focusing on report-grounded factuality, readability, and reasoning quality.

Objective Evaluation. This protocol is applied exclusively to laboratory-style reports (e.g., blood tests), where each report contains a list of structured test items represented as five-field tuples. Each item contains the test name, measured value, reference range, measurement unit, and an abnormality flag. Models are required to output these as a JSON array, as illustrated in Figure 2. We do not apply this objective protocol to examination reports because they are predominantly narrative and often lack a stable itemized schema; forcing field-level labels for such reports would introduce unreliable pseudo-labels. Instead, examination reports are covered by the subjective protocol, which evaluates report-grounded factuality, reasoning, and patient-facing clarity.

For each item, we define five target fields: name, value, unit, reference_range, and is_abnormal. The first four are directly extractable from the report, while is_abnormal is a derived label that requires comparing the measured value with the reference range and handling discrete labels such as 'positive/negative' or semi-quantitative levels. We expect models to infer is_abnormal rather than copying it from the text, and we treat it as a reasoning-oriented sub-task.

Performance is evaluated by comparing the model-generated JSON with the ground-truth JSON and measuring field-level recall. An item is matched to a reference only if the *name* field matches exactly (after normalization). Once matched, each ground-truth item is considered once. Recall is then computed for the remaining fields independently: *value*, *unit*, *range*, and *abnormal*. To account for minor formatting variation (e.g., 2–4 vs. 2∼4), fuzzy matching is permitted for the *range* field, while other fields require exact equivalence (either numerical or string).

Preprocessing includes whitespace trimming, irrelevant token removal, and fault-tolerant JSON parsing to ensure consistent evaluation. We report both per-field recall and averaged recall across all five attributes.

Subjective Evaluation. In this setting, models are asked to generate human-readable, report-grounded explanations based on the input report. For consistency, all models are evaluated under the no-OCR configuration, where the recognized text is not provided to the model.

We employ DeepSeek-R1 as the evaluator model, prompted to assess three key dimensions: factual accuracy, reasoning validity, and patient-facing clarity. Each interpretation is scored on a discrete 3-point scale:

- **0**: Severe factual errors or unsupported clinical claims, such as inventing absent findings, reversing abnormality status, or giving diagnosis/treatment advice not supported by the report.
- **1**: Minor inaccuracies or stylistic issues, without major unsupported clinical advice.
- **2**: Accurate, well-structured, and clearly grounded in the report.

We compute two metrics: **acceptability rate** (proportion of samples with score ≥ 1), and **excellence rate** (proportion with score 2). To ensure robustness, all evaluations use deterministic decoding (temperature=0). A subset of responses is re-evaluated across multiple seeds to confirm evaluation stability.

Table 1: Structured field-level recall (%) of VLMs on laboratory-style reports. no-OCR denotes end-to-end image-only inference; OCR-asst. denotes OCR-assisted inference where recognized text is provided as additional input.

Model	$R_{name} \uparrow$		$R_{value} \uparrow$		$R_{unit} \uparrow$		$R_{range} \uparrow$		$R_{abnormal} \uparrow$	
	no-OCR	OCR-asst.	no-OCR	OCR-asst.	no-OCR	OCR-asst.	no-OCR	OCR-asst.	no-OCR	OCR-asst.
InternVL2.5-8B	71.38	85.64	55.71	70.14	60.63	77.85	58.88	79.40	45.06	58.59
InternVL2.5-38B	84.83	95.20	68.85	85.02	74.93	89.00	72.37	89.44	66.01	81.52
InternVL3-8B	88.21	95.25	71.29	84.87	75.56	88.00	70.63	88.80	67.92	77.97
InternVL3-14B	88.70	95.25	67.69	86.53	74.63	89.52	67.96	88.81	70.12	83.56
InternVL3-38B	88.72	94.69	64.83	85.64	75.63	88.79	69.36	89.88	66.74	77.55
Qwen2.5-VL-7B	89.90	93.94	73.17	82.46	77.31	81.61	78.33	85.07	68.36	76.39
Qwen2.5-VL-32B	90.13	95.89	76.32	84.72	79.30	87.71	81.61	89.15	59.66	83.95
LLaMA-4-Scout	87.90	92.68	65.69	78.86	74.29	82.94	67.87	77.91	63.24	69.23
LLaMA-4-Maverick	88.51	96.68	72.26	87.35	78.83	90.23	77.58	90.74	75.83	86.56

4 Results

4.1 Objective Evaluation (Structured Field Recall)

We evaluate structured extraction on laboratory-style reports using field-level recall over five attributes: *name*, *value*, *unit*, *reference range*, and *abnormality*. Predicted entries are matched to ground truth via normalized *name* alignment, after which recall is computed independently for the remaining attributes. Results are averaged over three decoding runs with different random seeds; standard deviations are below 5%, indicating stable evaluation.

Table 1 reports VLM performance in our primary *no-OCR* setting (end-to-end, vision-grounded extraction) and the corresponding *OCR-assisted* setting for reference. Several models, notably the InternVL3 family and Qwen2.5-VL-32B, achieve meaningful recall even without explicit text input. Providing OCR text consistently improves recall across models and fields—especially for *value* and *reference range*—and serves as an approximate upper bound when visual perception errors are minimized. Nevertheless, as discussed later, OCR-driven pipelines remain vulnerable to cascading errors and are less grounded in layout, motivating further progress toward robust end-to-end VLMs.

Although the OCR+LLM pipeline performs well on both objective and subjective metrics, its performance remains highly dependent on OCR quality. Recognition errors, such as missing entries, structural misalignment, and incorrect segmentation, can propagate through the pipeline and affect downstream reasoning, resulting in factual hallucinations such as incorrectly identifying normal findings as abnormal. Such cascading errors are hard to correct due to the non-end-to-end nature of the pipeline. Furthermore, by discarding visual layout cues, OCR+LLM models lack spatial grounding, limiting their robustness and clinical reliability in real-world settings.

Figure 3 visualizes average recall across all five fields with respect to model size. In both no-OCR and OCR-assisted settings, performance generally scales with model capacity. However, several models deviate from this trend, suggesting that alignment strategies and training data play crucial roles. Notably, our RL-optimized model (Ours-GRPO) outperforms much larger baselines under no-OCR conditions, highlighting the potential of benchmark-driven optimization.

Table 2: Field-level recall (%) for LLMs under OCR-assisted input.

Model	$R_{name} \uparrow$	$R_{value} \uparrow$	$R_{unit} \uparrow$	$R_{range} \uparrow$	$R_{abnormal} \uparrow$
Qwen3-1.7B	92.88	70.50	79.42	80.91	69.04
Qwen3-4B	94.64	74.59	86.68	87.40	64.29
Qwen3-8B	94.93	79.48	88.36	88.26	68.50
Qwen3-14B	95.75	87.43	89.69	90.18	85.43
Qwen3-32B	95.66	87.46	90.92	90.68	85.77
Qwen3-30B-A3B	96.08	86.48	88.88	90.08	82.67
Qwen3-235B-A22B	94.96	87.27	89.09	90.56	84.54
Qwen3-235B-A22B-2507	96.42	89.40	91.14	92.48	88.03
DeepSeek-R1-Qwen-7B	91.68	70.60	75.83	75.13	67.13
DeepSeek-R1-Qwen-14B	95.50	84.61	88.07	88.38	83.01
DeepSeek-R1-Qwen-32B	92.93	83.40	86.36	87.53	82.29
DeepSeek-V3	96.48	88.83	90.75	92.60	89.06

Table 3: LLM-based subjective evaluation scores under no-OCR input.

Model	Accept. ↑	Excel. ↑
InternVL2.5-8B	33.67	17.30
InternVL2.5-38B	57.97	42.45
InternVL3-8B	48.24	31.83
InternVL3-14B	47.00	28.55
InternVL3-38B	51.15	32.70
Qwen2.5-VL-7B	26.21	5.58
Qwen2.5-VL-32B	67.83	51.40
LLaMA-4-Scout	46.84	26.18
LLaMA-4-Maverick	60.78	42.74

These results confirm the difficulty of end-to-end report interpretation, especially in the absence of OCR. On average, field-level recall drops by 10–20% when OCR input is removed. This gap reveals substantial headroom for improving visual-text alignment and structured reasoning in VLMs, further motivating the need for targeted benchmarks such as MedRepBench.

No-OCR Failure Modes. Manual inspection shows three recurring error patterns in end-to-end VLM outputs: (1) character-level mistakes on fine-print numeric values and units, (2) layout misalignment in nested tables or multi-column reports, and (3) missed abnormality flags when visual cues are subtle or separated from the corresponding value. These failures explain why OCR-assisted pipelines remain strong while also motivating models that can jointly preserve text, layout, and visual grounding.

4.2 Subjective Evaluation (Interpretability Score)

We additionally conduct a pilot human study to validate the reliability of the LLM-based subjective scores. Beyond factual field extraction, we evaluate models

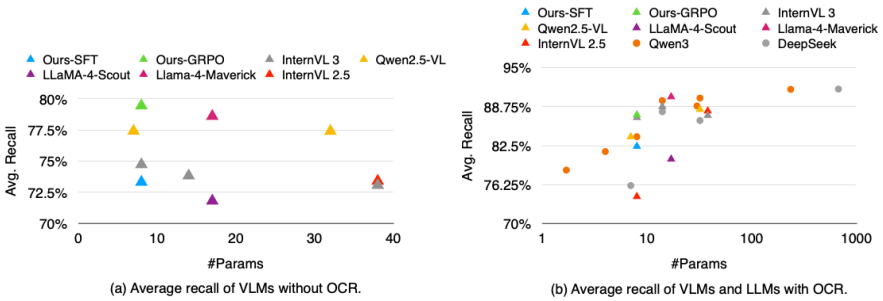

Fig. 3: Average field-level recall (%) vs. model size. (a) no-OCR input; (b) OCR-assisted input. VLMs (△) and LLMs (○) are distinguished by markers. Ours-GRPO substantially improves 8B VLM performance, and is competitive with larger baselines in the no-OCR setting. Y-axis in (a) is limited to 70–80% for clarity; InternVL2.5-8B (58.33%) is excluded.

from the perspective of clinical usability: each model-generated explanation is scored by DeepSeek-R1 along three dimensions—factuality, reasoning quality, and patient-facing clarity. The scores are mapped to a three-level scale, and we report two summary metrics: **Acceptability Rate** (scores ≥ 1) and **Excellence Rate** (scores $= 2$).

Table 3 presents the subjective evaluation results in the no-OCR setting. Larger VLMs such as Qwen2.5-VL-32B and LLaMA-4-Maverick generally achieve stronger acceptability and excellence rates, while smaller or insufficiently aligned models often generate vague or factually inaccurate interpretations. These results highlight the need for visual grounding and domain-aware alignment in patient-facing medical report explanations.

Agreement Between LLM-Based Evaluator and Human Experts. To assess the reliability of our LLM-based evaluator, we conduct a consistency study against expert human judgments. We randomly sample 20 cases from each of the three rating categories (score 0, 1, and 2) from the 1,925 evaluation set, resulting in a total of 60 samples. Each case is independently rated by three experienced annotators (licensed medical reviewers or clinically trained annotators). Final human scores are determined via majority voting. If consensus could not be reached, the case was discarded and replaced with another randomly sampled instance from the same score category.

Figure 4 shows the confusion matrix between the LLM-based evaluator and the aggregated human scores. Overall agreement is high, with an accuracy of **88.3%** and Cohen's κ of **0.82**, suggesting substantial consistency under our evaluation rubric. Most disagreements occur between adjacent ratings (e.g., 1 vs. 2), indicating borderline cases rather than clear-cut factual contradictions. We emphasize that this pilot study serves as a sanity check rather than a substitute for expert clinical assessment; the LLM judge is used as a *diagnostic* and scalable signal for ranking and qualitative comparison, not as an independent clinical assessment.

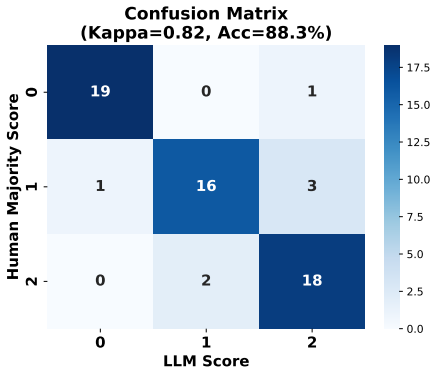

Fig. 4: Confusion matrix between LLM-based evaluator scores and human expert majority vote. High agreement is observed (Cohen's $\kappa = 0.82$, Accuracy = 88.3%).

4.3 A Practical Alignment Baseline (GRPO)

To reduce the performance gap in the *no-OCR* setting, where vision-grounded structured extraction remains challenging even for strong VLMs, we further optimize a supervised fine-tuned InternVL3-8B model using GRPO. In our GRPO experiments, the policy input follows the no-OCR setting (images only), and the reward is computed against the released JSON labels. MedRepBench provides a rigorous objective protocol and an explicit field-level recall signal, which naturally serves as a reward for this alignment.

We present GRPO here as a lightweight and reproducible baseline to demonstrate that MedRepBench induces a meaningful reward landscape for improving structured extraction; we do *not* claim novelty of the underlying RL algorithm.

Concretely, we first perform 2 epochs of supervised fine-tuning (SFT) on InternVL3-8B using ∼15k human-annotated instruction–response pairs derived from real-world medical report interpretations, yielding **Ours-SFT**. For reinforcement learning, we additionally curate 800 laboratory-report samples formatted according to our benchmark schema and run 2 epochs of GRPO with LoRA (rank 128), producing **Ours-GRPO**.

Table 4: Field-level recall (%) for Ours-GRPO and baselines.

Model	OCR	#Param	Avg. Recall ↑
LLaMA-4-Maverick	No	17B	78.60
Qwen2.5-VL-32B	No	32B	77.41
Ours-SFT	No	8B	73.31
Ours-GRPO	No	8B	**79.45**
Qwen2.5-VL-7B	Yes	7B	83.89
InternVL2.5-8B	Yes	8B	74.32
Qwen3-8B	Yes	8B	83.90
InternVL3-8B	Yes	8B	86.98
Ours-SFT	Yes	8B	82.37
Ours-GRPO	Yes	8B	**87.41**

Table 5: Subjective scores for Ours-GRPO and baselines.

Model	Param	Accept. ↑	Excel. ↑
Ours-SFT	8B	56.27	38.86
Ours-GRPO	8B	60.64	42.45
LLaMA-4-Maverick	17B	60.78	42.74
Qwen2.5-VL-32B	32B	**67.83**	**51.40**
InternVL2.5-38B	38B	57.97	42.45
InternVL3-38B	38B	51.15	32.70

Table 4 and Table 5 compare **Ours-GRPO** against baselines in both objective and subjective settings, with Figure 3 illustrating recall improvements under no-/with-OCR conditions. In the no-OCR setting, **Ours-GRPO** achieves the highest average recall (79.45%), outperforming LLaMA-4-Maverick (78.60%) and Qwen2.5-VL-32B (77.41%), and surpassing its supervised baseline (73.31%) by +6%.

Although **Ours-GRPO** is not the top-ranked model in subjective interpretability, its excellence rate of 42.45% is on par with the much larger LLaMA-4-Maverick (42.74%) and exceeds that of InternVL3-38B (32.70%). These results confirm that MedRepBench offers a meaningful reward landscape for policy optimization and that targeted reinforcement learning can further enhance structured interpretation, even beyond supervised baselines.

5 Related Work

5.1 Medical Report Interpretation vs. Generation

Most prior work in medical AI has focused on report generation or VQA using vision–language models (VLMs). For example, Argus [15] and MedVAG [1] tackle radiology report generation, while reviews [1, 9] highlight that structured field extraction—crucial for report *interpretation*—is underexplored. In contrast, MedRepBench specifically targets the interpretive task: extracting structured findings and evaluating readability rather than generating narrative summaries.

5.2 Document Understanding and VQA Benchmarks

Recent benchmarks in document understanding span both general and medical domains. General-purpose benchmarks such as the MMDocBench [29] and BenchX [28] focus on layout analysis, structure parsing, and multi-task performance across business or legal documents. In the medical domain, visual question answering (VQA) benchmarks such as OmniMedVQA [12], GEMeX [14], MedFrameQA [26], and BESTMVQA [10] evaluate multimodal reasoning over radiology and pathology images. However, these datasets typically assume clean OCR or use synthetic templates, and they do not address the structured interpretation of noisy, real-world medical report images. Docopilot [6] recently explored end-to-end document reasoning with layout-aware VLMs, but was not

tailored to clinical interpretation or evaluation. In contrast, MedRepBench emphasizes OCR robustness, layout variance, and medically grounded extraction, filling a key gap in VLM evaluation.

5.3 LLMs for Structured Extraction and Evaluation

LLMs have shown potential for extracting structured information from clinical texts [7] and for evaluating generation tasks via automatic metrics [17]. MedRepBench uses LLMs in both roles when OCR is available: as information extractors in OCR-assisted pipelines and as subjective evaluators for patient-facing explanations. Unlike generic evaluation frameworks, our subjective evaluation emphasizes medical factuality and report-grounded interpretability, while remaining a diagnostic signal rather than an independent clinical assessment. Moreover, recent analysis [27] highlights the limitations of vision-only models in handling OCR-sensitive inputs, further justifying the dual no-OCR/OCR-assisted evaluation design in MedRepBench.

6 Conclusion

We present **MedRepBench**, a real-world benchmark for evaluating end-to-end, vision-grounded structured understanding of medical report images. MedRepBench provides two complementary protocols: an *objective* field-level recall metric for structured extraction and an *automated subjective* framework that uses an LLM judge to assess report-grounded interpretability and factuality. Benchmarking representative open-source models reveals that robust image-to-structure extraction remains challenging in the *no-OCR* setting, leaving substantial headroom for future VLMs. As a practical demonstration, we show that a lightweight RL-based alignment baseline using MedRepBench's objective reward can further improve a mid-scale VLM.

7 Limitations

We mainly focus on open-source models to promote reproducibility and privacy-preserving evaluation. Therefore, we do not report results from proprietary models such as GPT, Gemini, and Claude. MedRepBench is constructed from Chinese medical reports and does not claim cross-lingual or cross-country generalization, although the schema and evaluation protocols can be adapted to other languages and healthcare systems. Our current baseline suite also emphasizes general-domain VLMs; medical-domain VLMs and additional OCR systems can be benchmarked through the released resources.

References

1. Arısoy, M.V., et al.: A vision attention driven language framework for medical report generation. Scientific Reports (2025)

2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Chen, J., Gui, C., Gao, A., Ji, K., Wang, X., Wan, X., Wang, B.: Cod, towards an interpretable medical agent using chain of diagnosis. In: Annual Meeting of the Association for Computational Linguistics (2024), https://api.semanticscholar.org/CorpusID:271270605
4. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
5. Cheng, J., Liu, Y., Zhang, X., Fei, Y., Hong, W., Lyu, R., Wang, W., Su, Z., Gu, X., Liu, X., Bai, Y., Tang, J., Wang, H., Huang, M.: Glyph: Scaling context windows via visual-text compression. arXiv preprint arXiv:2510.17800 (2025)
6. Duan, Y., Chen, Z., Hu, Y., Wang, W., Ye, S., Shi, B., Lu, L., Hou, Q., Lu, T., Li, H., et al.: Docopilot: Improving multimodal models for document-level understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4026–4037 (2025)
7. Garcia-Carmona, A., et al.: Leveraging large language models for accurate retrieval of patient information from medical reports: Systematic evaluation study. JMIR AI (2025)
8. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
9. Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. Frontiers in artificial intelligence **7**, 1430984 (2024)
10. Hong, X., et al.: BESTMVQA: A benchmark evaluation system for medical visual question answering. arXiv preprint arXiv:2312.07867 (2023)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
12. Hu, Y., et al.: OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. arXiv preprint arXiv:2402.09181 (2024)
13. Lai, Y., et al.: Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. arXiv preprint arXiv:2503.13939 (2025)
14. Liu, B., et al.: GEMeX: A large-scale, groundable, and explainable medical VQA benchmark for chest x-ray diagnosis. arXiv preprint arXiv:2411.16778 (2024)
15. Liu, C., et al.: Argus: Benchmarking and enhancing vision–language models for 3D radiology report generation. arXiv preprint arXiv:2406.07146 (2025)
16. Liu, M., Hu, W., Ding, J., Xu, J., Li, X., Zhu, L., Bai, Z., Shi, X., Wang, B., Song, H., et al.: MedBench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. Big Data Mining and Analytics **7**(4), 1116–1128 (2024)
17. Liu, X., et al.: LLM-as-a-judge: Evaluating NLG with large language models. In: ACL (2023)
18. Meta AI: The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (2025), accessed: 2026-06-22

19. Moon, J.H., Choi, G., Rabaey, P., Kim, M.G., Hong, H.G., Lee, J.O., Yoon, H., Doe, E.W., Kim, J., Sharma, H., et al.: Lunguage: A benchmark for structured and sequential chest x-ray interpretation. arXiv preprint arXiv:2505.21190 (2025)
20. Pan, J., et al.: MedVLM-R1: Incentivizing medical reasoning capability of vision-language models via reinforcement learning. arXiv preprint arXiv:2502.19634 (2025)
21. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
22. Wei, H., Sun, Y., Li, Y.: DeepSeek-OCR: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
23. Xia, Y., Shi, Z., Zhou, J., Xu, J., Lu, C., Yang, Y., Wang, L., Huang, H., Zhang, X., Liu, J.: A speaker-aware co-attention framework for medical dialogue information extraction. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 4777–4786. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.315, https://aclanthology.org/2022.emnlp-main.315/
24. Xia, Y., Zhou, J., Shi, Z., Lu, C., ting Huang, H.: Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In: AAAI Conference on Artificial Intelligence (2020), https://api.semanticscholar.org/CorpusID:213593813
25. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
26. Yu, S., et al.: MedFrameQA: A multi-image medical VQA benchmark for clinical reasoning. arXiv preprint arXiv:2505.16964 (2025)
27. Zhang, P., Xu, H., Zhang, J., Xu, G., Zheng, X., Yang, Z., Liu, J., Zhang, Y., Jin, L.: Aesthetics is cheap, show me the text: An empirical evaluation of state-of-the-art generative models for OCR. arXiv preprint arXiv:2507.15085 (2025)
28. Zhou, Y., et al.: BenchX: A unified benchmark framework for medical vision–language pretraining on chest x-rays. NeurIPS Datasets and Benchmarks Track (2024)
29. Zhu, F., et al.: MMDocBench: Benchmarking large vision–language models for fine-grained visual document understanding. arXiv preprint arXiv:2410.21311 (2024)
30. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)