nature medicine

Article

https://doi.org/10.1038/s41591-026-04655-5

A clinical multimodal vision language foundation model with fine-grain explainability

Received: 29 September 2025

Accepted: 17 August 2026

Published online: 29 October 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

Despite their considerable promise, many current medical foundation models remain modality specific and lack transparent reasoning, limiting clinical adoption. To address this gap, we introduce Explainable Vision Language Foundation Model for Medicine (EVLF-FM), a multimodal vision language foundation model that combines broad diagnostic capabilities with fine-grained explainability. The development and testing of EVLF-FM encompassed more than 1.3 million total samples from 18 global datasets across 10 imaging modalities related to six clinical specialties: dermatology, hepatology, ophthalmology, pathology, pulmonology and radiology. External validation employed 13,443 independent test samples from 17 additional datasets across 10 imaging modalities. EVLF-FM supports disease diagnosis, visual question answering and pixel-level visual grounding with reasoning capabilities. In internal validation for disease diagnostics, EVLF-FM achieved the highest average accuracy, 0.895, outperforming leading generalist and specialist models. In medical visual grounding, EVLF-FM also achieved reasonable performance across nine modalities with average accuracy at 0.5 intersection over union (Acc@0.5) of 0.842. External testing further confirmed consistent zero-shot and few-shot performance, with competitive F1 scores despite a smaller model size. A hybrid training strategy combining supervised and visual reinforcement fine-tuning enables EVLF-FM to generate clinically interpretable explanations aligned with visual evidence. Human evaluation of these reasoning outputs using the S.C.O.R.E. framework yielded a high overall mean score of 4.25/5, supporting their clinical interpretability. Current limitations include the need for additional validation across diverse healthcare settings. Prospective clinical studies are required to evaluate the safety, effectiveness and clinical utility of EVLF-FM before routine clinical deployment.

The emergence of foundation models (FMs) has marked a paradigm shift in artificial intelligence (AI). Previously, AI in medicine generally referred to supervised deep learning models, which were limited to diagnosing a narrow set of diseases from prespecified labels[1]. By contrast, FMs are developed through large-scale unsupervised, semi-supervised or self-supervised pretraining and fine-tuning that confers more general abilities such as simultaneous detection of multiple diseases or accurate interpretation of images without specific training[2–5]. This versatility is compounded by a reduced requirement for labeled data to elicit useful performance relative to conventional deep learning. In addition, conversational AI systems extend FM capabilities by simulating clinician–patient interactions such as through structured history-taking and explanation of diagnosis and management[6–8]. These emerging applications illustrate the growing potential of FMs in medicine to support complex clinical reasoning and multimodal processing in diverse contexts, marking a major step toward general medical AI (GMAI)[9].

e-mail: daniel.ting@duke-nus.edu.sg

https://doi.org/10.1038/s41591-026-04655-5

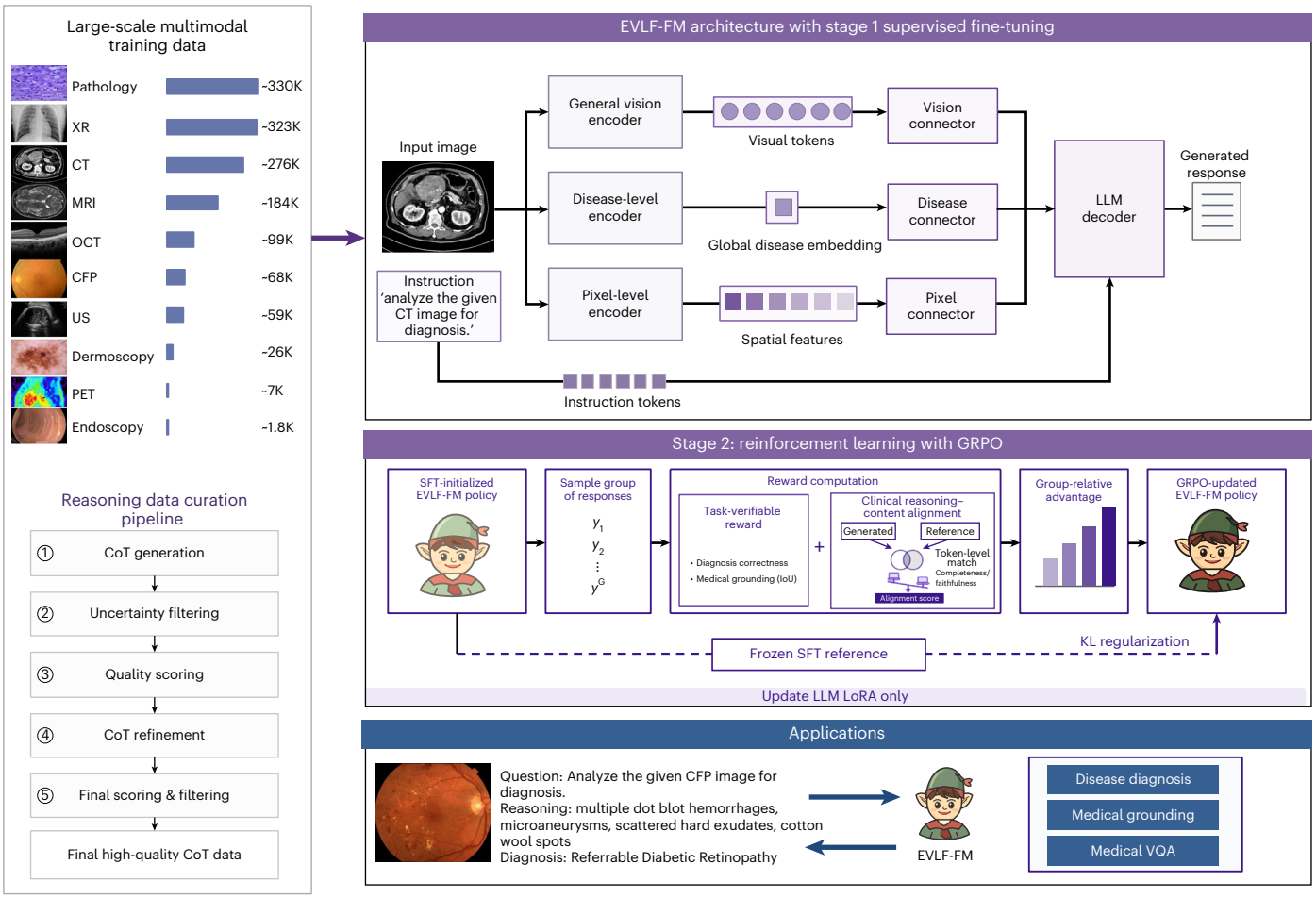


Fig. 1 | Overview of the EVLF-FM framework, training strategy and clinical capabilities. EVLF-FM was developed using approximately 1.3 million multimodal medical images spanning 10 imaging modalities. The model was trained through a two-stage pipeline consisting of SFT with disease-level and pixel-level encoders, followed by RFT using GRPO with task-specific reward functions. The resulting model supports three core clinical capabilities: disease diagnosis, VQA and medical visual grounding. ~, approximately; KL, Kullback–Leibler.

However, FMs continue to exhibit important limitations. First, unimodal FMs, although highly performant within their own domains, are computationally expensive and impractical to scale across the breadth of clinical practice[2,3]. Multimodal models tend to underperform on tasks that require pixel-level precision[10]. These tasks include segmentation, where anatomical regions (for example, organs, tumors or vasculature) are annotated to facilitate measurement and comparison, as well as medical visual grounding, which involves defining relevant parts of an image in response to a query or report[11,12]. The tradeoff between the multimodal versatility with pixel-level performance and portability raises concerns about whether these models can be used within high-stakes healthcare settings. A second persistent challenge concerns the trustworthiness of model outputs, encompassing factual correctness, explainability and reasoning faithfulness—that is, ensuring that generated rationales genuinely reflect diagnostic evidence rather than plausible but post hoc narratives. Although explainable AI tools such as saliency maps and SHapley Additive exPlanations (SHAP) values have been explored in conventional deep learning systems, most are not directly applicable to emerging FMs[13,14]. Most FMs still function as black boxes, limiting transparency and reducing clinician confidence in their output.

To address these challenges, we introduce EVLF-FM, a state-of-the-art medical vision language model (VLM) designed to integrate the breadth of generalist models with the pixel-level precision and explainability of specialist systems (Fig. 1). EVLF-FM supports three core clinical tasks—disease diagnosis, visual grounding and medical visual question answering (VQA)—while providing textual explanations for its predictions. EVLF-FM was trained on one of the most diverse and comprehensive medical vision language training corpora assembled to date, including over 1.3 million curated clinical images spanning six clinical specialties and 10 imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), X-ray (XR), ultrasound (US), positron emission tomography (PET), histopathology images, dermatoscopy, color fundus photography (CFP), optical coherence tomography (OCT) and endoscopic images. It incorporates supervised fine-tuning (SFT) and visual reinforcement fine-tuning (RFT) within a generalist FM framework, aligning high-level diagnostic reasoning with fine-grained spatial localization. Unlike conventional models that provide only a final prediction, EVLF-FM generates assistive explanations that highlight clinically relevant features and localize image regions associated with its outputs, linking predictions to visual evidence. These explanations are post hoc, and their faithfulness to the underlying decision process is explicitly evaluated. EVLF-FM enables both versatile performance across modalities and enhanced model interpretability, representing a step toward the development of clinically trustworthy GMAI.

Results

To evaluate the performance of EVLF-FM, we benchmarked the model against state-of-the-art generalist VLMs as well as specialist models across

core medical AI tasks, including image-based diagnosis, VQA and medical visual grounding, using a wide range of imaging modalities. Next, we assessed zero-shot and few-shot capabilities on external datasets not seen during training, comparing the performance of EVLF-FM with other VLMs. In addition, we conducted human evaluation of the model's explanatory outputs, along with subgroup performance, ablation and data efficiency analysis, to assess the fairness and robustness of model outputs.

Internal validation across core medical AI tasks

Medical image diagnostics. We benchmarked the performance of EVLF-FM in image diagnostics with four other generalist VLM models (LLaVA version 1.5 13B, Qwen2.5-VL 7B, LLava-Med V1.5 7B and InternVL 8B); six specialist FMs (Virchow, RETFound-CFP, RETFound-OCT, Rad-DINO, USFM and BiomedCLIP); and eight specialist deep learning models (VGG 16, ResNet 50, DenseNet 121, EfficientNet B0, ViT, DINO, CLIP and SAM)[2,3,5,15–28]. More than 1.3 million training images were assembled from datasets spanning dermatoscopy (HAM10000, BCN20000), histopathology images (MIDOG25, LC25000), XR (RSNA Pneumonia, NIH ChestX-ray14), US (BUSI, Fetal_Planes), CFP (PAPILA, OAEFD-CFP), OCT (OAEFD-OCT) and CT (RAPIER), together with the 12 datasets from MedMNIST (PathMNIST, ChestMNIST, DermaMNIST, OCTMNIST, PneumoniaMNIST, RetinaMNIST, BreastMNIST, BloodMNIST, TissueMNIST, OrganAMNIST, OrganCMNIST and OrganSMNIST)[29–39].

EVLF-FM demonstrated state-of-the-art diagnostic performance across multiple internal benchmarks, achieving the highest overall mean accuracy (0.895) among all evaluated models (Fig. 2a,b and Extended Data Table 1). Across the internal validation datasets, EVLF-FM consistently outperformed all generalist VLMs in both accuracy and F1 score. Compared to specialist models, it achieved the highest accuracy among specialist deep learning models on six of 11 datasets and among specialist FMs on four of 11 datasets. Specifically, EVLF-FM attained the highest accuracy among all models in key datasets, such as BCN20000 (0.949), BUSI (0.932), OAEFD-OCT (0.990) and MedMNIST (0.851), while also achieving the best F1 scores across multiple datasets, including BUSI (0.926), OAEFD-OCT (0.977) and MedMNIST (0.808). Although certain specialist models demonstrated competitive performance in isolated tasks—for example, ResNet50 in RSNA Pneumonia (accuracy 0.841) and LC25000 (accuracy 1.000)—these models exhibited reduced generalizability across datasets. By contrast, EVLF-FM maintained consistently strong performance across all tasks, highlighting its robustness as a generalist medical VLM.

Within the MedMNIST benchmark (Extended Data Table 2), EVLF-FM similarly achieved the highest mean accuracy (0.851) and F1 score (0.808), outperforming all generalist and specialist models overall. In terms of accuracy, EVLF-FM exceeded generalist VLMs in 11 of 12 tasks and achieved the highest accuracy among specialist models in five of 12 tasks, including ties. Similarly, EVLF-FM achieved higher F1 scores than all generalist VLMs across all 12 tasks and specialist models in five of 12 tasks. Specialist models generally retained advantages on tasks closely aligned with their training domains, particularly Virchow on TissueMNIST, Rad-DINO on PneumoniaMNIST and RETFound CFP on RetinaMNIST. The comparatively lower performance observed in the ChestMNIST dataset may be attributable to its multi-label structure and substantial class imbalance. To contextualize chest XR performance across datasets and comparator models, we conducted a per-class evaluation on two internal (NIH ChestX-ray14 and VinDr-CXR) and one external (AIMx) chest XR cohorts, benchmarking EVLF-FM against InternVL3-8B and MedGemma 27B on four representative classes (consolidation/pneumonia, pulmonary mass, pneumothorax and no-finding; Extended Data Table 3). EVLF-FM consistently achieved the highest mean F1 score and competitive mean accuracy across all three cohorts (NIH ChestX-ray14: 0.751/0.545; VinDr-CXR: 0.928/0.629; AIMx: 0.840/0.668) and obtained the best per-class performance on the majority of settings. Although other models achieved higher accuracy for certain pathologies, such as InternVL for consolidation/pneumonia on NIH ChestX-ray14 (0.690 versus 0.634 for EVLF-FM), EVLF-FM achieved a substantially higher corresponding F1 score (0.571 versus 0.340). This discrepancy between accuracy and F1 score suggests that the apparently higher accuracy of comparator models may partly be due to class imbalance, whereas EVLF-FM demonstrated more balanced classification performance.

VQA. VQA tasks challenge models to generate accurate answers by integrating visual and textual information. We benchmarked the performance of EVLF-FM in VQA against five specialized models: RadFM, LLaVA-Med, Med-Flamingo, MedDr and InternVL 8B[17,40–43]. Three datasets were used for internal validation: VQA-RAD, SLAKE and Path-VQA[44–46]. Model performance was evaluated in terms of closed accuracy with multiple-choice questions as well as open recall with open-ended questions.

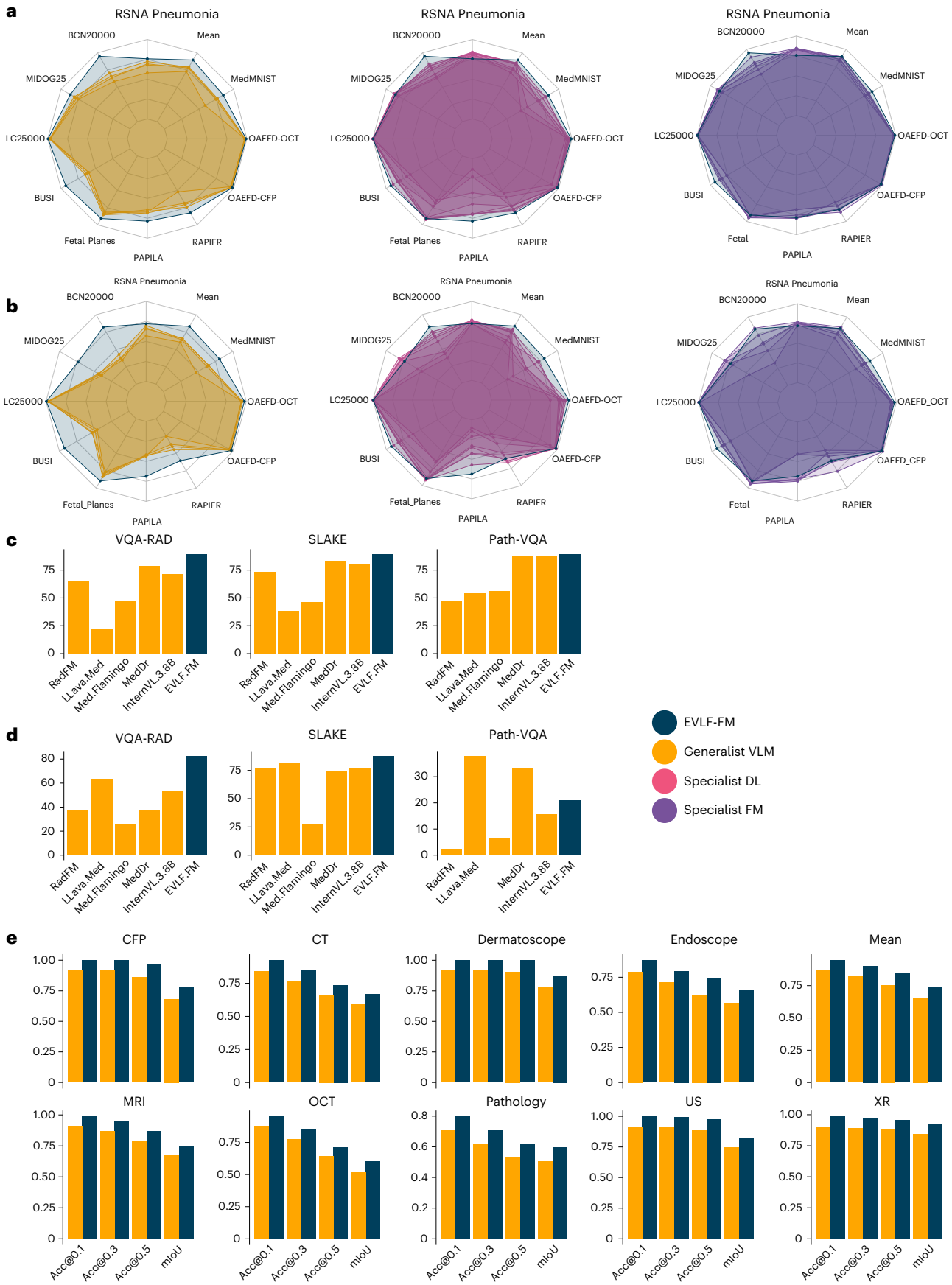
EVLF-FM demonstrated exceptional closed accuracy (Fig. 2c) and open recall (Fig. 2d). Closed accuracy was consistently high across all three datasets, with performance around 90% (Extended Data Table 4). Open recall, which assesses the ability of the model to generate contextually appropriate free-text responses, is inherently more challenging than closed-ended VQA. Nevertheless, EVLF-FM achieved the highest open recall of all tested models on both VQA-RAD and SLAKE, exceeding 80%, highlighting its strong reasoning capabilities. Moreover, EVLF-FM maintained consistently high performance across datasets and task formats. For example, although InternVL 8B achieved high open recall (77.5%) on the SLAKE dataset, its open recall dropped substantially to 52.9% on the VQA-RAD dataset, suggesting limited consistency across different formats and topics. By contrast, EVLF-FM consistently performed well on both closed-ended and open-ended VQA tasks.

Medical visual grounding. We assessed EVLF-FM's medical visual grounding capabilities on BiomedParse and VinDr-CXR datasets across nine different imaging modalities (MRI, pathology slides, XR, CT, OCT, CFP, US, dermatoscopy and endoscopy) using conventional localization metrics—Acc@0.1, Acc@0.3 and Acc@0.5—and mean intersection over union (mIoU)[47,48]. In these evaluations, EVLF-FM was compared against a strong multimodal comparator (InternVL 8B)[43]. EVLF-FM consistently outperformed InternVL 8B across all metrics (Fig. 2e). Averaged over all modalities, EVLF-FM achieved Acc@0.1 = 0.946, Acc@0.3 = 0.902, Acc@0.5 = 0.842 and mIoU = 0.741, surpassing the baseline models' corresponding scores of 0.865, 0.82, 0.755 and 0.656.

At the modality level, EVLF-FM consistently outperformed InternVL 3 8B across all nine imaging modalities and all four localization metrics (Extended Data Table 5). The largest gains in Acc@0.5 were observed for endoscopic images (0.742 versus 0.625), CFP (0.968 versus 0.86) and dermatoscopy (1 versus 0.905). Similarly, EVLF-FM demonstrated the largest improvements in localization overlap for CFP (mIoU: 0.781 versus 0.682) and endoscopic images (0.660 versus 0.564), indicating more accurate localization of clinically relevant regions. These consistent improvements across diverse imaging modalities highlight the strong generalizability of EVLF-FM for medical visual grounding.

Fig. 2 | Internal validation results of EVLF-FM compared against generalist VLMs, specialist deep learning and specialist FMs. a, Accuracy of EVLF-FM and comparator models across 11 diagnostic tasks, with separate radar plots comparing EVLF-FM against generalist VLMs, specialist deep learning models and specialist FMs. **b**, F1 score of EVLF-FM and comparator models across the same 11 diagnostic tasks, shown using the same three comparator groups. **c**, Closed-ended VQA accuracy of EVLF-FM and generalist VLM comparators across three benchmark datasets. **d**, Open-ended VQA recall of EVLF-FM and generalist VLM comparators across three benchmark datasets. **e**, Visual grounding performance of EVLF-FM compared to a generalist VLM across nine imaging modalities, evaluated using accuracy at IoU thresholds of 10%, 30% and 50% (Acc@0.1, Acc@0.3 and Acc@0.5), together with mIoU. DL, deep learning.

https://doi.org/10.1038/s41591-026-04655-5



External testing of zero-shot and few-shot performance

Comparison with VLMs for diagnostics. We compared the performance of EVLF-FM in image diagnostics against five other VLMs: Internvl3 8B, HuatuoGPT-Vision 7B, MedDr 32B MedGemma 4B and MedGemma 27B[42,43,49,50]. Twelve external datasets were used, spanning six different modalities (HRCTCov19 and SARS-CoV-2 for CT; CXR-Pneuomnia for XR; AMI-Br for pathology slices; BUS-UCLM and US3M for US; RetinalOCT and DRCR-OCT for OCT; and APTOS, Glaucoma fundus, SEED and DRCR-CFP for CFP)[51–60]. For EVLF-FM, we evaluated both its zero-shot performance and its few-shot performance, specifically eight-shot, 16-shot, 32-shot and 64-shot.

In the zero-shot setting, EVLF-FM achieved the highest F1 score and accuracy on eight benchmark datasets (Fig. 3a and Extended Data Table 6). Specifically, EVLF-FM attained the top performance on HRCT-Cov19 (0.972/0.969), CXR-Pneumonia (0.894/0.892), US3M (0.613/0.591) RetinalOCT (0.503/0.379), DRCR-OCT (0.861/0.861), DRCR-CFP (0.952/0.770), SEED (0.912/0.823) and Glaucoma fundus (0.746/0.743). By contrast, larger models demonstrated isolated strengths: MedDr-32B achieved superior accuracy on SARS-CoV-2 (0.781), and MedGemma-27B outperformed EVLF-FM on BUS-UCLM (0.679/0.519) in terms of both accuracy and F1 and APTOS in F1 (0.440). However, these gains were dataset specific and not consistently maintained across modalities. Notably, despite having a substantially smaller parameter size (8B), EVLF-FM remained competitive with or superior to larger models such as MedGemma 27B (27B) and MedDr 32B (32B), highlighting its efficiency and generalizability.

As anticipated, performance across diagnostic tasks improved with the incorporation of few-shot examples (Fig. 3b and Extended Data Table 6). The largest improvement in accuracy was observed in RetinalOCT, from 0.503 in the zero-shot setting to 0.797 in the 64-shot setting, whereas the largest improvement in F1 score was seen in BUS-UCLM, from 0.200 to 0.780. Improvement was observed across most datasets, including HRCT-Cov19 (0.972/0.969 to 0.995/0.994), CXR-Pneumonia (0.894/0.892 to 0.966/0.964) and Glaucoma fundus (0.746/0.743 to 0.776/0.776). Interestingly, increased shot count did not uniformly translate to better performance. On the DRCR-CFP dataset, the zero-shot model achieved the highest accuracy (0.952), and the highest F1 score was achieved by the eight-shot model (0.771). These results suggest that, although few-shot tuning enhances diagnostic accuracy, optimal performance is not strictly dependent on the number of examples provided.

Comparison with VLMs for grounding. We benchmarked the performance of EVLF-FM in visual grounding against BiRD and MedGround on the Med-GRIT-270k dataset, which spans eight imaging modalities, including CT, MRI, XR, PET, endoscopy, dermatoscopy, CFP and US[61,62]. This dataset also serves as the training set for BiRD, providing a strong fully supervised baseline for comparison.

EVLF-FM achieved strong performance on the Med-GRIT-270k benchmark (Fig. 3c and Extended Data Table 6). In the zero-shot setting without Med-GRIT fine-tuning, EVLF-FM attained an average Acc@0.5 of 0.360, outperforming MedGround (0.317). Notably, MedGround achieved similar performance only after few-shot fine-tuning. With few-shot fine-tuning using a small subset of the training data, EVLF-FM further improved to an average Acc@0.5 of 0.545, surpassing BiRD, while requiring only a fraction of the training data. At the modality level, EVLF-FM achieved superior performance to BiRD on MRI (0.332 versus 0.293), dermatoscopy (0.882 versus 0.756), CFP (0.928 versus 0.842) and US (0.530 versus 0.460). Although BiRD retained higher performance on CT, XR, PET and endoscopy, the overall results demonstrate that EVLF-FM achieves competitive visual grounding performance across diverse imaging modalities with substantially reduced training data requirements.

Human grading of model thinking

To assess the quality of EVLF-FM's generated thinking outputs, we sampled cases from OCT, CFP, chest XR, abdominal CT and pathology images. A total of 12 physicians, including four junior doctors and eight specialists across radiology, pathology and ophthalmology, evaluated the outputs using the S.C.O.R.E framework: safety (S), consensus and context (C), objectivity (O), reproducibility (R) and explainability (E). Each domain was scored from 1 (lowest) to 5 (highest) for each case[63]. Inter-grader agreement was further assessed using Gwet's AC2 coefficient and the percentage of ratings within one point between grader pairs.

Overall, EVLF-FM demonstrated consistently strong performance across all S.C.O.R.E domains, achieving an overall mean score of 4.25 ± 0.94 (Fig. 4a and Extended Data Table 7). Among the five domains, objectivity received the highest score (4.36 ± 0.93), suggesting that generated thinking outputs were generally well supported by identifiable imaging findings. Across imaging modalities, OCT achieved the highest overall performance (4.73 ± 0.50). Further subgroup analysis between junior doctors and specialists demonstrated broadly consistent evaluations across most modalities, although junior doctors generally assigned higher scores (Fig. 4b). The difference between groups reached statistical significance only for CT, where specialist reviewers noted the inherent limitations of interpreting complex volumetric imaging studies from a single representative slice. Inter-grader agreement ranged from substantial to almost perfect (Gwet's AC2: 0.5–0.93), with 74–91% of ratings falling within one point of each other for CT, CFP, OCT and chest XR.

Representative examples from the five evaluated imaging modalities are shown in Fig. 4c, including the input image, model-generated reasoning output, average S.C.O.R.E. domain scores assigned by graders and summarized comments from senior specialists. Overall, EVLF-FM generated structured, stepwise reasoning that explicitly linked visual observations to diagnostic conclusions and was generally well aligned with expert interpretation. Qualitative review from specialists highlighted occasional omission of clinically relevant but non-diagnosis-defining findings, including presence of vitreomacular traction or epiretinal membrane in OCT examples of diabetic macular edema or ancillary radiographic observations such as increased chest volume on chest XR. These findings suggest that, although EVLF-FM can reliably identify major pathological features and generate clinically meaningful reasoning outputs, further improvement in completeness and reporting granularity may enhance its alignment with specialist-level interpretation. Additional examples of thinking output for grounding tasks are included in Fig. 5.

Subgroup analysis based on age and sex

To evaluate performance consistency across demographic subgroups, we conducted stratified analyses by age and biological sex on four datasets with available metadata (HAM10000, RAPIER, PAPILA and AIMx). For AIMx, the analysis was restricted to the 381 samples with available age and sex annotations, benchmarking EVLF-FM against InternVL 3 8B and Qwen 3 VL 8B (Extended Data Table 8)[29,37,43,64].

Across all datasets, EVLF-FM not only achieved the highest overall performance but also consistently outperformed comparator models within each subgroup (male, female, age 0–60 years and age >60 years). For sex-based stratification, EVLF-FM exhibited particularly small subgroup differences on HAM10000 (0.006) and AIMx (0.005) while maintaining the highest subgroup accuracies across all datasets. Although a larger sex-related difference was observed on PAPILA (0.166), EVLF-FM maintained higher accuracy in both male (0.684) and female (0.850) subgroups than InternVL 3 8B (0.632 and 0.717) and Qwen 3 VL 8B (0.605 and 0.717), respectively. The larger gap, therefore, reflected particularly strong performance in the female subgroup rather than reduced performance in the male subgroup.

A similar trend was observed across age groups. EVLF-FM demonstrated substantially reduced age-related performance gaps on HAM10000 (0.018) and RAPIER (0.018) compared to larger disparities for InternVL 3 8B (0.270 and 0.048) and Qwen 3 VL 7B (0.223 and 0.006), respectively. On HAM10000, the age-related performance

https://doi.org/10.1038/s41591-026-04655-5

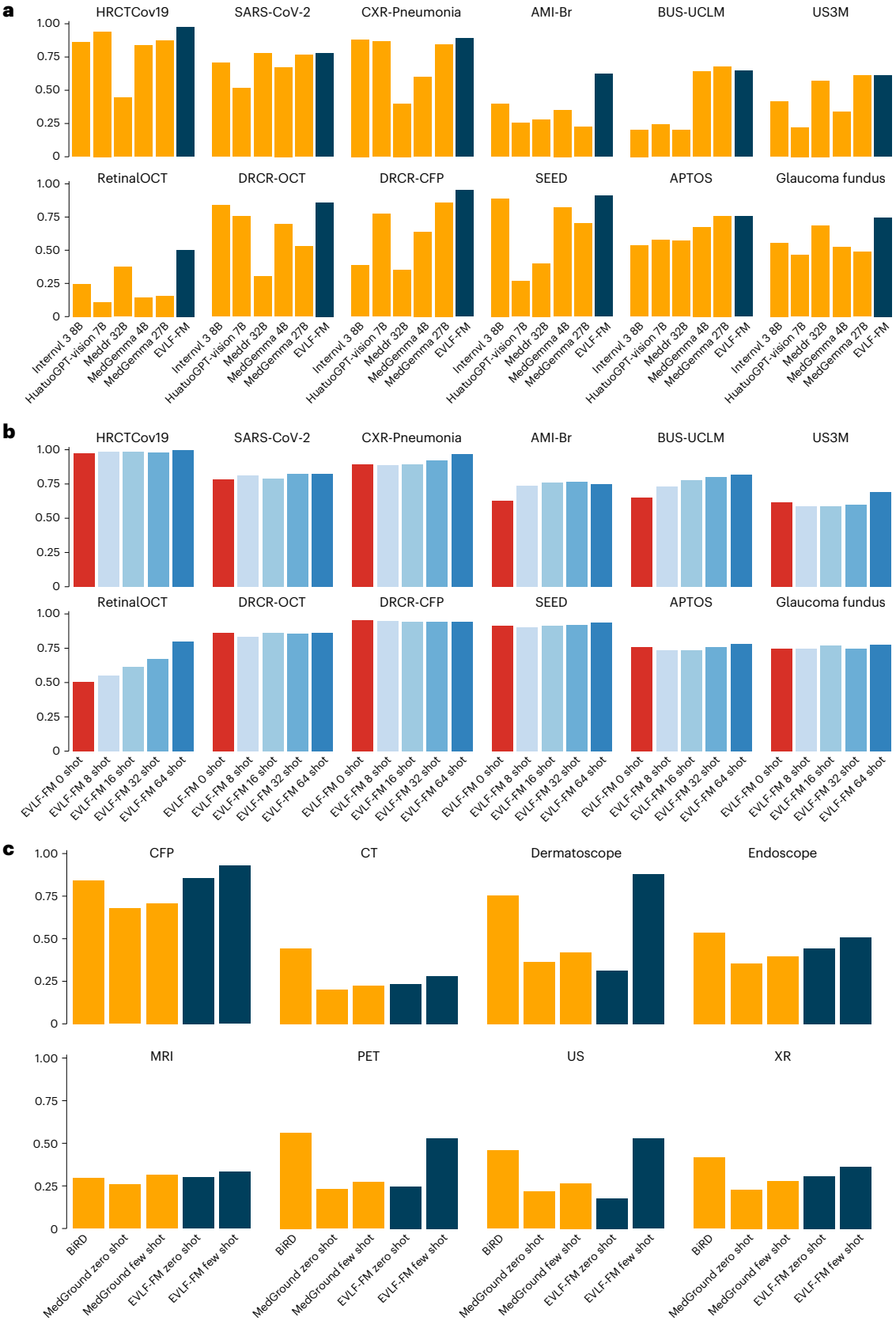


Fig. 3 | External validation performance of EVLF-FM in zero-shot and few-shot settings compared to VLM comparators. a, Zero-shot diagnostic performance, measured by accuracy, of EVLF-FM and comparator VLMs across 12 external datasets spanning diverse clinical imaging modalities. **b**, Zero-shot and few-shot diagnostic performance of EVLF-FM across the same 12 diagnostic tasks, demonstrating improved accuracy across most tasks as the number of in-context examples increases. **c**, Zero-shot and few-shot visual grounding performance of EVLF-FM and comparator models across eight imaging modalities, evaluated using accuracy at an IoU threshold of 50% (Acc@0.5).

https://doi.org/10.1038/s41591-026-04655-5

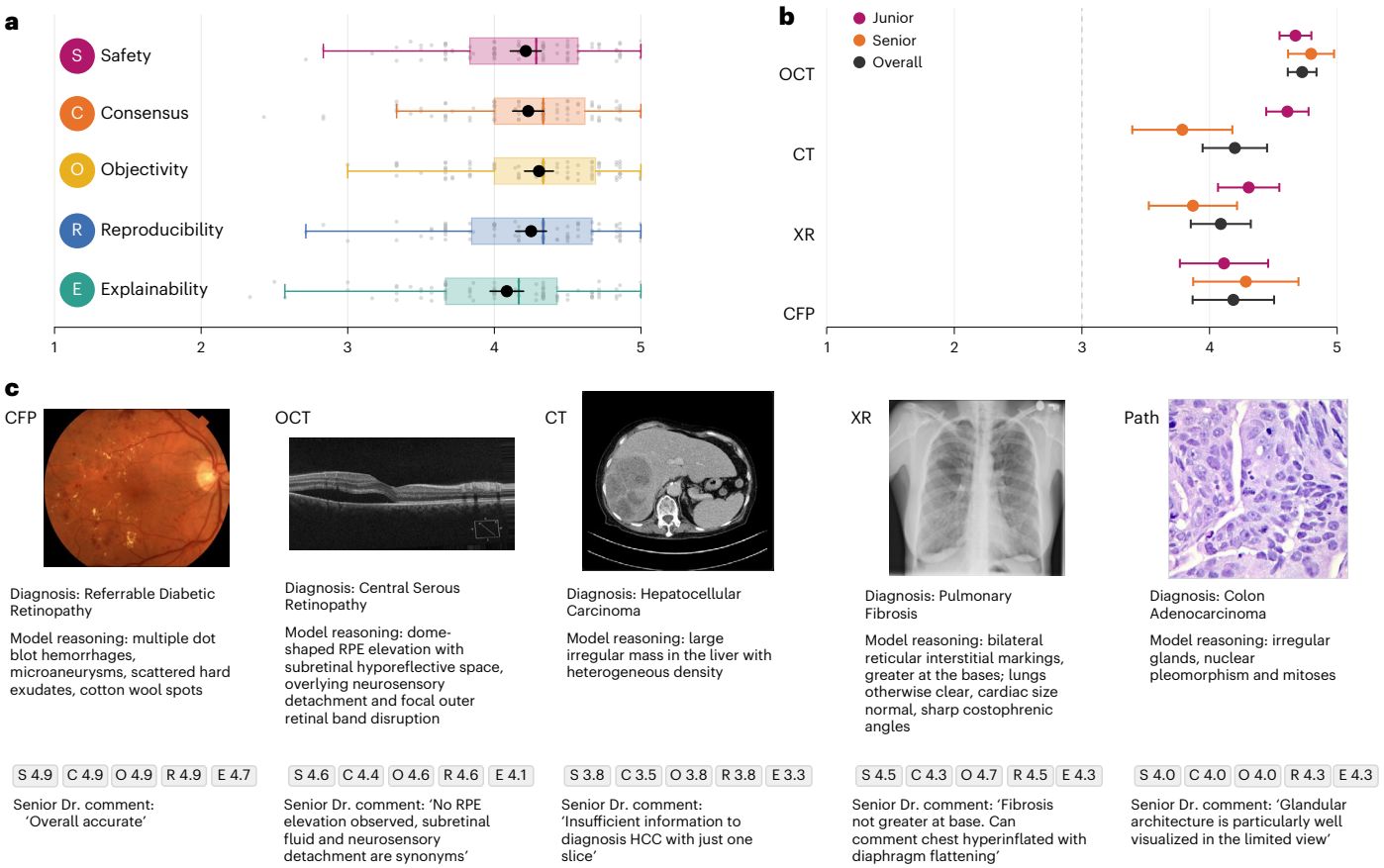


Fig. 4 | Evaluation of EVLF-FM thinking outputs using the S.C.O.R.E. framework across imaging modalities and representative examples. **a**, Domain-level S.C.O.R.E. evaluation of EVLF-FM thinking outputs aggregated across all imaging modalities, including Safety, Consensus, Objectivity, Reproducibility and Explainability. Each gray dot represents the score for one case averaged across the 3–7 clinicians who evaluated it. Ratings from different clinicians for the same case were repeated observer assessments and were not treated as independent replicates. Center lines indicate medians, boxes indicate the 25th–75th percentiles and whiskers extend to the most extreme values within 1.5× the interquartile range. The minimum and maximum are the lowest and highest displayed points. Black points and bars indicate means and two-sided 95% confidence intervals. **b**, Comparison of S.C.O.R.E. evaluation scores assigned by junior doctors, senior specialists and the overall average across imaging modalities. Pathology was excluded because it was evaluated only by senior specialists. Ratings from clinicians assessing the same case were averaged at the case level and were not counted as independent replicates. Points and error bars indicate mean values and two-sided 95% confidence intervals across independent cases. **c**, Representative examples of EVLF-FM thinking outputs across five imaging modalities, showing the input image, predicted diagnosis, model-generated reasoning, average S.C.O.R.E. domain scores assigned by graders and summarized comments from senior specialists. HCC, hepatocellular carcinoma; RPE, retinal pigment epithelium.

gap was reduced by approximately 15-fold relative to InternVL 3 8B (0.018 versus 0.270). Similarly, although larger age-related differences were observed on PAPILA (0.067) and AIMx (0.062), EVLF-FM maintained the highest absolute accuracies in both age groups (PAPILA: 0.743 and 0.810; AIMx: 0.876 and 0.815), exceeding those of the comparator models.

Training strategy ablation analysis

To understand the contribution of each individual component to overall model performance, we performed an ablation study evaluating different components within the EVLF-FM training pipeline, including individual components introduced during first-stage SFT—chain-of-thought (CoT) reasoning data, pixel-level encoder and disease-level encoder—as well as the sentence-level clinical reasoning alignment reward introduced during second-stage group relative policy optimization (GRPO) (Extended Data Table 9).

During SFT training, incorporating curated CoT reasoning data improved model performance compared to the base model (accuracy/F1: 0.811/0.713 versus 0.766/0.648), and further addition of the pixel-level encoder maintained similar overall performance (0.812/0.712). A further improvement was observed after incorporating the disease-level encoder, increasing performance to 0.845/0.769, highlighting its contribution to disease-specific representation learning and diagnostic performance. Applying GRPO alone without SFT initialization resulted in substantially degraded performance (accuracy/F1: 0.676/0.531), confirming that reinforcement learning requires a well-initialized model. SFT alone achieved strong performance (0.845/0.769), and the addition of GRPO improved results to 0.858/0.797, with notable gains in F1 in more challenging datasets such as Breast (0.080) and Derma (0.021). Finally, incorporating the sentence-level clinical reasoning alignment reward yielded the full EVLF-FM model. Although the absolute gains were modest because performance was already near ceiling on several benchmarks, the consistent directional improvements support the effectiveness of the reasoning-aware reward design.

Data efficiency analysis

To assess how the amount of data used for second-stage RFT affects model performance, we conducted a data efficiency analysis. Specifically EVLF-FM was trained using 20% (EVLF-20), 40% (EVLF-40), 60% (EVLF-60) and 80% (EVLF-80) of the original RFT dataset and evaluated on internal diagnostic benchmarks. This analysis provided insight into

https://doi.org/10.1038/s41591-026-04655-5

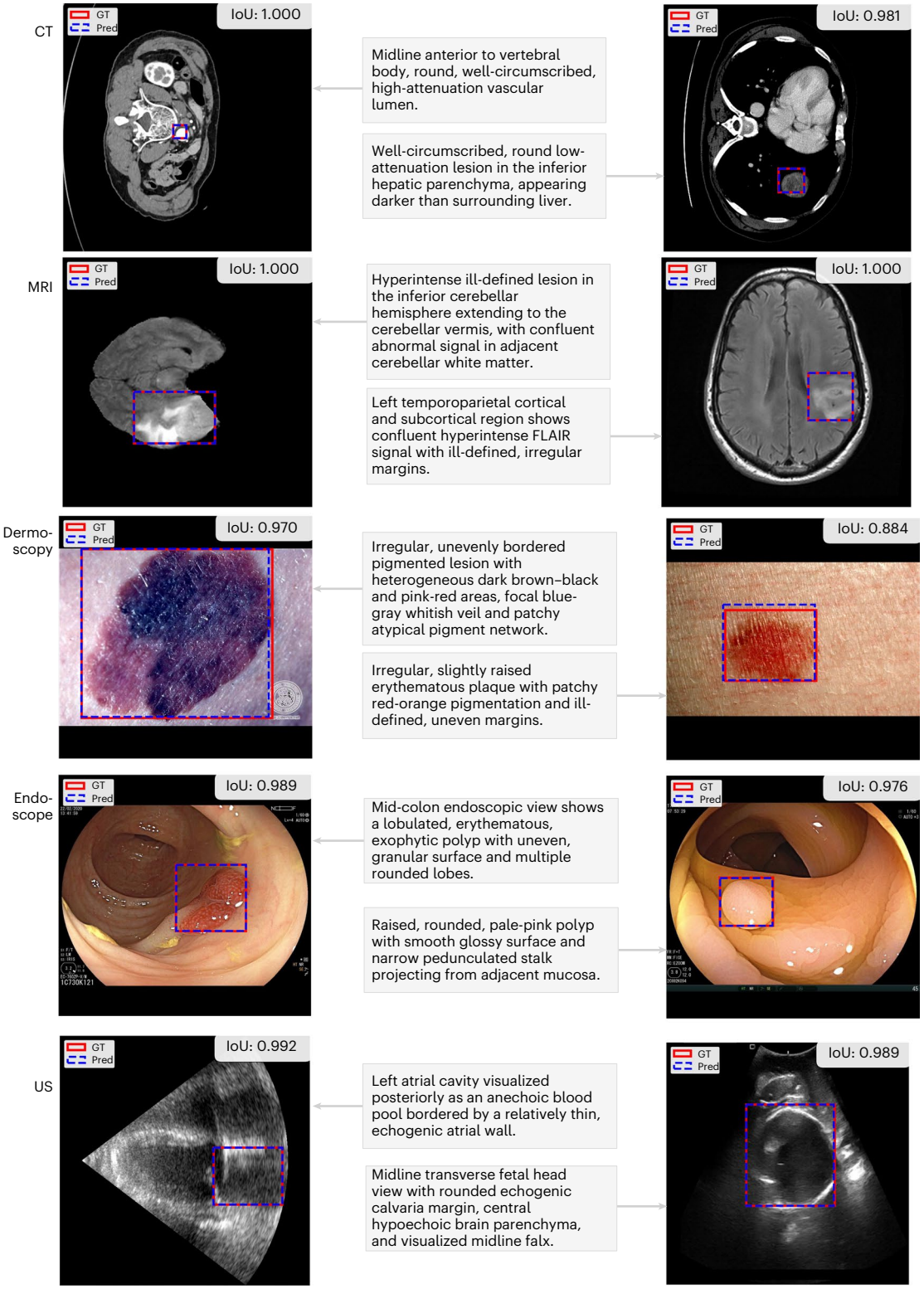


Fig. 5 | Representative examples of visual grounding performance of EVLF-FM across imaging modalities. Representative examples from CT, MRI, dermoscopy, endoscopy and US are shown, including the model-generated grounding boxes, corresponding textual descriptions and comparison with ground truth annotations. Grounding performance is illustrated by the IoU between predicted and ground truth regions. FLAIR, fluid-attenuated inversion recovery; GT, ground truth; Pred, predicted.

the sensitivity of EVLF-FM's performance to the amount of reinforcement data used, informing the scalability and efficiency of fine-tuning (Extended Data Table 9).

EVLF-FM demonstrated strong data efficiency, maintaining robust performance even with limited reinforcement data (Extended Data Table 9). EVLF-20 achieved a mean accuracy/F1 score of 0.823/0.766, compared to 0.851/0.808 for the fully trained model, with performance differences generally modest across datasets. Performance improved progressively with increased data, reaching near-saturation by EVLF-80 (0.849/0.806), suggesting diminishing

returns beyond moderate data scaling. Notably, gains were more pronounced in challenging tasks such as ChestMNIST (F1: 0.227 to 0.313) and BloodMNIST (F1: 0.888 to 0.983), whereas high-performing tasks such as PneumoniaMNIST showed only marginal improvements, indicating early convergence. Collectively, these results demonstrate that EVLF-FM achieves strong performance with efficient use of training data while benefiting from a synergistic combination of supervised initialization and reinforcement learning with clinically aligned objectives.

Repeat-prompt consistency evaluation

To assess the reliability of EVLF-FM, we evaluated its response consistency under different prompt formulations by rephrasing each clinical diagnostic question using 10 distinct prompt templates. This experiment was conducted on the internal validation datasets for diagnostic tasks. For each evaluation sample, one of the 10 prompt variants was randomly selected; the complete set of prompt templates is provided in Extended Data Table 10.

EVLF-FM demonstrated highly consistent performance across all 12 datasets despite variation in prompt phrasing (Extended Data Table 10). Using randomly selected prompt variants resulted in only minimal changes in overall performance (mean accuracy/F1 score: 0.845/0.801 versus 0.851/0.808 for the base prompt), with performance remaining stable across all imaging modalities. These findings demonstrate that EVLF-FM is robust to natural language variation in user prompts—an essential characteristic for dependable deployment in real-world clinical practice.

Discussion

EVLF-FM is a multimodal medical VLM designed with additional emphasis on reasoning to enhance explainability in clinical settings. After training with one of the largest and most diverse clinical datasets assembled, EVLF-FM performs three key tasks: disease classification, medical VQA and visual grounding. It demonstrates state-of-the-art performance and consistency in both internal and external validation, outperforming existing generalist VLMs as well as specialist architectures. Through a unique training strategy combining SFT and RFT, EVLF-FM also exhibits robust reasoning capabilities, aligning outputs with visual evidence. These findings demonstrate EVLF-FM's potential as a scalable, versatile and yet highly explainable clinical decision-support tool, supporting efforts to develop trustworthy, human-aligned AI for healthcare applications.

EVLF-FM introduces several key innovations and strengths. The three core tasks of EVLF-FM, comprising disease diagnosis, medical visual grounding and VQA, support a variety of downstream clinical applications. These tasks are designed to simulate real-life workflow: identifying anatomical structures on imaging, recognizing pathological findings and localizing them precisely with confidence. By incorporating explanation-driven outputs through both explicit stepwise reasoning and pixel-level grounding, EVLF-FM delivers operational interpretability and auditability far beyond traditional black-box models—an essential feature for clinical decision-making and trustworthiness[65]. Another major advance stems from EVLF-FM's hybrid training strategy, which combines FM pretraining with SFT and a novel visual RFT phase implemented via GRPO. SFT helps the model adapt to domain-specific tasks and ensures strong baseline performance while requiring less task-specific data and computational overhead compared to training multiple specialist models[42]. Visual RFT addresses key limitations of supervised schemata by encouraging the model to generate structured, explainable reasoning rather than relying on memorized answers[66]. To promote reasoning faithfulness during training, we introduce a sentence-level clinical reasoning alignment reward that encourages both completeness and faithfulness of generated rationales against expert references[67]. This reasoning-aware reward design likely contributes to the superior performance of EVLF-FM across all three tasks relative to other VLMs. Reasoning quality is then independently assessed through clinician evaluation using the S.C.O.R.E. framework. By leveraging verifiable, task-specific rewards for reasoning coherence and grounding accuracy, this method promotes generalization to unfamiliar clinical settings and enhances interpretability, making EVLF-FM both efficient and trustworthy for real-world deployment, and is robust to differences in image format.

The design and capabilities of EVLF-FM lend themselves to further development of applications that can change clinical practice and education. First, versatility across a multitude of clinical imaging media can support healthcare delivery in a range of settings, across many specialties. As a generalizable backbone model, EVLF-FM can be adapted to numerous downstream applications through task-specific fine-tuning, enabling optimization for specific clinical tasks, patient populations or healthcare environments. Crucially, fine-tuning may be undertaken locally to preserve privacy if using local data. In addition, EVLF-FM's ability to identify regions of abnormality and generate clinically relevant reasoning may provide additional support for image interpretation, workflow prioritization and decision-making. Such capabilities could help clinicians manage increasing imaging workloads by assisting with triage and prioritizing cases that require faster review[68]. Furthermore, its favorable evaluation by junior doctors in particular using the S.C.O.R.E framework suggests that EVLF-FM may have the potential to augment clinical capacity in resource-limited or high-volume settings, such as emergency departments where rapid preliminary assessment are especially valuable. This capability may also be beneficial in lower-income regions where access to specialist expertise and timely image interpretation remains limited[69].

Beyond clinical care, EVLF-FM holds considerable value for medical education. It can enable both medical students and junior doctors to explore cases through interactive diagnosis, grounding and VQA, with visual justifications that support self-directed learning and reinforce clinical reasoning skills[70,71]. Additionally, similar capabilities could be adapted for patient-facing education, helping individuals better understand their health conditions and care plans. Integration with chatbot applications could further improve patient communication by explaining diagnoses, prognosis and treatment options[10,72]. A previous study conducted in China found that patients with diabetic retinopathy who received care from both primary care providers and large language models (LLMs) exhibited better self-management behaviors and higher adherence to follow-up treatment compared to those managed solely by primary care providers[73]. These improvements are likely attributable to the LLM's capacity to deliver personalized, data-informed recommendations and to facilitate structured communication that supports shared decision-making between patients and clinicians.

Despite its strengths, EVLF-FM has several limitations that warrant further development work. Although the model demonstrated relatively stable performance across multiple datasets and tasks overall, its current accuracy does not yet reach the level typically required for direct clinical deployment (for example, consistently exceeding approximately 90% accuracy across datasets). In particular, performance on VQA tasks was generally lower than on diagnostic classification tasks, a pattern also observed in other medical FMs. This difference likely reflects the inherently greater complexity of VQA, which requires multimodal reasoning and language generation, as well as the relatively limited number of high-quality annotated VQA samples available for training compared to classification datasets. Similarly, further improvement in model reasoning remains an important future direction, as developing advanced reasoning capabilities requires large-scale datasets with high-quality medical images paired with expert annotations, reports and explanations that reflect clinical decision-making processes. Addressing this bottleneck will require closer collaboration between clinicians and engineers to generate richer multimodal training data, and emerging approaches, such as knowledge-enhanced pretraining, may further strengthen the reasoning capabilities of

medical FMs[74]. Rather than representing a deployment-ready system, EVLF-FM should, therefore, be viewed as a generalizable multimodal backbone that can serve as a starting point for downstream applications. With additional task-specific fine-tuning, domain adaptation and integration with clinical workflows, the model has the potential to support a wide range of medical AI applications across specialties.

Another important limitation is that EVLF-FM currently operates on static, two-dimensional image inputs and does not explicitly process volumetric medical imaging data. Although recent FMs have begun to incorporate volumetric medical imaging such as CT and OCT, these models are typically developed for specific modalities and clinical domains with limited generalizability across diverse imaging types[75,76]. By contrast, EVLF-FM was designed as a broadly applicable multimodal FM across multiple specialties and imaging modalities, highlighting an important tradeoff between versatility and modality-specific optimization in current medical AI development. Future work may explore extending EVLF-FM toward volumetric and time-resolved imaging modalities, provided that sufficiently large-scale, high-quality datasets become available to support training while preserving its broad multimodal capabilities.

Beyond task-specific differences, concerns about model generalizability and bias also remain. Like other large-scale models, EVLF-FM is susceptible to inheriting biases from its training data, which may lead to reduced performance in underrepresented patient populations, clinical environments or imaging modalities[77,78]. Although our training dataset is large by current standards, it still captures a limited portion of real-world diversity. If the open-source data used are not representative, particularly where certain demographic groups, imaging devices or rare pathologies are underrepresented, EVLF-FM's predictions in these contexts may be less reliable[77,78]. Addressing these limitations will require future efforts to curate more diverse, high-quality and representative datasets to ensure broader clinical applicability and fairness. Future work should prioritize curating more diverse and representative training datasets to mitigate this issue.

Additionally, the clinical integration of AI models raises important ethical and practical considerations. EVLF-FM is envisaged as the backbone of human-in-the-loop decision-support tools that augment, rather than replace, clinician judgment[79,80]. In this framework, although AI outputs can provide additional insights that may assist in pattern recognition, risk stratification or information synthesis, clinicians retain ultimate responsibility for interpretation and clinical decision-making[81]. Such a collaborative approach may help mitigate the risks associated with overreliance on automated systems while preserving clinical oversight. Rigorous validation in prospective clinical studies, transparency in model use and clear communication of confidence and uncertainty predictions will be necessary to build trust and minimize the risk of errors leading to harm.

In future work, several directions may further advance the development and clinical applicability of EVLF-FM. For instance, additional validation across real-world datasets and healthcare environments will be important to better assess the model's robustness and generalizability. Collaboration with international clinical and research partners may enable access to larger datasets that are not publicly available[82]. Fine-tuning on such datasets may further improve model performance in specialized clinical tasks while maintaining the advantages of the general representations learned during FM pretraining. Other groups may undertake local fine-tuning where data sharing is not possible, with the aim of achieving optimal performance in a particular setting. For any intervention designed to change the clinical decision-making process, prospective evaluation in clinical settings will be essential to understand the real-world impact of EVLF-FM. These studies may include controlled prospective studies or randomized controlled trials designed to evaluate how integrating EVLF-FM into clinical workflows influences outcomes such as diagnostic accuracy, workflow efficiency, clinician workload and user trust[83,84]. Evaluating the model within a human-in-the-loop framework will also help identify potential limitations, workflow challenges and opportunities for refinement. Insights from these studies may guide further model improvement and support the safe and effective integration of multimodal FMs into clinical practice.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-026-04655-5.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
3. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
4. Teo, Z. L. et al. Generative artificial intelligence in medicine. *Nat. Med.* **31**, 3270–3282 (2025).
5. Zhang, S. et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* **2**, AIoa2400640 (2025).
6. Tu, T. et al. Towards conversational diagnostic artificial intelligence. *Nature* **642**, 442–450 (2025).
7. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
8. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
9. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
10. Zhang, K. et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat. Med.* **30**, 3129–3141 (2024).
11. Wang, R. et al. Medical image segmentation using deep learning: a survey. *IET Image Process.* **16**, 1243–1267 (2022).
12. Yang, H. et al. Multimodal self-supervised learning for lesion localization. Preprint at https://doi.org/10.48550/arXiv.2401.01524 (2024).
13. Foo, L. L. et al. Deep learning system to predict the 5-year risk of high myopia using fundus imaging in children. *NPJ Digit. Med.* **6**, 10 (2023).
14. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. Preprint at https://doi.org/10.48550/arXiv.1705.07874 (2017).
15. Liu, H., Li, C., Li, Y. & Lee, Y. J. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 26286–26296 (IEEE, 2024).
16. Bai, S. et al. Qwen2.5-VL Technical Report. Preprint at https://doi.org/10.48550/arXiv.2502.13923 (2025).
17. Li, C. et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. Preprint at https://doi.org/10.48550/arXiv.2306.00890 (2023).
18. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at https://doi.org/10.48550/arXiv.1409.1556 (2014).
19. Pérez-García, F. et al. Exploring scalable medical image encoders beyond text supervision. *Nat. Mach. Intell.* **7**, 119–130 (2025).
20. Jiao, J. et al. USFM: a universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Med. Image Anal.* **96**, 103202 (2024).
21. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/CVPR.2016.90 (IEEE, 2016).

22. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. Densely connected convolutional cetworks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269 (IEEE, 2017).
23. Tan, M. & Le, Q.V. EfficientNet: rethinking model scaling for convolutional neural networks. Preprint at https://doi.org/10.48550/arXiv.1905.11946 (2019).
24. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2020).
25. Radford, A. et al. Learning transferable visual models from natural language supervision. Preprint at https://doi.org/10.48550/arXiv.2103.00020 (2021).
26. Caron, M., et al. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 9630–9640 (IEEE Computer Society, 2021).
27. Kirillov, A. et al. Segment anything. Preprint at https://doi.org/10.48550/arXiv.2304.02643 (2023).
28. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
29. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Nat. Sci. Data* **5**, 180161 (2018).
30. Hernández-Pérez, C. et al. BCN20000: dermoscopic lesions in the wild. *Sci. Data* **11**, 641 (2024).
31. Aubreville, M. et al. A comprehensive multi-domain dataset for mitotic figure detection. *Sci. Data* **10**, 484 (2023).
32. Borkowski, A. A. et al. Lung and colon cancer histopathological image dataset (LC25000). Preprint at https://doi.org/10.48550/arXiv.1912.12142 (2019).
33. Stein, A. et al. RSNA Pneumonia Detection Challenge https://kaggle.com/competitions/rsna-pneumonia-detection-challenge (Kaggle, 2018).
34. Wang, X. et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Preprint at https://doi.org/10.1109/CVPR.2017.369 (2017).
35. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **28**, 104863 (2020).
36. Burgos-Artizzu, X. P. et al. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Sci. Rep.* **10**, 10200 (2020).
37. Kovalyk, O. et al. PAPILA: dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Sci. Data* **9**, 291 (2022).
38. Ocular Disease Recognition https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k (Kaggle, 2020).
39. Yang, J. et al. MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data* **10**, 41 (2023).
40. Wu, C. et al. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nat. Commun.* **16**, 7866 (2025).
41. Moor, M. et al. Med-Flamingo: a multimodal medical few-shot learner. Preprint at https://doi.org/10.48550/arXiv.2307.15189 (2023).
42. He, S. et al. Towards generalizable AI in medicine via generalist–specialist collaboration. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-026-01653-3 (2026).
43. Chen, Z. et al. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 24185–24198 (IEEE, 2024).
44. Lau, J. J., Gayen, S., Ben Abacha, A. & Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* **5**, 180251 (2018).
45. Liu, B. et al. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 1650–1654 (IEEE, 2021).
46. He, X., Zhang, Y., Mou, L., Xing, E. & Xie, P. PathVQA: 30000+ questions for medical visual question answering. Preprint at https://doi.org/10.48550/arXiv.2003.10286 (2020).
47. Zhao, T. et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nat. Methods* **22**, 166–176 (2024).
48. Nguyen, H. Q. et al. VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. *Sci. Data* **9**, 429 (2022).
49. Chen, J. et al. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale. Preprint at https://doi.org/10.48550/arXiv.2406.19280 (2024).
50. Sellergren, A. et al. An open vision-language model for diverse medical applications. *Nat. Med.* https://doi.org/10.1038/s41591-026-04626-w (2026).
51. Abedi, I., Vali, M., Otroshi, B., Zamanian, M. & Bolhasani, H. HRCTCov19-a high-resolution chest CT scan image dataset for COVID-19 diagnosis and differentiation. *BMC Res. Notes* **17**, 32 (2024).
52. Soares, E., Angelov, P., Biaso, S., Froes, M. H. & Abe, D. K. SARS-CoV-2 CT-scan dataset: a large dataset of real patients CT scans for SARS-CoV-2 identification. Preprint at *medRxiv* https://doi.org/10.1101/2020.04.24.20078584 (2020).
53. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
54. Bertram, C. A. et al. Histologic dataset of normal and atypical mitotic figures on human breast cancer (AMi-Br). In *Bildverarbeitung für die Medizin 2025* (eds Palm, C. et al.) 113–118 (Springer Fachmedien Wiesbaden, 2025).
55. Vallez, N., Bueno, G., Deniz, O., Rienda, M. A. & Pastor, C. BUS-UCLM: breast ultrasound lesion segmentation dataset. *Sci. Data* **12**, 242 (2025).
56. Multimodal Breast Ultrasound Dataset US3M. https://www.kaggle.com/datasets/timesxy/multimodal-breast-ultrasound-dataset-us3m (Kaggle, 2026).
57. Naren, O. S. Retinal OCT Image Classification - C8. https://doi.org/10.34740/KAGGLE/DSV/2736749 (Kaggle, 2021).
58. Karthik, M. & Sohier, D. APTOS 2019 Blindness Detection. https://kaggle.com/competitions/aptos2019-blindness-detection (Kaggle, 2019).
59. Ahn, J. M. et al. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLoS ONE* **13**, e0207982 (2018).
60. Majithia, S. et al. Cohort profile: the Singapore Epidemiology of Eye Diseases study (SEED). *Int. J. Epidemiol.* **50**, 41–52 (2021).
61. Huang, X., et al. A refer-and-ground multimodal large language model for biomedicine. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (eds. Linguraru, M. G., et al.) 399–409 (Springer Nature Switzerland, 2024).
62. Zou, K. et al. Uncertainty-aware medical diagnostic phrase identification and grounding. *IEEE Trans. Pattern Anal. Mach. Intell.* **47**, 11315–11329 (2025).
63. Tan, T. F. et al. A S.C.O.R.E. framework for evaluating open-ended responses from large language models in healthcare. *Cell Rep. Med.* **7**, 102883 (2026).
64. Bai, S. et al. Qwen3-VL Technical Report. https://doi.org/10.48550/arXiv.2511.21631 (2025).

65. Rosenbacke, R., Melhus, Å, McKee, M. & Stuckler, D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *JMIR AI* **3**, e53207 (2024).
66. Liu, Z. et al. Visual-RFT: visual reinforcement fine-tuning. Preprint at https://doi.org/10.48550/arXiv.2503.01785 (2025).
67. Pan, J., et al. MedVLM-R1: incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025* (eds. Gee, J.C., et al.) 337–347 (Springer Nature Switzerland, 2026).
68. Annarumma, M. et al. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology* **291**, 196–202 (2019).
69. Tan, T. F. et al. Artificial intelligence and digital health in global eye health: opportunities and challenges. *Lancet Glob. Health* **11**, e1432–e1443 (2023).
70. Ng, F. Y. C. et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. *Cell Rep. Med.* **4**, 101230 (2023).
71. Ning, Y. et al. How can artificial intelligence transform the training of medical students and physicians? *Lancet Digit. Health* **7**, 100900 (2025).
72. Huo, B. et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw. Open* **8**, e2457879 (2025).
73. Li, J. et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat. Med.* **30**, 2886–2896 (2024).
74. Zhou, X. et al. Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis. *Cancer Cell* **44**, 777–791 (2026).
75. Liu, Z. et al. A three-dimensional multi-modal foundation model for optical coherence tomography. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-026-01662-2 (2026).
76. Blankemeier, L. et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature* **652**, 1318–1328 (2026).
77. Yu, A. C., Mohajer, B. & Eng, J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol. Artif. Intell.* **4**, e210064 (2022).
78. Yang, Y. et al. Demographic bias of expert-level vision-language foundation models in medical imaging. *Sci. Adv.* **11**, eadq0305 (2025).
79. Liu, M. et al. A translational perspective towards clinical AI fairness. *NPJ Digit. Med.* **6**, 172 (2023).
80. Yu, F. et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).
81. Thirunavukarasu, A. J. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J. R. Soc. Med.* **116**, 181–182 (2023).
82. Tham, Y. C. et al. Building the world's first truly global medical foundation model. *Nat. Med.* **31**, 3580–3585 (2025).
83. Wu, Y. et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nat. Med.* **31**, 3404–3413 (2025).
84. van Amsterdam, W. A. C. et al. Clinical trials for continuously monitored and updated AI systems. *Nat. Med.* **32**, 1972–1974 (2026).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

Yang Bai[1,41], Haoran Cheng[2,3,41], Yang Zhou[1,4,41], Jun Zhou[1], Arun Thirunavukarasu[5,6], Yuhe Ke[3,7], Jie Yao[2,3], Kanae Fukutsu[2,8], Chrystie Wan Ning Quek[2], Ashley Shuen Ying Hong[2], Laura Gutierrez[2], Alessandro Berni[2], Zhen Ling Teo[2,3], Darren Shu Jeng Ting[2,3,9,10], Brian T. Soetikno[11], Christopher S. Nielsen[12], Tobias Elze[13], Zengxiang Li[2,14], Sumit Sharma[15,16], Nicholas Ruthmann[17], Le Dinh Linh[1], Hiok Hong Chan[3], Victor Koh[18,19], Marcus Tan[18,19], Kelvin Z. Li[20,21], Leonard Yip[20,21], Ching Yu Cheng[2,19], Yih Chung Tham[2,19], Gavin Siew Wei Tan[2,3], Leopold Schmetterer[2,3,22,23,24,25,26,27], Rae-Young Kim[28], Damon Wong[2,3], Marcus Ang[2,3], Rahat Hussain[2,3], Jodhbir S. Mehta[2,3], Tin Aung[2,3], Lionel Tim-Ee Cheng[3,29,30], Nguyen Tuan Anh Tran[29], Chee Leong Cheng[3,31], Tien Yin Wong[2,32,33,34], Darius M. Moshfeghi[11], Nan Liu[35,36,37], Iain Beehuat Tan[3,38,39], Soon Thye Lim[3,38], Eyal Klang[40], Tony Kiat Hon Lim[31], Rick Siow Mong Goh[1], Yong Liu[1,42] & Daniel Shu Wei Ting[2,3,10,13,42] ✉

[1]Institute of Advanced Intelligence and Computing (IAIC), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [2]Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore. [3]Duke-NUS Medical School, National University of Singapore, Singapore, Singapore. [4]School of Computer Science, Wuhan University, Wuhan, China. [5]Nuffield Department of Clinical Neurosciences, Medical Sciences Division, University of Oxford, Oxford, UK. [6]International Centre for Eye Health, London School of Hygiene & Tropical Medicine, London, UK. [7]Department of Anesthesia and Perioperative Medicine, Singapore General Hospital, Singapore, Singapore. [8]Department of Ophthalmology, Faculty of Medicine and Graduate School of Medicine, Hokkaido University, Sapporo, Japan. [9]Department of Inflammation and Ageing, College of Medicine and Health, University of Birmingham, Birmingham, UK. [10]Birmingham and Midland Eye Centre, Sandwell and West Birmingham NHS Trust, Birmingham, UK. [11]Byers Eye Institute, Stanford University School of Medicine, Palo Alto, CA, USA. [12]Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada. [13]Massachusetts Eye and Ear, Department of Ophthalmology, Harvard Medical School, Boston, MA, USA. [14]SingHealth AI Office, Singapore, Singapore. [15]Cleveland Clinic Cole Eye Institute, Cleveland, OH, USA. [16]Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH, USA. [17]Heart, Vascular and Thoracic Institute, Cleveland Clinic, Cleveland, OH, USA. [18]Department of Ophthalmology, National University Health System, Singapore, Singapore. [19]Centre for Innovation and Precision Eye Health, Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [20]Department of Ophthalmology, National Healthcare Group, Singapore, Singapore.

[21]Lee Kong Chien School of Medicine, Nanyang Technology University, Singapore, Singapore. [22]SERI-NTU Advanced Ocular Engineering (STANCE) Program, Singapore, Singapore. [23]School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, Singapore, Singapore. [24]Department of Clinical Pharmacology, Medical University of Vienna, Vienna, Austria. [25]Aier Hospital Group, Changsha, China. [26]Fondation Ophtalmologique Adolphe De Rothschild, Paris, France. [27]Center for Medical Physics and Biomedical Engineering, Medical University Vienna, Vienna, Austria. [28]Department of Ophthalmology and Visual Science, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea. [29]Department of Cardiothoracic and Abdominal Radiology, Singapore General Hospital, Singapore, Singapore. [30]SingHealth Community Hospitals, Singapore Health Services, Singapore, Singapore. [31]Department of Anatomical Pathology, Singapore General Hospital, Singapore, Singapore. [32]School of Clinical Medicine, Beijing Tsinghua Changgung Hospital, Tsinghua Medicine, Tsinghua University, Beijing, China. [33]Beijing Visual Science and Translational Eye Research Institute, Beijing Tsinghua Changgung Hospital Eye Center, Tsinghua Medicine, Tsinghua University, Beijing, China. [34]School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University, Beijing, China. [35]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore. [36]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore. [37]NUS AI Institute, National University of Singapore, Singapore, Singapore. [38]Division of Medical Oncology, National Cancer Centre Singapore, Singapore, Singapore. [39]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [40]Generative AI Research Program, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [41]These authors contributed equally: Yang Bai, Haoran Cheng, Yang Zhou. [42]These authors jointly supervised this work: Yong Liu, Daniel Shu Wei Ting. ✉e-mail: daniel.ting@duke-nus.edu.sg

Methods

EVLF-FM aims to provide a single, unified solution for classification, detection and interpretative tasks in the medical domain. The framework comprises three main components. (1) Disease-level encoder: a vision encoder that captures global, high-level image features to support improved diagnostic reasoning. This encoder is fine-tuned on large-scale medical imaging data to embed expert-level disease context into its representations. (2) Pixel-level encoder: a vision encoder that retains dense, spatially detailed features for region-specific analysis. This component preserves fine-grained pixel-level information (for example, anatomical structures or lesion details) for precise medical grounding and detailed image understanding. (3) Training strategy: a multi-stage training regimen that first performs SFT (using low-rank adaptation (LoRA)) to align and specialize an LLM for medical tasks and then applies reinforcement learning via GRPO to directly optimize performance and reasoning[88]. This approach balances parameter efficiency, preservation of prior vision–language knowledge and domain-specific adaptation.

By unifying disease-level and pixel-level features within the model's architecture, EVLF-FM can address a wide variety of tasks without sacrificing either global domain insight or localized detail.

Disease-level exemplar prompting

To capture clinically salient patterns at a high level, EVLF-FM employs a disease-level encoder denoted $f_{\text{disease}}(\cdot)$, which is fine-tuned on large-scale medical imaging data. Given an input image I, this encoder produces a latent embedding vector e representing the image's global disease-related features: $e = f_{\text{disease}}(I)$.

A learnable vision–language connector $\phi_{\text{disease}}(\cdot)$ then projects this disease-level embedding into the latent space of the language model, yielding a compatible representation $\hat{e} = \phi_{\text{disease}}(e)$.

Through this mechanism, the model can incorporate high-level medical expertise relevant to disease diagnosis and classification tasks. In practice, $\hat{e}$ serves as an exemplar prompt that informs the LLM of the image's overall pathology context, ensuring that subsequent text generation attends to clinically relevant features.

Pixel-level exemplar prompting

To complement the global features, EVLF-FM integrates a pixel-level encoder $f_{\text{pixel}}(\cdot)$ that preserves spatially detailed information for localized analysis. Given the same image I, the pixel-level encoder produces a high-resolution feature map $p \in R^{H\times W\times C}$, where $H \times W$ is the spatial resolution and C is the number of channels in the feature map. Instead of spatially pooling these features into one global feature, we pass the full-resolution feature map through a second connector $\phi_{\text{pixel}}(\cdot)$ to obtain a projected feature map $\hat{p}$ in the LLM's embedding space: $\hat{p} = \phi_{\text{pixel}}(p)$.

By preserving the complete grid of image features, EVLF-FM can supply the language model with fine-grained visual information. This pixel-level prompt enables the framework to perform explicit region-level reasoning, making it well suited for tasks that require identifying specific anatomical locations or lesions (that is, medical visual grounding and detailed image explanation).

Unified exemplar prompting objective

EVLF-FM jointly leverages the disease-level and pixel-level representations during training through a unified learning objective. For each training sample consisting of an image I, an input instruction or question T and the corresponding ground truth response R, the model conditions on both the global feature $\hat{e}$ and the fine-grained feature $\hat{p}$ to generate the response. We train the model to maximize the likelihood of the correct response given these inputs. Formally, the training objective is to minimize the negative log-likelihood:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log p_\theta\left(R_i|I_i, T_i, \hat{e}_i, \hat{p}_i\right),$$

where N is the number of training examples and θ represents the trainable parameters of the model (including the LLM and the connector networks). In this formulation, the disease-level encoder $f_{\text{disease}}(\cdot)$ and pixel-level encoder $f_{\text{pixel}}(\cdot)$ provide crucial fixed representations (after their initial fine-tuning), and the connectors $\phi_{\text{disease}}(\cdot)$ and $\phi_{\text{pixel}}(\cdot)$ adapt those representations into the language model's space. By optimizing L, EVLF-FM learns to generate the appropriate response R (such as a diagnosis, answer or description) given both the image and the associated prompts, effectively uniting high-level and low-level visual cues in one learning process.

SFT with LoRA (stage 1)

In the first training stage, we perform SFT to align the multimodal model with medical domain knowledge and instruction-following capabilities. To preserve the pretrained knowledge of the FM and ensure parameter efficiency, we freeze the core vision–language model weights, including the vision encoders and the original LLM parameters, and introduce LoRA layers into the LLM as well as into the disease-level and pixel-level encoders. We allow only the newly introduced LoRA parameters in the LLM and the connector weights (ϕ_{disease} and ϕ_{pixel}) to be updated during fine-tuning. This way, the model learns to interpret the medical exemplar prompts and generate appropriate outputs for medical tasks, without overfitting or distorting the base model's general language and vision–language capabilities.

Reinforcement learning with GRPO (stage 2)

After SFT, we further refine EVLF-FM using reinforcement learning to directly optimize task performance metrics and encourage correct reasoning behavior. We adopt GRPO, a recent variant of the proximal policy optimization algorithm that is particularly efficient for language model fine-tuning. In conventional proximal policy optimization (PPO), a learned value function (critic) is used to estimate a baseline for advantage calculation, which adds substantial complexity and memory overhead, and it can struggle to assign per-token values in language generation. GRPO abrogates the requirement for a separate critic network; instead, it computes advantages by comparing a group of sampled outputs for the same input and using their average reward as an implicit baseline. This group-based relative advantage estimation aligns well with the comparative nature of reward signals in our setting (because our reward functions, described below, compare each output against the correct answer). It also eliminates the variance introduced by an absolute value baseline and simplifies the reinforcement learning pipeline.

In each reinforcement learning update, we start with the policy model from the SFT stage (treated as the initial policy $\pi_{\theta_{\text{old}}}$) and sample a set of G candidate outputs $\{o_i\}_{i=1}^{G}$ from it for a given input (for example, a particular question or case). We then compute a reward r_i for each output o_i based on task-specific criteria (described below). Let $\bar{r} = \frac{1}{G}\sum_{i=1}^{G} r_i$ be the mean reward for that group of outputs. We define the advantage of each output i as $A_i = r_i - \bar{r}$, which we further normalize by the group's standard deviation to obtain $\hat{A}_i$ (for numerical stability). Intuitively, outputs that score higher than the group average receive a positive advantage, and those below the average get a negative advantage. The model is then updated with policy gradient steps that encourage higher probability for tokens in outputs with positive $\hat{A}_i$ and lower probability for tokens in outputs with negative $\hat{A}_i$. To ensure stable learning, we apply the PPO-style clipping mechanism on the policy update and include a Kullback–Leibler divergence penalty to keep the updated policy from drifting too far from the pretrained/SFT policy. Formally, the GRPO objective for the policy network is adapted from the clipped PPO loss as follows:

$$L_{\text{GRPO}}(\theta) = -E\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right] + \beta D_{\text{KL}}\left(\pi_\theta|\pi_{\text{ref}}\right).$$

Here, π denotes the model's generation policy. $r_t(\theta) = \frac{\pi_\theta(a_t|h_t)}{\pi_{\text{old}}(a_t|h_t)}$ is the probability ratio at each token t between the current policy π_θ and the

the previous-iteration sampling policy π_{old}, and $\widehat{A}_t$ is the normalized advantage assigned to that token (in practice derived from the sequence-level A_i above). The hyperparameters ϵ and β control the clipping range and the strength of the Kullback–Leibler penalty, respectively. This objective encourages the policy to improve on outputs with above-average reward and avoid regressing on outputs with below-average reward while regularizing the policy update to prevent large deviations. Notably, GRPO achieves this without training a separate value function, substantially reducing the complexity and memory overhead of reinforcement learning training and making it feasible to fine-tune a large multimodal model with limited data and compute.

Reward design

We crafted verifiable reward functions for each task so that the model could be trained with feedback signals grounded in objective correctness, without requiring any human in the loop. In our reinforcement learning fine-tuning, we used only a very small fraction of the data, focusing on examples with reliable ground truth annotations to compute rewards. The reward for a given output was computed as follows for the key tasks:

Open-ended visual question answering. For general visual question-answering tasks requiring free-form explanations or reasoning, we did not apply reinforcement learning with direct numeric rewards as automatically grading open-ended answers goes well beyond simple exact-match criteria. Instead, we used F1 score (a standard evaluation metric in VQA that measures token-level overlap as the harmonic mean of precision and recall) to assess output quality—for example, treating the answer and ground truth as bags of tokens to compute F1 score.

Disease classification. The model's output in this task is typically a diagnosis (sometimes accompanied by a short explanation). We assign a reward of +1 if the output contains the correct diagnosis in the expected format and zero otherwise. This encourages the model to not only be correct but also to present the answer in a clear, standardized way for clinical utility.

Pathology detection and localization. In this task, the model's answer includes identification of a finding and an indicated location (for example, the model might output something like 'Detected finding: lung nodule in the left upper lobe; Location: (x, y, width, height)'). The reward is the localization accuracy. Specifically, we compute the IoU between the model's predicted region (for example, a bounding box or segmentation mask) and the ground truth region marked by experts; higher IoU yields a higher reward, and the reward is zero if IoU is below a low threshold (meaning the model pointed to the wrong location entirely).

Sentence-level clinical reasoning alignment

Beyond accuracy-based rewards, we introduced a reward that evaluates the quality of the model's CoT reasoning against reference reasoning. Our method decomposes both the generated and reference reasoning into individual sentences and scores each generated–reference sentence pair by a directional match that aligns their individual clinical terms—crediting partially overlapping findings at the level of terms rather than whole sentences—and gates each match by clinical polarity, so that a statement contradicting the reference earns little or no credit even when its wording is similar (the matching procedure is detailed in Supplementary Note 2, Implementation details of EVLF). For each pair of generated sentence s_i and reference sentence r_j, we obtain a directional match score. From these, we derive two clinically interpretable metrics: Completeness (recall), measuring whether the model's reasoning addresses all key clinical findings identified by the expert, computed as the average, over reference sentences, of the maximum match score to any generated sentence; and Faithfulness (precision), measuring whether every claim in the model's reasoning is grounded in the expert reference, computed as the average, over generated sentences, of the maximum match score to any reference sentence. Crucially, precision naturally penalizes verbose reasoning, as additional unsupported sentences directly reduce the precision score, eliminating the need for ad hoc length penalties. The final reasoning alignment reward is the harmonic mean (F1) of precision and recall, ensuring that neither metric can be trivially maximized at the expense of the other. This reward is combined additively with the task-specific accuracy reward: $r_{\text{total}} = r_{\text{accuracy}} + \alpha r_{\text{think}}$. Samples without reference reasoning receive $r_{\text{think}} = 0$, allowing graceful degradation to pure accuracy-based training. The sentence-level decomposition offers several advantages: it maps directly to established dimensions of clinical reasoning quality (Faithfulness and Completeness), provides natural verbosity control and is robust to stylistic variation in how clinical reasoning is expressed.

During GRPO training, the model generates multiple candidate answers per query and receives the above reward for each; the group-relative advantage formulation then guides the model to preferentially produce answers that maximize these rewards. Notably, this process does not require a human preference model or manual rubric; the reward for each output is computed automatically against the known correct answer or annotation, making the reinforcement learning procedure stable and reproducible. We ran the GRPO fine-tuning for only a small number of iterations on the selected subset of data (again emphasizing efficiency). Even with this limited exposure, the impact on EVLF-FM was notable: the model learned to avoid minor errors that it sometimes made after the SFT stage (for example, cases where it would output an almost-correct diagnosis but with slightly off terminology or omit a required detail), and it became more consistent in producing well-formatted, complete answers. This translates to improved predictive accuracy—for instance, we observed higher exact-match accuracy in diagnosis and better IoU scores in localization after reinforcement learning fine-tuning—as well as enhanced clinical interpretability of the model's outputs. The interpretability gains are especially evident in the detection task, where the model (when asked) can now pinpoint the finding on the image with greater precision and appropriately calibrated confidence, essentially providing a visual explanation for its diagnosis. We also observed that the model's open-ended explanations (when it provides reasoning or descriptions) became more aligned with clinical reasoning after reinforcement learning training, likely because the model learned to focus on the key image features that maximize the reward (which, by design, correspond to the ground truth findings). This outcome is in line with reports in other domains that policy optimization on a small subset of high-quality data can yield outsized improvements. By utilizing GRPO on the subset of our training data, we efficiently fine-tuned EVLF-FM to not only be more correct in its answers but also to present those answers in a manner that is trusted and useful to clinicians—a crucial aspect for real-world deployment.

In summary, our method blends architectural innovations with a comprehensive training regimen to produce a model with strong clinical reasoning abilities. The dual-encoder design provides EVLF-FM with a rich visual foundation: the disease-level encoder ensures broad diagnostic awareness of the image, and the pixel-level encoder provides granular visual evidence. The supervised training approach establishes the model's medical knowledge and basic vision–language alignment, with LoRA enabling effective adaptation on top of a powerful FM without overfitting and GPT-augmented rationales seeding the model with improved explainability. Finally, GRPO-based reinforcement learning directly optimizes the model's behavior toward correct and explainable outputs, honing the model's focus on decision-critical details. Each step is deliberately designed to bridge the gap between raw medical images and the expert reasoning of a physician. The result is a multimodal FM—EVLF-FM—that achieves high performance on diagnostic

tasks while also demonstrating interpretable reasoning and alignment with clinical norms. In practice, EVLF-FM has the potential to serve as a reliable AI assistant in medicine, capable of both answering 'What is it?' with accuracy and explaining 'Why?' in human-understandable terms, thereby maximizing its real-world utility for clinical decision support.

Reasoning data curation pipeline

To construct a high-quality CoT training dataset for medical visual reasoning, we designed a multi-stage data curation pipeline that systematically generates, validates and refines visual explanations for disease diagnosis. The pipeline comprises five stages: (1) CoT generation, (2) semantic uncertainty filtering, (3) GPT-based quality scoring, (4) CoT refinement with rescoring and (5) final quality filtering.

Dataset selection. We curated training data from 26 publicly available medical image classification datasets spanning seven imaging modalities: fundus photography (6), OCT (2), chest XR (5), CT (3), histopathology (4), dermoscopy (3) and US (3). These datasets are drawn from the training splits of established medical imaging benchmarks, including widely used public datasets such as AIROGS, EyePACS, HAM10000, RSNA Pneumonia and LC25000. This selection ensures broad coverage across anatomical regions, imaging protocols and disease categories while maintaining strict separation from the evaluation (test) splits.

Stage 1: CoT generation. For each training sample consisting of a medical image, a task-specific prompt and a ground truth label, we generated a CoT description that articulates the visual evidence supporting the diagnosis. We employed MedGemma-1.5 as the primary generator. Given an image and its ground truth label, the model was prompted to describe only the observable visual findings that support the answer, producing a structured response within <thinking> tags. For OCT images, where MedGemma's performance is limited, we used GPT-5-mini via the OpenAI Batch API as an alternative generator.

Stage 2: Semantic uncertainty filtering. To filter out unreliable CoT descriptions where the model lacks genuine visual understanding, we adapted vision-amplified semantic entropy (VASE), a hallucination detection method based on semantic entropy[85,86]. For each sample, we performed $M = 10$ stochastic inference passes using MedGemma: five passes with weak visual transformations applied to the original image (random cropping, rotation and brightness/contrast adjustment) and five passes with the image additionally distorted by Gaussian and Poisson noise. The sampled responses were clustered into semantic equivalence classes using bidirectional entailment assessed by a DeBERTa-Large-MNLI model[87]. The probability of each semantic class was estimated by aggregating the generation probabilities of its constituent responses. After VASE[85], we computed a contrastive semantic predictive distribution that amplifies the influence of visual input by contrasting the distributions derived from clean and distorted images[85]:

$$P_{\text{con}}\left(c \mid x_v, x_q\right) = \sigma\left((1+\alpha)P\left(c \mid x_v, x_q\right) - \alpha P\left(c \mid x_v', x_q\right)\right)$$

where α is the vision-amplified ratio. The VASE score is then defined as the entropy of this contrastive distribution:

$$\text{VASE}\left(x_v, x_q\right) = -\sum_{i=1}^{N} P_{\text{con}}\left(c_i | x_v, x_q\right) \log P_{\text{con}}\left(c_i | x_v, x_q\right)$$

A high VASE score indicates that the model's response varies significantly with visual perturbations, suggesting hallucination or insufficient visual grounding. We retained only samples with VASE < 0.5.

Stage 3: GPT-based quality scoring. Samples passing the VASE filter were submitted to GPT-5-mini for multi-dimensional quality assessment. Each sample's CoT was evaluated along four axes: (1) format correctness (proper tag usage), (2) grounding fidelity (descriptions limited to visible findings without hallucination), (3) relevance to the question and answer and (4) overall quality as a training sample. Each dimension was scored from 1 (very poor) to 5 (excellent). We retained samples with an overall score ≥3 for refinement.

Stage 4: CoT refinement and rescoring. For samples that passed the quality threshold, we performed a refinement step to further improve CoT quality. GPT-5-mini was prompted with the original CoT, its quality assessment and the identified issues and asked to produce a revised description that preserves accurate observations while correcting any deficiencies. The refined CoTs were then re-scored using the same multi-dimensional rubric to ensure that refinement did not degrade quality.

Stage 5: Final filtering. In the final stage, we applied a re-score threshold of ≥4 on the overall dimension to produce the curated training set.

Pipeline statistics. The pipeline processed 110,290 initial samples. After VASE filtering (<0.5), 78,049 samples remained (70.8% pass rate). After GPT quality scoring (≥3), 61,531 samples survived. After refinement and rescoring (≥4), the final curated dataset contained 60,455 high-quality CoT training samples spanning 26 datasets across seven imaging modalities, yielding an end-to-end retention rate of 55%. The VASE stage provided the most aggressive filtering, indicating that approximately 30% of initially generated CoTs contained hallucinated or insufficiently grounded content.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

This study used a combination of publicly available, controlled-access and access-restricted clinical datasets for the development and evaluation of EVLF-FM. Publicly available datasets were obtained from their original repositories; controlled-access datasets were obtained through the corresponding data access mechanisms and data use agreements. A complete list of all datasets, including dataset description and accession information, is provided in the Reporting Summary and the Supplementary Information. Access-restricted clinical datasets, including RAPIER, AIMx and SEED, are not publicly available because they contain identifiable clinical data and are subject to institutional review board approvals, patient privacy regulations and institutional data governance policies. Deidentified data from these datasets may be made available for non-commercial academic research from the corresponding author (daniel.ting@duke-nus.edu.sg) upon reasonable request, subject to approval by the relevant institutional data management committees, applicable ethics requirements and execution of a data transfer or data use agreement. Access to the DRCR Retina Network imaging data is administered by the DRCR Retina Network and the Jaeb Center for Health Research and should be requested directly through their established data access procedures. Source data are provided with this paper.

Code availability

The code repository can be found at https://github.com/baiyang4/EVLF. The code is also archived on Zenodo at https://doi.org/10.5281/zenodo.21772695.

References

85. Wang, C., Zhou, X., Fu, W. & Zhou, Y. Mitigating hallucinations in large vision-language models with internal fact-based contrastive decoding. Preprint at https://doi.org/10.48550/arXiv.2502.01056 (2025).

86. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
87. He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: decoding-enhanced BERT with disentangled attention. Preprint at https://doi.org/10.48550/arXiv.2006.03654 (2020).
88. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).

Acknowledgements

This study was approved by the SingHealth Centralised Institutional Review Board (CIRB; reference no. 2025-0836).

Author contributions

Y.B. and Y.Z. designed the methodology, algorithms and computational framework, curated the data and developed, trained and implemented the model. Y.B. and H.C. performed the formal analyses and validation and prepared the visualizations. Y.B. and H.C. wrote the original draft, with contributions from A.T., Y.K., L.G., Z.L.T., D.S.J.T., B.T.S. and C.S.N. A.T., J.Y., A.S.Y.H., C.W.N.Q., A.B., K.F., R.K., C.L.C., L.T.-E.C. and N.T.A.T. evaluated the model's reasoning outputs. A.T., Y.K., Z.L.T., D.S.J.T., B.T.S., C.S.N., N.R., T.Y.W., D.M.M., Y.L. and D.S.W.T. reviewed and edited the manuscript. D.S.W.T. and Y.L. conceptualized the study, administered the project, supervised the research, provided resources and acquired funding. All authors reviewed and approved the final manuscript.

Funding

This study is supported by grants from the National Medical Research Council, Singapore (MOH-001689-00, MOH-000655-00 and MOH-001014-00); Duke-NUS Medical School (Duke-NUS/RSF/2021/0018, 05/FY2020/EX/15-A58 and 05/FY2022/EX/66-A128); the Agency for Science, Technology and Research, Singapore (A20H4g2141 and H20C6a0032); Research to Prevent Blindness; and National Institutes of Health P30 EY003790. A.J.T. is supported by National Institute for Health & Care Research (ACF-2025-20-001). Y.Z. is supported by the National Natural Science Foundation of China (62506270). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

D.S.W.T. holds a patent on a deep learning system for the detection of retinal diseases (10201706186V) and a computer-implemented method for training an image classifier using weakly annotated training data (10201901083Y) and stocks at EyRIS, Singapore and aISIGHT. A.J.T. has received funding from HealthSense to support research concerning biomedical applications of large language models. All other authors declare no conflicts of interest.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-026-04655-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-026-04655-5.

Correspondence and requests for materials should be addressed to Daniel Shu Wei Ting.

Peer review information *Nature Medicine* thanks Jansen Seheult and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Mattia Andreoletti and Joao Monteiro, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

Extended Data Table 1 | Detailed results on the performance EVLF-FM for medical image disease diagnostics in internal validation against generalist and specialist models in internal validation

	RSNA Pneumonia	BCN20000	MIDOG25	LC25000	BUSI	Fetal_Planes	PAPILA	RAPIER	OAEFD-CFP	OAEFD-OCT	MedMNIST	Mean
Generalist VLM												
LLaVA v1.5 13B	0.693/0.668	0.638/0.345	0.773/0.418	0.971/0.971	0.624/0.487	0.841/0.823	0.643/0.417	0.685/0.371	0.968/0.963	0.975/0.942	0.759/0.624	0.779/0.639
Qwen2.5 VL 7B	0.580/0.570	0.584/0.377	0.748/0.392	0.967/0.967	0.598/0.533	0.827/0.765	0.673/0.431	0.520/0.254	0.955/0.947	0.975/0.940	0.589/0.469	0.729/0.604
LLava-Med V1.5 7B	0.723/0.698	0.705/0.428	0.791/0.430	0.968/0.968	0.598/0.456	0.815/0.793	0.663/0.426	0.742/0.447	0.961/0.956	0.970/0.938	0.743/0.652	0.789/0.654
InternVL 3 8B	0.678/0.655	0.660/0.378	0.810/0.442	0.977/0.977	0.650/0.523	0.854/0.838	0.684/0.435	0.704/0.403	0.977/0.974	0.980/0.953	0.766/0.648	0.795/0.657
EVLF-FM	0.757/0.720	0.949/0.822	0.865/0.732	0.996/0.996	0.932/0.926	0.908/0.898	0.786/0.687	0.826/0.606	0.981/0.978	0.990/0.977	0.851/0.808	0.895/0.832
Specialist DL												
VGG16	0.830/0.763	0.758/0.607	0.873/0.780	0.998/0.998	0.795/0.778	0.890/0.886	0.571/0.328	0.726/0.520	0.977/0.974	0.975/0.941	0.767/0.660	0.833/0.749
ResNet50	0.841/0.762	0.808/0.690	0.889/0.813	1.000/1.000	0.838/0.824	0.910/0.909	0.694/0.349	0.800/0.640	0.984/0.982	0.956/0.898	0.455/0.275	0.834/0.740
DenseNet121	0.825/0.743	0.817/0.726	0.872/0.779	1.000/1.000	0.829/0.820	0.922/0.920	0.694/0.572	0.784/0.664	0.974/0.970	0.966/0.915	0.707/0.546	0.854/0.787
EfficientNet-B0	0.827/0.752	0.838/0.752	0.862/0.778	0.999/0.999	0.897/0.889	0.920/0.916	0.704/0.431	0.771/0.604	0.984/0.984	0.970/0.928	0.548/0.354	0.847/0.762
ViT-B/16	0.840/0.742	0.779/0.667	0.884/0.789	0.997/0.997	0.829/0.821	0.892/0.902	0.694/0.426	0.724/0.555	0.974/0.971	0.951/0.874	0.807/0.712	0.852/0.769
CLIP	0.823/0.718	0.651/0.427	0.837/0.542	0.989/0.989	0.650/0.608	0.683/0.663	0.133/0.102	0.543/0.298	0.899/0.895	0.916/0.743	0.696/0.531	0.711/0.592
DINOv2	0.804/0.719	0.634/0.369	0.820/0.477	0.988/0.988	0.701/0.666	0.643/0.643	0.429/0.212	0.599/0.244	0.974/0.975	0.931/0.846	0.813/0.730	0.758/0.624
SAM	0.835/0.750	0.689/0.455	0.869/0.742	0.994/0.994	0.821/0.808	0.731/0.801	0.225/0.142	0.732/0.386	0.961/0.959	0.946/0.865	0.721/0.555	0.775/0.678
Specialist FM												
Virchow	0.845/0.769	0.886/0.836	0.895/0.812	1.000/1.000	0.855/0.848	0.947/0.942	0.755/0.722	0.867/0.759	0.997/0.996	0.975/0.942	0.800/0.743	0.893/0.852
RETFound-CFP	0.842/0.756	0.756/0.624	0.848/0.697	0.996/0.996	0.761/0.726	0.911/0.897	0.796/0.738	0.724/0.497	0.987/0.985	0.970/0.929	0.762/0.683	0.850/0.775
RETFound-OCT	0.830/0.731	0.640/0.327	0.827/0.453	0.985/0.985	0.761/0.729	0.915/0.905	0.684/0.406	0.774/0.538	0.981/0.978	0.985/0.965	0.739/0.604	0.829/0.693
Rad-DINO	0.830/0.750	0.789/0.605	0.890/0.790	0.999/0.999	0.872/0.858	0.943/0.937	0.776/0.731	0.817/0.625	0.984/0.981	0.980/0.954	0.808/0.744	0.881/0.816
USFM	0.840/0.752	0.709/0.505	0.875/0.742	0.998/0.998	0.838/0.823	0.942/0.936	0.684/0.406	0.776/0.451	0.974/0.971	0.970/0.932	0.784/0.704	0.854/0.747
BiomedCLIP	0.845/0.759	0.809/0.731	0.896/0.817	1.000/1.000	0.829/0.812	0.948/0.943	0.796/0.750	0.830/0.595	0.984/0.982	0.980/0.954	0.796/0.727	0.883/0.824

Results are reported as accuracy/F1-score. Highest score for each evaluation matrix is highlighted in bold (including ties).

https://doi.org/10.1038/s41591-026-04655-5

Extended Data Table 2 | Detailed results on the performance EVLF-FM for medical image disease diagnostics in internal validation against generalist and specialist models in internal validation in the MedMNIST datasets

	Path	Chest	Derma	OCT	Pneumonia	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	Mean
Generalist VLM													
LLaVA v1.5 13B	0.935/0.905	**0.535**/0.073	0.731/0.355	0.788/0.786	0.881/0.864	0.557/0.279	0.750/0.671	0.951/0.832	0.613/0.497	0.878/0.855	0.796/0.750	0.689/0.621	0.759/0.624
Qwen2.5 VL 7B	0.823/0.754	0.510/0.051	0.716/0.384	0.738/0.729	0.438/0.383	0.280/0.166	0.494/0.510	0.286/0.166	0.575/0.411	0.807/0.777	0.724/0.681	0.672/0.618	0.589/0.469
LLava-Med V1.5 7B	0.939/0.915	0.513/0.088	0.800/0.556	0.868/0.868	0.910/0.900	0.542/0.280	0.212/0.382	0.975/0.856	0.642/0.540	0.916/0.908	0.865/0.843	0.738/0.687	0.743/0.652
InternVL 8B	0.938/0.911	0.520/0.045	0.755/0.487	0.798/0.788	0.886/0.872	0.575/0.394	0.788/0.699	0.960/0.841	0.594/0.465	0.865/0.838	0.817/0.783	0.698/0.650	0.766/0.648
EVLF-FM	**0.958/0.944**	0.369/0.313	**0.974/0.955**	**0.966/0.966**	0.971/0.969	**0.723**/0.473	**0.929/0.905**	0.981/0.983	0.693/0.618	0.944/0.937	0.900/0.881	0.798/**0.755**	**0.851/0.808**
Specialist DL													
vgg16	0.936/0.918	0.533/0.054	0.727/0.452	0.851/0.851	0.782/0.722	0.560/0.409	0.840/0.767	0.959/0.956	0.610/0.495	0.916/0.901	0.793/0.759	0.695/0.631	0.767/0.660
resnet50	0.911/0.890	0.532/0.000	0.491/0.128	0.714/0.696	0.625/0.385	0.435/0.121	0.333/0.317	0.165/0.135	0.456/0.215	0.402/0.272	0.220/0.078	0.177/0.064	0.455/0.275
densenet	0.954/0.937	0.533/0.018	0.662/0.211	0.915/0.916	0.729/0.637	0.443/0.139	0.737/0.504	0.962/0.955	0.643/0.541	0.882/0.869	0.567/0.493	0.456/0.337	0.707/0.546
efficientnet	0.886/0.830	0.532/0.000	0.648/0.150	0.736/0.726	0.627/0.389	0.435/0.121	0.609/0.549	0.351/0.306	0.489/0.284	0.675/0.618	0.324/0.185	0.258/0.085	0.548/0.354
VIT	0.950/0.939	0.530/0.097	0.783/0.552	0.941/0.942	0.942/0.937	0.510/0.274	0.833/0.770	**0.987/0.988**	0.673/0.583	0.952/0.950	0.851/0.835	0.732/0.674	0.807/0.712
CLIP	0.940/0.917	0.532/0.021	0.661/0.264	0.935/0.935	0.889/0.878	0.435/0.121	0.731/0.422	0.926/0.908	0.651/0.551	0.651/0.575	0.551/0.471	0.448/0.307	0.696/0.531
DINO	0.947/0.927	0.532/0.115	0.813/0.632	0.915/0.916	0.954/0.949	0.570/0.393	0.859/0.809	**0.987**/0.985	0.678/0.593	0.931/0.928	0.865/0.857	0.703/0.652	0.813/0.730
SAM	0.957/0.937	0.532/0.003	0.647/0.175	0.890/0.890	0.888/0.880	0.450/0.154	0.731/0.422	0.961/0.956	0.619/0.486	0.886/0.868	0.586/0.504	0.501/0.385	0.721/0.555
Specialist FM													
Virchow	0.933/0.918	0.404/0.310	0.838/0.675	0.930/0.929	0.872/0.851	0.640/0.439	0.609/0.605	0.981/0.980	**0.719/0.634**	**0.955/0.954**	**0.918/0.908**	**0.800**/0.719	0.800/0.743
RETFound-CFP	0.921/0.911	0.349/0.224	0.782/0.542	0.767/0.744	0.926/0.920	0.642/**0.504**	0.782/0.595	0.965/0.965	0.650/0.535	0.898/0.895	0.806/0.806	0.655/0.561	0.762/0.683
RETFound-OCT	0.928/0.905	0.379/0.174	0.706/0.215	0.933/0.933	0.883/0.871	0.435/0.121	0.731/0.422	0.930/0.918	0.632/0.512	0.863/0.868	0.818/0.795	0.626/0.513	0.739/0.604
Rad-DINO	0.914/0.884	0.415/**0.366**	0.776/0.525	0.912/0.913	**0.981/0.979**	0.593/0.395	0.859/0.793	0.959/0.958	0.690/0.603	0.937/0.935	0.898/0.881	0.763/0.696	0.808/0.744
USFM	0.940/0.916	0.384/0.236	0.753/0.436	0.889/0.891	0.933/0.926	0.517/0.295	0.846/0.779	0.973/0.973	0.672/0.579	0.900/0.896	0.872/0.859	0.731/0.664	0.784/0.704
BiomedCLIP	**0.958**/0.942	0.400/0.256	0.800/0.535	0.873/0.874	0.962/0.958	0.590/0.414	0.872/0.816	0.980/0.980	0.660/0.567	0.893/0.898	0.890/0.876	0.679/0.609	0.796/0.727

Results are reported as accuracy/F1-score. Highest score for each evaluation matrix is highlighted in bold (including ties).

https://doi.org/10.1038/s41591-026-04655-5

https://doi.org/10.1038/s41591-026-04655-5

Extended Data Table 3 | Per-class comparison on chest X-ray classification across internal (NIH ChestX-ray14, VinDr-CXR) and external (AIMx) evaluation sets

	Model	Cons./Pneu.	Pulm. Mass	Pneumothorax	No Finding	Mean
NIH ChestX-ray14 (Internal)	InternVL	**0.690**/0.340	0.860/0.130	0.870/0.104	0.598/**0.700**	**0.754**/0.318
	Medgemma 27B	0.530/0.480	0.810/0.280	0.815/0.420	0.585/0.480	0.685/0.415
	EVLF-FM	0.634/**0.571**	**0.877**/**0.508**	**0.872**/**0.570**	**0.622**/0.533	0.751/**0.545**
VinDr-CXR (Internal)	InternVL	**0.908**/0.300	0.920/0.082	0.975/0.000	0.870/0.920	0.918/0.325
	Medgemma 27B	0.815/0.360	0.905/0.320	0.970/0.380	0.830/0.875	0.880/0.484
	EVLF-FM	0.886/**0.496**	**0.937**/**0.523**	**0.995**/**0.563**	**0.892**/**0.936**	**0.928**/**0.629**
AIMx (External)	InternVL	0.640/0.290	0.780/0.080	0.910/0.000	0.575/0.680	0.726/0.263
	Medgemma 27B	0.680/0.595	0.820/0.210	0.945/0.440	0.660/0.655	0.776/0.475
	EVLF-FM	**0.759**/**0.679**	**0.865**/**0.529**	**0.987**/**0.709**	**0.750**/**0.757**	**0.840**/**0.668**

Results reported as Accuracy / Macro-F1 on four representative classes. Cons./Pneu. = Consolidation/Pneumonia; Pulm. = Pulmonary. Highest score is highlighted in bold.

https://doi.org/10.1038/s41591-026-04655-5

Extended Data Table 4 | Detailed results on the performance of EVLF-FM for visual question-answering (VQA) tasks in internal validation against other VLMs

	VQA-RAD		SLAKE		Path-VQA	
	Closed Accuracy	Open recall	Closed Accuracy	Open recall	Closed Accuracy	Open recall
RadFM	65.3	36.8	74.1	77.3	48.9	2.5
LLava-Med	22.3	63.4	38.9	81.6	56	**37.9**
Med-Flamingo	47.4	25.7	46.5	27.2	57.4	6.6
MedDr 32B	78.9	37.5	83.4	74.2	90.2	33.5
InternVL 3 8B	71.3	52.9	81.4	77.5	90.1	15.7
EVLF-FM	**89.2**	**82.5**	**89.6**	**87.6**	**91.3**	21.1

Results are reported as closed accuracy and open recall. Highest score for each evaluation matrix is highlighted in bold (including ties).

https://doi.org/10.1038/s41591-026-04655-5

Extended Data Table 5 | Detailed results on the performance of EVLF-FM for medical visual grounding tasks in internal validation against other VLMs

		MRI	Path	XR	CT	OCT	CFP	US	Dermatoscope	Endoscope	Mean
EVLF-FM	Acc@0.1	0.987	0.797	0.985	0.924	0.951	1.000	0.998	1.000	0.871	0.946
	Acc@0.3	0.952	0.708	0.972	0.844	0.855	1.000	0.991	1.000	0.793	0.902
	Acc@0.5	0.870	0.618	0.955	0.737	0.714	0.968	0.975	1.000	0.742	0.842
	mIOU	0.747	0.595	0.922	0.667	0.603	0.781	0.827	0.869	0.660	0.741
InternVL 3 8B	Acc@0.1	0.908	0.712	0.904	0.839	0.878	0.92	0.917	0.92	0.788	0.865
	Acc@0.3	0.871	0.618	0.893	0.766	0.775	0.918	0.907	0.92	0.713	0.82
	Acc@0.5	0.791	0.533	0.884	0.663	0.644	0.86	0.89	0.905	0.625	0.755
	mIOU	0.67	0.506	0.846	0.589	0.523	0.682	0.746	0.783	0.564	0.656

Results are reported as Acc@0.1, Acc@0.3 and Acc@0.5 (proportion of predictions with intersection-over-union values exceeding thresholds of 0.1, 0.3 and 0.5, respectively) and mean intersection-over-union (mIOU).

Extended Data Table 6 | Detailed results on the performance of EVLF-FM for medical image disease diagnostics and medical visual grounding in zero and few shot external validation against generalist VLMs

Diagnostic	CT		XR	Pathology	US		OCT		CFP			
	HRCTCov19	SARS-COV-2	CXR-Pneumonia	AMI-Br	BUS-UCLM	US3M	RetinalOCT	DRCR-OCT	DRCR-CFP	SEED	APTOS	Glaucoma fundus
Zero Shot												
InternVL 3 8B	0.862/0.850	0.708/0.691	0.885/0.875	0.399/**0.391**	0.204/0.156	0.417/0.295	0.246/0.215	0.839/0.838	0.385/0.318	0.889/0.743	0.536/0.216	0.557/0.530
HuatuoGPT-Vision 7B	0.941/0.934	0.518/0.387	0.870/0.865	0.253/0.080	0.241/0.089	0.222/0.107	0.108/0.005	0.756/0.743	0.775/0.240	0.267/0.130	0.578/0.296	0.462/0.130
MedDr 32B	0.443/0.177	**0.781**/0.431	0.400/0.100	0.280/0.050	0.200/0.080	0.574/0.116	0.374/0.118	0.304/0.122	0.350/0.120	0.400/0.150	0.570/0.070	0.688/0.156
MedGemma 4B	0.835/0.799	0.676/0.427	0.601/0.396	0.352/0.351	0.642/0.397	0.339/0.081	0.143/0.058	0.695/0.670	0.640/0.035	0.822/0.054	0.676/0.445	0.527/0.413
MedGemma 27B	0.874/0.854	0.766/0.759	0.846/0.838	0.225/0.184	**0.679/0.519**	**0.613**/0.501	0.157/0.101	0.530/0.366	0.859/0.677	0.702/0.609	**0.759**/0.440	0.486/0.485
EVLF-FM	**0.972/0.969**	0.780/**0.768**	**0.894/0.892**	**0.624**/0.282	0.650/0.200	**0.613/0.591**	**0.503/0.379**	**0.861/0.861**	**0.952/0.770**	**0.912/0.823**	0.756/**0.558**	**0.746/0.743**
Few Shot												
EVLF-FM 8 Shot	0.985/0.983	0.809/0.806	0.886/0.884	0.735/0.296	0.730/0.511	0.587/0.582	0.549/0.445	0.834/0.834	**0.950/0.771**	0.900/0.803	0.738/0.581	0.746/0.737
EVLF-FM 16 Shot	0.986/0.985	0.790/0.782	0.891/0.889	0.761/0.303	0.774/0.725	0.587/0.582	0.616/0.541	**0.864/0.863**	0.941/0.747	0.915/0.825	0.734/0.575	0.768/0.768
EVLF-FM 32 Shot	0.981/0.980	0.821/0.816	0.918/0.916	**0.767**/0.297	0.796/0.748	0.596/0.589	0.671/0.608	0.855/0.853	0.942/0.751	0.921/0.835	0.759/0.605	0.748/0.744
EVLF-FM 64 Shot	**0.995/0.994**	**0.823/0.819**	**0.966/0.964**	0.747/**0.691**	**0.818/0.780**	**0.691/0.666**	**0.797/0.752**	0.862/0.862	0.944/0.754	**0.935/0.858**	**0.781/0.620**	**0.776/0.776**
Grounding	**CT**	**MRI**	**XR**	**PET**	**Endoscope**	**Dermatoscope**	**CFP**	**US**	**Mean**			
BiRD (SOTA)	**0.445**	0.293	**0.417**	**0.565**	**0.536**	0.756	0.842	0.46	0.539			
MedGround Zero Shot	0.204	0.259	0.228	0.231	0.356	0.364	0.677	0.218	0.317			
MedGround Few Shot	0.225	0.312	0.28	0.273	0.398	0.421	0.707	0.266	0.36			
EVLF-FM Zero Shot	0.236	0.299	0.308	0.247	0.442	0.312	0.855	0.178	0.36			
EVLF-FM Few Shot	0.282	**0.332**	0.363	0.531	0.51	**0.882**	**0.928**	**0.53**	**0.545**			

Results are reported as accuracy/F1-score for diagnostics and acc@0.5 for medical visual grounding. Highest score for each evaluation matrix is highlighted in bold (including ties).

https://doi.org/10.1038/s41591-026-04655-5

https://doi.org/10.1038/s41591-026-04655-5

Extended Data Table 7 | Detailed results of the human evaluation of EVLF-FM thinking outputs using the S.C.O.R.E. framework

Modality	S	C	O	R	E	Mean	Gwet's AC2	%Withthin -1
CT	4.28 (0.99)	4.23 (1.01)	4.28 (1.01)	4.26 (1.03)	3.94 (1.09)	4.20 (0.98)	0.72 (0.66, 0.78)	75
CFP	4.14 (1.09)	4.14 ((1.20)	4.51 (0.87)	4.15 (1.24)	3.98 (1.16)	4.19 (0.98)	0.69 (0.60, 0.77)	74
OCT	4.78 (0.50)	4.69 (0.68)	4.78 (0.54)	4.79 (0.57)	4.59 (0.70)	4.73 (0.50)	0.93 (0.91, 0.95)	91
Path	3.96 (1.22)	4.09 (1.02)	4.03 (1.08)	4.09 (1.09)	4.07 (1.12)	4.05 (1.07)	0.5 (0.45, 0.55)	42
Xray	4.10 (0.95)	4.12 (0.98)	4.16 (0.93)	4.11 (1.00)	3.95 (1.14)	4.09 (0.93)	0.7 (0.65, 0.76)	76
Mean	4.26 (1.00)	4.26 (1.01)	4.36 (0.93)	4.29 (1.04)	4.10 (1.08)	4.25 (0.94)		

S.C.O.R.E. domain scores and overall scores are presented as mean (standard deviation). Gwet's AC2 is presented as the mean agreement coefficient (95% confidence interval), and %Within ±1 represents the proportion of ratings differing by no more than one point between evaluators.

https://doi.org/10.1038/s41591-026-04655-5

Extended Data Table 8 | Detailed results on the performance of EVLF-FM for medical image disease diagnosis stratified by age and sex, compared with generalist VLMs

Subgroup	HAM10000			RAPIER			PAPILA			AMIx		
	InternVL 3 8B	Qwen3 VL 8B	EVLF-FM	InternVL 3 8B	Qwen3 VL 8B	EVLF-FM	InternVL 3 8B	Qwen3 VL 8B	EVLF-FM	InternVL 3 8B	Qwen3 VL 8B	EVLF-FM
Overall	0.832	0.845	**0.985**	0.704	0.52	**0.826**	0.684	0.673	**0.786**	0.79	0.797	**0.836**
Male	0.827	0.836	**0.988**	0.724	0.533	**0.814**	0.632	0.605	**0.684**	0.78	0.784	**0.838**
Female	0.837	0.854	**0.982**	0.692	0.533	**0.849**	0.717	0.717	**0.85**	0.802	0.812	**0.833**
Difference by sex	0.01	0.018	**0.006**	0.031	**0.001**	0.035	0.085	0.111	0.166	0.022	0.028	**0.005**
Age 0-60	0.894	0.896	**0.99**	0.678	0.537	**0.842**	0.686	0.686	**0.743**	0.837	0.845	**0.876**
Age >60	0.625	0.674	**0.972**	0.727	0.531	**0.824**	0.683	0.667	**0.81**	0.766	0.772	**0.815**
Difference by age	0.27	0.223	**0.018**	0.048	0.006	0.018	**0.003**	0.019	0.067	0.071	0.072	**0.062**

Results are reported as accuracy. The highest accuracy for each subgroup and the smallest performance difference between subgroups are highlighted in bold.

Extended Data Table 9 | Detailed results on ablation study of EVLF-FM performance on training strategy components and varying proportions of RFT data

	Path	Chest	Derma	OCT	Pneumonia	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	Mean
Base	0.938/0.911	0.520/0.045	0.755/0.487	0.798/0.788	0.886/0.872	0.575/0.394	0.788/0.699	0.960/0.841	0.594/0.465	0.865/0.838	0.817/0.783	0.698/0.650	0.766/0.648
Stage 1													
Base+CoT	0.947/0.923	0.507/0.121	0.812/0.625	0.853/0.852	0.950/0.947	0.667/0.446	0.808/0.756	0.975/0.856	0.632/0.532	0.921/0.911	0.890/0.866	0.773/0.726	0.811/0.713
Base + CoT+ PE	0.949/0.926	0.511/0.121	0.801/0.603	0.852/0.851	0.946/0.941	0.640/0.392	0.853/0.808	0.979/0.860	0.629/0.527	0.925/0.915	0.894/0.872	0.766/0.726	0.812/0.712
Base + CoT +PE+ DE (Full SFT)	0.963/0.950	0.521/0.148	0.935/0.872	0.942/0.942	0.955/0.950	0.628/0.462	0.878/0.856	0.984/0.863	0.743/0.665	0.933/0.928	0.885/0.868	0.778/0.728	0.845/0.769
Stage 2													
Base + GRPO	0.891/0.845	**0.532**/0.038	0.648/0.275	0.712/0.705	0.803/0.775	0.425/0.178	0.545/0.432	0.942/0.815	0.528/0.372	0.785/0.752	0.691/0.642	0.612/0.538	0.676/0.531
Base + Full SFT + GRPO (r_acc)	**0.969**/**0.960**	0.490/0.253	0.944/0.893	0.951/0.951	0.962/0.958	0.662/**0.506**	**0.949**/**0.936**	**0.988**/0.868	**0.747**/**0.675**	**0.949**/**0.945**	0.896/0.879	0.790/0.738	**0.858**/0.797
EVLF-FM (Base + Full SFT +GRPO + r_think)	0.958/0.944	0.369/**0.313**	**0.974**/**0.955**	**0.966**/**0.966**	**0.971**/**0.969**	**0.723**/0.473	0.929/0.905	0.981/**0.983**	0.693/0.618	0.944/0.937	**0.900**/**0.881**	**0.798**/**0.755**	0.851/**0.808**
Data Efficiency													
EVLF-20	0.952/0.937	0.350/0.227	0.946/0.840	0.944/0.944	0.965/0.963	0.693/0.472	0.923/0.897	0.918/0.888	0.685/0.611	0.915/0.904	0.855/0.824	0.730/0.686	0.823/0.766
EVLF-40	0.957/0.944	0.354/0.286	0.963/0.922	0.963/0.963	0.962/0.959	0.710/0.464	**0.942**/**0.924**	0.966/0.962	0.700/0.624	0.942/0.932	0.891/0.868	0.784/0.743	0.845/0.799
EVLF-60	**0.960**/**0.947**	**0.370**/0.299	0.971/0.946	0.963/0.963	0.962/0.960	0.710/**0.487**	0.923/0.896	0.979/0.980	**0.701**/**0.628**	0.942/0.935	0.896/0.876	0.792/0.750	0.847/0.805
EVLF-80	0.958/0.944	0.367/0.311	0.972/0.951	0.962/0.962	0.966/0.964	0.718/0.467	0.929/0.905	**0.981**/**0.984**	0.690/0.618	0.942/0.935	0.898/0.878	**0.799**/**0.757**	0.849/0.806
EVLF-FM	0.958/0.944	0.369/**0.313**	**0.974**/**0.955**	**0.966**/**0.966**	**0.971**/**0.969**	**0.723**/0.473	0.929/0.905	**0.981**/0.983	0.693/0.618	**0.944**/**0.937**	**0.900**/**0.881**	0.798/0.755	**0.851**/**0.808**

Results are reported as accuracy/F1-score on internal validation (MedMNIST). Highest score for each evaluation matrix is highlighted in bold (including ties). CoT: Chain of Thought; PE: Pixel-level Encoder; DE: Disease-level Encoder; SFT: Supervised Fine-Tuning; GRPO: Group Relative Policy Optimization.

https://doi.org/10.1038/s41591-026-04655-5

Extended Data Table 10 | Detailed results on the performance of EVLF-FM for medical image disease diagnostics tasks using different prompts in internal validation (MedMNIST)

	Path	Chest	Derma	OCT	Pneumonia	Retina	Breast	Blood	Tissue	OrganA	OrganC	OrganS	Mean
Base Prompt	0.958/0.944	0.370/0.313	0.974/0.955	0.966/0.966	0.971/0.969	0.723/0.473	0.930/0.905	0.981/0.983	0.693/0.618	0.944/0.937	0.900/0.881	0.798/0.755	0.851/0.808
Variant Prompt	0.959/0.946	0.369/0.292	0.970/0.946	0.967/0.967	0.962/0.959	0.708/0.476	0.923/0.896	0.978/0.979	0.687/0.612	0.939/0.932	0.888/0.866	0.789/0.744	0.845/0.801

Results are reported as accuracy/F1-score. Base Prompt: Analyze the given {modality} image for diagnosis. Variant Prompt 1: “Please perform diagnostic analysis on the provided {modality} image for diagnosis. Variant Prompt 2: "Given a {modality} scan, determine the correct diagnosis" Variant Prompt 3: "Evaluate the following {modality} image for diagnosis" Variant Prompt 4: "Assess the {modality} image and provide a diagnosis" Variant Prompt 5: "Based on the {modality} image, identify the diagnosis" Variant Prompt 6: "For this {modality} image, specify its diagnosis" Variant Prompt 7: "Analyze the provided {modality} scan for diagnosis" Variant Prompt 8: "Interpret the {modality} image to determine its diagnosis" Variant Prompt 9: "Diagnose the given {modality} image" Variant Prompt 10: "Use the {modality} image to establish its diagnosis.

https://doi.org/10.1038/s41591-026-04655-5

nature portfolio

Corresponding author(s): Daniel Shu Wei Ting & Yong Liu

Last updated by author(s): May 18, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Scripts for dataset preparation and annotation were implemented in Python (version 3.10) using the libraries pandas (version 1.5.3), numpy (version 1.23.5), and Pillow (version 9.3.0).
Data analysis	Model development, training and evaluation were performed in Python with PyTorch, using two separate pinned environments. Stage 1 (supervised fine-tuning and inference/evaluation): Python 3.9, PyTorch 2.1.0 (CUDA 12.1), Hugging Face Transformers 4.37.2, PEFT 0.10.0 (LoRA), DeepSpeed 0.15.4, Accelerate, timm 0.9.12, FlashAttention 2.4.0, Segment Anything (MedSAM pixel encoder), with albumentations 2.0.7 and OpenCV for image processing. Stage 2 (Group Relative Policy Optimization reinforcement learning): Python 3.10, PyTorch 2.6.0, Hugging Face Transformers 4.49.0, TRL 0.16.0, PEFT 0.15.2, DeepSpeed 0.16.9, Accelerate 1.7.0, timm 1.0.15, FlashAttention 2, with a DeBERTa-v2 MNLI model served via FastAPI/uvicorn for the reasoning-alignment reward. Statistical analysis used scikit-learn (version 1.2.1); figures were produced with matplotlib (version 3.7.1) and seaborn (version 0.12.2). The custom code has been deposited in Zenodo with a citable DOI (https://doi.org/10.5281/zenodo.21772695) and cited in the reference list, with the corresponding GitHub repository (https://github.com/baiyang4/EVLF) also provided.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

EVLF-FM was trained on a large multi-task corpus comprising 1,375,034 samples across ten imaging modalities related to six clinical specialties (dermatology, hepatology, ophthalmology, pathology, pulmonology, and radiology), supplemented by 60,455 chain-of-thought samples.
Disease classification (751,426 samples): MedMNIST (PathMNIST, ChestMNIST, DermaMNIST, OCTMNIST, PneumoniaMNIST, RetinaMNIST, BreastMNIST, BloodMNIST, TissueMNIST, OrganAMNIST, OrganCMNIST, OrganSMNIST; https://medmnist.com), HAM10000 (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T), BCN20000 (https://api.isic-archive.com/collections/249/), MIDOG25 (https://midog.deepmicroscopy.org/), LC25000 (https://github.com/tampapath/lung_colon_image_set), RSNA Pneumonia (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge), NIH ChestX-ray14 (https://nihcc.app.box.com/v/ChestXray-NIHCC), VinDr-CXR (https://vindr.ai/datasets/cxr), BUSI (https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset), Fetal Planes (https://zenodo.org/records/3904280), PAPILA (https://doi.org/10.6084/m9.figshare.14798004.v1), ODIR-5K (OAEFD-CFP and OAEFD-OCT; https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k), BRSET (https://physionet.org/content/brazilian-ophthalmological/), AIROGS (https://airogs.grand-challenge.org/), EyePACS (https://www.kaggle.com/c/diabetic-retinopathy-detection), and RAPIER (private liver-lesion CT dataset from Singapore General Hospital and the National Cancer Centre Singapore; access on reasonable request).
Visual grounding (415,378 samples): VinDr-CXR (https://vindr.ai/datasets/cxr), BiomedParse (https://huggingface.co/datasets/microsoft/BiomedParseData), and curated multi-modal grounding samples covering CT, MRI, X-ray, ultrasound, pathology, endoscopy, fundus, OCT, mammography, and dermoscopy.
Visual Question Answering (208,230 samples): Path-VQA (https://github.com/UCSD-AI4H/PathVQA), VQA-RAD (https://osf.io/89kps), and SLAKE (https://www.med-vqa.com/slake).
External validation datasets:
CT: HRCT-Cov19 (https://doi.org/10.5281/zenodo.10252424), SARS-CoV-2 CT (https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset).
X-ray: CXR Pneumonia (https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia), and AIMx (private chest X-ray dataset from Singapore General Hospital; access on reasonable request).
Pathology: AMi-Br (https://github.com/DeepMicroscopy/AMi-Br).
Ultrasound: BUS-UCLM (https://www.kaggle.com/datasets/orvile/bus-uclm-breast-ultrasound-lesion-segmentation), US3M (https://www.kaggle.com/datasets/timesxy/multimodal-breast-ultrasound-dataset-us3m).
OCT: RetinalOCT-C8 (https://www.kaggle.com/datasets/obulisainaren/retinal-oct-c8), and DRCR_OCT_DME (DRCR.net clinical trial dataset; access via the DRCR Retina Network).
Fundus: APTOS (https://www.kaggle.com/competitions/aptos2019-blindness-detection), Glaucoma fundus (https://doi.org/10.7910/DVN/1YRRAC), DRCR_CFP_DR (DRCR.net clinical trial dataset; access via the DRCR Retina Network), and SEED (Singapore Epidemiology of Eye Diseases; private dataset, access restricted and available upon reasonable request from the corresponding author).
Visual grounding: Med-GRIT-270k (https://github.com/ShawnHuang497/BiRD).
Data availability statement: All publicly available datasets are listed above with accession links. The RAPIER, AIMx and SEED datasets are proprietary and subject to institutional ethics, patient-privacy and data-sharing restrictions; access may be requested from the corresponding author in line with institutional data-sharing policies. Access to the DRCR_CFP_DR and DRCR_OCT_DME datasets is governed by the DRCR Retina Network and may be requested from that network.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex and gender were not used as covariates during model training. Post-hoc subgroup analyses by sex were performed on datasets where such metadata were available (HAM10000, RAPIER, PAPILA, AIMx) and showed consistent performance across sex subgroups. For SEED, RAPIER, and AIMx, only de-identified data were used.
Reporting on race, ethnicity, or other socially relevant groupings	No covariates relating to race, ethnicity, or other socially relevant groupings were collected, analyzed, or reported in this study. All datasets were de-identified prior to our use.
Population characteristics	Where human participant data were included (e.g., in fundus or chest X-ray datasets), only de-identified imaging data and corresponding clinical labels were used. Demographic or socioeconomic covariates were not accessed or analyzed.
Recruitment	No new patient recruitment was performed. All datasets used were either publicly available or part of existing retrospective cohorts (SEED, RAPIER, AIMx, and DRCR.net clinical trials). For these private datasets, only retrospective, de-identified images were used.
Ethics oversight	The approving committee (SingHealth Centralised Institutional Review Board, CIRB; reference no. 2025-0836) is stated in the manuscript and in the Reporting Summary.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal sample-size calculation was performed. Instead, we included all available samples from the selected public datasets and the private RAPIER, AIMx, and SEED datasets to maximize statistical power and diversity. The final training corpus comprised 1,375,034 samples across 10 imaging modalities and 6 clinical specialties, supplemented by 60,455 chain-of-thought reasoning samples used in the second-stage reinforcement fine-tuning. External validation employed independent test samples across 13 datasets and 5 imaging modalities. This large sample size exceeds those used in comparable prior multimodal foundation model studies and was considered sufficient to support reproducible conclusions.
Data exclusions	No additional exclusions were applied beyond the inherent curation of each dataset (e.g., removal of corrupted images or incomplete metadata performed by the dataset providers). For SEED, RAPIER, and AIMx, only de-identified images meeting the quality-control standards of the source cohort or institution were included.
Replication	Replication of full model training was not performed due to the high computational cost of training large vision–language foundation models. Instead, reproducibility was supported by evaluating EVLF-FM across multiple independent datasets, diverse imaging modalities, and tasks, which consistently demonstrated stable performance.
Randomization	Randomization is not applicable, as this is a retrospective computational evaluation rather than a controlled intervention, so data were not randomly allocated into intervention groups. Data were partitioned into training and held-out internal test sets using predefined/official splits (or a fixed split with a fixed seed) with no patient/image overlap, and generalization was further assessed on independent external test datasets. No validation set was used, as training was performed for a single epoch without model selection.
Blinding	Blinding was not relevant to this study. Model training and evaluation were conducted on de-identified, digitized medical images and structured labels. No investigator blinding was required as no subjective human scoring or allocation was involved.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	Not applicable. No plant materials were used in this study.
Novel plant genotypes	Not applicable. No novel plant genotypes were generated or analyzed in this study.
Authentication	Not applicable. No plant genotypes or seed stocks were used in this study.