CentraleSupélec | université PARIS-SACLAY

FOCUS

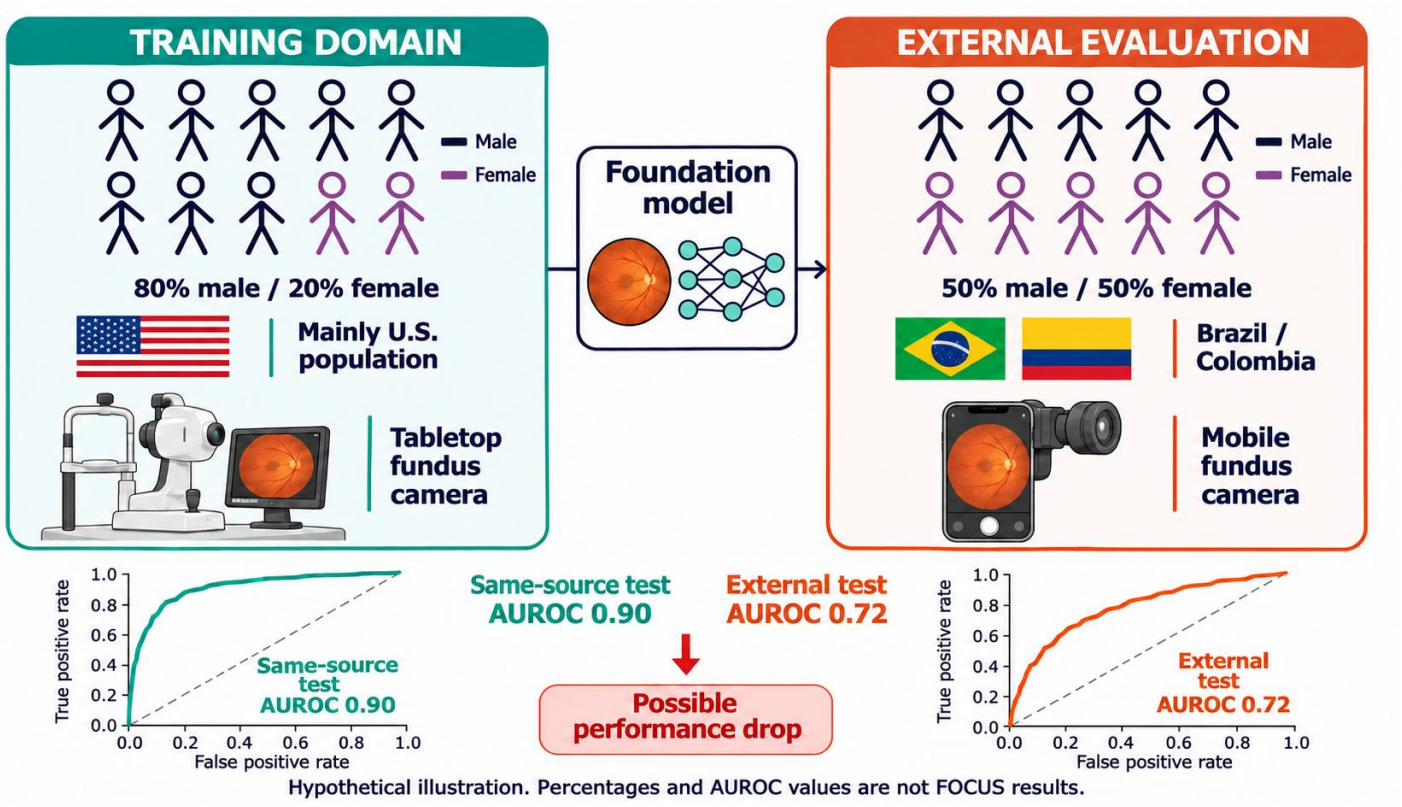
FOCUS: Benchmarking Retinal Model Generalization from Foundation Vision Encoders to Multimodal LLMs

David Restrepo, Chenwei Wu, Luis Filipe Nakayama, Miguel L. Martins,
Stergios Christodoulidis, Maria Vakalopoulou, Enzo Ferrante

INSTITUTE DATAIA
Data Science, Intelligence & Society

Co-funded by the European Union

Retina Foundation Models



A strong score on one dataset leaves important questions unanswered

Does performance transfer to another dataset?

Who experiences errors, and under which image conditions?

FOCUS Benchmark Pipeline

Foundation Ophthalmic Cross-Dataset Understanding under Shift

1 Dataset Registry & Task Harmonization

10 Complementary Fundus Datasets Across 5 Countries

- BRSET — Brazil
- mBRSET — Brazil
- PAPILA — Spain
- RFMiD — India
- RFMiD 2.0 — India
- IDRiD — India
- Messidor-2 — France
- REFUGE — China
- G1020 — Germany
- JSIEC1000 — China

Harmonized Binary Tasks

- Any Diabetic Retinopathy (DR)
- Referable Diabetic Retinopathy (Ref. DR)
- Glaucomatous Optic Neuropathy (GON)

Available Metadata (When Provided)

- Age
- Sex
- Image Quality
- Geography
- Acquisition Device
- Ethnicity (if available)

Cross-Region · Diverse Devices · Label Heterogeneity
Demographics · Quality Annotations

2 Model Interfaces (Cross-Paradigm Coverage)

VMs (Vision-only Encoders)

Generalist VM Encoders (ViT, DINOv2, DINOv3)

Ophthalmic VM Encoders (RETFound, VisionFM)

Interface: Embeddings / Logits

VLM-Encoders (Vision-Language Encoders)

Generalist VLM Encoders (CLIP, SigLIP2, MedSigLIP)

Ophthalmic VLM Encoders (EyeCLIP, RET-CLIP, FLAIR)

Interface: Image-Text Similarity / Embeddings

MLLMs (Multimodal LLMs)

Generalist MLLMs (Qwen3-VL, Gemma-3, LLaVA-NeXT)

Medical MLLMs (MedGemma Variants)

Interface: Generated Text + Token Probabilities

Evaluation Methods

- Linear Probing (VM & VLM)
- Zero-Shot Matching (VLM)
- Prompted Generation (MLLM)

3 Evaluation Axes (Multidimensional Analysis)

Performance (Ranking)

AUROC · AUPRC · Accuracy · F1

Calibration

ECE (Positive-Class)

Fairness (When Available)

Age Gap · Sex Gap
(Absolute AUROC / ECE Gaps)

Robustness (Image Quality)

Quality-Stratified Gaps
(AUROC / ECE)

Coverage & Diagnostics

Dataset / Task Coverage
Sample Sizes · Missingness

All metrics computed per Model × Dataset × Task × Method with consistent thresholds & mappings

4 Adaptation & Transfer Tests (MLLM LoRA Fine-Tuning)

LoRA Fine-Tuning Strategy

SFT (Supervised Fine-Tuning)

Low-Rank Adaptation

Training Settings

Train Datasets: BRSET / mBRSET

Tasks: DR · Ref. DR · Glaucoma

Evaluation Protocol

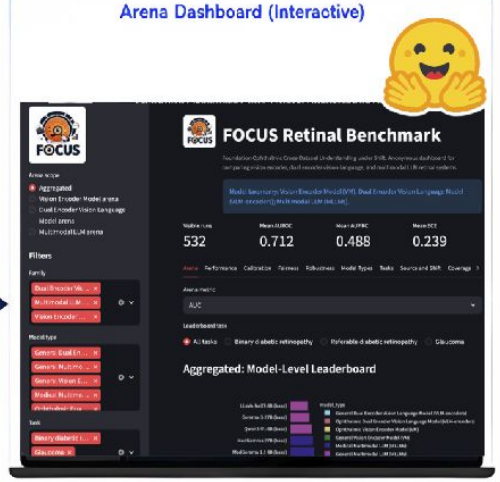
Same-Task Transfer to All Datasets (When Label Mapping Exists)

Compare to Matching Base MLLM (ΔAUROC, ΔAcc, ΔECE)

228 LoRA Adapter Evaluations
(228 SFT per Dataset × 1 Dataset)

5 Arena Outputs & Reporting

Arena Dashboard (Interactive)



Reports & Tables

- Leaderboards (by Task / Model Type)
- Base Results (532) VM / VLM / MLLM
- Performance · Calibration Fairness · Robustness
- FT Results (228) SFT (LoRA)
- Coverage · Diagnostics Model & Dataset Views
- Downloadable CSVs & Plots

Pipeline Summary

10 datasets across 5 countries → 3 harmonized binary tasks (DR, Ref. DR, GON) → 6 model families 22 base models → 3 evaluation axes tailored to interfaces → 7 evaluation axes (multidimensional) → 532 base evaluations + 228 LoRA (SFT) evaluations → One unified arena for generalization beyond one dataset

Legend: 1 Datasets · 2 Models · 3 Evaluation · 4 Adaptation · 5 Reporting · = Images · = Demographics · = Quality · = Geography · = Transfer · = Metrics · = Data · = Fine-Tuning

Original Figure 2

General encoders are strong retinal baselines

Table 3: Method-aware mean metrics by model type and task. Higher is better except for ECE.

Task	Model type	Method	Accuracy	AUROC	AUPRC	ECE	Fairness	Quality	Shift
Overall									
	General VM	Linear probing	**0.803**	0.828	0.617	0.209	0.899	0.885	0.654
	Ophthalmic VM	Linear probing	0.733	0.667	0.414	0.231	0.917	0.899	0.762
	General VLM-encoders	Linear probing	0.802	**0.829**	0.614	0.201	0.923	0.863	0.661
	General VLM-encoders	Zero-shot	0.552	0.549	0.303	0.275	0.934	0.891	**0.800**
	Ophthalmic VLM-encoders	Linear probing	0.763	0.773	0.592	**0.182**	0.915	0.911	0.732
	Ophthalmic VLM-encoders	Zero-shot	0.636	0.647	0.395	0.240	0.948	0.918	0.729
	General MLLMs	Zero-shot prompting	0.575	0.658	0.423	0.388	**0.950**	0.950	0.603
	Medical MLLMs	Zero-shot prompting	0.791	0.803	**0.642**	0.194	0.939	**0.966**	0.567

General VLM probes and general VM probes lead mean AUROC

Original Table 3, excerpt: header and Overall rows

General encoders are strong retinal baselines

Table 3: Method-aware mean metrics by model type and task. Higher is better except for ECE.

Task	Model type	Method	Accuracy	AUROC	AUPRC	ECE	Fairness	Quality	Shift
Overall									
	General VM	Linear probing	**0.803**	0.828	0.617	0.209	0.899	0.885	0.654
	Ophthalmic VM	Linear probing	0.733	0.667	0.414	0.231	0.917	0.899	0.762
	General VLM-encoders	Linear probing	0.802	**0.829**	0.614	0.201	0.923	0.863	0.661
	General VLM-encoders	Zero-shot	0.552	0.549	0.303	0.275	0.934	0.891	**0.800**
	Ophthalmic VLM-encoders	Linear probing	0.763	0.773	0.592	**0.182**	0.915	0.911	0.732
	Ophthalmic VLM-encoders	Zero-shot	0.636	0.647	0.395	0.240	0.948	0.918	0.729
	General MLLMs	Zero-shot prompting	0.575	0.658	0.423	0.388	**0.950**	0.950	0.603
	Medical MLLMs	Zero-shot prompting	0.791	0.803	**0.642**	0.194	0.939	**0.966**	0.567

General VLM probes and general VM probes lead mean AUROC

Original Table 3, excerpt: header and Overall rows

The largest VM gap occurs on mobile mBRSET

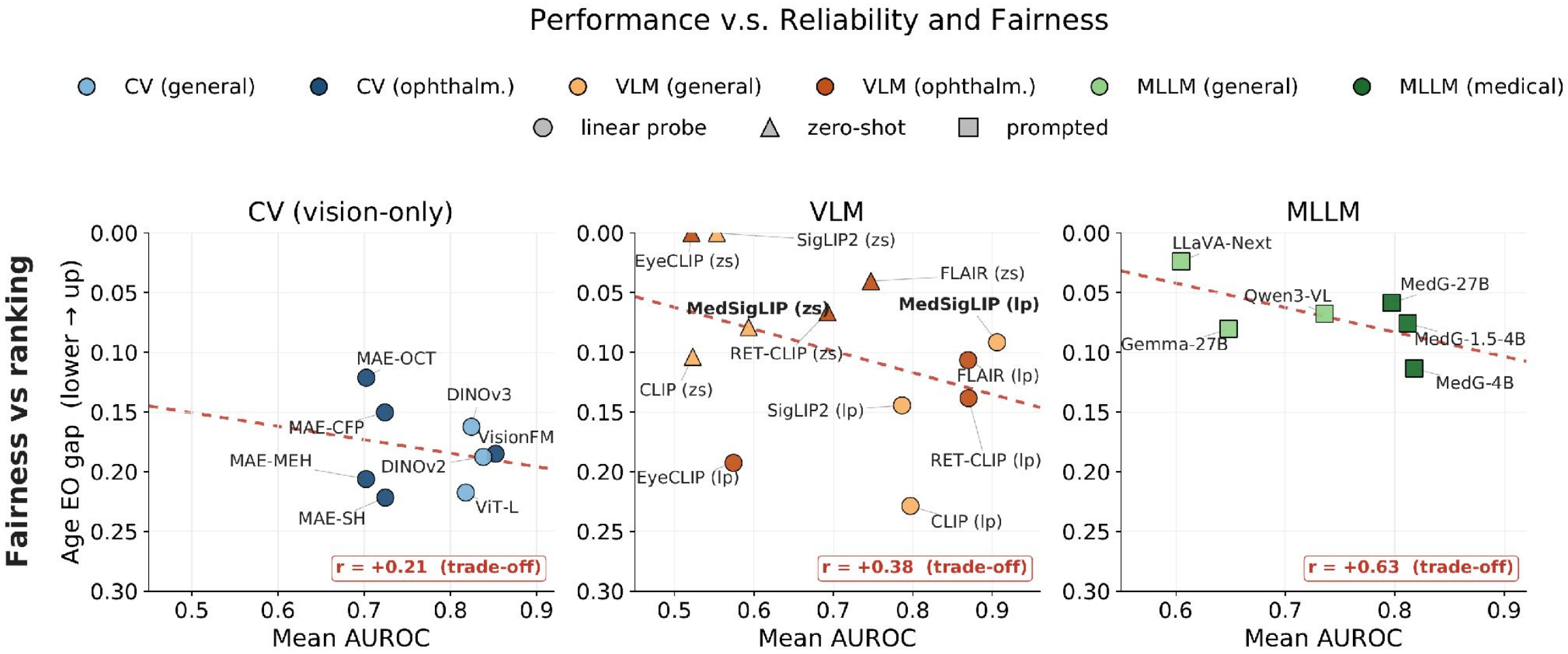
Table 4: Per-dataset mean AUROC for VM encoders. $\Delta = \mathrm{AUROC}_{\mathrm{gen}} - \mathrm{AUROC}_{\mathrm{oph}}$.

Dataset	Country	Device	Gen.	Oph.	Δ
mBRSET	Brazil	**mobile**	0.754	0.592	**+0.162**
IDRiD	India	desktop	0.825	0.692	+0.133
BRSET	Brazil	desktop	0.822	0.700	+0.122
JSIEC1000	China	desktop	0.958	0.851	+0.106
Messidor-2	France	desktop	0.836	0.759	+0.077
RFMiD	India	desktop	0.914	0.849	+0.066
PAPILA	Spain	desktop	0.766	0.710	+0.056
G1020	Germany	desktop	0.631	0.614	+0.017
RFMiD-2	India	desktop	0.846	0.918	-0.072

Population and label differences accompany acquisition differences.

Original Table 4

Performance does not imply readiness for clinical deploy



Original Figure 3

Conclusions

Strong general encoders remain essential baselines

Retinal model evaluation must assess cross-dataset performance, calibration, subgroup disparities, and image-quality robustness.

Code and interactive results

github.com/dsrestrepo/FOCUS-Retinal-Benchmark
huggingface.co/spaces/dsrestrepo/FOCUS-Retinal-Benchmark

INSTITUTE DATAIA
Data Science, Intelligence & Society

Co-funded by
the European Union