MEDIVERSE: A UNIFIED MEDICAL VISION–LANGUAGE MODEL WITH ZERO-FORGETTING EXPANSION

Anonymous authors
Paper under double-blind review

ABSTRACT

A unified medical foundation model should support many imaging modalities, provide a language interface, and expand without losing existing capabilities. Current models usually satisfy only part of this goal. Vision-only models support several modalities but cannot perform text-prompted zero-shot classification or image–text retrieval. Vision–language models provide these abilities but often require large paired corpora and may forget earlier modalities when adapted to new ones. We introduce MediVerse, a pair-efficient and extensible medical vision–language model. MediVerse uses a Multiway Transformer with a *Domain-Aware Mixture-of-Experts*. A deterministic router sends each image to its modality-specific expert, while a separate text pathway interacts with the image pathway through contrastive alignment. We first pretrain on 19.4M unpaired images and then align the model using 3.2M image–text pairs. With a ViT-Large backbone, MediVerse wins 10 of 11 zero-shot benchmarks (mean 0.674 vs. 0.594 for the strongest baseline), is best or tied-best on all 10 linear probes (0.856 vs. 0.824), and wins all 4 cross-modal retrieval benchmarks (20.70 vs. 12.09) across nine imaging modalities. To add a new modality, we train a new routed expert while freezing the existing computation paths. This design preserves old-modality embeddings when the existing computation paths and inference conditions remain unchanged. Across fundus and OCT expansion, adapted baselines lose 2.97–6.11 points of pathology retrieval, whereas MediVerse shows no change on the evaluated pathology task.

1 INTRODUCTION

Clinical imaging is inherently multimodal, spanning radiology, pathology, endoscopy, and other domains that provide complementary views of disease (Cercek et al., 2022). In practice, examinations are selected according to clinical need rather than acquired as a fixed multimodal set (Kasivisvanathan et al., 2018), leading to heterogeneous and often incomplete combinations of imaging evidence across patients (Liu et al., 2023; Qian et al., 2025), and limiting the availability of large-scale patient-matched data across modalities (Vanguri et al., 2022). Moreover, many medical images lack paired text or corresponding images from other modalities, further increasing the prevalence of incompletely paired data. These characteristics call for a medical foundation model that can learn jointly from heterogeneous imaging sources without complete pairing, while remaining sensitive to modality-specific structure and learning representations that transfer across tasks and can be extended to new modalities.

Existing approaches address only parts of this goal. MerMED-FM (Zhou et al., 2026) learns visual representations across multiple imaging modalities but lacks a native vision–language interface, so it supports linear probing but not zero-shot classification or retrieval. BiomedCLIP(BMC) (Zhang et al., 2025), BMC-Clip(BMCA) (Lozano et al., 2025), ConceptCLIP (Nie et al., 2026), and UniMed-CLIP (Khattak et al., 2024) provide image–text alignment for text-prompted classification and retrieval, but rely on paired supervision for that alignment. MUSK (Xiang et al., 2025) combines unpaired masked pretraining with paired contrastive alignment, though only within pathology. These advances leave an important gap: a unified medical image foundation model should exploit

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

abundant image-only data across modalities, use paired data only where available, and incorporate new modalities without disrupting representations already learned.

Learning from image-only data at scale. A contrastive VLM's breadth is capped by the image–text pairs available for each modality, and high-quality medical pairs are expensive to collect. Expert annotation is costly, while literature-derived pairs often contain plots, schematics, or multi-panel figures rather than clinical images. Large paired corpora are especially limited for fundus, OCT, and ultrasound. BMC uses 15M PMC figure–caption pairs (Zhang et al., 2025), and BMCA uses a 24M-pair archive (Lozano et al., 2025). In contrast, masked pretraining can use images without captions. We pretrain on 19.4M image rows and use 3.23M image–text pairs to align the seven-modality core. This separation makes it easier to support modalities for which paired data are scarce.

Preserving representations during expansion. A deployed medical encoder must be *extensible*—new modalities arrive continually—but updating shared components can alter previously learned representations. RetCoP mitigates this through rehearsal and distillation (Yao et al., 2025), while progressive architectures freeze earlier computations (Rusu et al., 2016). Fine-tuning shared parameters can change the representations of previously served modalities. In our experiments, sequential LoRA adaptation also reduces performance on earlier tasks (Table 4). We instead add modality-specific capacity while freezing the existing model. We assume that the image modality is known at inference time, a setting that naturally reflects routine clinical practice.

We present MediVerse, a medical vision–language foundation model based on a Multiway Transformer (Wang et al., 2023) whose feed-forward sublayers form a *Domain-Aware Mixture-of-Experts* (DA-MoE), each modality routed to its own expert, with a separate, unrouted text pathway that interacts with the vision towers only through contrastive alignment. Training proceeds through a modality-aware VQ-KD tokenizer, masked image/text pretraining and contrastive alignment on image–text pairs. After the tokenizer is fixed, training uses no VLM distillation. We integrate a new modality by adding a routed expert and freezing the existing model. Because old-modality inputs continue to use only their original frozen paths, their representations are unchanged (Prop. 1). We evaluate 25 benchmark settings across nine modalities and three task families using one common evaluation protocol within each family.

Our contributions are threefold.

- **Pair-efficient multimodal learning with incomplete pairing.** MediVerse scales to 19.4M unpaired images across seven medical imaging modalities and 3.2M image–text pairs, without requiring cross-modal pairing or paired text for every image. Modality-aware VQ-KD tokenization and DA-MoE enable efficient, selective use of available modalities, making the framework well suited to heterogeneous clinical data and providing a natural basis for continual expansion to new modalities. Controlled ViT-B ablations further validate the contribution of DA-MoE and its modality-specific experts.
- **Continual expansion to unseen modalities without forgetting.** MediVerse extends to modalities not seen during pretraining by adding new modality-specific experts while keeping the shared backbone, text encoder, and existing experts frozen. Matched-data sequential experiments assess both adaptation to new modalities and retention of existing representations, while independently trained full-data experts are integrated into the final modality model. Parameter and output consistency checks further verify that the existing pathways remain unchanged during expansion.
- **Comprehensive evaluation with strict data separation.** We evaluate MediVerse on zero-shot classification, linear probing, and image–text retrieval across diverse medical imaging modalities, using benchmark-specific protocols and excluding evaluation test samples from pretraining. Across these settings, MediVerse achieves the best or tied-best performance in 24 of 25 evaluations, including $10/11$ classification, $10/10$ linear-probe, and 4/4 retrieval settings.

2 Related Work

Biomedical Multimodal Foundation models. CLIP (Radford et al., 2021) inspired biomedical vision–language models such as BMC , BMCA , and ConceptCLIP . These models support text-prompted classification and image–text retrieval, but do not explicitly separate the computation paths

of different imaging modalities. Vision-only models such as MerMED-FM cover several modalities, but lack a text interface. Another strategy is to connect existing domain-specific models: CTF (Chen et al., 2026) integrates separately trained radiology and pathology encoders through clinical concepts and cross-domain prompting. In contrast, MediVerse combines broad modality coverage and a language interface within one routed encoder.

Multiway pretraining and visual tokenization. MediVerse builds on the Multiway Transformer of BEiT-3 (Wang et al., 2023) and the VQ-KD tokenizer of BEiT-v2 (Peng et al., 2022). MUSK (Xiang et al., 2025) applies this framework to pathology using masked image and language modeling followed by contrastive alignment. MediVerse extends the design to multiple modalities through one tokenizer supervised by six visual teachers and visual FFNs with modality-specific experts. The teachers supervise tokenizer training; after the tokenizer is frozen, the VLM receives no additional teacher-feature or teacher-logit loss.

Mixture-of-Experts and continual expansion. MoE models route inputs through selected parameters (Shazeer et al., 2017; Fedus et al., 2022). MediVerse uses deterministic modality routing and therefore needs no load-balancing loss. Unlike learned token routing, each image uses a known modality label. LoRA (Hu et al., 2022), adapters (Houlsby et al., 2019), regularization (Kirkpatrick et al., 2017), and rehearsal can reduce interference, while progressive networks preserve earlier computations through model expansion (Rusu et al., 2016). MedQwen (Nejati Manzari et al., 2026) routes low-rank experts in a generative medical VLM; MediVerse instead uses full-rank modality experts in a contrastive encoder. Its preservation guarantee follows from keeping existing computation paths fixed, not from routing alone.

3 METHOD

MediVerse has three ingredients: a Multiway backbone whose feed-forward stack is a *Domain-Aware Mixture-of-Experts* (§3.1); a two-stage training pipeline (§3.2); and a continual-expansion rule that preserves existing embedding functions under explicit conditions (§3.3). Throughout, $m \in \mathcal{M}$ denotes an input's modality. Figure 1 gives the end-to-end overview.

3.1 BACKBONE AND DOMAIN-AWARE MIXTURE-OF-EXPERTS

Multiway backbone. MediVerse builds on a Multiway Transformer (Wang et al., 2023): an image and a caption are each patch- or token-embedded and encoded through their own forward pass, and every layer applies token-type-specific self-attention followed by a token-type-specific feed-forward branch. The pre-normalized block is

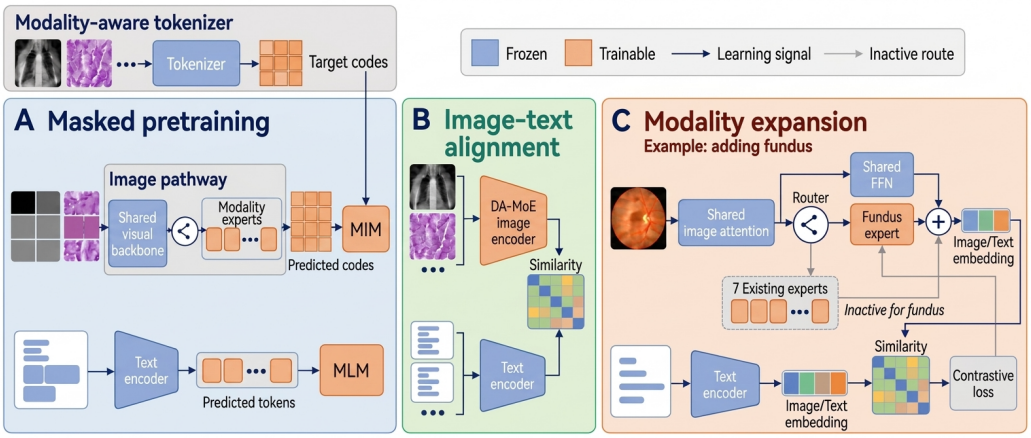
$$\begin{aligned} \boldsymbol{h}_t^{(\ell)} &= \boldsymbol{u}_t^{(\ell-1)} + \mathrm{Attn}_t^{(\ell)}\Big(\mathrm{LN}_{t,1}^{(\ell)}(\boldsymbol{u}_t^{(\ell-1)})\Big), \\ \boldsymbol{u}_t^{(\ell)} &= \boldsymbol{h}_t^{(\ell)} + \mathrm{FFN}_t^{(\ell)}\Big(\mathrm{LN}_{t,2}^{(\ell)}(\boldsymbol{h}_t^{(\ell)})\Big), \qquad t \in \{\text{vision}, \text{language}\}. \end{aligned} \tag{1}$$

Attention and FFN denote the corresponding implementation blocks, including their internal projections and any implementation-specific operations in addition to the displayed pre-normalization; dropout and stochastic depth are omitted for clarity. The subscript t selects an independently parameterized attention and FFN branch, and the two normalization subscripts distinguish the attention and FFN operations. The two towers share no attention weights and interact only through the Stage-2 contrastive loss on their pooled embeddings, never through a joint attended sequence. It is the vision FFN branch that we make *domain-aware*.

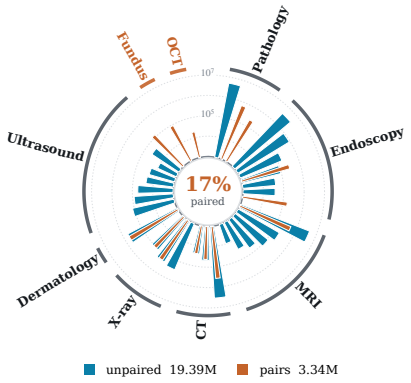
Domain-Aware MoE. We replace the single vision FFN at a set of layers $\mathcal{L} \subseteq \{1, \ldots, 24\}$ (backbone depth for the reported ViT-Large model) with a bank of E domain experts $\{f_1^{(\ell)}, \ldots, f_E^{(\ell)}\}$ plus a shared expert $s^{(\ell)}$. A *hard* router maps each image to the single expert of its modality,

$$\mathrm{FFN}_{\mathrm{vision}}^{(\ell)}(\boldsymbol{x}; m) = s^{(\ell)}(\boldsymbol{x}) + f_{r(m)}^{(\ell)}(\boldsymbol{x}), \qquad \ell \in \mathcal{L}, \quad r : \mathcal{M} \to \{1, \ldots, E\}, \tag{2}$$

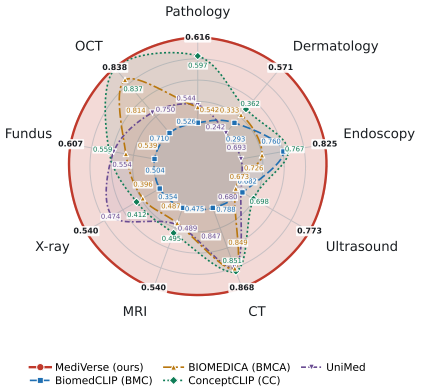
so pathology, CT, MRI, ultrasound, X-ray, dermatology, and endoscopy each have an expert while $s^{(\ell)}$ carries shared visual information. Routing is deterministic and specified by the image modality ID. Only the selected modality-specific expert contributes alongside the shared FFN. Different



(a) the MediVerse pipeline



(b) training corpus



(c) modality-level results

Figure 1: **MediVerse: pipeline, training data, and evaluation breadth.** **(a)** A modality-aware tokenizer supplies discrete targets for masked pretraining; contrastive alignment then connects the separate image and text pathways. New modalities use paired data to train new routed experts while existing paths remain frozen. **(b)** Each spoke is a source dataset grouped by modality. On a shared $\log_{10}$ scale, wide bars show Stage-1 image rows and narrow bars show Stage-2 pairs; fundus and OCT use paired data only. The core pair count is approximately 17% of the masked-pretraining image-row count. **(c)** Raw mean performance by modality across applicable task families, with retrieval percentages divided by 100. Each axis uses its own range; markers are data and curves are visual guides.

modalities retain separate expert parameters while jointly training the shared visual pathway. No load-balancing objective is used in the reported configuration.

Expert placement. We index the 24 Transformer layers as $1, \ldots, 24$. $\mathcal{L}$ controls how much dedicated capacity a modality receives. *Integrate*-mode routes every fourth layer, $\mathcal{L} = \{4, 8, 12, 16, 20, 24\}$, giving depth-6 experts. *Expand*-mode routes every layer, $\mathcal{L} = \{1, \ldots, 24\}$, giving depth-24 experts. Layers outside $\mathcal{L}$ use their ordinary visual FFN. The seven-modality ViT-L core uses the six integrate-mode routed layers. For the full-data fundus and OCT extensions, the remaining layers are converted with function-preserving FFN remapping and zero-output branches, and each new expert spans all 24 layers. This adds capacity for the new modalities while preserving the existing computation paths (§4.6).

3.2 Two-stage training pipeline

Modality-aware tokenizer. A single VQ-KD tokenizer (Peng et al., 2022) q : image $\rightarrow \{1, \ldots, K\}^N$ ($K = 8192$) maps each image to a grid of discrete codes, trained to reconstruct per-

modality teacher features through a quantized bottleneck. External teacher features supervise tokenizer training. The resulting tokenizer is then frozen and produces discrete targets for MediVerse's masked-image objective. Gradients from this objective update MediVerse but are not propagated through the tokenizer or its teachers.

Stage 1: masked unimodal pretraining. With the tokenizer frozen we pretrain the backbone with two independent objectives, each on its own forward pass: masked image modeling against the VQ-KD codes, and masked language modeling on unpaired text. The image and text streams are unpaired. Let $\mathcal{I}_M$ and $\mathcal{T}_M$ contain masked sample–token pairs across the image and text microbatches, respectively, and let $c_{b,i}$ denote the frozen tokenizer target at image position i in sample b. Routing each image through its modality expert (Eq. 2),

$$\mathcal{L}_{\mathrm{MIM}} = -\frac{1}{|\mathcal{I}_M|} \sum_{(b,i)\in\mathcal{I}_M} \log p_\theta\big(c_{b,i} \mid \widetilde{\boldsymbol{x}}_b^{\mathrm{v}}, m_b\big), \tag{3}$$

$$\mathcal{L}_{\mathrm{MLM}} = -\frac{1}{|\mathcal{T}_M|} \sum_{(q,j)\in\mathcal{T}_M} \log p_\theta\big(t_{q,j} \mid \widetilde{\boldsymbol{x}}_q^{\mathrm{t}}\big), \tag{4}$$

$$\boxed{\mathcal{L}_{\mathrm{stage1}} = \mathcal{L}_{\mathrm{MIM}} + \mathcal{L}_{\mathrm{MLM}}.} \tag{5}$$

where $\widetilde{\boldsymbol{x}}_b^{\mathrm{v}}$ and $\widetilde{\boldsymbol{x}}_q^{\mathrm{t}}$ are masked image and text inputs. The indices b and q refer to independent microbatches and do not indicate paired samples; MLM has no access to image tokens. Stage 1 is the equally weighted sum of the mean MIM and mean MLM losses. If no text tokens are masked, the implementation sets $\mathcal{L}_{\mathrm{MLM}} = 0$. Stage 1 trains the shared visual pathway and the original modality-specific experts before image–text alignment. During subsequent expansion, a new expert learns from paired data against the frozen text embedding space.

Stage 2: contrastive alignment. For core Stage 2 training, we attach linear vision/text heads, pool and ℓ_2-normalize a global image embedding $\boldsymbol{z}^{\mathrm{v}}$ and text embedding $\boldsymbol{z}^{\mathrm{t}}$, and align them over 3.23M image–text pairs. Let a be the learned logit-scale parameter and define

$$s_{ik} = \gamma\,(\boldsymbol{z}_i^{\mathrm{v}})^\top \boldsymbol{z}_k^{\mathrm{t}}, \qquad \gamma = \min\{\exp(a), 100\}. \tag{6}$$

The effective temperature is $\tau = 1/\gamma$. The symmetric core-training objective is

$$\boxed{\mathcal{L}_{\mathrm{CLIP}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii})}{\sum_{k=1}^{B}\exp(s_{ik})} + \log\frac{\exp(s_{ii})}{\sum_{k=1}^{B}\exp(s_{ki})}\right].} \tag{7}$$

B is the number of pairs participating in one contrastive calculation. With cross-GPU gathering, seven GPUs and 32 pairs per GPU give $B = 224$. Eighteen accumulation steps give a $4{,}032$-pair optimizer batch, not a $4{,}032$-pair negative pool. The implementation's contrastive helper additionally divides the returned loss by the distributed world size; this affects optimizer scaling but not the positive or negative sets in the equation. The matched-data expansion experiment instead uses an equal-caption multiple-positive contrastive objective and is therefore not described exactly by the diagonal-only equation above.

3.3 Zero-forgetting continual expansion

A deployed encoder must absorb new modalities without disturbing the ones it serves. MediVerse does this by *growing the expert bank on a frozen base*.

Definition 1 (Integration operator)**.** *Given MediVerse with modality experts $\{f_j\}_{j=1}^{E}$ and a new modality $m^\star$, integration $\mathcal{A}$ (i) appends a new expert f_{E+1} whose parameters are independent of the existing experts; (ii) freezes all parameters on the existing image and text embedding paths, including embeddings, attention blocks, normalization layers, shared FFNs, existing modality experts, and projection heads; and (iii) trains only the new expert using paired image–text data and a contrastive objective. The reported fundus and OCT extensions use no additional Stage-1 masked pretraining. The new expert comprises modality-specific FFNs across the routed layers. Integration assigns $r(m^\star) = E + 1$ while preserving all existing routing assignments and embedding computations.*

Proposition 1 (Preservation of existing embedding functions)**.** *Let g_θ^{v} and g_θ^{t} denote the image and text embedding functions before expansion, and let θ' denote the parameters after expansion. If expansion preserves all original parameters, persistent state, routing, and inference operations, then*

$$g_{\theta'}^{\mathrm{v}}(\boldsymbol{x}, m) = g_\theta^{\mathrm{v}}(\boldsymbol{x}, m), \qquad \forall\, \boldsymbol{x},\ m \in \mathcal{M}_{\mathrm{old}}, \tag{8}$$

$$g_{\theta'}^{\mathrm{t}}(\boldsymbol{t}) = g_\theta^{\mathrm{t}}(\boldsymbol{t}), \qquad \forall\, \boldsymbol{t}. \tag{9}$$

Proof sketch. The embedding of an input with modality m depends only on parameters on its *computation path*: shared attention and embeddings, the shared experts $s^{(\ell)}$, and—by the hard router (Eq. 2) with $r(m) = j \leq E$ unchanged—exactly one old expert f_j per DA-MoE layer. Every one of these lies in the frozen base; the only trainable parameters (f_{E+1}) do not affect the embedding of an old-modality input. The text pathway also remains unchanged. With the same persistent state, routing, and inference operations, both existing embedding functions are therefore preserved. For the 6-to-24-layer conversion, this argument additionally requires the stated function-preserving remapping. □

Proposition 1 states preservation by construction under the specified conditions. Methods that alter the effective mapping used by existing modalities, including shared LoRA adapters, do not automatically satisfy these conditions. The guarantee arises from preserving existing computation paths, not from routing alone. Unchanged predictions and metrics additionally require unchanged preprocessing, scoring rules, candidate sets, and downstream classifiers. Bitwise equality is an empirical verification on the evaluated inputs, not an unconditional consequence for every possible output. §4.5 reports this verification, and Section 4.6 tests whether the new experts learn useful representations.

4 Experiments

4.1 Setup

Models and data. MediVerse uses a ViT-Large/16 backbone with seven base DA-MoE experts (1.280B total stored parameters[1]). Baselines are (BMC, ViT-B/16, 196M), (BMCA, ViT-L/14, 454M), ConceptCLIP (SigLIP-SO400M, 543M), UniMed (ViT-B/16, 196M) (Khattak et al., 2024) and, on linear probing only, the vision-only MerMED-FM (ViT-B, 86M). MediVerse has the largest total parameter count in this comparison. Unless noted otherwise, all results below use this ViT-Large backbone, including the continual-expansion experiments of §4.5–§4.6. Stage 1 pretrains on 19.39M *unpaired* images across the seven base modalities; stage 2 then aligns only 3.23M pairs, rising to 3.34M pairs after the expansions of §4.6. Fundus and OCT are integrated using *paired data only*: unlike the seven base modalities, neither receives Stage-1 unpaired pretraining, so adding them leaves the Stage-1 image total unchanged at 19.39M. For the seven-modality core, the 3.23M alignment-pair count is approximately 17% of the 19.39M Stage-1 image-row count (Fig. 1(b)). For the seven core modalities, masked pretraining shapes visual features before contrastive alignment. The added fundus and OCT experts instead learn directly from paired data against the frozen text embedding space. Spread over nine modalities, the total pair count is roughly a fifth of BMC's 15M pairs and an eighth of the 24M-pair BMCA archive, where MUSK spends 1M on pathology alone. After tokenizer training, no additional teacher-feature or teacher-logit distillation loss is used. The tokenizer obtains visual targets from six teachers: UNI for pathology, MedDINOv3 for CT, USFM for ultrasound, RAD-DINO for X-ray, GastroNet-5M for endoscopy, and BMC for MRI and dermatology (supplementary material).

Evaluation. Training and evaluation sources are separated; Appendix A lists every dataset and leakage-control decision. We report 11 zero-shot benchmarks using balanced accuracy, 10 linear probes on frozen features, and 4 retrieval benchmarks using mean bidirectional $R@\{1, 10, 50\}$. Retrieval uses cosine similarity and one fixed gallery per benchmark. Every model within a table uses

[1] A single modality's image pathway contains 355.40M parameters. This count includes the shared visual pathway and the selected expert, but excludes the text encoder and unselected experts; see the supplementary material.

Table 1: **Zero-shot classification**, balanced accuracy. **Best**, second. MediVerse wins ten of eleven rows outright, losing only Pneumonia to BMCA's radiology-specific pretraining. ‡PanNuke is a 19-class tissue panel on which every model scores well below saturation.

Benchmark	Modality	OURS	BMC	BMCA	ConceptCLIP	UniMed-CLIP
BreaKHis	pathology	**.665**	.532	.553	.637	.601
Pneumonia	X-ray	.722	.732	**.911**	.859	.786
CASIA-CXR	X-ray	**.720**	.630	.590	.680	.670
Thyroid	ultrasound	**.744**	.604	.579	.612	.599
EDD-polyp	endoscopy	**.824**	.752	.644	.714	.623
HAM10000	dermatology	**.819**	.300	.365	.389	.178
LC25000	pathology	**.881**	.664	.694	.879	.709
PanNuke‡	pathology	**.246**	.071	.093	.142	.112
SARS-CoV-2 CT	CT	**.743**	.626	.730	.717	.731
APTOS-DR5	fundus	**.589**	.377	.382	.457	.447
ColorFundus10	fundus	**.460**	.375	.419	.445	.459
Mean (n=11)		**.674**	.515	.542	.594	.538

the same prompts, probes, galleries, and scoring code. The matched-data expansion comparison also uses the same fundus/OCT subsets and checkpoint-selection rules for all methods.

4.2 Zero-shot classification

MediVerse wins 10 of 11 zero-shot benchmarks outright, for a mean of 0.674 (+0.080 over ConceptCLIP, the strongest baseline at 0.594). The largest margins occur on dermatology (0.819 vs. 0.389), thyroid ultrasound (0.744 vs. 0.612), and endoscopic polyp detection (0.824 vs. 0.752). MediVerse also leads on the newly added APTOS-DR5 fundus benchmark (0.589 vs. 0.457). Its narrowest win is on ColorFundus10 (0.460 vs. 0.459 for UniMed). BMCA performs best on Pneumonia (0.911 vs. 0.722 for MediVerse). All models obtain low scores on the 19-class PanNuke benchmark.

4.3 Linear probe

Table 2 probes frozen features with logistic regression. MediVerse has the best mean (0.856 over all ten rows; 0.906 on the seven benchmarks common to every baseline) and is best or tied-best on *all* 10 rows. Ties are CT-Kidney, where MerMED-FM matches our 0.993 at the reported precision, and the newly added ColorFundus10, where ConceptCLIP matches our 0.712 at the reported precision. Linear probing is also the *only* axis the vision-only MerMED-FM supports; it has no fundus/OCT probe, so its 0.867 mean stays over the original seven benchmarks, where it remains the strongest baseline, narrowly ahead of ConceptCLIP's 0.866 mean on that same subset (ConceptCLIP's mean over all ten rows is lower, 0.824, since its fundus/OCT scores trail its performance on the original seven).

4.4 Cross-modal retrieval

Under the same retrieval protocol, MediVerse achieves the highest mean score (20.70 vs. 12.09, a difference of 8.61; Table 3) and leads all four benchmarks. The largest gains occur on OpenI (35.88 vs. 21.97 for UniMed) and MR-RATE (11.52 vs. 5.68 for BMCA). The margins are smaller on derm7pt (9.64 vs. 8.74 for ConceptCLIP) and PathGen-600 (25.75 vs. 25.61 for ConceptCLIP).

Figure 1(c) summarizes performance across all nine modalities. The smallest gaps occur on OCT and CT: MediVerse and ConceptCLIP differ by 0.001 on OCT and 0.017 on CT. The largest gaps occur on dermatology and X-ray.

4.5 Zero-forgetting continual expansion

We add independently trained fundus and OCT experts to the frozen ViT-Large core and test an old modality after each addition. On a balanced 4096-image PCam probe, the core, +fundus, and +fundus+OCT checkpoints give identical AUROC (0.8691), balanced accuracy (64.26%), features,

Table 2: **Linear probe** on frozen features, 5-fold cross-validation (seed 0), balanced accuracy. **Best**, second. MediVerse is best or tied-best on all ten rows and leads the mean despite no longer being the smallest encoder in the comparison. Ties are CT-Kidney, where MediVerse and MerMED-FM both reach 0.993, and ColorFundus10, where MediVerse and ConceptCLIP both reach 0.712. †MerMED-FM has no reported fundus/OCT probe; its mean is computed over the original seven benchmarks, where MediVerse scores 0.906 on the same subset.

Benchmark	OURS	BMC	BMCA	ConceptCLIP	MerMED	UniMed-CLIP
HEST (pathology)	**.994**	.978	.986	.981	.984	.979
BrainTumor (MRI)	**.972**	.866	.894	.926	.917	.903
CT-Kidney (CT)	**.993**	.950	.967	.986	**.993**	.964
TN5000 (ultrasound)	**.802**	.761	.768	.783	.782	.762
RadImageNet (MRI)	**.956**	.940	.942	.953	.932	.949
EDD2020 (endoscopy)	**.825**	.767	.808	.820	.824	.763
HAM10000 (dermatology)	**.797**	.519	.559	.610	.626	.515
APTOS-DR5 (fundus)	**.665**	.610	.654	.624	–†	.620
ColorFundus10 (fundus)	**.712**	.653	.700	**.712**	–†	.689
OCTDL (OCT)	**.838**	.710	.814	.837	–†	.750
Mean (n=10)	**.856**	.776	.809	.824	–†	.793

Table 3: **Cross-modal retrieval**, mean $R@\{1, 10, 50\}$ over image→text and text→image, with one fixed gallery per benchmark for all models. All benchmarks use cosine similarity; PathGen-600 treats rows with identical captions as positives. **Best**, second. MediVerse leads all four rows and the mean.

Benchmark	Modality	OURS	BMC	BMCA	ConceptCLIP	UniMed-CLIP
OpenI	X-ray	**35.88**	2.60	4.23	5.47	21.97
MR-RATE	MRI	**11.52**	4.72	5.68	5.12	5.12
derm7pt	dermatology	**9.64**	6.00	7.37	8.74	3.30
PathGen-600	pathology	**25.75**	17.64	19.39	25.61	17.97
Mean (n=4)		**20.70**	7.74	9.17	11.24	12.09

Table 4: **PathGen pathology retention** after fundus→OCT adaptation. Baselines use sequential LoRA and report the mean over three seeds; MediVerse uses one assembled-expert sequence. Full trajectories are in Appendix Figure 4.

Model (adaptation)	Δ PathGen
MediVerse ViT-L (routed expert)	**0.0**
BMCA + LoRA	−3.22
BMC + LoRA	−6.11
UniMed-CLIP + LoRA	−2.97
ConceptCLIP + LoRA	−5.39

cosine scores, and predictions. Parameter and seven-modality retention checks are reported in Appendices F and G.

4.6 Adapting unseen modalities

The assembled experts learn the new modalities: in Tables 1 and 2, MediVerse is best or tied-best on all five fundus/OCT rows. We also measure pathology retention on PathGen-600. MediVerse remains at 25.75% mean bidirectional recall after both additions, with byte-identical saved image and text features. Four baselines adapted by sequential LoRA lose 2.97–6.11 points (Table 4). This comparison is descriptive: MediVerse's full-data experts are trained independently and assembled, whereas the baselines use sequential LoRA on matched-data subsets. Appendix E provides a controlled matched-data comparison and the full per-stage results.

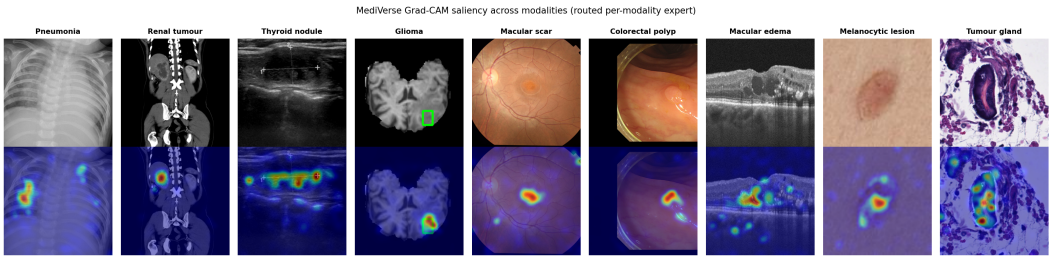


Figure 2: **Qualitative saliency across nine modalities.** Similarity-gradient Grad-CAM through each modality's routed expert; top row original, bottom row saliency. Left to right: pneumonia, renal tumour, thyroid nodule, glioma on FLAIR with the ground-truth box, macular scar, colorectal polyp, macular edema on OCT, melanocytic lesion, and malignant gland on H&E. Fundus and OCT use experts added to a frozen model.

4.7 Analysis

What routing buys. A model with *no* experts (matched random initialization, tokenizer, training data, and training seed) performs worse on 15 of 18 benchmarks, ties MediVerse on BreaKHis zero-shot classification, and beats MediVerse on Pneumonia zero-shot classification and HAM10000 linear probing. On average it is 0.043 worse on zero-shot balanced accuracy, 0.020 worse on linear-probe balanced accuracy, and 0.133 points worse on retrieval. Both variants train from random initialization under a fixed, matched five-epoch budget per stage, with no early stopping or validation-based checkpoint selection, so this isolates the marginal effect of routing capacity under a short, controlled schedule rather than at converged, full-scale training; per-benchmark rows and the full training configuration are in the supplementary material. Additional tokenizer results, matched-data experiments, and seven-modality retention appear in the appendix.

Why does MediVerse perform better. The compared baselines and our no-MoE variant all rely on a shared visual pathway across modalities, although they differ in architecture, pretraining data, scale, and optimization. Within the controlled MediVerse setting, replacing this shared-only design with DA-MoE improves 15 of 18 benchmarks under matched initialization, data, optimization, training duration, and seed, suggesting that modality-specific capacity contributes to the performance gains. MediVerse also learns from 19.4M unpaired images and uses a multi-teacher VQ-KD tokenizer to provide domain-relevant visual targets before contrastive alignment. Together, these results support the benefits of domain-aware tokenization, unpaired visual pretraining, and modality-specific experts.

4.8 Qualitative saliency analysis

Figure 2 shows similarity-gradient Grad-CAM maps for all nine modalities. We back-propagate the image–text cosine similarity for each finding prompt through the routed expert to the final block's patch tokens. The highlighted regions overlap clinically relevant structures, including in fundus and OCT images.

5 Conclusion

MediVerse combines masked pretraining on 19.4M image rows with alignment on 3.2M image–text pairs. It performs strongly across nine modalities and supports expansion by adding modality-specific experts to a frozen base. This design preserves the existing embedding functions under the stated conditions. On the evaluated pathology retrieval task, MediVerse shows no change after fundus and OCT expansion, whereas sequentially adapted LoRA baselines lose 2.97–6.11 points. These results support modular expert expansion as a practical way to extend a medical vision–language model without changing its existing representations.

Reproducibility Statement

We provide the architecture and objectives in Section 3, the evaluation protocols in Section 4, and dataset sources, splits, training configurations, checkpoint-selection rules, and implementation details in the appendix. We will release the code, tokenizer, checkpoints, plotting scripts, numerical results, and evaluation manifests.

Ethics Statement

This work follows the ICLR Code of Ethics and uses publicly available datasets under their respective access conditions. MediVerse is a research model, not a clinical device; clinical use would require prospective, site-specific validation and regulatory review.

AI Use Statement

During the preparation of this manuscript, we used publicly available large language models solely for limited English-language editing. All research ideas, methodological designs, experimental implementations, analyses, and interpretations of the results were independently developed by the authors.

References

Andrea Cercek, Melissa Lumish, Jenna Sinopoli, Jill Weiss, Jinru Shia, Michelle Lamendola-Essel, Imane H. El Dika, Neil Segal, Marina Shcherba, Ryan Sugarman, et al. PD-1 blockade in mismatch repair–deficient, locally advanced rectal cancer. *New England Journal of Medicine*, 386: 2363–2376, 2022. doi: 10.1056/NEJMoa2201445.

Yihang Chen, Yanyan Huang, Fuying Wang, Maximus Yeung, Yuming Jiang, Shujun Wang, and Lequan Yu. Bridging radiology and pathology foundation models via concept-based multimodal co-adaptation. In *International Conference on Learning Representations (ICLR)*, 2026.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning (ICML)*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Veeru Kasivisvanathan, Antti S. Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A. Mynderse, Markku H. Vaarala, Alberto Briganti, Lars Budäus, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *New England Journal of Medicine*, 378:1767–1777, 2018. doi: 10.1056/NEJMoa1801993.

Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of Biomedical Informatics*, 145:104466, 2023. doi: 10.1016/j.jbi.2023.104466.

Alejandro Lozano, Min Woo Sun, James Burgess, et al. Biomedica: An open biomedical image–caption archive, dataset, and vision–language models derived from scientific literature. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Omid Nejati Manzari, Hojat Asgariandehkordi, Taha Koleilat, Yiming Xiao, and Hassan Rivaz. Sparse spectral lora: Routed experts for medical vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35351–35362, 2026.

Yuxiang Nie, Sunan He, Yequan Bie, Yihui Wang, Zhixuan Chen, Shu Yang, Zhiyuan Cai, Linshan Wu, Hongmei Wang, Xi Wang, Ngai Shing Cheng, Luyang Luo, Mingxiang Wu, Haibo Jin, Xian Wu, Ronald Cheong Kin Chan, Yuk Ming Lau, Zhengyu Zhang, Sushan Xiao, Can Yang, Yinghua Zhao, Xiaohui Duan, Li Zhang, Li Liang, Yefeng Zheng, Pranav Rajpurkar, and Hao Chen. An explainable biomedical foundation model via large-scale concept-enhanced vision-language pre-training. *Nature Biomedical Engineering*, 2026.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. In *arXiv preprint arXiv:2208.06366*, 2022.

J. Qian, A. Joshi, H. Li, N. A. Parikh, J. R. Dillman, and L. He. Utility of deep learning to address missing modalities from multi-modal medical imaging studies: A systematic review. *Artificial Intelligence and Applications*, 2025. doi: 10.47852/bonviewAIA52026392.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Rami S. Vanguri, Jia Luo, Andrew T. Aukerman, Jacklynn V. Egger, Christopher J. Fong, Natally Horvat, Andrea Pagano, João A. Bosco Araujo-Filho, Luke Geneslaw, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nature Cancer*, 3:1151–1164, 2022. doi: 10.1038/s43018-022-00416-8.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638:769–778, 2025.

Yuang Yao, Ruiqi Wu, Yi Zhou, and Tao Zhou. Continual retinal vision-language pre-training upon incremental imaging modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image–text pairs. *NEJM AI*, 2025.

Yang Zhou, Chrystie Wan Ning Quek, Jun Zhou, Yan Wang, Yang Bai, Laura Gutierrez, Yuhe Ke, Jie Yao, Zhen Ling Teo, Darren Shu Jeng Ting, et al. Mermed-fm: Multimodal, multi-disease medical imaging foundation model. *The Lancet Digital Health*, 8(7), 2026.