Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An open AI system for dermatology tasks

Hong-Yu Zhou

hongyu.zhou.ai@gmail.com

Tsinghua University

Linwei Chen

School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University; School of Computing and Data Science, The University of Hong Kong; Baichuan Inc.

Lijun Liu

Baichuan Inc.

Ying Fu

Baichuan Inc.

Meng Tian

Baichuan Inc.

Zhishou Zhang

Baichuan Inc.

Hengfu Cui

Baichuan Inc.

Qiang Ju

Baichuan Inc. https://orcid.org/0009-0000-0795-952X

Qianxi Li

Baichuan Inc.;Department of Dermatology, Peking University First Hospital;Beijing Key Laboratory of Molecular Diagnosis on Dermatoses

Jungong Han

Department of Automation, Tsinghua University

Yizhou Yu

The University of Hong Kong https://orcid.org/0000-0002-0470-5548

Article

Keywords:

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10631734/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

An open AI system for dermatology tasks

Linwei Chen[1,2,3], **Lijun Liu**[3], **Ying Fu**[4,✉], **Meng Tian**[3], **Zhishou Zhang**[3], **Hengfu Cui**[3], **Qiang Ju**[3], **Qianxi Li**[3,5,6], **Jungong Han**[7,✉], **Yizhou Yu**[2,✉] **and Hong-Yu Zhou**[1,3,✉]

[1]School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University
[2]School of Computing and Data Science, The University of Hong Kong
[3]Baichuan Inc.
[4]School of Computer Science and Technology, Beijing Institute of Technology
[5]Department of Dermatology, Peking University First Hospital
[6]Beijing Key Laboratory of Molecular Diagnosis on Dermatoses
[7]Department of Automation, Tsinghua University

ABSTRACT

Skin diseases impose a major global disease burden, yet dermatologists remain scarce, especially in low- and middle-income countries. Foundation models could extend specialist-level care, but leading systems depend on private cohorts that most institutions cannot access, reproduce, afford or audit. Here we present UniDerm, a dermatology foundation model trained entirely on publicly available image–text data through supervision-denoising contrastive learning. Public skin images each carry several heterogeneous texts, and standard contrastive learning averages them into one representation, letting each text add noise to the others. UniDerm instead uses natural-language instructions to disentangle these targets and recover the fine-grained supervision that averaging discards. Unlike private-cohort systems, UniDerm is released as a fully open system: its model weights, code and the data recipe needed to reconstruct its training corpus are all public. Across 11 benchmarks spanning five continents, eight countries and the full Fitzpatrick range, UniDerm matches or exceeds private-cohort models: it reaches their full-label diagnostic accuracy using roughly a third of the labels, and detects malignancy at an AUROC of 0.946, above a 302-reader expert consensus. Its accuracy holds across skin tones, including the darker-skin lesions on which such systems most often fail. Because the approach is not specific to skin, it charts a route to specialist-level medical AI from public data that under-resourced communities can build, audit and own rather than purchase.

Introduction

Skin diseases are a leading cause of non-fatal disease burden worldwide, yet trained dermatologists remain scarce, especially in low- and middle-income countries [1–4]. Foundation models for dermatology could extend specialist-level care, with recent systems approaching board-certified-dermatologist performance in reader studies [5–8].

The leading dermatology foundation models published to date, however, are not fully open. With few exceptions, the state-of-the-art systems depend on private clinical cohorts that most research groups cannot access, and most release no weights that the community can download and test (Extended Data Table ED1). PanDerm [8], the leading published example, was trained on more than two million proprietary images sourced from eleven institutions across four imaging modalities. Earlier dermatologist-level classifiers [6, 9, 10] and recent multimodal systems [11] followed the

Corresponding author: Hong-Yu Zhou (hongyu.zhou.ai@gmail.com), Yizhou Yu (yizhouy@acm.org), Ying Fu (fuying@bit.edu.cn), and Jungong Han (jghan@tsinghua.edu.cn)

Table ED1 | **Openness of representative dermatology-diagnosis deep-learning models, 2017–2026.** Landmark and state-of-the-art skin-disease-diagnosis systems grouped into three release tiers. **Closed**: weights not downloadable. **Open-weight**: weights downloadable but training data private or licence restrictive, so the model cannot be independently retrained. **Fully open**: public training data *and* downloadable weights, hence reproducible without any private cohort. **Weights**, ✓ if openly downloadable; **License** summarises reuse terms. Compiled from each model's paper, repository and licence as of June 2026. Abbreviations: CNN, convolutional neural network; SSL, self-supervised learning; MLLM, multimodal large language model; VLM, vision–language model; ViT, vision transformer; DLS, deep-learning system.

Model	Venue	Task & architecture	Training data	Weights	License
(a) Closed: weights not downloadable for independent testing					
Esteva et al. [6]	Nature 2017	Skin-cancer classification (CNN)	129K clinical photos (private)	✗	—
Haenssle et al. [12]	Ann. Oncol. 2018	Melanoma recognition (CNN)	>100K dermoscopy (private)	✗	—
ModelDerm (Build2017) [13]	J. Invest. Dermatol. 2018	Skin-tumour classification (CNN)	~19K clinical photos (private)	✗ (demo)	—
Brinker et al. [14]	Eur. J. Cancer 2019	Melanoma vs. nevus (CNN)	12K dermoscopy (public)	✗	—
Tschandl et al. [15]	Lancet Oncol. 2019	Pigmented lesion, 7-class (CNN)	10K dermoscopy (public)	✗	—
DERM (Skin Analytics) [16]	JAMA Netw. Open 2019	Melanoma detection (CNN, commercial)	~7K dermoscopy (private)	✗	—
Haenssle et al. [17]	Ann. Oncol. 2020	Multi-class lesions (CNN, commercial)	Undisclosed (private)	✗	—
Liu et al. (DLS) [9]	Nat. Med. 2020	Differential dx, 419-class (CNN)	16K telederm cases (private)	✗	—
ModelDerm (Build2020) [10]	J. Invest. Dermatol. 2020	Multi-class, 134 disorders (CNN)	220K clinical photos (private)	✗ (demo)	—
Google AI tool [18]	JAMA Netw. Open 2021	Teledermatology triage (CNN)	16K telederm cases (private)	✗	—
Soenksen et al. [19]	Sci. Transl. Med. 2021	Wide-field lesion triage (CNN)	34K crops (mixed)	✗	—
Combalia et al. [20]	Lancet Digit. Health 2022	Dermoscopy, 8-class (CNN)	25K dermoscopy (public)	✗	—
SkinGPT-4 [11]	Nat. Commun. 2024	Diagnostic dialogue (MLLM)	53K images (mixed)	✗ (withheld)	—
(b) Open-weight: weights downloadable, but data private and/or licence restrictive					
PanDerm [8]	Nat. Med. 2025	Multi-task foundation (ViT, SSL)	2M images (private)	✓	CC-BY-NC-ND
(c) Fully open: public training data and openly downloadable weights					
MONET [21]	Nat. Med. 2024	Concept annotation (CLIP)	105K images (public)	✓	CC-BY-NC-SA
DermLIP [22]	ICCV 2025	Zero-shot diagnosis (CLIP)	404K images (public)	✓	CC-BY-NC-ND
MAKE [23]	MICCAI 2025	Zero-shot diagnosis (CLIP)	404K images (public)	✓	MIT
UniDerm (ours)	This work	Multi-task foundation (VLM)	506K images (public)	✓	**Apache-2.0**

"Demo" marks systems testable only through a web demo, with no downloadable checkpoint. "Training data" gives the approximate training-corpus size and its public/private/mixed access. Here "mixed" denotes a public subset combined with an essential private cohort, so the model still cannot be independently retrained. The two ModelDerm rows are successive builds of the same system (Build2017 and Build2020). DermLIP and MAKE release weights that download without gating. DermLIP's model card carries inconsistent licence fields (a YAML field of CC-BY-4.0 versus a CC-BY-NC-ND-4.0 body note). We report the most restrictive (CC-BY-NC-ND). **Licence scope.** The **License** column refers to the model-weight licence under which each release distributes its trained parameters.

same private-cohort route. Privacy and consent constraints place such cohorts beyond the reach of most research groups. We envision a fully open dermatology foundation model, one whose model weights, training data and training recipe are all public, trained from publicly available image–text data alone and extensible through federated private cohorts when those become available. Full openness is not merely a matter of access, it is what makes the model transparent, so that any group can audit it for bias, reproduce every result and hold its behaviour accountable, which private-cohort systems structurally cannot allow. Because such a model is also openly accessible, the community can iteratively refine it on newly captured clinical photographs, emerging conditions and imaging devices, keeping it current with medical knowledge and steadily improving diagnostic reliability, particularly in low- and middle-income countries where dermatological expertise is scarcest.

In this paper we present UniDerm, a fully open dermatology foundation model trained from publicly available image–text data alone. Public dermatology data usually pair each image with several texts at once, *e.g.*, a clinical caption, a body site, a visual concept and a diagnosis. Standard contrastive learning cannot exploit these diverse signals. Because a single fixed image embedding must stay close to all of them at once, it behaves like a low-pass filter on the supervision: it keeps only the lowest-frequency component, the average across the texts, and discards the high-frequency detail that distinguishes one target from another. Seen from each target, this averaging injects noise: the signal that target carries is corrupted by the others sharing its embedding, washing out the distinct supervision each provides [24–26]. UniDerm resolves this one-image-many-text degradation through

supervision-denoising contrastive learning, which treats the heterogeneous texts paired with each image not as one blurred target but as distinct supervision signals to be recovered, and acts at two levels (model architecture in Fig. 8a,b, contrastive objective in Fig. 1c, Methods §4.2, §4.5). First, at the architectural level, the image and a natural-language clinical instruction are encoded jointly in a unified vision–language transformer. A different instruction therefore steers the same image toward a different, task-specific embedding. Second, at the objective level, each image-plus-instruction input is matched against the text that answers that specific instruction. The model therefore reads out the attribute the instruction asks for, recovering the differentiated, high-frequency information that plain averaging discards rather than collapsing every target onto its low-frequency average.

Evaluated on 11 public benchmarks (Methods §4.4), UniDerm demonstrates that a fully open model trained solely on public data can match or exceed the private-cohort state of the art across the clinical spectrum of dermatology, challenging the premise that clinical-grade dermatology AI requires private clinical cohorts. Where expert labels are scarce, it exceeds the private-cohort PanDerm [8] on skin-cancer and skin-condition diagnosis by 20.2 percentage points (46.8% relative improvement) and matches PanDerm's full-label accuracy using only ~30% of the labels (Fig. 1a and Fig. 2a). Without task-specific training it ranks first on every zero-shot diagnosis benchmark and on concept-to-image and cross-modal evidence search (Fig. 2b,c). In two independent reader studies, it exceeds a 302-reader consensus on malignancy detection (AUROC 0.946) and approaches the board-certified-dermatologist reference on the skin-tone-stratified DDI benchmark [5, 7], holding up across skin tones. These gains are independent of model scale: UniDerm-VE-lite, a 0.1-billion-parameter model distilled from UniDerm's visual encoder, exceeds PanDerm at a third of its size while running 3.1 times faster on four CPU threads with no GPU (Methods §4.9).

Full openness changes what such a model can be. Because UniDerm runs on modest computational resources, and needs no accelerator once trained, a clinic can deploy it on-site and adapt it locally, so that patient images never leave the institution. With its weights, code and data recipe all public, its results are independently reproducible and its behaviour auditable for bias, a level of scrutiny that the confidentiality constraints of private-cohort systems cannot readily support. And because the model is open, the community can continue to refine it on newly captured images, emerging conditions and imaging devices, allowing performance to keep advancing after release rather than freezing at the point of publication. These properties reframe clinical-grade dermatology AI as a resource that under-resourced communities can build and improve, rather than a proprietary product they can only purchase. Because the paradigm is not inherently specific to skin, it may point toward a broader route to specialist-level medical AI in other data-rich, expertise-scarce domains such as radiology and pathology.

Results

We evaluate UniDerm across the capabilities that the introduction identifies as deployment-critical (Fig. 1): label efficiency and diagnostic generalisation under limited or no supervision; the use of patient context at inference time; competitiveness with board-certified dermatologists across skin tones; and concept-to-image and cross-modal evidence search, linking images to clinical concepts and to their written descriptions.

For each capability we compare UniDerm against the leading recent models, grouped into three families. The first is the private-cohort dermatology foundation model PanDerm [8], the current state of the art for dermatology AI. The second is a set of large general-purpose vision foundation models, spanning roughly one to six billion parameters, that anchor the label-efficiency comparison: DINOv2-G [37], AIMv2-3B [38] and InternViT-6B [39]. The third is the set of publicly trained

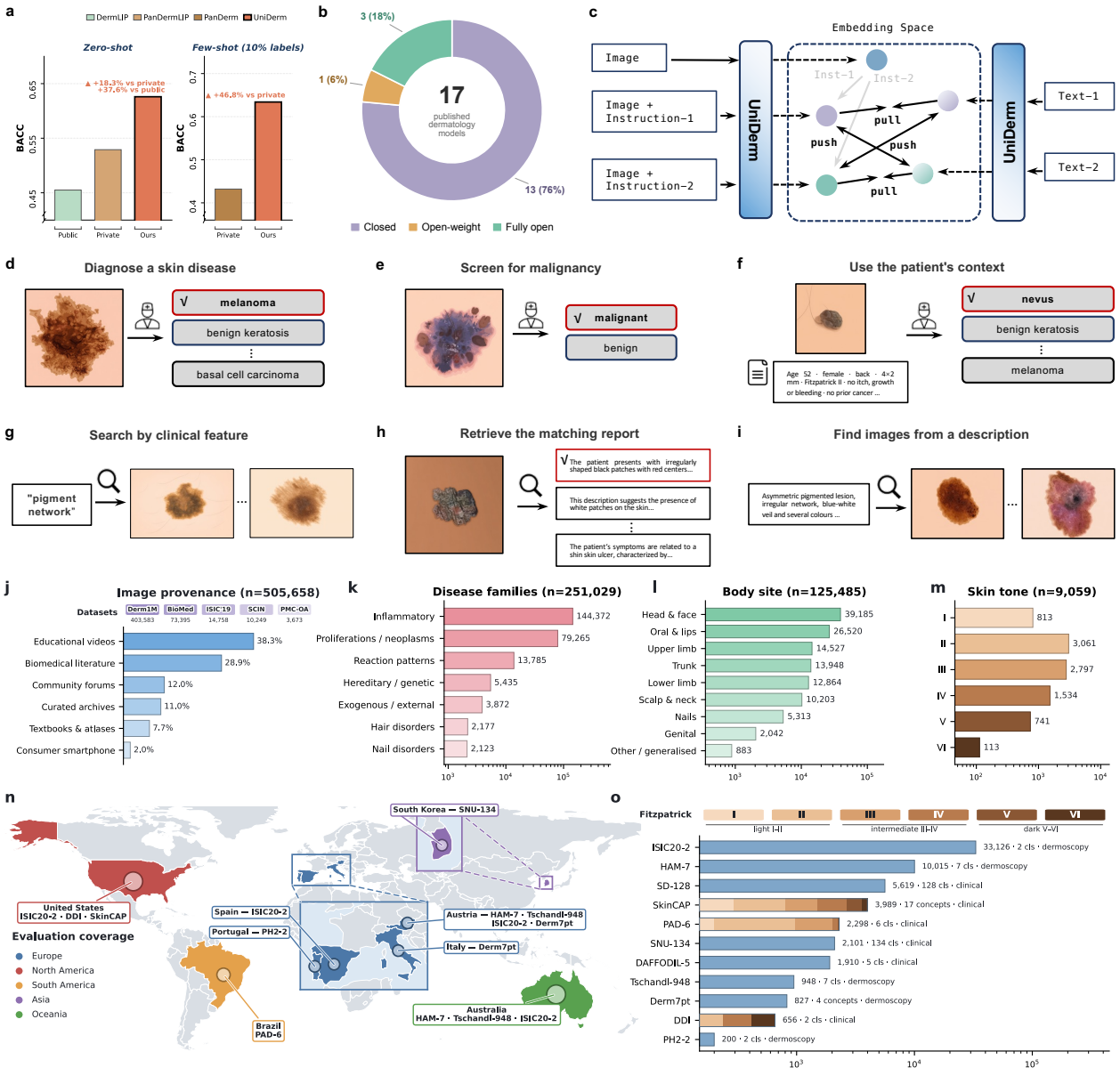


Figure 1 | **UniDerm: a fully open dermatology foundation model trained by supervision-denoising contrastive learning. a**, Mean diagnostic accuracy of the public-data DermLIP [22], private-cohort PanDermLIP [22] and UniDerm in strict zero-shot (left), and of DermLIP, private-cohort PanDerm [8] and UniDerm in 10%-label few-shot linear probing (right); UniDerm's relative gain over each comparator is annotated. **b**, Openness tiers of the 17 prior dermatology-diagnosis models (2017–2026; Extended Data Table ED1, UniDerm excluded): *Closed* (13/17), *Open-weight* (1/17, PanDerm) and *Fully open* (3/17; MONET [21], DermLIP, MAKE [23]), with UniDerm extending the fully open tier under Apache-2.0. **c**, Supervision-denoising contrastive learning: each image is paired with multiple system instructions (S_1, S_2) and each instruction-conditioned embedding is pulled toward its target text (T_1, T_2) and away from other in-batch targets (Methods §4.2, §4.5). **d–i**, Six clinical capabilities performed directly without task-specific training: **d**, *diagnose a skin disease*; **e**, *screen for malignancy*; **f**, *use the patient's context*, where a free-text caption (Fitzpatrick phototype, symptoms and history) refines the diagnosis; **g**, *search by clinical feature* (concept-to-image search); **h**, *retrieve the matching report*; **i**, *find images from a description*; **h**,**i** are the two directions of cross-modal evidence search. Quantitative results are in Fig. 2 and the Extended Data. **j–m**, Composition of the fully public training corpus (505,658 images; Methods §4.3, Extended Data Table ED17): **j**, the five source datasets by image count (top bar) and image provenance by content source; **k**, disease families; **l**, body site; **m**, Fitzpatrick skin tone. **k–m** are tallied over the labelled subset only (n per panel, well below the full corpus; $\log_{10}$ axes), with skin tone contributed predominantly by SCIN [27]. **n**,**o**, The held-out evaluation suite, disjoint from training (Methods §4.4): **n**, geographic coverage of the 11 benchmarks (five continents, eight documented countries); **o**, per-benchmark size ($\log_{10}$), diagnostic classes or annotated concepts, imaging modality and Fitzpatrick composition where recorded.

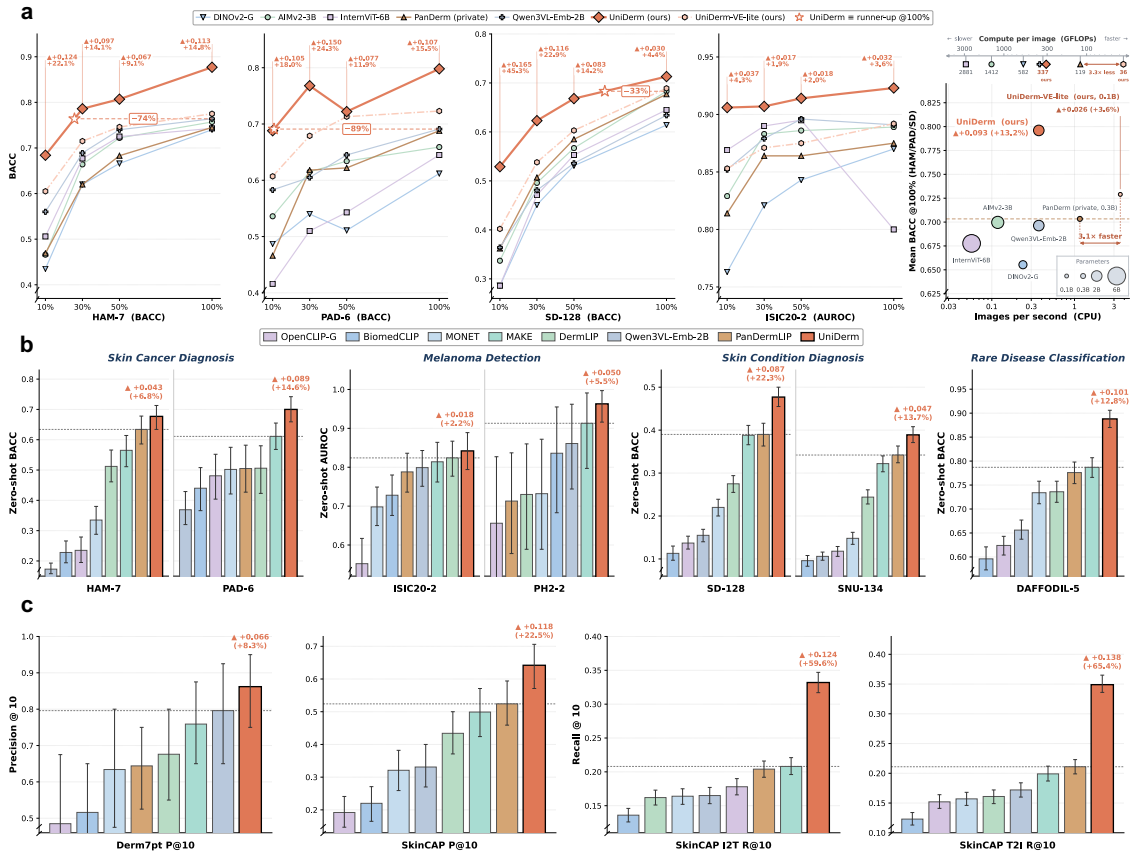


Figure 2 | **Benchmark performance of UniDerm across capabilities. a**, Few-shot linear probing on HAM-7 [28], PAD-6 [29], SD-128 [30] and ISIC20-2 [31] at 10/30/50/100% labels. The right-most panel plots 100%-label mean balanced accuracy over the three multiclass benchmarks against inference throughput (images per second, log scale), with bubble area proportional to total parameter count, for UniDerm (2 billion parameters), UniDerm-VE-lite (0.1 billion parameters, pruned and distilled from UniDerm's visual encoder; Methods §4.9), 1- to 6-billion general-domain models, Qwen3VL-Emb-2B [32] and PanDerm [8] (Methods §4.12). Throughput is measured with no GPU, on four CPU threads (Methods §4.10), the constrained hardware an under-resourced clinic would have to work from; it is a deliberate floor, and a single GPU lifts the same models above a hundred images per second. The horizontal dashed line marks the best comparator, PanDerm; the bracket below it reads off the throughput ratio between PanDerm and UniDerm-VE-lite, the only model in the panel that is both faster and more accurate than PanDerm. The ruler above the panel carries the hardware-independent half of that cost, the arithmetic each model performs on one 224-pixel image (GFLOPs, log scale, cheapest to the right; Methods §4.10). **b**, Zero-shot balanced accuracy on seven benchmarks in four task groups (multi-class skin-cancer diagnosis: HAM-7, PAD-6; binary melanoma detection: ISIC20-2, PH2-2 [33]; skin-condition diagnosis: SD-128, SNU-134 [10]; rare-disease classification: DAFFODIL-5 [34]). **c**, Concept-to-image search (P@10) on the Derm7pt melanoma–nevus subset [35] and SkinCAP [36], and cross-modal evidence search (R@10) on SkinCAP. Whiskers in **a**–**c** are 95% bootstrap confidence intervals; per-bar values are in Extended Data Tables ED2–ED4 for **a**,**b** and Extended Data Tables ED13–ED16 for **c**.

dermatology and general vision–language models, which we evaluate under matched protocols using each model's published image and text encoders (Methods §4.8): BiomedCLIP [40], OpenCLIP-G [41], MONET [21], DermLIP [22], MAKE [23] and Qwen3VL-Emb-2B [32]. We additionally evaluate PanDermLIP [22], a private-cohort CLIP variant that pairs the DermLIP framework with the private-cohort PanDerm vision encoder (Methods §4.12). Short names follow the mapping in Methods §4.12.

2.1. Matching private-cohort foundation models at low expert-label cost

We first ask whether a publicly trained model can match the diagnostic accuracy of models trained on private cohorts, and at how low a label cost it can do so. Because expert, biopsy-confirmed annotation is the binding constraint in dermatology, we probe UniDerm's frozen embeddings across the label budget: at small training-label fractions, with no labels at all, and at full supervision. For few-shot linear probing we fit a probe at four label fractions (10%, 30%, 50%, 100%) across four benchmarks: HAM-7 (7 classes, dermoscopic images) [28], PAD-6 (6 classes, smartphone images) [29], SD-128 (128 long-tail classes) [30] and ISIC20-2 (binary melanoma detection) [31]. We report balanced accuracy (BACC) on the three multiclass benchmarks and binary AUROC on ISIC20-2, whose ~2% melanoma prevalence pins BACC to the majority class. The full grid is given in Fig. 2a and Extended Data Table ED2; for zero-shot we score prompt-only diagnosis across seven benchmarks (Fig. 2b).

Matching private-cohort foundation models with ~70% fewer expert labels. The label-efficiency picture is especially favourable to UniDerm (Fig. 2a and Extended Data Table ED2). At only 10% of training labels, the three-benchmark mean BACC of UniDerm is 0.634, exceeding the private-cohort PanDerm at the same label fraction (0.432) by +20.2 percentage points, a +46.8% relative gain. It exceeds every large general-domain comparator by +18.8 to +23.1 percentage points (AIMv2-3B 0.446, DINOv2-G and InternViT-6B 0.403), a +42.1 to +57.3% relative gain in this label-scarce regime. At 30% of training labels, UniDerm reaches 0.726, already exceeding 100%-label PanDerm (0.703) on the mean and on HAM-7 (0.786 vs. 0.745). At 50% of training labels, UniDerm reaches 0.732, further surpassing 100%-label PanDerm on the three-benchmark mean. In dermatology the binding constraint is expert, biopsy-confirmed annotation, so matching the private-cohort reference's full-label accuracy from roughly a third of the public labels, and outscoring it at every matched fraction, is what makes such an embedding practical to adapt to a new clinic, imaging device or locally prevalent condition rather than merely competitive on a benchmark.

Exceeding the private-cohort foundation model at zero label cost. At the zero-label extreme, a deployable model must recognise disease from the image alone, with no labelled examples from the target clinic. In strict zero-shot diagnosis across the four task families (Fig. 2b): multi-class skin-cancer diagnosis (HAM-7, PAD-6), binary melanoma detection (ISIC20-2, PH2-2), general skin-condition diagnosis (SD-128, SNU-134) and rare-disease classification (DAFFODIL-5). UniDerm ranks first on every one of the seven benchmarks, ahead of both the public-data comparators and the private-cohort PanDermLIP [22]. This follows directly from resolving the one-image-many-text limitation by supervision denoising: a model that no longer averages over heterogeneous text targets is a stronger zero-shot classifier with no new image-side supervision.

The five benchmarks scored by balanced accuracy span multi-class skin-cancer diagnosis (HAM-7, PAD-6), general skin-condition diagnosis (SD-128, SNU-134) and rare-disease classification (DAFFODIL-5). Across them (Extended Data Table ED3), UniDerm reaches a mean BACC of 0.626, +0.091 above the leading public-data comparator MAKE [23] (0.535, a +17.0% relative gain) and +0.097 above the private-cohort PanDermLIP (0.529). The remaining comparators fall much further back: DermLIP [22] 0.455, MONET [21] 0.388 and BiomedCLIP [40] 0.295, which UniDerm exceeds

by +112.2% in relative terms. The two binary melanoma-detection benchmarks (PH2-2, ISIC20-2) are scored by AUROC; across them (Extended Data Table ED4) UniDerm again leads at mean AUROC 0.903, ahead of MAKE (0.863) and PanDermLIP (0.751). The contrast with the linear-probing picture is informative. An embedding this class-discriminative needs only a small linear probe, or none at all under prompt-only inference, to reach a state-of-the-art result on each benchmark.

Diagnostic accuracy with full expert labels. For *multi-class skin-cancer diagnosis*, the everyday task of assigning a lesion to its specific diagnostic category, UniDerm leads every comparator at 100% labels. On HAM-7 [28] it reached BACC 0.877 (95% CI 0.854–0.901), exceeding the private-cohort PanDerm [8] (0.745) by +13.2 percentage points and the leading large general-domain comparator AIMv2-3B [38] (0.757) by +12.0 percentage points; on the six-class PAD-6 [29] it reached 0.798 (0.757–0.836), +11.0 percentage points above PanDerm (0.688) and +13.9 and +15.3 percentage points above the much larger AIMv2-3B (0.659) and InternViT-6B [39] (0.645), a 21.1 to 23.7% relative gain over the latter two. For *general skin-condition diagnosis*, tested on the fine-grained 128-class long-tail SD-128 [30], UniDerm reached 0.713 (0.690–0.736), ahead of AIMv2-3B (0.683, +3.0 percentage points) and PanDerm (0.677, +3.6 percentage points). Averaged over these three diagnostic benchmarks UniDerm reached BACC 0.796 versus PanDerm 0.703 (+9.3 percentage points), with every comparator trailing well behind (next-best PanDerm 0.703, then AIMv2-3B 0.700, Qwen3VL-Emb-2B 0.696, InternViT-6B 0.678 and DINOv2-G [37] 0.655, Extended Data Table ED2e). For *binary melanoma detection*, scored by AUROC on the class-imbalanced ISIC20-2 [31], UniDerm leads outright at 0.923, ahead of the pruned UniDerm-VE-lite (0.892), Qwen3VL-Emb-2B (0.891), AIMv2-3B (0.889) and PanDerm (0.875), confirming that the embedding stays competitive on this high-imbalance melanoma slice.

Exceeding PanDerm at a third of its size. The headline comparison above uses the full 2.07-billion-parameter UniDerm, roughly seven times the parameter count of the ViT-L PanDerm, which leaves backbone scale unisolated. UniDerm-VE-lite settles that question from below. It is UniDerm's visual encoder structurally pruned to 0.1 billion parameters, a third of PanDerm's 0.3 billion, and then distilled against that unpruned encoder on unlabelled public images, no diagnostic label and no image beyond the public training corpus enters either step (Methods §4.9). Under matched protocols (Fig. 2a, Extended Data Table ED2) it exceeds PanDerm on all four benchmarks at all four label fractions. At 100% training labels it reaches BACC 0.775 on HAM-7 (vs. PanDerm 0.745), 0.723 on PAD-6 (vs. 0.688) and 0.689 on SD-128 (vs. 0.677), giving a three-benchmark mean of 0.729 versus PanDerm's 0.703. The margins widen as labels grow scarce: at the 10% few-shot level it reaches 0.605 on HAM-7 and 0.607 on PAD-6, +13.6 and +14.1 percentage points above PanDerm, and 0.538 on the three-benchmark mean against 0.432 (+10.6 percentage points). A 0.1-billion-parameter model fine-tuned on publicly available image–text data with supervision-denoising contrastive learning therefore clears the private-cohort foundation-model band with a third of its parameters, which places the lead reported above with the training recipe rather than with backbone scale.

Accuracy against the cost of running the model. A frozen embedding is only deployable if a clinic can afford to compute it, so the right-most panel of Fig. 2a sets accuracy against both axes of that cost. The score is the mean balanced accuracy over the three multiclass benchmarks (HAM-7, PAD-6, SD-128) at 100% labels. The horizontal axis is throughput on four CPU threads with no GPU, the hardware floor for a clinic without dedicated accelerators, and bubble area is total parameter count. The ruler above the panel gives the arithmetic each model performs on one image, the part of that cost no choice of hardware can change (Methods §4.10). The throughput axis is a deliberately punishing floor rather than a working speed. With a single H20 GPU, the same 224-pixel input is encoded at more than one hundred images per second by full UniDerm, two orders of magnitude above the figures plotted here. The full 2-billion-parameter UniDerm attains the leading score, ahead

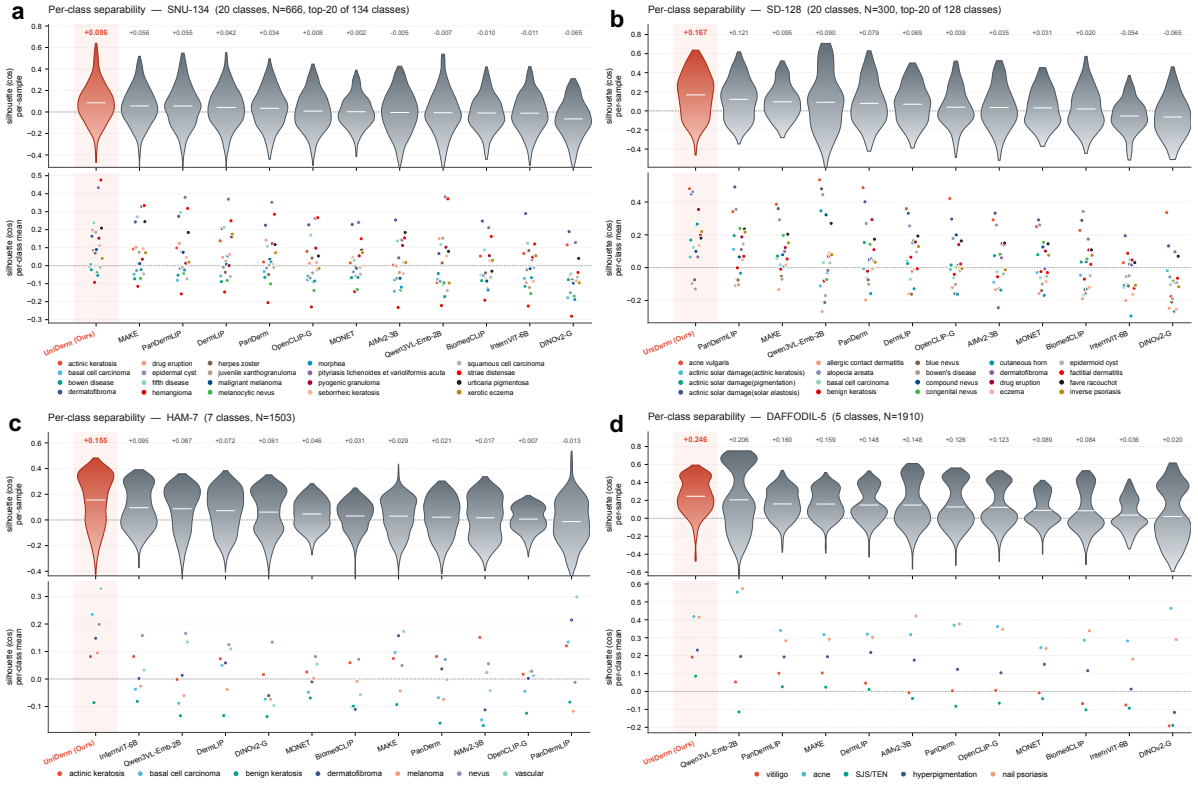


Figure 3 | **Embedding-space class structure measured by per-sample silhouette score (cosine).** Silhouette is computed in each model's native embedding space. Top sub-panel: per-sample distribution as a violin per model (mean annotated, white median bar). Bottom: per-class mean silhouette, one dot per disease class. Models are sorted by descending mean silhouette; UniDerm is in red. **a**, SNU-134 [10]. **b**, SD-128 [30]. **c**, HAM-7 [28]. **d**, DAFFODIL-5 [34].

of the next-best comparator, the private-cohort PanDerm, by 0.093. It runs its language tower as well and encodes 0.37 images per second on four threads, 3.1 and 6.5 times faster than AIMv2-3B [38] and InternViT-6B [39] under the same constraint. UniDerm-VE-lite then converts its smaller size into an advantage on both axes of cost. It scores 0.729, second only to UniDerm and 0.026 above PanDerm, and outscores every 1- to 6-billion-parameter general-domain vision model (AIMv2-3B 0.700, InternViT-6B 0.678, DINOv2-G [37] 0.655). It reaches that score on 35.9 GFLOPs per image, 3.3 times less arithmetic than PanDerm and 16 to 80 times less than those three, and runs at 3.58 images per second, 3.1 times faster than PanDerm and fifteen times faster than DINOv2-G, the quickest general-domain model. It is the only model in the panel that is both faster and more accurate than the private-cohort system. A publicly trained 0.1-billion-parameter model thus surpasses general-domain models more than fifty times its size, and the private-cohort PanDerm, at the lowest cost per image in the comparison.

Embedding-space class separability. The quality of a frozen representation is fixed before any classifier is fitted. The more cleanly it separates disease classes in its raw space, the less work a downstream linear head has to do, which is what underwrites the label efficiency reported above. We measure this geometry directly with the per-sample silhouette score (cosine, Fig. 3a–d), a head-free, non-parametric measure of how tightly each image sits with its own class relative to the nearest other class. UniDerm leads on mean silhouette across all four benchmarks (HAM-7 [28], SD-128 [30], SNU-134 [10] and the rare-disease DAFFODIL-5 [34]), with the widest margin on HAM-7 and a smaller but consistent lead on the other three. The per-class panels show the gain is broad-based, lifting class-mean silhouettes across the class spectrum.

Table ED2 | **Few-shot linear-probing balanced accuracy at** $\{10, 30, 50, 100\}$% **training-label fractions for HAM-7 [28], PAD-6 [29], SD-128 [30] (BACC) and ISIC20-2 [31] (binary AUROC), across seven models.** The model set matches Fig. 2a; our two models are listed last, in ascending order of accuracy. Cells are mean (lo–hi) of a 1,000-resample percentile bootstrap; best per column in **bold**. All models use the same probe on the same frozen embeddings, with inverse-regularisation $C = 1$ for the multiclass panels (a)–(c) and $C = 0.01$ for the binary panel (d) (Methods §4.8). Panel (e) is the per-model, per-fraction mean of (a)–(c) (no separate bootstrap interval); each comparator cell gives UniDerm's lead in parentheses (absolute, and relative to the cell). UniDerm and UniDerm-VE-lite are marked Best and Second. UniDerm-VE-lite is UniDerm's visual encoder (its vision tower taken on its own, hence the *-VE* suffix) structurally pruned to 0.1 billion parameters and distilled against that unpruned encoder without labels (Methods §4.9).

Model	10%	30%	50%	100%
(a) HAM-7 (BACC)				
DINOv2-G	0.435 (0.390–0.486)	0.620 (0.560–0.676)	0.666 (0.607–0.719)	0.740 (0.687–0.788)
AIMv2-3B	0.466 (0.427–0.507)	0.664 (0.613–0.714)	0.722 (0.670–0.773)	0.757 (0.706–0.809)
InternViT-6B	0.506 (0.465–0.547)	0.677 (0.623–0.732)	0.725 (0.674–0.777)	0.743 (0.693–0.796)
PanDerm *(ViT-L, private-cohort)*	0.469 (0.427–0.510)	0.620 (0.566–0.672)	0.683 (0.632–0.736)	0.745 (0.696–0.794)
Qwen3VL-Emb-2B	0.560 (0.510–0.609)	0.689 (0.642–0.737)	0.740 (0.690–0.786)	0.764 (0.716–0.809)
UniDerm-VE-lite *(ours)*	0.605 (0.555–0.654)	0.715 (0.665–0.763)	0.746 (0.698–0.790)	0.775 (0.727–0.815)
UniDerm (ours)	**0.684 (0.630–0.735)**	**0.786 (0.739–0.829)**	**0.807 (0.762–0.849)**	**0.877 (0.854–0.901)**
(b) PAD-6 (BACC)				
DINOv2-G	0.487 (0.424–0.556)	0.540 (0.475–0.614)	0.511 (0.471–0.555)	0.612 (0.537–0.683)
AIMv2-3B	0.536 (0.458–0.606)	0.615 (0.538–0.691)	0.634 (0.556–0.708)	0.659 (0.581–0.729)
InternViT-6B	0.416 (0.376–0.458)	0.510 (0.473–0.550)	0.543 (0.480–0.614)	0.645 (0.564–0.724)
PanDerm *(ViT-L, private-cohort)*	0.466 (0.398–0.539)	0.618 (0.542–0.693)	0.622 (0.549–0.701)	0.688 (0.623–0.749)
Qwen3VL-Emb-2B	0.583 (0.512–0.652)	0.605 (0.534–0.677)	0.645 (0.575–0.716)	0.691 (0.649–0.732)
UniDerm-VE-lite *(ours)*	0.607 (0.538–0.673)	0.679 (0.614–0.738)	0.713 (0.653–0.769)	0.723 (0.663–0.776)
UniDerm (ours)	**0.688 (0.649–0.726)**	**0.768 (0.727–0.809)**	**0.722 (0.658–0.778)**	**0.798 (0.757–0.836)**
(c) SD-128 (BACC)				
DINOv2-G	0.286 (0.264–0.308)	0.451 (0.424–0.478)	0.531 (0.503–0.558)	0.614 (0.589–0.639)
AIMv2-3B	0.337 (0.314–0.361)	0.496 (0.470–0.523)	0.567 (0.539–0.592)	0.683 (0.659–0.707)
InternViT-6B	0.286 (0.264–0.308)	0.471 (0.444–0.499)	0.553 (0.526–0.579)	0.645 (0.619–0.670)
PanDerm *(ViT-L, private-cohort)*	0.362 (0.337–0.386)	0.507 (0.481–0.532)	0.585 (0.559–0.610)	0.677 (0.652–0.702)
Qwen3VL-Emb-2B	0.364 (0.341–0.389)	0.481 (0.457–0.505)	0.536 (0.511–0.561)	0.634 (0.609–0.658)
UniDerm-VE-lite *(ours)*	0.402 (0.380–0.426)	0.538 (0.515–0.563)	0.603 (0.580–0.627)	0.689 (0.667–0.712)
UniDerm (ours)	**0.529 (0.506–0.554)**	**0.623 (0.601–0.646)**	**0.668 (0.645–0.692)**	**0.713 (0.690–0.736)**
(d) ISIC20-2 (binary AUROC)				
DINOv2-G	0.763 (0.715–0.807)	0.821 (0.780–0.856)	0.843 (0.804–0.878)	0.870 (0.837–0.898)
AIMv2-3B	0.829 (0.794–0.863)	0.883 (0.855–0.907)	0.886 (0.860–0.910)	0.889 (0.861–0.912)
InternViT-6B	0.869 (0.835–0.900)	0.890 (0.861–0.916)	0.895 (0.865–0.918)	0.800 (0.760–0.838)
PanDerm *(ViT-L, private-cohort)*	0.814 (0.774–0.850)	0.864 (0.832–0.893)	0.864 (0.832–0.892)	0.875 (0.846–0.901)
Qwen3VL-Emb-2B	0.852 (0.818–0.885)	0.879 (0.848–0.908)	0.896 (0.867–0.923)	0.891 (0.862–0.919)
UniDerm-VE-lite *(ours)*	0.853 (0.815–0.890)	0.871 (0.836–0.905)	0.875 (0.840–0.908)	0.892 (0.861–0.921)
UniDerm (ours)	**0.906 (0.878–0.932)**	**0.907 (0.877–0.935)**	**0.914 (0.887–0.938)**	**0.923 (0.900–0.944)**
(e) Three-benchmark mean of (a)–(c): HAM-7 + PAD-6 + SD-128 (BACC)				
DINOv2-G	0.403 (+0.231, +57.3%)	0.537 (+0.189, +35.1%)	0.569 (+0.163, +28.6%)	0.655 (+0.141, +21.5%)
AIMv2-3B	0.446 (+0.188, +42.1%)	0.592 (+0.134, +22.6%)	0.641 (+0.091, +14.1%)	0.700 (+0.096, +13.7%)
InternViT-6B	0.403 (+0.231, +57.3%)	0.553 (+0.173, +31.2%)	0.607 (+0.125, +20.5%)	0.678 (+0.118, +17.4%)
PanDerm *(ViT-L, private-cohort)*	0.432 (+0.202, +46.8%)	0.582 (+0.144, +24.7%)	0.630 (+0.102, +16.1%)	0.703 (+0.093, +13.2%)
Qwen3VL-Emb-2B	0.502 (+0.132, +26.2%)	0.592 (+0.134, +22.6%)	0.640 (+0.092, +14.3%)	0.696 (+0.100, +14.3%)
UniDerm-VE-lite *(ours)*	0.538 (Second)	0.644 (Second)	0.687 (Second)	0.729 (Second)
UniDerm (ours)	**0.634** (Best)	**0.726** (Best)	**0.732** (Best)	**0.796** (Best)

Table ED3 | **Strict zero-shot balanced accuracy on the five multiclass benchmarks: HAM-7 [28], PAD-6 [29], SD-128 [30], SNU-134 [10] and DAFFODIL-5 [34] (BACC).** Values are mean (lower–upper) of a 1,000-resample percentile bootstrap, best per column in **bold**. The **Mean** column is the per-model average over the five benchmarks (point value, no interval); its parenthesis gives UniDerm's lead over that model (absolute, and relative to the cell), with UniDerm marked Best. Ordering follows Fig. 2b (UniDerm last); inference and prompt configurations in Methods §4.8. The two binary AUROC benchmarks are in Extended Data Table ED4.

Model	HAM-7	PAD-6	SD-128	SNU-134	DAFFODIL-5	Mean
BiomedCLIP	0.228 (0.194–0.266)	0.440 (0.366–0.508)	0.113 (0.097–0.130)	0.096 (0.083–0.108)	0.596 (0.573–0.621)	0.295 (+0.331, +112.2%)
MONET	0.335 (0.288–0.380)	0.502 (0.421–0.575)	0.220 (0.200–0.239)	0.148 (0.134–0.162)	0.734 (0.711–0.758)	0.388 (+0.238, +61.3%)
MAKE	0.565 (0.511–0.614)	0.611 (0.568–0.655)	0.388 (0.366–0.411)	0.322 (0.303–0.340)	0.787 (0.766–0.807)	0.535 (+0.091, +17.0%)
DermLIP	0.512 (0.461–0.566)	0.506 (0.423–0.580)	0.275 (0.255–0.294)	0.244 (0.228–0.261)	0.736 (0.714–0.758)	0.455 (+0.171, +37.5%)
OpenCLIP-G	0.235 (0.195–0.279)	0.481 (0.404–0.552)	0.137 (0.123–0.153)	0.118 (0.106–0.129)	0.624 (0.604–0.643)	0.319 (+0.307, +96.2%)
Qwen3VL-Emb-2B	0.173 (0.158–0.193)	0.369 (0.320–0.429)	0.155 (0.140–0.169)	0.106 (0.097–0.116)	0.656 (0.637–0.677)	0.292 (+0.334, +114.3%)
PanDermLIP	0.634 (0.586–0.678)	0.505 (0.427–0.582)	0.390 (0.363–0.416)	0.342 (0.324–0.363)	0.776 (0.753–0.798)	0.529 (+0.097, +18.3%)
UniDerm (ours)	**0.677 (0.634–0.713)**	**0.700 (0.659–0.742)**	**0.477 (0.455–0.500)**	**0.389 (0.371–0.408)**	**0.888 (0.870–0.906)**	**0.626** (Best)

Table ED4 | **Strict zero-shot binary melanoma detection on the two AUROC benchmarks: PH2-2 [33] and ISIC20-2 [31] (AUROC).** Values are mean (lower–upper) of a 1,000-resample percentile bootstrap, best per column in **bold**. The **Mean** column is the per-model average of the two AUROCs (point value); its parenthesis gives UniDerm's lead (absolute, and relative to the cell), with UniDerm marked Best. Companion to the five balanced-accuracy benchmarks in Extended Data Table ED3; models, ordering and protocol as there.

Model	PH2-2	ISIC20-2	Mean
BiomedCLIP	0.836 (0.683–0.955)	0.728 (0.676–0.780)	0.782 (+0.121, +15.4%)
MONET	0.732 (0.589–0.872)	0.698 (0.650–0.749)	0.715 (+0.188, +26.2%)
MAKE	0.913 (0.797–0.991)	0.814 (0.762–0.864)	0.863 (+0.040, +4.6%)
DermLIP	0.730 (0.589–0.860)	0.824 (0.777–0.867)	0.777 (+0.126, +16.2%)
OpenCLIP-G	0.656 (0.471–0.827)	0.552 (0.486–0.617)	0.604 (+0.299, +49.5%)
Qwen3VL-Emb-2B	0.861 (0.744–0.962)	0.799 (0.751–0.843)	0.830 (+0.073, +8.7%)
PanDermLIP	0.713 (0.578–0.837)	0.788 (0.736–0.836)	0.751 (+0.152, +20.2%)
UniDerm (ours)	**0.963 (0.916–0.997)**	**0.842 (0.794–0.889)**	**0.903** (Best)

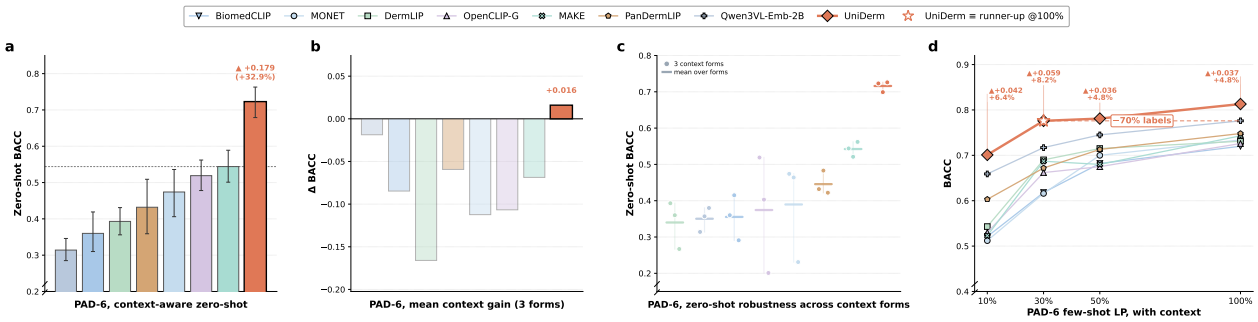


Figure 4 | **Patient-context-conditioned diagnosis on PAD-6 [29].** Per-case clinical metadata are rendered into a patient-context caption supplied to the model alongside the lesion image and compared with a matched no-context setting under the same image-side fusion. Panels **a**, **b** and **d** use the simplest form, a compact field list (Methods §4.8). **a**, Zero-shot BACC with the patient-context caption; the dashed line marks the leading comparator and the bold "▲" gives UniDerm's absolute and relative gain over that runner-up. **b**, Per-model Δ zero-shot BACC averaged over the three caption forms (mean over forms minus no context); the positive bar is in the model's colour, negative bars are dimmed. UniDerm is the only model with a positive mean Δ (+0.016); every comparator degrades on average. **c**, Robustness to the *form* of the context: zero-shot BACC under three caption forms of increasing verbosity (field list, natural prose and a templated clinical report; example forms in Extended Data Table ED5, per-form accuracies in Extended Data Table ED6); per model the three markers are the three forms, the bar their mean and the thin line their range. **d**, Linear-probing BACC versus training-label fraction with the patient-context caption; the dashed line marks the runner-up's 100%-label level and the white star the fraction (~30%) at which UniDerm crosses it. Zero-shot values are in Extended Data Table ED6, few-shot values with 95% bootstrap confidence intervals in Extended Data Table ED7.

2.2. Reading the whole patient case, the way a clinician does, sharpens diagnosis

A dermatologist never reads an image in isolation. Age, site, symptoms and history all shape the diagnosis, and a clinically useful model should sharpen its reading when given this patient context.

We tested this on PAD-6 [29], whose per-case clinical fields (age, sex, anatomical site, lesion size, Fitzpatrick phototype, symptom flags and personal and family history) we render into a compact patient-context caption. Because UniDerm is a unified transformer, the lesion image and the caption enter as one shared token stream and fuse into a single context-conditioned embedding in one forward pass. The two-tower CLIP-family comparators cannot perform this fusion, so for them the caption is encoded separately and averaged with the image embedding (mean-fusion, Methods §4.8). Unless stated otherwise, the zero-shot and few-shot evaluations both use the simplest of these renderings, a compact field list (Extended Data Table ED5). Only the robustness analysis below sweeps the two longer forms.

Patient context sharpens zero-shot diagnosis. Supplying the patient-context caption moves UniDerm from BACC 0.700 (no context) to 0.723 under the field-list form, a +2.3 percentage-point gain, whereas the same caption *lowers* the leading CLIP-family comparator MAKE [23] from 0.611 to 0.544 under mean-fusion, leaving UniDerm +0.179 BACC ahead of the best comparator with context (Fig. 4a). This gain is not generic. Averaged over the three caption forms, the patient-context caption *lowers* the zero-shot BACC of all seven comparators (BiomedCLIP [40], MONET [21], DermLIP [22], OpenCLIP-G, MAKE, PanDermLIP and Qwen3VL-Emb-2B [32]), by between 0.019 and 0.166 BACC, because mean-fusion dilutes the image-side similarity or injects noisy text-side terms. UniDerm is the only model whose accuracy *rises*: its single-stream fusion turns the same metadata into a net +0.016 BACC gain averaged across the three forms (Fig. 4b).

Robustness to the form of the context. The advantage does not depend on how the context is phrased. Across three caption forms that render the same metadata at increasing verbosity, namely a field list, natural prose and a templated clinical report (Extended Data Table ED5), UniDerm's zero-shot BACC stays within a narrow 0.699–0.726 band, so that its worst caption form still exceeds every comparator's best form (0.562). The comparators instead swing by up to 0.318 BACC and several collapse under a single phrasing (for example OpenCLIP-G ranges 0.201–0.519) (Fig. 4c and Extended Data Table ED6).

Patient context compounds under few-shot probing. With the patient-context caption, UniDerm's linear-probing curve reaches BACC 0.701, 0.776, 0.781 and 0.813 at the 10/30/50/100% label fractions. Its gain over the matched no-context curve ranges from +0.8 to +5.9 percentage points across the four fractions, and it matches the leading comparator's full-label accuracy using only ~30% of the labels (Fig. 4d and Extended Data Table ED7). Across these settings UniDerm reads the whole case the way a clinician does: the structured patient context that two-tower systems can only bolt on after the fact, and that degrades most of them, becomes through single-stream fusion a net gain in diagnostic accuracy at no additional labelling cost.

2.3. Reaching board-certified-dermatologist performance equitably across skin tones

We next test whether the embedding-level advantages above translate to reader-level diagnostic performance. To avoid the self-organized-reader-cohort concern that affects in-house reader studies, all reader comparisons in this submission are against *independently published* reader references under matched protocols.

Two dermatology benchmarks in our suite have published large-scale independent reader studies at matching label granularity; we compare UniDerm with those published reader results on the same test sets, without recruiting our own in-house reader cohort.

Approaching human experts, and human–AI collaboration. On this 948-image subset the average of the 302-reader panel (a mix of board-certified dermatologists, dermatology residents and

Table ED5 | **The three patient-context caption forms evaluated on PAD-6, shown in full for one case** (a 52-year-old female, lesion on the back, 4 × 2 mm, Fitzpatrick phototype II, all symptom and history flags negative). The same 17 of the 22 patient clinical metadata fields in PAD-UFES-20 are rendered three ways of increasing verbosity (the diagnostic label, biopsy status, parental-background fields, household-infrastructure fields and image identifiers are withheld), all deterministic string templates (no model in the loop, fully reproducible). Panels **a**,**b**,**d** of Fig. 4 use the field-list form. Panel **c** sweeps all three (per-form accuracies in Extended Data Table ED6). Examples are complete (not excerpted). In the field-list form the original line breaks are shown as semicolons.

Field list deterministic template

Patient context: Age: 52; Sex: female; Region: back; Diameter: 4x2 mm; Fitzpatrick phototype: II; Itching: no; Growth: not reported; Pain: no; Change: not reported; Bleeding: no; Elevation: no; Smoker: no; Alcohol use: no; Pesticide exposure: no; Personal skin cancer history: no; Family cancer history: no

Natural prose deterministic template

The patient is a 52-year-old female with Fitzpatrick phototype II, presenting with a cutaneous lesion on the back. The lesion measures approximately 4 × 2 mm. The patient denies pruritus / itching, pain, bleeding, elevation above skin surface. No history of smoking, alcohol use, pesticide exposure, personal history of skin cancer, family history of cancer.

Templated clinical report deterministic template

Patient Demographics and Phototype. The patient is a 52-year-old female individual with Fitzpatrick skin phototype II. Fitzpatrick phototype provides important context regarding baseline melanin content and susceptibility to ultraviolet-induced cutaneous changes, which informs the interpretation of pigmented and erythematous lesions. **Lesion Anatomic Site and Dimensions.** The cutaneous lesion is localized to the back region, which is a clinically relevant anatomic site for risk stratification. The lesion's recorded dimensions are approximately 4 mm along the long axis and 2 mm along the short axis. Diameter is a primary morphological descriptor used to monitor growth and assess clinical significance. **Symptom Profile.** The patient explicitly denies: pruritus / itching; pain at the lesion site; spontaneous or contact bleeding; elevation above the surrounding skin. Absence of these features is recorded for completeness. **Patient History and Exposures.** Negative history includes tobacco use (smoking); alcohol consumption; occupational or environmental pesticide exposure; personal history of cutaneous malignancy; family history of malignancy, recorded to characterize the patient's risk profile. **Integrated Clinical Summary.** The clinical context comprises a 52-year-old female patient of Fitzpatrick phototype II, with a 4 × 2 mm cutaneous lesion on the back.

Table ED6 | **Robustness of context-conditioned zero-shot BACC to the caption form on PAD-6 [29] (source data for Fig. 4c).** Zero-shot balanced accuracy (mean with 95% bootstrap CI, 1,000 resamples) with no context and with each of the three caption forms of Extended Data Table ED5. The penultimate column reports the across-form (min, max) of BACC and the last column its spread, max – min across the three forms (smaller is more robust, ↓). UniDerm row in **bold**.

Model	No context	Field list	Natural prose	Templated clinical report	Across-form (min, max)	Spread (max – min) ↓
BiomedCLIP	0.440 (0.366–0.508)	0.360 (0.310–0.419)	0.415 (0.385–0.446)	0.291 (0.247–0.347)	(0.291, 0.415)	0.124
MONET	0.502 (0.421–0.575)	0.474 (0.406–0.536)	0.464 (0.426–0.501)	0.231 (0.204–0.262)	(0.231, 0.474)	0.243
DermLIP	0.506 (0.423–0.580)	0.393 (0.356–0.431)	0.360 (0.278–0.435)	0.267 (0.207–0.318)	(0.267, 0.393)	0.126
OpenCLIP-G	0.481 (0.404–0.552)	0.519 (0.478–0.562)	0.403 (0.325–0.474)	0.201 (0.180–0.224)	(0.201, 0.519)	0.318
MAKE	0.611 (0.568–0.655)	0.544 (0.501–0.589)	0.521 (0.480–0.563)	0.562 (0.519–0.606)	(0.521, 0.562)	0.041
PanDermLIP	0.505 (0.427–0.582)	0.432 (0.359–0.509)	0.483 (0.411–0.546)	0.422 (0.345–0.495)	(0.422, 0.483)	0.061
Qwen3VL-Emb-2B	0.369 (0.320–0.429)	0.314 (0.285–0.346)	0.357 (0.289–0.427)	0.380 (0.344–0.419)	(0.314, 0.380)	0.066
UniDerm (ours)	**0.700** (0.659–0.742)	**0.723** (0.679–0.763)	**0.699** (0.640–0.754)	**0.726** (0.683–0.770)	(**0.699**, **0.726**)	**0.027**

general practitioners) reaches macro-mean accuracy 0.643, ranging from 0.340 on actinic keratosis / intraepithelial carcinoma to 0.829 on vascular lesions [5]. UniDerm clears the unaided-rater bar without a single expert label: its zero-shot HAM-7 macro-mean BACC is 0.703 (95% CI 0.673–0.735), +6.0 percentage points above the average rater ($p < 0.001$), though below the AI-assisted readers (0.776). Fitting the full HAM-7 expert labels lifts it to 0.763 (0.725–0.801, Extended Data Table ED10), still short of the AI-assisted readers but well above the unaided rater (per-class values in Fig. 5b and Extended Data Table ED11). The picture is stronger for binary any-malignancy detection (Fig. 5a). UniDerm's zero-shot ROC (AUC 0.917) sits level with the reconstructed reader-consensus ROC, and at 100% labels its ROC (AUC 0.946) exceeds that consensus on both axes, reaching sensitivity 0.93 (95% CI 0.90–0.97) at the reader's specificity and specificity 0.93 (0.90–0.96) at the reader's sensitivity (Extended Data Table ED9). Against a reader-consensus ROC reconstructed from the per-image votes

Table ED7 | **Patient-context-conditioned few-shot linear probing on PAD-6 [29] with the headline patient-context caption (the compact structured field-list form), across eight models.** Few-shot linear-probing balanced accuracy at {10, 30, 50, 100}% training-label fractions, each as mean (lower–upper) of a 1,000-resample percentile bootstrap. The caption is supplied to the model alongside the image and the image-side embedding is recomputed; for the two-tower CLIP-family comparators it is fused on the text side under the matched protocol (Methods §4.8). Matched zero-shot results are in Extended Data Table ED6. Models and ordering follow Fig. 4 (UniDerm last). UniDerm and Qwen3VL-Emb-2B use the canonical multi-view text-prototype probe, the CLIP-family comparators the matched mean-fusion head on the same frozen embeddings. UniDerm row in **bold**.

Model	**Few-shot** 10%	**Few-shot** 30%	**Few-shot** 50%	**Few-shot** 100%
BiomedCLIP	0.521 (0.446–0.599)	0.618 (0.544–0.690)	0.682 (0.608–0.743)	0.720 (0.654–0.773)
MONET	0.512 (0.440–0.591)	0.616 (0.546–0.689)	0.700 (0.638–0.755)	0.734 (0.670–0.788)
DermLIP	0.543 (0.483–0.619)	0.690 (0.611–0.750)	0.715 (0.637–0.775)	0.731 (0.651–0.790)
OpenCLIP-G	0.530 (0.470–0.606)	0.662 (0.586–0.738)	0.675 (0.604–0.744)	0.726 (0.652–0.794)
MAKE	0.523 (0.460–0.597)	0.687 (0.621–0.738)	0.680 (0.606–0.742)	0.743 (0.699–0.784)
PanDermLIP	0.603 (0.528–0.679)	0.672 (0.604–0.737)	0.713 (0.665–0.754)	0.748 (0.706–0.789)
Qwen3VL-Emb-2B	0.659 (0.604–0.708)	0.717 (0.658–0.769)	0.745 (0.704–0.790)	0.776 (0.736–0.820)
UniDerm (ours)	**0.701** (0.662–0.738)	**0.776** (0.733–0.819)	**0.781** (0.738–0.821)	**0.813** (0.770–0.853)

All values are with-context few-shot BACC. UniDerm's matched no-context few-shot curve reaches BACC 0.688, 0.768, 0.722 and 0.798, so the patient-context caption adds between +0.8 and +5.9 percentage points across the four fractions and UniDerm matches the leading comparator's full-label accuracy at ~30% of the labels. The matched zero-shot results (with-context and no-context, all three caption forms) are in Extended Data Table ED6.

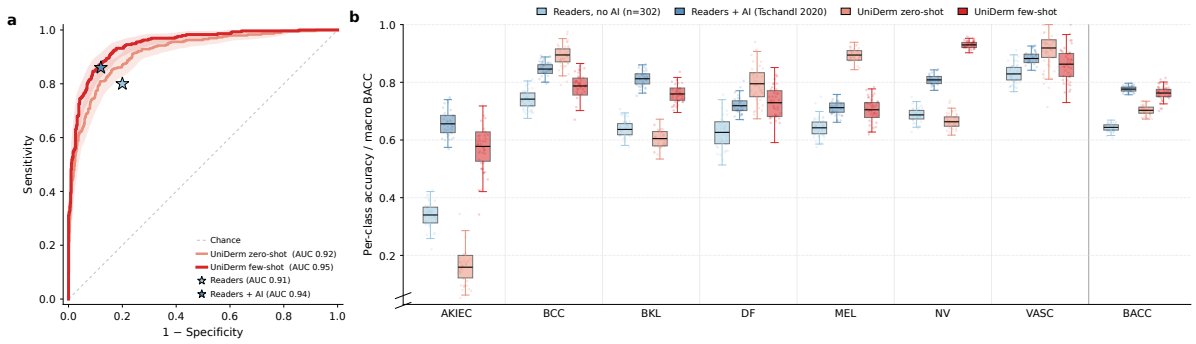


Figure 5 | **Head-to-head against the 302-reader panel of the Tschandl-948 reader subset of ISIC2018 task 3.** UniDerm was evaluated on the exact 948-image dermoscopic subset of the reader study of Tschandl et al. (2020) [5]. **a**, Binary any-malignancy ROC for UniDerm zero-shot (light salmon, AUC 0.917) and 100%-label few-shot linear probing (crimson, AUC 0.946), with 95% bootstrap bands. Stars mark two published reader operating points (n = 302; 95% CIs in brackets): average reader without AI (light blue) and with AI assistance (medium blue), each with its reconstructed reader-consensus AUC in the legend. Matched-operating-point and paired DeLong comparisons against the reader-consensus ROC are in Extended Data Tables ED8 and ED9. **b**, Per-class top-1 accuracy and macro-mean BACC at 100% labels, as box plots for readers without AI (light blue), readers with AI (medium blue), UniDerm zero-shot (salmon) and UniDerm few-shot (crimson). UniDerm boxes are the bootstrap distribution of per-class recall (2,000 resamples of the 948 test images); reader boxes are Monte-Carlo distributions from the published per-class mean±standard error (ref. [5], Supplementary Table 3). Boxes mark the interquartile range and median, whiskers the 2.5–97.5 percentile and overlaid dots are a 40-point subsample; the rightmost group is the macro-mean BACC. The y-axis is truncated at 0.03 (axis break at lower left).

Table ED8 | **Binary any-malignancy detection on the Tschandl-948 reader subset of ISIC2018 task 3 (corresponds to Fig. 5a).** Reader sensitivity and specificity (unaided and AI-assisted) are the published operating points [5] (95% CI in parentheses). † marks reader-consensus ROC AUCs reconstructed from per-image votes of the 302 raters. "Curve" is UniDerm's threshold-dependent ROC (matched operating points in Extended Data Table ED9). The last column gives the paired DeLong ΔAUC (two-sided p, 948 images) of each UniDerm ROC against the reader-consensus reference. The Readers + AI row is context only (those readers used a different system). Best AUC in **bold**; consensus and UniDerm AUCs carry 95% bootstrap CIs (5,000 resamples).

Method	Sensitivity	Specificity	Binary AUC	ΔAUC vs Readers (p)
Readers (n = 302)	0.80 (0.78–0.83)	0.80 (0.77–0.82)	0.914 (0.895–0.931)†	(reference)
Readers + AI (n = 302)	0.86 (0.84–0.88)	0.88 (0.87–0.90)	0.941 (0.926–0.955)†	— (context only)
UniDerm zero-shot	(curve)	(curve)	0.917 (0.898–0.936)	+0.003 (p=0.77)
UniDerm few-shot (100%)	(curve)	(curve)	**0.946 (0.930–0.961)**	+0.032 (p=0.002)

Table ED9 | **Matched operating-point comparison of UniDerm few-shot (100%-label linear probing, AUC 0.946) against reader and AI-assisted reference points on the Tschandl-948 reader subset of ISIC2018 task 3 (corresponds to Fig. 5a).** At each reference operating point we report UniDerm's sensitivity at the reference specificity and its specificity at the reference sensitivity, each with a 95% non-parametric bootstrap CI (5,000 resamples). A value is in **bold** when its CI clears the reference. UniDerm is above the reader point on both axes and approximately matches the AI-assisted readers (cf. Extended Data Table ED8).

Reference operating point (Sens / Spec)	UniDerm Sens @ ref Spec	UniDerm Spec @ ref Sens
Readers (0.80 / 0.80)	**0.932 (0.902–0.965)**	**0.930 (0.898–0.956)**
Readers + AI (0.86 / 0.88)	0.871 (0.824–0.917)	0.893 (0.844–0.929)

Table ED10 | **Label efficiency on the Tschandl-948 reader subset of ISIC2018 task 3 [5].** UniDerm macro-mean BACC (mean, 95% bootstrap CI) and binary any-malignancy AUC at zero-shot (0%) and 100% labels, versus the average rater without and with AI assistance (n = 302; no published BACC CI). †, reader-consensus ROC reconstructed from per-image votes. Each p tests UniDerm against the average reader (BACC, bootstrap two-distribution test; AUC, paired DeLong test on the same 948 images; uncorrected for multiple comparisons). Best UniDerm value per column in **bold**.

Training-label fraction	Macro BACC (p vs Readers)	Binary AUC (p vs Readers)
UniDerm zero-shot (0%)	0.703 (0.673–0.735, p<0.001)	0.917 (p=0.77)
UniDerm few-shot 100%	**0.763 (0.725–0.801, p<0.001)**	**0.946 (p=0.002)**
Readers (n = 302)	0.643	0.914†
Readers + AI (n = 302)	0.776	0.941†

of all 302 raters, UniDerm's few-shot AUC exceeds the consensus AUC (0.914) by a paired DeLong test (ΔAUC +0.032, p = 0.002, Extended Data Table ED8), whereas its zero-shot AUC (0.917) is statistically indistinguishable from that consensus (ΔAUC +0.003, p = 0.77). The 100%-label AUC (0.946) also reaches the AI-assisted reader consensus (0.941), which we report as context rather than a formal test because those readers were aided by a different system.

Diverse Dermatology Images (DDI), fairness across skin tones. DDI [7] (Daneshjou et al., *Science Advances* 2022) is a 656-image, biopsy-confirmed clinical-photograph set with binary malignant/benign labels, stratified by Fitzpatrick skin type (FST) and paired with a majority-vote reference from three board-certified dermatologists under the same stratification. We benchmarked UniDerm against this reference at two expert-label budgets (Methods §4.8): *strict zero-shot*, which uses no expert-labelled image and fits no parameter on DDI, scoring malignancy as the summed softmax over six per-disease PAD-6 text prototypes; and *cross-dataset few-shot*, a logistic-regression linear probe fitted on the fully expert-labelled PAD-6 training partition (1,493 smartphone clinical images, six classes collapsed to malignant/benign) and applied to DDI. As PAD-6 shares no image with DDI and

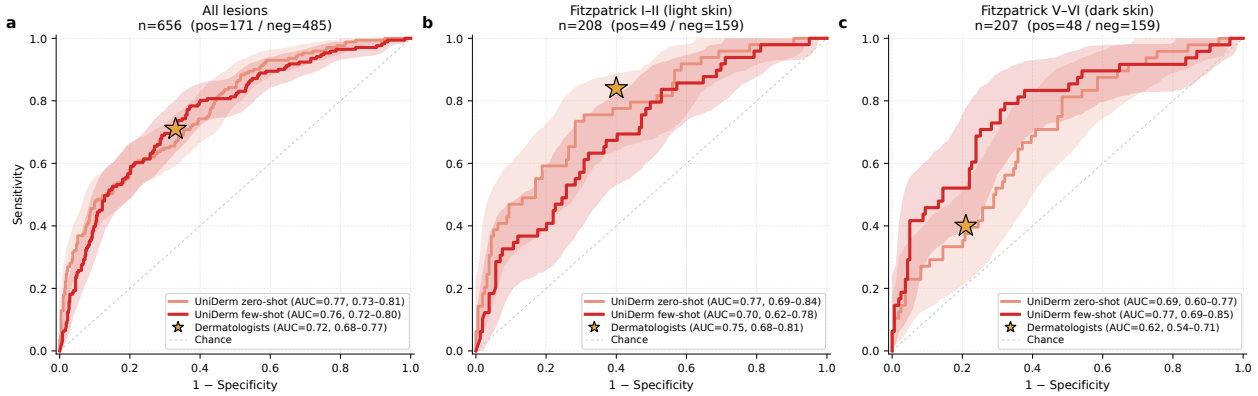


Figure 6 | **Skin-tone-stratified malignancy detection on the Diverse Dermatology Images benchmark (DDI). a–c**, ROC curves of UniDerm under two protocols (Methods §4.8): *zero-shot* (light salmon; no expert-labelled image and no parameter fit on DDI) and *few-shot* (dark crimson; a linear-probe logistic-regression head fitted on the fully expert-labelled PAD-6 training partition, six classes collapsed to malignant/benign, and applied to DDI test images). Subgroups follow Daneshjou et al. (2022) [7]: **a**, all lesions (n = 656, 171/485 positive/negative); **b**, Fitzpatrick I–II light skin (n = 208, 49/159); **c**, Fitzpatrick V–VI dark skin (n = 207, 48/159). Bands are 95% non-parametric bootstrap CIs (1,000 resamples); the gold star in each panel is the majority-vote operating point of three board-certified dermatologists, and the legend reports each AUC with its 95% CI. Matched operating points and significance tests are in Extended Data Table ED12.

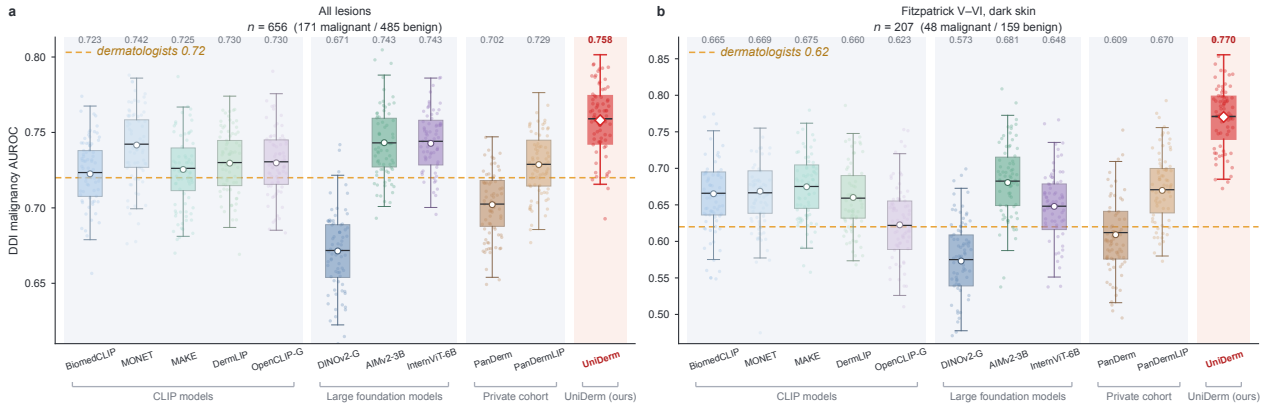


Figure 7 | **UniDerm detects malignancy most accurately on darker skin.** All eleven models received the same expert labels under one identical few-shot protocol (Methods §4.8). The comparators are state-of-the-art transformer-based models spanning every family. **a**, malignancy AUROC on all lesions (n = 656, 171 malignant and 485 benign). **b**, the Fitzpatrick V–VI dark-skin slice (n = 207, 48 malignant and 159 benign), the subgroup this benchmark exists to expose. UniDerm leads in both panels and is the only model that does not lose accuracy from **a** to **b**. Each column is the full 1,000-resample non-parametric bootstrap distribution of that model's AUROC, shown as a box (interquartile range), centre line (median), whiskers (2.5th to 97.5th percentiles, the 95% CI) and 72 randomly drawn resample dots jittered horizontally. The white marker and printed value are the point AUROC on the full subgroup, and the gold dashed line is the dermatologist AUROC of that subgroup.

Table ED11 | **Per-class top-1 accuracy and macro-mean BACC at** 100% **labels on the Tschandl-948 reader subset of ISIC2018 task 3 (corresponds to Fig. 5b).** n = 302 readers without and with AI assistance (ref. [5], Supplementary Table 3) versus UniDerm under zero-shot and 100%-label few-shot linear probing. Classes: AKIEC, actinic keratosis / intraepithelial carcinoma; BCC, basal cell carcinoma; BKL, benign keratosis-like lesions; DF, dermatofibroma; MEL, melanoma; NV, melanocytic nevus; VASC, vascular lesions. Each UniDerm value carries the two-sided p for that class against the average reader (exact binomial for class rows, bootstrap for macro BACC; uncorrected) and 95% bootstrap CIs in parentheses. Best UniDerm macro-mean in **bold**.

Class	**n**	**Readers**	**Readers +AI**	**UniDerm zero-shot**	**UniDerm few-shot** 100%
AKIEC	43	0.340 (0.259–0.421)	0.655 (0.569–0.741)	0.160 (0.062–0.286, p=0.02)	0.581 (0.436–0.732, p=0.002)
BCC	93	0.742 (0.677–0.807)	0.845 (0.801–0.889)	0.892 (0.822–0.951, p=0.001)	0.784 (0.697–0.870, p=0.41)
BKL	192	0.638 (0.582–0.694)	0.812 (0.763–0.861)	0.605 (0.532–0.674, p=0.33)	0.760 (0.700–0.819, p<0.001)
DF	44	0.626 (0.515–0.737)	0.720 (0.670–0.770)	0.792 (0.667–0.905, p=0.02)	0.726 (0.585–0.848, p=0.21)
MEL	158	0.642 (0.584–0.700)	0.712 (0.664–0.761)	0.893 (0.846–0.938, p<0.001)	0.702 (0.633–0.770, p=0.12)
NV	383	0.687 (0.643–0.731)	0.808 (0.772–0.845)	0.664 (0.618–0.711, p=0.32)	0.930 (0.903–0.954, p<0.001)
VASC	35	0.829 (0.766–0.892)	0.883 (0.840–0.925)	0.916 (0.815–1.000, p=0.26)	0.858 (0.730–0.967, p=0.82)
Macro BACC	948	0.643 (0.616–0.670)	0.776 (0.756–0.797)	0.703 (0.673–0.735, p<0.001)	**0.763 (0.725–0.801,** p<0.001**)**

Table ED12 | **Skin-tone-stratified malignancy detection on DDI: UniDerm versus three board-certified dermatologists (Daneshjou et al. 2022 [7]).** Per-subgroup dermatologist operating point (majority-vote sensitivity and specificity) and ROC AUC (ref. [7], Supplementary Table 2, n = 3), and UniDerm AUC under strict zero-shot (no expert-labelled image) and cross-dataset few-shot linear probing on the fully expert-labelled PAD-6 training partition. Rows are Fitzpatrick skin-type subgroups; values are mean (95% percentile bootstrap CI, 1,000 resamples). The higher of the two UniDerm AUCs per subgroup is in **bold**. Sens and Spec, the dermatologists' majority-vote sensitivity and specificity; n, subgroup size; Malignant, number of biopsy-confirmed malignant lesions.

			Dermatologists (n = 3)			**UniDerm AUC**	
Subgroup	n	**Malignant**	**Sens**	**Spec**	**AUC**	**zero-shot**	**few-shot (PAD-6)**
All lesions	656	171	0.71	0.67	0.72 (0.68–0.77)	**0.773 (0.730–0.813)**	0.758 (0.717–0.801)
FST I–II light	208	49	0.84	0.60	0.75 (0.68–0.81)	**0.771 (0.690–0.843)**	0.702 (0.616–0.778)
FST V–VI dark	207	48	0.40	0.79	0.62 (0.54–0.71)	0.689 (0.603–0.768)	**0.770 (0.685–0.851)**

contains no FST V–VI labelling, that probe is a genuine cross-dataset transfer.

UniDerm is reader-competitive on every subgroup at both label budgets (Fig. 6a–c, Extended Data Table ED12). On all lesions it reaches AUC 0.77 with no expert label and 0.76 with the PAD-6 labels, against the readers' 0.72. The two budgets diverge by skin tone. On FST I–II light skin the label-free setting leads (0.77 versus readers' 0.75) whereas the PAD-6-fitted probe regresses to 0.70. On the FST V–VI dark-skin slice, where reader accuracy is lowest, the label-free setting essentially matches the reader (0.69 versus 0.62, within bootstrap). The fitted probe instead rises to 0.77, exceeding the reader on both axes at its operating point (sensitivity 0.52 versus 0.40, specificity 0.95 versus 0.79).

Exceeding every comparator on all and on darker skin. Given the same expert labels, UniDerm leads all ten comparators on both DDI slices, by its widest margin on darker skin (Fig. 7). On all lesions the strongest of them, AIMv2-3B and InternViT-6B, tie at 0.743, against 0.758 for UniDerm. On the FST V–VI dark-skin slice every comparator degrades, losing between 0.051 and 0.107 AUROC against its own all-lesion value, and the best of them, AIMv2-3B, drops to 0.681. UniDerm instead improves, to 0.770, which is +0.089 above that best comparator. It is the only one of the eleven models whose accuracy does not fall on darker skin.

In low-resource settings, where specialists are scarcest, what matters is an openly trained model that reaches the published board-certified-dermatologist reference and holds up on dark skin. These remain same-test-set comparisons with published reader results, and cannot replace prospective validation.

Table ED13 | **Concept-to-image search on the Derm7pt melanoma–nevus subset [35] (827 images; four dermoscopic concepts: pigment network, dots and globules, streaks, pigmentation), macro precision@K for $K \in \{5, 10, 25, 50\}$.** $P@K$ is macro-averaged over the four concepts; values are mean (lower–upper) of the 95% image-paired bootstrap CI (n = 1,000), best per column in **bold**. Models and ordering follow Fig. 2c (UniDerm last): the public-data BiomedCLIP [40], MONET [21], MAKE [23], DermLIP [22], OpenCLIP-G [41], the private-cohort PanDermLIP [22] and the Qwen3VL-Emb-2B backbone [32]. UniDerm uses the dual target-text-instruction setting (Extended Data Table ED20).

Model	$P@5$	$P@10$	$P@25$	$P@50$
BiomedCLIP	0.538 (0.350–0.700)	0.516 (0.400–0.650)	0.565 (0.480–0.650)	0.588 (0.530–0.650)
MONET	0.652 (0.450–0.850)	0.634 (0.475–0.800)	0.627 (0.540–0.710)	0.627 (0.565–0.680)
MAKE	0.719 (0.600–0.850)	0.759 (0.650–0.875)	0.736 (0.660–0.810)	0.722 (0.665–0.775)
DermLIP	0.715 (0.550–0.900)	0.676 (0.550–0.800)	0.698 (0.610–0.780)	0.698 (0.640–0.755)
OpenCLIP-G	0.502 (0.250–0.750)	0.485 (0.325–0.675)	0.511 (0.410–0.600)	0.519 (0.450–0.585)
Qwen3VL-Emb-2B	0.833 (0.650–1.000)	0.796 (0.650–0.925)	0.720 (0.628–0.808)	0.682 (0.619–0.744)
PanDermLIP	0.598 (0.400–0.750)	0.644 (0.525–0.750)	0.669 (0.590–0.740)	0.662 (0.600–0.720)
UniDerm (ours)	**0.936 (0.800–1.000)**	**0.862 (0.750–0.950)**	**0.806 (0.740–0.870)**	**0.794 (0.735–0.845)**

2.4. Searching skin images with clinical language, for evidence retrieval and transparency

Beyond assigning a diagnosis, a dermatology model is clinically useful when images and clinical language share one searchable space, so that a clinician can search images by a named clinical feature and move freely between a lesion and its written description. UniDerm supports both, and leads every comparator on each (Fig. 2c).

Concept-to-image search. A clinical *concept* is a named visual feature that composes a diagnosis, such as a pigment network or streaks, rather than the diagnosis itself. Concept-to-image search takes such a concept name as a free-text query and returns the images that display it, ranking every image by the similarity of its embedding to the concept text so that a high-ranked result is a lesion that genuinely shows the feature. We measured concept-to-image macro precision-at-10 (macro $P@10$) on the Derm7pt melanoma–nevus subset [35] (four main concepts) and SkinCAP [36] (17 concepts with more than 50 positive samples). UniDerm reached macro $P@10$ = 0.862 on Derm7pt and 0.642 on SkinCAP, ahead of every comparator (Extended Data Tables ED13, ED14). It outranks even MONET [21], the dermatology foundation model purpose-built for concept annotation (0.634 on Derm7pt, 0.321 on SkinCAP), by +0.23 and +0.32 macro $P@10$, and also leads the private-cohort PanDermLIP (0.644, 0.524). Because each query ranks images by embedding similarity to a free-text concept name, this precision means a clinician can search by a named dermoscopic concept, such as pigment network or streaks, and reliably surface the lesions that display it, without any concept-level supervision. That capability is the basis for transparent, auditable concept-level annotation of dermatology images and models, the use case MONET was introduced to serve [21].

Cross-modal evidence search. A clinician also needs to move between an image and its written evidence. On SkinCAP image–caption retrieval at $R@10$, UniDerm again leads. Its image-to-text $R@10$ is 0.332 versus the next-best public-data comparator MAKE (0.208), and its text-to-image $R@10$ is 0.350 versus the private-cohort PanDermLIP runner-up (0.211) (Extended Data Tables ED15, ED16). An image–text alignment this tight lets a query image retrieve its matching free-text clinical description and a written description retrieve the lesions it depicts, the substrate for case-based decision support, literature linkage and education from either an image or a clinical query.

Discussion

UniDerm is a fully open dermatology foundation model trained on publicly available image–text data alone. Across diagnosis at low expert-label cost, from a fraction of the labels down to none, patient-

Table ED14 | **Concept-to-image search on SkinCAP [36] (17 concepts with more than 50 positive samples), macro precision@K for $K \in \{5, 10, 25, 50\}$.** Second of four search sub-tables (Fig. 2c). $P@K$ is macro-averaged over the per-concept precision values. Models, ordering and bootstrap as in Extended Data Table ED13. Best value per column in **bold**.

Model	$P@5$	$P@10$	$P@25$	$P@50$
BiomedCLIP	0.234 (0.165–0.318)	0.220 (0.165–0.271)	0.208 (0.179–0.240)	0.188 (0.166–0.209)
MONET	0.325 (0.247–0.412)	0.321 (0.259–0.382)	0.295 (0.261–0.334)	0.272 (0.245–0.300)
MAKE	0.527 (0.435–0.624)	0.499 (0.423–0.571)	0.430 (0.386–0.473)	0.386 (0.355–0.415)
DermLIP	0.486 (0.400–0.577)	0.434 (0.371–0.500)	0.380 (0.341–0.421)	0.335 (0.307–0.363)
OpenCLIP-G	0.210 (0.141–0.294)	0.192 (0.147–0.241)	0.177 (0.148–0.209)	0.173 (0.152–0.194)
Qwen3VL-Emb-2B	0.354 (0.281–0.434)	0.331 (0.270–0.400)	0.310 (0.273–0.349)	0.298 (0.273–0.324)
PanDermLIP	0.563 (0.471–0.659)	0.524 (0.459–0.594)	0.447 (0.402–0.492)	0.392 (0.361–0.421)
UniDerm (ours)	**0.698 (0.612–0.776)**	**0.642 (0.571–0.706)**	**0.533 (0.489–0.579)**	**0.489 (0.457–0.522)**

Table ED15 | **Cross-modal evidence search on SkinCAP [36]: image-to-text recall@K for $K \in \{5, 10, 50\}$.** Third of four search sub-tables (Fig. 2c). For each query image, $R@K$ is the fraction of relevant captions retrieved in the top K over the full SkinCAP pool (n = 3,989). The cross-modal evaluation grid is $K \in \{5, 10, 50\}$ (no $K = 25$). Models, ordering and bootstrap as in Extended Data Table ED13. Best value per column in **bold**.

Model	$R@5$	$R@10$	$R@50$
BiomedCLIP	0.085 (0.076–0.093)	0.136 (0.126–0.146)	0.344 (0.330–0.357)
MONET	0.102 (0.093–0.111)	0.164 (0.152–0.175)	0.389 (0.374–0.405)
MAKE	0.135 (0.125–0.145)	0.208 (0.196–0.221)	0.444 (0.430–0.459)
DermLIP	0.103 (0.094–0.113)	0.162 (0.151–0.173)	0.366 (0.351–0.380)
OpenCLIP-G	0.117 (0.106–0.126)	0.178 (0.166–0.190)	0.393 (0.379–0.409)
Qwen3VL-Emb-2B	0.112 (0.102–0.122)	0.165 (0.153–0.177)	0.373 (0.358–0.388)
PanDermLIP	0.134 (0.123–0.144)	0.204 (0.192–0.216)	0.434 (0.419–0.449)
UniDerm (ours)	**0.226 (0.214–0.239)**	**0.332 (0.317–0.347)**	**0.591 (0.576–0.606)**

Table ED16 | **Cross-modal evidence search on SkinCAP [36]: text-to-image recall@K for $K \in \{5, 10, 50\}$.** Fourth of four search sub-tables (Fig. 2c). For each query caption, $R@K$ is the fraction of relevant images retrieved in the top K over the full SkinCAP pool (n = 3,989). The cross-modal evaluation grid is $K \in \{5, 10, 50\}$ (no $K = 25$). Models, ordering and bootstrap as in Extended Data Table ED13. Best value per column in **bold**.

Model	$R@5$	$R@10$	$R@50$
BiomedCLIP	0.076 (0.068–0.084)	0.123 (0.113–0.134)	0.330 (0.316–0.344)
MONET	0.098 (0.089–0.106)	0.157 (0.146–0.168)	0.364 (0.349–0.379)
MAKE	0.132 (0.121–0.143)	0.199 (0.187–0.212)	0.442 (0.427–0.457)
DermLIP	0.104 (0.094–0.114)	0.161 (0.149–0.172)	0.361 (0.346–0.375)
OpenCLIP-G	0.098 (0.089–0.107)	0.152 (0.141–0.164)	0.360 (0.344–0.375)
Qwen3VL-Emb-2B	0.106 (0.096–0.117)	0.172 (0.160–0.184)	0.362 (0.347–0.377)
PanDermLIP	0.140 (0.129–0.150)	0.211 (0.199–0.223)	0.461 (0.447–0.475)
UniDerm (ours)	**0.244 (0.230–0.257)**	**0.350 (0.336–0.365)**	**0.626 (0.610–0.641)**

context-conditioned diagnosis, two independent reader studies, and concept-to-image and cross-modal evidence search, one finding recurs. A model built without any private clinical cohort matches and surpasses systems trained on millions of proprietary images. We discuss in turn what makes UniDerm open, accurate, label-efficient, context-aware and reader-competitive, why its recipe transfers beyond dermatology, and where each property stops holding. Their common thread, developed in the closing section, is that the openness making UniDerm reproducible is also what makes specialist-level dermatology AI accessible rather than something to be purchased.

UniDerm is open. Its model weights, training configurations, training pipeline and data recipe are public, no private clinical cohort was used, and even the synthesised training text is produced with open-weights models alone (§4.3). The field's leading systems instead depend on proprietary clinical cohorts, bounded by the privacy and consent constraints on patient data, and most release no downloadable weights (Fig. 1b). As a result, they cannot readily be reproduced, audited for bias, or extended by the community. Full openness removes each barrier: any group can reproduce every result, audit the model for bias, and build on it freely. Because it depends on no private cohort and runs on local hardware, a clinic can retrain and deploy it on-site, so its own patients' images never leave the institution. That hardware need not include an accelerator. UniDerm-VE-lite, a third the size of PanDerm, clears that floor on four CPU threads with no GPU at all, running 3.1 times faster than PanDerm while scoring above it, so a site without dedicated compute is not excluded. Openness also lets the community iteratively refine the model on new photographs, emerging conditions and devices, so the foundation steadily improves rather than freezing at release.

UniDerm is accurate. It ranks first across zero-shot diagnosis, label-efficient linear probing, and concept-to-image and cross-modal evidence search, ahead of all open models and the leading private-cohort models PanDerm [8] and PanDermLIP [22] (§2.1, §2.4). The field had reached this accuracy only from large private cohorts, as PanDerm did on more than two million proprietary images; the public-data models that avoided them have tended to trail behind. Yet each leaves part of the supervision the public data already carries unused: the image-only self-supervised PanDerm never sees the text paired with each image, and two-tower CLIP, whose single fixed image embedding must stay close to all of an image's texts at once, collapses them to their average, so each text's distinctive signal becomes noise on the others, and loses the detail that distinguishes one from another. By instead learning from that text through supervision-denoising contrastive learning, UniDerm reaches private-cohort accuracy from public data alone, challenging the premise that clinical-grade dermatology accuracy depends on private data. This is more likely a floor than a ceiling: adding a private cohort, or simply more public data, could lift UniDerm further.

UniDerm is label-efficient. Because it keeps this class-distinguishing detail via supervision denoising, UniDerm's embedding is more cleanly separated by disease class in its raw space (Fig. 3), so a small linear probe already reaches the accuracy that competitors need their full label budget to approach (Fig. 2a). In dermatology, expert and biopsy-confirmed annotations are the binding cost. An embedding that attains state-of-the-art accuracy from a small probe therefore lowers the burden of adapting to a new clinic, imaging device or locally prevalent condition.

UniDerm is context-aware. A dermatological diagnosis rests not on the image alone but on the patient's history, body site and prior findings. UniDerm fuses this free-text context with the image *early*, in one token stream, so it reshapes the image embedding. Two-tower CLIP-style models can only fuse it *late*, averaging it onto a fixed image embedding, so for them the same context acts as noise and degrades most of them even as it raises UniDerm (Fig. 4). This split is architectural, a single unified transformer lets image and text tokens attend jointly, whereas two separate encoders meet only at the final average. The model therefore uses the whole case the way a clinician does, turning the patient narrative that image-only systems discard into added accuracy at the point of care.

UniDerm is reader-competitive, within bounds. On an independent malignancy-detection study it matches a large reader consensus without a single expert label, and fitting that benchmark's expert labels lifts it significantly above the consensus (§2.3). It approaches the reader reference even on the Fitzpatrick V–VI dark-skin slice of DDI [7], and exceeds every comparator on darker skin (Fig. 7b). Approaching the reference on darker skin is itself a health-equity signal, since skin-tone accuracy gaps are what can make dermatology AI risk widening, rather than narrowing, disparities in care. That an openly trained model reaches the dermatologist reference is what makes it clinically meaningful where specialists are scarce. We frame this as reaching the clinical-reference range, not parity, because both comparisons use published references rather than prospective matched studies, and multiclass diagnosis, while above unaided readers, only approaches the AI-assisted level.

The recipe is transferable. What makes supervision denoising work is structural, not specific to dermatology. Most medical imaging fields release the same one-image-many-text data, which a unified model can fuse but a two-tower objective cannot (Fig. 8a,b and Fig. 1c). The same open, public-data-only recipe is therefore likely to extend to radiology (MIMIC-CXR, CheXpert), pathology (TCGA) and ophthalmology (EyePACS), plausibly needing only a new visual front-end and field-specific training text. Dermatology would then be a template rather than a special case, suggesting more broadly that specialist-level medical AI need not depend on private data.

Limitations. Two caveats bound these results. First, the DDI [7] and Tschandl-948 [5] reader comparisons are anchored to independently published references, not to prospective matched studies, and the DDI subgroup intervals are wide. Multi-site prospective validation is therefore needed before any claim of clinical readiness. Second, the five training-image sources (Extended Data Table ED17) carry heterogeneous licences, and the aggregate corpus is bounded by its most restrictive component, CC BY-NC 4.0. Several sources are non-redistributable, so rather than the corpus itself we release a data recipe, the per-source download and assembly instructions (§4.13), that reconstructs it from each original distributor.

Toward democratized dermatology AI. UniDerm shows that a fully open model, trained on public data alone and runnable on four CPU threads without a GPU, reaches the performance the field had obtained only from private cohorts. But its deeper significance is access. Dermatological expertise is scarcest in low- and middle-income settings, which are also least able to license proprietary models, buy the accelerators those models assume, or assemble private cohorts of their own. An open foundation inverts that constraint: the populations least served by closed systems can run, scrutinise and improve this one themselves, rather than wait to purchase it. Because the recipe is not specific to skin, the route it charts from public data to specialist-level medical AI extends well beyond dermatology. We hope this moves medical AI from a resource that must be bought toward one that communities everywhere can build and own.

Methods

UniDerm is a fully open dermatology foundation model that learns from public image–text data through supervision-denoising contrastive learning. The subsections below detail its motivation and architecture, training data and objective, and the datasets, metrics and protocols used to evaluate it.

4.1. Background and motivation

The one-image-many-text limitation. Public dermatology image–text datasets pair each image with several heterogeneous targets at once. Derm1M [22], the largest, ships about 1 million image–text pairs from roughly 0.4 million unique images, each image associated on average with a refined clinical

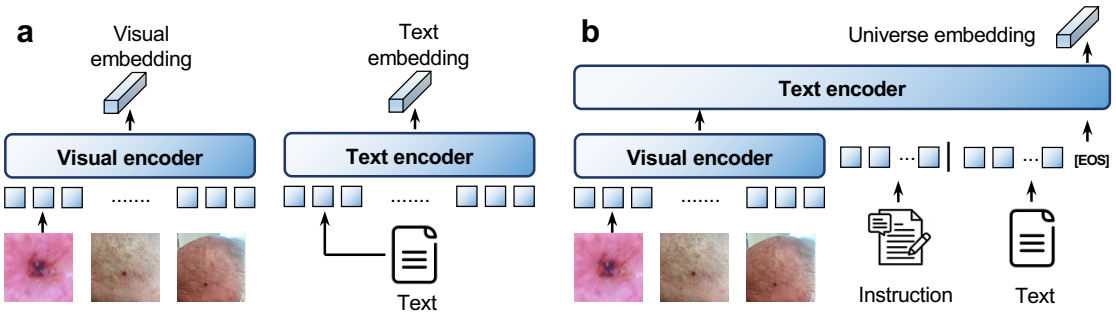


Figure 8 | **From two-tower CLIP to instruction-conditioned embedding. a**, The conventional two-tower CLIP paradigm: separate visual and text encoders, so the image embedding is fixed regardless of the queried attribute and, when one image is paired with several texts (diagnosis, body site, concept), is pulled toward their average. **b**, UniDerm: the image I and a natural-language system instruction S are concatenated into a single token stream consumed by a unified vision–language transformer, and the L_2-normalised final-hidden state of the end-of-sequence ([EOS]) token is taken as the instruction-conditioned embedding $\mathbf{e} = \text{Embed}(I, S) \in \mathbb{R}^{2048}$, contrasted against the embedding of the target text T. The same image under different instructions yields different task-specific embeddings.

caption, patient context and, when applicable, a visual-concept or disease label. Under the standard two-tower CLIP objective [21, 23, 40, 42, 43], a single fixed image embedding must lie close to all of these simultaneously, so its optimum collapses toward their average [24], a low-pass filter that keeps the shared low-frequency component and attenuates the attribute-specific detail separating a diagnosis from a body site or a concept [25, 44]. The failure is architectural: because the two towers interact only at the final dot product, supplying a task instruction to the text tower cannot steer a fixed image embedding, and CLIP performance instead *drops* by 29.4% [26]. UniDerm therefore adopts *supervision-denoising contrastive learning*, a unified-transformer instruction-conditioned contrastive objective in which the image and a natural-language clinical instruction enter the model as one shared token stream; the same image under different instructions then yields different task-specific embeddings that recover the attribute-specific content instead of averaging it, and the patient's clinical context can shape the image embedding itself rather than only the text-side prompt.

Supervision denoising. We formalise the effect as follows. Let a single image I be paired with K heterogeneous target texts $\{T_1, \ldots, T_K\}$, for example a diagnosis, a body site, a visual concept and a clinical caption. A two-tower model must place its one fixed image embedding $\mathbf{e}_I$ close to all K targets at once. Under the normalised-similarity objective, its optimum therefore approaches the mean direction of the target embeddings $\{\text{Embed}(T_k)\}_{k=1}^{K}$, their centroid on the unit sphere. The attribute-specific component that separates one target from another cancels in this average, and each target's fine-grained supervision survives only as a shared, low-frequency residual. Viewed from the supervision side, this collapse is a noising operation: the distinct signal each target carries is corrupted by the others into their common mean. Supervision denoising removes this interference by conditioning the embedding on an instruction. For target T_k it forms $\mathbf{e}(I, S_k) = \text{Embed}(I, S_k)$ and matches it against T_k alone, so the single many-target compromise is decomposed into K single-target problems. Each embedding then need only recover the one supervision signal its instruction selects, and the instruction acts as a selection operator that denoises the mixed supervision back into its per-target components. Its two data-side instantiations, multi-task instruction synthesis and ICD-11 ontology augmentation (§4.3), supply the instruction–target pairs that realise this decomposition; the architecture and the contrastive objective that optimise it follow in §4.2 and §4.5.

4.2. UniDerm architecture

UniDerm is an instruction-aware multimodal unified vision–language transformer [26, 32]. It comprises a 24-layer Vision Transformer encoder with hidden dimension 1,024 and patch size 16×16, followed by a 2×2 spatial-merger producing 32×32-pixel visual tokens, and a 28-layer language-model transformer with hidden dimension 2,048 and rotary positional embedding extended to 32,768 tokens. Image tokens and instruction tokens enter the language transformer as a single shared token stream. The final-hidden state of the end-of-sequence ([EOS]) token is taken as the embedding and is L_2-normalised. The complete model has 2.07 billion parameters.

Given an image I and a natural-language system instruction S, the unified transformer produces the instruction-conditioned image embedding

$$\mathbf{e} \;=\; \mathrm{Embed}(I, S) \;\in\; \mathbb{R}^d, \qquad d = 2048, \tag{1}$$

with $\|\mathbf{e}\|_2 = 1$. The same image with two different system instructions $S_1 \neq S_2$ produces $\mathbf{e}_1 \neq \mathbf{e}_2$. The target text T is encoded by the same transformer in text-only mode (no image tokens) to yield $\mathrm{Embed}(T) \in \mathbb{R}^d$. The contrastive objective then pulls $\mathrm{Embed}(I, S)$ towards its matching $\mathrm{Embed}(T)$. Fine-tuning the vision tower together with the language transformer lets the dermatology signal reshape the visual representation itself, which is what leaves UniDerm's visual encoder a dermatology-adapted ViT-L rather than the off-the-shelf vision tower, and is what UniDerm-VE-lite is distilled from (§4.9).

4.3. Training data

We trained UniDerm exclusively on publicly available dermatology image–text pairs from five public image sources (Fig. 1j–m, Extended Data Table ED17). The corpus spans a broad range of content provenance, disease families, body sites and Fitzpatrick skin tones (Fig. 1j–m); because per-image metadata are sparse and inconsistently recorded across sources, the disease-family, body-site and skin-tone tallies (Fig. 1k–m) are computed only over the labelled subset and are correspondingly smaller than the full image count, with skin-tone labels contributed predominantly by SCIN [27]. The dominant source is Derm1M [22], a public dermatology dataset of 403,583 unique images released with 1,029,761 image–text pairs; it supplies about 80% of our training images. To broaden coverage of dermoscopy, smartphone photography and published-literature figure panels, we added four further public sources: the ISIC2019 dermoscopy archive [20] (14,758 images, with HAM10000 and ISIC2020 images blacklisted to protect the evaluation partitions); the SCIN smartphone-image collection [27] (10,249 images, deduplicated upstream against Derm1M); dermatology figure panels filtered from the PubMed Central Open Access Subset [45] (3,673 images); and dermatology figure panels from the BIOMEDICA archive [46] (73,395 images). In total the five sources contribute 505,658 images, all deduplicated against the evaluation partitions. No private clinical cohort was used, and every training image and its source text is publicly available.

From images to instruction–target pairs. The training signal is not the raw image–text pairs but task-specific (instruction, target) pairs synthesised from them. Each image is expanded into several pairs, each coupling a natural-language system instruction S with the target text T that the instruction requests. This expansion turns Derm1M's 1,029,761 native pairs into 3,436,649 image-paired instruction–target rows, and provides the two data-side components of supervision denoising, layered on the off-the-shelf VLM2Vec instruction-conditioned objective.

(i) Multi-task instruction synthesis. For each image we synthesised up to four instruction–target categories, conditional on which metadata fields were populated: *disease diagnosis* (all images), *ICD-11 coding* (images whose diagnosis matched a WHO ICD-11 dermatology code [47]), *morphological*

Table ED17 | **Composition of the UniDerm training dataset.** Five publicly available dermatology image sources contribute 505,658 unique images and 3,606,540 image-paired instruction–target rows. An additional 2,163,542 text-only rows (WHO ICD-11 dermatology ontology [47], caption-to-disease) bring the total pair count to 5,770,082.

Source	Images	Pairs	Modality / provenance
Derm1M [22]	403,583	3,436,649	Clinical + dermoscopy + educational
BIOMEDICA [46]	73,395	73,871	PMC-OA figure panels (BIOMEDICA)
ISIC2019 [20]	14,758	57,510	Dermoscopy; HAM10000-overlap and ISIC2020-overlap images removed [48]
SCIN [27]	10,249	34,837	Consumer smartphone photographs
PMC-OA dermatology [45]	3,673	3,673	Figure panels from PubMed Central Open Access
Subtotal (image-paired)	505,658	3,606,540	
Text-only pairs	n/a	2,163,542	WHO ICD-11 dermatology ontology [47], caption → disease
Total pairs	n/a	5,770,082	

description (images carrying ≥ 1 visual-concept annotation) and *clinical caption* (all images). Each category draws from a small fixed pool of natural-language instruction templates (for example, "Determine the dermatological diagnosis from this image." or "Identify the visual morphological features of this skin lesion."), so that the model sees the same image under several distinct diagnostic intents. A bare disease name is short and clinically ambiguous, so the *clinical caption* task instead pairs each image with a longer, more informative target: a detailed clinical caption and a structured diagnostic report generated by the open-weights Qwen3.5-397B multimodal large language model, whose caption and clinical-assessment sections serve as the long-form target. Patient context (age, sex and body location) is not rendered into the training instruction; the model is instruction-aware, so such structured context is instead supplied at inference time. This is the configuration probed by the patient-context-conditioned diagnosis experiments on PAD-6.

(ii) ICD-11 ontology augmentation. We drew two complementary training signals from the WHO ICD-11 dermatology ontology [47]. First, we matched each diagnosis to an ICD-11 code with a large language model and used the code's official ICD-11 definition and description as the target text for the *ICD-11 coding* task, replacing the short, clinically ambiguous bare disease name with an ontology-grounded definition. Second, for each pair we mined contrastive hard negatives from the same ontology: two stem codes that share their first three characters fall under a common parent grouping and differ only at the fourth-character category, so they denote closely related but distinct diagnoses, and sampling such same-prefix siblings (about three per pair) forces the model to separate clinically adjacent diagnoses rather than only easy, unrelated ones.

Across all components the corpus totals 5,770,082 pairs: 3,606,540 image-paired rows and 2,163,542 text-only rows (caption-to-disease and ICD-11 ontology pairs). Representative per-family instruction–target examples are given in Extended Data Table ED18.

Test-set contamination audit. To rule out train–test leakage, we audited every one of the 61,689 downstream evaluation images against the 505,658 training images with a three-layer pipeline of increasing stringency: two complementary perceptual hashes required to match simultaneously and bit-exactly (pHash and dHash, random-pair collision probability $\sim 3 \times 10^{-39}$), and an independent-model check using PanDerm [8] cosine similarity above 0.99. At the strictest threshold (both hashes *and* PanDerm), only 2 near-duplicate pairs survive across the entire pool, both in the SkinCAP retrieval set. None of the seven classification benchmarks, the two reader cohorts or the Derm7pt retrieval set retains a single flagged pair. Full per-dataset counts for both partitions are given in Extended Data Table ED19.

Table ED18 | **Representative instruction–target training examples, one per synthesised task family.** Each example is quoted verbatim from the UniDerm training corpus, one field per line. The query is a system *Instruction* with either an *Image* (rendered as the `<image>` token) or query *Text*. Training pulls the query embedding towards its *Target* and away from an ICD-11-sibling hard *Negative* (about three are sampled per pair, one shown; ICD-11 codes follow the WHO dermatology taxonomy [47]). Families (1)–(4) are image-paired; family (5) is text-only.

(1) Disease diagnosis	
Instruction	Determine the dermatological diagnosis from this image.
Image	dermatology image
Text	— (none; image-only query)
Target	An image of dermatitis, a skin condition.
Negative	This image depicts chronic cutaneous eruption.
(2) ICD-11 coding	
Instruction	Assign the ICD-11 classification code for the skin condition shown.
Image	dermatology image
Text	— (none; image-only query)
Target	Common acquired melanocytic naevus (ICD-11: 2F20.0)
Negative	Seborrhoeic keratosis (ICD-11: 2F21.0)
(3) Morphological description	
Instruction	Identify the visual morphological features of this skin lesion.
Image	dermatology image
Text	— (none; image-only query)
Target	erythema, plaque, nodule
Negative	macule, plaque, erosion
(4) Clinical caption	
Instruction	Generate a clinical description of this skin image.
Image	dermatology image
Text	— (none; image-only query)
Target	Dermoscopic image of invasive melanoma
Negative	An image of acral naevus, nonpolarised dermoscopy view, a skin condition.
(5) Text-only matching (no image)	
Instruction	What skin disease is described in this clinical text?
Image	— (none)
Text	Ill-defined erythematous to hyperpigmented hyperkeratotic scaly plaques with a rough surface (eczematous type, hyperkeratotic).
Target	Miscellaneous specified eczematous dermatoses
Negative	Seborrhoeic dermatitis and related conditions

Table ED19 | **Three-layer training–test contamination audit of the 11 downstream benchmarks, at the pHash ∧ dHash and pHash ∧ dHash ∧ PanDerm thresholds.** For each benchmark we report flagged-image counts for the non-test pool (training plus validation images) and the test partition; the few-shot probe is fitted on the training partition alone, so the validation images are included here only for completeness of the leakage audit. **pHash ∧ dHash** counts images bit-exact to a training image on two 64-bit perceptual hashes ($pH = 0$ **and** $dH = 0$). **pHash ∧ dHash ∧ PanDerm** additionally requires PanDerm cosine > 0.99 on that pair. Em-dashes mark partitions a dataset does not provide (test-only benchmarks).

	Linear-probe train+val			Test partition		
Dataset	**N**	**pHash ∧ dHash**	**pHash ∧ dHash ∧ PanDerm**	**N**	**pHash ∧ dHash**	**pHash ∧ dHash ∧ PanDerm**
HAM-7	8,512	3	**0**	1,503	**0**	**0**
PAD-6	1,837	0	**0**	461	**0**	**0**
SD-128	4,214	0	**0**	1,405	**0**	**0**
SNU-134	—	—	—	2,101	**0**	**0**
PH2-2	122	0	**0**	78	**0**	**0**
DAFFODIL-5	—	—	—	1,910	**0**	**0**
ISIC2020-2	28,157	6	**0**	4,969	**3**	**0**
DDI	—	—	—	656	**0**	**0**
SkinCAP	—	—	—	3,989	**56**	**2**
Derm7pt	—	—	—	827	**0**	**0**
Tschandl-948	—	—	—	948	**0**	**0**
Total	**42,842**	**9**	**0**	**18,847**	**59**	**2**

4.4. Evaluation datasets

We describe below the evaluation benchmarks used in this paper. All datasets are publicly available and used in their standard published partition unless otherwise noted. Together the evaluation suite spans **five continents and eight documented countries of image acquisition** (Australia, Austria, Brazil, Italy, Portugal, South Korea, Spain and the United States; several atlas-derived benchmarks add clinical images of undocumented national origin), two imaging modalities (dermoscopy and clinical photography, the latter including consumer-smartphone-captured images), and the Fitzpatrick I–VI skin-tone scale grouped into three strata (light, intermediate and dark) (Fig. 1n,o). This geographic and modality breadth is the closest approximation to multi-institutional external validation that is achievable using publicly released benchmarks, and is intended to expose the model to acquisition shifts well beyond the source distribution of the training data.

HAM-7 (HAM10000; Austria and Australia). A dermoscopic dataset of 10,015 common pigmented-lesion images acquired at two sites, the Department of Dermatology of the Medical University of Vienna (Austria) and the skin-cancer practice of Cliff Rosendahl in Queensland (Australia) [28], labelled into seven diagnostic categories: actinic keratoses / intraepithelial carcinoma (AKIEC), basal-cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevus (NV) and vascular lesion (VASC). HAM-7 denotes this seven-class slice, used in our linear-probing and zero-shot evaluations. Linear probing sub-samples the training partition at $\{10, 30, 50, 100\}$%, with the official test partition held out.

PAD-6 (PAD-UFES-20; Brazil). A clinical, smartphone-acquired skin-lesion dataset of 2,298 images in six diagnostic categories, actinic keratosis (ACK), basal-cell carcinoma (BCC), melanoma (MEL), nevus (NEV), squamous-cell carcinoma (SCC) and seborrheic keratosis (SEK), collected by the Dermatological and Surgical Assistance Program of the Federal University of Espírito Santo (Brazil) [29]. Each lesion carries 22 patient clinical metadata fields in the PAD-UFES-20 release [29] (age, sex, geographic region, two lesion diameters, Fitzpatrick skin type, six symptom flags, nine history and exposure fields, and biopsy status); we render 17 of them into the patient-context caption, excluding the diagnostic label, biopsy status, parental-background fields, household-infrastructure fields and image identifiers. We adopt the standard 60/20/20 patient-wise partition.

SD-128 (SD-198 128-class split). A clinical (non-dermoscopic) benchmark for general skin-condition diagnosis, spanning 128 fine-grained diagnostic categories drawn from routine, multi-centre dermatology practice. It is the subset of the 198-class SD-198 benchmark of Sun et al. [30] whose classes retain at least 20 images each (the 70 rarest classes, fewer than 20 images, are dropped), leaving 5,619 clinical images. The images were harvested from the DermQuest online dermatology atlas, contributed by an international community of patients and dermatologists across diverse imaging conditions and patient demographics, so the per-image country of acquisition is not recorded. We use a stratified 70/5/25 train/validation/test partition (3,933/281/1,405 images).

ISIC20-2 (SIIM-ISIC 2020 challenge; United States, Australia, Austria and Spain). The binary malignant/benign melanoma-detection slice of the ISIC 2020 challenge dataset, 33,126 dermoscopic images with ~2% melanoma prevalence, contributed through the International Skin Imaging Collaboration from multiple clinical centres across four countries (Memorial Sloan Kettering Cancer Center, United States; Melanoma Institute Australia and the Sydney Melanoma Diagnostic Centre, Australia; The University of Queensland, Australia; Medical University of Vienna, Austria; and Hospital Clínic de Barcelona, Spain) [31]. We denote this binary benchmark ISIC20-2 and report AUROC rather than BACC given the extreme class imbalance.

SNU-134 (South Korea). A clinical-photograph benchmark for general skin-condition diagnosis spanning 134 common disease categories encountered in routine dermatology practice, ranging from inflammatory disorders to infectious diseases and neoplasms. The images were collected across three South Korean university hospitals (Seoul National University Bundang Hospital, Inje University Sanggye Paik Hospital and Hallym University Dongtan Hospital) and emphasise presentations in East-Asian patients [10], a demographic complementary to the dermoscopic and Brazilian-smartphone benchmarks. We evaluate on the 134-class, 2,101-image partition, given the large number of categories the full class list is not enumerated here.

PH2-2 (PH2; Portugal). The binary (melanocytic nevus versus melanoma) slice of PH2, a dermoscopic dataset of 200 images acquired at the Dermatology Service of Hospital Pedro Hispano, Matosinhos (Portugal) [33]. We denote this binary benchmark PH2-2.

DAFFODIL-5 (Daffodil skin-disease dataset). A five-class rare-disease set of 1,910 clinical images, vitiligo, acne, Stevens–Johnson syndrome / toxic epidermal necrolysis (SJS/TEN), hyperpigmentation and nail psoriasis, drawn from the Skin Disease Classification Dataset of Khushbu and Akter (Daffodil International University, Bangladesh) [34]. That dataset was self-collected from unnamed hospitals and publicly available online resources spanning multiple countries, so the per-image country of acquisition is undocumented. It probes rare-disease generalisation under acquisition shift relative to the more curated benchmarks.

DDI (Diverse Dermatology Images; United States). A 656-image biopsy-confirmed clinical-photograph set retrospectively curated from Stanford Clinics pathology reports (United States; 2010–2020), Fitzpatrick-stratified with a binary malignant/benign label [7]. It is accompanied by a three-board-certified-dermatologist reader study (binary decisions ensembled by majority vote) with the same Fitzpatrick stratification, which we use as the literature-anchor reader reference. The FST V–VI dark-skin slice (n = 207) is the most acquisition-shifted partition in our suite and our principal cross-skin-tone-fairness probe.

Tschandl-948 (ISIC2018 task 3 reader subset; Austria and Australia). The 948-image dermoscopic reader-study subset of ISIC2018 task 3 [5], drawn from the same two sources as HAM10000 (Medical University of Vienna, Austria; Cliff Rosendahl, Queensland, Australia). Its per-class composition (AKIEC = 43, BCC = 93, BKL = 192, DF = 44, MEL = 158, NV = 383, VASC = 35) is byte-for-byte aligned with ref. [5] Supplementary Table 3. We use it for the reader head-to-head in Fig. 5.

Derm7pt (seven-point-checklist dataset; Italy and Austria). The melanoma–nevus subset (827 images) of the seven-point-checklist dermoscopic dataset [35], whose images derive from the Interactive Atlas of Dermoscopy (Universities of Naples and Florence, Italy, and the Medical University of Graz, Austria) and were assembled into the machine-learning dataset at Simon Fraser University (Canada). We report macro P@10 concept-to-image search over the four main dermoscopic concepts (pigment network, dots and globules, streaks and pigmentation).

SkinCAP (United States). A 3,989-image clinical-photograph dataset with expert-written multi-sentence captions [36], re-drawn from two public sources: DDI (Stanford Clinics, United States) and Fitzpatrick17k, the latter scraped from the DermaAmin and Atlas Dermatologico teaching atlases with per-image patient origin undocumented. In the subset we use, 17 concepts carry more than 50 positive samples; we evaluate both macro P@10 concept-to-image search and Recall@5/10/50 cross-modal evidence search.

Together the 11 evaluation benchmarks span dermoscopy (HAM-7, Tschandl-948, PH2-2, ISIC20-2, Derm7pt) and clinical photography (PAD-6, SD-128, SNU-134, DDI, DAFFODIL-5, SkinCAP), the latter spanning capture devices from clinical cameras to consumer smartphones (PAD-6 at the source-modality level). They span the full Fitzpatrick I–VI skin-tone range, grouped into three strata (light I–II, intermediate III–IV and dark V–VI), with explicit per-stratum analysis on DDI. They span eight documented countries of acquisition. This distribution is broader than the training-data source distribution along every axis, and the comparisons reported here are therefore informative about out-of-distribution generalisation, not only in-distribution performance.

4.5. Training objective

Supervision-denoising contrastive learning optimises UniDerm with an instruction-conditioned InfoNCE loss. The loss pulls each image–instruction embedding onto the one target its instruction selects, recovering that target's supervision instead of the cross-target average. For a batch of B samples $\{(I_i, S_i, T_i)\}$ (image, system instruction, target text),

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle \mathbf{e}_i^{\text{img}}, \mathbf{e}_i^{\text{tgt}} \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle \mathbf{e}_i^{\text{img}}, \mathbf{e}_j^{\text{tgt}} \rangle / \tau) + \sum_{k=1}^{H_i} \exp(\langle \mathbf{e}_i^{\text{img}}, \mathbf{e}_{i,k}^{\text{neg}} \rangle / \tau)}, \tag{2}$$

where $\mathbf{e}_i^{\text{img}} = \text{Embed}(I_i, S_i)$, $\mathbf{e}_i^{\text{tgt}} = \text{Embed}(T_i)$, $\{\mathbf{e}_{i,k}^{\text{neg}}\}_{k=1}^{H_i}$ are the embeddings of the H_i per-pair hard-negative targets of anchor i, and τ is a fixed softmax temperature of 0.02. For each anchor the contrastive negatives are the other in-batch targets together with its per-pair hard negatives. The in-batch term additionally applies false-negative masking: any in-batch target $j \neq i$ whose similarity to the anchor exceeds that of the true positive by more than a margin $m = 0.1$ ($\langle \mathbf{e}_i^{\text{img}}, \mathbf{e}_j^{\text{tgt}} \rangle > \langle \mathbf{e}_i^{\text{img}}, \mathbf{e}_i^{\text{tgt}} \rangle + m$) is treated as a likely false negative and dropped from the first denominator sum.

4.6. Implementation and training schedule

Training was performed on 64 NVIDIA H20 GPUs (8 nodes × 8 GPUs per node, 97.9 GB each) using the ms-swift [49] framework with DeepSpeed ZeRO-2 [50] optimiser-state and gradient partitioning. We fine-tuned both the vision encoder and the language transformer (§4.2). The released UniDerm is a linear weight-space merge of three weight-decomposed low-rank adaptation (DoRA) [51] fine-tunes that share the schedule below but differ in adapter rank and dropout: rank 16 with LoRA-dropout 0.1, rank 32 with dropout 0.2 and rank 64 with dropout 0.2. Each DoRA fine-tune is merged into the backbone weights and the three resulting models are combined with equal weight by a linear

parameter-space average. We optimised with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, gradient clipping at L_2 norm 1.0), a per-GPU micro-batch of 16 and gradient accumulation of 1 step, giving an effective contrastive batch of 1,024 image–instruction pairs per optimiser step. The learning rate was decayed linearly from a 100-step warmup peak of 5×10^{-5} over 10,000 optimiser steps (about three passes over the instruction-augmented Derm1M dataset). Computation used bfloat16 with fp32 master weights, and the maximum per-sample token length was 4,096. Each training image had its long side drawn uniformly from 512 to 1,536 pixels and rounded so that height and width were integer multiples of 32 (the Qwen3VL-Emb-2B patch size after the 2×2 spatial merger), with pixel area capped at 1,310,720. Augmentation comprised random horizontal and vertical flips, ±15° rotation, random translation of up to 5% of the image size and zoom-in cropping, plus mild brightness and contrast jitter (±0.1). Saturation and hue were left unchanged to preserve dermoscopic colour, itself a diagnostic cue. Total wall-clock training time was approximately 24.6 hours, corresponding to about 1,573 GPU-hours.

4.7. Evaluation metrics

We report a single primary metric per task family, chosen to match the label structure of each benchmark, and quantify all uncertainty by bootstrap resampling.

Classification. For the multiclass benchmarks we report *balanced accuracy* (the mean of the per-class recalls), which weights every disease class equally and is therefore insensitive to the heavy class imbalance of the long-tail dermatology benchmarks. For the two binary melanoma-detection benchmarks (PH2-2 and ISIC20-2), whose malignant prevalence is low (about 2% for ISIC20-2), balanced accuracy and top-1 accuracy are pinned to the majority class, so we instead report the *area under the receiver operating characteristic curve* (AUROC, equivalently the area under the curve, AUC), a threshold-free ranking metric. The Tschandl-948 and Diverse Dermatology Images reader head-to-heads are likewise scored by binary any-malignancy AUROC.

Search. Concept-to-image search is scored by *macro precision-at-K* (the precision among the top-K retrieved images, averaged over concepts). Cross-modal evidence search is scored by *recall-at-K* (the fraction of relevant items retrieved within the top K). The K values are those tabulated in each search table.

Embedding geometry. To characterise a frozen representation independently of any fitted classifier, we report the per-sample *silhouette score* under cosine distance, a head-free measure of how tightly each image sits with its own disease class relative to the nearest other class. Positive values indicate genuine class structure.

Skin-tone strata. Skin tone is reported on the six-level *Fitzpatrick skin type* (FST) scale (FST I–II light, III–IV intermediate, V–VI dark), used for the per-stratum fairness analysis on Diverse Dermatology Images.

Statistical reporting. Unless stated otherwise, every reported value is the mean over a 1,000-resample percentile bootstrap of the test partition, with 95% confidence intervals (CIs), and every comparator is run through the identical harness. We did not adjust for multiple comparisons across the main tables, because each row corresponds to a distinct evaluation axis (linear probing versus zero-shot versus search) chosen *a priori*. The per-class reader p values in Fig. 5b and Extended Data Table ED11 are likewise uncorrected and, being computed against the reader mean, are anti-conservative. Because per-reader responses are not published, the per-class reader comparisons were obtained by Monte-Carlo-sampling each class accuracy from the reader-mean±standard-error summary of Tschandl et al. (2020) Supplementary Table 3 (one draw per simulated reader, $n = 302$),

which gives a conservative lower bound on the true reader spread. Paired AUROC comparisons against a reconstructed reader-consensus ROC use the DeLong test on the same images.

4.8. Inference and evaluation protocols

UniDerm is instruction-aware, so each image embedding is conditioned on a system instruction and one image yields several task-specific embeddings rather than a single fixed vector. At inference the per-task image-side instruction(s), target-text instruction(s) and templates are given in Extended Data Table ED20: few-shot linear probing ensembles six image-side instructions (I1–I6), whereas zero-shot, cross-modal evidence search and concept-to-image search each select a single image-side instruction. Each class prototype is formed as the L_2-normalised mean over the selected target-text-instruction and template pairs. Each image is rendered once at a single long side of 1,536 pixels (768 pixels for the lower-resolution SNU-134 and ISIC20-2) using the backbone's native dynamic resolution, whereas every fixed-grid comparator is encoded at its single training-optimal resolution. Forcing such a comparator onto UniDerm's 1,536-pixel rendering instead degrades it. Under a matched probe, rendering PanDerm at 1,536 pixels rather than its native input lowers its 100%-label score on all four linear-probing benchmarks, by 0.4 to 1.4 percentage points. The verbatim UniDerm instructions and templates, the matched CLIP-comparator prompt ensemble [22], and the exact per-task axis settings are given in Extended Data Tables ED21, ED22 and ED20, respectively.

Linear probing uses one multinomial logistic-regression probe (scikit-learn, LBFGS, max-iter 1,000) on the frozen embeddings, at label fractions $\{10, 30, 50, 100\}$%. For every benchmark these label fractions are stratified sub-samples of the *training* partition alone; no validation split is used, and the official test partition is held out for evaluation throughout. The inverse-regularisation constant is fixed *a priori* rather than tuned per dataset: its default $C = 1$ on the multiclass benchmarks and the stronger $C = 0.01$ on the class-imbalanced binary ISIC20-2. The value quoted here and in the figure captions is a fixed multiplier; the effective scikit-learn inverse-regularisation is $C_{\text{sklearn}} = (d\,K/100)\,C$, with d the embedding dimension and K the number of classes, a deterministic normalisation that keeps the regularisation strength comparable across embeddings of different width and is applied with no per-model or per-dataset search. Bootstrap, confidence-interval and multiple-comparison conventions follow Methods §4.7, and every comparator is run through this identical harness. Multi-view ensembling and the auxiliary class-name text prototypes are intrinsic capabilities of instruction-aware models, not add-on tricks. The same-compute Qwen3VL-Emb-2B comparator is therefore run through the *identical* multi-view, text-prototype-augmented probe, isolating the medical recipe as the only difference. The vision-only comparators (PanDerm, DINOv2, AIMv2-3B, InternViT-6B, UniDerm-VE and UniDerm-VE-lite) have no text tower and appear in the linear-probing comparison only.

A few task-specific choices are not obvious from the table. Zero-shot prediction is the argmax cosine to the class prototypes. DDI is scored by AUROC, so there the malignancy score is instead the temperature-0.1 softmax mass over the four malignant of the six PAD-6 class prototypes (malignant: basal-cell carcinoma, actinic keratosis, squamous-cell carcinoma and melanoma; benign: nevus and seborrheic keratosis). Both strict zero-shot and a PAD-6-fitted cross-dataset probe are evaluated, with no DDI image seen in training. The ten comparators of Fig. 7 repeat that cross-dataset probe on their own frozen embeddings, fitted on the same PAD-6 training partition without class weighting. Patient-context diagnosis re-extracts the image embedding with the case metadata appended as a caption, fused in one forward pass for UniDerm and by image–text mean-fusion for the two-tower comparators. The three caption forms render the same structured PAD-6 metadata as deterministic string templates and never describe the lesion. For cross-modal evidence search on SkinCAP we report Recall@5/10/50. The recipe-component ablation (Extended Data Table ED23) omits the auxiliary

Table ED20 | **Per-task inference configuration for UniDerm.** Each evaluation task fixes the image-side instruction(s), target-text instruction(s), template set and metric. The I1–I6, T1–T2 and P1–P3 identifiers are defined verbatim in Extended Data Table ED21. †Multiclass benchmarks are scored by BACC and the binary PH2-2 and ISIC20-2 by AUROC. Image rendering, linear-probing regularisation, and the patient-context and DDI cross-dataset protocols follow Methods §4.8.

Task	Image instr.	Text instr.	Templates	Metric
Zero-shot classification	I2	T1	P1, P2	BACC†
Few-shot linear probing	I1–I6	T1	P1, P2	BACC†
Cross-modal evidence search	I5	T1	caption	R@5/10/50
Concept-to-image search	I4	T1, T2	P3	macro *P*@5/10/25/50

Table ED21 | **Verbatim wording of the image-side system instructions, target-text instructions and templates used by UniDerm at inference.** The bracketed slot "[name]" is filled with the candidate class name ("[concept]" with the candidate concept phrase). The I/T/P identifiers are referenced by the per-task configuration in Extended Data Table ED20. The matched zero-shot prompt ensemble used by the CLIP-family comparators is listed separately in Extended Data Table ED22.

ID	Role	Verbatim text
Image-side system instructions		
I1	Diagnosis	"Classify the skin disease shown in this dermatological image."
I2	ICD-11 coding	"Assign the ICD-11 code for the disease shown in this image."
I3	Description	"Describe the clinical features of this skin condition."
I4	Morphology	"Identify the visual morphological features of this skin lesion."
I5	Generic representation	"Represent the user's input."
I6	Generic representation	"Represent the given image in one word."
Target-text instructions		
T1	Description matching	"Represent this skin disease description for matching with dermatological images."
T2	Generic representation	"Represent the given text in one word."
Templates		
P1	Diagnosis template	"The diagnosis is [name]."
P2	Full clinical description	a detailed clinical and morphological description of the condition
P3	Concept prompt	"a photo of a skin lesion with [concept]."

text prototypes so that all four arms share one probe configuration. The Tschandl-948 head-to-head uses the exact 948-image reader subset of ISIC2018 task 3 [5].

4.9. Structured pruning and label-free distillation

UniDerm-VE-lite is derived in two steps from UniDerm's visual encoder, by which we mean the ~0.3-billion-parameter ViT-L vision tower of UniDerm taken on its own after supervision-denoising contrastive fine-tuning, with the language tower discarded; the *-VE* suffix records that origin. Neither step uses a diagnostic label or any image beyond the public training corpus of §4.3. That encoder is first pruned structurally rather than by sparsifying individual weights, so that what remains is a smaller dense transformer that runs fast on stock kernels. Depth goes from 24 to 20 blocks (retaining source blocks 0–16, 18, 22 and 23), hidden width from 1,024 to 768, feed-forward width from 4,096 to 1,280 and patch-merger width from 4,096 to 3,072. Attention heads go from 16 to 12 at an unchanged head dimension of 64, so the rotary tables carry over untouched. The output width is held at 2,048, so the pruned model emits an embedding of the same dimension as its teacher and the linear-probing protocol of §4.8 applies to it unchanged. The three DeepStack merger heads are

Table ED22 | **Verbatim wording of the shared zero-shot prompt ensemble used by the CLIP-family comparator models.** These are the seven prompts of the dermatology zero-shot ensemble of Derm1M [22] (Supplementary Table 11), applied to every two-tower CLIP-family comparator under the matched zero-shot protocol (Methods §4.8). The bracketed slot "[name]" is filled with the candidate class name. The corresponding image-side instructions, target-text instruction and templates used by UniDerm are in Extended Data Table ED21.

ID	Verbatim text
C1	"This is a skin image of [name]."
C2	"A skin image of [name]."
C3	"An image of [name], a skin condition."
C4	"[name], a skin disorder, is shown in this image."
C5	"The skin lesion depicted is [name]."
C6	"The skin cancer in this image is [name]."
C7	"This image depicts [name], a type of skin cancer."

dropped, since they serve the language tower alone (§4.10). What remains is 0.1 billion parameters, 31.8% of the 331.4 million it started from, and 35.9 GFLOPs per 224-pixel image against 121.5 (§4.10).

The pruned model is then distilled against the unpruned one, which supplies the training signal in place of any annotation. The student is trained to match the teacher's pooled embedding on unlabelled images from the same public corpus; no diagnostic label, caption or human annotation enters the objective, so the step adds no supervision beyond what UniDerm has already seen. We distil for 4,500 steps at 12 images per device on four H20 GPUs, some 216 thousand images in total, with learning rate 2×10^{-4} and weight decay 0.02. The input resolution is drawn per batch from a mixture spanning 224 to 1,120 pixels, so the student inherits the teacher's behaviour across the scales at which it is evaluated. The run takes about 40 minutes. UniDerm-VE-lite is then evaluated exactly as every other model in this work.

4.10. Inference throughput and compute per image

Throughput is the number of images a model's image encoder completes per second in a single forward pass. The input tensor is already resident on the device, so image loading and preprocessing fall outside the timed region. Every model was timed on the same host: an AMD EPYC 9K84 (Zen 4 "Genoa") server whose virtualised environment (KVM guest) exposes 128 logical cores, arranged as two sockets of 32 cores with two hardware threads each across two NUMA nodes, at a nominal clock of 2.60 GHz, with 256 MiB of L3 cache, AVX-512 support including the BF16 and VNNI extensions, and running PyTorch 2.6.0. Each measurement was pinned with `taskset` to four of those logical cores (cores 0–3) and both thread pools were capped to match (`torch.set_num_threads(4)` and `OMP_NUM_THREADS=4`), with a batch size of one. No accelerator is used at any point in this measurement. Four threads is a small fraction of the 128 available on that host, and is meant to stand for the constrained, GPU-free hardware an under-resourced clinic would deploy on rather than for a well-provisioned research machine. Each figure is the mean of 20 timed iterations after 3 warm-up iterations, or of 5 timed iterations for DINOv2-G, InternViT-6B and UniDerm, which were measured in a later pass. This four-thread setting is a deployment floor chosen to bound the worst case, not a representative speed. On a single H20 GPU, with a batch of eight and bf16 autocast at the same 224-pixel input, UniDerm encodes more than one hundred images per second, so an accelerator of any kind moves it into the hundreds of images per second. Every model receives the same 224-pixel input. The fixed-grid comparators resize any input to their own native resolution, which is 224 pixels for

DINOv2-G and PanDerm. AIMv2-3B and InternViT-6B are the exception: their position embeddings admit no grid smaller than 448 pixels, so they cannot be run at 224 pixels at all. We time those two at 448 pixels and divide by the ratio of token counts, which is four, since a 448-pixel input yields four times as many patches as a 224-pixel one, and report the resulting 224-pixel equivalent (8.5 and 17.7 seconds per image, against 33.9 and 70.8 seconds measured at 448 pixels).

Throughput is a property of the model and the host together, so we report alongside it the arithmetic each model performs on one 224-pixel image, which no host can change. Floating-point operations are counted on the dispatched forward graph with PyTorch's `FlopCounterMode` rather than derived from an analytic formula: every matrix multiplication, convolution and attention call is charged its actual arithmetic, a multiply–accumulate counting as two operations. Element-wise work (normalisation, activation, softmax) is excluded, the usual convention, and accounts for under 2%. Each model is driven through the same code path the throughput measurement times, so the two numbers describe one computation. UniDerm and Qwen3VL-Emb-2B accordingly include the language tower, run over the 74-token sequence the benchmark feeds it (49 merged image tokens and a 26-token instruction template, in which the image placeholder counts once), at 337.4 GFLOPs against 121.5 for the vision tower alone; UniDerm-VE-lite is a vision tower throughout and costs 35.9, against 119.3 for PanDerm and 582.5 for DINOv2-G. AIMv2-3B and InternViT-6B are again the exception, and are counted on the same architecture rebuilt on a 224 grid (1,411.8 and 2,881.2 GFLOPs, against 5,879.2 and 11,944.7 at their native 448).

4.11. Ablation studies

To attribute UniDerm's accuracy gain to specific recipe components rather than to backbone scale or general-domain pretraining, we ran a two-factor leave-one-out ablation on HAM-7 (Extended Data Table ED23), holding the Qwen3VL-Emb-2B backbone, Derm1M training data and fine-tuning configuration fixed; the two factors are the multi-task instruction synthesis (the "Instruction" branch) and the ICD-11 ontology augmentation (the "ICD" branch). Relative to the fine-tuning-only reference (zero-shot BACC 0.661, 10% few-shot 0.674), the two branches are roughly additive in zero-shot (+0.3 and +1.4 percentage points alone, +1.6 combined). At the 10% few-shot level the ICD branch dominates (+6.4 percentage points alone, against +1.3 for Instruction and +1.0 for the full recipe), consistent with its ICD-11 definitions and sibling hard negatives sharpening fine-grained class separation before the probe has seen the full disease vocabulary. As supervision grows the full recipe pulls ahead, reaching BACC 0.877 at 100% labels (against 0.825, 0.831 and 0.815 for the ICD-only, Instruction-only and reference runs) and uniquely winning the "All avg" column (0.766 versus 0.730, 0.752 and 0.743). Together with the UniDerm-VE-lite comparison in Results §2.1, which clears PanDerm at a third of its parameter count, this attributes the gain to the dermatology-specific recipe and the instruction-aware architecture, not to backbone scale or general-domain pretraining.

4.12. Comparator models and naming

Throughout this article we refer to comparator models by short names. Every comparator is a transformer-based (ViT) model evaluated at its publicly released checkpoint. The mapping to each model's original published designation and the specific weight checkpoint we evaluate is as follows. *BiomedCLIP* [40] refers to BiomedCLIP-PubMedBERT-256/ViT-B-16 trained on PMC-15M. *OpenCLIP-G* [41] refers to the OpenCLIP ViT-G/14 weights trained on LAION-2B (laion2b-s34b-b88k checkpoint). *MONET* [21] is the dermatology vision–language model of Kim et al. *DermLIP* [22] refers to the ViT-B/16 DermLIP weights trained on the Derm1M dataset (originally released as DermLIP-ViTB16). *MAKE* [23] refers to the ViT-B/16 MAKE weights trained on Derm1M (originally released as MAKE-

Table ED23 | **Recipe-component 2×2 ablation on HAM-7 [28]: balanced accuracy (BACC) across zero-shot and few-shot linear probing for four configurations sharing one Qwen3VL-Emb-2B backbone [32] and Derm1M training split [22].** The two factors are the Instruction branch (multi-task instruction synthesis) and the ICD branch (ICD-11 ontology augmentation), giving the fine-tuning-only reference (neither), ICD-only, Instruction-only and full UniDerm (both). "All avg" is the mean over zero-shot plus the four few-shot fractions; few-shot cells use the canonical linear-probing sub-sampling. Best per column in **bold**; the grey subscript on each lower row is its change relative to the fine-tuning-only reference (top row). Linear probing uses the multi-view probe *without* text-prototype augmentation so all four arms share one configuration (Methods §4.8).

ICD	Inst.	Zero-shot	Few-shot 10%	Few-shot 30%	Few-shot 50%	Few-shot 100%	All avg
✗	✗	0.661	0.674	0.739	0.760	0.815	0.730
✓	✗	0.664$_{+0.003}$	0.738$_{+0.064}$	0.756$_{+0.017}$	0.777$_{+0.017}$	0.825$_{+0.010}$	0.752$_{+0.022}$
✗	✓	0.675$_{+0.014}$	0.687$_{+0.013}$	0.762$_{+0.023}$	0.760$_{+0.000}$	0.831$_{+0.016}$	0.743$_{+0.013}$
✓	✓	**0.677**$_{+0.016}$	0.684$_{+0.010}$	**0.786**$_{+0.047}$	**0.807**$_{+0.047}$	**0.877**$_{+0.062}$	**0.766**$_{+0.036}$

ViTB16). *PanDermLIP* [22] is our short name for the DermLIP-PD-PMC variant released alongside Derm1M, which combines the DermLIP framework with the PanDerm vision encoder and is trained on Derm1M plus the dermatology subset of PMC-OA. We adopt this more compact name in place of the longer original label. *PanDerm* [8] refers to the ViT-L self-supervised dermatology foundation model of Yan et al. *Qwen3VL-Emb-2B* [32] refers to the off-the-shelf Qwen3VL-Embedding-2B general-domain multimodal-embedding model. *UniDerm-VE-lite* is UniDerm's visual encoder, its ~0.3-billion-parameter ViT-L vision tower taken on its own, structurally pruned to 0.1 billion parameters and distilled against that unpruned encoder without labels (§4.9); the *-VE* suffix records that origin. It is the deployment-oriented model of the accuracy–efficiency comparison, a third the size of PanDerm and 3.1 times faster on the same four CPU threads. The general-domain vision-only models in the accuracy–efficiency comparison (Fig. 2a) retain their original names: *DINOv2* (ViT-G/14) [37], *AIMv2-3B* [38] and *InternViT-6B* [39].

4.13. Data availability

Training data. UniDerm was trained on five publicly available dermatology image sources (Extended Data Table ED17): the Derm1M dataset [22] (~80% of training images), the ISIC2019 dermoscopy archive [20] (with HAM10000 and ISIC2020 images blacklisted to protect the evaluation partitions), the SCIN smartphone-image collection [27], dermatology figure panels filtered from the PubMed Central Open Access Subset [45], and dermatology figure panels drawn from the BIOMEDICA archive [46]. No private clinical cohort was used, and every training image and its source captions are publicly available. The aggregate training corpus is not redistributed by the authors; instead, a released data recipe lists each source and, together with the training pipeline (§4.14), reconstructs the corpus once each source is obtained from its original distributor.

Evaluation data. The eleven evaluation benchmarks are obtained from their original distributors: HAM10000 and the Tschandl-948 reader subset of ISIC2018 task 3 [5, 28], PAD-UFES-20 [29], ISIC2020 [31], PH2 [33], the Diverse Dermatology Images set [7], Derm7pt [35], SkinCAP [36], and the SD-128, SNU-134 and DAFFODIL-5 partitions [10, 30, 34]. Repository DOIs, accession codes and download URLs for every training source and evaluation benchmark are consolidated in the *Dataset access* list below and in the reproducibility release; each dataset is obtained from, and redistributed only by, its original provider under that provider's licence.

Dataset access. *Training-image sources:* Derm1M [22] at https://huggingface.co/datasets/redlessone/Derm1M; the BIOMEDICA dermatology subset [46] at https://huggingface.co/datasets/BIOMEDICA/biomedica_dermatology_subset_parquet; the ISIC2019

archive [20] at https://challenge.isic-archive.com/data/ (BCN20000 component at DOI 10.6084/m9.figshare.24140028); SCIN [27] at https://github.com/google-research-datasets/scin; and the PMC-OA dermatology panels [45] at https://huggingface.co/datasets/axiong/pmc_oa. *Evaluation benchmarks:* HAM10000 and the Tschandl-948/ISIC2018 task-3 reader subset [5, 28] at Harvard Dataverse DOI 10.7910/DVN/DBW86T (https://doi.org/10.7910/DVN/DBW86T); PAD-UFES-20 [29] at Mendeley Data DOI 10.17632/zr7vgbcyr2.1 (https://doi.org/10.17632/zr7vgbcyr2.1); the SIIM-ISIC 2020 dataset [31] at ISIC Archive DOI 10.34970/2020-ds01 (https://doi.org/10.34970/2020-ds01); PH2 [33] at https://www.fc.up.pt/addi/ph2%20database.html; the Diverse Dermatology Images set [7] at Stanford AIMI DOI 10.71718/kqee-3z39 (https://ddi-dataset.github.io); Derm7pt [35] at https://derm.cs.sfu.ca; SkinCAP [36] at https://huggingface.co/datasets/joshuachou/SkinCAP (DOI 10.57967/hf/2256); SD-198 [30] on request from the original authors; the SNU-134 partition [10] at Figshare DOI 10.6084/m9.figshare.6454973; and DAFFODIL-5 [34] at Mendeley Data DOI 10.17632/3hckgznc67.1 (https://doi.org/10.17632/3hckgznc67.1). All URLs were last accessed in July 2026.

Model weights and data. The trained UniDerm weights (https://huggingface.co/Linwei-Chen/UniDerm-model; Apache License 2.0) and the training instruction–target pairs (https://huggingface.co/datasets/Linwei-Chen/UniDerm-data) are publicly released; the released data inherit their source licences, bounded by the most restrictive component, CC BY-NC 4.0 (see Limitations). The base Qwen3VL-Emb-2B backbone is publicly available at https://huggingface.co/Qwen/Qwen3-VL-Embedding-2B under the Apache License 2.0. During peer review the model-weight and dataset repositories are private; referees can access them with the read-only Hugging Face token `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`, set as the `HF_TOKEN` environment variable. A step-by-step reviewer download guide, with the exact `huggingface-cli` commands, is provided in the demo repository at https://github.com/Linwei-Chen/UniDerm-demo. All repositories will be made public under their stated licences upon publication.

4.14. Code availability

The code (training recipe, data-rebuild tooling and evaluation harness) is available at https://github.com/Linwei-Chen/UniDerm, mirrored at https://huggingface.co/Linwei-Chen/UniDerm-code, and released under the Apache License 2.0. An interactive demonstration of UniDerm is provided at https://github.com/Linwei-Chen/UniDerm-demo. The code is shared with referees during peer review and will be public upon publication, with redistribution and commercial reuse supported under the upstream Apache 2.0 obligations.

References

[1] Chante Karimkhani, Robert P Dellavalle, Luc E Coffeng, Carsten Flohr, Roderick J Hay, Sinéad M Langan, Elaine O Nsoesie, Alize J Ferrari, Holly E Erskine, Jonathan I Silverberg, et al. Global skin disease morbidity and mortality: an update from the global burden of disease study 2013. *JAMA Dermatol.*, 153(5):406–412, 2017.

[2] Alize J Ferrari, Damian Francesco Santomauro, Amirali Aali, Yohannes Habtegiorgis Abate, Cristiana Abbafati, Hedayat Abbastabar, Samar Abd ElHafeez, Michael Abdelmasseh, Sherief Abd-Elsalam, Arash Abdollahi, et al. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021. *Lancet*, 403(10440): 2133–2161, 2024.

[3] Junpeng Zhang, Fan Zhong, Kaiqiao He, Mengqi Ji, Shuli Li, and Chunying Li. Recent advancements and perspectives in the diagnosis of skin diseases using machine learning and deep learning: a review. *Diagnostics*, 13(23):3506, 2023.

[4] Aobuliaximu Yakupu, Rehanguli Aimaier, Bo Yuan, Bin Chen, Jia Cheng, Yaohua Zhao, Yinbo Peng, Jiaoyun Dong, and Shuliang Lu. The burden of skin and subcutaneous diseases: findings from the global burden of disease study 2019. *Front. Public Health*, 11:1145513, 2023.

[5] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.*, 26(8):1229–1234, 2020.

[6] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[7] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A C Allerup, Utako Okata-Karigane, James Zou, and Albert S Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.*, 8(32):eabq6147, 2022.

[8] Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nat. Med.*, 31(8):2691–2702, 2025.

[9] Yuan Liu, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.*, 26(6):900–908, 2020.

[10] Seung Seog Han, Ilwoo Park, Sung Eun Chang, Woohyung Lim, Myoung Shin Kim, Gyeong Hun Park, Je Byeong Chae, Chang Hun Huh, and Jung-Im Na. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J. Invest. Dermatol.*, 140(9):1753–1761, 2020.

[11] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, and Xin Gao. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.*, 15(1):5649, 2024.

[12] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, and L. Uhlmann. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.*, 29(8):1836–1842, 2018.

[13] Seung Seog Han, Myoung Shin Kim, Woohyung Lim, Gyeong Hun Park, Ilwoo Park, and Sung Eun Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J. Invest. Dermatol.*, 138(7):1529–1538, 2018.

[14] Titus J. Brinker, Achim Hekler, Alexander H. Enk, Carola Berking, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer*, 113:47–54, 2019.

[15] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P. Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol.*, 20(7):938–947, 2019.

[16] Michael Phillips, Helen Marsden, Wayne Jaffe, Rubeta N. Matin, Gorav N. Wali, Jack Greenhalgh, Emily McGrath, RoShan James, Eleni Ladoyanni, Anthony Bewley, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw. Open*, 2(10):e1913436, 2019.

[17] H. A. Haenssle, C. Fink, F. Toberer, J. Winkler, W. Stolz, T. Deinlein, R. Hofmann-Wellenhof, A. Lallas, S. Emmert, T. Buhl, M. Zutt, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann. Oncol.*, 31 (1):137–143, 2020.

[18] Ayush Jain, David Way, Vishakha Gupta, Yi Gao, Guilherme de Oliveira Marinho, Jay Hartford, Rory Sayres, Kimberly Kanada, Clara Eng, Kunal Nagpal, et al. Development and assessment of an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA Netw. Open*, 4(4):e217249, 2021.

[19] Luis R. Soenksen, Timothy Kassis, Susan T. Conover, Berta Marti-Fuster, Judith S. Birkenfeld, Jason Tucker-Schwartz, Asif Naseem, Robert R. Stavert, Caroline C. Kim, Maryanne M. Senna, et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Sci. Transl. Med.*, 13(581):eabb3652, 2021.

[20] Marc Combalia, Noel Codella, Veronica Rotemberg, Cristina Carrera, Stephen Dusza, David Gutman, et al. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 international skin imaging collaboration grand challenge. *Lancet Digit. Health*, 4(5):e330–e339, 2022.

[21] Chanwoo Kim, Soham Uday Gadgil, Alex J DeGrave, Jesutofunmi A Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nat. Med.*, 30(4):1154–1165, 2024.

[22] Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. Derm1M: a million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12681–12690, 2025.

[23] Siyuan Yan, Xieji Li, Ming Hu, Yiwen Jiang, Zhen Yu, and Zongyuan Ge. MAKE: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964 of *Lecture Notes in Computer Science*, pages 369–379. Springer, 2025.

[24] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: evaluating CLIP-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[26] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations (ICLR)*, 2025.

[27] Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, et al. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Netw. Open*, 7(11):e2446615, 2024.

[28] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5(1): 180161, 2018.

[29] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fabio CR Alves Jr, Jose GM Esgario, Alana C Simora, Pedro BC Castro, et al. PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data Brief*, 32:106221, 2020.

[30] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision (ECCV)*, pages 206–222. Springer, 2016.

[31] International Skin Imaging Collaboration. SIIM-ISIC 2020 challenge dataset. ISIC Archive, 2020.

[32] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: a unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv*, 2601.04720, 2026.

[33] Teresa Mendonça, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marcal, and Jorge Rozeira. PH2 — a dermoscopic image database for research and benchmarking. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5437–5440, 2013.

[34] Sharun Akter Khushbu. Skin disease classification dataset. Mendeley Data, 2024.

[35] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. 7-point checklist and skin lesion classification using multi-task multi-modal neural nets. *IEEE J. Biomed. Health Inform.*, 23(2):538–546, 2019.

[36] Yuhao Shen, Liyuan Sun, Yan Xu, Wenbin Liu, Shuping Zhang, Shawn Afvari, Zhongyi Han, Jiaoyan Song, Yongzhi Ji, Tao Lu, Xiaonan He, Xin Gao, and Juexiao Zhou. SkinCAP: a multimodal dermatology dataset annotated with rich medical captions. *arXiv*, 2405.18004, 2024.

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

[38] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, et al. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[39] Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: exploring advanced training and test-time recipes for open-source multimodal models. *arXiv*, 2504.10479, 2025.

[40] Sheng Zhang, Yanbo Xu, Naoto Usuyama, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2025.

[41] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2023.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[43] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.*, 6:1399–1406, 2022.

[44] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023.

[45] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-CLIP: contrastive language-image pre-training using biomedical documents. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 14227 of *Lecture Notes in Computer Science*, pages 525–536. Springer, 2023.

[46] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J. Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, Anita Rau, Xiaohan Wang, Yuhui Zhang, Alfred Seunghoon Song, Robert Tibshirani, and Serena Yeung-Levy. BIOMEDICA: an open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[47] Jonathan M L White, Harvey Lui, Christopher G Chute, Robert Jakob, and Rachel J G Chalmers. The WHO ICD-11 classification of dermatological diseases: a new comprehensive online skin disease taxonomy designed by and for dermatologists. *Br. J. Dermatol.*, 186(1):178–179, 2022.

[48] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the ISIC image datasets: usage, benchmarks and recommendations. *Med. Image Anal.*, 75:102305, 2022.

[49] Yuze Zhao, Jintao Huang, Jinghan Hu, et al. SWIFT: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, No. 28*, pages 29733–29735, 2025.

[50] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3505–3506, 2020.

[51] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.