How Medical VLMs Underutilize Their Vision Encoders: A Dermatology Perspective

Anonymous authors
Paper under double-blind review

Abstract

Medical Vision-Language Models (VLMs) show significant promise for clinical image understanding, offering accurate diagnosis with interpretable reasoning. However, a critical performance gap exists between their strong vision encoders and the full multimodal model: in dermatology, the MedSigLIP encoder outperforms MedGemma by an average of 10.26 percentage points even when both use zero target-task labels; few-shot linear probing provides further evidence of strong visual representations. This gap motivates an investigation of how visual information is used in end-to-end diagnosis and why plausible-sounding predictions can lack grounding in image evidence. Using dermatology as our primary testbed, we systematically investigate three hypotheses for this phenomenon. We further provide a mechanistic analysis of the model's internal attention patterns, showing that a simple describe-then-decide prompting strategy increases vision attention by 30–40% during generation. Task-specific fine-tuning improves dermatology classification but reduces cross-domain medical question-answering performance in our evaluation. To address these challenges, we combine label-free prompting with low-label encoder-assisted reranking while keeping the VLM frozen. We validate the interventions across five VLM backbones in dermatology and provide supporting representation and attention analyses across additional medical modalities.

1 Introduction

Diagnosing conditions from dermatological images is a challenging task due to inherent complexities, such as subtle variations in disease presentation and the lack of standardized image quality. While advances in computer vision have enabled diagnostic models to achieve expert-level accuracy (Brinker et al., 2019; Esteva et al., 2017; Liu et al., 2020; Soenksen et al., 2021; Wang et al., 2024), their deployment in safety-critical domains like healthcare requires more than just precision. It is crucial that a model's decisions are interpretable, ensuring that they are grounded in clinical evidence rather than spurious correlations. Recent emergence of Vision-Language Models (VLMs) in dermatology addresses this need, aiming not only for accurate diagnosis but also for explainable reasoning, thereby paving the way for the reliable, real-world application of medical AI.

The development of dermatological VLMs has often relied on pre-training with large-scale, specialized datasets and advanced architectures. For example, DermLIP is a CLIP-based VLM (Radford et al., 2021) that was pre-trained on Derm1M, a dataset with over one million image-text pairs, enabling tasks like zero-shot classification and concept identification (Yan et al., 2025a). Similarly, MONET (Kim et al., 2024) has fine-tuned a CLIP model on over 100K dermatological images paired with natural language descriptions from a large collection of medical literature, aiming for interpretable diagnoses. Additionally, SkinVL (Zeng et al., 2025) and SkinGPT-4 (Zhou et al., 2024) have integrated large language models into their dermatological VLMs and leveraged large dermatological corpora to facilitate classification with nuanced disease interpretation and visual question answering. More recently, MedGemma, a family of medical foundation models, has emerged to demonstrate superior medical reasoning capabilities (Sellergren et al., 2025). Built upon the powerful Gemma 3 (Team et al., 2025) architecture and incorporating a medically tuned vision encoder MedSigLIP, MedGemma excels at transparent reasoning, making it an ideal testbed for clinical applications.

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

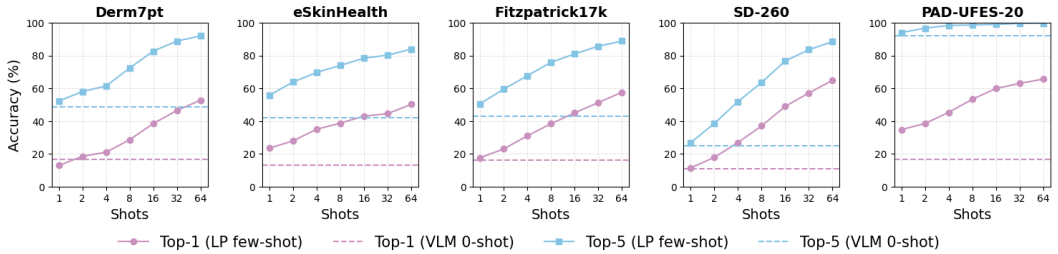


Figure 1: A comparison of the zero-shot performance of the MedGemma-4B VLM against the linear probe performance of its vision encoder, MedSigLIP. LP denotes a linear probe on the vision encoder, while VLM denotes a direct query for diagnosis to the full vision-language model. The comparison shows the diagnostic information accessible through linear probing; the label-free control is reported in Table 1.

In this work, we identify a critical paradox: while MedGemma's vision encoder, MedSigLIP, possesses exceptional discriminative power, this strength fails to translate into the zero-shot diagnostic accuracy of the full multimodal model. As shown in Fig. 1, we observe that a simple few-shot linear probe on MedGemma's vision encoder, which is medically tuned from SigLIP (Zhai et al., 2023a), achieves remarkable classification accuracy across five skin disease datasets. Direct comparisons reveal that MedSigLIP is comparable to, and at times superior to, PanDerm (Yan et al., 2025b), a specialized dermatology vision encoder, as shown in Appendix Table 11. The zero-label comparison in Table 1 further shows that this gap persists without fitting a classifier on target-task labels. Although fine-tuning improves in-domain accuracy, applying LoRA (Hu et al., 2022) to MedGemma still underperforms a linear probe (Appendix Table 11). The Fitzpatrick17k-adapted checkpoint also loses 5.04 and 15.00 percentage points on closed- and open-ended VQA-RAD accuracy, respectively (Appendix C).

This discrepancy leads us to two fundamental questions: **(1)** What factors contribute to the large performance gap between MedGemma's highly capable vision encoder and its end-to-end zero-shot diagnostic performance? **(2)** Can we develop inference-time strategies to narrow this gap while keeping the VLM frozen? To answer these questions, our work makes the following contributions:

1. Using dermatology as our primary testbed, we document a vision–language performance gap under zero target-task labels and compare multimodal image ICL with linear probing and encoder-assisted inference. Additional medical modalities provide supporting evidence of strong frozen visual representations.
2. We investigate three hypotheses for the gap (*training distribution mismatch*, *vision branch under-reliance*, and *encoder-decoder objective misalignment*), using controlled prompts and attention analysis to relate targeted interventions to changes in model behavior.
3. Building on these observations, we combine clinical descriptions, DtD, and low-label candidate reranking without fine-tuning the VLM, with improvements across MedGemma, SkinVL, InternVL3, InternVL3.5, and biomed-Qwen in the evaluated dermatology settings.

Our investigation is grounded in dermatology using MedGemma-4B, with additional backbones and medical modalities testing the scope of the observed behavior. The over-reliance on language priors and modality misalignment are fundamental issues facing the broader field of vision-language research (Zhai et al., 2023b; Yang et al., 2025; Hu et al., 2024; Tong et al., 2024). Therefore, our analysis provides valuable insights into a generalizable approach to improving the reliability and visual grounding of VLMs in safety-critical applications beyond dermatology or healthcare.

2 Related Work

Foundation Medical Vision-Language Models for Dermatology. Recent advancements in dermatological VLMs span specialized curricula and architecture choices (Zhang et al., 2024; Lin et al., 2023). DermLIP was trained on Derm1M, a one-million image-text pair dataset, enabling zero-shot classification and concept identification (Yan et al., 2025a); MONET fine-tuned CLIP on over 100K image-caption pairs from medical literature for transparent, concept-based diagnoses (Kim et al., 2024). Similar curriculum strategies have improved feature learning in other medical VLMs such as

retinal image analysis (Holland et al., 2025), while in-context learning has recently been shown to enable multimodal LLMs to classify cancer pathology images without fine-tuning (Ferber et al., 2024). SkinVL (Zeng et al., 2025) and SkinGPT-4 (Zhou et al., 2024) integrate LLMs for more nuanced disease interpretation and visual question answering: SkinVL is trained on $\sim$10K image–captions and 27K textbook-derived QA pairs, while SkinGPT-4 is aligned with Llama-2 (Touvron et al., 2023) on over 52K skin disease images. While these models perform strongly on dermatology-specific tasks, our work centers on MedGemma (Team et al., 2025; Sellergren et al., 2025), a generalist medical foundation model whose medically-tuned vision encoder (MedSigLIP) and advanced language backbone (Gemma 3) make it an ideal testbed for the vision–language alignment challenges relevant to next-generation medical AI.

Misalignment of VLMs and Over-reliance on Language Priors. A central challenge in current VLMs is the misalignment between their powerful vision encoders and the end-to-end multimodal system (Xing et al., 2025), often manifesting as over-reliance on language priors. Studies show that as a VLM generates its response, attention to visual input gradually diminishes, sometimes having negligible influence on the final output (Yang et al., 2025)—a tendency to "look away" from the image that yields plausible-sounding but visually ungrounded responses and hallucinations. (Tong et al., 2024) further showed that even advanced models like GPT-4V exhibit elementary visual shortcomings traceable to weaknesses in the underlying CLIP encoder, struggling with basic patterns of orientation, counting, and object attributes. This misalignment is exacerbated during instruction tuning: Zhai et al. (2023b) reported that fine-tuning often degrades the model's core visual perception, eroding the classification ability of its original vision encoder. To mitigate such inconsistencies, recent work has proposed automated structured reporting to standardize outputs and reduce hallucination (Delbrouck et al., 2025). Taken together, these challenges underscore the need for strategies that enforce strong visual grounding in VLMs' final reasoning.

3 Evaluation Setup and Baseline Comparisons

Models. The MedGemma family of multimodal models is available in two variants: a 4B and a 27B parameter version. Both models employ the MedSigLIP image encoder, which was pre-trained on a diverse corpus of de-identified medical data, including chest X-rays, dermatology images, ophthalmology images, and histopathology slides. The primary distinction between the two is that the 27B variant was also pre-trained on an additional medical corpus that includes electronic health records (EHRs). As our research focuses on visual diagnosis from images rather than EHR data, MedGemma-4B is a more suitable choice for this study. We formalize the interaction with the VLM $\mathcal{M}$ as $\mathcal{M}[\text{Prompt}, \text{Image}] \rightarrow \text{Response}$, where the model receives a textual prompt and an image of a skin condition as inputs and generates a response containing a diagnosis and an explanation.

Datasets. To ensure a comprehensive and robust evaluation, our study considers five distinct datasets selected to cover a wide spectrum of diseases and patient demographics. For broad coverage, we include SD-260 (Yang et al., 2019), which contains clinical images across 260 skin conditions, and Fitzpatrick17k (Groh et al., 2021), which features light-skinned images annotated with Fitzpatrick Skin Types. Following the methodology of Wang et al. (2025b) to leverage their disease checklists, we use the same subset of Fitzpatrick17k as was used in their work. To assess performance on more specific diseases and diverse populations, we incorporate Derm7pt (Kawahara et al., 2019), focusing on its clinical photos of skin cancer, and eSkinHealth (Wang et al., 2025a), a specialized dataset of Neglected Tropical Diseases (NTDs) in West African populations. Finally, to test the model's performance on familiar data, we also include PAD-UFES-20 (Pacheco et al., 2020), an in-domain dataset of skin lesions that was used during MedGemma's pre-training. For datasets that do not provide an official train-test split, we partitioned the data by case and class with a 0.5 split ratio, with further details provided in Table 11 and Appendix Table 23. Implementation details and hyperparameter tuning results for VLM fine-tuning and linear probing are provided in Appendix I. Additional backbones are InternVL3-4B (Zhu et al., 2025), InternVL3.5-4B (Wang et al., 2025c), and biomed-Qwen2.5-VL-3B (Cheng et al., 2025).

Label requirements. Clinical descriptions and DtD use no labeled target images. The Top-5-to-Top-1 route trains a lightweight linear probe on a small labeled support set to shortlist candidates, then uses the frozen VLM for prediction. This route is low-label and VLM-fine-tuning-free. Multimodal image ICL supplies labeled image–class pairs as demonstrations and is distinct from the clinical-description context used in our prompt-only experiments.

Table 1: **The encoder–VLM gap persists without target-task labels.** Top-1 accuracy (%). The zero-label columns use no fitted classifier; the 8-shot columns compare image ICL, LP, and LP-assisted Top-5-to-Top-1 prediction. A dash marks an unreported setting. LP and reranking variability and the 1-shot results are retained in Appendix Table 9.

	Zero target labels		8-shot		
Dataset	VLM	MedSigLIP	Image ICL	LP	Reranking
Derm7pt	16.46	24.18	19.44	28.66	35.56
eSkinHealth	13.35	26.53	18.15	38.80	42.05
Fitzpatrick17k	16.73	28.37	20.72	38.43	37.82
SD-260	10.81	—	—	37.12	42.33
PAD-UFES-20	46.91	55.42	49.58	53.33	80.31

Zero-label and image-ICL comparisons. Table 1 compares MedGemma with MedSigLIP evaluated directly through image–text similarity using the same class labels. MedSigLIP exceeds the VLM on all four evaluated datasets by 7.72–13.18 percentage points (10.26 on average), showing that the gap persists without target-task supervision. The 8-shot image-ICL baseline improves over zero-shot MedGemma by 2.67–4.80 points, while the 8-shot LP has higher accuracy on all four datasets. Appendix A includes BiomedCLIP (Zhang et al., 2025) and DermLIP (Yan et al., 2025a) zero-shot comparisons and further protocol details.

4 Methods

4.1 Hypothesis 1: Train-test Distribution Mismatch

We first hypothesize that the performance gap stems from a train-test distribution mismatch. That is, the model may not have learned sufficient representations for certain conditions if they were rare or absent in its training data. This distribution mismatch prevents the model from generalizing to these "tail" or out-of-distribution classes during zero-shot inference.

Evidence. While the exact composition of MedGemma's training data is not public, we can investigate this hypothesis by analyzing the model's prediction bias. The confusion matrix in Fig. 2 clearly demonstrates this bias: the model shows a strong tendency to over-predict certain common conditions, such as basal cell carcinoma (BCC). For instance, nearly all cases of lupus erythematosus (LE) are misclassified as BCC. This raises a critical question: is this error due to a bias in the visual pre-training data (a vision branch issue) or from an incomplete conceptual understanding in the language model (a language branch issue)? By leveraging the model's reasoning capabilities, we can probe its conceptual knowledge. We prompted the model to describe the characteristics of diseases it consistently fails to diagnose (i.e., those with 0% zero-shot accuracy). As shown in Appendix Table 25, the model generates detailed descriptions for both BCC and LE, indicating access to relevant disease concepts. These examples motivate the distribution-mismatch hypothesis, while the controlled prompts below test whether class-aligned clinical information improves diagnosis.

Solution. Although the model may lack sufficient visual examples of rare diseases, it possesses a foundational knowledge of general dermatological terms (e.g., "papules," "inflammation," and "reddish"). We can leverage this by re-framing the task from recognizing a disease name to matching visual evidence with a clinical description. Following the methodology of Wang et al. (2025b); Ferber et al. (2024), our solution provides the model with contextual information in the prompt. Alongside each potential disease name in the list of options, we include a description of its key visual characteristics using a set of generic dermatological vocabularies. Instead of forcing the model to map visual cues to a potentially unfamiliar disease label, this approach encourages it to map visual cues to a provided description. As shown in the improved confusion matrix in Fig. 2 (b), this strategy effectively guides the model toward more accurate and diverse predictions. See visual descriptions in Appendix Table 24 and complete prompt examples in Appendix Table 16.

Context controls. Class-aligned descriptions outperform both length-matched generic dermatology text and descriptions shuffled across class labels (Table 2). On Fitzpatrick17k, the three accuracies are

Table 2: **Clinical content matters beyond prompt length.** Top-1 accuracy (%) with zero labeled target images. The complete comparison appears in Appendix Table 10.

Prompt context	Fitzpatrick17k	PAD-UFES-20
Names only	16.73	46.22
Class-aligned descriptions	24.01	52.23
Length-matched generic text	16.68	36.36
Shuffled descriptions	12.71	15.64

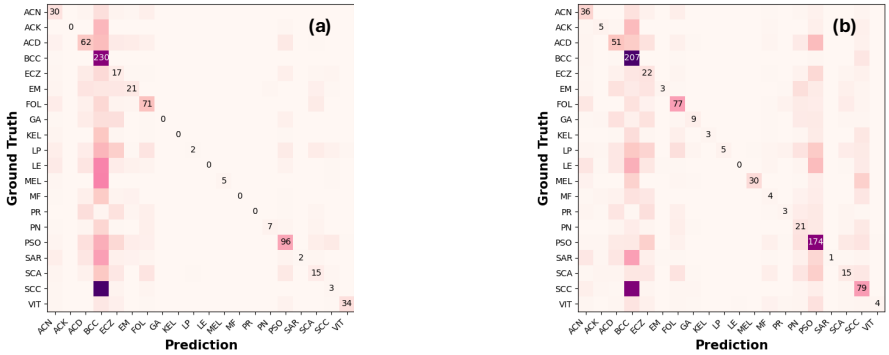


Figure 2: **Clinical context changes the pattern of diagnostic errors.** Fitzpatrick17k confusion matrices for (a) direct answering and (b) clinical descriptions. The full four-strategy comparison, including DtD and 2-shot candidate filtering, appears in Appendix Fig. 5; class abbreviations are listed in Appendix Table 22.

24.01%, 16.68%, and 12.71%; on PAD-UFES-20, they are 52.23%, 36.36%, and 15.64%. Correctly aligned clinical content therefore contributes beyond simply lengthening the prompt.

4.2 Hypothesis 2: Under-reliance on the Vision Branch

We hypothesize that the VLM's powerful, heavy-weight language model component leads it to over-rely on textual priors when making a diagnosis. This behavior, documented in prior works (Yang et al., 2025; Tong et al., 2024), causes the model to generate a plausible-sounding diagnosis without sufficiently grounding its prediction and reasoning in the visual evidence of the image.

Evidence. We first observe that the model is prone to hallucinated reasoning, generating explanations that are not grounded in the image. As shown in Fig. 3, when presented with an image of "sarcoidosis," MedGemma incorrectly identifies the condition as BCC and provides a factually incorrect explanation, as there is no visual evidence of a "pearly or waxy" lesion. Similar observations are made for an image of melanoma, even when the lesion is intentionally masked. In these cases, the model tends to provide a generic, post hoc explanation for BCC, indicating that its diagnosis is not derived from the image's visual features. To quantify this behavior, we masked 200 images and tested whether the model would abstain from making a diagnosis when the primary lesion was hidden. As shown in Appendix Fig. 9(a), the model failed to abstain in 88% of cases, instead defaulting to a BCC diagnosis. This provides evidence that the model makes predictions without using its "eyes."

Solution. To mitigate this over-reliance on language priors, we introduce a "describe-then-decide" (DtD) prompting strategy. This method asks the model to articulate the visible features of the lesion before making a final diagnosis (Appendix Table 16). The premise is that an accurate diagnosis should be preceded by an accurate visual description. This simple intervention yields improved classification accuracy, as shown in the enhanced confusion matrix in Appendix Fig. 5(c) and Table. 3. Additionally, when presented again with the 200 masked images, the model's abstention rate increased from 12% to 72%, evidencing that this prompt encourages visual grounding.

To examine how DtD changes visual attention, we compare heatmaps for the final diagnosis tokens and for the entire generated response. The examples show more lesion-focused attention under DtD; the heatmaps and their construction are provided in Appendix E. We complement this qualitative comparison with the layer-wise analysis in Sec. 5.

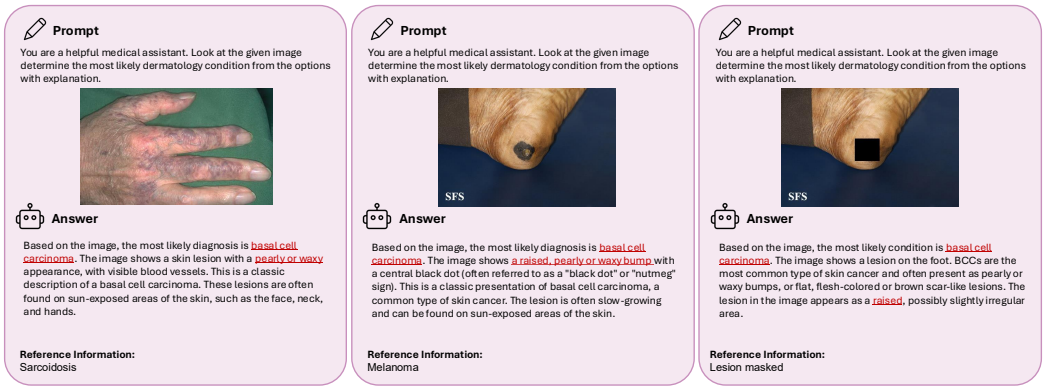


Figure 3: Examples of incorrect answers and hallucinated explanations from MedGemma-4B. The model incorrectly diagnoses sarcoidosis (left) and melanoma (middle) as basal cell carcinoma, providing descriptions of features that are not present in the images. In the third example (right), the lesion is manually masked, yet the model generates the same incorrect diagnosis and explanation, indicating it may generate a prediction first and rationalize it with a post-hoc explanation.

4.3 Hypothesis 3: Misalignment Between the Language and Vision Branches

While DtD improves visual grounding, it does not fully close the performance gap. This leads to our third hypothesis: a fundamental objective misalignment between the VLM's components. The vision encoder is optimized for discriminative tasks (distinguishing classes), whereas the large language model is optimized for generative tasks (explaining and reasoning). We hypothesize that the VLM is better suited for explanation and description than for forced-choice classification and that leveraging these distinct strengths is key to maximizing performance.

Evidence. Previous sections have already established that the vision encoder's discriminative performance surpasses that of the VLM. To complete our evidence, we now evaluate the VLM's primary strength: its ability to generate high-quality, descriptive text. Following the methodology of Wang et al. (2025b), we prompted MedGemma-4B to describe lesions using a structured format covering five key clinical criteria: [*Location Site, Lesion Type, Shape/Border, Color, Texture*]. An example of a generated description is shown in Appendix Fig. 11. To quantitatively assess quality, the model was instructed to review the generated text against the original image and verify the accuracy of each of the five requested criteria. The resulting self-evaluation score (0–5) counts the criteria that the model judges to be correctly described. As shown in Fig. 4c, the model assigns high scores (predominantly 4 or 5) to its descriptions. This performance stands in stark contrast to the hallucinated and visually ungrounded explanations observed during the standard zero-shot diagnosis (see Fig. 3), suggesting a difference between descriptive and classification behavior that motivates encoder-assisted candidate selection.

Solution. Our solution is a two-stage pipeline that leverages the complementary strengths of the vision and language components. Instead of forcing the VLM to perform a task it is not optimized for, we delegate the initial discriminative work to the vision encoder. First, we use the powerful vision encoder with a few-shot linear probe to identify the top-5 most likely diagnosis candidates for a given image. This narrows the field to a small set of high-probability candidates. Then, we feed these top-5 candidates to the full VLM and prompt it to make the final diagnosis from only that reduced set. We term this strategy as "Top-5 to Top-1."

As shown in Table 1, this approach is highly effective, especially in low-data regimes. With 1 to 8 shots, the two-stage pipeline improves over a linear probe in most evaluated settings; at 8 shots on Fitzpatrick17k, it obtains 37.82% compared with 38.43% for LP. The benefit diminishes as the number of shots increases (e.g., 16 or more), as the linear probe becomes powerful enough on its own (see Appendix I.4). This makes our approach particularly valuable for real-world medical applications where labeled data is scarce.

This strategy offers three additional benefits. First, by providing the VLM with only 5 options instead of 20 or more, it reduces the prompt length, saving token space for more context. Second,

Table 3: Ablation study on the effectiveness of each proposed strategy. This table presents Top-1 accuracy for five inference configurations. Clinical descriptions and DtD are label-free; candidate filtering uses target labels. The "DA" column serves as the baseline, using a standard classification prompt. The "in-context" strategy augments the prompt with clinical descriptions; "DtD" requires the model to describe visual features before diagnosis; and "Top-5 to Top-1" is our two-stage strategy where the VLM reranks candidates from a 8-shot linear probe. The final "all combined" column integrates all proposed strategies.

Dataset	DA	in-context	DtD	top-5 to top-1	all combined
Derm7pt	16.46 (± 0.17)	26.72 (± 0.82)	30.98 (± 1.37)	35.10 (± 1.95)	38.27 (± 0.81)
eSkinHealth	13.35 (± 0.23)	30.07 (± 1.51)	29.77 (± 0.92)	43.94 (± 1.03)	48.00 (± 0.86)
Fitzpatrick17k	16.73 (± 0.25)	24.01 (± 2.01)	24.75 (± 1.53)	28.32 (± 1.67)	40.97 (± 1.41)
SD-260 (subset)	24.31 (± 0.15)	34.29 (± 1.06)	33.80 (± 1.37)	40.01 (± 2.18)	44.90 (± 0.91)
PAD-UFES-20	46.22 (± 0.17)	52.23 (± 1.26)	62.33 (± 1.58)	75.23 (± 2.07)	84.68 (± 1.80)

Table 4: Generalizability on SkinVL (LLaVA-based). Top-1 Accuracy (%) comparison. Our proposed inference pipeline consistently improves performance over the direct answer baseline, demonstrating robustness across different VLM architectures.

Inference Strategy	Derm7pt	eSkinHealth	Fitzpatrick17k
Direct Answer (Baseline)	12.77 (± 0.12)	23.76 (± 0.21)	10.05 (± 0.08)
+ Top-5 Filtering	29.66 (± 1.45)	40.25 (± 1.82)	58.37 (± 1.15)
+ In-Context Description	31.19 (± 1.33)	44.03 (± 2.01)	62.97 (± 1.28)
+ DtD (Ours)	**32.56** (± 0.95)	**44.89** (± 1.67)	**63.91** (± 1.05)

the vision encoder is a plug-and-play component; practitioners can substitute MedSigLIP with any other powerful, domain-specific vision encoder to potentially boost performance further. Third, and perhaps most importantly, we still have the model's reasoning capabilities at our disposal.

4.4 Ablation Study and Combined Performance

Having investigated three hypotheses for the performance gap and proposed a corresponding intervention for each, we now conduct an ablation study to evaluate the individual and combined effects of our strategies. We compare the following five inference configurations, with the results presented in Table. 3. For the "Top-5 to Top-1" and "all combined" configurations, we use a 8-shot linear probe to generate the initial candidates. We choose 8 shots because, as shown in Table 20, it represents a point of high efficacy in the low-data regime before performance gains begin to saturate. For the SD-260 dataset, providing expert-verified clinical descriptions for all 260 classes was infeasible. Therefore, for this study, we use a subset of SD-260 containing only the classes that overlap with the other datasets. Example prompts are provided in Appendix Table 16.

The results of the ablation study provides evidence for the effectiveness of the proposed fine-tuning-free pipeline. We observe that each of the three strategies (in-context, DtD, and "Top-5 to Top-1") independently yields a significant improvement in Top-1 accuracy over the direct answer baseline across all five datasets. These improvements support the practical utility of the three complementary interventions. Most importantly, the "all combined" design consistently achieves the highest performance, substantially outperforming any single strategy. The combined results show that the interventions can work together to narrow the performance gap and enhance diagnostic accuracy.

4.5 Generalization to Other VLMs for Dermatology

To demonstrate that the identified challenges and proposed solutions are not specific to the MedGemma architecture, we extended our evaluation to SkinVL (Zeng et al., 2025). As a LLaVA-based model, SkinVL represents a distinct architectural lineage from MedGemma. First, we confirmed that the performance gap is a pervasive issue: consistent with MedGemma, a simple linear probe on SkinVL's vision encoder significantly outperforms its zero-shot VLM accuracy (see Appendix J for full comparison). Second, we evaluated our fine-tuning-free pipeline on SkinVL. Due to its stricter context window, we adapted the pipeline to perform "Top-5 to Top-1" filtering first. As shown in Table 4, our strategies yield consistent performance gains across multiple datasets, even improving

Table 5: **The interventions transfer to additional VLM backbones.** Top-1 accuracy (%) on Derm7pt (D7) and Fitzpatrick17k (F17). DA: direct answer; DtD: describe-then-decide; Top-5: 8-shot candidate-assisted Top-5-to-Top-1 inference. Full results are listed in Appendix Table 12.

Model	Dataset	DA	DtD	Top-5
InternVL3-4B	D7	5.68	6.20	14.83
	F17	10.10	12.23	19.70
InternVL3.5-4B	D7	7.21	7.85	18.32
	F17	9.25	13.15	22.16
biomed-Qwen2.5-VL-3B	D7	20.93	38.50	41.24
	F17	12.14	19.96	22.35

upon the high baselines of Fitzpatrick17k (a dataset included in SkinVL's pre-training). This supports transfer to the LLaVA-based SkinVL architecture. We further evaluate InternVL3-4B, InternVL3.5-4B, and biomed-Qwen2.5-VL-3B on Derm7pt and Fitzpatrick17k (Table 5). DtD improves over direct answering in all six model–dataset pairs, and candidate-assisted inference gives the highest accuracy among the three settings. On InternVL3.5, Top-5-to-Top-1 improves over DA by 11.11 and 12.91 percentage points, respectively. Appendix B provides the extended comparison.

5 Mechanistic Analysis of Vision Attention

Inference-time strategies narrow the vision–language gap, but they do not explain why the gap exists in the first place. Why does the model underutilize its vision encoder? To answer this, we analyze internal attention patterns during autoregressive generation.

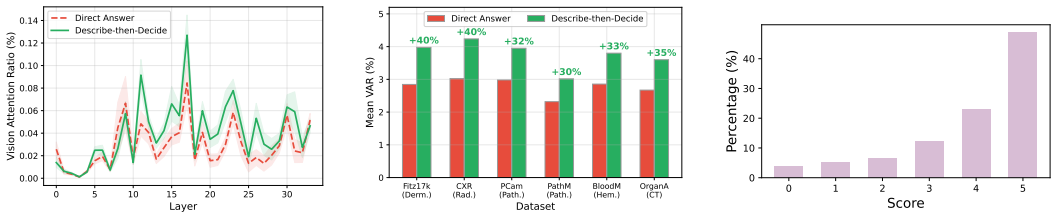
Vision Attention Ratio. We define the Vision Attention Ratio (VAR) as the fraction of attention allocated to vision tokens during generation. For decoder layer l and generation step t:

$$\mathrm{VAR}_{l,t} = \frac{\sum_{i \in \mathcal{V}} A_{l,t,i}}{\sum_j A_{l,t,j}},$$

where $A_{l,t}$ is the head-averaged attention distribution of the newly generated token over all key–value positions, and $\mathcal{V}$ indexes vision-token positions. Per-layer VAR is averaged over all generation steps and test images, with higher values indicating more attention allocated to visual tokens. Two prompting conditions are compared: **Direct Answer (DA)**, where the model classifies directly, and **Describe-then-Decide (DtD)**, where the model first describes the image and then classifies. Full attention weights are extracted using eager attention across all 34 transformer layers of MedGemma-4B, with 100 test images per dataset.

DtD Increases Vision Attention Across Modalities. On Fitz17k (Fig. 4a), DA allocates only 2.85% of generation-time attention to vision tokens, indicating that the model largely ignores the image when classifying directly. DtD raises this to 3.98%, a relative gain of 40%. The effect generalizes: Fig. 4b shows DtD producing 30–40% higher mean VAR than DA across all six modalities, with the largest gains in radiology (CheXpert, +40%) and dermatology (Fitz17k, +40%). The mechanism is straightforward. The description phase forces the model to continuously consult vision tokens, whereas under DA a classification label can be produced from language priors alone. The increased visual attention accompanies the accuracy gains reported in Sec. 4.2. The per-layer profile (Fig. 4a) reveals periodic VAR spikes at layers 5, 10, 17, 22, 27, and 33. These match the global attention layers in Gemma 3's sliding-window architecture (Team et al., 2025); the remaining layers use local sliding-window attention and attend minimally to vision tokens. DtD increases VAR most at these global layers. Attention entropy tells the same story: DtD produces lower entropy than DA across all six modalities, indicating more focused and selective attention (Appendix G).

Generality Across Medical Modalities. We further evaluate five additional medical modalities: CheXpert (CXR) (Irvin et al., 2019), a 5-class chest X-ray dataset; PatchCamelyon (PCam) (Veeling et al., 2018), binary histopathology; and three MedMNIST (Yang et al., 2023) benchmarks, PathMNIST (PathM, 9-class colorectal pathology), BloodMNIST (BloodM, 8-class hematology), and OrganAMNIST (OrganA, 11-class abdominal CT). *The gap extends across the evaluated modalities.* Table 6 shows that 8-shot linear probing of the frozen encoder outperforms zero-shot MedGemma on all five additional datasets. These comparisons extend the representation-level observation beyond



(a) Fitz17k per-layer VAR (b) Mean VAR across modalities (c) Description quality scores

Figure 4: Vision Attention Ratio (VAR) and description quality scores. **(a)** Per-layer VAR for Fitz17k shows that Describe-then-Decide (DtD) increases vision attention across all layers, with the most pronounced gains at global attention layers 5, 10, 17, 22, 27, and 33. **(b)** Mean VAR across six medical imaging datasets shows that DtD increases attention allocated to visual tokens by 30–40%. **(c)** Distribution of 0–5 quality scores for 100 generated descriptions, based on the five clinical criteria described in Sec. 4.3.

Table 6: Few-shot linear probe accuracy (%) on frozen MedSigLIP features vs. MedGemma-4B VLM zero-shot classification across medical imaging modalities. LP results are averaged over 5 runs. At 8 shots, LP outperforms the zero-shot VLM on every dataset.

			Linear Probe				VLM
Dataset	**Modality**	**#Cls**	**1-shot**	**4-shot**	**8-shot**	**16-shot**	**0-shot**
CXR	Radiology	5	24.6 (± 2.1)	54.6 (± 1.5)	59.3 (± 1.2)	60.9 (± 0.9)	36.1 (± 0.5)
PCam	Pathology	2	58.2 (± 2.4)	78.9 (± 1.3)	82.2 (± 0.8)	81.5 (± 1.0)	70.9 (± 0.8)
PathM	Pathology	9	78.2 (± 1.6)	90.9 (± 0.9)	93.5 (± 0.6)	94.3 (± 0.5)	40.6 (± 0.6)
BloodM	Hematology	8	61.3 (± 2.3)	78.1 (± 1.4)	86.3 (± 0.9)	89.0 (± 0.7)	26.8 (± 0.4)
OrganA	CT	11	49.2 (± 2.5)	63.9 (± 1.6)	71.5 (± 1.1)	76.2 (± 0.8)	21.4 (± 0.3)

dermatology; the zero-label control is evaluated on the dermatology datasets in Table 1. *DtD increases visual attention across modalities.* The same layer-wise VAR pattern observed in dermatology holds across all modalities (Appendix Fig. 8). Radiology and dermatology see the largest gains (+40%), while pathology and hematology see +30–35%. This likely reflects the fact that modalities requiring whole-image spatial reasoning benefit most from forced visual description.

6 Conclusion and Discussion

In this work, we investigated a critical paradox in medical foundation models: the significant performance gap between a VLM's powerful vision encoder and its end-to-end zero-shot diagnostic accuracy. Focusing on MedGemma-4B in the context of dermatology, we systematically investigated three hypotheses for this discrepancy: a train-test distribution mismatch, an over-reliance on language priors, and a fundamental objective misalignment between the vision and language components. To address these challenges, we combined three VLM-fine-tuning-free strategies: providing in-context clinical descriptions, enforcing a DtD reasoning process, and leveraging the vision encoder for candidate selection in a "Top-5 to Top-1" framework. Our experiments demonstrate that these interventions, both individually and combined, improve classification accuracy, reduce diagnostic bias, and enhance the model's visual grounding, while retaining the frozen VLM and using a lightweight probe for candidate selection. Experiments on SkinVL, InternVL3, InternVL3.5, and biomed-Qwen support intervention transfer across backbones, while the cross-modality experiments extend the representation and attention analyses. These findings motivate evaluating medical VLMs at both the representation and decision levels, and developing inference strategies that better connect available visual evidence with the final diagnosis.

Limitations. Our detailed analysis centers on MedGemma-4B and dermatology. Additional backbone and modality comparisons broaden the evidence, but the complete pipeline requires evaluation across medical specialties and prospective validation of its clinical utility. Establishing causal visual reliance requires additional controlled interventions.

AI USE STATEMENT

We used a Large Language Model to assist with language polishing and improving the readability of this manuscript. The authors are fully responsible for all research ideas, experimental results, and claims presented in this paper.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. All data are public and de-identified, and no protected health information is included. We do not expect our approach to introduce direct ethical risks or harmful societal outcomes.

REPRODUCIBILITY STATEMENT

The appendix provides implementation details and hyperparameter configurations for fine-tuning and linear probing, together with evaluation prompts, dataset statistics, and clinical descriptions. It also reports detailed results for the baseline comparisons, additional backbones, cross-domain evaluation, and shot ablations.

REFERENCES

Titus Brinker, Achim Hekler, Alexander Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Jochen Utikal, Christof Kalle, Stefan Fröhling, and Michael Weichenthal. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 03 2019. doi: 10.1016/j.ejca.2019.02.005.

Daixuan Cheng, Shaohan Huang, Ziyu Zhu, Xintong Zhang, Wayne Xin Zhao, Zhongzhi Luan, Bo Dai, and Zhenliang Zhang. On domain-adaptive post-training for multimodal large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 274–296, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.17. URL https://aclanthology.org/2025.findings-emnlp.17/.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, et al. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26813–26829, 2025.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin M. 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.

Dyke Ferber, Georg Wölflein, Isabella C Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar SM El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, et al. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1):10104, 2024.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1820–1828, 2021.

Robbie Holland, Thomas RP Taylor, Christopher Holmes, Sophie Riedl, Julia Mai, Maria Patsiamanidi, Dimitra Mitsopoulou, Paul Hager, Philip Müller, Johannes C Paetzold, et al. Specialized

curricula for training vision language models in retinal image analysis. *NPJ Digital Medicine*, 8 (1):532, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, mar 2019. ISSN 2168-2194. doi: 10.1109/JBHI.2018.2824327.

Chanwoo Kim, Soham U 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. *Nature medicine*, 30(4):1154–1165, 2024.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

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. *Nature medicine*, 26(6):900–908, 2020.

Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G De Angelo, Fábio CR Alves Jr, José 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 in brief*, 32:106221, 2020.

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*, pp. 8748–8763. PmLR, 2021.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

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. *Science Translational Medicine*, 13(581):eabb3652, 2021.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018.

Janet Wang, Yunbei Zhang, Zhengming Ding, and Jihun Hamm. Achieving reliable and fair skin lesion diagnosis via unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5157–5166, 2024.

Janet Wang, Xin Hu, Yunbei Zhang, Diabate Almamy, Vagamon Bamba, Konan Amos Sébastien Koffi, Yao Koffi Aubin, Zhengming Ding, Jihun Hamm, and Rie R Yotsu. eskinhealth: A multimodal dataset for neglected tropical skin diseases. *arXiv preprint arXiv:2508.18608*, 2025a.

Janet Wang, Yunbei Zhang, Zhengming Ding, and Jihun Hamm. Doctor approved: Generating medically accurate skin disease images through ai-expert feedback. *arXiv preprint arXiv:2506.12323*, 2025b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025c. URL https://arxiv.org/abs/2508.18265.

Shuo Xing, Yuping Wang, Peiran Li, Ruizheng Bai, Yueqi Wang, Chan-wei Hu, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. Re-align: Aligning vision language models via retrieval-augmented direct preference optimization. *arXiv preprint arXiv:2502.13146*, 2025.

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. *arXiv preprint arXiv:2503.14911*, 2025a.

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. *Nature Medicine*, pp. 1–12, 2025b.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Jufeng Yang, Xiaoping Wu, Jie Liang, Xiaoxiao Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Self-paced balance learning for clinical skin disease recognition. *IEEE transactions on neural networks and learning systems*, 31(8):2832–2846, 2019.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in mllm reasoning. *arXiv preprint arXiv:2507.03019*, 2025.

Wenqi Zeng, Yuqi Sun, Chenxi Ma, Weimin Tan, and Bo Yan. Mm-skin: Enhancing dermatology vision-language model with an image-text dataset derived from textbooks. *arXiv preprint arXiv:2505.06152*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023a.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023b.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. URL https://arxiv.org/abs/2303.00915.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

Appendix

We provide additional baseline and prompt-control results, evaluations on other backbones and VQA-RAD, and extended confusion-matrix and attention analyses. The appendix also contains qualitative examples, evaluation prompts, implementation details, shot ablations, SkinVL comparisons, dataset statistics and clinical descriptions, and a discussion of limitations.

A Additional Baselines and Controls

A.1 Zero-Label Encoder Comparison

We evaluate MedSigLIP through image–text similarity using the same candidate class names provided to MedGemma. This evaluation uses no labeled target images, parameter updates, or fitted classifier. Table 7 also reports BiomedCLIP (Zhang et al., 2025) and DermLIP (Yan et al., 2025a) as additional contrastive baselines. MedSigLIP exceeds MedGemma on each of the four datasets, with an average accuracy difference of 10.26 percentage points.

Table 7: **Zero-label encoder baselines outperform MedGemma in most comparisons.** Top-1 accuracy (%) without fitting a target-task classifier. MedSigLIP outperforms MedGemma on all four datasets.

Dataset	MedGemma	MedSigLIP	BiomedCLIP	DermLIP
Derm7pt	16.46	24.18	24.55	24.03
eSkinHealth	13.35	26.53	28.22	29.28
Fitzpatrick17k	16.73	28.37	25.71	31.55
PAD-UFES-20	46.91	55.42	46.39	53.06

A.2 Multimodal Image In-Context Learning

The image-ICL baseline supplies MedGemma with labeled image–class pairs as demonstrations. Table 8 compares this 8-shot setting with the zero-shot VLM, the 8-shot linear probe, and encoder-assisted reranking. Image ICL improves accuracy on all four datasets, by 3.61 percentage points on average. The 8-shot LP remains stronger in each comparison. Reranking gives the highest accuracy on Derm7pt, eSkinHealth, and PAD-UFES-20; on Fitzpatrick17k its accuracy is 0.61 points below LP. The full 1-shot and 8-shot pipeline comparison, including the original reported variability, is retained in Table 9.

Table 8: **Image ICL improves the VLM baseline while leaving a gap to LP.** Top-1 accuracy (%). The image-ICL, LP, and reranking columns use the 8-shot settings.

Dataset	VLM 0-shot	Image ICL	LP	Reranking
Derm7pt	16.46	19.44	28.66	35.56
eSkinHealth	13.35	18.15	38.80	42.05
Fitzpatrick17k	16.73	20.72	38.43	37.82
PAD-UFES-20	46.91	49.58	53.33	80.31

A.3 Clinical-Description Controls

We compare class names alone with three types of additional text: class-aligned clinical descriptions, length-matched generic dermatology text without class-discriminative information, and descriptions assigned to incorrect classes. Table 10 shows that correctly aligned descriptions perform best among these four prompt-only settings on both datasets. Generic text does not improve over names alone, and shuffled descriptions reduce accuracy. The comparison supports a role for the alignment of clinical information with the candidate labels.

Table 9: Effectiveness of the proposed two-stage inference pipeline. Comparison of the **Baseline VLM** (MedGemma zero-shot), the standalone **Linear Probe** (MedSigLIP), and **Our Pipeline** which uses the Linear Probe to filter candidates (Top-5) for the VLM (Top-1). All accuracy values are Top-1 (%). Ablation study on different number of shots can be found in Appendix I.4.

Dataset	**Baseline VLM** *Zero-Shot*	**Linear Probe** *(Vision Encoder Only)*		**Our Pipeline** *(LP Top-5 → VLM Top-1)*	
	0-shot	**1-shot**	**8-shot**	**1-shot**	**8-shot**
Derm7pt	16.46 (± 0.01)	12.96 (± 1.75)	28.66 (± 1.38)	20.02 (± 2.39)	35.56 (± 0.99)
eSkinHealth	13.35 (± 0.23)	23.51 (± 1.91)	38.80 (± 1.06)	30.98 (± 2.85)	42.05 (± 1.48)
Fitzpatrick17k	16.73 (± 0.19)	17.63 (± 3.13)	38.43 (± 1.03)	25.35 (± 2.41)	37.82 (± 1.37)
SD-260	10.81 (± 0.13)	11.48 (± 2.27)	37.12 (± 1.46)	22.31 (± 1.95)	42.33 (± 0.92)
PAD-UFES-20	46.91 (± 0.20)	34.85 (± 1.83)	53.33 (± 1.34)	58.33 (± 2.24)	80.31 (± 1.23)

Table 10: **Class-aligned clinical information improves over prompt-length controls.** Top-1 accuracy (%). The first four rows use no labeled target images; the last two include an 8-shot probe.

Setting	Fitzpatrick17k	PAD-UFES-20
Names only	16.73	46.22
Class-aligned descriptions	24.01	52.23
Length-matched generic text	16.68	36.36
Shuffled descriptions	12.71	15.64
Top-5-to-Top-1	28.32	75.23
All combined	40.97	84.68

Table 11: Classification performance across five dermatology datasets. The table compares the performance for MedGemma-4B under three conditions: zero-shot inference, full-data fine-tuning, and a linear probe on its vision encoder (MedSigLIP). Performance is compared against a linear probe on a competitive vision encoder, PanDerm. Linear probing on both vision encoders significantly outperforms the zero-shot with the VLM, even after fine-tuning. Results are averaged over five runs.

Data Details				**MedGemma-4B**						**PanDerm**	
Datasets	**#Class**	**#Train**	**#Test**	**Zero-Shot**		**Fine-tune**		**Linear Probe**		**Linear Probe**	
				ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
Derm7pt	14	413	395	16.46 (± 0.02)	17.77	39.50 (± 2.15)	22.49	57.22 (± 0.88)	30.07	58.99 (± 0.76)	57.93
eSkinHealth	24	2,714	2,676	13.35 (± 0.15)	16.33	60.26 (± 1.42)	56.17	65.35 (± 0.54)	42.51	60.64 (± 0.61)	59.83
Fitzpatrick17k	20	3,100	3,100	16.73 (± 0.08)	11.13	47.32 (± 1.10)	45.16	66.32 (± 0.45)	66.90	64.87 (± 0.39)	64.88
SD-260	260	10,362	10,238	10.81 (± 0.05)	7.94	39.45 (± 0.95)	34.63	75.35 (± 0.32)	60.10	72.08 (± 0.28)	60.24
PAD-UFES-20	6	1,134	1,164	46.91 (± 0.12)	32.98	66.75 (± 1.85)	61.53	76.03 (± 0.91)	66.22	74.21 (± 0.84)	64.98

A.4 Full-Data Classification Comparison

B Additional Backbone Evaluation

We evaluate the general-purpose InternVL3-4B (Zhu et al., 2025) and InternVL3.5-4B (Wang et al., 2025c), together with the medical biomed-Qwen2.5-VL-3B (Cheng et al., 2025), on Derm7pt and Fitzpatrick17k. The InternVL3.5 evaluation uses the same settings as InternVL3. Table 12 retains all six model–dataset comparisons. DtD improves direct-answer accuracy in every comparison, and Top-5-to-Top-1 gives the strongest result among the evaluated inference settings. These results test transfer of the interventions across backbones.

C Cross-Domain VQA after Task-Specific Fine-Tuning

We evaluate the base MedGemma-4B and its Fitzpatrick17k QLoRA checkpoint on VQA-RAD (Lau et al., 2018). The adapted checkpoint is trained on the full Fitzpatrick17k training set and improves in-domain classification from 16.73% to 47.32%. Both checkpoints are evaluated on the same 451 VQA-RAD questions with identical prompts and greedy decoding. Table 13 shows decreases of 5.04

Table 12: **DtD and candidate-assisted inference improve all six added backbone–dataset pairs.** Top-1 accuracy (%). Top-5-to-Top-1 uses the 8-shot configuration.

Model	Dataset	DA	DtD	Top-5-to-Top-1
InternVL3-4B	Derm7pt	5.68	6.20	14.83
	Fitzpatrick17k	10.10	12.23	19.70
InternVL3.5-4B	Derm7pt	7.21	7.85	18.32
	Fitzpatrick17k	9.25	13.15	22.16
biomed-Qwen2.5-VL-3B	Derm7pt	20.93	38.50	41.24
	Fitzpatrick17k	12.14	19.96	22.35

points in closed-ended accuracy, 15.00 points in open-ended accuracy, and 22.53 points in token recall after adaptation. This comparison measures the cross-domain medical question-answering trade-off associated with task-specific fine-tuning.

Table 13: **Dermatology adaptation improves in-domain classification but reduces cross-domain VQA performance.** Accuracy and token recall (%) on VQA-RAD; higher is better.

Model	Closed-ended	Open-ended	Token recall
Base MedGemma-4B	70.92	37.50	49.96
Fitzpatrick17k QLoRA	65.88	22.50	27.43

D Full Confusion-Matrix Comparison

Figure 5 presents the full confusion matrices for the four inference settings compared in our dermatology analysis.

E Attention Heatmap Construction

To examine how DtD changes attention to the lesion area, we generate and analyze attention heatmaps. Here, we define two metrics: **Label-Token Heatmap** ($\mathbf{m}_{\text{label}}$): This measures the average attention the model pays to the input image tokens specifically when generating the words of the final diagnosis (e.g., "basal cell carcinoma"). It helps us see what the model attends to at the moment of decision. **All-Tokens Heatmap** ($\mathbf{m}_{\text{all}}$): This measures the average attention paid to the image tokens across the entire generated response, including both the description and the diagnosis.

Specifically, let the generated sequence have S tokens. At step t, after averaging head layers, the decoder gives an attention row $\bar{\mathbf{a}}^{(t)} \in \mathbb{R}^{T_t}$. Let the image occupy a contiguous token block $\mathcal{I} = \{s_{\text{img}}, \ldots, s_{\text{img}} + L_{\text{img}} - 1\}$. Define the image-restricted vector, the slice of a token's attention only over the image tokens, with everything else (text tokens) dropped, as

$$\mathbf{r}^{(t)} \in \mathbb{R}^{L_{\text{img}}}, \qquad \mathbf{r}^{(t)}[j] = \begin{cases} \bar{a}^{(t)}_{s_{\text{img}}+j}, & s_{\text{img}} + j \le T_t, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

So $\mathbb{R}^{L_{\text{img}}}$ is just a length-L_{img} real vector, with one weight per visual token. To check how much the model attends to the final lesion prediction when generating responses (see Fig. 6 (b-c)), the label-token heat vector is $\mathbf{m}_{\text{label}} = \frac{1}{|\mathcal{S}_{\text{label}}|} \sum_{t \in \mathcal{S}_{\text{label}}} \mathbf{r}^{(t)} \in \mathbb{R}^{L_{\text{img}}}$, where $\mathcal{S}_{\text{label}}$ indexes the tokens that spell the predicted diagnosis. Reshape $\mathbf{m}_{\text{label}}$ to the token grid and upsample to the image to render the heatmap. For the entire-response attention maps (see Fig. 6 (d-e)), all-tokens heat vector is $\mathbf{m}_{\text{all}} = \frac{1}{S} \sum_{t=1}^{S} \mathbf{r}^{(t)} \in \mathbb{R}^{L_{\text{img}}}$. By comparing the heatmaps from a baseline prompt (that queries direct answer, Fig. 6 (b,d)) with those from our DtD prompt (Fig. 6 (c,e)), we can isolate the effect of our strategy. When the label-token heatmap $\mathbf{m}_{\text{label}}$ is more concentrated on the lesion, it indicates that the description prompt changes the distribution of attention during the final choice. Similarly, the DtD prompt yields broader yet meaningful attention in the $\mathbf{m}_{\text{all}}$ heatmap, as the model must scan the image to enumerate pixel-level attributes for the description.

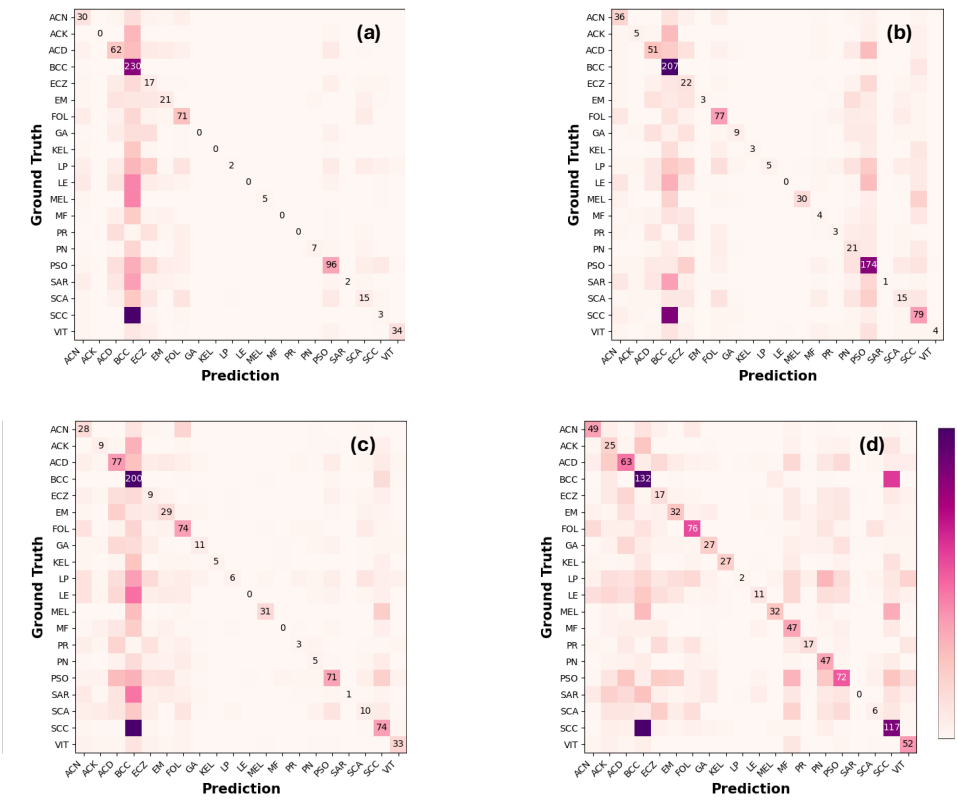


Figure 5: **Prompting and candidate selection change diagnostic confusion.** Original Fitzpatrick17k matrices for (a) direct answer, (b) clinical descriptions, (c) DtD, and (d) Top-5-to-Top-1 with a 2-shot probe. Panels retain the original counts and are enlarged for readability. Class abbreviations appear in Table 22.

F Per-Layer VAR Analysis Across Modalities

Fig. 8 presents the per-layer Vision Attention Ratio (VAR) for all six medical imaging datasets under Direct Answer (DA) and Describe-then-Decide (DtD) prompting. The consistent pattern across modalities shows that DtD elevates VAR at every layer, with particularly large gains at the global attention layers (layers 5, 10, 17, 22, 27, 33). Table 15 provides the numerical summary of mean VAR values and relative improvements.

G Attention Entropy Analysis

To complement the Vision Attention Ratio (VAR) analysis presented in the main text, we examine the entropy of the attention distributions across layers. While VAR measures *how much* attention is directed toward vision tokens, entropy measures *how focused* or *diffuse* that attention is. Lower entropy indicates that the model concentrates its attention on fewer tokens, suggesting more selective visual processing.

Table 14 reports the mean attention entropy under Direct Answer (DA) and Describe-then-Decide (DtD) prompting for six medical imaging datasets. DtD produces lower mean entropy than DA for all six modalities, with reductions ranging from -0.08 (PathMNIST, BloodMNIST) to -0.21 (OrganAMNIST). The largest decreases are observed for OrganAMNIST ($4.88 \rightarrow 4.67$), CheXpert ($4.81 \rightarrow 4.61$), and PatchCamelyon ($4.74 \rightarrow 4.58$).

Fig. 10 visualises the per-layer entropy for all six modalities under both prompting strategies. The heatmap confirms that the entropy reduction from DtD is not confined to a few layers but is distributed

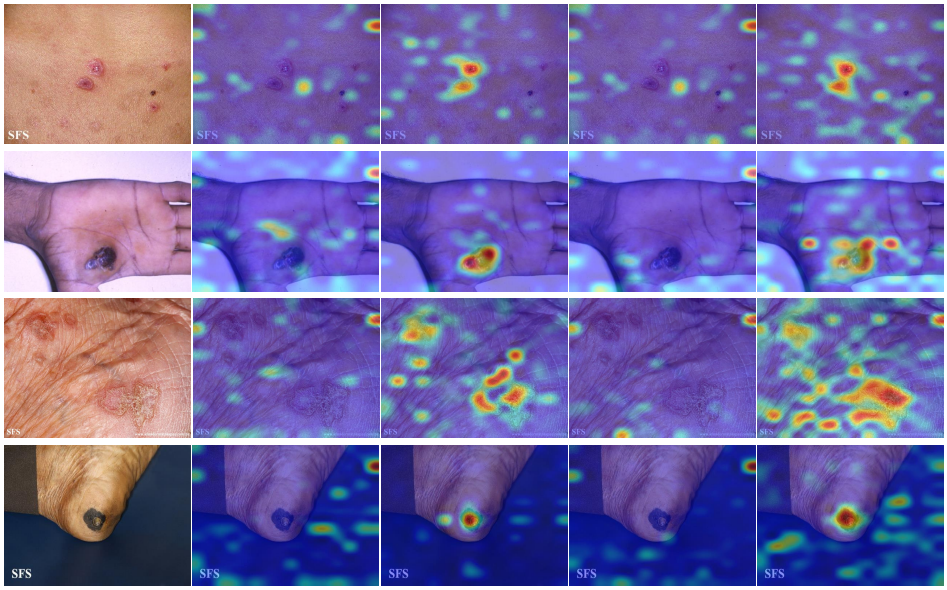


Figure 6: Attention heatmaps. Columns show (a) the original image, (b, d) attention from the DA prompt (baseline), and (c, e) attention from our DtD strategy. We visualize attention for the final diagnosis tokens (b, c) and for the entire response (d, e). The DtD prompt yields attention maps that are more accurately concentrated on the lesion area, both during the final decision (c) and across the full response (e), illustrating the change in attention allocation.

Table 14: Mean attention entropy under Direct Answer (DA) vs. Describe-then-Decide (DtD). Lower entropy indicates more focused attention. DtD reduces entropy for all six modalities.

Dataset	**Modality**	**DA**	**DtD**	**Δ**
Fitzpatrick17k	Dermatology	4.81	4.67	−0.14
CheXpert	Radiology	4.81	4.61	−0.20
PatchCamelyon	Pathology	4.74	4.58	−0.16
PathMNIST	Pathology	4.77	4.69	−0.08
BloodMNIST	Hematology	4.65	4.57	−0.08
OrganAMNIST	CT	4.88	4.67	−0.21

broadly, with the most pronounced decreases occurring in the global-attention layers that also exhibit the largest VAR increases (see Fig. 8).

Taken together with the VAR results, these findings show that DtD not only increases the proportion of attention directed to vision tokens (higher VAR) but also makes that attention more concentrated and selective (lower entropy). This dual effect characterizes how the description prompt changes attention allocation during generation.

Table 15: Mean Vision Attention Ratio (%) under Direct Answer (DA) vs. Describe-then-Decide (DtD). DtD increases the fraction of attention allocated to visual tokens during generation across the evaluated modalities.

Dataset	**Modality**	**DA**	**DtD**	**Δ (Rel.)**
Fitz17k	Dermatology	2.85	3.98	+39.8%
CXR	Radiology	3.02	4.24	+40.2%
PCam	Pathology	2.98	3.95	+32.5%
PathM	Pathology	2.32	3.02	+30.1%
BloodM	Hematology	2.86	3.80	+33.1%
OrganA	CT	2.67	3.60	+34.7%

Fig. 7 further examines the peak-layer VAR values for each dataset, highlighting which layers exhibit the strongest vision attention under each prompting strategy.

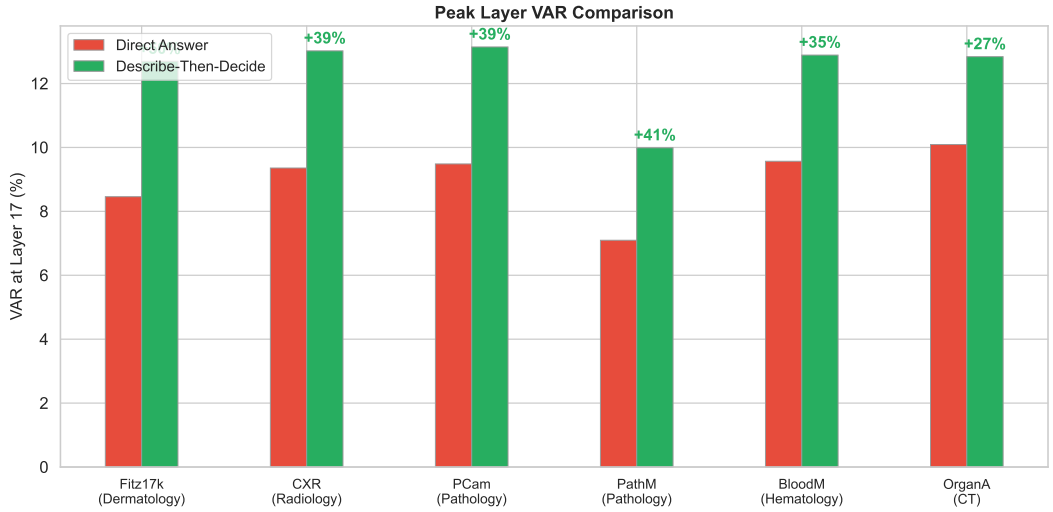


Figure 7: Peak-layer Vision Attention Ratio across datasets. DtD consistently achieves higher peak VAR values compared to DA, with the largest improvements observed in radiology and dermatology.

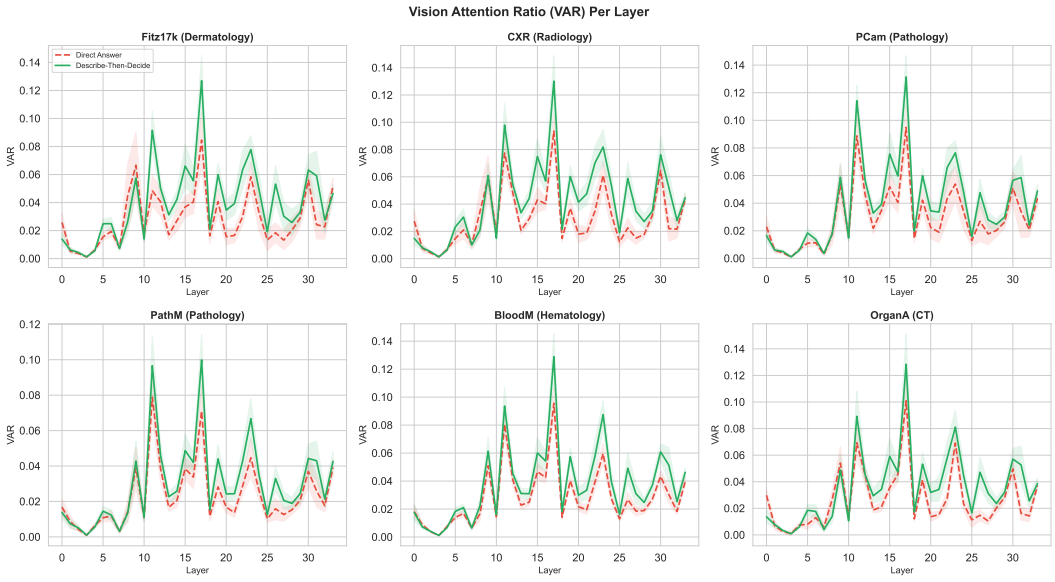


Figure 8: Vision Attention Ratio (VAR) per layer for six medical imaging datasets under Direct Answer (DA, red dashed) vs. Describe-then-Decide (DtD, green solid). DtD consistently increases vision attention across all layers and modalities, with the most pronounced gains at global attention layers (5, 10, 17, 22, 27, 33).

H Supplementary Experimental Details

H.1 Abstention Test on Masked Images

To quantify the model's reliance on language priors, we masked the primary lesion in 200 test images and evaluated whether MedGemma-4B would abstain from diagnosis. Fig. 9 shows the results under both Direct Answer and Describe-then-Decide prompting. Under DA, the model fails to abstain

in 88% of cases, defaulting to a BCC diagnosis. Under DtD, the abstention rate increases to 72%, showing that the description prompt increases abstention when the lesion is masked.

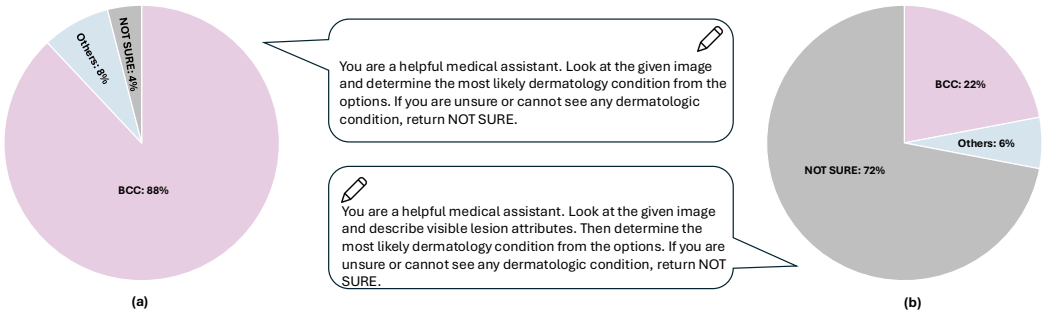


Figure 9: Abstention test on 200 masked images. (a) DA: the model diagnoses despite masked lesions. (b) DtD: the model abstains more frequently.

H.2 Description Quality Example

Fig. 11 shows an example of a clinical description generated by MedGemma-4B when prompted to describe lesion attributes before diagnosis. The description is detailed and clinically accurate, contrasting sharply with the hallucinated reasoning observed during standard zero-shot classification (cf. Fig. 3 in the main text).

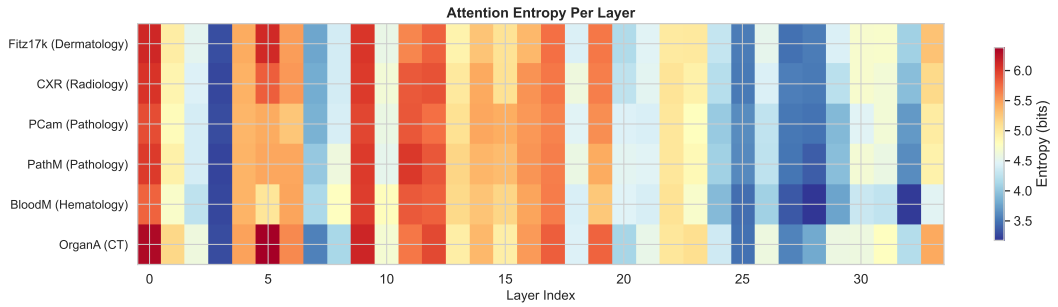


Figure 10: Attention entropy heatmap across layers for six medical imaging modalities under Direct Answer (DA) and Describe-then-Decide (DtD) prompting. DtD consistently produces lower entropy (more focused attention) across all modalities and layers.

Prompt

You are a helpful medical assistant. Look at the given image and describe the visible skin condition in terms of location, lesion type, shape, color and texture.



Answer

Location: The lesion is located on the dorsum (back) of the hand, specifically on the fingers.
Lesion Type: The lesion appears to be a combination of several different types, including:
* Telangiectasias: These are small, dilated blood vessels that appear as fine, red lines or streaks.
* Hyperkeratosis: This refers to thickening of the outer layer of skin (epidermis). It can appear as rough, scaly patches.
* Possible Erythema: There is some redness present, which could indicate inflammation or increased blood flow.
Shape: The lesion is irregular in shape, with a somewhat patchy distribution across the fingers. It doesn't have a distinct border.
Color: The lesion exhibits a range of colors, including:
* Red/Pink: This is due to the telangiectasias and possible erythema.
* Brownish/Yellowish: This could be due to hyperkeratosis or changes in the underlying skin.
Texture: The texture is rough and scaly, consistent with hyperkeratosis. The telangiectasias may also contribute to a slightly rough feel.
Overall Impression: The lesion is a complex one, likely representing a combination of vascular and epidermal changes. The presence of telangiectasias, hyperkeratosis, and possible erythema suggests a possible underlying condition.

Figure 11: Example description generated by MedGemma-4B. The model produces a detailed, clinically relevant analysis when prompted to describe visible features before diagnosing.

H.3 Zero-Shot Prompt Templates

Table 16 provides the full prompt templates used for each of the three zero-shot prompting strategies evaluated in the main text.

Table 16: Zero-shot prompts used for evaluation.

Type	Prompt
direct answer (baseline)	You are a helpful medical assistant. Look at the given image and determine the most likely dermatology condition from the options: (A) Granuloma annulare (B) Lupus erythematosus (C) Vitiligo. ... ...
in-context	You are a helpful medical assistant. Look at the given image and determine the most likely dermatology condition from the options: (A) Granuloma annulare, characterized by smooth skin-colored or pink papules arranged in an annular ring with firm ridge. (B) Lupus erythematosus, characterized by scaly plaques with follicular plugging and central scarring on sun-exposed skin; acute malar rash appears as flat red butterfly across cheeks sparing folds. (C) Vitiligo, characterized by sharply bordered milky white macules on face, hands, or genital skin with normal texture and pigment loss. ... ...
DtD	You are a helpful medical assistant. Look at the given image and describe visible lesion attributes (location, lesion type, shape/border, color, texture). Then determine the most likely diagnosis from the options: (A) Granuloma annulare (B) Lupus erythematosus (C) Vitiligo. ... ...

I Implementation Details

I.1 LoRA Fine-tuning

To ensure a rigorous evaluation of the model's adaptability, we fine-tuned MedGemma-4B-it using Low-Rank Adaptation (LoRA). Given the high memory requirements of 4B parameter models, we employed QLoRA (Dettmers et al., 2023), loading the base model in 4-bit precision with bfloat16 compute dtype. We applied LoRA adapters to all linear layers to maximize model expressivity while keeping the number of trainable parameters low. We also explicitly set the embed_tokens module to be trainable. The optimization was performed using the fused AdamW optimizer with a linear learning rate scheduler and a warmup ratio of 0.03. To determine the optimal configuration, we conducted a grid search over the LoRA rank (r), scaling factor (α), and learning rate. The search space and selected hyperparameters are detailed in Table 17. We trained for 10 epochs with a batch size of 4 and gradient accumulation steps of 4. The final selected configuration (Rank 8, Alpha 16, LR $1e^{-4}$) was chosen based on validation set performance and used for all reported fine-tuning results. Details can be found in Table. 17.

Table 17: Hyperparameter search space and selected configuration for LoRA fine-tuning.

Hyperparameter	Search Space	Selected Value
LoRA Rank (r)	$\{4, 8, 16\}$	8
LoRA Alpha (α)	$\{8, 16, 32\}$	16
Learning Rate	$\{5e^{-5}, 1e^{-4}, 2e^{-4}\}$	$1e^{-4}$
LoRA Dropout	-	0.05
Target Modules	-	All Linear Layers
Precision	-	4-bit (QLoRA)
Optimizer	-	AdamW (Fused)
Training Epochs	-	10

I.2 Linear Probe

For feature extraction, we utilized the pre-trained medgemma-4b-it vision encoder in bfloat16 precision. We extracted the global image representation by computing the mean of the last hidden state tokens rather than using a single [CLS] token. The head was trained using the AdamW optimizer and CrossEntropyLoss for 100 epochs with a batch size of 128, as shown in Table. 19.

Table 18: Full grid search results for LoRA fine-tuning on the Fitzpatrick17k dataset. The optimal configuration selected for the final model is highlighted in bold.

Rank (r)	**Alpha** (α)	**Learning Rate**	**Accuracy (%)**
4	8	5×10^{-5}	42.15
4	8	3×10^{-4}	46.12
4	8	4×10^{-4}	45.88
4	16	5×10^{-5}	41.80
4	16	3×10^{-4}	45.05
4	16	4×10^{-4}	44.20
4	32	5×10^{-5}	35.60
4	32	3×10^{-4}	39.40
4	32	4×10^{-4}	37.10
8	8	5×10^{-5}	43.50
8	8	3×10^{-4}	46.45
8	8	4×10^{-4}	46.10
8	16	5×10^{-5}	44.10
8	16	3×10^{-4}	**47.32**
8	16	4×10^{-4}	46.95
8	32	5×10^{-5}	38.90
8	32	3×10^{-4}	41.25
8	32	4×10^{-4}	40.50
16	8	5×10^{-5}	43.20
16	8	3×10^{-4}	45.90
16	8	4×10^{-4}	45.50
16	16	5×10^{-5}	44.80
16	16	3×10^{-4}	46.98
16	16	4×10^{-4}	46.40
16	32	5×10^{-5}	43.10
16	32	3×10^{-4}	46.15
16	32	4×10^{-4}	45.20

Table 19: Implementation details and hyperparameters for Linear Probing experiments.

Hyperparameter	**Value**
Vision Encoder	MedSigLIP (Frozen)
Feature Pooling	Global Average Pooling
Precision (Backbone)	`bfloat16`
Precision (Head)	`float32`
Optimizer	AdamW
Batch Size	128
Epochs	100
Learning Rate	$\{3 \times 10^{-5}, 3 \times 10^{-4}, 3 \times 10^{-3}\}$

I.3 Vision Attention Ratio Computation

We extract full attention weights using eager attention (disabling Flash Attention) across all 34 transformer layers of MedGemma-4B. For each dataset, we randomly sample 100 test images and compute VAR under both Direct Answer (DA) and Describe-then-Decide (DtD) prompting. MedGemma-4B uses a MedSigLIP vision encoder that produces 256 vision tokens (feature dimension 1152), which are projected into the language model's embedding space. The model uses 8 attention

heads per layer. Attention weights are averaged across all heads and all generation steps per layer to produce the final per-layer VAR. All computations are performed in bfloat16 precision on a single NVIDIA A100 GPU.

I.4 Ablation Study on the Number of Shots

We selected the 8-shot setting based on the analysis presented in Table 20, where 8-shot maximizes the utility of our pipeline. At 8-shot, the vision encoder is strong enough to capture the correct diagnosis in the Top-5 (high recall), but not yet perfect at Top-1 (low precision), creating the ideal scenario for the VLM to apply its reasoning for re-ranking. Beyond 8 shots, the vision encoder becomes self-sufficient, and the added value of VLM reasoning diminishes.

Table 20: Effectiveness of the "Top-5 to Top-1" reranking strategy. This table compares three methods: the baseline zero-shot VLM, a standard k-shot linear probe (LP Top-1), and our two-stage approach. In our method, a k-shot linear probe selects the Top-5 candidates, which are then used by the VLM to produce a final prediction. The results show that our two-stage strategy consistently outperforms the direct linear probe's Top-1 accuracy in the low-data regime.

Dataset	0-shot Top-1	1-shot			2-shot			4-shot			8-shot			16-shot		
		Top1	Top5	0-Shot	Top1	Top5	0-Shot	Top1	Top5	0-Shot	Top1	Top5	0-Shot	Top1	Top5	0-Shot
Derm7pt	16.46	12.96	52.25	20.02	18.53	58.03	25.73	21.06	61.42	32.80	28.66	72.35	35.56	38.53	82.68	37.95
eSkinHealth	13.35	23.51	55.75	30.98	27.99	63.84	33.67	35.04	69.70	41.22	38.80	73.96	42.05	43.04	78.41	43.17
Fitzpatrick17k	16.73	17.63	50.58	25.35	23.08	59.50	27.42	31.01	67.67	33.68	38.43	75.88	37.82	45.05	80.99	38.01
SD-260	10.81	11.48	26.89	22.31	17.90	38.61	25.02	26.97	51.92	33.50	37.12	63.61	42.33	49.01	76.72	44.72
PAD-UFES-20	46.91	34.85	93.99	58.33	38.56	96.70	67.60	45.36	98.33	77.85	53.33	98.63	80.31	59.90	98.99	91.11

J Performance Analysis on SkinVL

To test whether the vision–language gap generalises beyond MedGemma, we repeat the core analysis on SkinVL (Zeng et al., 2025), a dermatology VLM built on the LLaVA architecture. Table 21 compares zero-shot VLM performance against full fine-tuning and a linear probe on SkinVL's frozen vision encoder across three datasets. The pattern mirrors our MedGemma findings: the linear probe consistently outperforms both the zero-shot VLM and full fine-tuning on every dataset, with the largest gap on Fitzpatrick17k (58.88% LP vs. 10.05% zero-shot). This confirms that the under-utilisation of visual features is not specific to MedGemma but reflects a broader architectural issue in current medical VLMs.

The cumulative candidate-filtering, clinical-description, and DtD results for SkinVL are provided in Table 4.

Table 21: Performance Gap Analysis on SkinVL. Comparison of the zero-shot VLM performance against full Fine-Tuning (FT) and a Linear Probe (LP) on the vision encoder. Consistent with MedGemma, the Linear Probe significantly outperforms the zero-shot VLM, confirming the under-utilization of visual features.

	VLM Zero-Shot		**Fine-Tuning (FT)**		**Linear Probe (LP)**	
Dataset	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
Derm7pt	12.77	15.05	44.81	22.98	**50.16**	**26.59**
eSkinHealth	13.76	10.33	57.74	39.24	**62.77**	**42.90**
Fitzpatrick17k	10.05	8.27	42.84	30.49	**58.88**	**46.22**

K Limitations

Our main analysis covers five dermatology datasets, with additional backbone comparisons on Derm7pt and Fitzpatrick17k. The experiments in radiology, pathology, hematology, and CT extend the representation and attention analyses; evaluation of the complete pipeline across these specialties

remains a next step. The prompt-only interventions use no target labels, whereas Top-5-to-Top-1 depends on a labeled support set and a fitted linear probe. Description quality is measured through model self-evaluation, and independently verified description errors are an important target for further analysis. Attention allocation characterizes model behavior, while causal visual reliance requires additional controlled interventions. Furthermore, exploring how these principles could inform the development of new VLM architectures remains a promising direction.

A comprehensive clinical utility study with multiple experts and inter-rater reliability analysis is a necessary next step for evaluating real-world deployment.

Table 22: Skin Condition Distribution for Fitzpatrick17k

Skin Condition	**Real Training**	**Real Test**	**Synthetic**
Acne (ACN)	92	91	93
Actinic Keratosis (AK)	88	87	164
Allergic Contact Dermatitis (ACD)	215	215	181
Basal Cell Carcinoma (BCC)	234	234	154
Eczema (ECZ)	102	102	166
Erythema Multiforme (EM)	118	118	155
Folliculitis (FOL)	171	171	114
Granuloma Annulare (GA)	106	105	148
Keloid (KEL)	78	78	135
Lichen Planus (LP)	246	245	151
Lupus Erythematosus (LE)	205	205	172
Melanoma (MEL)	130	131	155
Mycosis Fungoides (MF)	91	91	165
Pityriasis Rosea (PR)	96	97	156
Prurigo Nodularis (PN)	85	85	152
Psoriasis (PSO)	326	327	165
Sarcoidosis (SAR)	174	175	162
Scabies (SCA)	170	169	176
Squamous Cell Carcinoma (SCC)	290	291	175
Vitiligo (VIT)	83	83	161
Total	3100	3100	3100

Table 23: **Dataset details.**

Dataset	**# Cls**	**Train**	**Test**	**Class Names**
Derm7pt	14	413	395	clark nevus, melanoma, reed or spitz nevus, seborrheic keratosis, basal cell carcinoma, vascular lesion, lentigo, blue nevus, dermal nevus, dermatofibroma, combined nevus, congenital nevus, miscellaneous, recurrent nevus
eSkinHealth	24	2,714	2,676	buruli ulcer, scabies, yaws, prurigo nodularis, tinea corporis, erysipelas, tinea capitis, leprosy, impetigo, necrotizing fasciitis, contact dermatitis, lichen planus, tinea versicolor, folliculitis, chickenpox, acne, vitiligo, abscess, keratosis, herpes zoster, atopic dermatitis, eczema, lipome, mycetoma
Fitzpatrick17k	20	3,100	3,100	psoriasis, squamous cell carcinoma, lichen planus, basal cell carcinoma, allergic contact dermatitis, lupus erythematosus, sarcoidosis, folliculitis, scabies, melanoma, erythema multiforme, granuloma annulare, eczema, pityriasis rosea, mycosis fungoides, acne, actinic keratosis, prurigo nodularis, vitiligo, keloid

Dataset	# Cls	Train	Test	Class Names
SD-260	260	10,362	10,238	abrasion, acne excoriee, acne keloidalis nuchae, acne vulgaris, acrokeratosis verruciformis, actinic solar damage, actinex treatment, actinic cheilitis, actinic keratosis, cutis rhomboidalis nuchae, actinic favre-racouchot, pigmentation, actinic solar elastosis, solar purpura, telangiectasia, actinic wrinkles, acute eczema, allergic contact dermatitis, alopecia areata, anagen effluvium, androgenetic alopecia, angiofibroma, angiokeratoma, angioma, angular cheilitis, aphthous ulcer, apocrine hydrocystoma, arsenical keratosis, atopic dermatitis, balanitis xerotica obliterans, basal cell carcinoma, beaus lines, beckers nevus, behcets syndrome, benign keratosis, blue nevus, bowenoid papulosis, bowens disease, cafe au lait macule, callus, candidiasis, cellulitis, chalazion, cherry angioma, clubbing of fingers, combined nevus, compound nevus, congenital nevus, contact dermatitis, crowes sign, cutanea larva migrans, cutaneous horn, cutaneous leishmaniasis, cutaneous t-cell lymphoma, cutis marmorata, darier-white disease, dermatofibroma, dermatomyositis, dermatosis papulosa nigra, desquamation, digital fibroma, dilated pore of winer, discoid lupus erythematosus, disseminated actinic porokeratosis, drug eruption, dry skin eczema, dyshidrosiform eczema, dysplastic nevus, eccrine poroma, eczema, epidermal nevus, epidermoid cyst, epithelioma adenoides cysticum, erythema ab igne, erythema annulare centrifigum, erythema craquele, erythema multiforme, exfoliative erythroderma, factitial dermatitis, favre-racouchot, fibroma, fibroma molle, fixed drug eruption, follicular mucinosis, follicular retention cyst, fordyce spots, frictional lichenoid dermatitis, ganglion, geographic tongue, granulation tissue, granuloma annulare, green nail, guttate psoriasis, hailey-hailey disease, half and half nail, halo nevus, hand foot mouth disease, herpes gestationis, herpes simplex virus, herpes zoster, hidradenitis suppurativa, hirsutism, histiocytosis x, hyperkeratosis palmaris et plantaris, hypertrichosis, ichthyosis, ichthyosis vulgaris, id reaction, impetigo, infantile atopic dermatitis, insect bite, intradermal nevus, inverse psoriasis, ischemia, junction nevus, keloid, keratoacanthoma, keratolysis exfoliativa of wende, keratosis pilaris, kerion, koilonychia, kyrles disease, leiomyoma, lentigo maligna melanoma, lentigo simplex, leprosy, leukemia cutis, leukocytoclastic vasculitis, leukonychia, lichen planus, lichen sclerosis et atrophicus, lichen simplex chronicus, lichen spinulosis, linear epidermal nevus, lipoma, livedo reticularis, lymphangioma circumscriptum, lymphocytic infiltrate of jessner, lymphocytoma cutis, lymphomatoid papulosis, mal perforans, malignant melanoma, median nail dystrophy, melasma, metastatic carcinoma, milia, molluscum contagiosum, morphea, mucha-habermann disease, mucous membrane psoriasis, myxoid cyst, nail cosmesis, nail dystrophy, nail nevus, nail psoriasis, nail ridging, nail trauma, neurodermatitis, neurofibroma, neurotic excoriations, nevus cell nevus, nevus comedonicus, nevus incipiens, nevus sebaceous of jadassohn, nevus spilus, nummular eczema, onychogryphosis, onycholysis, onychomycosis, onychoschizia, paronychia, pearly penile papules, pediculosis pubis, pemphigus foliaceus, perioral dermatitis, photodermatitis, pilomatrixoma, pincer nail syndrome, pitted keratolysis, pityriasis alba, pityriasis rosea, pityriasis rubra pilaris, pityriasis versicolor, pityrosporum folliculitis, poikiloderma atrophicans vasculare, pomade acne, porokeratosis of mibelli, port wine stain, pseudofolliculitis barbae, pseudorhinophyma, psoriasis, pterygium inversum unguis, pustular psoriasis, pyoderma gangrenosum, pyogenic granuloma, racquet nail, radiodermatitis, rhinophyma, rosacea, scabies, scalp psoriasis, scar, scarring alopecia, schambergs disease, sebaceous gland hyperplasia, seborrheic dermatitis, seborrheic keratosis, skin tag, solar elastosis, solar lentigo, spindle cell nevus, squamous cell carcinoma (scc), stasis dermatitis, stasis edema, stasis ulcer, steroid acne, steroid atrophy, steroid striae, steroid use, stomatitis, strawberry hemangioma, striae, subacute cutaneous lupus erythematosus, subungual hematoma, superficial actinic porokeratosis, syringoma, systemic lupus erythematosus, terrys nails, thermal burn, tick bite, tinea capitis, tinea corporis, tinea cruris, tinea faciale, tinea incognito, tinea manus, tinea pedis, tinea versicolor, toe deformity, traction alopecia, trichilemmal cyst, trichoepithelioma, trichofolliculoma, trichostasis spinulosa, trichotillomania, tuberous sclerosis, twenty nail dystrophy, ulcer, urticaria, uvl burn, varicella, verruca vulgaris, viral exanthem, virilization, vitiligo, von recklinghausens disease, wart, wound infection, xerosis, x-linked ichthyosis
PAD-UFES-20	6	1,134	1,164	actinic keratosis, basal cell carcinoma, melanoma, nevus, seborrheic keratosis, squamous cell carcinoma

Table 24: **Skin condition descriptions used in this study.**

Condition	Description
acne	Face/chest/back comedones and inflamed papules/pustules ±nodules; red or skin-colored with black/white heads; oily, ±crust.
actinic keratosis	Sun-exposed rough, scaly flat/slightly raised papule <1 cm; pink/red/brown; gritty sandpaper feel.
allergic contact dermatitis	At contact sites, ill-defined pink-red (darker on dark skin) patches/plaques ±vesicles/edema; weepy/crusty/scaly surface.
basal cell carcinoma	Sun-exposed pearly/waxy papule or thin scaly patch with rolled edge ±central ulcer; translucent or brown/black; smooth/shiny ±crust.
eczema	Flexural dry itchy ill-defined patches/plaques ±tiny vesicles; red/pink or purple/gray on dark skin; flaky ±lichenified.
erythema multiforme	Acral target lesions—round 1–3 cm with dark center, pale ring, red outer rim—mostly flat ±blister.

Condition	Description
folliculitis	Hair-bearing sites with clustered 2–5 mm follicle-centered pustules/red papules; red/darker base with white/yellow pus; dome-shaped ±crust.
granuloma annulare	Hands/feet/wrists/ankles smooth firm non-scaly papules forming annular rings; skin-colored/pink/red (purple on dark skin).
keloid	Over scars on chest/shoulders/earlobes etc., shiny firm hairless raised irregular growth extending beyond wound; pink/red or darker.
lichen planus	Wrists/ankles etc. flat-topped polygonal 2–10 mm violaceous papules/plaques with fine white Wickham striae; shiny ±scale.
lupus erythematosus	Malar 'butterfly' smooth pink rash ±discoid coin-shaped scaly scarred plaques on scalp/ears; red or hyperpigmented; rough if discoid.
melanoma	Anywhere (palms/soles/nails in dark skin) asymmetric lesion with irregular borders, color variegation, often >6 mm; becomes raised/crusted/ulcerated.
mycosis fungoides	Non-sun areas with dry scaly patches → thicker scaly plaques (±smooth tumor nodules); pink-red to brown/darker; irregular.
pityriasis rosea	Trunk herald patch then multiple smaller ovals along skin lines; pink/salmon (gray/brown/purple on dark skin) with fine collarette scale.
prurigo nodularis	Reachable areas with multiple very itchy 1–3 cm firm nodules, often crusted/scabbed on top; pink/red/brown/black; thick/rough with excoriations.
psoriasis	Elbows/knees/scalp/lower back well-demarcated plaques with thick silvery/gray scale; pink/red or purple/dark brown; dry/flaky (Auspitz sign).
sarcoidosis	Face/shins/scars with smooth firm plaques/nodules/patches; purplish/red-brown or lighter/darker areas; rubbery; shin nodules are tender.
scabies	Finger webs/wrists/waist/genitals etc. with 5–15 mm wavy burrows plus clustered 1–2 mm itchy papules/vesicles; excoriated/crusted.
squamous cell carcinoma	Sun-exposed or scarred sites with firm scaly/crusted nodule/plaque/ulcer ±raised border/central depression; pink/red or darker; rough, may bleed.
vitiligo	Sharply bordered depigmented patches (often symmetric) on face/hands/feet/genitals; chalk-white contrast; normal texture without scale.
nevus	Anywhere well-circumscribed round/oval macule/papule/nodule with smooth borders; uniform tan/brown/black or skin-colored; smooth flat or dome-raised.
seborrheic keratosis	Anywhere except palms/soles 'stuck-on' waxy papule/plaque with sharp borders; tan to dark brown/black or mixed; warty/greasy with keratin plugs.
buruli ulcer	Limb lesion evolves from painless nodule/plaque/edema to undermined necrotic ulcer; circular/irregular with yellow-white base and violaceous/brown edge; moist ±slough.
yaws	Leg/foot raspberry-like papilloma ±ulcer then multiple papules/plaques/shallow ulcers; dome-lobulated; bright red → red-brown/yellow; verrucous/granular ±crust.
tinea capitis	Scalp alopecic scaly ring-like patch with broken hairs/black dots ±boggy kerion or yellow scutula; gray-white/red-brown; fine scale or boggy crust.
tinea corporis	Exposed skin annular plaques with raised scaly active border and central clearing; polycyclic/serpiginous; rim scaly, center smoother.
erysipelas	Lower legs/face sharply demarcated lobulated cellulitic plaque ±tense bullae; fiery red/violaceous; smooth tense shiny peau d'orange.
leprosy	Cool sites hypopigmented/coppery patches with sensory loss or plaques/nodules/diffuse infiltration; round/oval with raised rim; dry hairless/anhidrotic or thick shiny.
necrotizing fasciitis	Rapidly enlarging ill-defined limb/trunk plaque with dusky patches, ecchymoses and flaccid bullae → black necrotic eschar; tense shiny then leathery.
impetigo	Perioral/exposed erosions with honey-colored crust after vesicles/pustules (bullous form flaccid blisters); round/oval, coalescent; moist → sticky crust.
tinea versicolor	Upper trunk/shoulders/neck flat hypo- or hyperpigmented macules/patches with fine powdery 'bran' scale accentuated by scraping; coalescing map-like.
varicella	Face/trunk then scalp/limbs crops of itchy 2–4 mm thin-walled vesicles on pink/dark base → crusted scabs; discrete ±coalescent; glistening then flaky.
abscess	Axillae/buttocks/groin etc. tender dome-shaped red-violaceous nodule that becomes fluctuant with pointing yellow/white center; shiny tense skin; purulent drainage after rupture.
atopic dermatitis	Flexures/hands/eyelids etc. itchy ill-defined papules→plaques with acute weeping/crust or chronic lichenification; pink-red or purple-brown/gray; dry rough.
keratosis pilaris	Outer arms/thighs/cheeks tiny follicular keratin plugs ('goose-bump' papules) in indistinct patches; skin-colored to red; rough sandpapery dry feel.
herpes zoster	Single-dermatome band of clustered clear vesicles on pink/purple base not crossing midline → pustules/crusts; 2–4 mm; tense/glossy then adherent crust.
lipoma	Trunk/neck/limbs soft smooth mobile subcutaneous dome/oval nodule with normal overlying color; rubbery/doughy consistency.
mycetoma	Foot/leg/hand firm lobulated subcutaneous mass with multiple draining sinus openings extruding colored grains; overlying skin normal→hyperpigmented; crusted seepage.
lymphatic filariasis	Lower legs/feet (±arms/genitals) chronic lymphedema to elephantiasis with column-like limb, bulbous foot and verrucous corrugated plaques; hyper/hypopigmented; thick 'mossy' hyperkeratosis.

Table 25: Visual checklists for Basal Cell Carcinoma (BCC) and Lupus Erythematosus (LE), used to probe MedGemma's conceptual knowledge of misclassified conditions.

Condition	Feature	Details
BCC	Location	Sun-exposed areas: face (nose, cheeks, forehead), ears, neck, scalp, upper chest/shoulders, back, hands, arms
	Morphology	"Pearly"/waxy bump; papillary (dome-shaped); nodular (firm); superficial infiltrating (flat, scaly patch); morpheaform (scar-like plaque); micropapillary; infiltrating (flat, scaly, mimics other dermatoses)
	Shape	Often irregular with poorly defined borders; typically asymmetrical; evolving in size, shape, or color
	Color	Pink or red; may be white/waxy; brown/tan or black; blue/purple hues if ulcerated
	Texture	Smooth or waxy; may be scaly or crusted; ulcerated surface can occur
LE	Location	Sun-exposed areas: face, ears, neck, chest, upper arms; mucous membranes (oral cavity, nose, eyelids); scalp, nails, genital area
	Morphology	Malar (butterfly) rash: flat/raised erythema across cheeks and nasal bridge; discoid: coin-shaped scarring plaques with hypo-/hyperpigmentation; photosensitivity rash; oral ulcers; livedo reticularis; vasculitis (purpura, petechiae); alopecia; Raynaud's phenomenon
	Shape	Malar: butterfly-shaped with central clearing; discoid: coin-shaped plaques; oral ulcers: round/oval; livedo: reticular (net-like); purpura: pinpoint to macular
	Color	Malar: red with central clearing; discoid: red, heals with pigment changes; oral ulcers: red to violaceous; livedo: reddish-blue; purpura: red or purple
	Texture	Malar: flat or slightly raised; discoid: raised, scaly, may crust or scar; oral ulcers: smooth base; livedo: smooth with reticular pattern