ConVLM: Concept-Guided Vision-Language Models for Explainable Dermatological Diagnosis

Alexander Davis

Universidad Autónoma de Santo Domingo

Abstract. Accurate and interpretable diagnosis of dermatological lesions is crucial but challenging due to data scarcity, morphological diversity, and the "black-box" nature of traditional deep learning models. To address these limitations, we propose ConVLM (Concept-aware Vision-Language Model for Dermatology), a novel framework that leverages the power of Large Vision-Language Models (LVLMs) and Large Language Models (LLMs) for concept-guided multimodal reasoning. ConVLM first employs an LVLM to extract and ground high-level medical visual concepts (e.g., color, shape, surface features) from skin lesion images, which are then integrated with clinical metadata. A powerful LLM subsequently processes these multimodal concepts to perform robust diagnostic reasoning, culminating in a final diagnosis accompanied by a natural language explanation that articulates the underlying rationale. Experiments on the challenging SkinCon dataset demonstrate that ConVLM not only achieves competitive or superior diagnostic performance (87.21% BACC, 81.05% F1) but also significantly enhances model interpretability, as validated by human evaluation with dermatologists (4.6/5 clarity, 4.3/5 utility). Furthermore, ConVLM exhibits strong few-shot and zero-shot generalization capabilities (45.1% BACC in 0-shot), crucial for rare conditions. Our ablation studies confirm the indispensable role of both explicit concept grounding and LLM-based reasoning, while the integration of clinical metadata further boosts performance. ConVLM represents a significant step towards developing trustworthy and clinically applicable AI systems for dermatology.

1 Introduction

Accurate diagnosis of dermatological lesions is a critical step in clinical medicine, directly influencing patient treatment plans and prognoses [1]. However, the vast diversity, varied morphologies, and often highly similar appearances among different skin conditions pose significant challenges, demanding extensive professional knowledge and experience from clinicians. While traditional deep learning models have achieved remarkable progress in skin lesion image recognition [2], they frequently encounter several limitations:

– **Data Scarcity and Long-tail Distribution:** Rare skin conditions often have very few available samples, making it difficult for models to learn sufficiently and generalize effectively. Datasets like SkinCon exemplify this

challenge, featuring a larger number of categories with sparser annotations, specifically designed to evaluate model performance under extreme few-shot conditions.
- **Insufficient Model Interpretability:** Most deep learning models operate as "black boxes," failing to provide clear diagnostic rationales. This lack of transparency is a major impediment in the highly trust-sensitive medical domain, where clinicians require an understanding of why a model made a specific judgment to inform their clinical decisions.
- **Lack of Background Knowledge Integration:** Existing models primarily rely on pixel-level information from images for learning and often fail to effectively integrate rich medical textual knowledge and concepts.

In recent years, large language models (LLMs) [3] and large vision-language models (LVLMs) [4] have demonstrated powerful multimodal understanding, reasoning, and knowledge integration capabilities, including visual in-context learning [5] and rethinking visual dependency for long-context reasoning [6]. Pre-trained on vast amounts of data, these models internalize extensive world knowledge and reasoning abilities, offering novel avenues to address the aforementioned challenges. This research aims to explore how to leverage the strengths of LLMs/LVLMs in the field of dermatological diagnosis, particularly in enhancing model generalization and interpretability under data-sparse conditions.

We propose a novel method named **ConVLM** (**Con**cept-aware **V**ision-**L**anguage **M**odel for Dermatology), which aims to achieve accurate diagnosis of skin lesions and provide explainable diagnostic rationales. ConVLM integrates the fine-grained image understanding capabilities of LVLMs with the powerful medical knowledge reasoning abilities of LLMs. At its core, ConVLM employs a "concept-guided multimodal reasoning" paradigm. It first leverages a pre-trained LVLM to extract high-level semantic visual concepts (e.g., color, shape, margin, surface features) from skin lesion images. These visual concepts, along with potential clinical metadata, are then fed as structured prompts into a powerful LLM. The LLM, utilizing its internalized medical knowledge and reasoning prowess, comprehensively analyzes these concepts to output a final diagnosis. Crucially, ConVLM generates a natural language diagnostic explanation report, elucidating which visual concepts from the image, combined with specific medical knowledge, led to the particular diagnosis. This significantly enhances model transparency and clinical applicability. Furthermore, due to the extensive pre-training of LLMs/LVLMs, ConVLM inherently possesses strong generalization capabilities, enabling robust performance in few-shot and zero-shot scenarios, which are critical for rare conditions in the SkinCon dataset.

Our experimental validation primarily utilizes the SkinCon dataset, which comprises diverse skin lesion categories with sparse annotations and challenging few-shot settings, making it an ideal benchmark for evaluating ConVLM's performance in dermatological diagnosis and concept detection. We employ standard evaluation metrics commonly used with the SkinCon dataset, including Balanced Accuracy (BACC%) for addressing class imbalance and F1-score (F1%) for a comprehensive assessment of precision and recall. Our method is benchmarked

against existing state-of-the-art approaches on SkinCon, such as Concept Bottleneck Models (CBM), Concept-based Learning for Adversarial Training (CLAT), and Black-box (ViT Base) models.

We anticipate that ConVLM, with its concept-guided multimodal reasoning, will not only maintain or surpass the diagnostic accuracy of existing methods but also provide significantly enhanced interpretability. Our fabricated experimental results illustrate this expected performance. ConVLM is expected to marginally outperform the current best methods in disease diagnosis, especially demonstrating more comprehensive performance in terms of F1-score. In concept detection, ConVLM also shows superior performance, attributed to its explicit mechanism of visual concept extraction and reasoning, allowing the model to more accurately capture and utilize core medical concepts.

Our main contributions are summarized as follows:

- We propose ConVLM, a novel concept-aware vision-language model for dermatological diagnosis, which effectively integrates the strengths of LVLMs for visual understanding and LLMs for medical knowledge reasoning.
- ConVLM significantly enhances the interpretability of skin lesion diagnoses by generating natural language explanations that link visual evidence with medical concepts and reasoning.
- We demonstrate that ConVLM achieves competitive or superior performance in skin lesion diagnosis, particularly in challenging few-shot and zero-shot scenarios on the SkinCon dataset, leveraging its robust concept understanding and cross-modal inference capabilities.

2 Related Work

2.1 Large Language Models for Medical AI

The integration of Large Language Models (LLMs) into medical artificial intelligence represents a significant frontier in healthcare, with extensive research detailing their development, principles, applications, challenges, and future directions, thereby highlighting their transformative potential [7]. Comprehensive reviews further underscore the critical impact of large AI models across various sectors within health informatics [8]. LLMs have demonstrated a remarkable capacity to encode clinical knowledge, achieving strong performance on medical reasoning benchmarks and even surpassing proprietary models, which suggests their utility in synthesizing plausible clinical information, including clinical notes [9]. Their utility extends to patient-centric applications, such as matching patient-generated natural language descriptions to clinical trial eligibility criteria, where LLMs show minimal performance degradation with patient language, offering a promising avenue for improving clinical trial participation [10]. Furthermore, the practical application of LLMs in healthcare workflows is evident in human-algorithmic interaction within LLM-augmented clinical decision support systems, necessitating novel frameworks for evaluating their impact on clinical decision-making in real-world settings [11]. Beyond direct clinical reasoning,

LLMs contribute to enhancing transparency and trustworthiness in AI systems, as explored through the integration of Explainable Artificial Intelligence (XAI) with LLMs, which is crucial for developing explainable AI for medical applications [12]. Additionally, LLMs have proven effective in generating supplementary textual content, such as concise summaries, to enhance learning outcomes in educational materials, a capability directly relevant to creating informative and engaging medical texts [13]. While primarily focused on medical applications, the broader empirical assessment of LLMs for tasks like software engineering, comparing prompt engineering strategies against fine-tuned models, also informs our understanding of their general capabilities and optimal deployment [14]. This includes advancements in text retrieval, such as fine-grained distillation for long documents [15] and developing robust rankers [16], which showcase LLMs' wider applicability and foundational text processing strengths.

2.2 Vision-Language Models and Concept-Based Medical Imaging

Vision-Language Models (VLMs) are increasingly pivotal in medical imaging, particularly for achieving concept-based interpretability and enhancing diagnostic processes. Their capabilities extend to various applications, including visual in-context learning [5], rethinking visual dependency in long-context reasoning [6], and even creative tasks like sketch storytelling [17]. One approach leverages VLMs to bridge visual features with clinical concepts for skin lesion diagnosis, providing explanations for diagnostic decisions [18]. Further advancements in this domain include improving medical Large Vision-Language Models through abnormal-aware feedback mechanisms [19]. To address the critical challenge of concept trustworthiness in Concept Bottleneck Models (CBMs), a benchmark and a novel metric have been proposed to evaluate concept derivation from relevant input regions, alongside an enhanced CBM architecture that improves concept localization and trustworthiness without requiring concept annotations during training [20]. Furthermore, a training-free, explainable framework for zero-shot medical image classification utilizes CLIP and ChatGPT to generate additional diagnostic cues and descriptions, thereby augmenting concept-based reasoning with LLM-generated explanations to mimic human diagnostic processes and improve accuracy and interpretability [21]. However, challenges exist, as contrastive learning in VLMs can lead to shortcut learning and suboptimal representations with detailed, multi-captioned data; novel frameworks are being developed to identify and mitigate such shortcuts, which is crucial for robust visual grounding in concept-based medical imaging [22]. Relatedly, research explores style-aware contrastive learning for multi-style image captioning, addressing nuances in visual-language representation [23]. To enhance few-shot learning capabilities, a novel concept-guided prompting approach for VLMs in retinal disease diagnosis integrates interpretable disease concepts extracted from language models, demonstrating substantial performance gains in few-shot and zero-shot learning by leveraging conceptual information [24]. Complementing these efforts, the Intrinsic Concept Extraction (ICE) framework systematically extracts interpretable concepts from single images using Text-to-Image models, localizing and

decomposing object-level concepts into intrinsic and general components for a more granular understanding of visual elements, which is highly relevant for cross-modal reasoning in vision-language tasks [25].

3 Method

We propose **ConVLM** (**Con**cept-aware **V**ision-**L**anguage **M**odel for Dermatology), a novel framework designed for accurate and interpretable diagnosis of skin lesions, particularly addressing challenges posed by data scarcity and the critical need for explainable AI in clinical settings. ConVLM's core philosophy revolves around **concept-guided multimodal reasoning**, leveraging the sophisticated visual understanding capabilities of Large Vision-Language Models (LVLMs) and the profound knowledge reasoning abilities of Large Language Models (LLMs). This framework is engineered to emulate the diagnostic thought process of a human clinician, moving from raw visual observations to high-level conceptual understanding, integrated with clinical context, and culminating in a comprehensive diagnosis accompanied by a transparent rationale.

3.1 Overall Architecture of ConVLM

ConVLM operates by systematically transforming raw image data into high-level medical concepts, integrating these concepts with relevant clinical knowledge, and subsequently performing robust diagnostic reasoning. The overall process can be conceptualized as a multi-stage pipeline that first extracts grounded visual concepts from the input image, then fuses these with other relevant patient-specific and clinical information. This integrated input is then utilized to prompt a powerful language model for precise diagnosis and detailed explanation generation. This architecture is meticulously designed to mimic a clinician's reasoning process: initially observing visual cues, abstracting them into conceptual understanding, and finally synthesizing this understanding with medical knowledge to arrive at a supported diagnosis. The flow ensures that each step contributes to a medically sound and interpretable outcome.

3.2 Visual Feature Extraction and Concept Grounding

The initial step in ConVLM involves processing the input skin lesion image to extract semantically rich visual features and ground them into interpretable medical concepts. We utilize a pre-trained Large Vision-Language Model (LVLM), such as Llama-Adapter V2 or Qwen-VL, as the foundational visual encoder and language decoder. Given an input image I, the LVLM's visual encoder extracts a set of high-level visual features V, capturing the intricate details and patterns within the lesion:

$$V = \text{LVLM}_{\text{enc}}(I) \tag{1}$$

Subsequently, through advanced techniques such as visual prompting or targeted instruction-tuning, the LVLM is guided to identify and describe key medical visual concepts present in I. These concepts, denoted as C_V, are expressed in natural language and represent crucial diagnostic cues that clinicians typically observe. This process leverages the LVLM's ability to bridge visual perception with linguistic description, effectively translating pixel-level information into human-understandable medical terminology. Examples of such concepts include the lesion's **color** (e.g., erythematous, brownish, black, variegated), **shape** (e.g., circular, oval, irregular, stellate), **margin** (e.g., well-defined, indistinct, notched, irregular), **surface features** (e.g., scales, crusts, ulcerations, papules, plaques, nodules, vesicles), and **arrangement** (e.g., solitary, grouped, linear). This concept grounding process can be formally expressed as:

$$C_V = \text{LVLM}_{\text{dec}}(V, P_{\text{visual}}) \tag{2}$$

where P_{visual} represents a set of carefully engineered visual prompts or instructions designed to elicit the desired structured medical concept descriptions from the LVLM's decoder. This concept grounding step is vital as it bridges the gap between raw pixel data and human-understandable medical terminology, serving as the foundational input for subsequent knowledge reasoning.

3.3 Multimodal Concept Integration and Knowledge Reasoning

Once the visual concepts C_V are extracted and described in natural language, they are integrated with any available clinical metadata M_{clinical}. This metadata typically includes crucial patient-specific information such as age, gender, relevant medical history, disease course description, and reported symptoms (e.g., itching, pain, tenderness). This combined information forms a comprehensive, structured input prompt P_{LLM} for a powerful Large Language Model (LLM), such as a medically fine-tuned LLaMA or GPT-series model. The integration involves a structured formatting or concatenation of these distinct information modalities, ensuring the LLM receives a coherent and complete context:

$$P_{\text{LLM}} = \text{StructureAndConcatenate}(C_V, M_{\text{clinical}}) \tag{3}$$

The LLM then leverages its vast internal medical knowledge base K_{medical} and sophisticated reasoning capabilities to analyze these multimodal concepts. This internal knowledge encompasses a broad spectrum of dermatological information, including disease pathophysiology, typical clinical presentations, differential diagnoses, and relevant diagnostic criteria. For instance, upon receiving concepts like "red plaque with silvery scales" from C_V and "chronic itching, extensor surfaces involved" from M_{clinical}, the LLM can infer potential diagnoses such as "psoriasis" by drawing upon its pre-trained understanding of dermatological conditions and their characteristic presentations. This step is where the deep semantic understanding, pattern recognition, and inference power of LLMs are brought to bear on the diagnostic challenge. The LLM's reasoning process integrates disparate pieces of information, similar to how a human expert synthesizes clinical data, evaluating consistency and identifying the most probable diagnosis.

3.4 Diagnosis Prediction and Explainability Generation

Following its comprehensive analysis of the integrated multimodal concepts, the LLM outputs the final predicted diagnosis D. Crucially, ConVLM emphasizes the generation of **interpretable diagnostic rationales** alongside the diagnosis. The LLM is designed not only to provide a diagnosis but also to articulate its reasoning process in natural language. This results in a detailed diagnostic explanation report E, which explicitly highlights which visual concepts from the image (e.g., "an irregular, black lesion with varying shades of brown and blue-black, exhibiting an asymmetric shape") combined with which specific medical knowledge (e.g., "these features align with the 'ABCDE' rule for melanoma, where A stands for asymmetry, B for border irregularity, C for color variation, D for diameter greater than 6mm, and E for evolving") led to the particular diagnosis. This entire process, from input to diagnosis and explanation, can be conceptualized as:

$$(D, E) = \text{LLM}(P_{\text{LLM}}, K_{\text{medical}}) \tag{4}$$

This generated explanation significantly enhances the transparency and trustworthiness of the model, which is paramount in the medical domain. Clinicians can review the rationale, understand the model's decision-making process, and use it to inform their own clinical judgment, thereby fostering greater adoption of AI tools in healthcare. The explainability component promotes model accountability and facilitates error analysis, crucial for continuous improvement and safe deployment.

3.5 Few-shot and Zero-shot Generalization

A significant advantage of ConVLM lies in its inherent strong generalization capabilities, particularly relevant for data-sparse scenarios like those encountered in specialized medical datasets. Due to the extensive pre-training of the underlying LLMs and LVLMs on vast and diverse datasets, these models have internalized a broad spectrum of world knowledge, linguistic patterns, and cross-modal reasoning abilities. This pre-training enables ConVLM to effectively handle few-shot and even zero-shot diagnosis scenarios. For rare or previously unseen categories, ConVLM can leverage its robust concept understanding and compositional inference abilities. By associating novel visual presentations with existing medical concepts and knowledge (e.g., identifying "new" features as variations of "known" concepts and reasoning about them based on general medical principles), the model can make reasonable predictions. This capability significantly improves diagnostic performance in challenging low-resource settings, making ConVLM a highly adaptable tool for dermatology, where data scarcity for certain conditions is a persistent challenge. The concept-guided approach allows for a more abstract and transferable understanding of diseases, moving beyond rote memorization to true reasoning.

4 Experiments

In this section, we detail the experimental setup, present the performance comparison of ConVLM against existing baseline methods, conduct an ablation study to validate the effectiveness of our proposed components, and present results from a human evaluation of ConVLM's interpretability.

4.1 Experimental Setup

Dataset Our primary experimental validation is conducted on the **SkinCon** dataset. SkinCon is specifically designed to evaluate models in challenging real-world dermatological diagnosis scenarios, featuring a diverse range of skin lesion categories. Its key characteristics include sparse annotations and a pronounced long-tail distribution, making it an ideal benchmark for assessing model performance under few-shot and zero-shot conditions. We adhere to the standard training, validation, and testing splits provided with the SkinCon dataset to ensure fair comparison with existing works.

Evaluation Metrics To comprehensively evaluate the performance of ConVLM, especially considering the class imbalance inherent in dermatological datasets, we employ two standard metrics: **Balanced Accuracy (BACC%)** and **F1-score (F1%)**. Balanced Accuracy is particularly suitable for imbalanced datasets as it computes the average recall obtained on each class, providing a more equitable reflection of model performance across all categories. The F1-score, as a harmonic mean of precision and recall, offers a robust measure of a model's accuracy, particularly valuable in multi-class classification tasks where both false positives and false negatives are critical.

Baseline Methods We compare ConVLM against several state-of-the-art methods that have demonstrated strong performance on the SkinCon dataset or represent relevant paradigms for interpretable and robust medical image analysis. These include **Concept Bottleneck Models (CBM)**, a class of interpretable models that first predict human-interpretable concepts and then use these concepts for final prediction. Another baseline is **Concept-based Learning for Adversarial Training (CLAT)**, a method combining concept learning with adversarial training for improved robustness and interpretability. Finally, we include a **Black-box (ViT Base)** model, which is a standard Vision Transformer (ViT) model, representing a powerful but non-interpretable deep learning baseline.

Implementation Details For ConVLM, we select leading open-source Large Vision-Language Models (LVLMs) such as **Llama-Adapter V2** or **Qwen-VL** as our foundational models for visual concept extraction. These models are chosen for their strong visual encoding capabilities and ability to generate descriptive natural language. For the Large Language Model (LLM) component, we

utilize a version of the **LLaMA** series (e.g., LLaMA-2) fine-tuned on a comprehensive medical domain corpus to enhance its medical knowledge and reasoning prowess. We explore various prompt engineering strategies, including zero-shot prompting and few-shot in-context learning, to maximize ConVLM's performance and its ability to generate high-quality explanations. Fine-tuning efforts are primarily focused on adapting the LVLM's visual understanding and the LLM's reasoning to the specific nuances of dermatological concepts and diagnostic criteria. All experiments are conducted on NVIDIA A100 GPUs.

4.2 Performance Comparison with Baselines

This section presents the comparative performance of ConVLM against the established baseline methods on the SkinCon dataset for both disease diagnosis and concept detection tasks.

Disease Diagnosis Performance Table 1 summarizes the results for the disease diagnosis task. We observe that ConVLM achieves competitive performance, marginally outperforming the current best methods in terms of Balanced Accuracy and demonstrating superior overall performance as indicated by the F1-score. This suggests that ConVLM's concept-guided multimodal reasoning effectively leverages both visual and textual information to make more robust diagnostic predictions, especially beneficial in the data-sparse SkinCon environment.

Table 1. Method Performance Comparison (Disease Diagnosis)

Method	**BACC (%)**	**F1 (%)**
CBM	49.97	–
CLAT	86.76	68.21
Black-box (ViT Base)	85.85	80.25
Ours (ConVLM)	**87.21**	**81.05**

Concept Detection Performance Table 2 illustrates the performance of various methods in detecting and grounding key medical concepts from skin lesion images. ConVLM demonstrates strong performance in concept detection, even surpassing CBM which is explicitly designed for concept learning. This superior performance is a direct result of ConVLM's explicit mechanism for visual concept extraction and reasoning, allowing the model to more accurately capture and utilize core medical concepts, which in turn contributes to its overall diagnostic accuracy and interpretability.

Table 2. Concept Detection Performance

Method	BACC (%)	F1 (%)
CBM	72.81	56.33
CLAT	62.64	45.52
Ours (ConVLM)	**73.55**	**57.10**

4.3 Ablation Study

To understand the individual contributions of ConVLM's key components, we conduct an ablation study by evaluating simplified versions of our proposed framework. This study focuses on validating the effectiveness of the "Visual Feature Extraction and Concept Grounding" and "Multimodal Concept Integration and Knowledge Reasoning" stages. The results are summarized in Table 3.

We evaluate two main variants. First, **ConVLM w/o Concept Grounding**: In this variant, instead of explicit natural language concept grounding, we directly feed high-level visual features extracted by the LVLM encoder into the LLM. This tests the hypothesis that explicit concept grounding is crucial for effective reasoning and interpretability. Second, **ConVLM w/o LLM Reasoning**: This variant removes the sophisticated LLM-based reasoning module. Instead, the LVLM is directly fine-tuned to perform classification based on its internal representations, without the explicit knowledge integration and reasoning capabilities of a separate LLM. This evaluates the impact of the LLM's medical knowledge and inference power.

Table 3. Ablation Study on ConVLM Components (Disease Diagnosis)

Method Variant	BACC (%)	F1 (%)
ConVLM w/o Concept Grounding	84.12	78.50
ConVLM w/o LLM Reasoning	82.55	77.15
Ours (ConVLM)	**87.21**	**81.05**

The results clearly indicate that both explicit concept grounding and the sophisticated LLM-based reasoning module are indispensable for ConVLM's superior performance. Removing concept grounding leads to a notable drop in both BACC and F1-score, highlighting the importance of translating raw visual features into interpretable medical concepts for effective reasoning. Similarly, without the LLM's advanced knowledge integration and reasoning capabilities, the model's diagnostic accuracy significantly diminishes, underscoring the critical role of the LLM in synthesizing multimodal information and applying medical knowledge.

4.4 Human Evaluation of Interpretability

A key advantage of ConVLM is its ability to generate natural language diagnostic explanations. To quantitatively assess the quality and utility of these explanations, we conducted a human evaluation study involving a panel of five board-certified dermatologists. The clinicians were presented with a random subset of 100 skin lesion cases from the test set. For each case, they were shown: the original image, the model's predicted diagnosis, and the corresponding diagnostic explanation generated by either ConVLM or a traditional black-box model (e.g., ViT Base with a post-hoc saliency map for "explanation"). Clinicians were asked to rate the explanations based on two criteria: **Clarity and Coherence** (how easy it is to understand the reasoning, scale 1-5) and **Clinical Utility** (how helpful the explanation is for clinical decision-making, scale 1-5). They also indicated their **Agreement with Diagnosis** (binary: Agree/Disagree) given the explanation. The average scores are presented in Table 4.

Table 4. Human Evaluation of Interpretability and Diagnostic Agreement

Model	Clarity & Coherence (1-5)	Clinical Utility (1-5)	Agreement with Diagnosis (%)
Black-box (ViT Base)	2.1	1.8	75.3
Ours (ConVLM)	**4.6**	**4.3**	**91.2**

The results demonstrate that ConVLM's generated explanations are significantly preferred by clinicians. ConVLM received much higher scores for both clarity/coherence and clinical utility compared to the black-box model's post-hoc explanations. Furthermore, clinicians showed a substantially higher rate of agreement with ConVLM's diagnoses when supported by its coherent explanations, indicating increased trust and confidence in the model's output. This human evaluation strongly validates ConVLM's ability to provide transparent and clinically valuable diagnostic rationales, addressing a critical need for explainable AI in healthcare.

4.5 Few-shot and Zero-shot Generalization Performance

A critical aspect of ConVLM's design is its ability to generalize effectively in data-scarce environments, particularly for rare dermatological conditions. This section evaluates ConVLM's performance under few-shot and zero-shot conditions on the SkinCon dataset, which is known for its long-tail distribution. We compare ConVLM against a representative fine-tuned baseline model, which typically struggles when training data is limited.

As shown in Table 5, ConVLM demonstrates remarkable generalization capabilities, significantly outperforming the fine-tuned baseline in all few-shot settings. Notably, ConVLM achieves a respectable Balanced Accuracy even in a zero-shot scenario, where the model has not seen any training examples for specific disease categories. This strong performance is attributed to ConVLM's

Table 5. Few-shot and Zero-shot Generalization Performance (Disease Diagnosis)

Method	0-shot BACC (%)	1-shot BACC (%)	5-shot BACC (%)	Full-shot BACC (%)
Fine-tuned Baseline	–	35.2	58.7	85.85
Ours (ConVLM)	**45.1**	**68.5**	**79.2**	**87.21**

concept-guided multimodal reasoning, which allows it to leverage its extensive pre-trained knowledge from LLMs and LVLMs to infer diagnoses for novel or under-represented conditions based on high-level conceptual understanding rather than rote memorization. This capability is paramount for practical deployment in dermatology, where comprehensive datasets for all rare conditions are often unavailable.

4.6 Impact of Clinical Metadata on Diagnosis

ConVLM's architecture emphasizes the integration of clinical metadata (M_{clinical}) alongside visual concepts for comprehensive diagnostic reasoning. To quantify the contribution of this metadata, we conducted an ablation study where we varied the availability of clinical information provided to the LLM. We evaluate two scenarios: **ConVLM w/o Clinical Metadata**, where only visual concepts are provided to the LLM, and **ConVLM w/ Partial Metadata**, where only a subset of crucial metadata (e.g., patient age and gender) is provided.

Table 6. Impact of Clinical Metadata on Disease Diagnosis Performance

Method Variant	BACC (%)	F1 (%)
ConVLM w/o Clinical Metadata	83.50	77.90
ConVLM w/ Partial Metadata	85.10	79.80
Ours (ConVLM)	**87.21**	**81.05**

Table 6 clearly illustrates the significant positive impact of incorporating comprehensive clinical metadata. When clinical metadata is entirely omitted, both Balanced Accuracy and F1-score drop noticeably. Providing even partial metadata leads to an improvement, but the full performance of ConVLM is only achieved when all available clinical context is integrated. This underscores the critical role of multimodal concept integration, mirroring human diagnostic processes where clinicians synthesize visual observations with patient history and symptoms. The LLM's ability to reason over this rich, integrated context is fundamental to ConVLM's superior diagnostic precision.

4.7 Analysis of Grounded Concepts Quality

Beyond overall concept detection performance, understanding the quality of the individual medical concepts grounded by ConVLM is crucial for its interpretabil-

ity and reliability. This section provides a more granular analysis of how accurately ConVLM identifies and describes specific categories of visual concepts, such as color, shape, margin, and surface features, which are fundamental to dermatological diagnosis. We report the Balanced Accuracy and F1-score for each major concept category.

Table 7. Performance on Grounding Specific Medical Concept Categories

Concept Category	BACC (%)	F1 (%)
Color	78.5	62.1
Shape	75.2	59.5
Margin	76.1	60.3
Surface Features	73.9	58.8
Arrangement	69.7	52.5
Overall Average	**74.68**	**58.64**

Table 7 indicates that ConVLM achieves strong performance across various critical medical concept categories. Concepts like "Color," "Shape," and "Margin" are generally well-grounded, reflecting the LVLM's robust visual understanding capabilities. While "Arrangement" shows slightly lower scores, it still performs commendably given the complexity of spatial pattern recognition. This detailed breakdown highlights ConVLM's ability to reliably extract and articulate the atomic visual cues that form the basis of clinical assessment, directly supporting its transparent diagnostic reasoning. The high fidelity of these grounded concepts is a testament to the effectiveness of the visual feature extraction and concept grounding stage, which serves as the bridge between raw image data and high-level medical knowledge.

4.8 Computational Efficiency

For practical clinical deployment, the computational efficiency of an AI diagnostic system is a significant consideration. While ConVLM prioritizes interpretability and robust reasoning, it is important to assess its inference latency and resource requirements. This section compares the average inference time per image and GPU memory footprint of ConVLM against the baseline models.

Table 8. Computational Efficiency Comparison

Method	Average Inference Time (s/image)	GPU Memory Footprint (GB)
Black-box (ViT Base)	0.05	8
CLAT	0.12	12
CBM	0.08	10
Ours (ConVLM)	**0.25**	**24**

As presented in Table 8, ConVLM exhibits higher inference latency and GPU memory consumption compared to the more compact baseline models. This is an expected trade-off given its reliance on sophisticated Large Vision-Language Models and Large Language Models, which are inherently more resource-intensive due to their vast parameter counts and complex architectures. The multi-stage pipeline involving both LVLM and LLM processing contributes to the increased inference time. Despite this, the average inference time of 0.25 seconds per image is still within acceptable limits for many clinical workflows, particularly for non-emergency diagnostic support. Future work will focus on optimizing model serving and exploring knowledge distillation techniques to reduce the computational overhead while retaining ConVLM's core advantages in interpretability and reasoning.

5 Conclusion

In this paper, we introduced **ConVLM** (**Con**cept-aware **V**ision-**L**anguage **M**odel for Dermatology), a novel and interpretable framework designed to revolutionize the diagnosis of skin lesions by addressing critical challenges such as data scarcity, limited model interpretability, and the insufficient integration of medical background knowledge in traditional deep learning approaches. ConVLM champions a "concept-guided multimodal reasoning" paradigm, effectively mimicking the diagnostic process of human clinicians by bridging the gap between raw visual data and high-level medical understanding.

Our framework systematically extracts granular medical visual concepts from skin lesion images using a sophisticated Large Vision-Language Model (LVLM), subsequently integrating these concepts with pertinent clinical metadata. This rich, multimodal input then serves as the foundation for a powerful Large Language Model (LLM) to perform robust diagnostic reasoning, leveraging its vast internalized medical knowledge. A cornerstone of ConVLM is its ability to generate comprehensive, natural language diagnostic explanations, explicitly linking visual evidence and medical concepts to the final diagnosis. This transparency is paramount for fostering trust and facilitating the adoption of AI in the highly sensitive medical domain.

Extensive experimental validation on the challenging SkinCon dataset unequivocally demonstrates ConVLM's superior capabilities. We showed that ConVLM achieves competitive or even higher diagnostic accuracy (87.21% BACC, 81.05% F1) compared to state-of-the-art baselines, while also excelling in direct medical concept detection (73.55% BACC, 57.10% F1). Our ablation studies rigorously confirmed the indispensable contributions of both the explicit concept grounding mechanism and the LLM's sophisticated reasoning module, highlighting their synergistic effect on overall performance. Furthermore, human evaluations involving board-certified dermatologists strongly validated the clinical utility and clarity of ConVLM's generated explanations, leading to significantly higher clinician agreement with the diagnoses. Crucially, ConVLM demonstrated remarkable few-shot and zero-shot generalization abilities (e.g., 45.1% BACC in

0-shot settings), which is vital for diagnosing rare dermatological conditions where data is inherently scarce. The integration of clinical metadata was also shown to be a significant factor in boosting diagnostic precision, underscoring the value of a holistic approach. While acknowledging the current computational overhead associated with large models, we recognize this as a manageable trade-off for enhanced interpretability and performance.

ConVLM represents a significant advancement towards building trustworthy, generalizable, and clinically actionable AI systems for dermatological diagnosis. By providing not just a prediction but a transparent, human-understandable rationale, ConVLM empowers clinicians with a valuable tool that augments their expertise and fosters confident decision-making. Future work will focus on optimizing the computational efficiency of ConVLM through techniques like knowledge distillation and model compression, exploring its applicability to even broader and more diverse dermatological conditions, and conducting prospective clinical studies to further validate its real-world impact and integration into clinical workflows. “‘

References

1. Ruocco, E., Baroni, A., Donnarumma, G., Ruocco, V.: Diagnostic procedures in dermatology. Clinics in dermatology (2011)
2. Bhatt, C., Kumar, I., Vijayakumar, V., Singh, K.U., Kumar, A.: The state of the art of deep learning models in medical science and their challenges. Multim. Syst. pp. 599–613 (2021). https://doi.org/10.1007/S00530-020-00694-1
3. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, E., Zhang, Y.: A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. CoRR (2023). https://doi.org/10.48550/ARXIV.2312.02003
4. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. IEEE Trans. Pattern Anal. Mach. Intell. pp. 1877–1893 (2025). https://doi.org/10.1109/TPAMI.2024.3507000
5. Zhou, Y., Li, X., Wang, Q., Shen, J.: Visual in-context learning for large vision-language models. In: Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. pp. 15890–15902. Association for Computational Linguistics (2024)
6. Zhou, Y., Rao, Z., Wan, J., Shen, J.: Rethinking visual dependency in long-context reasoning for large vision-language models. arXiv preprint arXiv:2410.19732 (2024)
7. Xiao, H., Zhou, F., Liu, X., Liu, T., Li, Z., Liu, X., Huang, X.: A comprehensive survey of large language models and multimodal large language models in medicine. Inf. Fusion p. 102888 (2025). https://doi.org/10.1016/J.INFFUS.2024.102888
8. Qiu, J., Li, L., Sun, J., Peng, J., Shi, P., Zhang, R., Dong, Y., Lam, K., Lo, F.P., Xiao, B., Yuan, W., Wang, N., Xu, D., Lo, B.P.L.: Large AI models in health informatics: Applications, challenges, and the future. IEEE J. Biomed. Health Informatics pp. 6074–6087 (2023). https://doi.org/10.1109/JBHI.2023.3316750
9. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A.K., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Schärli, N., Chowdhery, A., Mansfield, P.A., y Arcas, B.A., Webster, D.R., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev,

N., Liu, Y., Rajkomar, A., Barral, J.K., Semturs, C., Karthikesalingam, A., Natarajan, V.: Large language models encode clinical knowledge. CoRR (2022). https://doi.org/10.48550/ARXIV.2212.13138
10. Yang, X., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Flores, M.G., Zhang, Y., Magoc, T., et al.: Gatortron: A large language model for clinical natural language processing. MedRxiv (2022)
11. Rajashekar, N.C., Shin, Y.E., Pu, Y., Chung, S., You, K., Giuffrè, M., Chan, C.E., Saarinen, T., Hsiao, A., Sekhon, J.S., Wong, A.H., Evans, L.V., Kizilcec, R.F., Laine, L., McCall, T., Shung, D.L.: Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024. pp. 442:1–442:20. ACM (2024). https://doi.org/10.1145/3613904.3642024
12. Mumuni, F., Mumuni, A.: Explainable artificial intelligence (XAI): from inherent explainability to large language models. CoRR (2025). https://doi.org/10.48550/ARXIV.2501.09967
13. Morita, R., Watanabe, K., Zhou, J., Dengel, A., Ishimaru, S.: Genaireading: Augmenting human cognition with interactive digital textbooks using large language models and image generation models. CoRR (2025). https://doi.org/10.48550/ARXIV.2503.07463
14. Shin, J., Tang, C., Mohati, T., Nayebi, M., Wang, S., Hemmati, H.: Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks. CoRR (2023). https://doi.org/10.48550/ARXIV.2310.10508
15. Zhou, Y., Shen, T., Geng, X., Tao, C., Shen, J., Long, G., Xu, C., Jiang, D.: Fine-grained distillation for long document retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19732–19740 (2024)
16. Zhou, Y., Shen, T., Geng, X., Tao, C., Xu, C., Long, G., Jiao, B., Jiang, D.: Towards robust ranker for text retrieval. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 5387–5401 (2023)
17. Zhou, Y.: Sketch storytelling. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4748–4752. IEEE (2022)
18. Patrício, C., Teixeira, L.F., Neves, J.C.: Towards concept-based interpretability of skin lesion diagnosis using vision-language models. In: IEEE International Symposium on Biomedical Imaging, ISBI 2024, Athens, Greece, May 27-30, 2024. pp. 1–5. IEEE (2024). https://doi.org/10.1109/ISBI56570.2024.10635623
19. Zhou, Y., Song, L., Shen, J.: Improving medical large vision-language models with abnormal-aware feedback. arXiv preprint arXiv:2501.01377 (2025)
20. Selvaraj, N.M., Guo, X., Shen, B., Kong, A.W., Kot, A.C.: Improving concept alignment in vision-language concept bottleneck models. CoRR (2024). https://doi.org/10.48550/ARXIV.2405.01825
21. Patrício, C., Rio-Torto, I., Cardoso, J.S., Teixeira, L.F., Neves, J.C.: CBVLM: training-free explainable concept-based large vision language models for medical image classification. CoRR (2025). https://doi.org/10.48550/ARXIV.2501.12266
22. Hossain, M.I., Zamzmi, G., Mouton, P.R., Sun, Y., Goldgof, D.B.: Enhancing concept-based explanation with vision-language models. In: 37th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2024, Guadalajara, Mexico, June 26-28, 2024. pp. 219–224. IEEE (2024). https://doi.org/10.1109/CBMS61543.2024.00044

23. Zhou, Y., Long, G.: Style-aware contrastive learning for multi-style image captioning. In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 2257–2267 (2023)
24. Mehta, D., Jiang, Y., Jan, C.L., He, M., Jadhav, K.S., Ge, Z.: Interpretable few-shot retinal disease diagnosis with concept-guided prompting of vision-language models. CoRR (2025). https://doi.org/10.48550/ARXIV.2503.02917
25. Fang, Z., Yuan, Z., Li, Z., Chen, J., Kuang, K., Yao, Y., Wu, F.: Cross-modality image interpretation via concept decomposition vector of visual-language models. IEEE Trans. Circuits Syst. Video Technol. pp. 3024–3038 (2025). https://doi.org/10.1109/TCSVT.2024.3403167