FiLMed: Fine-Grained Visual Tokens Align with Localized Semantics

Zhuohao Ni[1], Xiaoxiao Li[1,2]
1. Department of ECE, University of British Columbia
2. Vector Institute

ICML International Conference On Machine Learning
Actionable Interpretability Workshop

Background

Deep learning models match dermatologists in skin-lesion classification but operate as opaque "**black boxes**." There is a pressing need for explanations that align with how clinicians reason about images.

Recent works have begun bridging this gap:
ExpLICD [1]: Maps diagnostic to vision-language embeddings and aligns image features to these concept for explainable classification.

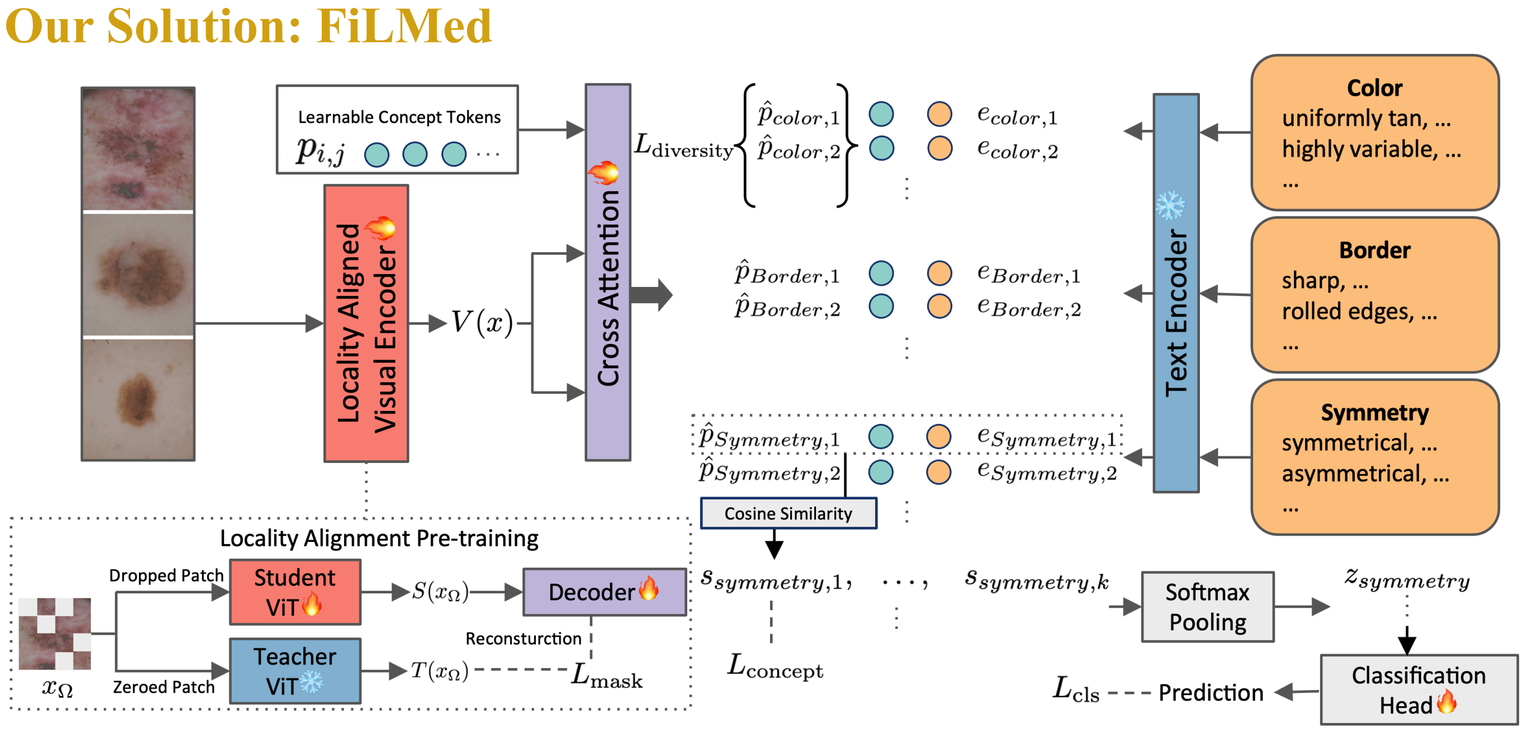
Locality Alignment [2]: Applies masked-patch self-distillation to refine Vision Transformer embeddings, strengthening local semantic encoding and spatial reasoning.

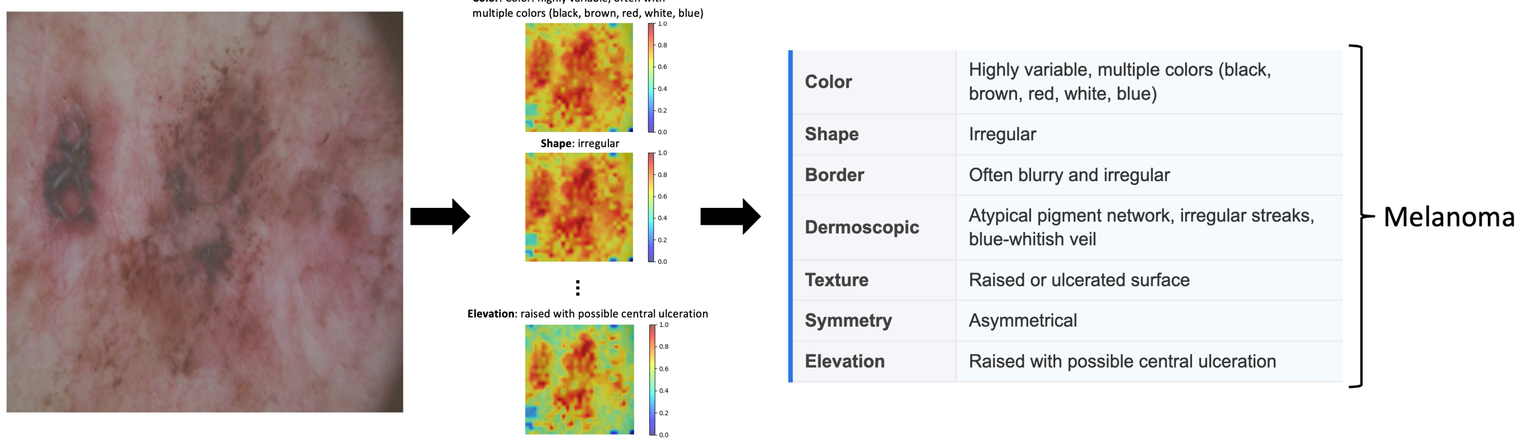
Motivation

Challenges:
- **Dermatology models → black boxes**.
- **Saliency maps** → show **where** but not **what**.
- **Concept models** → give **what** but not **where** and need heavy annotation.

Motivation:
Design an end-to-end model that combine attribute-level interpretability with spatial localization of features, all while maintaining diagnostic performance.

Our Solution: FiLMed





Color	Highly variable, multiple colors (black, brown, red, white, blue)
Shape	Irregular
Border	Often blurry and irregular
Dermoscopic	Atypical pigment network, irregular streaks, blue-whitish veil
Texture	Raised or ulcerated surface
Symmetry	Asymmetrical
Elevation	Raised with possible central ulceration

Melanoma

- **Locality-Aligned ViT**: Masked-patch distillation → patch tokens know "what + where". [2]
- **Attribute Tokens Embeddings**: Multiple learnable tokens per diagnostic axis to capture diverse looks.
- **Cross Attention Map**: Each token lights up its own region → attribute-level heat-maps.
- **Softmax Pooling**: Emphasizes the most salient attribute per axis while still aggregating all token scores → yields one concept logit per axis.
- **Concept-Guided Diagnosis**: Concatenate all concept logits → linear classifier → final lesion label.
- Bottom figure illustrates how FiLMed pinpoints each detected attribute on the lesion and lists them as explicit evidence—mirroring a clinician's step-by-step evaluation before making a diagnosis.

Quantitative Result

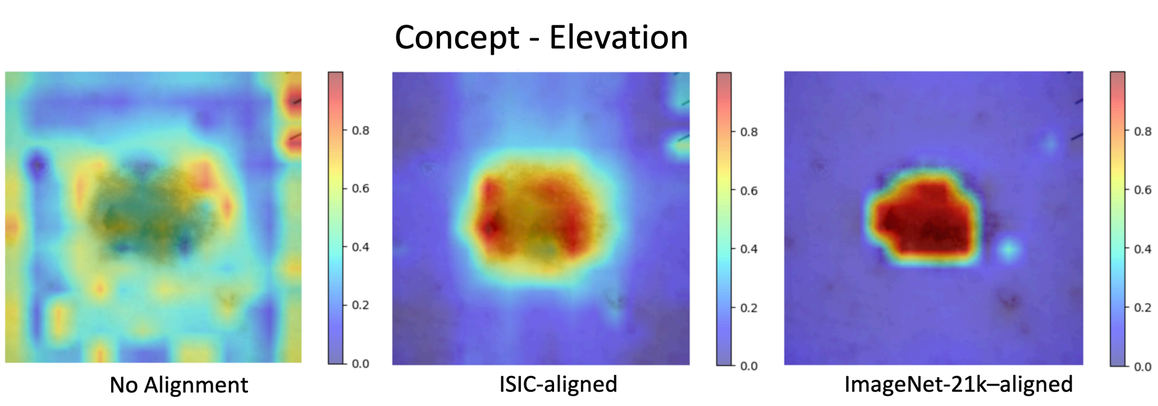
Table 1. ISIC 2018 test performance with different backbones.

Model	Backbone	Acc (%)	Macro-Sen	Macro-Spe	Macro-F1
FiLMed	ViT–SO400M-14–SigLIP@384	**91.2**	**86.9**	97.5	**86.0**
Black-box	ViT–SO400M-14–SigLIP@384	84.8	82.6	97.0	79.5
ExpLICD†	ViT–SO400M-14–SigLIP@384	86.1	76.6	96.4	76.3
ExpLICD	BiomedCLIP/16@224	90.5	85.6	**97.8**	84.6

Ablation Study

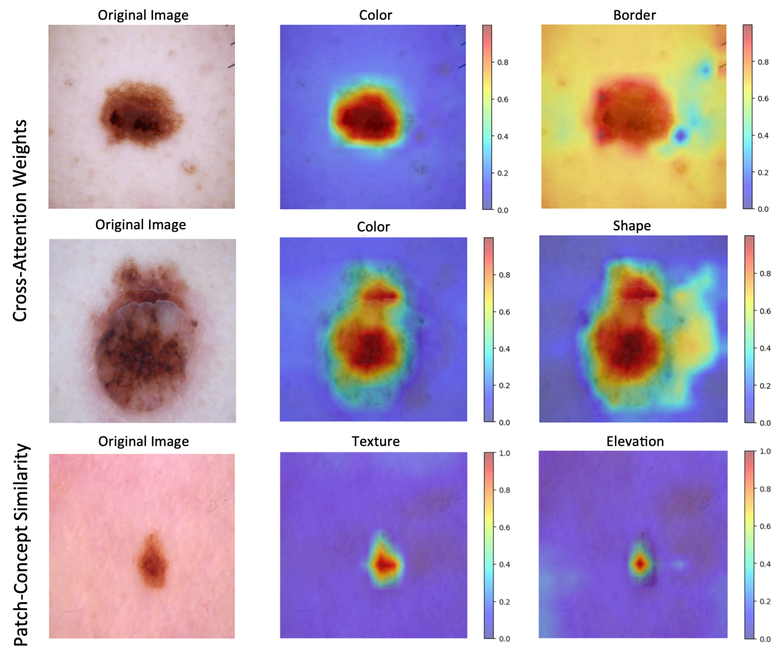
Table 2. Ablation of locality alignment data on classification performance (CLIP ViT-B/16).

Variant	Alignment Data	Accuracy (%)	BMAC
Baseline (no alignment)	w/o	85.1	72.8
+ Locality Alignment	ISIC 2018 (10 015 images)	86.4	78.3
+ Locality Alignment	ISIC 2020 (33 126 images)	86.1	76.8
+ Locality Alignment	ImageNet-21k (∼14 M images)	**86.7**	**79.4**



Locality alignment boosts spatial grounding. Without alignment, the "Elevation" heatmap is diffuse; ISIC-2018 alignment (10,015 images) concentrates on raised lesion areas; ImageNet-21k alignment (~14 M images) produces the sharpest, most focused activation—showing large-scale tuning yields the best attribute localization.

Qualitative Result



Patch-level explanations from FiLMed. Top: Cross-attention maps for "Color" and "Border" highlight each attribute's region on two lesion examples. Bottom: Patch concept similarity maps for "Texture" and "Elevation" confirm that **FiLMed's tokens focus on the clinically relevant subregions, offering faithful visual explanations for each attribute**.

References

[1] Y. Gao, D. Gu, M. Zhou & D. Metaxas, "Aligning human knowledge with visual concepts towards explainable medical image classification," MICCAI, 2024.
[2] I. Covert, T. Sun, J. Zou & T. Hashimoto, "Locality Alignment improves vision–language models," ICLR, 2024.