Quantitative Skin Attribute Grounding and Retrieval with Multimodal Large Language Models

Max Torop[1], Masih Eskandar[1], Jinyang Liu[1], Octavia Camps[1], Veronica Rotemberg[2], Jennifer Dy[1], Kivanc Kose[2]

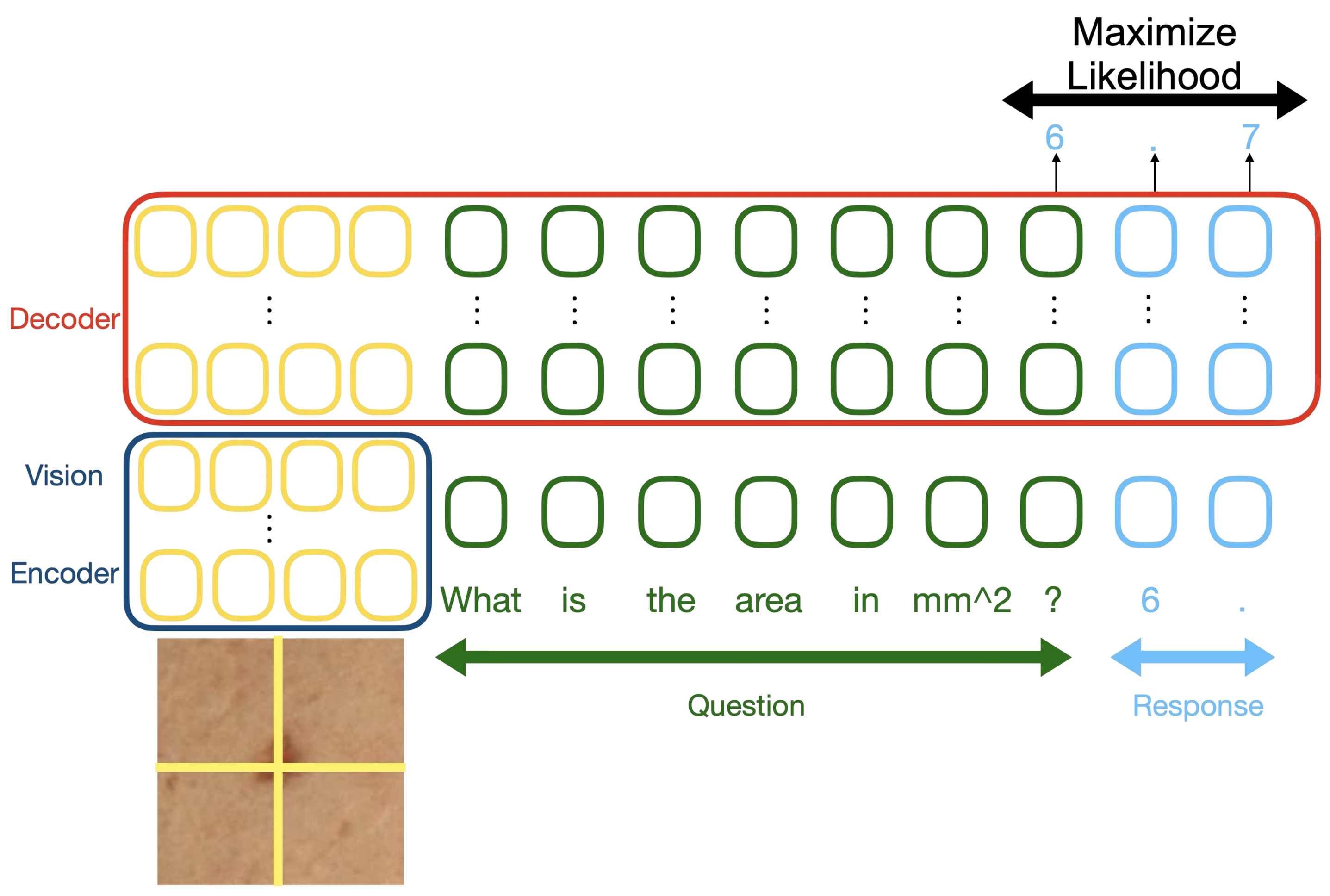
1. Northeastern University, Boston, MA, USA **2. Memorial Sloan Kettering Cancer Center, New York, NY, USA**

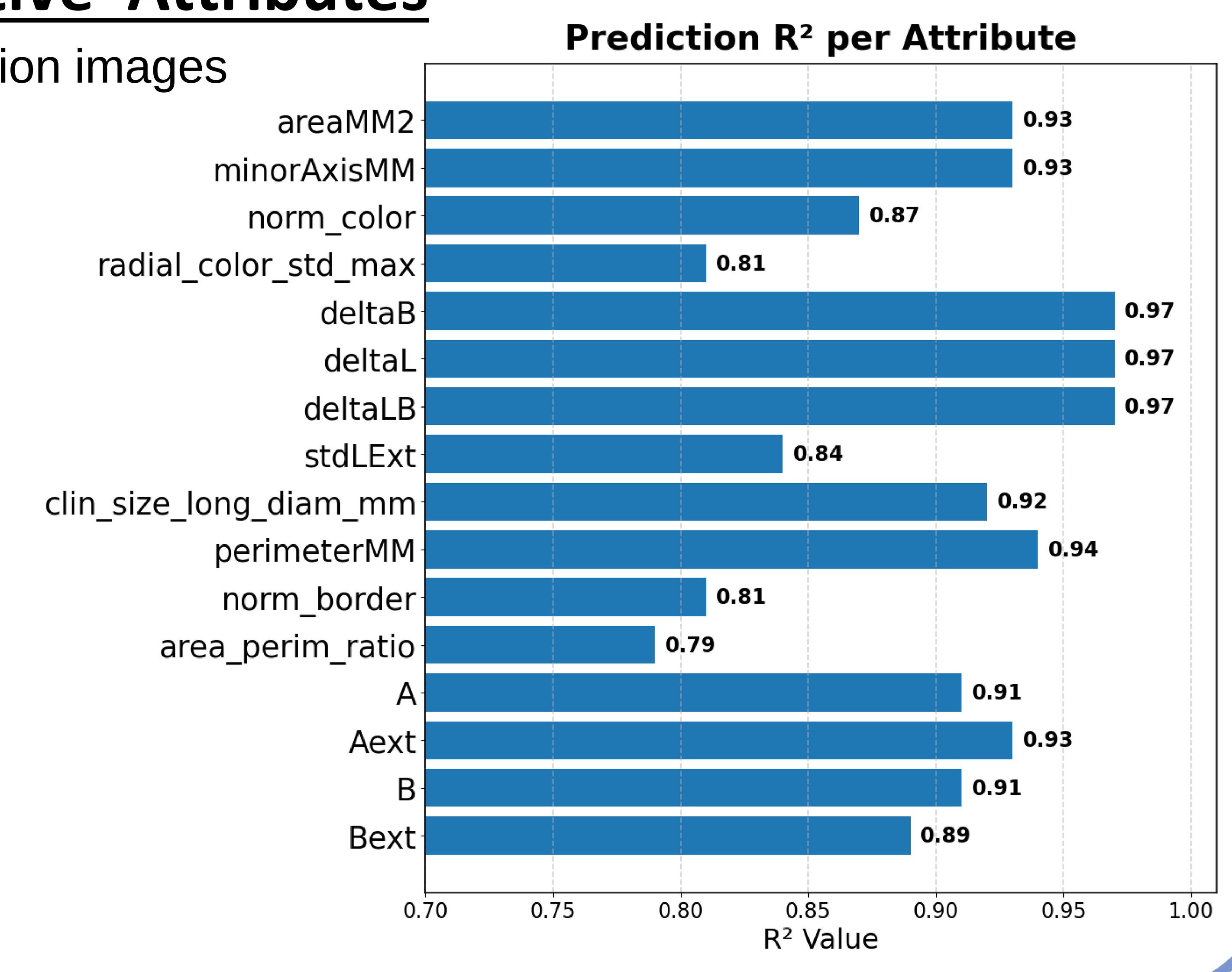
Background

- **Multimodal Large Language Models (MLLMs):** Promising avenue for interpretable interaction with data for clinicians
- **Quantitative Attributes**: Some phenotypical lesion attributes are predictive of malignancy[1]
  - These structural and visible features are inherently interpretable
- **Goal 1**: Assess if these quantitative attribute information can be learned by MLLMs
- **Goal 2**: Use the trained VLM for attribute-grounded content-based image search
- **Dataset**: SLICE-3D[2], ~400,000 TBP skin images with numerical quantitative attribute labels
- **Model**: Qwen2-VL[3]

Learning Quantitative Attributes

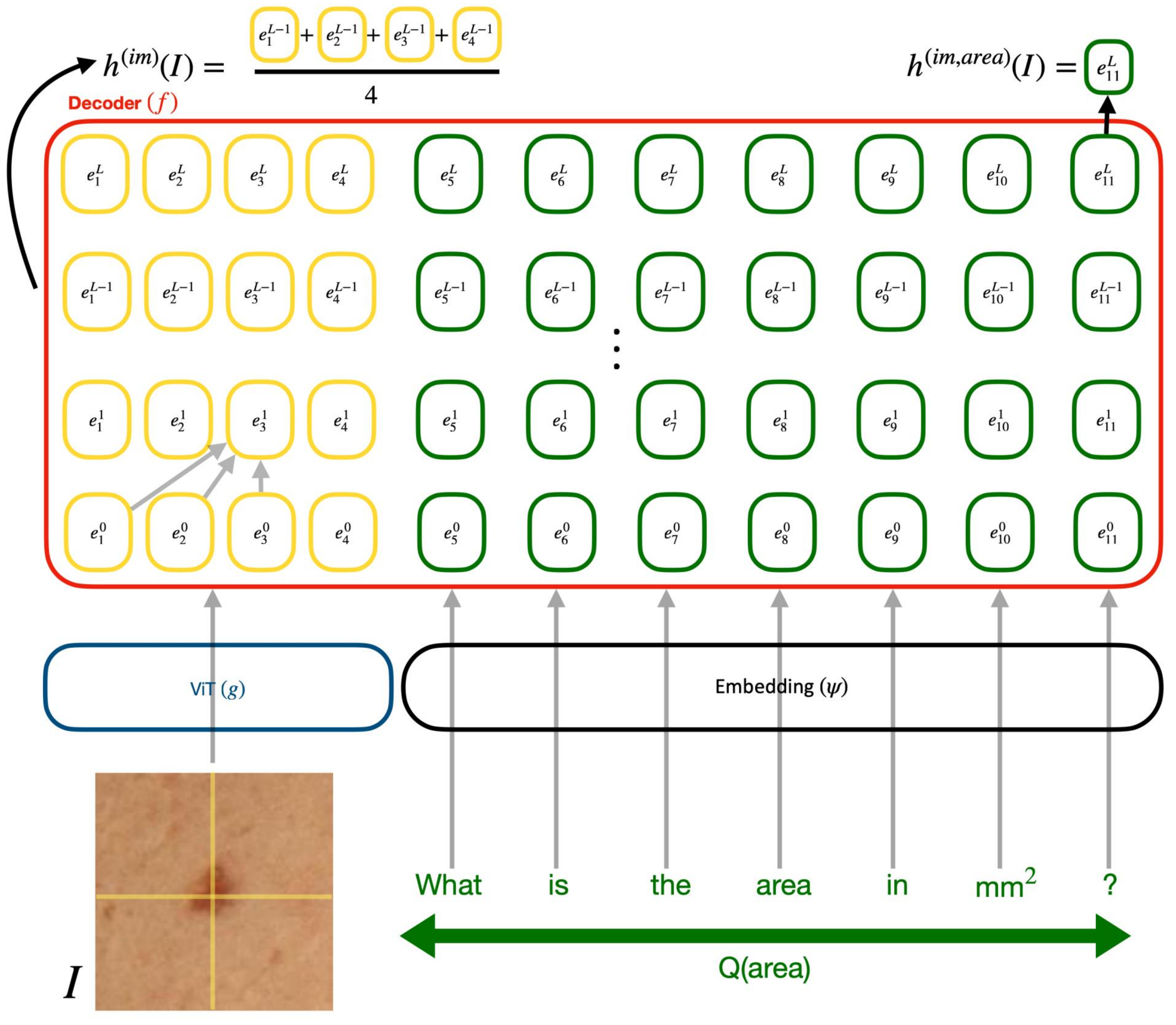
Finetune Qwen2-VL to predict 16 quantitative attributes from lesion images



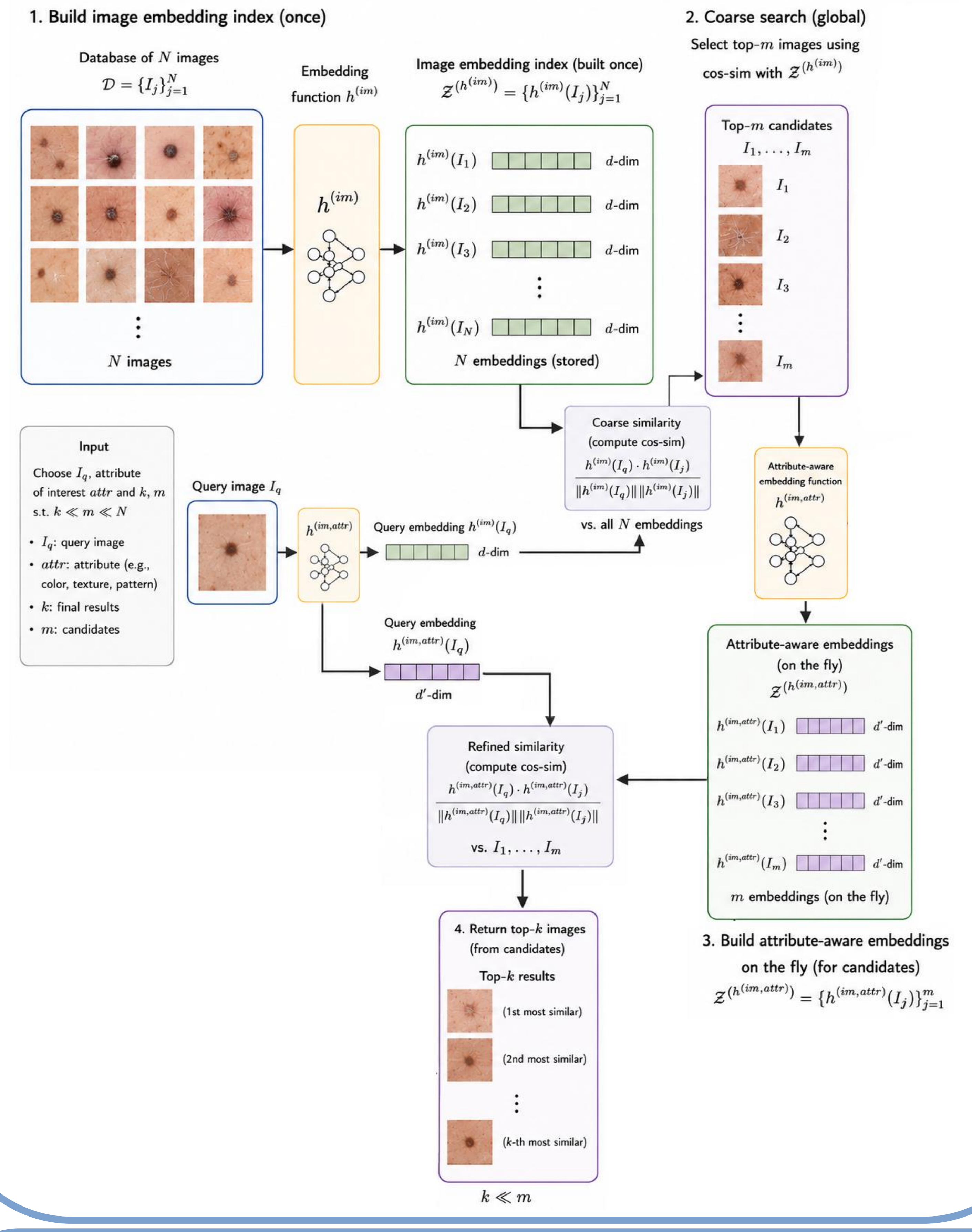


Embedding Strategies

- $h^{(im)}(I)$: comes from image
  - Captures general visual appearance
- $h^{(im,attr)}(I)$: comes from image and attribute-specific question text
  - Captures general appearance + attribute specific properties



Retrieval



Quantitative Search Results

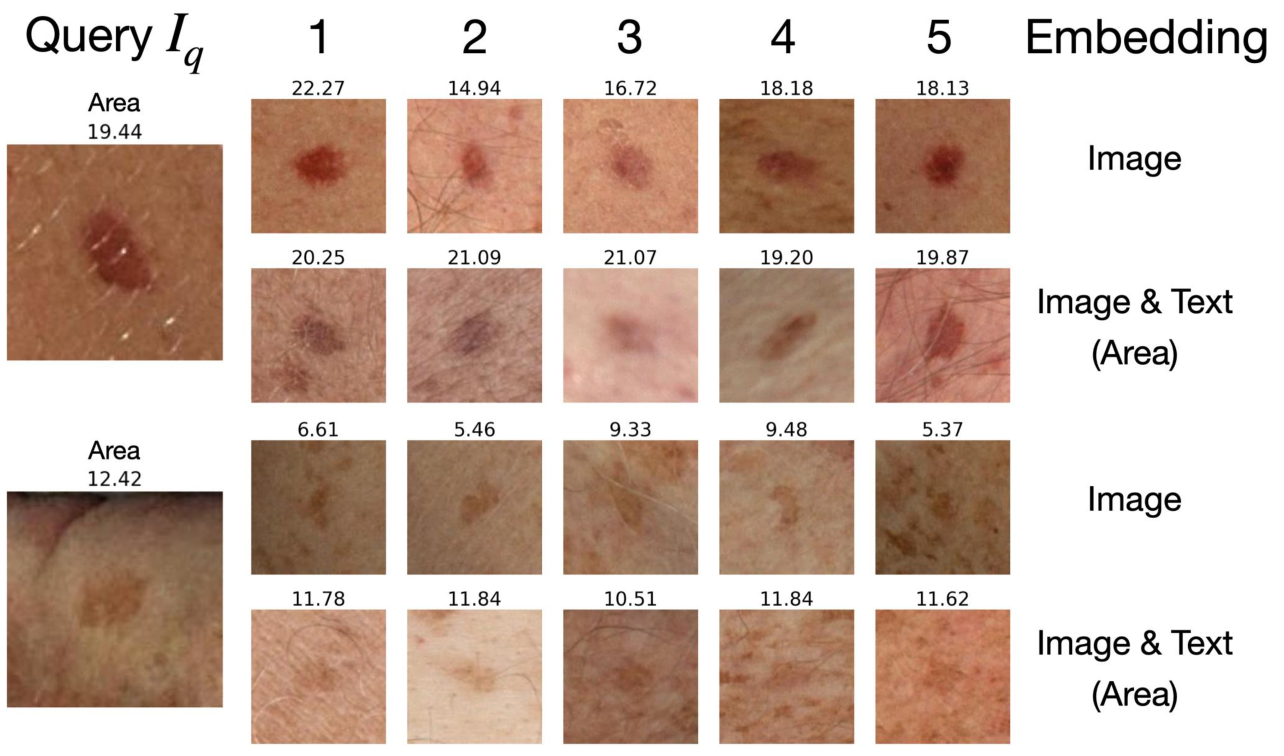
Median absolute attribute distance

Attribute / Method	Im & Text	Im.	Im. Untuned	MONET	PanDerm
areaMM2	**0.19 (0.26)**	0.29 (0.39)	0.64 (0.91)	0.55 (0.79)	0.52 (0.73)
norm_color	**0.48 (0.71)**	0.76 (1.07)	1.33 (1.75)	1.21 (1.61)	1.14 (1.53)
radial_color_std_max	**0.20 (0.29)**	0.28 (0.40)	0.45 (0.61)	0.42 (0.57)	0.40 (0.54)
deltaB	**0.32 (0.41)**	0.74 (0.94)	1.48 (1.92)	1.15 (1.50)	1.15 (1.49)
deltaL	**0.50 (0.67)**	0.66 (0.89)	1.70 (2.30)	1.34 (1.81)	1.30 (1.72)
deltaLB	**0.39 (0.53)**	0.68 (0.91)	1.71 (2.29)	1.33 (1.79)	1.29 (1.71)
stdLExt	**0.15 (0.20)**	0.23 (0.30)	0.40 (0.52)	0.33 (0.43)	0.33 (0.43)
clin_size_long_diam_mm	**0.25 (0.37)**	0.36 (0.55)	0.83 (1.41)	0.72 (1.21)	0.69 (1.14)
perimeterMM	**0.61 (0.89)**	1.05 (1.58)	2.70 (4.50)	2.34 (3.91)	2.19 (3.66)
norm_border	**0.50 (0.73)**	0.63 (0.93)	1.06 (1.71)	1.00 (1.64)	0.96 (1.61)
area_perim_ratio	**1.26 (2.07)**	1.55 (2.63)	2.87 (5.11)	2.63 (4.88)	2.54 (4.8)
A	**0.75 (1.11)**	1.27 (1.63)	3.10 (3.91)	2.53 (3.21)	2.43 (3.11)
Aext	**0.51 (0.83)**	1.20 (1.53)	2.83 (3.56)	2.19 (2.79)	2.11 (2.68)
B	**0.93 (1.49)**	1.65 (2.13)	4.13 (5.20)	3.08 (3.96)	3.10 (4.00)
Bext	**0.82 (1.40)**	1.67 (2.14)	3.72 (4.69)	2.75 (3.55)	2.76 (3.57)

Diagnostic Results

Score / Method	Im. (Ours)	Im. Untuned	MONET [7]	PanDerm [20]
AUROC	**0.928** (0.916 – 0.939)	0.881 (0.864 – 0.897)	0.890 (0.873 – 0.905)	0.909 (0.894 – 0.922)
pAUC	**0.143** (0.133 – 0.152)	0.112 (0.102 – 0.124)	0.122 (0.111 – 0.133)	0.131 (0.119 – 0.141)

Qualitative Search Results



This work is supported by National Cancer Institute under Grant **U24CA264369**, and Memorial Sloan Kettering Cancer Center institutional National Institutes of Health/National Cancer Institute Cancer Center Support Grant **P30 CA008748.**

References

[1] Marchetti et al, 3D Whole-body skin imaging for automated melanoma detection. JEADV 2023.

[2] Kurtansky et al, The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection. Scientific Data 2024

[3] Wang et al, Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. ArXiv 2024.

[4] Kim et al, Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine 2024.