10th International Skin Imaging Collaboration (ISIC) Workshop
on Skin Image Analysis @ MICCAI 2025

Retrieval-Augmented VLMs for Multimodal Melanoma Diagnosis

Jihyun Moon , Charmgil Hong

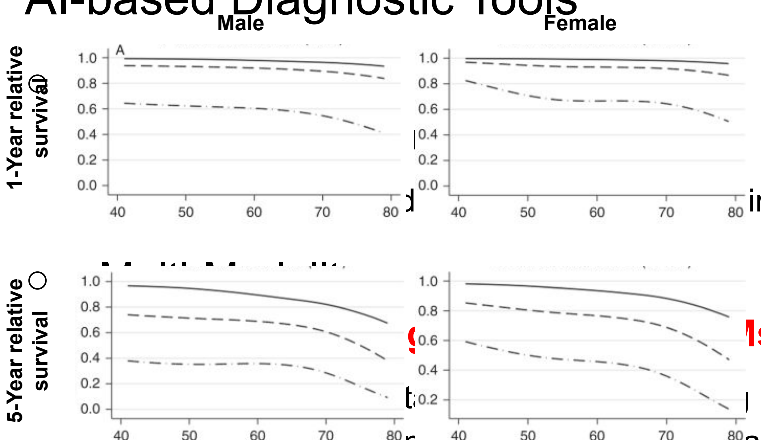
{**jhmoon**, charmgil}@handong.ac.kr

Handong Global University

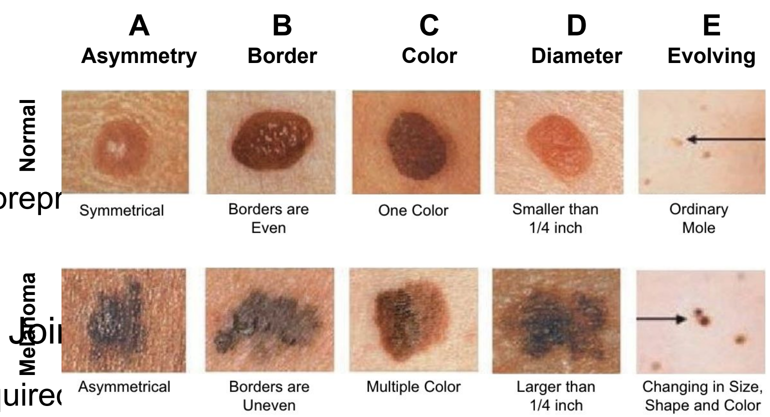
Background

- **Malignant Melanoma**
  - Early Detection Critical
    - 99% vs < 35% survival rate
  - Traditional Approach: ABCDE rule
    - Clinical **expertise**: **Pattern recognition** based on experience

- AI-based Diagnostic Tools



1- and 5-year relative survival curves as a function of age at diagnosis [Aiden et al., 2020]



The ABCDEs of Detection Melanoma [Alfghani, 2018]

Challenges

- **VLM**s for Medical Domain
  - General-purpose training: Lack medical domain specificity
  - Fine-tuning limitation: Resource intensive, privacy constraints, data variability

- **Example-based explainability** - Find similar cases to justify decision
  - More effective at decisions if they mimicked a dermatologist's experience
    - **AI**: Classification based on **content-based image retrieval**
    - **Human**: Compare with similar cases with structured analysis

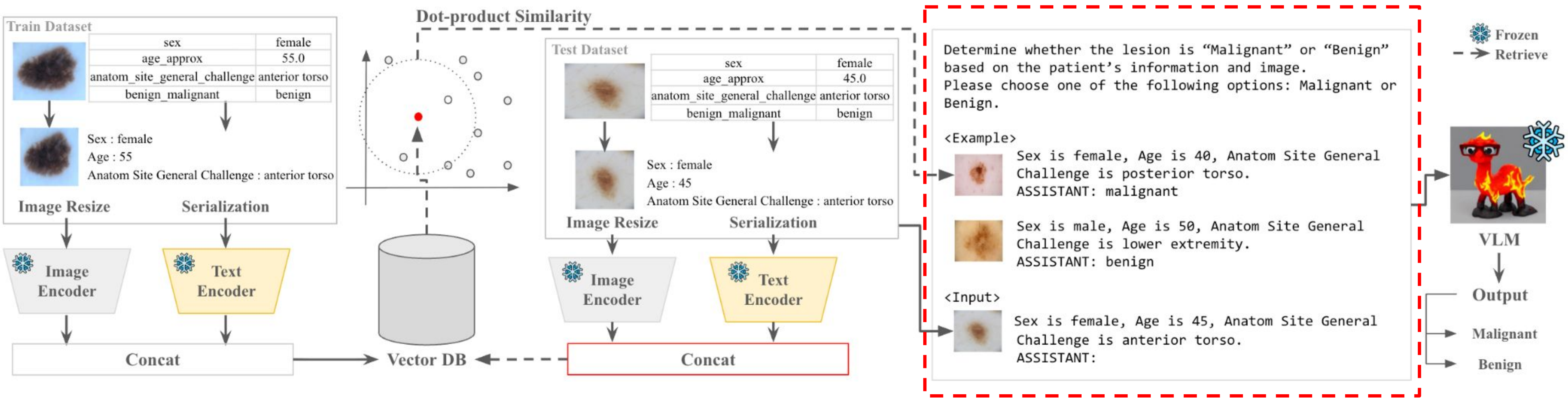
Motivation

- **Retrieval-Augmented Generation (RAG)** for Medical **Reasoning**
  - Clinical insight: Physicians compare new cases with similar historical cases
  - External knowledge: Incorporates relevant examples without fine-tuning
  - Medical application: Retrieve similar patient cases

- Retrieval-augmented VLM-based diagnostic framework
  - Can **VLM** be **effectively** used for dermoscopic image **classification**?
  - Does RAG **improve** performance through **example-based reasoning** without fine-tuning?

Proposed Approach

Multimodal Embedding and Case Indexing (*Indexing*)	**Semantically-Guided Retrieval** (*Retrieval*)	**Prompt Construction and VLM Inference** (*Generation*)



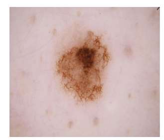
Proposed Approach

- **Prompt** Construction
  - **Task** Definition
    - Clear instruction to classify
  - Constrained **Output**
  - Contextual **Examples**
    - Infer the label
    - Top-*K* (*K*-shot) similar cases
  - **Target** query
    - Zero-shot cases

Determine whether the lesion is “Malignant” or “Benign” based on the patient’s information and image.

Please choose one of the following options: Malignant or Benign.

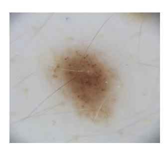
<Example>



Sex is female, Age is 40, Anatom Site General Challenge is posterior torso.
ASSISTANT: malignant

Sex is male, Age is 50, Anatom Site General Challenge is lower extremity.
ASSISTANT: benign

<Input>



Sex is female, Age is 45, Anatom Site General Challenge is anterior torso.
ASSISTANT:

Proposed Approach

- **Template**-Based Sentence Transformation
  - **3 serialization** strategies:
    1. **HTML**: Preserve tabular structure
    2. **Attribute-Value pair**: Reduce prompt length and improve parsing
    3. **Sentence**: VLMs training style

Attribute	Value
sex	female
age_approx	55.0
anatomic_site_general_challenge	anterior torso
benign_malignant	benign

Raw Clinical Metadata

```
<table>
 <tr>
  <th>Sex</th>
  <th>Age</th>
  <th>Anatomic Site General Challenge</th>
  <td>female</td>
  <td>55</td>
  <td>anterior torso</td>
 </tr>
</table>
```

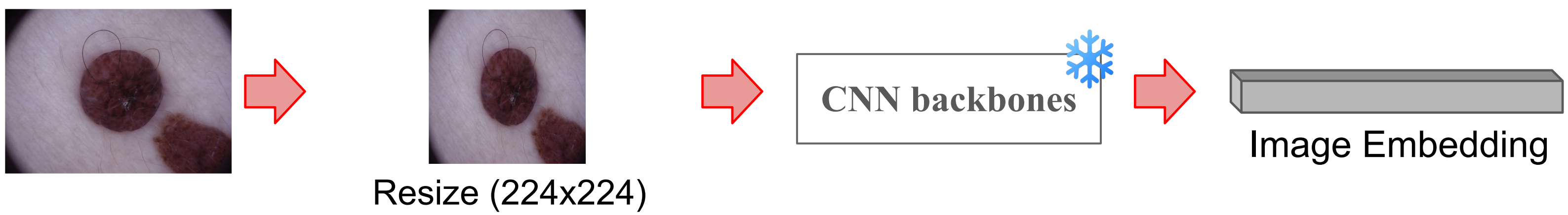
Sex: female,
Age: 55,
Anatomic Site General Challenge:
anterior torso

Sex is female, Age is 55,
Anatomic Site General Challenge is
anterior torso.

HTML / **Attribute-Value Pair** / **Sentence**

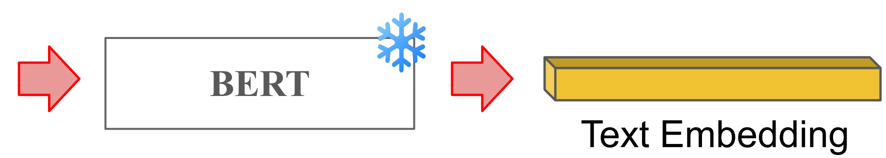
Proposed Approach

- **Multimodal Embedding** and Case Indexing
  - Use **modality-specific** encoders
    - **Image**: Resized to 224x224 and encoded using CNN backbones
      - ResNeXt-50, EfficientNet-V2-M
    - **Text**: Serialized into text and embedding using a pre-trained language model



Attribute	Value
sex	female
age_approx	55.0
anatomic_site_general_challenge	anterior torso
benign_malignant	benign

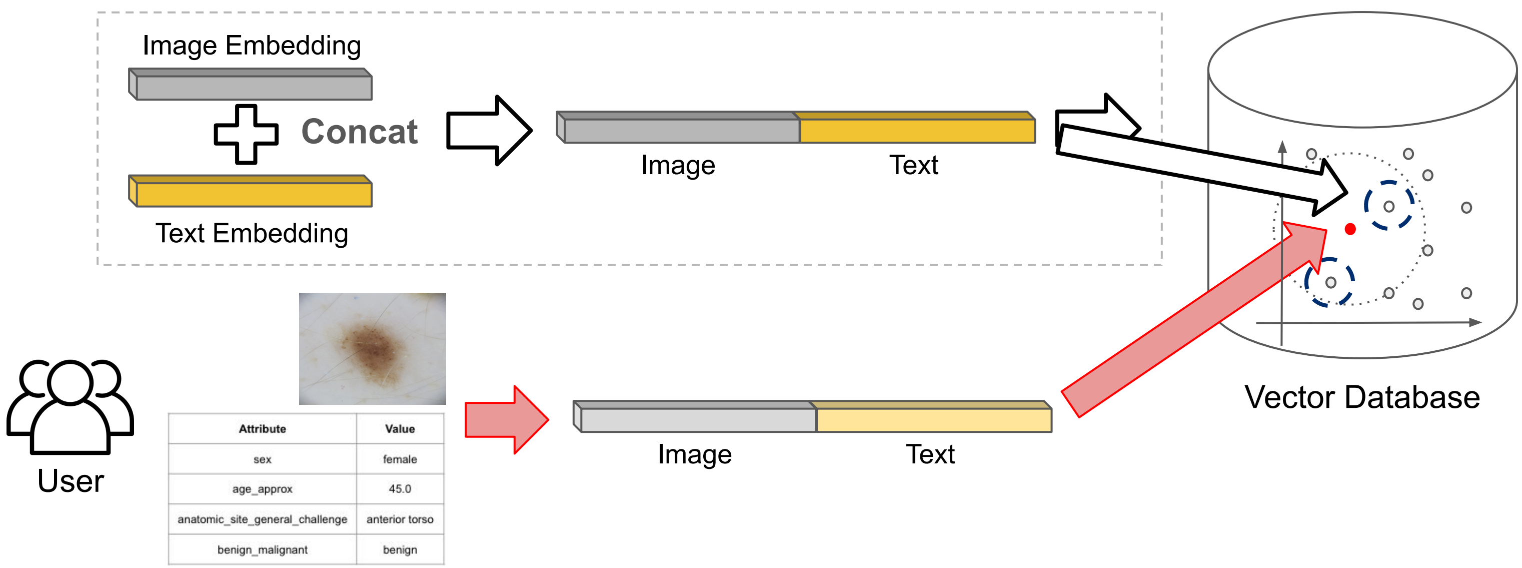
Sex is female,
Age is 55,
Anatomic Site General Challenge is anterior torso.



Proposed Approach

- Case **Indexing** and **Retrieval**
  - Stored in **FAISS-based** database
  - Similarity is computed using **dot-product**

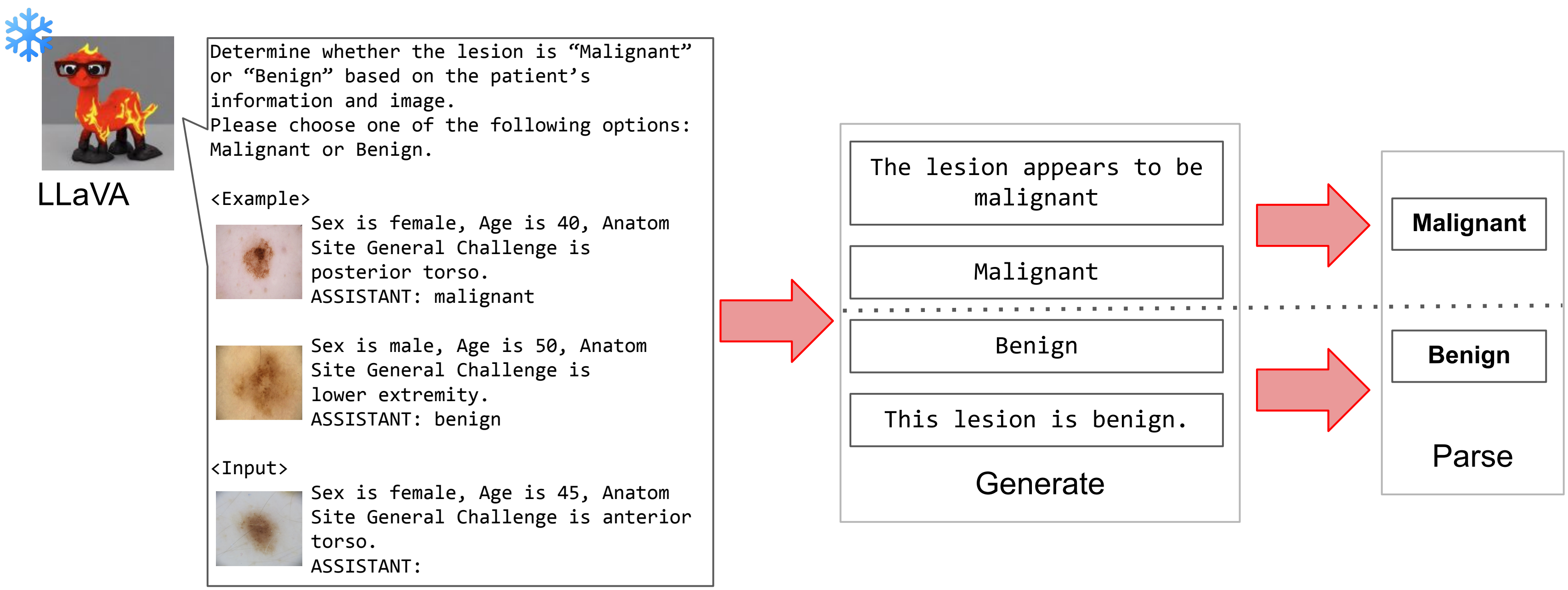
➡ **Top-*K* (*K*-shot)** most similar patient cases retrieved as contextual examples



Attribute	Value
sex	female
age_approx	45.0
anatomic_site_general_challenge	anterior torso
benign_malignant	benign

Proposed Approach

- **Classification** using VLMs
  - **Generate** diagnosis results in natural language text form
  - **Parse** to extract sentence containing the keywords “malignant” or “benign”



Experimental Setup

- ISIC 2019 dataset
  - **Binary** classification task: Malignant vs. Benign
  - Dermoscopic images with corresponding patient metadata (age, sex, anatomical site)
  - 70/30 split for train / test
- Evaluation metrics
  - Accuracy, Balanced accuracy, **F1 score**
- **Baselines**
  - **Image-based**: ResNeXt-50, EfficientNet-V2-M
  - **Text-based**: Random Forest, Vicuna-7B v1.5
  - **Multimodal early-fusion**: Classified via Random Forest, ReLU-activated FNN
  - **Zero-shot VLM**: LLaVA v1.5
- **Ours**
  - **Training data** (16,756 image–text pairs) indexed with FAISS
  - Retrieve **Top-2** neighbors (*K* = 1,**2**,3,4)

Results

- Can **VLM** be **effectively** used for dermoscopic image **classification**?
  - Single-Modality Limitations: Image-based and text-based achieve **< 30%** F1 score
    - Multimodal advantage: **42.7%** improvement
  - VLM advantage: **74.9%** improvement
  - RAG effectiveness:
    - Achieve **44%** improvement over best baseline
    - **1.8x** better than zero-shot VLM

	Modality					Balanced	
	Image	Metadata	Model	Serialization	Accuracy	Accuracy	F1 score
Image-based	✓	-	EfficientNet-V2-M	-	0.6954	0.5061	0.2001
Text-based	-	✓	Vicuna 7B v1.5	Sentence	0.6063	0.5152	0.2613
Multimoal Early-Fusion	✓	✓	BERT + ResNeXT-50 + FNN	HTML	0.6819	0.5079	0.2132
Zero-Shot VLM	✓	✓	LLaVA 7B v1.5 hf	Attribute-Value pair	0.7126	0.6128	0.3729
Ours (K = 2)	✓	✓	BERT + ResNeXt-50 + LLaVA 7B v 1.5 hf	Attribute-Value pair	0.8876	0.797	0.6864

Baselines

+ 42.7%

+ 74.9%

+ 44%

Results

- Does RAG **improve** performance through **example-based reasoning** without fine-tuning?
    - Similar **lesions** with corresponding patient's **metadata**

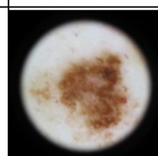
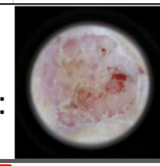
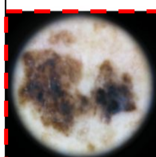
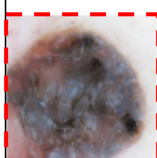
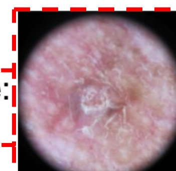
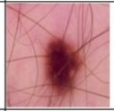
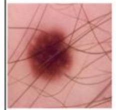
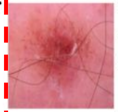
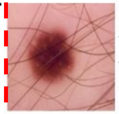
	Ground Truth	Malignant		Benign	
	Input		Sex: male Age: 75.0 Anatom Site General Challenge: anterior torso		Sex: female Age: 85.0 Anatom Site General Challenge: anterior torso
Retrieved Similar Cases	Ours at $K = 1$		Sex: male Age: 75.0 Anatom Site General Challenge: anterior torso ASSISTANT: malignant		Sex: female Age: 85.0 Anatom Site General Challenge: anterior torso ASSISTANT: benign
	Ours at $K = 2$		Sex: female Age: 65.0 Anatom Site General Challenge: anterior torso ASSISTANT: malignant		Sex: male Age: 70.0 Anatom Site General Challenge: anterior torso ASSISTANT: benign

Fig 2. (a) Misclassified case by all baselines (LLM, early-fusion, zero-shot VLM) correctly classified by our method (K = 2).

Results

- Effect of Input Serialization
  - **Structured** metadata encoding enhances VLM's clinical understanding

Ground Truth	Benign		Malignant	
Serialization	HTML	Attribute-value pair	Sentence	Attribute-value pair
Input	 <table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>male</td><td>5.0</td><td> lower extremity</td></tr></table>	 Sex: male Age: 5.0 Anatom Site General Challenge: lower extremity	 Sex is male, Age is 40.0, Anatom Site General Challenge is upper extremity.	 Sex: male Age: 40.0 Anatom Site General Challenge: upper extremity
Prediction	Malignant	Benign	Benign	Malignant
Ours at K =1	 <table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>male</td><td>5.0</td><td> anterior torso</td></tr></table> ASSISTANT: benign	 Sex: male Age: 35.0 Anatom Site General Challenge: lower extremity ASSISTANT: benign	 Sex is male, Age is 45.0, Anatom Site General Challenge is head/neck. ASSISTANT: benign	 Sex: male Age: 55.0 Anatom Site General Challenge: anterior torso ASSISTANT: benign
Ours at K = 2	 <table><tr><th>Sex</th><th>Age</th><th> Anatom Site General Challenge</th></tr> <tr><td>female</td><td>55.0</td><td> anterior torso</td></tr></table> ASSISTANT: benign	 Sex: male Age: 5.0 Anatom Site General Challenge: anterior torso ASSISTANT: benign	 Sex is male, Age is 55.0, Anatom Site General Challenge is anterior torso. ASSISTANT: benign	 Sex: male Age: 40.0 Anatom Site General Challenge: upper extremity ASSISTANT: malignant

Conclusion

- Proposed a **retrieval-augmented VLM framework** to improve melanoma classification using retrieved similar cases
- Provide **example-based explanations** via retrieved similar cases
- Achieve improved diagnostic performance without fine-tuning
- Future work
  - Extend to multi-class skin lesion classification and other multimodal clinical tasks
- Limitations
  - Depends on curated training data
  - Retrieval speed needs improvement for real-time use

Thank you for your attention

Title	**Retrieval-Augmented VLMs for Multimodal Melanoma Diagnosis**

Presenter Jihyun Moon (jhmoon@handong.ac.kr)
Advisor Charmgil Hong (charmgil@handong.ac.kr)