VLEER: Vision and Language Embeddings for Explainable Whole Slide Image Representation

Sep. 27th, 2025

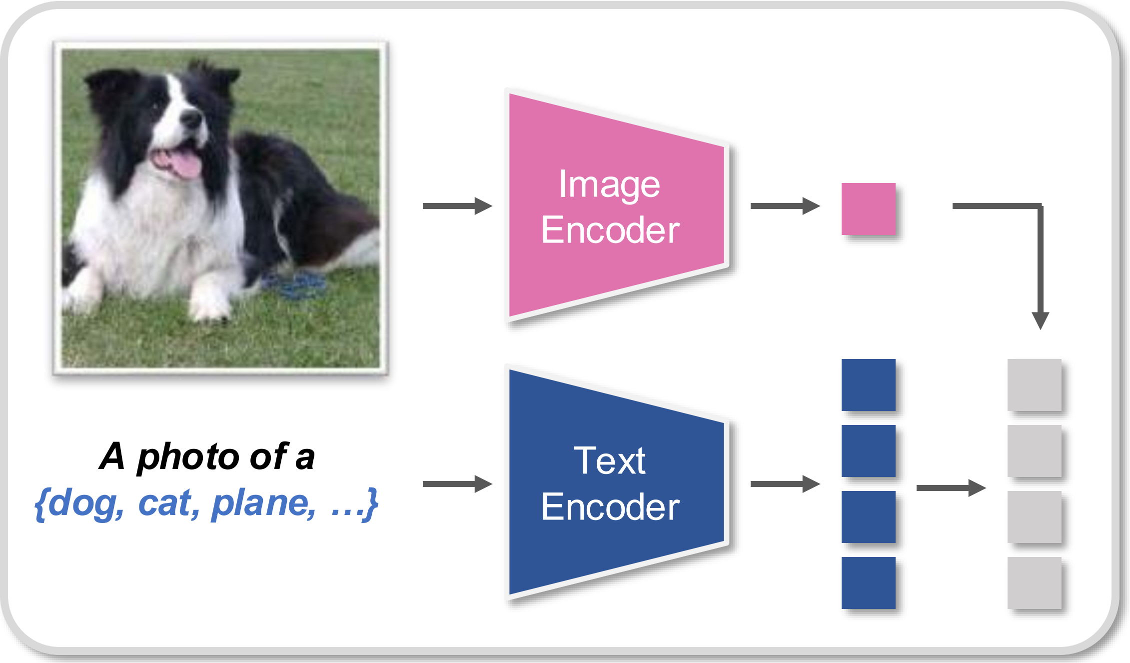
Anh Tien Nguyen[1], **Keunho Byeon**[1], Kyungeun Kim[2], and Jin Tae Kwak[1]

[1] School of Electrical Engineering, Korea University, Seoul, South Korea
[2] Seegene Medical Foundation, Seoul, South Korea

1 Introduction

❑ Vision-Language Model

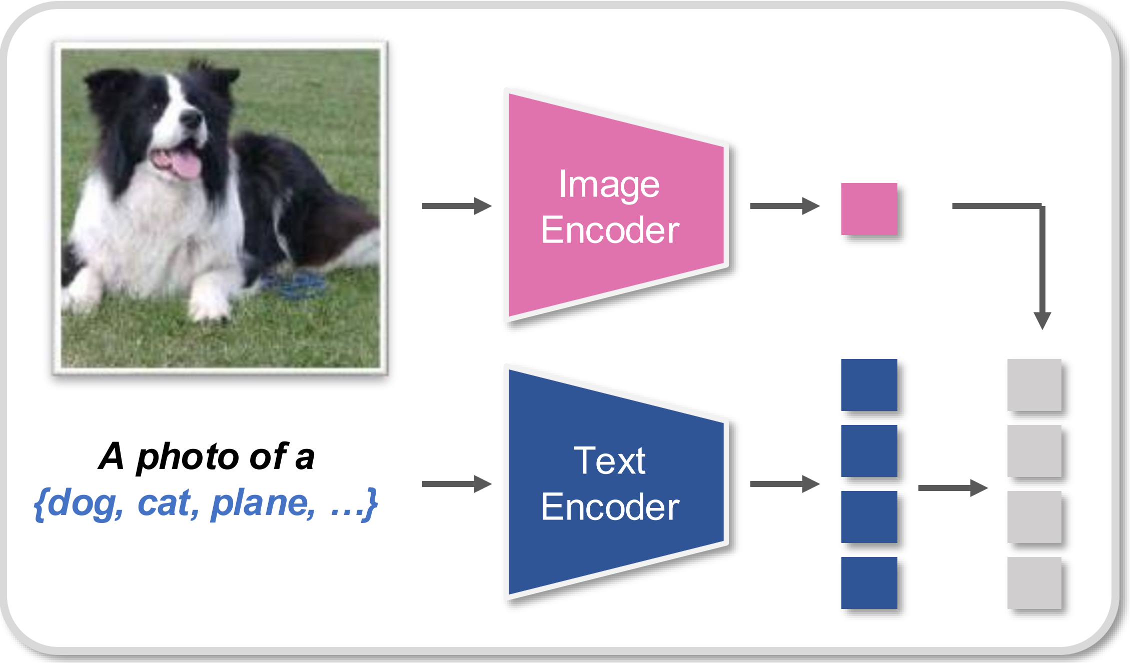
- There has been growing interest in vision-language models (VLMs), which **integrate vision and language** modalities by jointly learning from image-text datasets.



1 Introduction

❑ Vision-Language Model

- There has been growing interest in vision-language models (VLMs), which **integrate vision and language** modalities by jointly learning from image-text datasets.



CLIP

Radford et al. (2021)

CoCa

Yu et al. (2022)

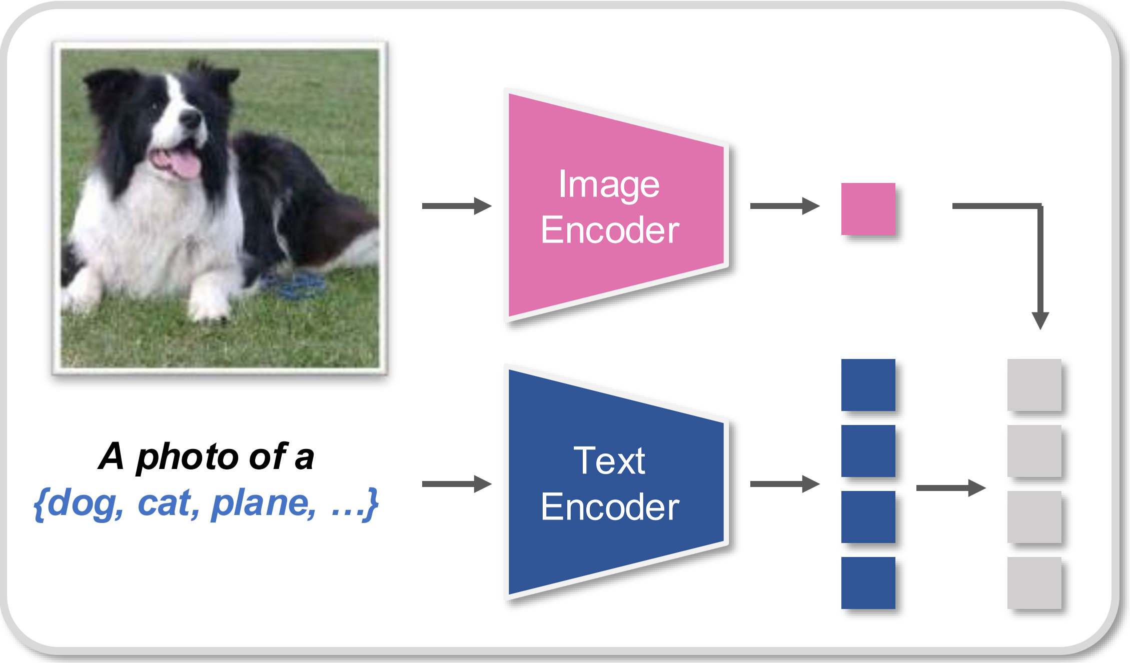
FLAVA

Singh et al. (2022)

1 Introduction

❑ Vision-Language Models in Pathology

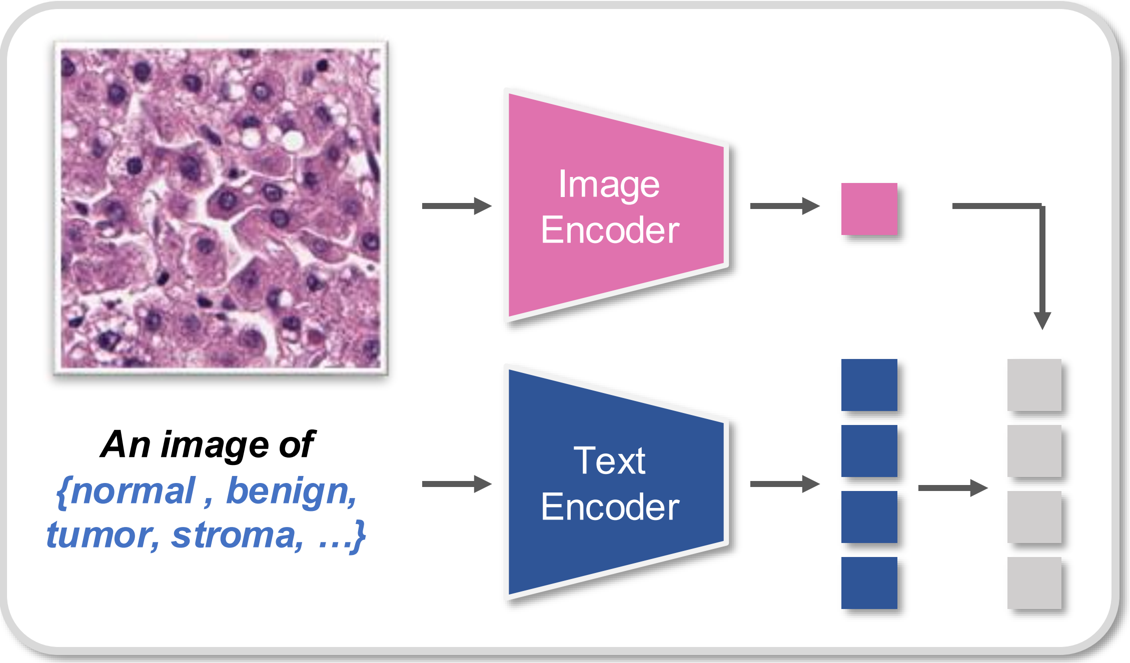
- In computational pathology, this approach has been adapted to domain-specific datasets, resulting in **pathology VLMs**



1 Introduction

❑ Vision-Language Models in Pathology

- In computational pathology, this approach has been adapted to domain-specific datasets, resulting in **pathology VLMs**



PLIP

Huang et al. (2023)

QUILT-NET

Ikezogwo et al. (2023)

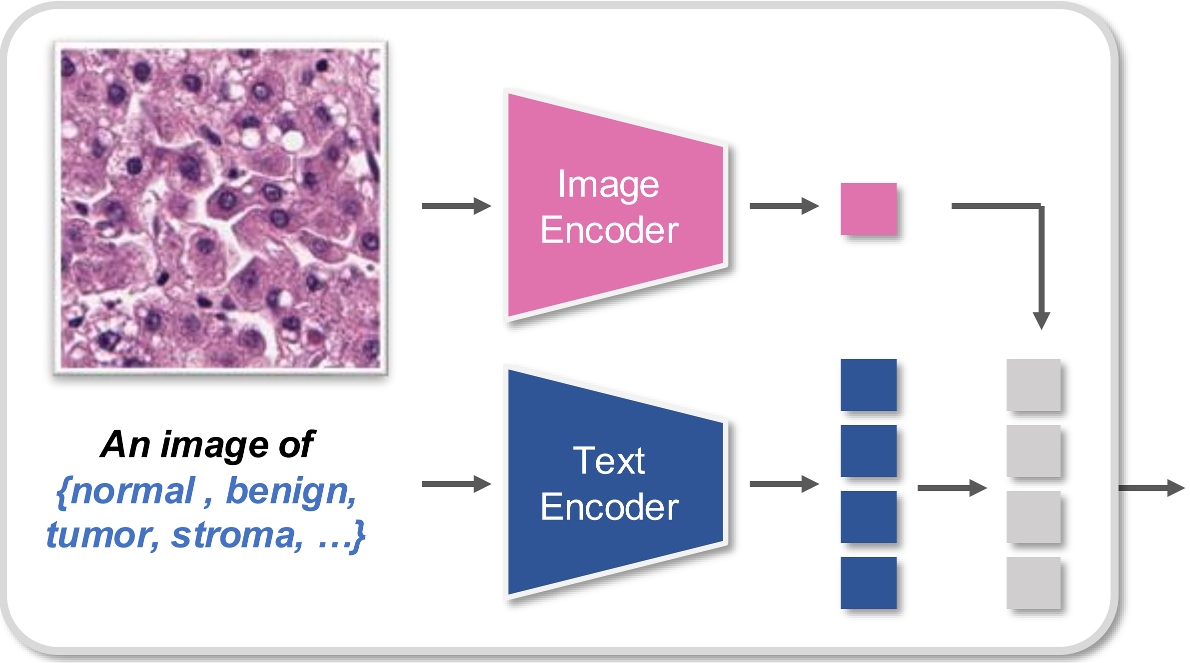
CONCH

Lu et al. (2024)

1 Introduction

❑ Vision-Language Models in Pathology

- They have achieved remarkable results in various classification tasks, often **without** requiring **further training** or **fine-tuning**.



PLIP

Huang et al. (2023)

QUILT-NET

Ikezogwo et al. (2023)

CONCH

Lu et al. (2024)

Various Classification Tasks

Tissue phenotyping

Gleason grading

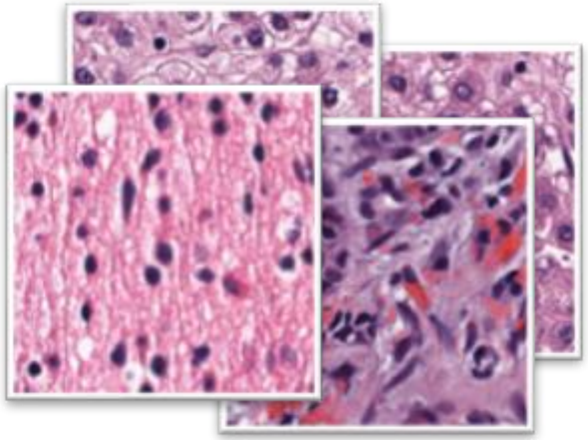
Lymph node metastasis detection

…

1 Introduction

❑ Motivation

- Most studies have primarily focused on pre-training VLMs and their direct application to downstream tasks, overlooking **two key limitations**.



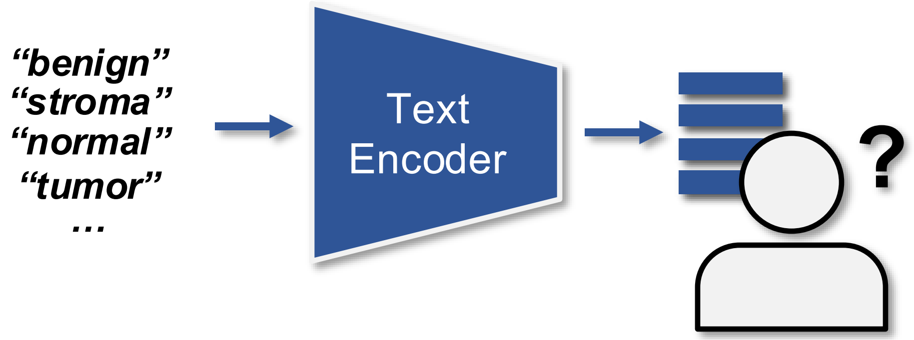
Most prior works mainly focus on **patch-level tasks**,
while **WSI-level** applications remain largely unexplored.

1

Introduction

❑ Motivation

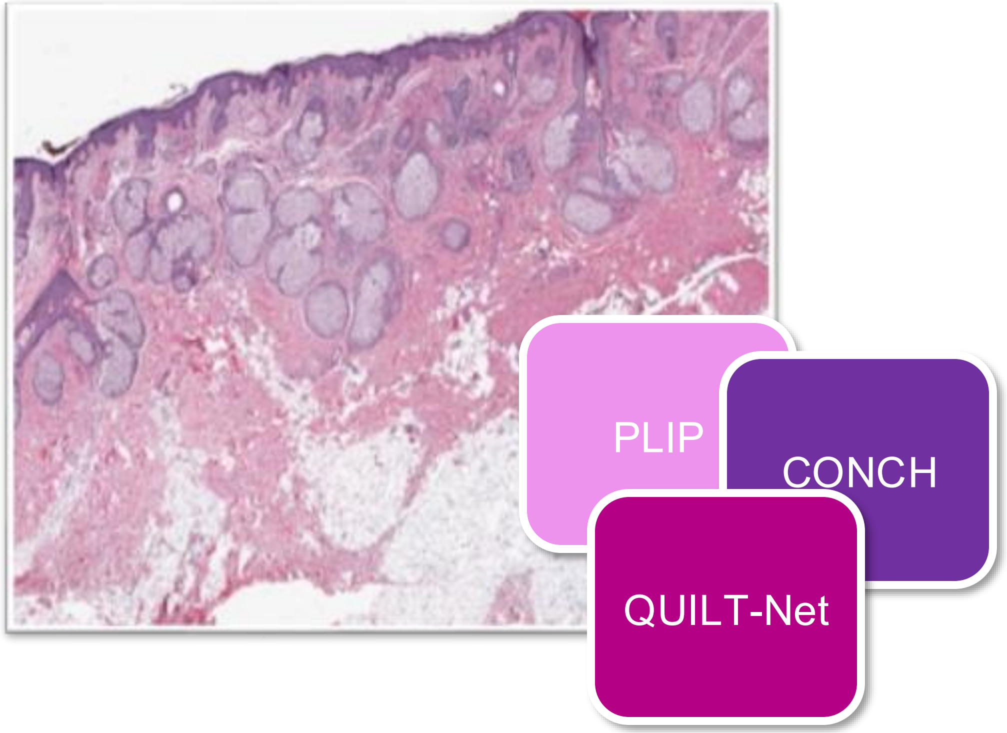
- Most studies have primarily focused on pre-training VLMs and their direct application to downstream tasks, overlooking **two key limitations**.



The **interpretability of textual embeddings** in VLM has not been thoroughly explored.

❑ Our Approach

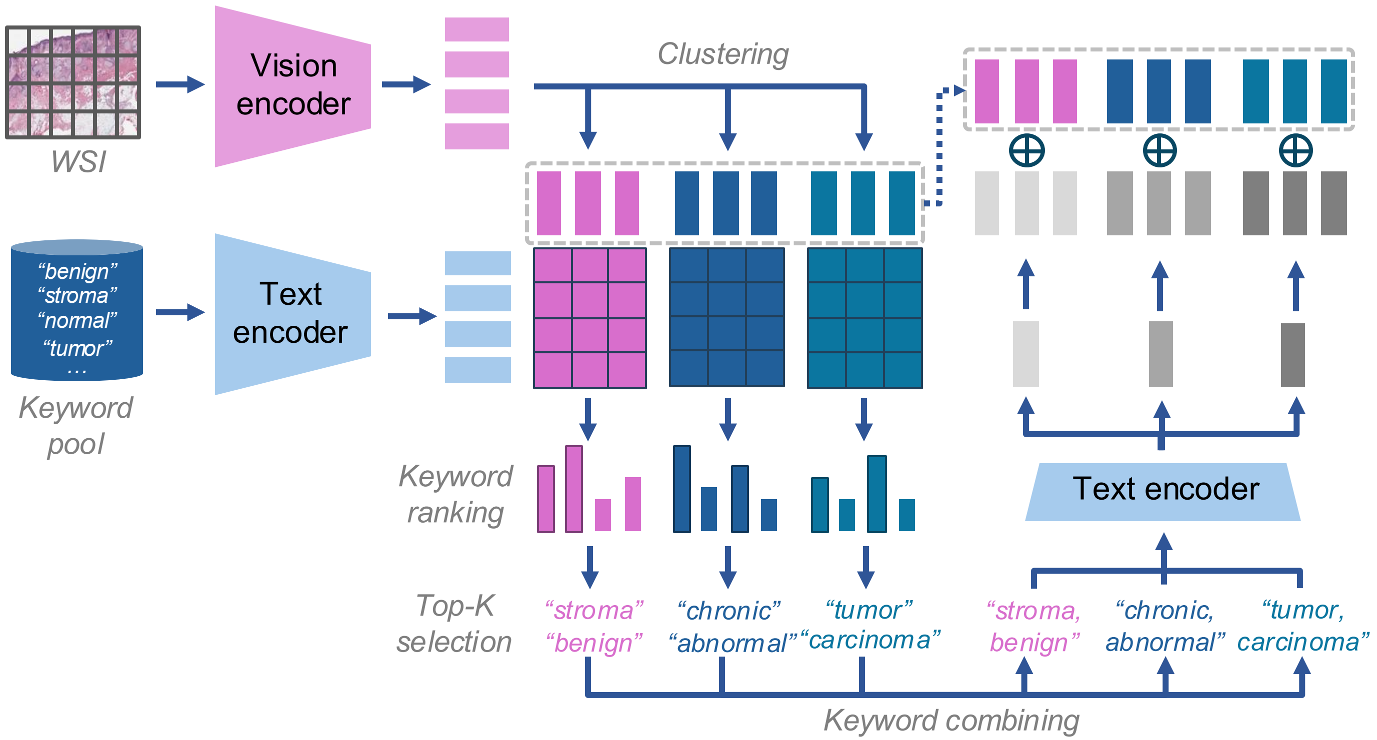
- We hypothesize that **pre-trained VLMs can inherently represent WSIs** in a quantitative and interpretable manner.



Introduction

❑ Our Approach

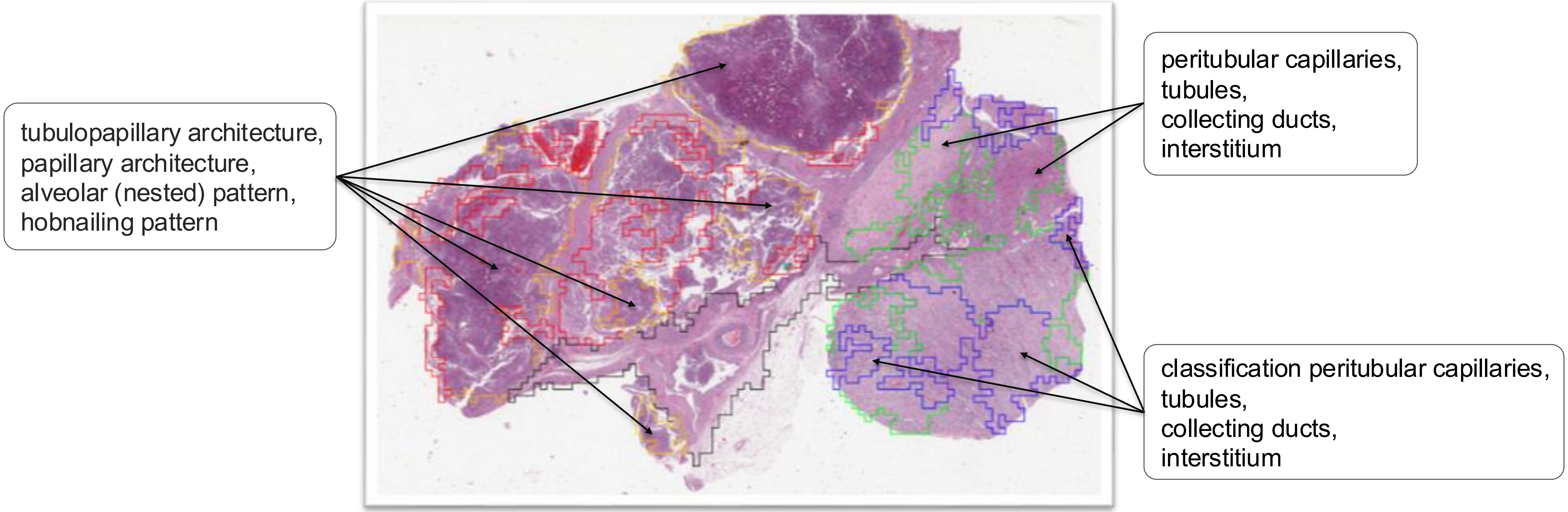
- We introduce **Vision and Language Embeddings for Explainable WSI Representation (*VLEER*)**.



1 Introduction

❑ Our Approach

- *VLEER* facilitates direct interpretation of results through **human-readable** and **understandable** textual representations.



2

Methodology

❑ Overall,

- *VLEER* utilizes two components to learn explainable WSI embeddings:

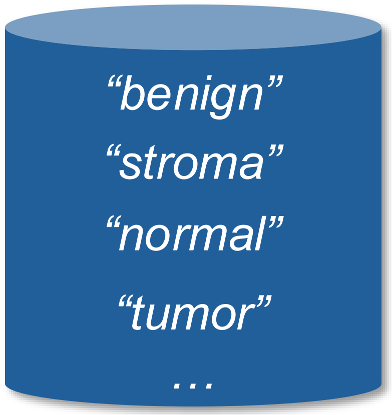
  *a task-related **text pool** of pathology keywords*

  *a pre-trained **pathology VLM***

2 Methodology

❑ Task-related pathology text pool

- We collect **task-specific keywords** illustrating the histology of tissues for each task.

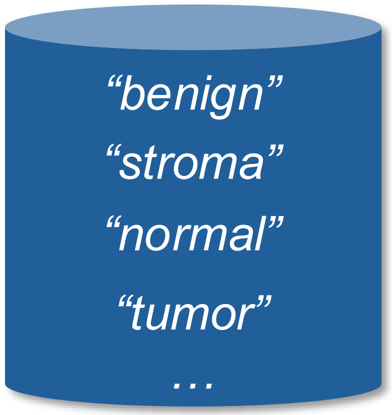


Keyword pool

2 Methodology

❑ Task-related pathology text pool

- These keywords include pathological terms that are relevant to **both normal and abnormal** conditions.
- All collected keywords are then **reviewed and validated** by a board-certified, experienced pathologist.

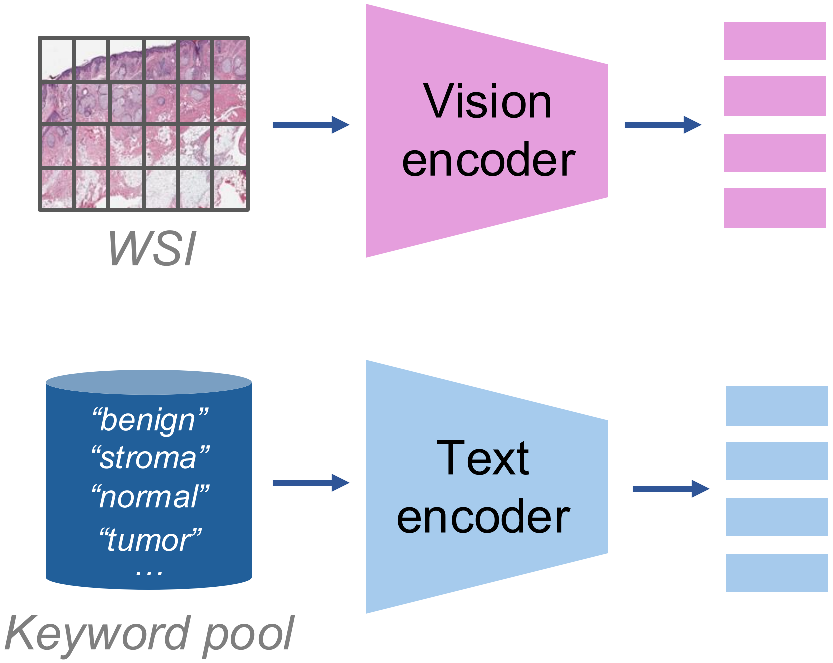


Keyword pool

2 Methodology

❑ Textual and visual embedding extraction

- WSI is tiled into a bag of patches, the **vision encoder** transforms these patches into **vision embeddings**.
- We adopt the **text encoder** to embed all keywords in the pool into **textual embeddings**.



2 Methodology

❑ Textual and visual embedding extraction

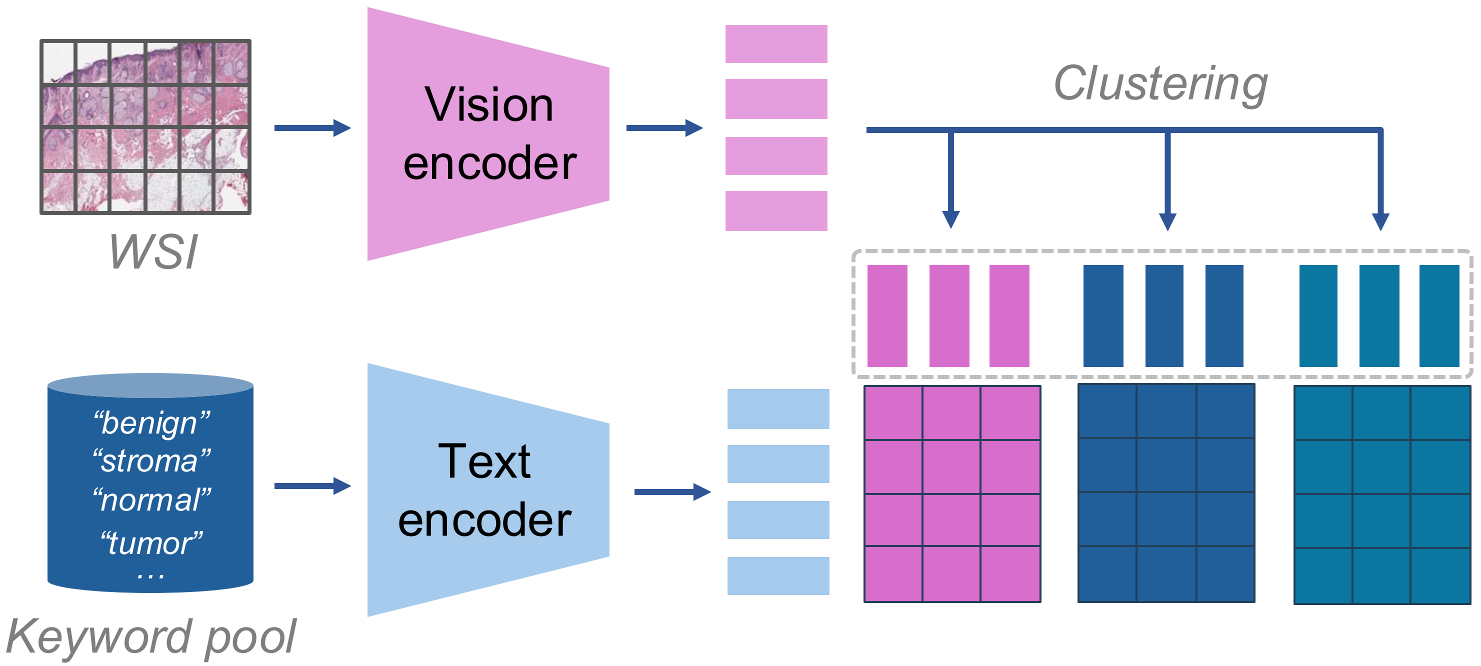
- We employ **various templates** to generate diverse text prompts for each keyword.

an image of CLASSNAME. *an image showing CLASSNAME.* *an example of CLASSNAME.* *a histopathological image showing CLASSNAME.* *a histopathological image of CLASSNAME.* *CLASSNAME is shown.* *this is CLASSNAME.* *there is CLASSNAME.* *a histopathological photograph of CLASSNAME.* *a histopathological photograph showing CLASSNAME.* …

Lu, Ming Y., et al. "A visual-language foundation model for computational pathology." Nature medicine 30.3 (2024): 863-874.

Vision-Text alignment

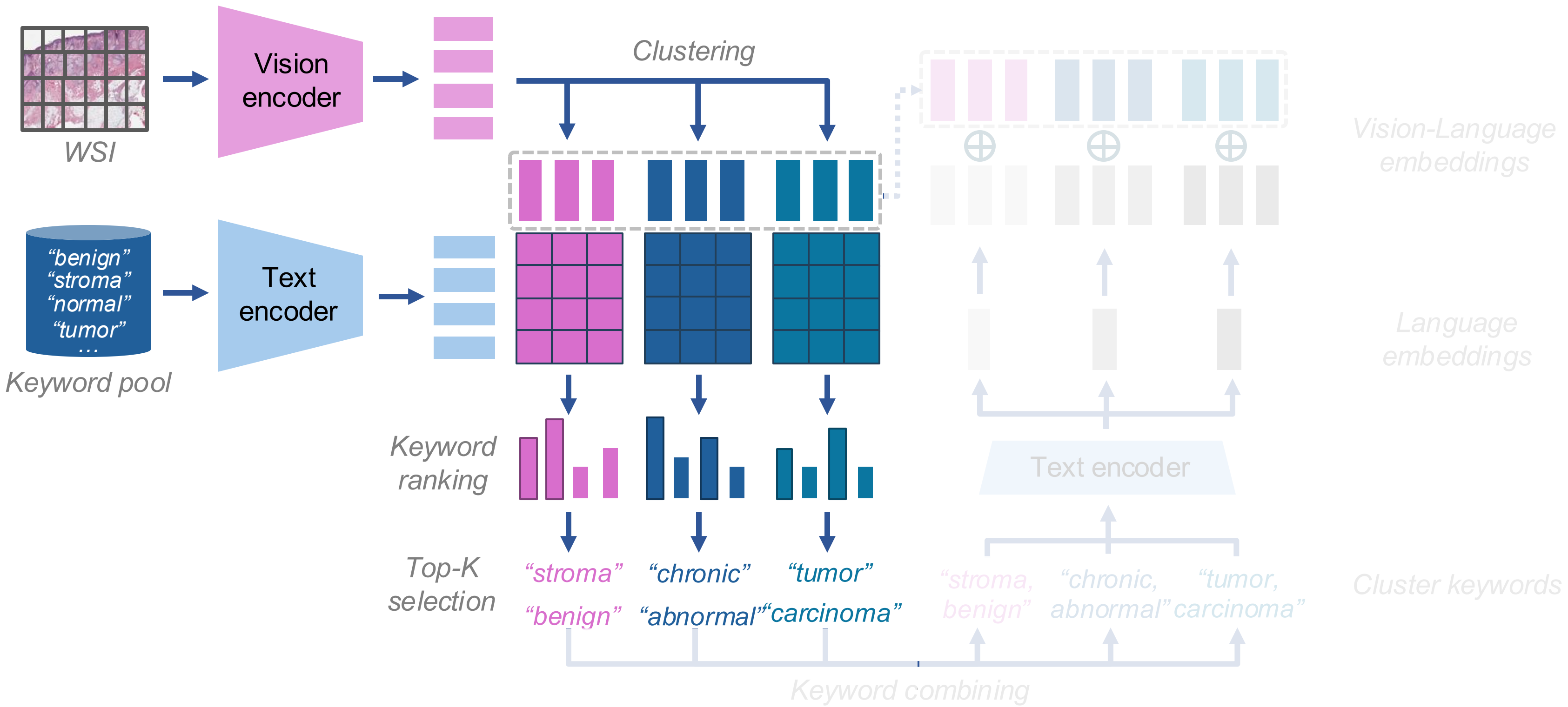
- We **cluster patches** into distinct groups to improve the semantic meaning.
- For each cluster, **similarity scores** are calculated between all visual and textual embeddings.



2 Methodology

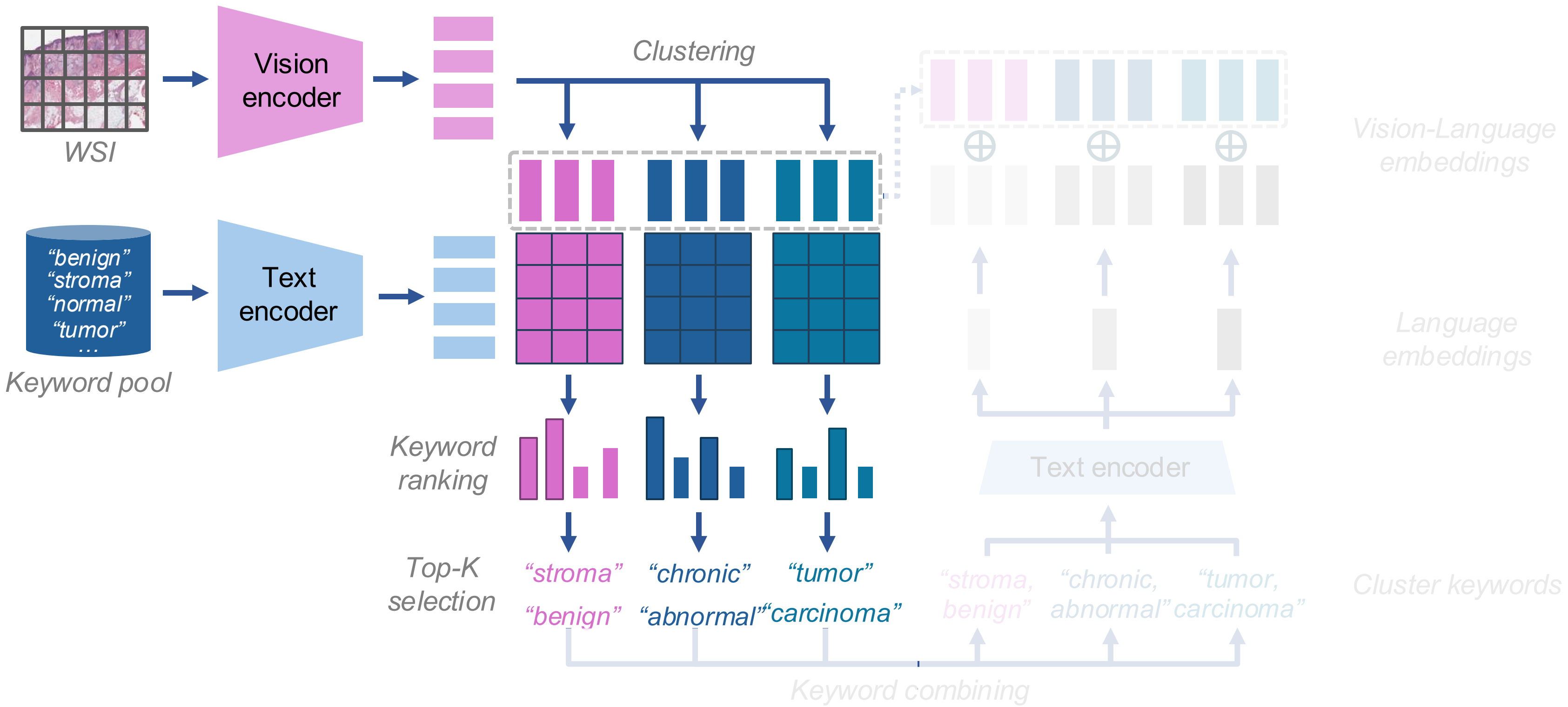
❑ Cluster representative keywords retrieval

- We **rank all keywords** based on their similarity scores with each patch image, **aggregate** these rankings and retrieve the **most representative keywords** for each cluster.



❑ Language embedding generation

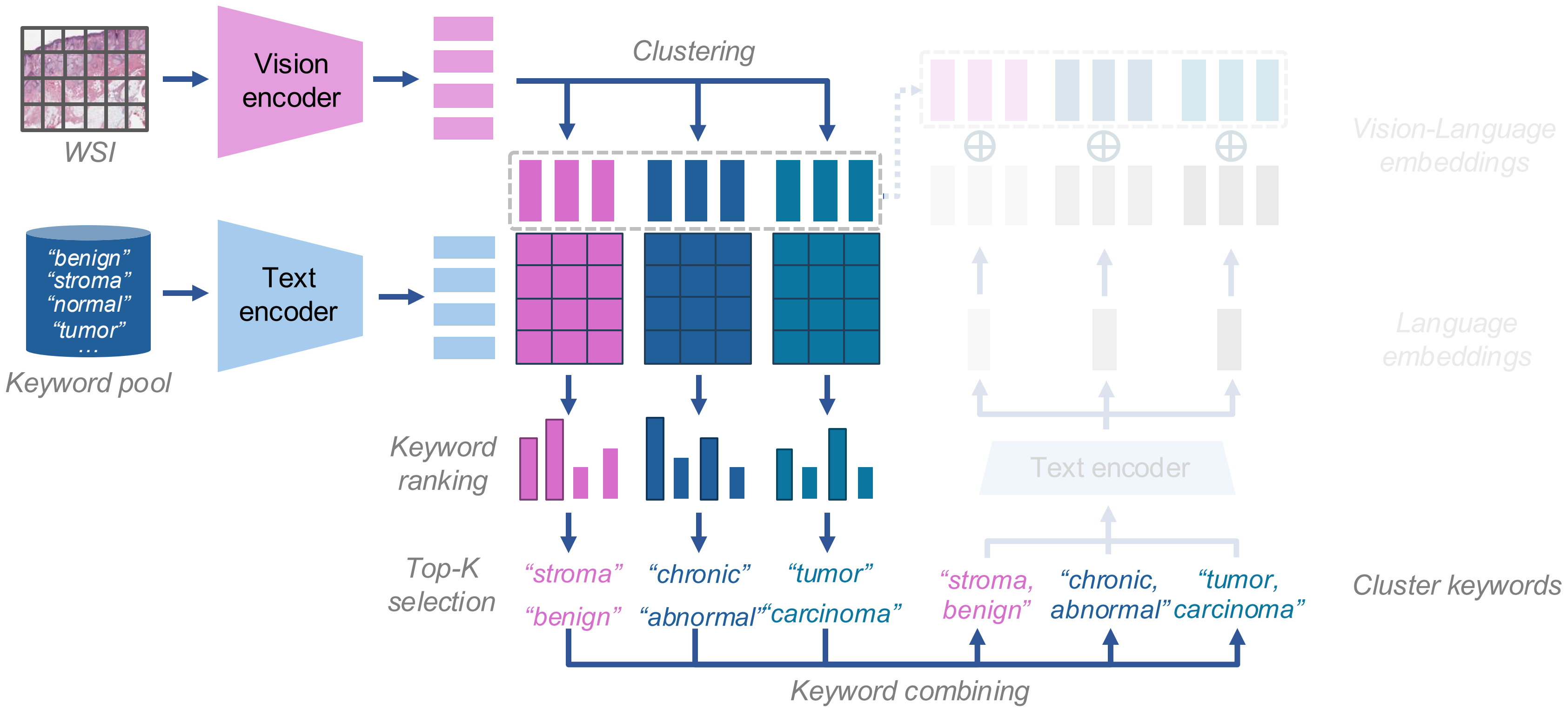
- The representative **keywords are concatenated** by commas



2 Methodology

❑ Language embedding generation

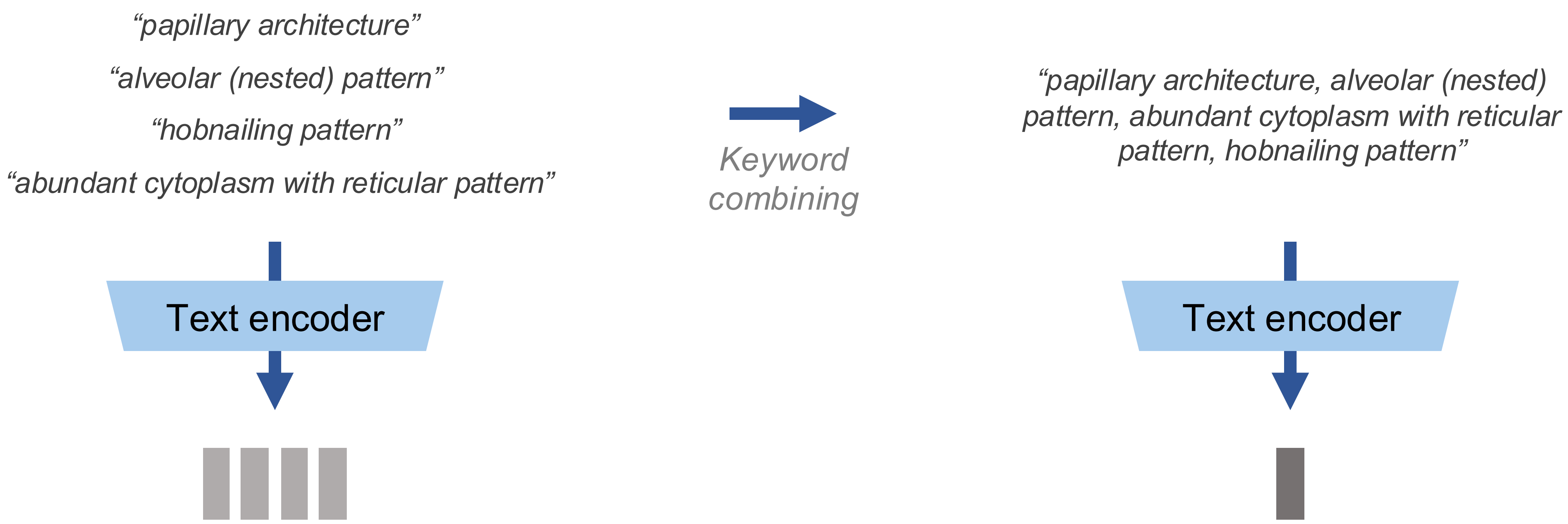
- and forwarded through the text encoder to obtain the **cluster-level language embeddings**.



2 Methodology

❑ Language embedding generation

- With combined keywords, the model can understand their **contextual relationships**.

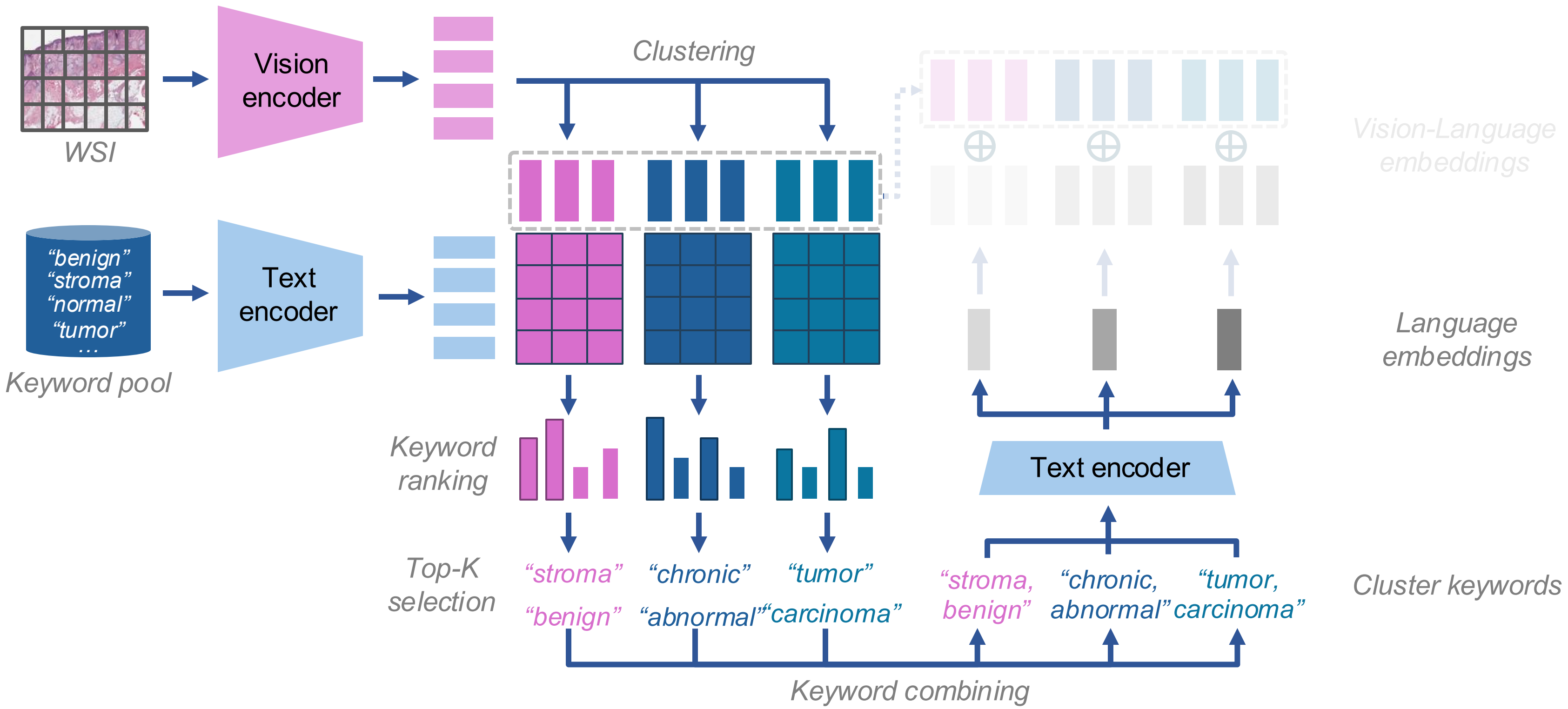


2

Methodology

❑ Vision-language embedding generation

- **Language** embeddings are then **concatenated** with the corresponding **vision** embeddings.
- These embeddings are then aggregated into a WSI-level embedding using a trainable MIL aggregator.



3 Experiments

❑ MIL aggregators

- We compare vision-only and vision-language embeddings using **four MIL aggregators**,

  - *ABMIL, CLAM-SB, CLAM-MB, and TransMIL.*

❑ Datasets

- **Three public TCGA datasets** are used for evaluation.

  - *TCGA-NSCLC: Lung cancer subtyping*

  - *TCGA-RCC: Renal cell carcinoma subtyping*

  - *TCGA-BRCA: Breast invasive carcinoma subtyping*

Experiments

❑ For qualitative analysis,

- We generate **heatmaps** using the normalized attention scores from the MIL aggregator.
- Following clustering, adjacent patches within the same cluster are merged into a region of interest (RoI).
- Each RoI is annotated with the **representative keywords**,

  which is **re**gion-specific and is generated using **V**ision-**L**anguage embeddings (**ReVL annotation**).

4 Results

❑ Quantitative evaluation

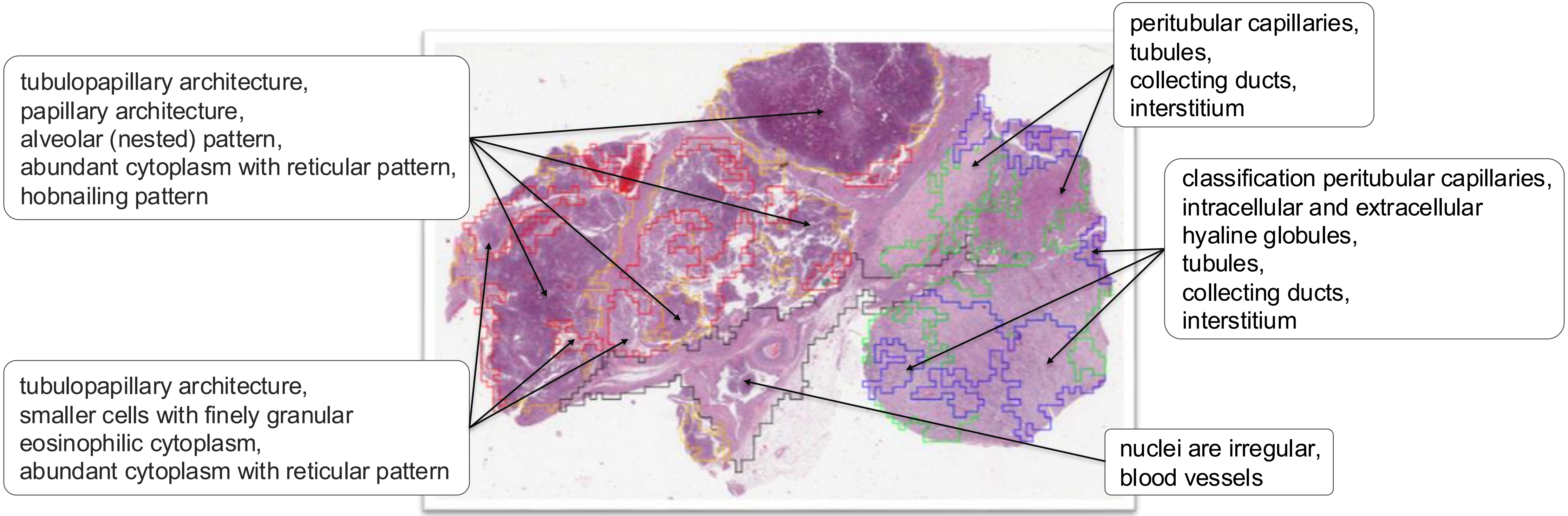
Aggregator	Emb.	TCGA-NSCLC			TCGA-RCC			TCGA-BRCA			Average		
		Acc	*F1*	*AUC*	*Acc*	*F1*	*AUC*	*Acc*	*F1*	*AUC*	*Acc*	*F1*	*AUC*
ABMIL	V	**0.9035**	**0.9033**	**0.9739**	**0.9425**	**0.9338**	0.9949	0.9313	0.9005	0.9707	0.9257	0.9126	0.9798
	V-L	0.8989	0.8988	0.9679	0.9310	0.9205	**0.9961**	**0.9479**	**0.9225**	**0.9762**	**0.9259**	**0.9139**	**0.9800**
CLAM-SB	V	0.9012	0.9011	**0.9715**	**0.9402**	**0.9282**	0.9951	0.9271	0.8915	0.9649	0.9228	0.9069	0.9771
	V-L	**0.9058**	**0.9056**	0.9696	0.9333	0.9181	**0.9954**	**0.9479**	**0.9196**	**0.9770**	**0.9290**	**0.9144**	**0.9806**
CLAM-MB	V	0.8989	0.8988	0.9691	**0.9333**	**0.9191**	0.9949	0.9292	0.8949	0.9650	0.9205	0.9043	0.9764
	V-L	**0.9103**	**0.9103**	**0.9699**	0.9287	0.9131	**0.9964**	**0.9479**	**0.9196**	**0.9780**	**0.9290**	**0.9143**	**0.9814**
TransMIL	V	0.8736	0.8729	0.9603	0.9333	**0.9247**	0.9882	**0.9146**	**0.8517**	**0.9741**	0.9072	0.8831	0.9759
	V-L	**0.8920**	**0.8918**	**0.9688**	**0.9379**	0.9236	**0.9903**	0.9125	0.8479	0.9728	**0.9141**	**0.8878**	**0.9773**

- On average, vision-language embeddings consistently achieved **higher performance** than vision-only embeddings across all evaluation metrics, aggregators, and datasets.

Results

❑ Qualitative evaluation of *VLEER*

- The ReVL annotations of a papillary renal cell carcinoma in TCGA-RCC

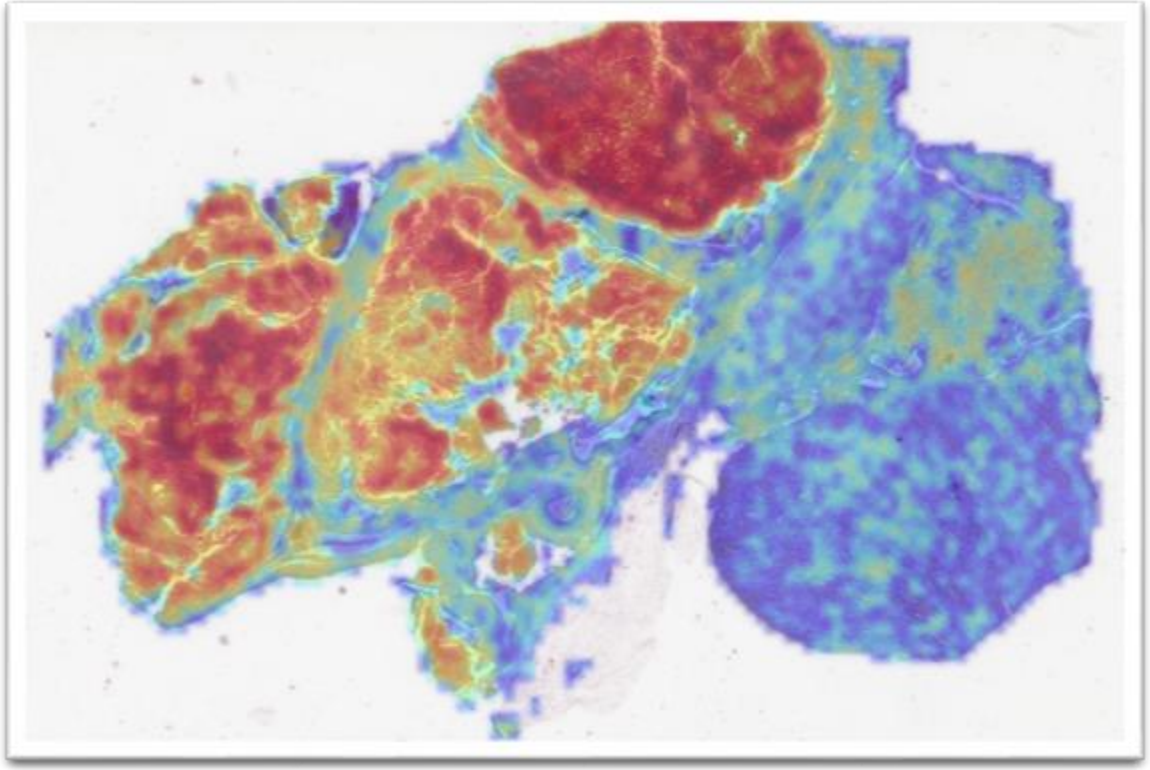


ReVL annotation
(a papillary renal cell carcinoma in TCGA-RCC)

4 Results

❑ Qualitative evaluation of *VLEER*

- The highly attended regions (red) in the heatmap are closely related to the patterns of **papillary cancer**, whereas the low attended regions (green and blue) are **normal histology of renal tissues**.

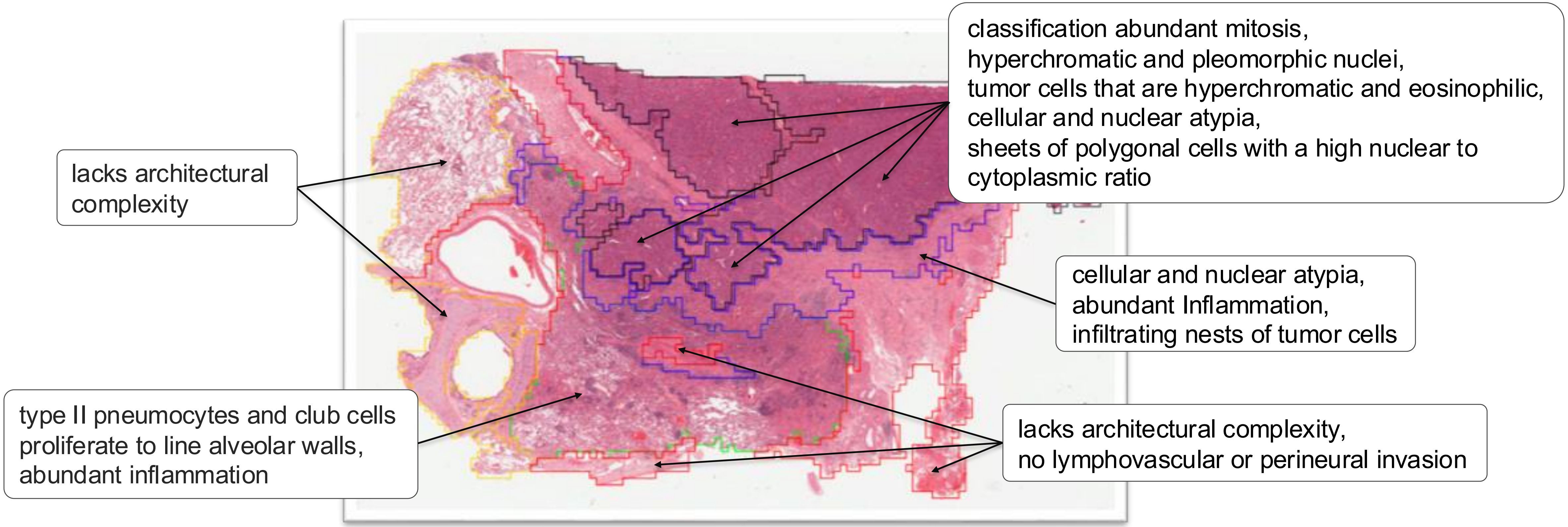


Attention heatmap
(a papillary renal cell carcinoma in TCGA-RCC)

4 Results

❑ Qualitative evaluation of *VLEER*

- *VLEER* enhances transparency by providing text-based justifications that align with established pathology knowledge.

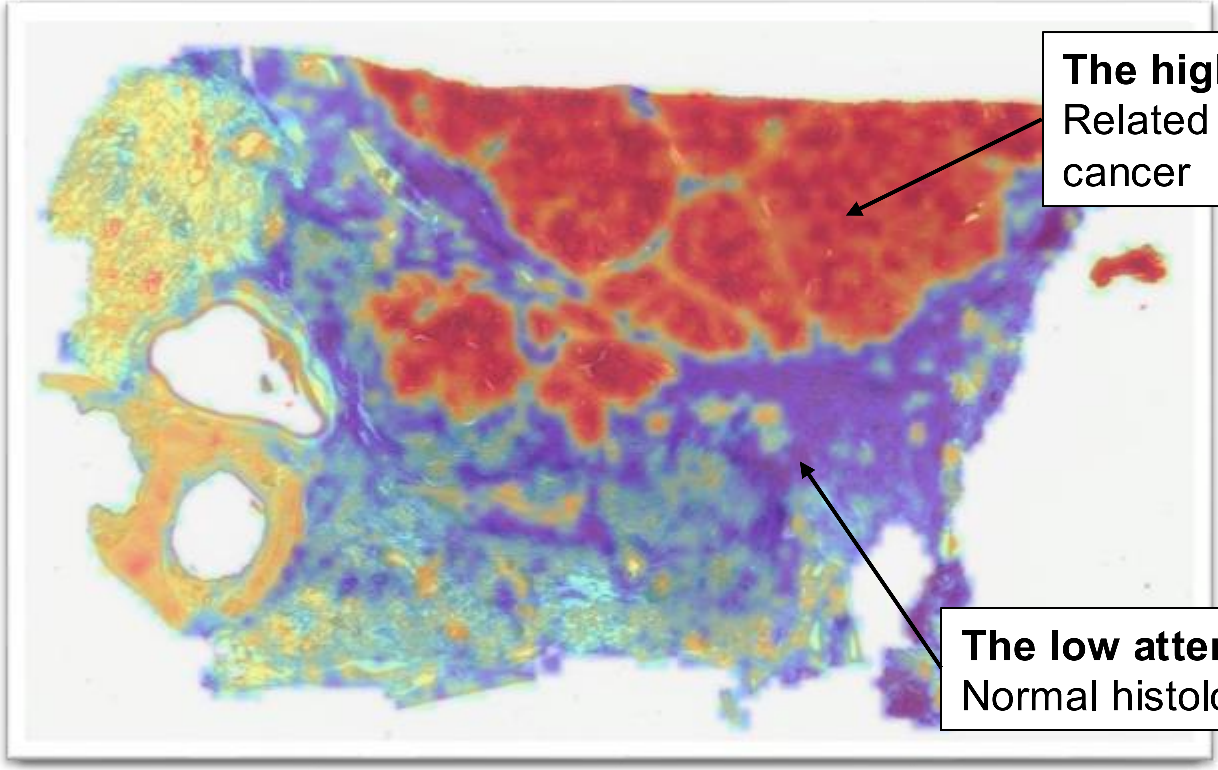


ReVL annotation
(a lung squamous cell carcinoma in TCGA-NSCLC)

4 Results

❑ Qualitative evaluation of *VLEER*

- The combination of ReVL annotations and heatmaps enables the accurate detection of abnormal regions within a WSI along with their pathological patterns.



Attention heatmap
(a lung squamous cell carcinoma in TCGA-NSCLC)

5 Conclusion

❑ *VLEER*

- We propose *VLEER* for generating vision-language embeddings in WSI representation.
- The method not only increases the performance on downstream tasks but also provides **explainability** of the prediction.
- The combination of **ReVL annotations** and **attention heatmaps** forms a powerful and interpretable framework for WSI analysis.

Thank you