NEURAL INFORMATION PROCESSING SYSTEMS

Few-Shot Learning from Gigapixel Images via Hierarchical Vision-Language Alignment and Modeling

Bryan Wong[1], Jongwoo Kim[1], Huazhu Fu[2], Mun Yong Yi[1]

[1]KAIST [2]IHPC, A*STAR

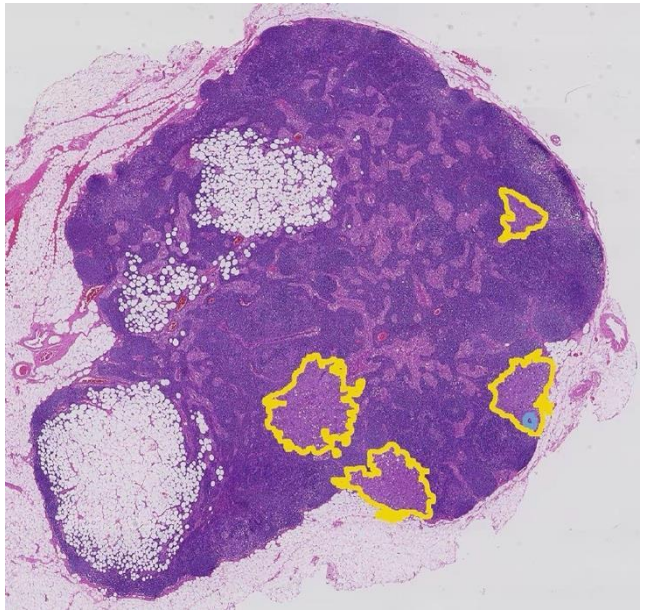
KAIST

Agency for Science, Technology and Research SINGAPORE

Background

NEURAL INFORMATION PROCESSING SYSTEMS

Whole Slide Image (WSI)



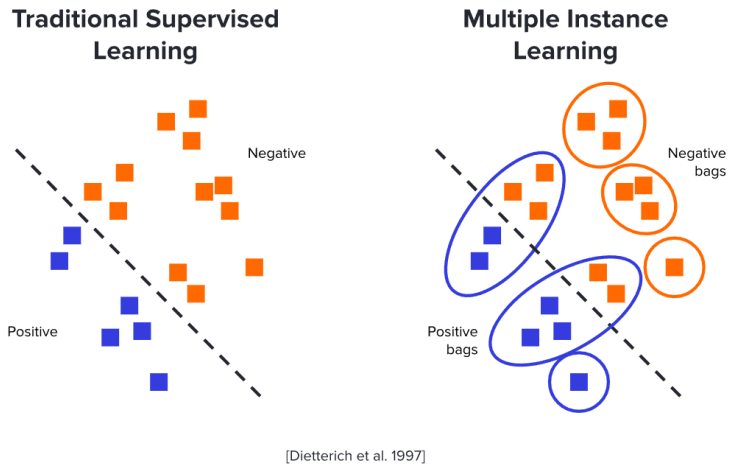
- ***Gigapixel size***
  (≈100,000 X 100,000 pix)
- ***Hierarchical structure***
  (5x, 20x)

Fully-supervised methods

(Fine-grained annotations is expensive)

Multiple Instance Learning (MIL)

(Utilizes only WSI label)



[Dietterich et al. 1997]

Motivation

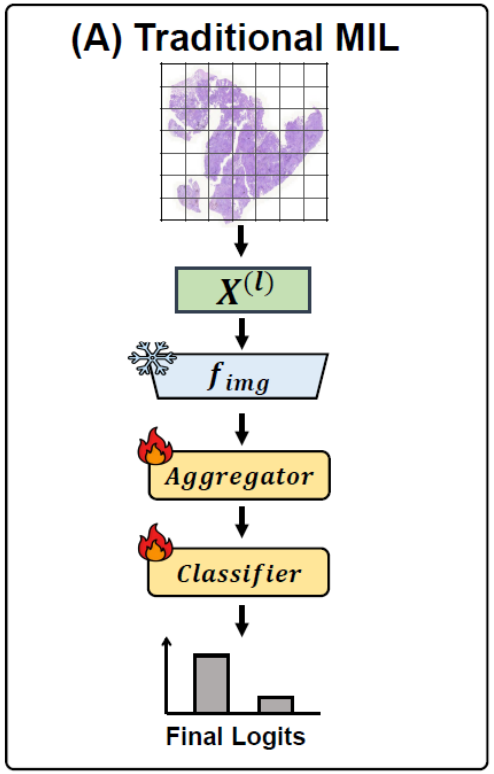
NEURAL INFORMATION PROCESSING SYSTEMS

(A) Traditional MIL

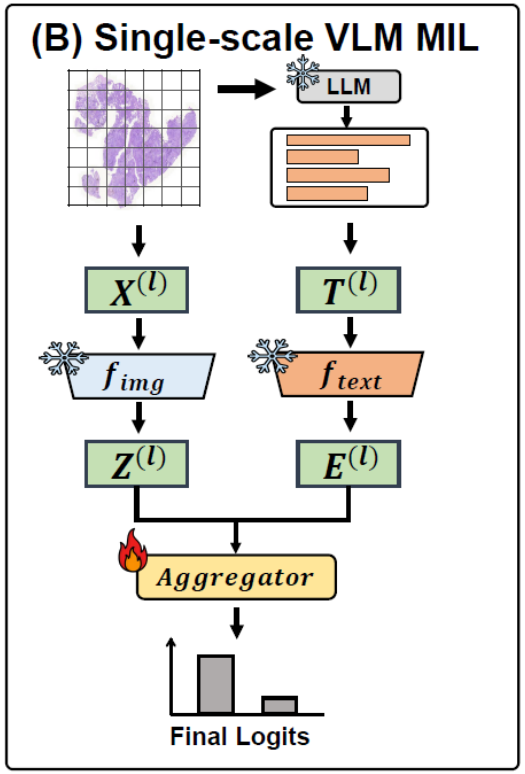
- Requires large labeled WSI datasets
  (privacy & rare diseases)
- Learns only from the original slides
  (staining variability & domain shifts)

(B) Single-scale Vision-Language MIL

- Adds LLM-generated text
  (prior-domain knowledge for data efficiency)
- Lacks of contextual & scale awareness
  (ignores hierarchical structure)

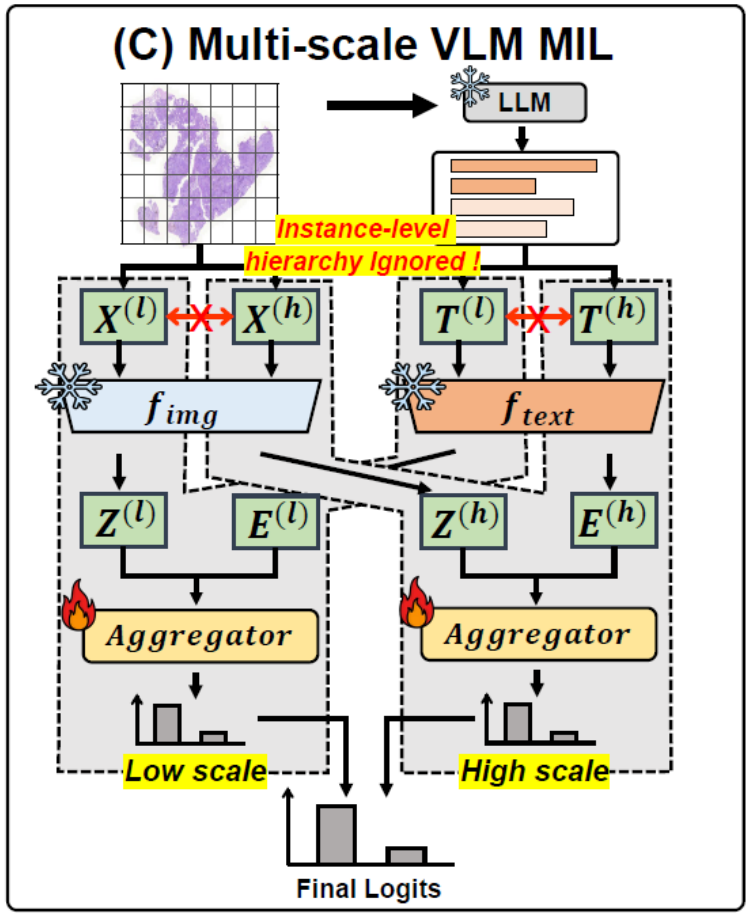


Lack of textual understanding



Lack of contextual understanding

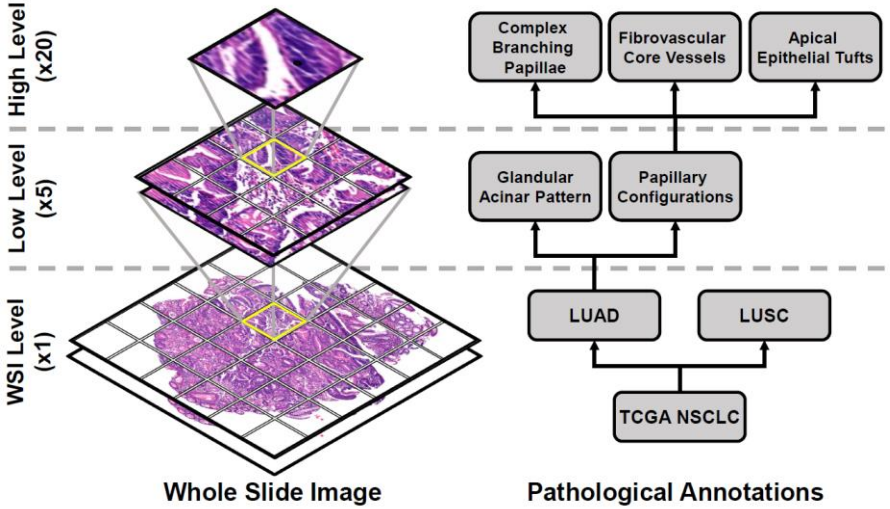
Motivation



Lack of instance hierarchy and modality types

(C) Multi-scale Vision-Language MIL

- Lacks hierarchy modeling within the same modalities *(late-fusion approach)*



- Inadequate alignment between modalities on the same scale

How can we transfer VLM knowledge to gigapixel WSIs for better hierarchical modeling and multimodal integration?

HiVE-MIL

NEURAL INFORMATION PROCESSING SYSTEMS

(D) Hierarchical Vision-Language MIL (HiVE-MIL)

Cross-Scale Hierarchical Interaction

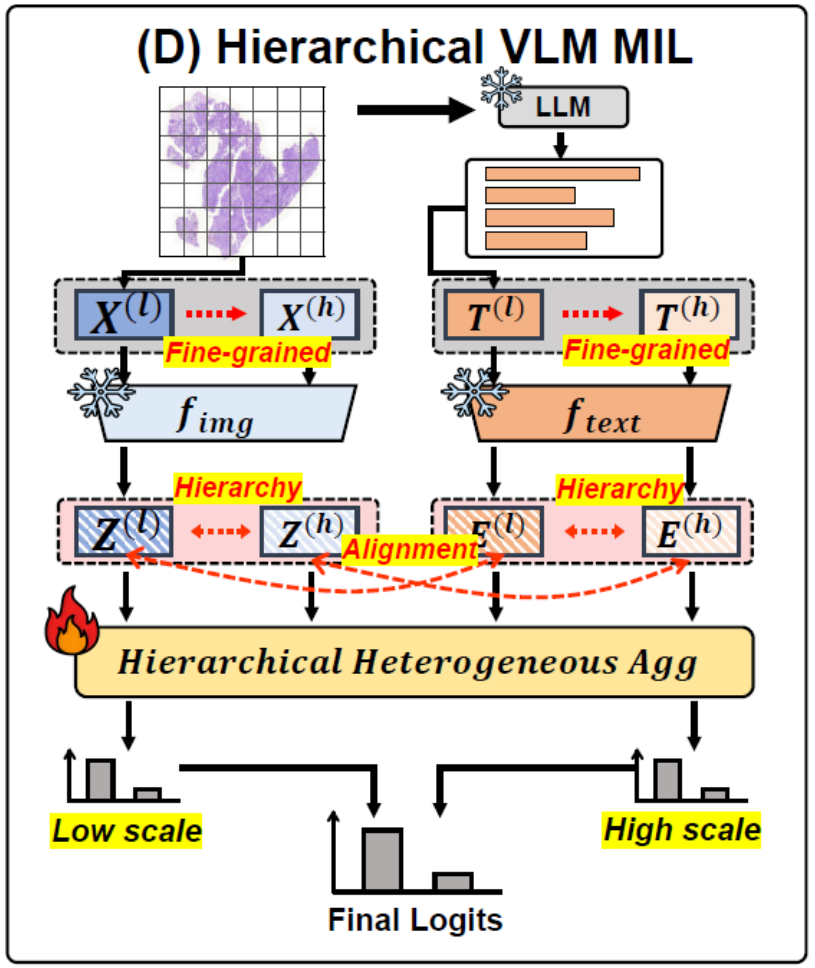
- Hierarchical Graph
  *(builds **parent-child edges** between 5x and 20x visual/text nodes)*

- Hierarchical Text Contrastive Loss (HTCL)
  *(enforces **semantic text consistency** across scales)*

Intra-Scale Multimodal Interaction

- Heterogeneous Graph
  *(**models visual-text relations** on the same scale)*

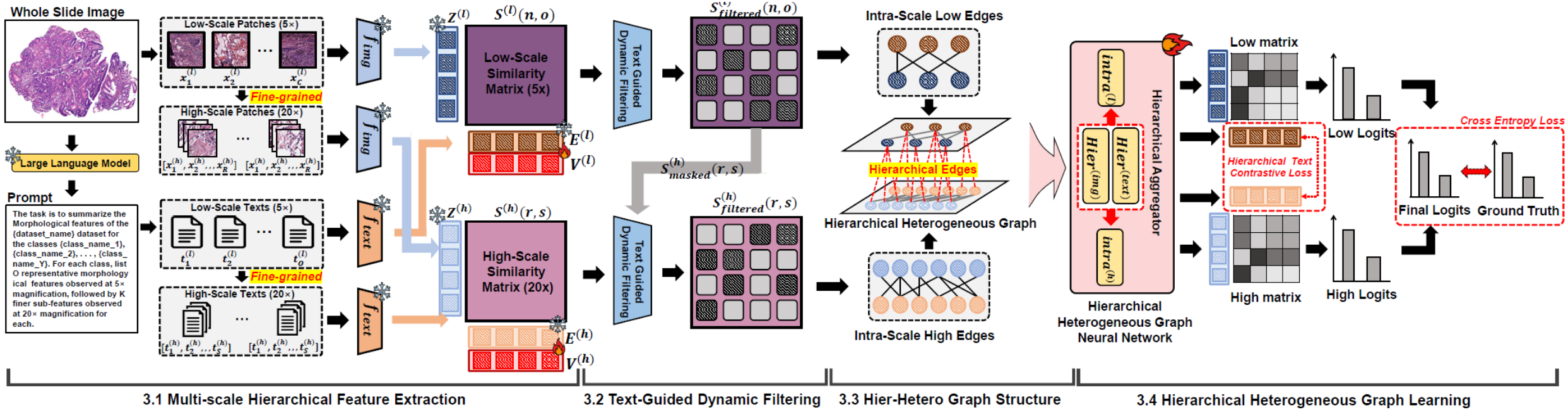
- Text-Guided Dynamic Filtering (TGDF)
  *(**filters out irrelevant or weakly matched** patch-text pairs)*



Captures hierarchical instance relations and modality types

HiVE-MIL

NEURAL INFORMATION PROCESSING SYSTEMS

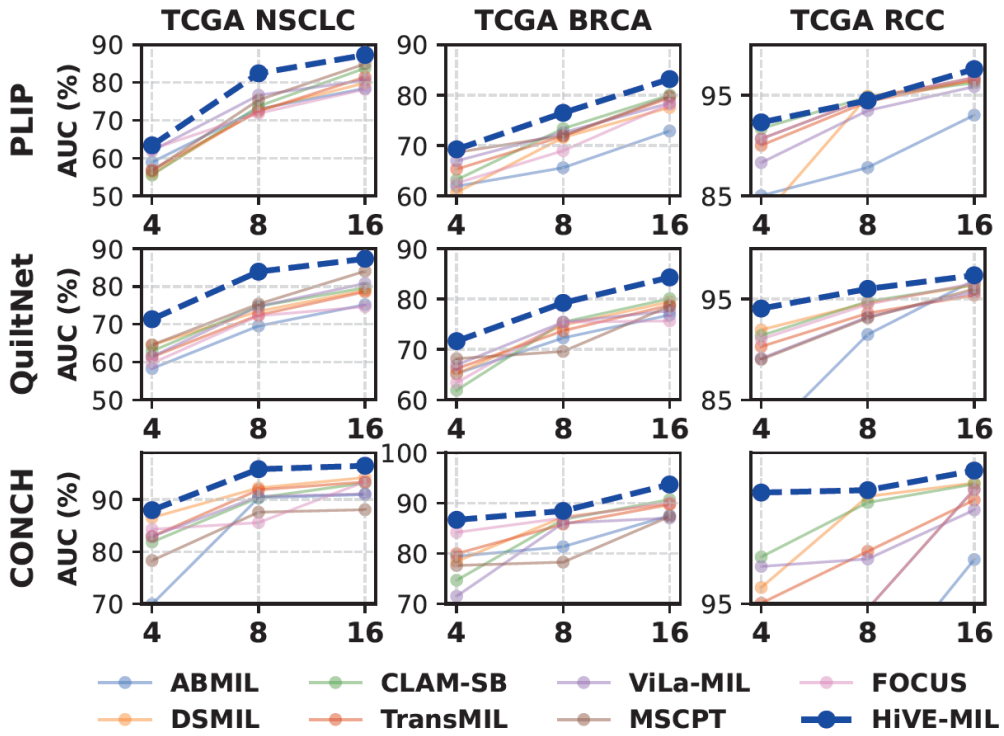


SOTA Performance

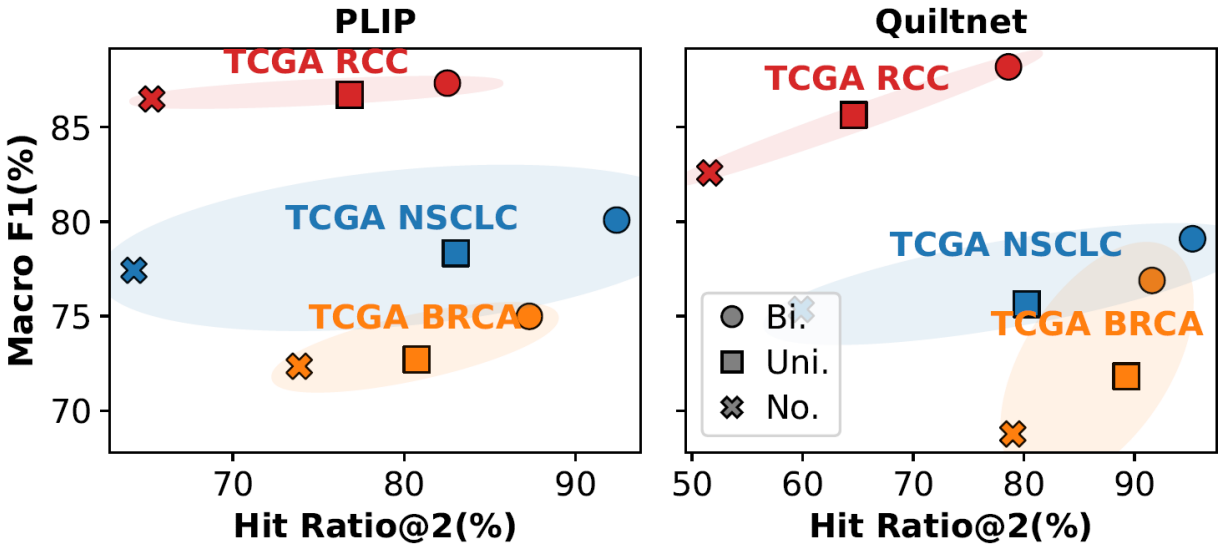
NEURAL INFORMATION PROCESSING SYSTEMS

Table 1: **16-shot** results on three datasets using three pathology VLMs. The best and second-best results are highlighted in **bold** and underlined. HiVE-MIL outperforms all baselines in all settings.

	Dataset	TCGA NSCLC			TCGA BRCA			TCGA RCC		
	Model	ACC	AUC	Macro F1	ACC	AUC	Macro F1	ACC	AUC	Macro F1
PLIP [25] 208K Pathology Image-Text Pairs	Max Pooling	55.00 ±3.88	57.33 ±4.63	53.96 ±4.86	57.29 ±3.23	62.33 ±2.94	53.68 ±6.91	66.82 ±6.94	80.58 ±6.68	61.38 ±8.68
	Mean Pooling	61.73 ±5.65	65.29 ±7.55	61.15 ±6.16	65.25 ±4.40	70.83 ±3.93	64.04 ±4.42	79.62 ±3.51	92.09 ±1.92	76.67 ±3.49
	ABMIL [27]	70.64 ±2.98	78.44 ±3.63	70.37 ±3.09	65.83 ±5.33	72.87 ±7.88	65.29 ±5.78	80.00 ±3.71	93.01 ±1.53	77.95 ±3.43
	DSMIL [33]	72.63 ±3.88	79.88 ±4.60	72.48 ±3.96	71.38 ±3.20	77.55 ±1.62	71.04 ±3.40	86.74 ±1.23	96.44 ±0.63	84.63 ±1.51
	CLAM-SB [38]	75.96 ±2.60	83.79 ±3.21	75.94 ±2.61	71.75 ±3.57	80.00 ±2.59	71.49 ±3.60	85.98 ±1.51	96.22 ±0.48	83.35 ±1.54
	CLAM-MB [38]	73.46 ±3.15	82.13 ±3.41	73.42 ±3.13	72.50 ±2.92	78.39 ±2.95	72.20 ±2.87	86.97 ±1.03	96.53 ±0.78	84.92 ±1.03
	TransMIL [47]	73.21 ±3.02	81.44 ±2.75	72.98 ±2.95	72.08 ±3.32	79.47 ±3.71	71.94 ±3.34	87.05 ±1.52	96.51 ±0.56	84.96 ±1.32
	DTFD-MIL [54]	72.95 ±3.40	79.79 ±4.65	72.91 ±3.39	71.25 ±2.68	78.91 ±3.16	70.86 ±2.76	86.74 ±0.79	95.94 ±0.62	84.86 ±1.45
	WiKG [34]	67.89 ±3.66	75.54 ±4.05	67.51 ±3.62	67.71 ±2.19	74.92 ±4.16	67.15 ±2.42	83.07 ±0.89	94.34 ±0.76	80.32 ±1.40
	ViLa-MIL [48]	74.17 ±1.01	80.63 ±2.37	73.90 ±1.15	71.04 ±6.92	78.42 ±5.86	70.56 ±6.98	85.06 ±2.13	95.53 ±0.97	82.51 ±2.30
	MSCPT [22]	76.86 ±1.85	84.93 ±1.59	76.82 ±1.89	72.71 ±2.90	79.78 ±4.14	72.58 ±2.81	86.21 ±0.54	95.84 ±0.45	84.20 ±0.81
	FOCUS [19]	71.73 ±5.52	78.21 ±5.93	71.65 ±5.51	71.66 ±5.60	78.19 ±4.51	71.36 ±5.69	87.82 ±1.69	96.73 ±0.70	85.54 ±1.87
	HiVE-MIL	**80.13 ±4.73**	**87.28 ±2.76**	**80.08 ±4.73**	**75.21 ±3.51**	**83.19 ±4.72**	**74.99 ±3.67**	**88.89 ±1.36**	**97.58 ±0.41**	**87.18 ±1.78**
	Δ from 2nd-best	**(+3.27)**	**(+2.35)**	**(+3.26)**	**(+2.50)**	**(+3.19)**	**(+2.41)**	**(+1.07)**	**(+0.85)**	**(+1.64)**
QuiltNet [26] 1M Pathology Image-Text Pairs	Max Pooling	53.59 ±3.66	57.24 ±5.97	51.36 ±5.39	55.83 ±4.04	56.64 ±4.36	53.75 ±4.57	68.28 ±6.77	81.33 ±7.72	61.31 ±10.73
	Mean Pooling	60.77 ±4.86	65.68 ±6.04	60.48 ±4.87	65.96 ±2.32	72.41 ±3.86	64.33 ±2.27	79.62 ±3.15	92.09 ±1.92	76.67 ±3.49
	ABMIL [27]	67.31 ±4.64	75.18 ±5.13	66.81 ±5.22	68.96 ±4.86	76.84 ±4.27	68.42 ±5.45	88.89 ±1.71	96.86 ±0.84	87.11 ±2.44
	DSMIL [33]	72.76 ±3.42	78.99 ±3.90	72.53 ±3.41	72.29 ±3.64	79.46 ±2.20	72.06 ±3.54	88.89 ±1.71	96.86 ±0.01	87.11 ±2.44
	CLAM-SB [38]	72.82 ±2.68	79.47 ±2.93	72.58 ±2.74	71.46 ±3.82	80.09 ±1.80	71.24 ±4.00	88.66 ±2.17	97.58 ±0.01	87.00 ±2.98
	CLAM-MB [38]	73.27 ±3.56	80.53 ±3.76	73.25 ±3.55	72.29 ±2.43	78.42 ±2.75	72.24 ±2.47	88.74 ±1.62	97.34 ±0.01	86.83 ±2.50
	TransMIL [47]	71.60 ±4.62	78.59 ±4.86	71.21 ±5.00	71.67 ±3.75	78.77 ±2.92	71.56 ±3.73	86.97 ±1.83	96.71 ±0.01	85.01 ±2.65
	DTFD-MIL [54]	70.51 ±5.77	77.38 ±5.26	70.33 ±5.89	72.71 ±2.02	79.28 ±1.81	72.66 ±1.99	88.66 ±1.65	96.74 ±0.71	87.06 ±1.99
	WiKG-MIL [34]	68.20 ±3.47	75.08 ±4.66	67.98 ±3.56	68.75 ±3.16	75.51 ±2.16	68.59 ±3.07	83.99 ±1.70	95.13 ±0.70	81.54 ±3.14
	ViLa-MIL [48]	73.27 ±5.54	80.82 ±6.41	73.24 ±5.52	72.50 ±3.93	77.67 ±3.12	72.35 ±3.92	84.60 ±1.04	95.67 ±0.70	81.42 ±1.04
	MSCPT [22]	76.15 ±3.83	84.06 ±3.02	76.13 ±3.82	72.08 ±5.16	78.59 ±4.21	71.82 ±5.21	87.20 ±1.90	96.89 ±0.87	85.33 ±2.41
	FOCUS [19]	69.04 ±3.54	74.64 ±4.29	69.00 ±3.56	68.75 ±4.42	75.66 ±2.86	68.47 ±4.70	89.12 ±1.23	97.13 ±0.46	87.43 ±1.68
	HiVE-MIL	**79.23 ±2.70**	**87.34 ±4.08**	**79.09 ±2.75**	**77.08 ±3.90**	**84.31 ±4.22**	**76.80 ±4.15**	**89.97 ±0.85**	**98.32 ±0.45**	**88.18 ±1.25**
	Δ from 2nd-best	**(+3.08)**	**(+3.28)**	**(+2.96)**	**(+4.37)**	**(+4.22)**	**(+4.14)**	**(+0.85)**	**(+0.74)**	**(+0.75)**
CONCH [39] 1.17M Pathology Image-Text Pairs	Max Pooling	78.85 ±1.78	87.43 ±1.69	78.82 ±1.77	71.25 ±2.99	78.46 ±4.53	70.91 ±3.14	80.15 ±4.86	91.95 ±2.76	78.11 ±4.60
	Mean Pooling	79.55 ±2.73	87.90 ±2.78	79.47 ±2.74	76.67 ±2.92	86.08 ±4.43	76.47 ±2.81	87.74 ±0.69	96.76 ±0.47	86.06 ±0.46
	ABMIL [27]	84.30 ±2.22	90.97 ±0.60	84.28 ±2.21	81.04 ±3.05	87.50 ±5.38	80.93 ±3.04	88.43 ±1.95	96.17 ±0.76	86.95 ±2.33
	DSMIL [33]	85.83 ±2.78	94.23 ±1.20	85.76 ±2.84	82.08 ±3.92	89.91 ±5.46	81.99 ±3.89	91.95 ±1.95	98.20 ±0.23	90.87 ±2.00
	CLAM-SB [38]	85.83 ±4.25	93.19 ±2.39	85.80 ±4.29	82.29 ±7.42	90.70 ±6.73	82.24 ±7.41	92.11 ±0.52	98.17 ±0.33	90.76 ±0.85
	CLAM-MB [38]	86.92 ±3.39	94.01 ±2.16	86.91 ±3.40	81.88 ±4.82	90.41 ±5.14	81.84 ±4.81	91.42 ±1.13	98.15 ±0.22	89.96 ±1.11
	TransMIL [47]	85.90 ±3.36	93.38 ±2.11	85.88 ±3.36	82.50 ±5.37	89.69 ±4.54	82.38 ±5.36	89.27 ±2.34	97.75 ±0.69	87.66 ±2.95
	DTFD-MIL [54]	88.40 ±3.54	95.36 ±1.52	88.37 ±3.56	83.54 ±3.86	91.22 ±3.39	83.48 ±3.83	91.65 ±1.44	97.99 ±0.09	90.38 ±1.52
	WiKG [34]	82.24 ±3.13	91.17 ±1.62	82.15 ±3.21	79.58 ±6.17	87.42 ±6.54	79.44 ±6.39	89.73 ±2.37	97.65 ±0.67	87.84 ±3.12
	ViLa-MIL [48]	83.08 ±3.63	91.10 ±2.43	83.04 ±3.64	77.08 ±6.69	87.03 ±8.01	76.98 ±6.73	89.27 ±2.32	97.48 ±0.79	87.91 ±2.88
	MSCPT [22]	80.06 ±5.20	88.06 ±6.28	79.95 ±5.24	79.79 ±8.22	87.33 ±6.78	79.69 ±8.21	92.03 ±1.52	98.03 ±0.35	90.89 ±1.94
	FOCUS [22]	85.32 ±2.54	93.43 ±1.45	85.24 ±2.60	82.50 ±5.57	90.10 ±4.50	82.20 ±5.77	91.57 ±1.14	98.13 ±0.54	90.21 ±1.37
	HiVE-MIL	**90.39 ±1.57**	**96.49 ±0.56**	**90.37 ±1.58**	**87.29 ±2.83**	**93.86 ±0.89**	**87.24 ±2.85**	**92.34 ±1.33**	**98.53 ±0.13**	**91.32 ±1.68**
	Δ from 2nd-best	**(+1.99)**	**(+1.13)**	**(+2.00)**	**(+3.75)**	**(+2.64)**	**(+3.76)**	**(+0.23)**	**(+0.33)**	**(+0.43)**



Hierarchical Text Semantic Alignment



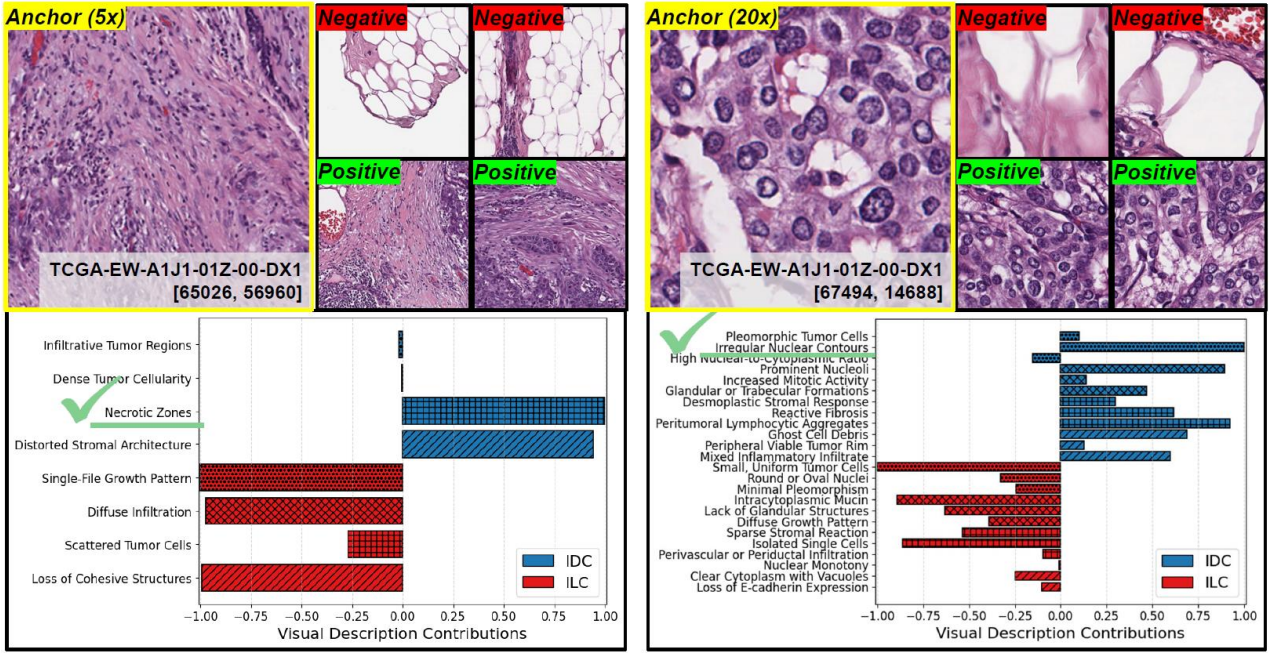
Hit Ratio Evaluation

- **Step 1:** Find the **2 most similar parent texts** for each low-scale patch
- **Step 2:** Collect their child texts as **candidate children**
- **Step 3:** Check if any **high-scale patch matches a candidate child**. If not, check other children under the same parent
- **Step 4:** Record a ***hit*** when both alignments (low ↔ parent and high ↔ child) are correct

Hit Ratio has **strong correlation** with Macro F1

(bidirectional message passing preserves hierarchical text consistency)

Interpretability Analysis



- **Positive:** patch with **text distributions most similar** to the Anchor → ***similar*** **morphology**
- **Negative:** patch with the **most dissimilar distributions**→ ***distinct*** **morphology**

Provides interpretable evidence based on the description of the contributing text

NEURAL INFORMATION PROCESSING SYSTEMS

Questions?
Come visit our poster!

Thu, Dec 4 2025, 11 a.m. — 2 p.m. PST

Paper

Code