Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images

Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F.K. Williamson,
Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, Faisal Mahmood

Massachusetts Institute of Technology,
Harvard University,
Mass General Brigham

CVPR 2023

Report: Yu-Chen Lai

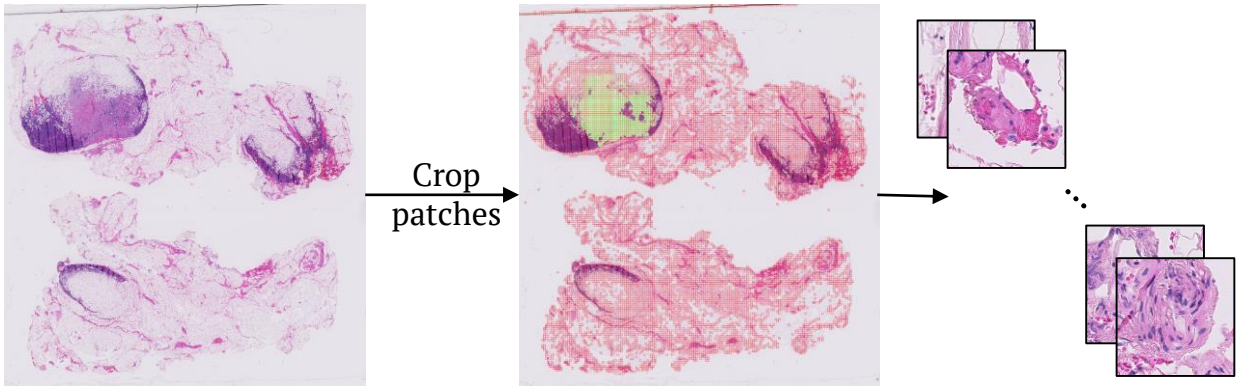
Data: 2023.08.03.

Outline

- Introduction
  - Histopathology Image
  - Zero-shot Learning
- Methods
  - Image caption dataset
  - Unsupervised pretraining of unimodal encoders
  - Aligning vision and language embeddings
  - Zero-shot transfer for image classification
  - Zero-shot transfer for gigapixel WSIs
- Experiments and results
- Conclusion

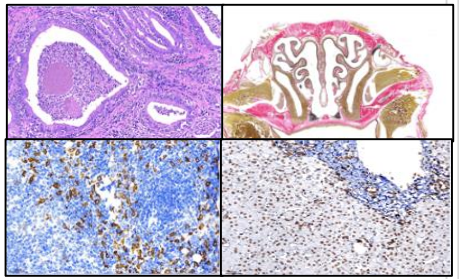
1. Introduction - Histopathology Image

- Challenges of histopathology whole slide images (WSIs)
  1. **High resolution**: Each image can span up to 100,000 × 100,000 pixels.
  2. **Heterogeneous**
     - Different tissue types
     - Tumor characteristics
     - Staining techniques



WSI

patches



Different staining techniques

1. Introduction - Zero-shot Learning

- **Motivation** - Lack of task-agnostic model development
  1. Most tumor types under-represented in public datasets or having inadequate samples for model development.
  2. Developing the slide classifier still requires supervision, which may not be possible for disease types with small sample sizes.

- **Problem** - Zero-shot transfer for pathology has not yet been studied
  - The lack of large-scale, publicly available datasets of paired images and captions in the highly specialized field of pathology.
  - Fundamental computational challenges associated with WSIs(High resolution) and do not routinely come with textual descriptions, bounding box annotations or even region of interest labels.

1. Introduction - Zero-shot Learning

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

} Different Task

x^s

y^s

Hourse Panda Leopard

x^t

1. Introduction - Zero-shot Learning

Training

- Representing each class by its attributes.

x^s

Model

1 0 0

y^s Hourse

Model

0 1 0

Panda

Model

0 0 1

Leopard

Database

	hoofed	Black-and-white	brindle
Hourse	o	x	x
Panda	x	o	x
Leopard	x	x	o
Zebra	o	o	o

1. Introduction - Zero-shot Learning

Testing

- Find the class with the most similar attributes.

x^t

Model

1	1	1

Database

	hoofed	Black-and-white	brindle
Hourse	o	x	x
Panda	x	o	x
Leopard	x	x	o
Zebra	o	o	o

1. Introduction - Zero-shot Learning

x2

x1

x3

f(x2) **D**(Panda) = (0, 1, 0)

f(x1) **D**(Hourse) = (1, 0, 0)

f(x3) **D**(Leopard) = (0, 0, 1)

f(xt)

D(Zebra) = (1, 1, 1)

f: NN model (img→embedding)
D: Database (class→attributes)

1. Introduction - Zero-shot Learning

x2

x1

x3

f(x2) **g**(Panda)

f(x1) **g**(Hourse)

f(x3) **g**(Leopard)

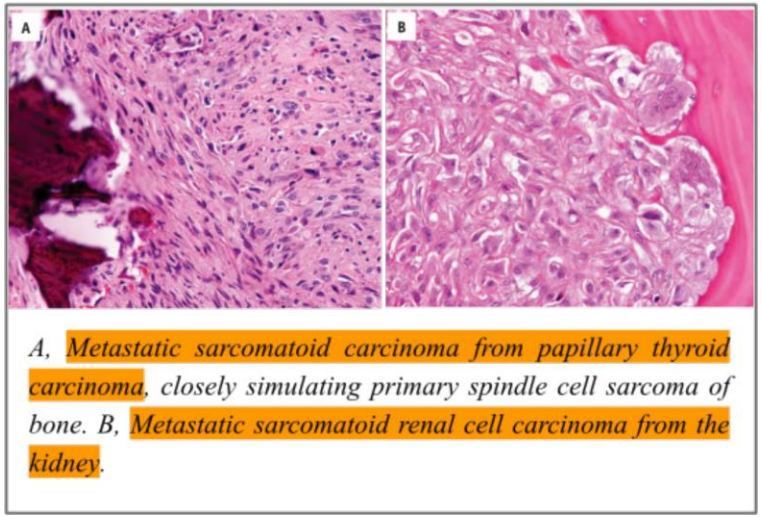
f(xt)

g(Zebra)

f: NN model (img→embedding)
g: NN model (class→word vector)

2.1. Image caption dataset

- Scraping from publicly available educational resources and incorporating the existing **ARCH** dataset [**22**].
- Perform cleaning and filtering, yielding a highly diverse dataset of 33,480 **image-caption pairs** covering a diverse set of tissue sites and morphologies.



- **ARCH dataset**: containing 15,164 histopathology image-caption pairs from pathology textbooks and PubMed research articles.

*[**22**] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology text19772 books and articles. CVPR 2021.*

2.2. Unsupervised pretraining of unimodal encoders

- While our paired dataset currently represents the largest of its kind in the domain of histopathology, it is **still** considerably **smaller** than MIMIC-CXR [34] (radiology, 217k pairs), LiT [86], (general, 4B pairs).
- Therefore, we initialize our encoders using **pretrained weights** before aligning their latent space using paired examples.
    - **Text encoders**
        1. **HistPathGPT**
            - Data: the final diagnosis section of over 550k surgical pathology reports from Massachusetts General Hospital and over 400k histopathology-relevant PubMed abstracts.
            - Model architecture: GPT2-medium [57]
        2. **BioClinicalBert** [2] (biomedical and clinical corpora)
            - Data: MIMIC-III v1.4 database
            - Model architecture: BERT
        3. **PubMedBert** [23]
            - Data: PubMed abstracts
            - Model architecture: BERT

2.2. Unsupervised pretraining of unimodal encoders

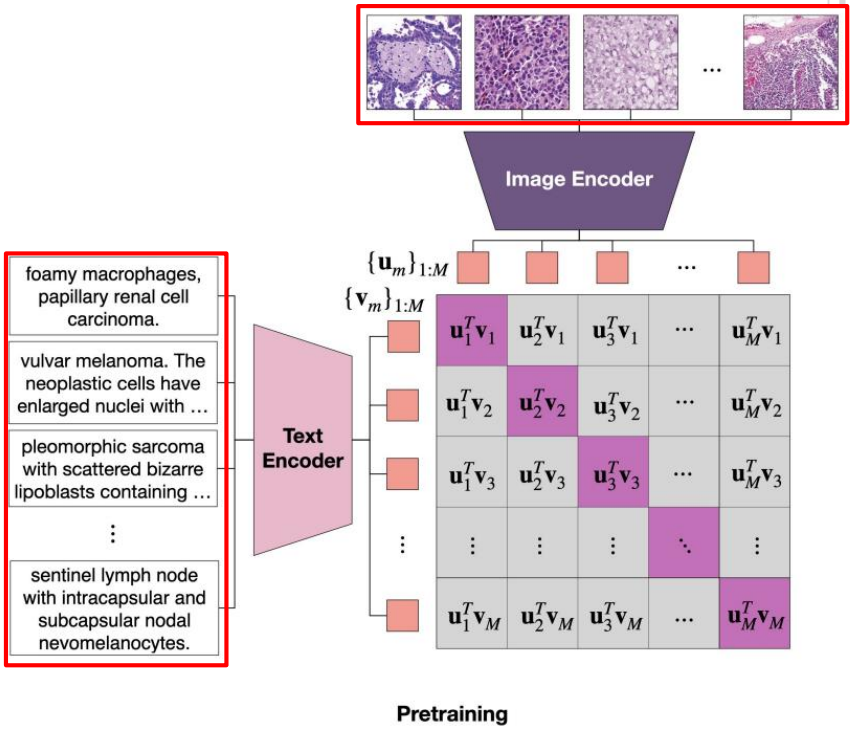
- While our paired dataset currently represents the largest of its kind in the domain of histopathology, it is **still** considerably smaller **than** MIMIC-CXR [34] (radiology, 217k pairs), LiT [86], (general, 4B pairs).
- Therefore, we initialize our encoders using **pretrained weights** before aligning their latent space using paired examples.
    - **Image encoder**
        1. **ViT-S**
            - **ImageNet** pretrained weights
        2. **CTransPath [77] (CTP)**
            - **SOTA** publicly available **encoder** trained using self-supervised representation learning on a total of 15.5M unlabeled **histopathology** image patches.

2.2. Unsupervised pretraining of unimodal encoders

[2] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019.
[23] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 2021
[34] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a deidentified publicly available database of chest radiographs with free-text reports. Scientific data, 2019.
[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019
[77] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis, 2022.
[86] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. CVPR, 2022.

2.3. Aligning vision and language embeddings.

- Align the latent space of our visual and language encoders using the cross-modal contrastive loss formulated as a temperature scaled M-way classification, where M is the global batch-size of **image-text pairs** participating in the loss computation.
- Given a batch of M paired image and text samples $\{(\mathbf{x}_m, \mathbf{t}_m)\}_{m=1,\ldots,M}$.



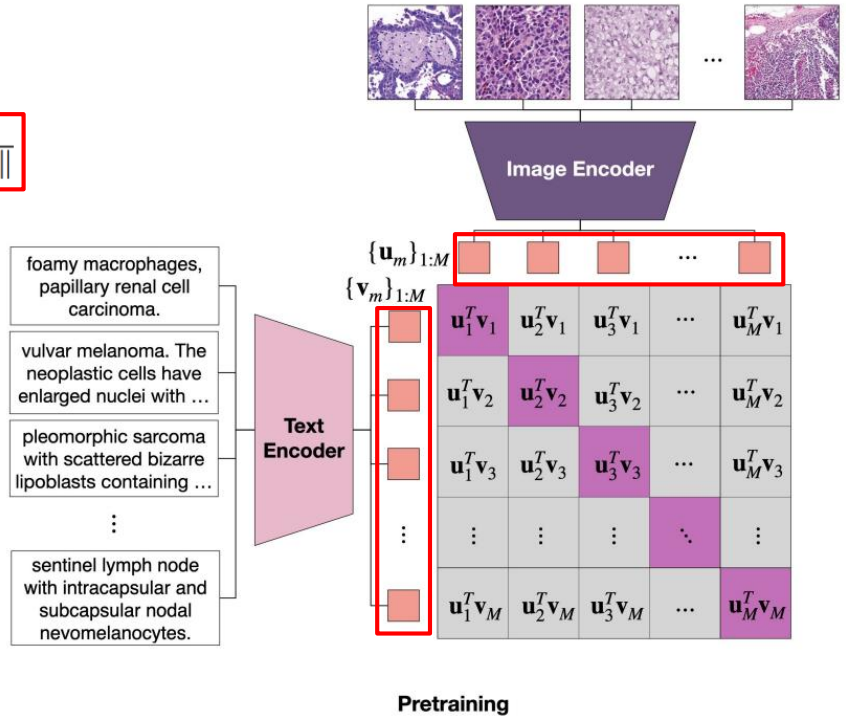
2.3. Aligning vision and language embeddings.

- ℓ_2-normalized visual and text embeddings are computed via the visual and text encoders $f(\cdot;\theta)$ and $g(\cdot;\phi)$ respectively as $\mathbf{u}_m = \frac{f(\mathbf{x}_m;\theta)}{\|f(\mathbf{x}_m;\theta)\|}$ and $\mathbf{v}_m = \frac{g(\mathbf{t}_m;\phi)}{\|g(\mathbf{t}_m,\phi)\|}$.

- The two directions of **contrastive learning** are viewed as symmetric and used jointly to optimize the model during training, where τ is a temperature parameter:

$$\mathcal{L}_{i2t}(\theta,\phi) = -\sum_{i=1}^{M} \log \frac{\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_i\right)}{\sum_{j=1}^{M} \exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j\right)}$$

$$\mathcal{L}_{t2i}(\theta,\phi) = -\sum_{j=1}^{M} \log \frac{\exp\left(\tau \boldsymbol{v}_j^T \boldsymbol{u}_j\right)}{\sum_{i=1}^{M} \exp\left(\tau \boldsymbol{v}_j^T \boldsymbol{u}_i\right)}$$



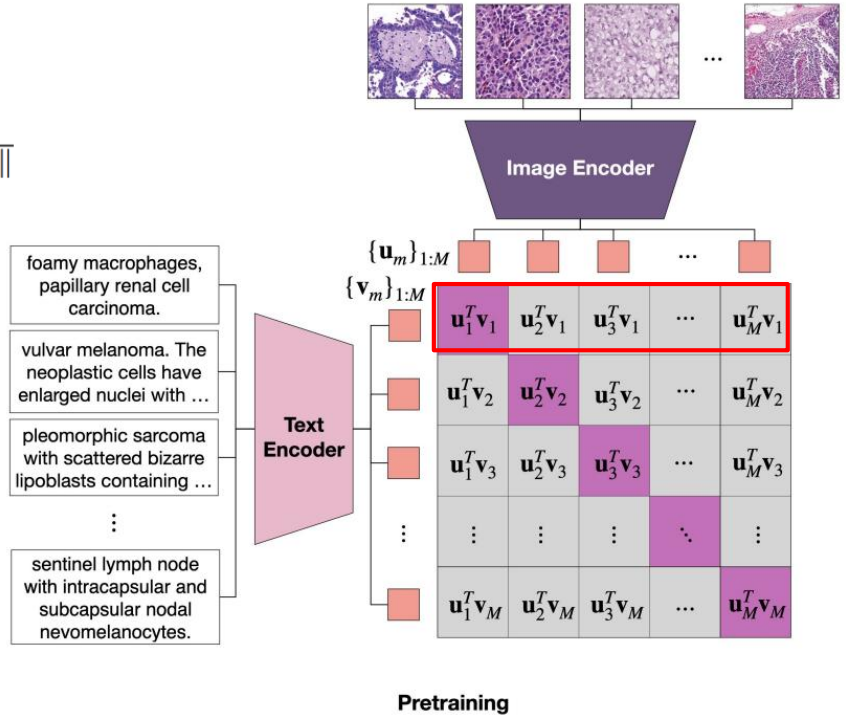
2.3. Aligning vision and language embeddings.

- ℓ_2-normalized visual and text embeddings are computed via the visual and text encoders
- $f(\cdot;\theta)$ and $g(\cdot;\phi)$ respectively as $\mathbf{u}_m = \frac{f(\mathbf{x}_m;\theta)}{\|f(\mathbf{x}_m;\theta)\|}$ and $\mathbf{v}_m = \frac{g(\mathbf{t}_m;\phi)}{\|g(\mathbf{t}_m,\phi)\|}$.

- The two directions of **contrastive learning** are viewed as symmetric and used jointly to optimize the model during training, where τ is a temperature parameter:

$$\mathcal{L}_{i2t}(\theta,\phi) = -\sum_{i=1}^{M} \log \frac{\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_i\right)}{\sum_{j=1}^{M} \exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j\right)}$$

$$\mathcal{L}_{t2i}(\theta,\phi) = -\sum_{j=1}^{M} \log \frac{\exp\left(\tau \boldsymbol{v}_j^T \boldsymbol{u}_j\right)}{\sum_{i=1}^{M} \exp\left(\tau \boldsymbol{v}_j^T \boldsymbol{u}_i\right)}$$



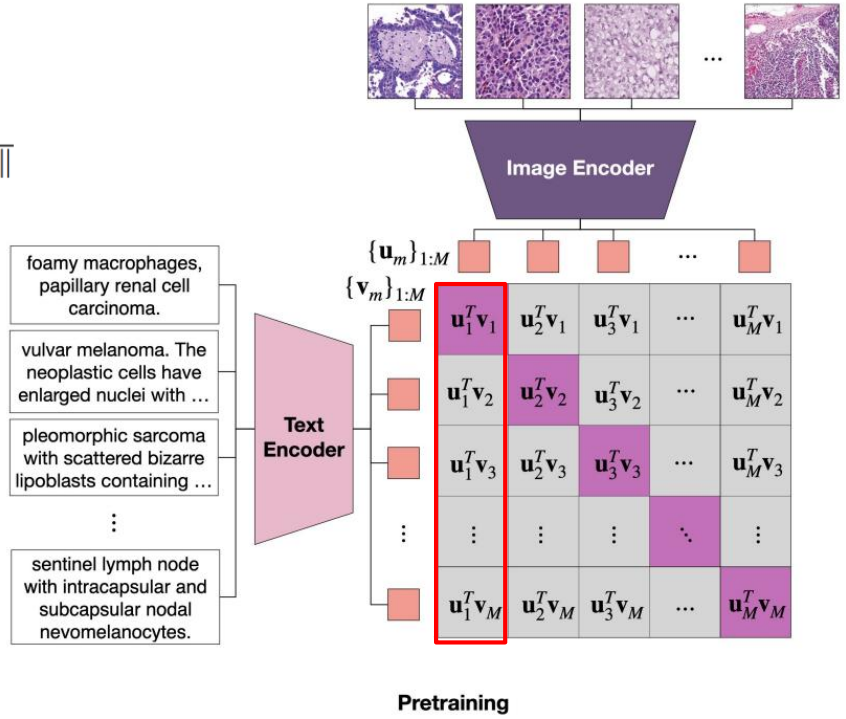
2.3. Aligning vision and language embeddings.

- ℓ_2-normalized visual and text embeddings are computed via the visual and text encoders $f(\cdot;\theta)$ and $g(\cdot;\phi)$ respectively as $\mathbf{u}_m = \frac{f(\mathbf{x}_m;\theta)}{\|f(\mathbf{x}_m;\theta)\|}$ and $\mathbf{v}_m = \frac{g(\mathbf{t}_m;\phi)}{\|g(\mathbf{t}_m,\phi)\|}$.

- The two directions of **contrastive learning** are viewed as symmetric and used jointly to optimize the model during training, where τ is a temperature parameter:

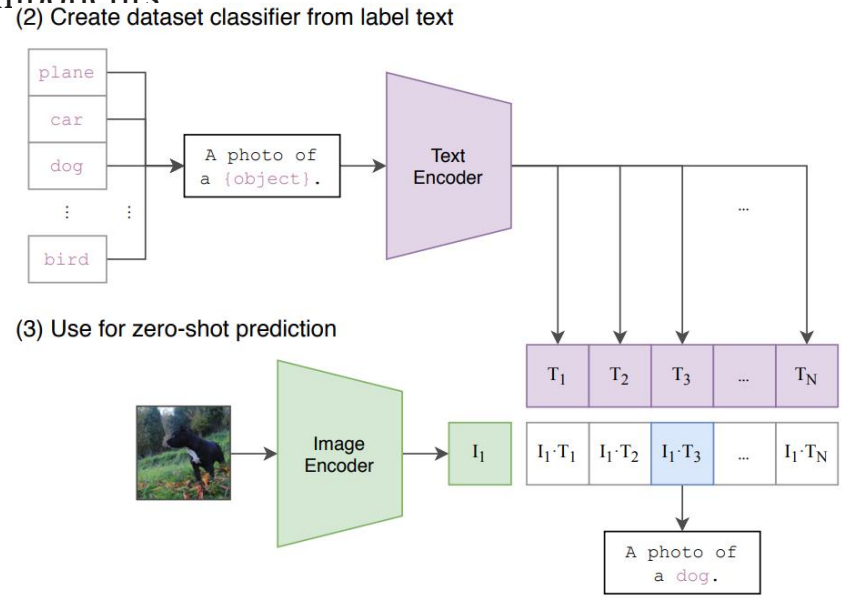
$$\mathcal{L}_{i2t}(\theta,\phi) = -\sum_{i=1}^{M} \log \frac{\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_i\right)}{\sum_{j=1}^{M} \exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j\right)}$$

$$\mathcal{L}_{t2i}(\theta,\phi) = -\sum_{j=1}^{M} \log \frac{\exp\left(\tau \boldsymbol{v}_j^T \boldsymbol{u}_j\right)}{\sum_{i=1}^{M} \exp\left(\tau \boldsymbol{v}_j^T \boldsymbol{u}_i\right)}$$



2.4. Zero-shot transfer for image classification.

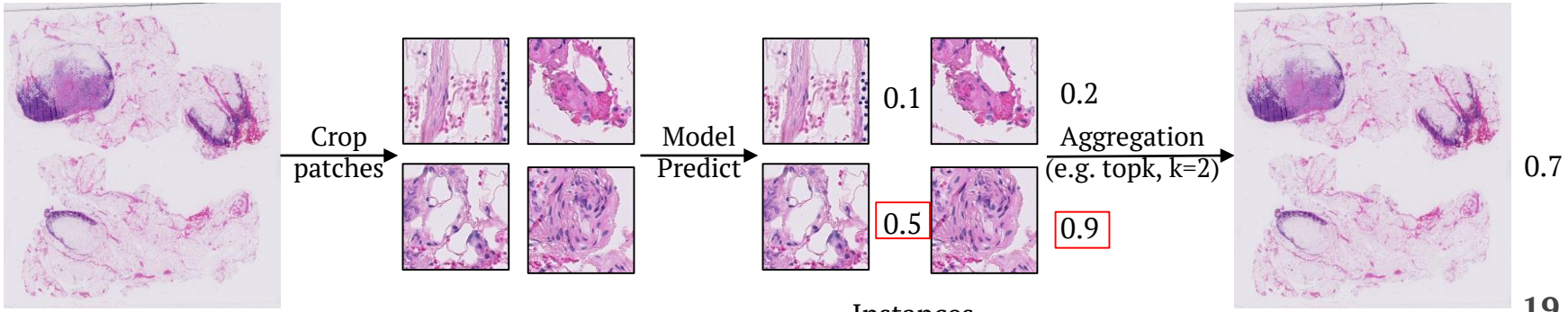
- We briefly describe the prompt-based approach to zero-shot classification popularized by **CLIP**[**55**].
- For each class of interest, a prompt has two components
  - the classname (e.g. "dog")
  - the template (e.g. "A photo of a {}.")
    ⇒ "A photo of a dog."



*[**55**] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. ICML. PMLR, 2021.*

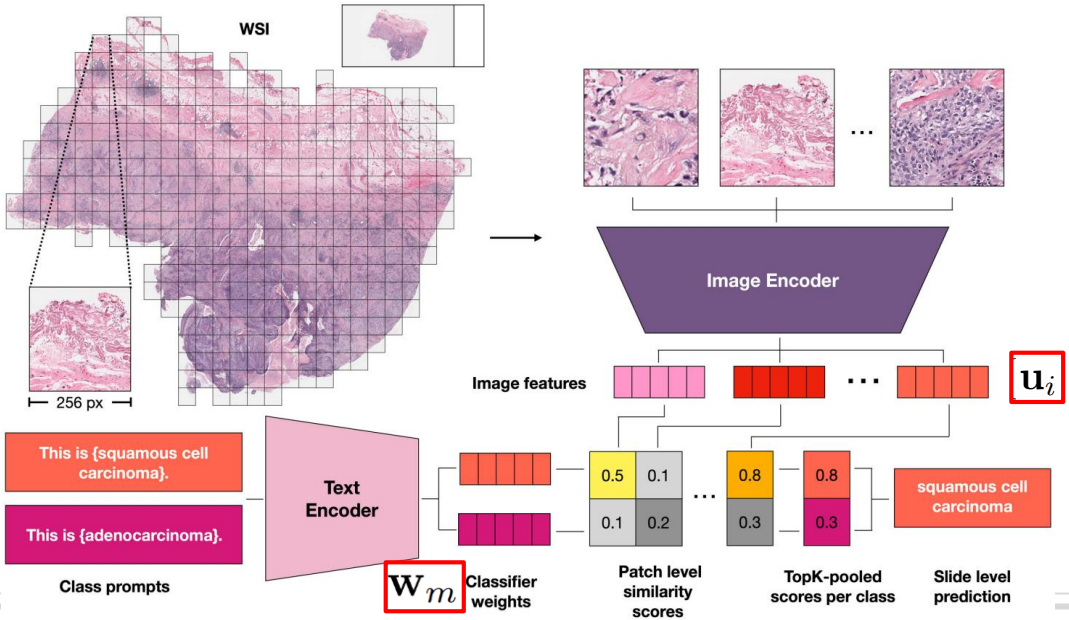
2.5. Zero-shot transfer for gigapixel WSIs.

- Key challenges in performing zero-shot transfer for WSIs
  1. **High resolution**: It's intractable to directly compute an embedding vector using the image encoder.
  2. **Heterogeneous:** Various tissue and cell types that can interact to form higher level architectures of both normal and diseased morphological patterns.
- We propose **MI-Zero**, a zero-shot transfer framework for classifying WSIs inspired by the success of **multiple instance learning (MIL)** for solving weakly-supervised learning tasks.
- The approach entails first dividing each WSI (called **bags**) into smaller tiles (called **instances**) more amenable to processing via our image encoder.



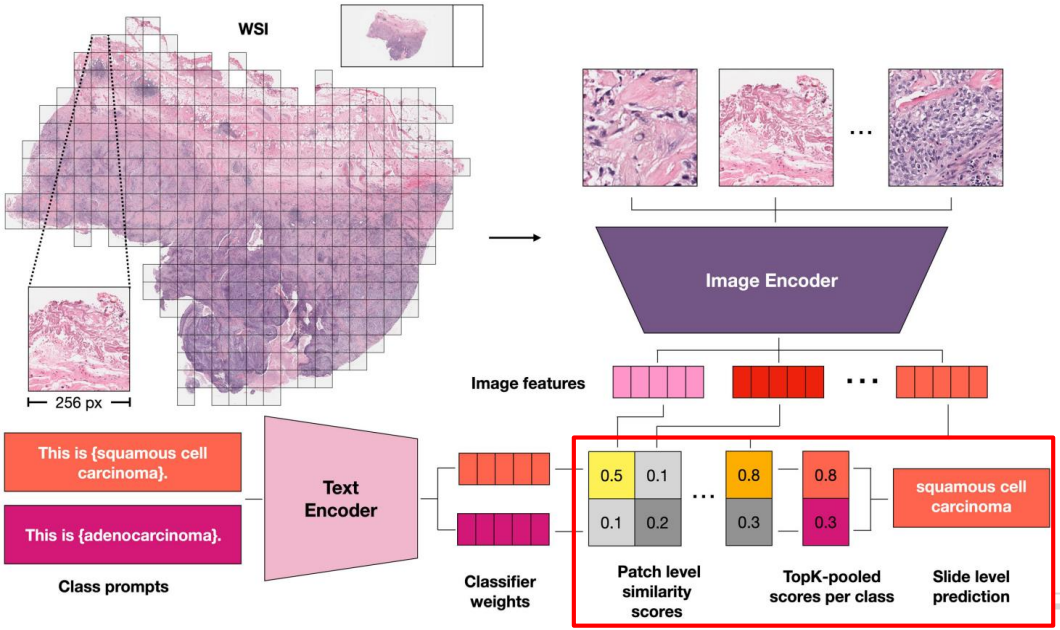
2.5. Zero-shot transfer for gigapixel WSIs.

- $\{\mathbf{u}_i\}_{i=1,\ldots,N}$: the embeddings of **each patch**, where N varies depending on how large each WSI is.
- $\{\mathbf{w}_m\}_{m=1,\ldots,C}$: the prompt embeddings of **each class**, where C is the total number of classes.
- $\{\mathbf{s}_i\}_{i=1,\ldots,N}$: the cosine similarity scores between each patch embedding and prompt embeddings , where $\mathbf{s}_i = \mathbf{u}_i^T[\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C]$



2.5. Zero-shot transfer for gigapixel WSIs.

- $h(\cdot)$: any permutation invariant operator to produce the slide-level prediction scores, such as **mean operator**, **topK max pooling** operator , where $\mathcal{S} = \{\mathbf{s}_i\}_{i=1,\ldots,N}$, $\mathcal{S}^c_{\text{topK}} = \{\tilde{s}^c_i\}_{i=1,\ldots,K}$ is the set of the K largest score values from S for class c



$$h_{\text{mean}}(\mathcal{S}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i$$

$$h_{\text{topK}}(\mathcal{S}) = \frac{1}{K}\left[\sum_{i=1}^{K}\tilde{s}^1_i, \sum_{i=1}^{K}\tilde{s}^2_i, \ldots, \sum_{i=1}^{K}\tilde{s}^C_i\right]^T$$

2.5. Zero-shot transfer for gigapixel WSIs.

- In the **graph-based representation**, we take into account the spatial positions of each patch.
  1. Build a directed **KNN graph** G = {M, E} connecting each patch (node) to its spatial neighbors, where the value at node i is its scores $\mathbf{s}_i$.
  2. We **spatially smooth** (e.g. average) the score values, by **replacing** $\mathbf{s}_i$ with $h_{\text{mean}}(\mathcal{S}_{\text{neighbors}})$, where $\mathcal{S}_{\text{neighbors}} = \{\mathbf{s}_j : j \in \{i\} \cup \mathcal{N}(i)\}$ and $\mathcal{N}(i) = \{j : (i,j) \in \mathcal{E}\}$ for each node i in the graph.
  3. Applying one of the permutation invariant **pooling** operators to the set of smoothed scores in the graph, $\mathcal{S}_{\text{smoothed}}$, and arrive at the slide-level prediction scores.
- Note that this is equivalent to applying a **mean-filter** with the receptive field size covering each patch's k-nearest neighbors.

4.1. Downstream datasets

- Zero-shot transfer performance for cancer subtype classification was evaluated on **3** Whole Slide Image (WSI) datasets from Brigham and Women's Hospital.
- We used **in-house independent datasets** for zero-shot transfer evaluation to avoid information leakage.

Dataset	# of WSIs	# of types
Independent BRCA	200	2
Independent NSCLC	200	2
Independent RCC	150	3

4.2. Supervised baselines

- **Supervised baselines**: weakly-supervised attention-based MIL (ABMIL) [**30**]
- **Trainset**: publicly available TCGA cohort of each task(~1000 WSIs each).
- Due to the relatively small size of these datasets we follow the study design of other weakly-supervised classification studies by performing 5-fold Monte Carlo cross-validation.
- Each cross-validation training set includes on average 836 slides for **BRCA**, 838 for **NSCLC**, and 739 for **RCC**.

*[**30**] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. ICML. PMLR, 2018.*

4.3. Zero-shot transfer

- **Zero-shot evaluation methodology**
  - Due to the reliance on prompts for zero-shot transfer, evaluation results vary with the choice of **class names** and **prompt templates**.
  - For each task, we first curate a pool of relevant prompt templates and classnames. We then evaluate each model configuration on each task by randomly sampling **50 prompts** and measuring the performance of each prompt.
  - **16 prompt templates**:
    - CLASSNAME.
    - a photomicrograph showing CLASSNAME.
    - a photomicrograph of CLASSNAME.
    - an image of CLASSNAME.
    - an image showing CLASSNAME.
    - an example of CLASSNAME.
    - ⋮

- **class names**

Task	Class	Class names
BRCA	IDC	invasive ductal carcinoma carcinoma of the breast, ductal pattern
	ILC	invasive lobular carcinoma carcinoma of the breast, lobular pattern
NSCLC	LUAD	adenocarcinoma lung adenocarcinoma adenocarcinoma of the lung pulmonary adenocarcinoma adenocarcinoma, lepidic pattern adenocarcinoma, solid pattern adenocarcinoma, micropapillary pattern adenocarcinoma, acinar pattern adenocarcinoma, papillary pattern
	LUSC	squamous cell carcinoma lung squamous cell carcinoma squamous cell carcinoma of the lung pulmonary squamous cell carcinoma
RCC	CCRCC	clear cell renal cell carcinoma renal cell carcinoma, clear cell type renal cell carcinoma of the clear cell type clear cell RCC
	PRCC	papillary renal cell carcinoma renal cell carcinoma, papillary type renal cell carcinoma of the papillary type papillary RCC
	CHRCC	chromophobe renal cell carcinoma renal cell carcinoma, chromophobe type renal cell carcinoma of the chromophobe type chromophobe RCC

4.3. Zero-shot transfer

- **Zero-shot transfer for WSIs**
  - **Image encoder**: CTP
  - **SS**: spatial smoothing

Model	Text Encoder & Pretraining	SS	Pooling	BRCA	NSCLC	RCC	Average
ABMIL (1% Data)	None	✗	attention	0.510	0.709	0.557	0.592
ABMIL (100% Data)	None	✗	attention	0.843	0.893	0.855	0.864
MI-Zero (Ours)	HistPathGPT (None)	✗	topK	0.625	0.680	0.653	0.653
	HistPathGPT (In-domain)	✗	topK	**0.673**	0.700	0.733	**0.702**
	PubmedBert (Out-of-domain)	✗	topK	0.570	0.693	**0.777**	0.680
	BioclinicalBert (Out-of-domain)	✗	topK	0.660	**0.742**	0.697	0.700
MI-Zero (Ours)	HistPathGPT (None)	✓	topK	0.623	0.700	0.653	0.659
	HistPathGPT (In-domain)	✓	topK	0.615	0.705	0.733	0.684
	PubmedBert (Out-of-domain)	✓	topK	0.577	0.725	**0.760**	0.688
	BioclinicalBert (Out-of-domain)	✓	topK	**0.660**	**0.770**	0.663	**0.698**
MI-Zero (Ours)	HistPathGPT (None)	✗	mean	0.655	0.593	0.577	0.608
	HistPathGPT (In-domain)	✗	mean	0.620	0.590	0.633	0.614
	PubmedBert (Out-of-domain)	✗	mean	0.585	0.650	**0.727**	**0.654**
	BioclinicalBert (Out-of-domain)	✗	mean	**0.672**	**0.680**	0.543	0.632
MI-Zero (Ours)	HistPathGPT (None)	✓	mean	0.655	0.595	0.573	0.608
	HistPathGPT (In-domain)	✓	mean	0.625	0.590	**0.637**	0.617
	PubmedBert (Out-of-domain)	✓	mean	0.587	0.650	0.730	**0.656**
	BioclinicalBert (Out-of-domain)	✓	mean	**0.675**	**0.682**	0.543	0.634

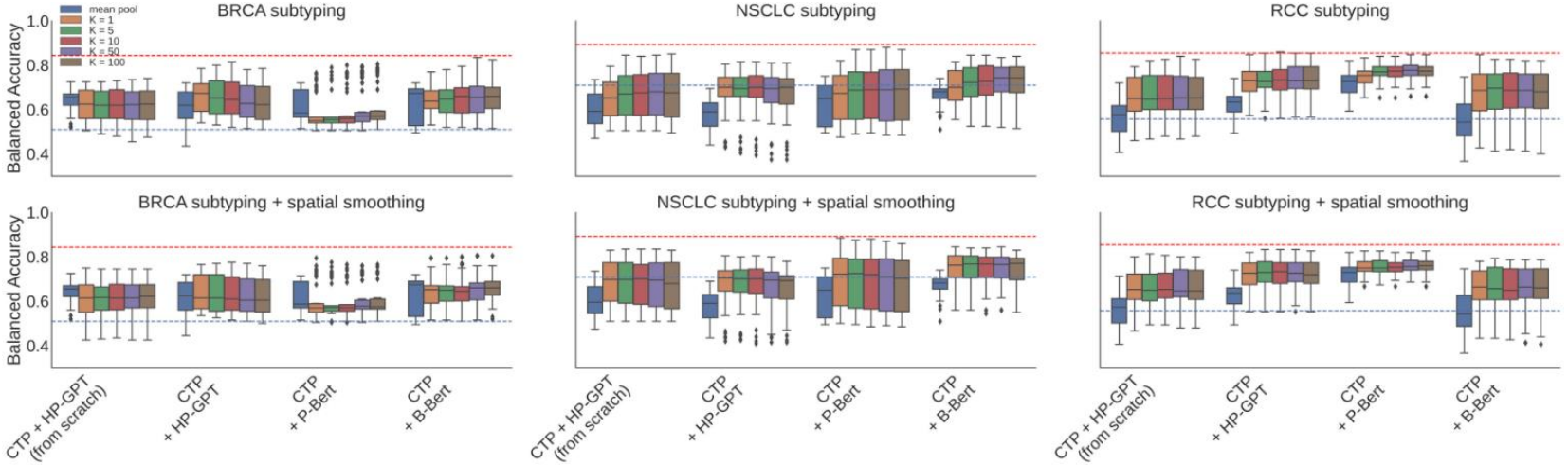
4.3. Zero-shot transfer

- **Zero-shot transfer for WSIs**
  - **Image encoder**: CTP
  - **SS**: spatial smoothing

Model	Text Encoder & Pretraining	SS	Pooling	BRCA	NSCLC	RCC	Average
ABMIL (1% Data)	None	✗	attention	0.510	0.709	0.557	0.592
ABMIL (100% Data)	None	✗	attention	0.843	0.893	0.855	0.864
MI-Zero (Ours)	HistPathGPT (None)	✗	topK	0.625	0.680	0.653	0.653
	HistPathGPT (In-domain)	✗	topK	**0.673**	0.700	0.733	**0.702**
	PubmedBert (Out-of-domain)	✗	topK	0.570	0.693	**0.777**	0.680
	BioclinicalBert (Out-of-domain)	✗	topK	0.660	**0.742**	0.697	0.700
MI-Zero (Ours)	HistPathGPT (None)	✓	topK	0.623	0.700	0.653	0.659
	HistPathGPT (In-domain)	✓	topK	0.615	0.705	0.733	0.684
	PubmedBert (Out-of-domain)	✓	topK	0.577	0.725	**0.760**	0.688
	BioclinicalBert (Out-of-domain)	✓	topK	**0.660**	**0.770**	0.663	**0.698**
MI-Zero (Ours)	HistPathGPT (None)	✗	mean	0.655	0.593	0.577	0.608
	HistPathGPT (In-domain)	✗	mean	0.620	0.590	0.633	0.614
	PubmedBert (Out-of-domain)	✗	mean	0.585	0.650	**0.727**	**0.654**
	BioclinicalBert (Out-of-domain)	✗	mean	**0.672**	**0.680**	0.543	0.632
MI-Zero (Ours)	HistPathGPT (None)	✓	mean	0.655	0.595	0.573	0.608
	HistPathGPT (In-domain)	✓	mean	0.625	0.590	**0.637**	0.617
	PubmedBert (Out-of-domain)	✓	mean	0.587	0.650	0.730	**0.656**
	BioclinicalBert (Out-of-domain)	✓	mean	**0.675**	**0.682**	0.543	0.634

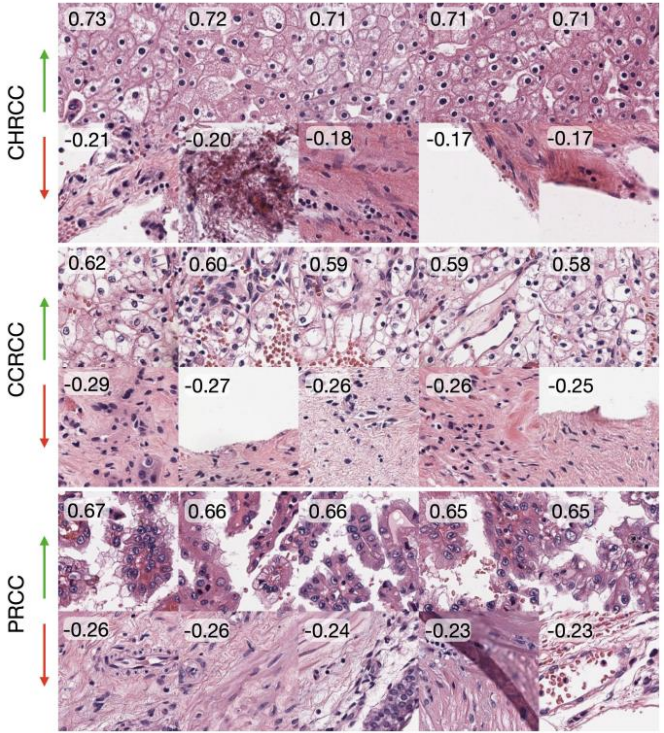
4.3. Zero-shot transfer

- **Zero-shot transfer for WSIs**
  - **Red dashed line:** ABMIL trained on 100% of training data.
  - **Blue dashed line:** ABMIL trained on 1% of training data.
  - **HP-GPT:** HistoPathGPT
  - **P-Bert:** PubMedBert
  - **B-Bert:** BioClinicalBert



4.3. Zero-shot transfer

- **Visualization of similarity scores.**



4.4. Ablation study

- **Training data comparison**
  - **OpenAI's CLIP model [55]**
    - Trained on 400M generic imagetext pairs.
  - **ARCH** (7,562 pathology pairs) [22]
    - A subset of our training data (33,480 pathology pairs).

Dataset	BRCA	NSCLC	RCC	Average
CLIP [55]	0.500	0.500	0.333	0.444
ARCH [22]	0.625	0.593	0.540	0.586
Ours	**0.672**	**0.700**	**0.733**	**0.702**

*[**55**] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. ICML. PMLR, 2021.*
*[**22**] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology text19772 books and articles. CVPR 2021.*

4.4. Ablation study

- **Image encoder pretraining**
  - **CTransPath [77] (CTP)**
    - **SOTA** publicly available **encoder** trained using self-supervised representation learning on a total of 15.5M unlabeled **histopathology** image patches.

Image Encoder	Text Encoder	Image Pretraining	Text Pretraining	BRCA	NSCLC	RCC	Average
CTP	HistPathGPT	SSL	In-domain	**0.672**	**0.700**	**0.733**	**0.702**
ViT-S	HistPathGPT	SSL	In-domain	0.617	0.625	0.673	0.639
ViT-S	HistPathGPT	ImageNet	In-domain	0.660	0.525	0.600	0.595
CTP	HistPathGPT	None	None	0.535	0.520	0.297	0.451
ViT-S	HistPathGPT	None	None	0.500	0.510	0.290	0.433

[77] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. Medical Image Analysis, 2022.

4.4. Ablation study

- **Locked-image tuning (LiT)**
  - Zhai et al. [86] recently showed that "locking" a well-pretrained image encoder outperforms its unlocked counterpart during contrastive tuning.

Text Encoder & Pretraining	LIT	SS	Pooling	BRCA	NSCLC	RCC	Average
HistPathGPT (In-domain)	✗	✗	topK	**0.672**	0.700	0.733	0.702
	✓			0.690	0.670	0.760	**0.707**
PubMedBert (Out-of-domain)	✗	✗	topK	0.570	0.693	**0.777**	0.680
	✓			0.597	0.615	0.643	0.619
BioClinicalBert (Out-of-domain)	✗	✗	topK	0.660	**0.742**	0.697	0.700
	✓			0.575	0.623	0.547	0.581
HistPathGPT (In-domain)	✗	✓	topK	0.615	0.705	0.733	0.684
	✓			**0.688**	0.675	0.740	**0.701**
PubMedBert (Out-of-domain)	✗	✓	topK	0.577	0.725	**0.760**	0.688
	✓			0.595	0.625	0.647	0.622
BioClinicalBert (Out-of-domain)	✗	✓	topK	0.660	**0.770**	0.663	0.698
	✓			0.600	0.635	0.543	0.593
HistPathGPT (In-domain)	✗	✗	mean	0.620	0.590	0.633	0.614
	✓			0.603	0.557	0.600	0.587
PubMedBert (Out-of-domain)	✗	✗	mean	0.585	0.650	**0.727**	**0.654**
	✓			0.573	0.557	0.543	0.558
BioClinicalBert (Out-of-domain)	✗	✗	mean	**0.672**	**0.680**	0.543	0.632
	✓			0.607	0.575	0.533	0.572
HistPathGPT (In-domain)	✗	✓	mean	0.625	0.590	0.637	0.617
	✓			0.605	0.557	0.600	0.588
PubMedBert (Out-of-domain)	✗	✓	mean	0.587	0.650	**0.730**	**0.656**
	✓			0.575	0.560	0.543	0.559
BioClinicalBert (Out-of-domain)	✗	✓	mean	**0.675**	**0.682**	0.543	0.634
	✓			0.613	0.577	0.533	0.574

*[**86**] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. CVPR, 2022.*

4.5. Conclusion

- **MI-Zero:** The first method for zero-shot transfer in pathology.
- **Future directions**
  - Collect **additional image caption datasets.**
  - Explore methods that may improve the sample efficiency of visual language pretraining.
  - Evaluate on a large and diverse set of computational pathology benchmarks.

Thanks For Listening !