Supplementary Material

A. Additional dataset details

A.1. Image-caption dataset preprocessing

Our cleaning and filtering workflow consist of first filtering out invalid image-caption pairs (e.g. either if the caption is empty/nonsensical, or if the corresponding image is missing). We then, to the best of our knowledge, removed non-histopathology images (including gross images, cytology images, X-ray/CT images, electron microscopy images, fluorescent microscopy images, schematic diagrams, etc.) and cropped multipanel figures into individual images, cleaning the paired caption accordingly. The resulting dataset is highly diverse and is expected to cover all major biopsy sites and morphologies of both diseased and normal tissue.

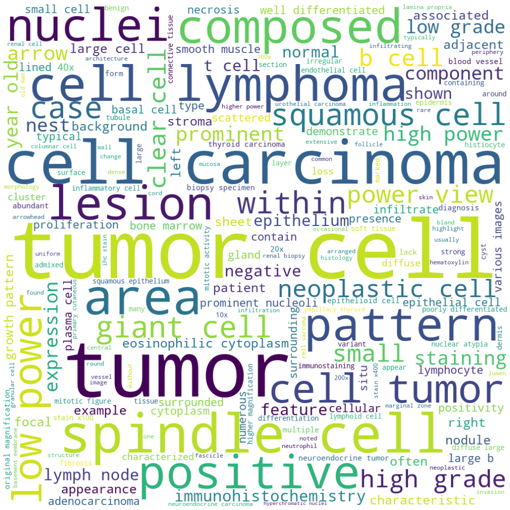

Figure 1. **Word cloud of captions in the paired image-caption dataset used for pretraining.** For clarity, we excluded some common verbs and combined spelling variations of the same word.

A.2. Color variation and diversity of images

Recent large scale studies have demonstrated that stain normalization (SN) is not required to achieve robust generalization [4] , especially when training with large, diverse, multi-institutional datasets that cover a wide range of staining profiles [1]. Therefore we did not perform SN to avoid computational overhead and choosing between SN algorithms. We investigated staining variations in the datasets used (Fig. 2). We find TCGA and our dataset show wider coverage due to sourcing from many institutions and/or diverse tissue sites compared to our in-house independent test set, but we observe substantial overlap across datasets.

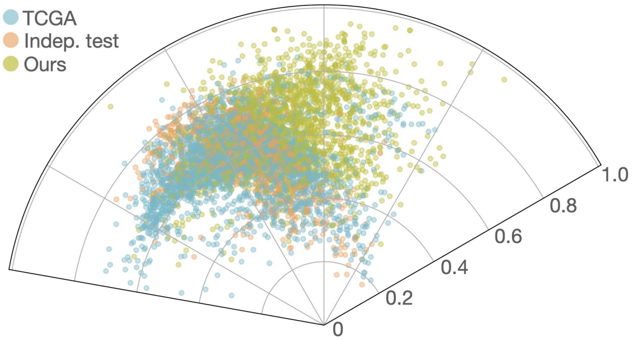

Figure 2. Plot of mean hue (angle) and saturation (radius) of each image (dot) after RGB to Hue-Saturation-Density (HSD) transformation. Sampling was performed to avoid over clutter.

B. Additional training details

B.1. HistPathGPT pretraining

See Table 1.

Hyperparameter	Value
Architecture	gpt2-medium
Max. sequence length	512
Vocabulary size	32000
Batch size	64
Gradient accumulation	4
Weight decay	0.01
AdamW β	(0.9, 0.999)
Peak learnng rate	1e-4
Learning rate schedule	Linear
Warmup steps	750
Training steps	15000

Table 1. **Hyperparameters used in pretraining the text encoder (HistPathGPT)**. In-house pathology reports were first deidentified using regex pattern matching before tokenization. 4 × 80GB NVIDIA A100 GPUs were used for training. Batch size refers to the total batch size across GPUs. Effective batch size used for optimization is batch size × gradient accumulation steps. The sequence length of training examples was set to the maximum sequence length supported by the model (*i.e.* 512).

B.2. Visual language pretraining

See Table 2.

B.3. Supervised baselines

To establish a supervised baseline of comparison for zeroshot transfer, we use the ABMIL weakly-supervised learning algorithm [2]. Same as MI-Zero, instances (*i.e.* patches of size 256 × 256 at 20×-equivalent magnification) are embedded using the SOTA SSL encoder CTransPath

Hyperparameter	Value
Batch Size	512
Weight decay	0.2
AdamW β	(0.9, 0.999)
Temperature	0.07
Peak learning rate	1e-4
Learning rate schedule	Cosine
Warmup steps	100
Epochs	50

Table 2. **Hyperparameters used in visual language pretraining**. 8 × 80GB NVIDIA A100 GPUs were used for training. The maximum sequence length for captions is set to 128.

Hyperparameter	Value
Weight decay	1e-5
AdamW β	(0.9, 0.999)
Learning rate	1e-4

Table 3. **Hyperparameters used in weakly supervised baselines**. Each experiment is performed on a single 24GB NVIDIA 3090 GPU.

[5] (CTP). Due to the relatively small size of the TCGA datasets ($\sim$ 1000 slides or fewer for each task), we follow the study design of previous works by performing 5-fold Monte Carlo cross-validation (CV). In each fold, each dataset is randomly partitioned at the patient level into training (80%), validation (10%) and testing (10%), stratified by the class label. The validation set is used to early stop training and model selection and performance on the test set is reported (Appendix C.4). Each model is trained for a minimum of 20 epochs and an early stopping patience of 10 epochs based on the balanced accuracy measured on the validation set. The 5 models trained using 5-fold CV are then evaluated on our independent, in-house datasets described in **Section 4.2**. For experiments using 1% and 10% of training labels, in each of 5 folds, we perform stratified sampling to select 1% and 10% of the full training set as the new training sets. All results are included in Figure 3 and Figure 4.

C. Additional zero-shot transfer details

C.1. Prompt pools

To evaluate zeroshot transfer, we sampled 50 sets of prompts using a predetermined prompt pool for each task. Each task draws from a unique pool of class names while the pool of templates are shared across tasks. A class name and the template collectively forms the prompt. For the templates, we have:

- `CLASSNAME.`
- `a photomicrograph showing CLASSNAME.`
- `a photomicrograph of CLASSNAME.`
- `an image of CLASSNAME.`
- `an image showing CLASSNAME.`
- `an example of CLASSNAME.`
- `CLASSNAME is shown.`
- `this is CLASSNAME.`
- `there is CLASSNAME.`
- `a histopathological image showing CLASSNAME.`
- `a histopathological image of CLASSNAME.`
- `a histopathological photograph of CLASSNAME.`
- `a histopathological photograph showing CLASSNAME.`
- `shows CLASSNAME.`
- `presence of CLASSNAME.`
- `CLASSNAME is present.`

The `CLASSNAME` is replaced by a sampled class name. The class names for each task are presented in Table 4. For each of 50 prompts, we sample a random number of templates and ensemble them in the embedding space in the same manner as performed by CLIP [3].

C.2. Additional results to ablation studies

Training data comparison. To assess the added value of our image-text pairs, we trained our best performing model configuration (CTP + HistPathGPT) on our full training dataset and compared to training only on ARCH (7,562 pathology pairs), which is a subset of our training data (33,480 pathology pairs). We find that for all pooling methods, training on our full dataset performs better than training on ARCH only. This trend did not change significantly across different pooling methods with MI-Zero. Results are presented in Table 5.

Image encoder pretraining. To assess the benefit of pretraining the image encoder, we compare our best performing model with a variation that uses ViT-S pretrained on ImageNet as well as starting with entirely randomly-initialized weights. We find that pretraining the image encoder and the text encoder on in-domain data performs the best across all 3 tasks. The same trend is maintained across different pooling methods. We report full results in Table 6.

Locked-image tuning. We assess whether locked image tuning (LiT) [6] improves performance on zeroshot transfer by freezing all parameters in the pretrained image encoder when performing visual-language pretraining. Results are reported in Table 7. We find that LiT improves performance slightly when the text encoder is pretrained on in-domain corpora (HistPathGPT), but significantly degrades performance otherwise. We hypothesize this is because when the image and text encoders are pretrained on different domains, the domain gap between the image and text latent space makes it challenging for the model to align via a linear projection and additional parameters and non-linearity might need to be needed.

C.3. Visualization of similarity scores

We randomly select a slide from each in-house test set and visualize patches with highest and lowest similarity scores (Figure 5 and Figure 6). A board-certified pathologist confirmed that relevant morphological patterns are selected with high similarity scores, which drive the model's zero-shot slide-level predictions in the case of topK pooling.

C.4. Additional zero-shot transfer on TCGA

As introduced in the main paper, we present additional results on the TCGA counterparts of our independent test set tasks (BRCA, NSCLC, and RCC). For each of the 3 cancer subtyping classification tasks, we preprocesses the WSIs the same way as presented in the main paper and use MI-Zero for zero-shot transfer. Classification results comparing several setups are summarized in Table 8 and boxplots showing the performance distribution of each model on the set of 50 sampled prompts can be found in Figure 7. Overall, we observe consistent trends between our independent test set results and TCGA results.

D. Runtime analysis

D.1. Comparing runtime of MI-Zero against ABMIL

We expect MI-Zero inference to outperform SOTA methods that utilize learned attention operators (typically require a forward pass through a multi-layer neural network), after patch embeddings are extracted, MI-Zero only requires a single linear projection to project them into the shared visual-language latent space and computing the cosine similarity scores between each patch and each class prompt (which can be efficiently implemented using matrix multiplication). Note while MI-Zero also needs to build the zero-shot classifier by embedding the class prompts using the text encoder and projecting them into the shared latent space, this step only needs to be computed once for each task and prompt, and can be cached for future use (i.e., the cost is amortized across all samples).

Using the in-house BRCA subtyping dataset (average bag size is 8767.42 patches per WSI), we benchmarked the run time of topK pooling MI-Zero, which we found to perform consistently well across all tasks. When not considering IO, MI-Zero inference took between 0.70 to 0.75 ms per WSI depending on the K used, which is nearly 2x the speed of ABMIL inference at 1.4 ms per WSI, in line with our expecation. We found on average, building the zero-shot classifier using the text encoder took around 70 ms per prompt (including time to ensemble multiple templates), which represents an additional amortized cost of 0.35 ms per WSI for our test set of 200 WSIs. Lastly, we note that in the current workflows of embedding-based MIL inference, IO (i.e. reading the embeddings from disk to memory and transferring the tensors to the GPU) still represents the primary bottleneck, which took nearly 20ms per WSI on our workstation despite using a fast SSD.

References

[1] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. 1

[2] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1

[3] 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. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[4] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019. 1

[5] 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*, 81:102559, 2022. 2

[6] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2

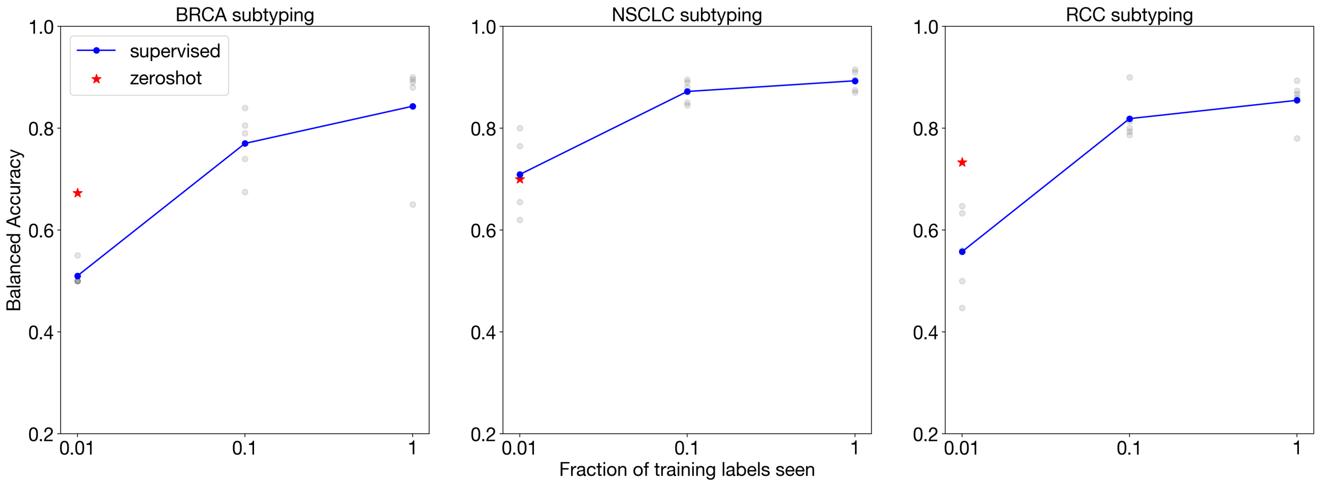

Figure 3. **ABMIL supervised baseline and MI-Zero zeroshot performance on the independent in-house datasets.** The topK pooling variant of MI-Zero is shown. For varying label fractions, each ABMIL model from the 5-fold CV run is represented by a gray dot and the 5-fold average is represented by the blue dot.

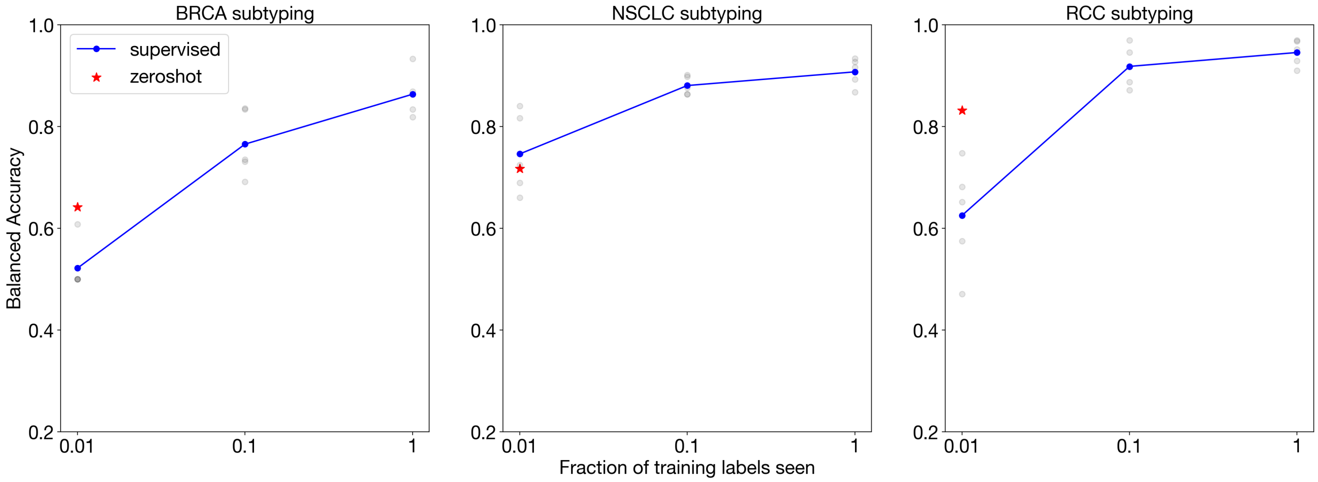

Figure 4. **ABMIL supervised baseline and MI-Zero zeroshot performance on TCGA test sets.** The topK pooling variant of MI-Zero is shown. For varying label fractions, each ABMIL model from the 5-fold CV run is represented by a gray dot and the 5-fold average is represented by the blue dot.

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

Table 4. **Class name pools** for each class in each task. Sampled subsets are substituted into sampled templates to form prompts.

Dataset	SS	Pooling	BRCA	NSCLC	RCC	Average
ARCH	✗	topK	0.625	0.593	0.540	0.586
Ours			**0.672**	**0.700**	**0.733**	**0.702**
ARCH	✓	topK	**0.635**	0.607	0.523	0.589
Ours			0.615	**0.705**	**0.733**	**0.684**
ARCH	✗	mean	**0.655**	0.515	0.533	0.568
Ours			0.620	**0.590**	**0.633**	**0.614**
ARCH	✓	mean	**0.650**	0.518	0.530	0.566
Ours			0.625	**0.590**	**0.637**	**0.617**

Table 5. **Training data comparison**. Balanced accuracies on in-house independent test sets are shown. To assess the added value of our image-text pairs, we trained our best performing model configuration (CTP + HistPathGPT) on our full training dataset and compared to training only on ARCH (7,562 pathology pairs), which is a subset of our training data (33,480 pathology pairs).

Image Encoder	Image Pretraining	Text Pretraining	SS	Pooling	BRCA	NSCLC	RCC	Average
CTP	SSL	In-domain	✗	topK	**0.672**	**0.700**	**0.733**	**0.702**
ViT-S	SSL	In-domain			0.617	0.625	0.673	0.639
ViT-S	ImageNet	In-domain			0.660	0.525	0.600	0.595
CTP	None	None			0.535	0.520	0.297	0.451
ViT-S	None	None			0.500	0.510	0.290	0.433
CTP	SSL	In-domain	✓	topK	0.615	**0.705**	**0.733**	**0.684**
ViT-S	SSL	In-domain			0.625	0.603	0.653	0.627
ViT-S	ImageNet	In-domain			**0.650**	0.512	0.657	0.606
CTP	None	None			0.528	0.527	0.313	0.456
ViT-S	None	None			0.495	0.495	0.310	0.433
CTP	SSL	In-domain	✗	mean	**0.620**	**0.590**	**0.633**	**0.614**
ViT-S	SSL	In-domain			0.590	0.515	0.543	0.549
ViT-S	ImageNet	In-domain			0.615	0.510	0.580	0.568
CTP	None	None			0.535	0.560	0.397	0.497
ViT-S	None	None			0.497	0.520	0.317	0.445
CTP	SSL	In-domain	✓	mean	**0.625**	**0.590**	**0.637**	**0.617**
ViT-S	SSL	In-domain			0.590	0.515	0.540	0.548
ViT-S	ImageNet	In-domain			0.615	0.510	0.573	0.566
CTP	None	None			0.542	0.562	0.393	0.499
ViT-S	None	None			0.495	0.520	0.320	0.445

Table 6. **Pretraining comparison**. Balanced accuracy of different pretraining configurations across all 3 tasks on in-house independent test set. All configurations above use HistPathGPT as the text encoder. **SS**: spatial smoothing.

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

Table 7. **Locked-image tuning**. Balanced accuracy of models using locked-image tuning across all 3 tasks on in-house independent test set. All configurations above use CTP as the image encoder. **LIT**: locked-image tuning, **SS**: spatial smoothing.

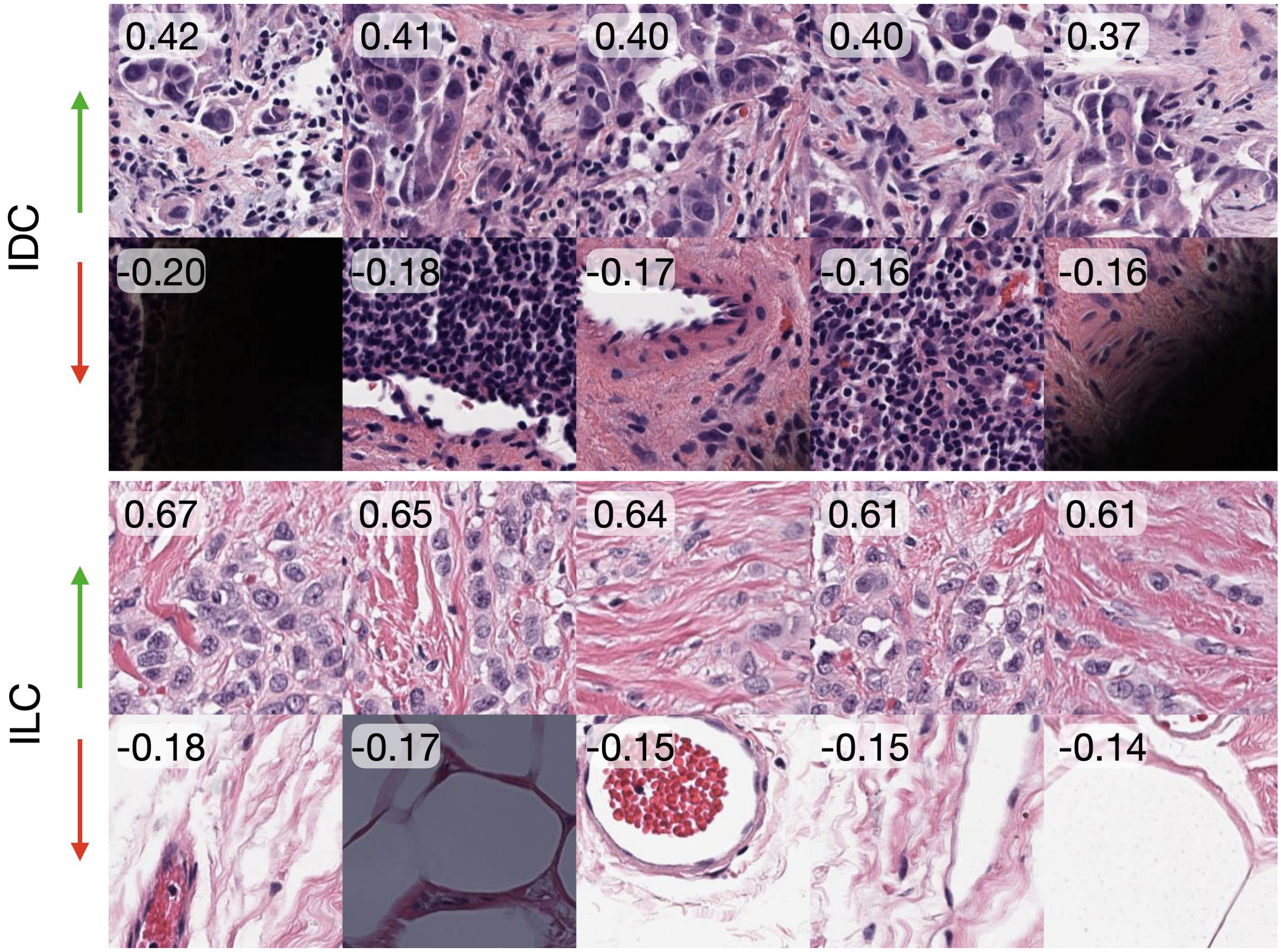

Figure 5. **Visualization of similarity scores for BRCA subtyping.** A WSI of each BRCA subtype (IDC, ILC) is randomly selected from the in-house independent test set, and patches are ranked by their cosine similarity score with the class prompt embedding. The top (highest similarity scores) and bottom (lowest similarity scores) patches are displayed for each WSI. A board-certified pathologist confirms relevant morphological patterns to each class embedding are selected by MI-Zero (high similarity scores). In IDC, high scores correspond to moderate grade tumor cells forming nests and abortive glands, typical of invasive ductal carcinoma of the breast. Low scores picked up instances covered by pen markings, connective tissue, and lymphocyte aggregates with no tumor cells present. In ILC, high scores correspond to low grade tumor cells invading as individual cells or as single files, typical of invasive lobular carcinoma of the breast. Low scores picked up instances of connective and adipose tissue with no tumor cells present.

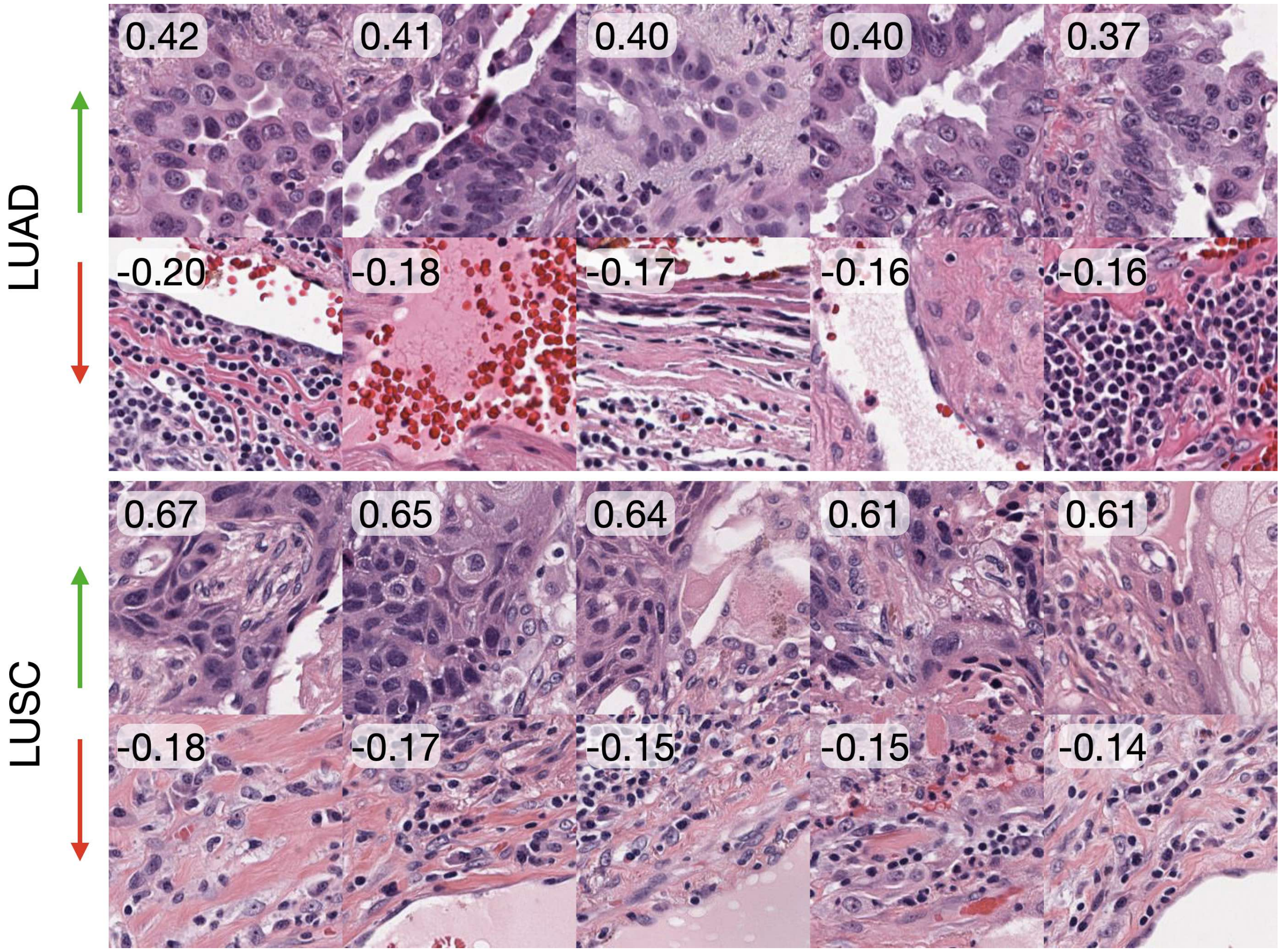

Figure 6. **Visualization of similarity scores for NSCLC subtyping.** A WSI of each NSCLC subtype (LUAD, LUSC) is randomly selected from the in-house independent test set, and patches are ranked by their cosine similarity score with the class prompt embedding. The top (highest similarity scores) and bottom (lowest similarity scores) patches are displayed for each WSI. A board-certified pathologist confirms relevant morphological patterns to each class embedding are selected by MI-Zero (high similarity scores). In LUAD, high scores correspond to tumor cells forming nests and lining spaces with prominent nucleoli, characteristic of adenocarcinoma. Low scores picked up connective tissue, lymphocytes, and blood. In LUSC, high scores correspond to aggregates of tumor cells with keratinization and intercellular bridges characteristic of squamous cell carcinoma. Low scores picked up connective tissue, muscle, and inflammatory cells.

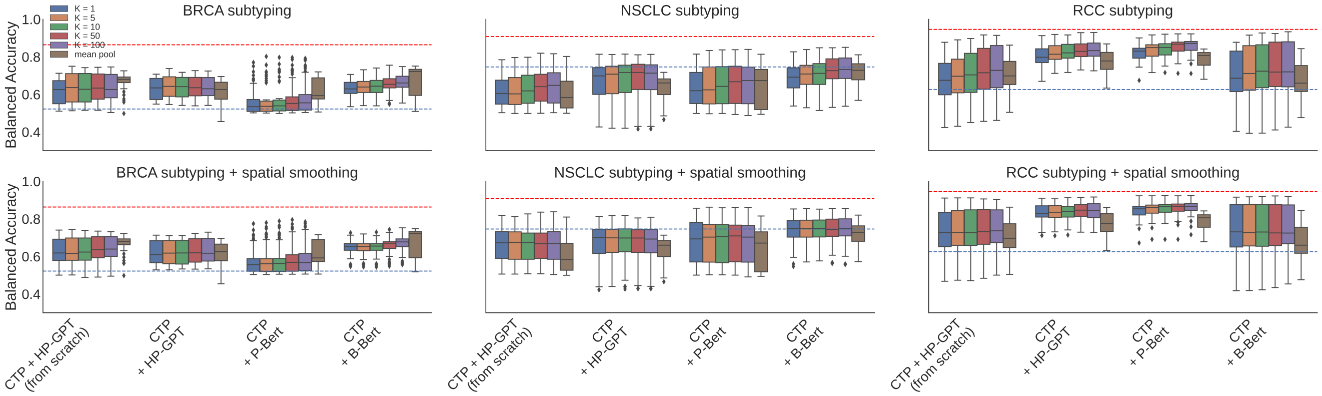

Figure 7. **Zero-shot transfer** performance of selected model configurations on TCGA subsets. Boxplots show distribution of 5-fold averaged model subsets performance for 50 randomly sampled prompts. Columns show different subtyping tasks, rows show the absence or presence of spatial smoothing before pooling, and colors within each boxplot group show pooling methods (K indicates the number of patches selected by topK pooling). Red dashed line shows balanced accuracy of ABMIL trained on 100% of the corresponding TCGA cancer subsets averaged across 5 folds. Blue dashed line shows ABMIL performance trained on 1% of training data instead. **HP-GPT**: HistoPathGPT, **P-Bert**: PubMedBert, **B-Bert**: BioClinicalBert.

Model	Text Encoder & Pretraining	SS	Pooling	BRCA	NSCLC	RCC	Average
ABMIL (1% Data)	None	✗	attention	0.522	0.746	0.625	0.631
ABMIL (100% Data)	None			0.863	0.907	0.946	0.905
MI-Zero (Ours)	HistPathGPT (None)	✗	topK	0.636	0.647	0.728	0.670
	HistPathGPT (In-domain)			0.642	0.717	0.832	**0.730**
	PubmedBert (Out-of-domain)			0.554	0.673	**0.871**	0.700
	BioclinicalBert (Out-of-domain)			**0.660**	**0.732**	0.723	0.705
MI-Zero (Ours)	HistPathGPT (None)	✓	topK	0.639	0.675	0.736	0.683
	HistPathGPT (In-domain)			0.619	0.701	0.846	**0.722**
	PubmedBert (Out-of-domain)			0.568	0.709	**0.867**	0.715
	BioclinicalBert (Out-of-domain)			**0.677**	**0.749**	0.731	0.719
MI-Zero (Ours)	HistPathGPT (None)	✗	mean	0.680	0.582	0.698	0.653
	HistPathGPT (In-domain)			0.626	0.660	0.777	0.688
	PubmedBert (Out-of-domain)			0.593	0.673	**0.806**	0.691
	BioclinicalBert (Out-of-domain)			**0.721**	**0.728**	0.659	**0.703**
MI-Zero (Ours)	HistPathGPT (None)	✓	mean	0.680	0.583	0.697	0.653
	HistPathGPT (In-domain)			0.625	0.660	0.776	0.687
	PubmedBert (Out-of-domain)			0.592	0.671	**0.806**	0.689
	BioclinicalBert (Out-of-domain)			**0.722**	**0.728**	0.660	**0.703**

Table 8. **Additional results on TCGA.** Balanced accuracies are shown. We observe similar trends compared to in-house independent test set results. Pretraining the text encoder on unpaired data yields the better results than no pretraining. Spatial smoothing does not yield consistent improvement. TopK pooling performs better than mean pooling for all models.