
Supplementary information

**A vision–language foundation model for
precision oncology**

In the format provided by the
authors and unedited

A Vision-Language Foundation Model for Precision Oncology

Jinxi Xiang^{1‡}, Xiyue Wang^{1‡}, Xiaoming Zhang², Yinghua Xi¹, Feyisope Eweje¹, Yijiang Chen¹, Yuchen Li¹, Colin Bergstrom³, Matthew Gopaulchan¹, Ted Kim¹, Kun-Hsing Yu⁴, Sierra Willens³, Francesca Maria Olguin³, Jeffrey J. Nirschl², Joel Neal³, Maximilian Diehn¹, Sen Yang^{*1}, Ruijiang Li^{*1,5}

¹*Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA, USA.*

²*Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA*

³*Department of Medicine (Oncology), Stanford University School of Medicine, Stanford, CA, USA*

⁴*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*

⁵*Stanford Institute for Human-Centered Artificial Intelligence, Stanford, CA, USA*

[‡] *equal contribution*

** Correspondence to Ruijiang Li (rli2@stanford.edu) and Sen Yang (seny@stanford.edu).*

1 Supplementary Results

Melanoma Relapse Prediction To facilitate the interpretation of the model’s prediction, we generated heatmaps of the attention scores overlaid on the whole slide images (WSIs) to identify features most relevant for relapse prediction (Extended Data Fig. 4e and Supplementary Fig. 6). The following general patterns were observed. For patients predicted to be at high risk of relapse, the images within high-attention areas show the presence of skin ulceration with abundant intratumoral macrophages accompanied by fibrosis, less infiltration of lymphocytes, and brisk mitotic activity. On the other hand, for patients predicted to be at low risk of relapse, high-attention areas demonstrate an intact epidermis without ulceration, abundant lymphocytes, and inconspicuous mitotic activity. These observations indicate that the model could automatically uncover relevant pathological features for predicting relapse.

Ablation Study on Methodology Adaptations In this study, we incorporated several domain-specific and general methodology adaptations to traditional approaches of masked pretraining and contrastive learning. There are four methodology adaptations as summarized below: 1. Pathology-specific image augmentations. We incorporate domain knowledge into the pre-training step, by adding stain augmentation and extracting image patches from multiple Field-of-VIEWS (FoVs). This can improve the model performance compared to applying traditional data augmentation techniques designed for natural images. 2. Pathology image tokenizers for Masked Image Modeling (MIM). Traditional image tokenizers, such as DALL-E¹ and BEiT-2², are trained from natural images. Since image tokenizers build up the learning targets of MIM, using natural image tokenizers is sub-optimal and may degrade model performance. For our MUSK model, we pretrain pathology-specific image tokenizers. 3. Fine-grained image-text decoder. Traditional image-text contrastive learning loss only considers the coarse alignment of class tokens at the *global* level. In our study, we incorporate an image-text decoder as an auxiliary loss to achieve fine-grained alignment of image patch tokens and language tokens at the *local* scale. 4. Bootstrapped contrastive learning. To tackle the problem of noisy image-text pairs (which is common for web-scraped datasets), we adopt a bootstrap approach for contrastive learning. We initially train a baseline model on the full Quilt1M dataset, and then filter out low-similarity image-text pairs based on this model. The final model is trained on the refined dataset with improved data quality.

We conducted ablation studies to evaluate the effects of various methodology adaptations. The training configurations are detailed in Supplementary Table 16. We use the “Base” settings in this ablation study. The models were evaluated across multiple benchmarks, including linear probe classification on 12 datasets, few-shot classification on 12 datasets, zero-shot classification on 6 datasets, VQA on the PathVQA dataset, image-text retrieval on BookSet, and image retrieval on 2 datasets. As shown in Supplementary Table 13 and Supplementary Fig. 11, BEiT-3 (without any proposed adaptations) consistently underperformed across all

tasks compared to MUSK. In the following sections, we break down these adaptations and investigate the effect of each adaptation.

Ablation 1: Pathology-Specific Image Augmentations. We compared two models: one incorporating pathology-specific image augmentations during mask pretraining and a baseline model without these augmentations. The baseline model used standard augmentations: random vertical flip ($p = 0.5$), color dropping ($p = 0.2$) to convert images to grayscale, weak color jittering ($p = 0.8$) with specific brightness, contrast, saturation, and hue adjustments, and all patches trained at $20\times$ magnification. The pathology-specific model added RandStainNA³ and FoVs⁴, i.e., mixed magnifications of $10\times$, $20\times$, and $40\times$. Both models were then trained using image-text contrastive learning and evaluated on benchmark tasks. As shown in Ablation 1 of Supplementary Table 14, the pathology-specific augmentations led to consistent improvements across all benchmarks. Vision tasks such as linear probe classification improved by $+2.64\%$, 10-shot classification by $+2.03\%$, and I2I retrieval by $+1.68\%$. Vision-language tasks, including zero-shot classification, VQA, and T2I/I2T retrieval, also saw consistent gains. These enhancements are attributed to the more realistic pathology images generated by stain augmentation and the multi-resolution training that simulates the hierarchical pathology images.

Ablation 2: Pathology Image Tokenizer for MIM. We trained two models: one using an ImageNet-pretrained BEiT-v2 as the image tokenizer and the other using a TCGA-pretrained BEiT-v2. Both models underwent image-text contrastive learning and were evaluated on the same benchmarks. Ablation 2 of Supplementary Table 14 reveals that the pathology-specific image tokenizer consistently outperformed the ImageNet-pretrained tokenizer. Linear probe classification increased by $+1.69\%$, 10-shot classification by $+1.66\%$, and I2I retrieval by $+1.01\%$. Vision-language tasks, including zero-shot classification, VQA, and T2I/I2T retrieval, also showed slight improvements. We attribute these gains to the pathology awareness embedded in the pathology-specific tokenizer, as visualized in the image reconstructions during mask pretraining.

Ablation 3: Fine-Grained Multimodal Decoder. We trained two models, both initialized from the same mask-pretrained model. One model was further trained using only contrastive loss, while the other incorporated both contrastive loss and mask decoder loss. Their performances were then compared on the benchmarks. Ablation 3 of Supplementary Table 14 demonstrates that the multimodal decoder yielded significant improvements in vision-language tasks, with zero-shot classification increasing by $+1.13\%$, VQA by $+1.59\%$, T2I retrieval by $+2.29\%$, and I2T retrieval by $+2.94\%$. Vision tasks such as linear probe classification, 10-shot classification, and I2I retrieval remained mostly unchanged. This suggests the multimodal decoder enhances fine-grained fusion between image and text data, leading to improvements in vision-language tasks.

Ablation 4: Bootstrapped Contrastive Learning. We first trained a baseline model using the original

Quilt1M for contrastive learning. This model was then used to score image-text pairs in Quilt1M, filtering out noisy samples with low similarities. The control model was retrained using the filtered data for image-text contrastive learning. Finally, we compared the performance of the baseline model with the bootstrapped model on the benchmark tasks. As shown in Supplementary Table 14, bootstrapped contrastive learning improved vision-language tasks, with zero-shot classification increasing by +2.20%, VQA by +2.81%, T2I retrieval by +6.33%, and I2T retrieval by +4.88%. Vision tasks showed slight improvements, including linear probe, 10-shot classification, and I2I retrieval. These results confirm the notion that eliminating noisy image-text pairs through bootstrap learning enhances vision-language alignment.

In summary, we employed four domain-specific and general methodology adaptations to traditional masked pretraining and contrastive learning framework. Through a series of ablation studies, we established that all of the four adaptations are essential for achieving high performance for precision oncology applications.

Ablation Study on Masked Pretraining We conducted a series of *controlled* ablation experiments to investigate the impact of masked pretraining, data distributions, data scale, and model capacity on the model performance. The ablation study includes five models pretrained on different datasets during the first unified masked pretraining phase. Still, all were trained on the same Quilt1M dataset during the second contrastive learning pretraining phase. Regarding the pretraining data used in the first phase, `model-1` had no additional masked pretraining and relied on randomly initialized weights. `model-2` was trained on 15 million natural images from ImageNet-21k and 500 million natural language tokens. `model-3` used the Quilt1M for unified masked pretraining. `model-4` employed 15 million pathology images from TCGA and 500 million pathology text tokens. Finally, `model-5` was trained on 50 million pathology images from TCGA and 1 billion pathology text tokens, which is referred to as the MUSK (base version).

It is important to note that in these ablation experiments, all model training configurations are maintained to be the same except for the parameter under investigation (training details are provided in Supplementary Table 16). We evaluated the different models on a variety of benchmarks, including linear probe classification on 12 datasets, few-shot classification on 12 datasets, zero-shot classification on 6 datasets, VQA on the PathVQA dataset, image-text retrieval on BookSet, and image retrieval on 2 datasets. Supplementary Table 15 lists the average performance metrics across benchmarks for different models.

Masked Pretraining. The impact of the unified masked pretraining phase is illustrated by the differences in performance between `model-1` and `model-5` shown in Supplementary Table 15. `model-1` is trained without the first unified masked pretraining phase, relying solely on the second pretraining on image-text pairs

using contrastive learning. In contrast, `model-5` represents the full MUSK (base version), leveraging both unified masked pertaining and contrastive learning pretraining phases.

The performance improvements from `model-1` to `model-5` are substantial across all evaluated tasks. In absolute terms, the linear probe classification accuracy increases by 14.66%, 10-shot classification by 19.41%, zero-shot classification by 32.04%, VQA by 17.90%, text-to-image retrieval by 67.98%, image-to-text retrieval by 66.69%, and image-to-image retrieval by 23.18%. The large performance gaps are not surprising. `model-1` is grossly underfitted and therefore is unable to learn robust image and text representations based on only 1 million image-text pairs. In contrast, the extensive pretraining of `model-5` enables it to develop more effective and generalizable features, resulting in improved performance across all tasks.

Data Distribution. The impact of data distribution is illustrated by the differences in performance between `model-2` and `model-4` in Supplementary Table 15. Specifically, in the unified masked pertaining phase, `model-2` utilizes ImageNet-21k (around 15 million images) and 500 million text tokens from BookCorpus, OpenWebText, and CC-News. In contrast, `model-4` employs 15 million pathology images from TCGA and 500 million text tokens from the PubMed corpus. Both models are trained using the same scale of datasets and the same training recipe, with the only variable being the data distribution (natural images/text vs. pathology images/text).

We observe that `model-4` outperforms `model-2` across all tasks. Specifically, with a pathology-related training dataset, `model-4` improves linear probe classification accuracy by 0.87%, 10-shot classification by 1.97%, zero-shot classification by 4.41%, VQA by 2.79%, text-to-image retrieval by 16.52%, image-to-text retrieval by 14.73%, and image-to-image retrieval by 2.49%. The advantage of `model-4` over `model-2` is most remarkable in vision-language tasks, such as image-text retrieval. Interestingly, the image representations learned from ImageNet-21k can be effectively transferred to certain pathology tasks, as evidenced by the decent performance of `model-2` in the linear probe and 10-shot classification. This is partly because the low-level features of natural images and pathology images share some commonalities. Despite the significant domain gap between natural and pathology images, the subsequent contrastive learning on Quilt1M bridges this gap to some extent.

Data Scale. We investigated the impact of data scale through three models: `model-3`, `model-4`, and `model-5`. In the unified masked pertaining phase, `model-3` used only Quilt1M, `model-4` used 15 million TCGA pathology image patches and 500 million PubMed text tokens, and `model-5` used 50 million TCGA pathology image patches plus 1 billion PubMed text tokens (Supplementary Table 15)

Overall, we observe that increasing the scale and diversity of data consistently boosts model performance. With approximately one million image-text pairs in Quilt1M, `model-3` demonstrates unsatisfactory performance, especially for challenging tasks like image-text retrieval, achieving only 25.01% and 22.22% at *Recall@50* on Bookset. This result is worse than all other models, except for `model-1`, which did not conduct masked pretraining at all. For `model-4`, the data scale for masked pretraining increases significantly to 15M image patches and 500 million text tokens, resulting in significant improvements across all benchmarks. Most notably, the improvement for zero-shot classification was 11.01%, text-to-image retrieval 49.18%, image-to-text retrieval 49.69%. The best performance is achieved by `model-5`, which utilizes all available data in our study, including 50 million pathology image patches and 1 billion text tokens. We also note that the performance gains achieved with 50 million pathology images over 15 million pathology images are largely incremental for evaluated tasks.

Model Capacity. We have investigated the impact of MUSK model capacity on the benchmark performance. For this analysis, we trained three different versions of MUSK of varying model sizes: small (72 million parameters, 15G FLOPS), base (222 million parameters, 64G FLOPS), and large (675 million parameters, 222G FLOPS). The model configurations and parameters are described in Supplementary Table 16. All three models are trained with the same pretraining dataset (50M TCGA image patches plus 1 billion text tokens for stage 1 mask pretraining; Quilt1M for stage 2 image-text contrastive learning) and evaluated on the same downstream benchmarks. The downstream benchmarks include linear probe classification on 12 datasets, few-shot classification on 12 datasets, zero-shot classification on 6 datasets, VQA on the PathVQA dataset, image-text retrieval on Bookset, and image retrieval on 2 datasets.

As shown in Supplementary Table 17 and Extended Data Fig. 8d, larger models consistently lead to higher performance. Notably, `MUSK-base` shows substantial improvements over `MUSK-small` in all benchmarks. In detail, `MUSK-base` improves linear probe classification accuracy by 5.86%, 10-shot classification by 6.53%, zero-shot classification by 8.79%, VQA by 6.15%, text-to-image retrieval by 21.71%, image-to-text retrieval by 18.92%, and image-to-image retrieval by 2.93% over `MUSK-small`. In comparison, the performance gain of `MUSK-large` over `MUSK-base` is much smaller at 1 – 2% for most benchmarks except for image-to-image retrieval at 2.6%. This suggests that the ‘large’ size model is most likely sufficient for learning pathology representations for which the performance has reached a plateau. Further increasing the model size beyond ‘large’ might yield marginal or even negligible improvements. However, the training cost and inference complexity may become prohibitive for practical applications. Also, more pretraining data will be required to properly train larger models. Therefore, in this study, we chose `MUSK-large` as our final model, which achieves high performance at an affordable computational cost.

2 Supplementary Methods

Multimodal Data Curation for Pretraining We used the PubMed Central Open Access Dataset to curate pathology-related journal articles. Searches were conducted on the PubMed open-access database from 2000 to 2023 using keywords such as “pathology”, “histopathology”, “whole-slide image”, “H&E”, and 148 terms from a histopathology glossary¹. We employed Entrez to retrieve the most pertinent articles for each keyword, resulting in 1,001,800 unique articles with PMCIDs. These articles were then autonomously downloaded using the PMCIDs, and the extracted text information was used for model pre-training.

For the pathology image data, we used publicly available hematoxylin and eosin (H&E)-stained WSIs from TCGA, including digital slides of flash-frozen and formalin-fixed paraffin-embedded (FFPE) tissue samples. Flash-frozen tissue sectioning allows better preservation of nucleic acids and proteins, making it suitable for molecular analysis. On the other hand, FFPE tissue sectioning allows excellent preservation of morphology and is widely used in routine diagnostic histopathology. Since it is important to have diverse large-scale data for pretraining foundation models, we utilized all available WSIs from TCGA, including 22,135 frozen and 11,763 FFPE slides from 11,577 unique patients representing 33 tumor types (Supplementary Fig. 1). These WSIs were preprocessed by segmenting the tissue and cropping them into non-overlapping patches, resulting in approximately 50 million image patches, each with a size of 512×512 pixels at $10\times$ magnification. We maintained the original image resolution without resizing to support training with multiple fields of view.

Image Representations. The image representations are obtained by vision Transformers⁵. Given a 2D image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, it is split and reshaped into $N = HW/P^2$ patches and each patch can be represented as $\mathbf{x}^p \in \mathbb{R}^{N \times (P^2 C)}$, where C is number of channels, (H, W) is the resolution of the input image, and (P, P) is the patch resolution. Then, these image patches are input into the image encoder and flattened into vectors, which are further linearly projected to obtain patch embeddings. In addition, we also prepend a learnable token $\mathbf{x}_{[\text{CLS}]}$ to the sequence, which is used as the global embedding of the image patches. Finally, the image representations are obtained via summing patch embeddings, learnable 1D position $\mathbf{X}_{pos} \in \mathbb{R}^{(N+1) \times D}$ (to denote patches position), and image type embeddings $\mathbf{X}_{type} \in \mathbb{R}^D$.

$$\mathbf{H}_0^x = [\mathbf{x}_{[\text{CLS}]}, \mathbf{x}_i^p, \dots, \mathbf{x}_N^p] + \mathbf{X}_{pos} + \mathbf{X}_{type}, \quad (1)$$

where image representations $\mathbf{H}_0^x \in \mathbb{R}^{(N+1) \times D}$; the subscribe 0 of $\mathbf{H}_0^x$ denotes that it is the input embedding.

Text Representations. Following BERT⁶, we tokenize the text to subword units by Word-Piece⁷. A start-of-sequence token $\mathbf{w}_{[\text{BOS}]}$ and a special boundary token $\mathbf{w}_{[\text{SEP}]}$ are added to the text sequence. Text

¹<https://lab-ally.com/histopathology-resources/histopathology-glossary/>

input representation $\mathbf{H}_0^w$ are computed vis summing the corresponding word embeddings, text position $\mathbf{W}_{pos}$, and text type embedding $\mathbf{W}_{type}$:

$$\mathbf{H}_0^w = [\mathbf{w}_{[\text{BOS}]}, \mathbf{w}_i, \dots, \mathbf{w}_M, \mathbf{w}_{[\text{SEP}]}] + \mathbf{W}_{pos} + \mathbf{W}_{type}, \quad (2)$$

where text representations $\mathbf{H}_0^w \in \mathbb{R}^{(M+2) \times D}$, M indicates the length of tokenized words; the subscribe 0 of $\mathbf{H}_0^w$ denotes that it is the input embedding; D is the embedding dimension.

Network Backbone. Each block captures modality-specific information by switching to different modality experts and employs a shared multi-head self-attention (MSA) module across modalities to connect visual and language information flows. Given the previous layer’s output vectors $\mathbf{H}_{l-1}$, $l \in [1, L]$, L represents the number of blocks, the output $\mathbf{H}_l$ is obtained by

$$\begin{aligned} \mathbf{H}'_l &= \text{MSA}(\text{LN}(\mathbf{H}_{l-1})) + \mathbf{H}_{l-1} \\ \mathbf{H}_l &= \text{MoME-FFN}(\text{LN}(\mathbf{H}'_l)) + \mathbf{H}'_l, \end{aligned} \quad (3)$$

where LN is short for layer normalization, $\mathbf{H}'_l$ represent the output from the shared MSA module, MoME-FFN represent the vision expert (V-FFN) or language expert (L-FFN) according to the input modality $\mathbf{H}'_l$.

Downstream Tasks Evaluation MUSK can be used for various downstream tasks with minimal or no further training. We specify the training configurations in this section.

Zero-shot cross-modal retrieval. In this application, MUSK is used as a multimodal encoder that can independently represent images and texts. We calculate the embeddings for all images, denoted as $\mathbf{u}$, and for all texts, denoted as $\mathbf{v}$. The similarity between every possible pair of image-to-text and text-to-image is determined using cosine similarity scores s_i , as expressed in the following equation:

$$s_i = \underset{j}{\operatorname{argmax}} \mathbf{u}_i^T \mathbf{v}_j \quad (4)$$

For text-to-image retrieval, we retrieve the top- K images closest to a given text query in the aligned latent space. Similarly, image-to-text retrieval follows an analogous procedure. We employ the Recall@ K metric to assess MUSK’s retrieval performance. This metric measures the percentage of candidates in the test set for which the ground truth result appears among the top- K retrieved samples. We selected $K \in \{1, 10, 50\}$, commonly used to evaluate retrieval performance.

Visual question answering. Visual Question Answering (VQA) aims to generate responses based on

an image and a corresponding question. In line with previous studies on PathVQA⁸⁻¹¹, it is formulated as a classification problem using unique plain-language answers across 4,092 classes, and we report accuracy on PathVQA for comparative purposes. In this context, MUSK is a multimodal encoder that processes images and texts. The image and question embeddings are concatenated to form a multimodal embedding. Based on this embedding, we construct a linear classifier with 2048 dimensions to predict answers from 4092 possible classes. The classifier is built on top of multimodal features, which are derived from the [CLS] token of the image embedding and the [BOS] token of the text embedding. Both embeddings have a dimension of 1024. We use a linear MLP to predict the result based on these multimodal features. We follow the training set of the PathVQA dataset¹² to optimize the classifier and then evaluate its performance on the test set. We finetune the whole model, including the backbone and the classification head. We use the Adam optimizer with a cosine learning rate scheduler, setting a top learning rate of $lr = 0.001$ with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ to train our model. The training process includes a batch size 64 for 50 epochs, with 4000 warmup steps.

Zero-shot image retrieval. In the first step of image retrieval, we compute the normalized embeddings for all images in the dataset using the MUSK image encoder. The datasets utilized are specifically UniToPatho¹³ and BRCAS¹⁴. For each query image, we calculate the cosine similarity between its embedding and those of all other images in the dataset. Then, we identify the pairs with the highest similarity and designate them as the retrieval results. A retrieval is considered correct when the class label of the retrieved image is identical to that of the query image. Importantly, this task is conducted in a zero-shot manner, meaning it is performed without training on the dataset.

Zero-shot image classification. We implemented a zero-shot classification approach similar to that used by CLIP¹⁵. Rather than using discrete numerical identifiers, the original class names are described using natural language based on well-designed prompts (Supplementary Fig. 3). This approach encodes class names by MUSK to obtain their respective embeddings. We use multiple templates for one class to accommodate description variations and aggregate multiple templates per class by averaging their embeddings. We then generate embeddings for the images and compute the cosine similarity between the image and text embeddings. The class corresponding to the highest image-text similarity is designated as the final classification result.

Supervised image classification. The supervised image classification experiments include linear probes and few-shot classification tasks. We use the vision encoder of MUSK for image representation and employ a lightweight linear classifier for the classification process. The classifier is optimized with the training set of the dataset and then evaluated on the test set, while the image embeddings are extracted from the frozen MUSK. While the test set remains unchanged in our few-shot classification setup, we vary the number of labeled examples per class used during training. These training variations, called ‘shots,’ range from $\{1, 2, 4, 8, 10\}$

labeled examples per class. We resize images to 384×384 for all tasks and normalize using mean and standard deviation parameters. We split the datasets into training and testing sets by stratifying for labels or using the official folds if available.

Slide-level image classification. To further compare MUSK with state-of-the-art pathology foundation models (PLIP¹⁶, UNI¹⁷, GigaPath¹⁸, Virchow¹⁹, and CONCH²⁰), we performed WSI classification for the biomarker prediction. We utilized the vision encoder of MUSK for tile-level image representation and employed the AbMIL²¹ weakly-supervised method for aggregating tile-level features. Additionally, the text encoder of MUSK was used to represent clinical information. Subsequently, the image and text features were fused for slide-level classification.

Benchmark Datasets We conducted downstream tasks using 16 benchmark datasets on multimodal retrieval, visual question answering, and histopathology image classification. We provide detailed data descriptions here.

BookSet²² comprises 4,265 image-text pairs, originally derived from figures and captions extracted from ten textbooks. This study utilized the BookSet dataset to evaluate cross-modal retrieval performance. This dataset and the following PathMMU contain more challenging cases than those used in previous studies with about 2,000 image-text pairs or less^{16,20}

PathMMU²³ is a comprehensive multimodal benchmark dataset with expert-level annotations. To ensure data quality, seven pathologists were recruited to meticulously review each question in the validation and test sets, adhering to strict standards. To prevent information leakage via the training set, Quilt1M, we excluded samples in the validation and test sets that included images from YouTube and PubMed. After thorough data curation, we combined the validation and test sets. Finally, we obtained 7,774 diverse and large-scale image-caption pairs, which were used as the multimodal retrieval benchmark in this study.

PathVQA¹² is a large-scale dataset for pathological visual question answering, containing 32,799 open-ended questions from 4,998 pathology images. These images are similar in style and level of difficulty to those used in the American Board of Pathology (ABP) certification exams. This dataset was created by extracting images and questions from pathology textbooks and the Pathology Education Informational Resource (PEIR) digital library. PathVQA was split into training, validation, and test sets, containing 19,654, 6,259, and 6,719 question-answer pairs, respectively. This dataset was used for evaluation of visual question answering.

PatchCamelyon²⁴ ($n_{class} = 2$) is a histopathology image classification dataset used in a grand challenge. It consists of 327,680 color images (96×96 pixels) extracted from histopathologic scans of lymph

node sections. Each image is annotated with a binary label indicating the presence of metastatic tissue. A positive label indicates that the patch’s center 32×32 pixel region contains at least one pixel of tumor tissue. Tumor tissue in the outer region of the patch does not influence the label. The dataset is divided into training (262,144), validation (32,768), and test sets (32,768).

NCT-CRC-HE-100K²⁵ ($n_{class} = 9$) consists of 107,180 images with 224×224 pixels at $0.5 \mu m$ per pixel (mpp), extracted from 50 patients with colorectal adenocarcinoma. Each image is categorized into one of nine classes: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, or colorectal adenocarcinoma epithelium. The supervised dataset uses the officially provided data splits of 100,000 images in the training set and 7,180 images in the test set. In our zero-shot classification experiment, we only used the official test set.

SICAPv2²⁶ ($n_{class} = 4$) contains 182 prostate histology WSIs, which are further divided into 12,081 (train $n = 9959$, test $n = 2122$) patches of 512×512 pixel. These WSIs are labeled with global Gleason scores, and the image patches are annotated with four classes: non-cancerous and tumors with Grade 3, Grade 4, and Grade 5.

Osteo²⁷ ($n_{class} = 3$) consists of 1,144 histopathology images of osteosarcoma samples, each with dimensions of 1024×1024 pixels at 10x magnification. The images are categorized into three classes: 536 (47%) non-tumor images, 263 (23%) necrotic tumor images, and 345 (30%) viable tumor images. This dataset is used to classify osteosarcoma tissue samples.

RenalCell²⁸ ($n_{class} = 6$) comprises 39,458 histopathology images of clear cell renal cell carcinoma (ccRCC), annotated into six distinct tissue texture types: empty, renal cancer, normal renal, stromal, red blood cells, and other textures. The dataset was designed to cover the tumor border regions and includes annotations from multiple clinical centers to ensure the generalizability of developed models.

SkinCancer²⁹ ($n_{class} = 16$) is a large collection of skin biopsies from 386 cases that were manually annotated on scanned histological slides. A total of 129,364 $100 \times 100 \mu m$ (395×395 pixels) image tiles were extracted and split into a training, validation, and test set. The dataset covers a range of skin conditions, including basal cell carcinoma (BCC), squamous cell carcinoma (SqCC), naevi (moles), and melanoma. In addition to these skin cancer categories, the images have been meticulously annotated for 16 different tissue types and structures, such as chondral tissue, dermis, epidermis, hair follicles, and various glands.

LC25000³⁰ ($n_{class} = 5$) contains 25,000 color histopathological images, each measuring 768×768

pixels in JPEG format, and is categorized into five classes, with each class comprising 5,000 images. The dataset is further divided into two subsets: LC25000 (Lung) ($n_{class} = 3$) with 15,000 and LC25000 (Colon) ($n_{class} = 2$) with 10,000 color images (768×768). The lung subset is labeled as lung adenocarcinomas, lung squamous cell carcinomas, and benign lung tissues, and the colon subset is labeled as colon adenocarcinomas and benign colonic tissues.

PanNuke³¹ ($n_{class} = 2$) is a comprehensive pan-cancer histology dataset that contains over 200,000 annotated nuclei across 19 different tissue types. These nuclei are categorized into five clinically important classes: neoplastic, non-neoplastic epithelial, inflammatory, connective, and dead. Experts semi-automatically annotated and quality-controlled the dataset. This dataset is processed into binary classification (benign versus malignant) according to the PLIP code². This resulted in 2,866 malignant images and 3,368 benign images.

UniToPatho¹³ ($n_{class} = 6$) consists of 9,536 H&E stained image patches extracted from 292 WSIs, designed specifically for the training of deep neural networks in colorectal polyps classification and adenomas grading. Each of these slides from a different patient was captured using a Hamamatsu Nanozoomer S210 scanner at $20\times$ magnification ($0.4415 \mu\text{m}/\text{px}$). Each slide was annotated by expert pathologists according to six classes: NORM (Normal tissue), HP (Hyperplastic Polyp), TA.HG (Tubular Adenoma, High-Grade dysplasia), TA.LG (Tubular Adenoma, Low-Grade dysplasia), TVA.HG (Tubulo-Villous Adenoma, High-Grade dysplasia), and TVA.LG (Tubulo-Villous Adenoma, Low-Grade dysplasia).

WSSS4LUAD³² ($n_{class} = 3$) is a collection of histopathological images ($n = 10,091$) of lung adenocarcinoma (LUAD) tissue samples. It comprises of image tiles ranging in size from 200 to 500 pixels, each categorized into one of three classes: tumor, tumor-associated stroma, or normal tissue. The evaluation was conducted on 2,917 images, using 7,174 images for training.

BRACS¹⁴ ($n_{class} = 6$) is a collection of annotated H&E histopathological images of breast tissue samples, containing a total of 547 WSIs and 4,539 labeled regions of interest (ROIs) from 189 patients. The dataset includes a diverse range of breast lesion types, such as Normal (N)/Pathological Benign (PB), Usual Ductal Hyperplasia (UDH), Flat Epithelial Atypia (FEA), Atypical Ductal Hyperplasia (ADH), Ductal Carcinoma in Situ (DCIS), and Invasive Carcinoma (IC). Furthermore, the dataset can be used as $n_{class} = 3$ by grouping N, PB, and UDH as benign tumors, FEA and ADH as atypical tumors, and DCIS and IC as malignant tumors.

BCNB³³ ($n_{class} = 2$) The Early Breast Cancer Core-Needle Biopsy WSI (BCNB) dataset comprises a total of 1,057 slides. The original authors divided the BCNB into training ($n=629$), validation ($n=210$), and

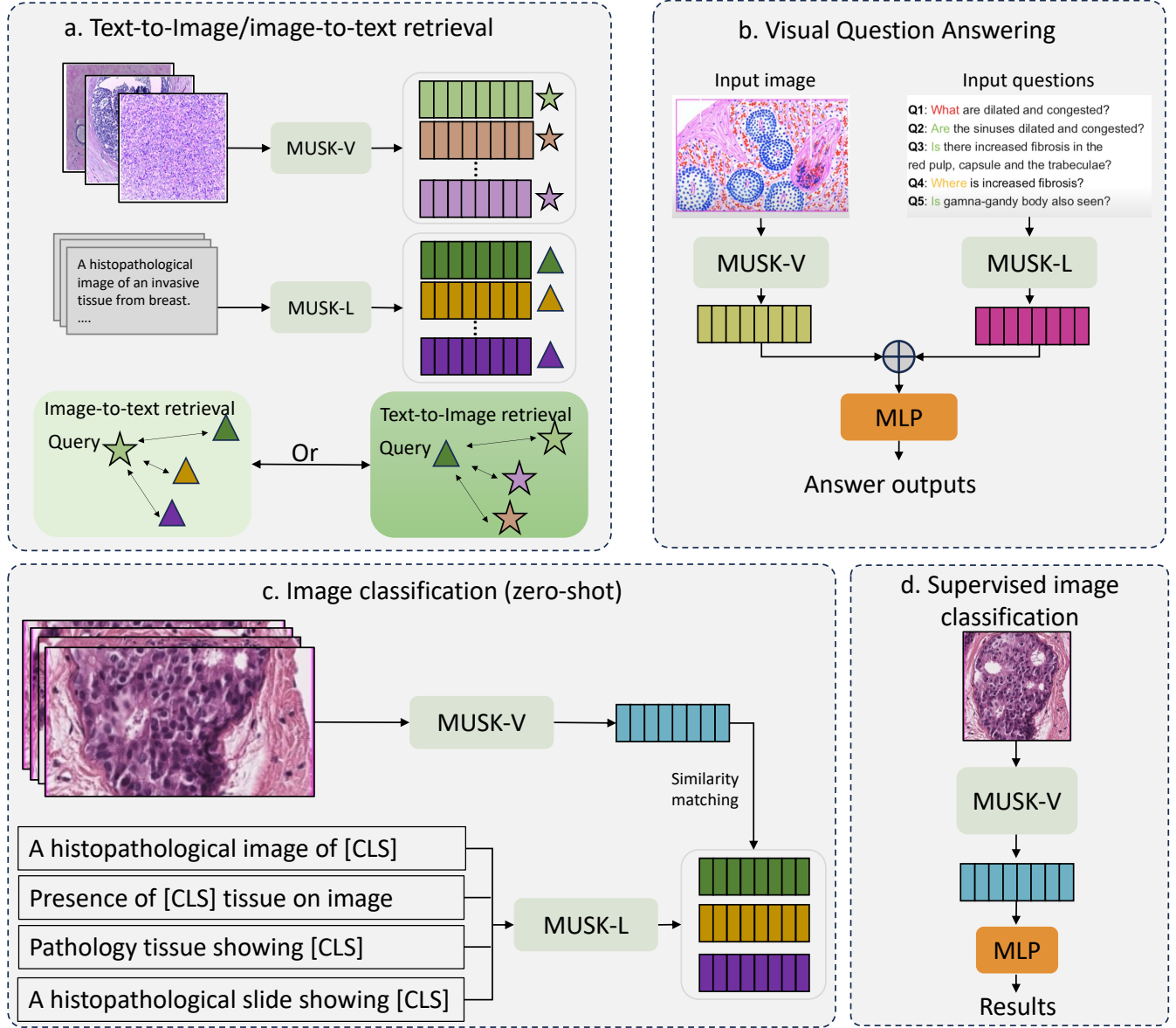
²<https://github.com/PathologyFoundation/plip/tree/main/reproducibility/evaluation>

test set ($n=218$). We developed models for predicting *ER*, *PR*, and *HER2* status using the training set. The best-performing model was selected based on its results on the validation set. Additionally, we incorporated clinical information including age, tumor size, grade, and surgical procedure for multi-modal classification.

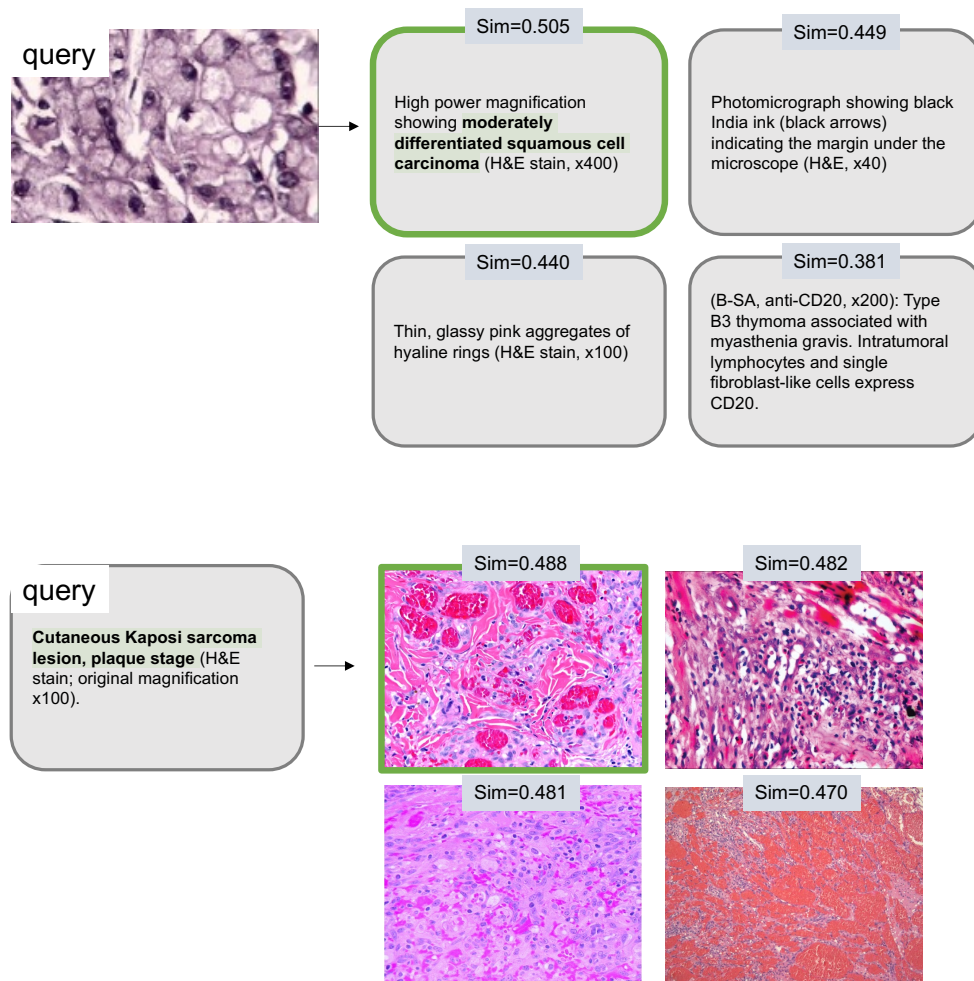
MUV-IDH³⁴($n_{class} = 2$) is used to evaluate models for predicting IDH mutation status in brain tumor patients using H&E-stained pathology images. These images were collected from the Medical University of Vienna (MUV) and included a total of 872 slides (333 IDH-mutant and 539 IDH-wild type). Additionally, we incorporated clinical information including age, sex, grade, and tumor location for multi-modal classification.

3 Computing Hardware and Software

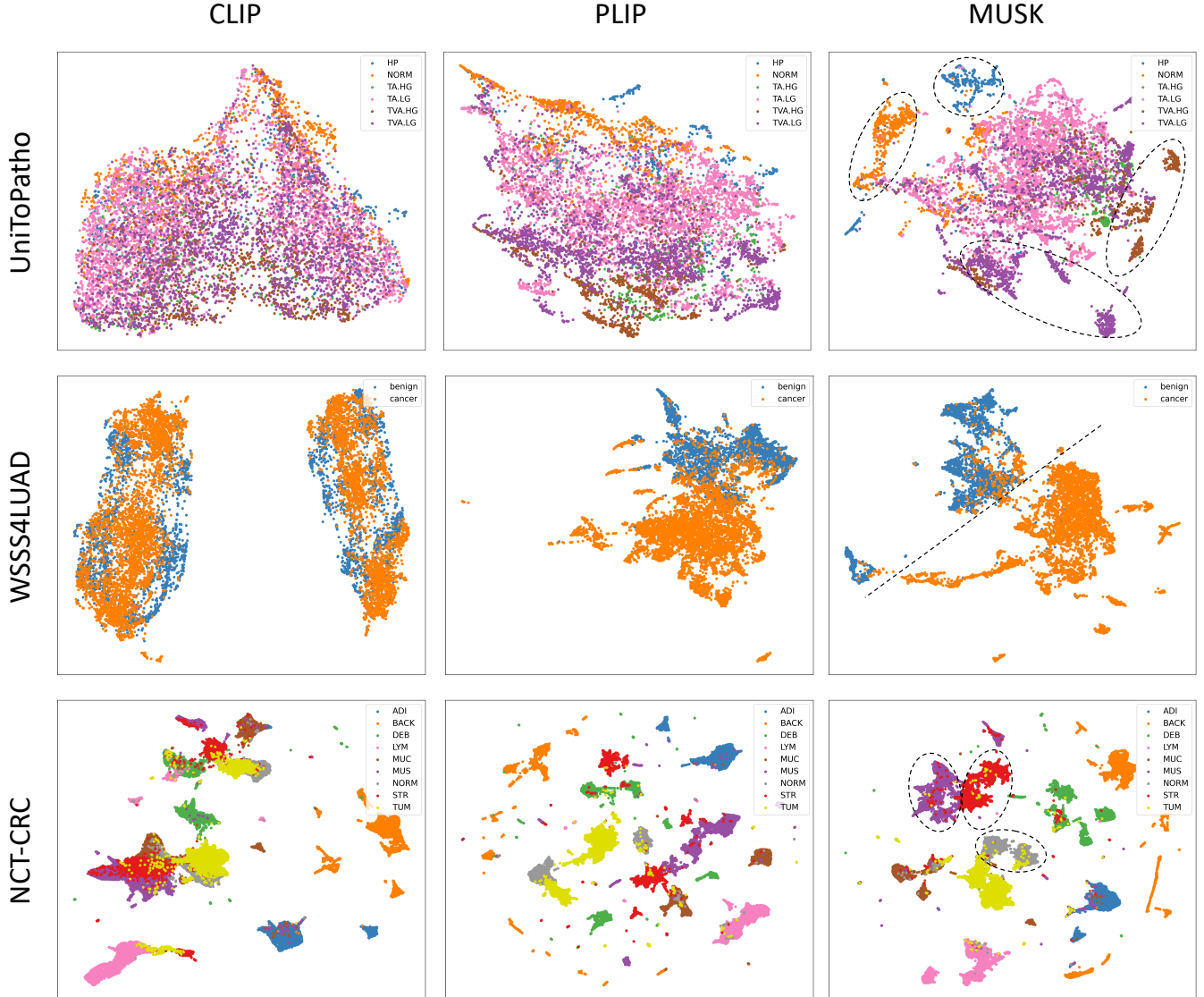
We used `Python` (version 3.10.12) and various open-source libraries for all experiments and analyses. For pretraining, we utilized 64 NVIDIA V100 GPUs each with 32-GB memory using `PyTorch` (version 2.0.1, CUDA 11.0) and the `Accelerator` from the `accelerate` library (<https://huggingface.co/docs/accelerate/en/index>). This setup enabled the `PyTorch` code to run across any distributed configuration for multi-GPU, task-agnostic pretraining. The first pretraining step with masked modeling took about 240 hours using 64 V100 GPUs. The second pretraining step with contrastive learning took about 10 hours using 64 V100 GPUs. All downstream experiments were conducted on a single 48-GB NVIDIA RTX6000 GPU. The multimodal transformer is based on implementations from `torchscale` (<https://github.com/microsoft/torchscale>) and the `BEiT3` library (<https://github.com/microsoft/unilm/tree/master/beit3>). The evaluation of downstream tasks, including VQA, cross-modal retrieval, image-to-image retrieval, and image classification (both zero-shot and supervised), was conducted within the unified framework `CLIP Benchmark` (https://github.com/LAION-AI/CLIP_benchmark). We employed Hugging Face’s `Transformers` (v4.38.1) and `Torchmetrics` (v0.7.3) for evaluation metrics (e.g., accuracy, AUC). The clinical tasks, including melanoma relapse prediction, prognosis, and immunotherapy prediction, were built on `PyTorch Lightning` (v2.2.1) (<https://lightning.ai/docs/pytorch/stable/>), allowing for fast and scalable evaluation with easy configuration. `OpenSlide` (v4.3.1) and `openslide-python` (v1.3.1) facilitated WSI processing, while `Scikit-learn` (v1.5.0) provided machine learning model evaluation metrics and logistic regression models. `Numpy` (v1.26.4), `Pandas` (v2.2.2), `Pillow` (v10.3.0), `OpenCv-python` (v4.9.0.80), `Matplotlib` (v3.9.0), and `Seaborn` (v0.13.2) were utilized for data handling, image processing, and visualization. Implementations of other visual-language models like `PLIP`, `BiomedCLIP`, `OpenAICLIP`, and `Quilt1M` were sourced from the Hugging Face model hub (<https://huggingface.co/models>).



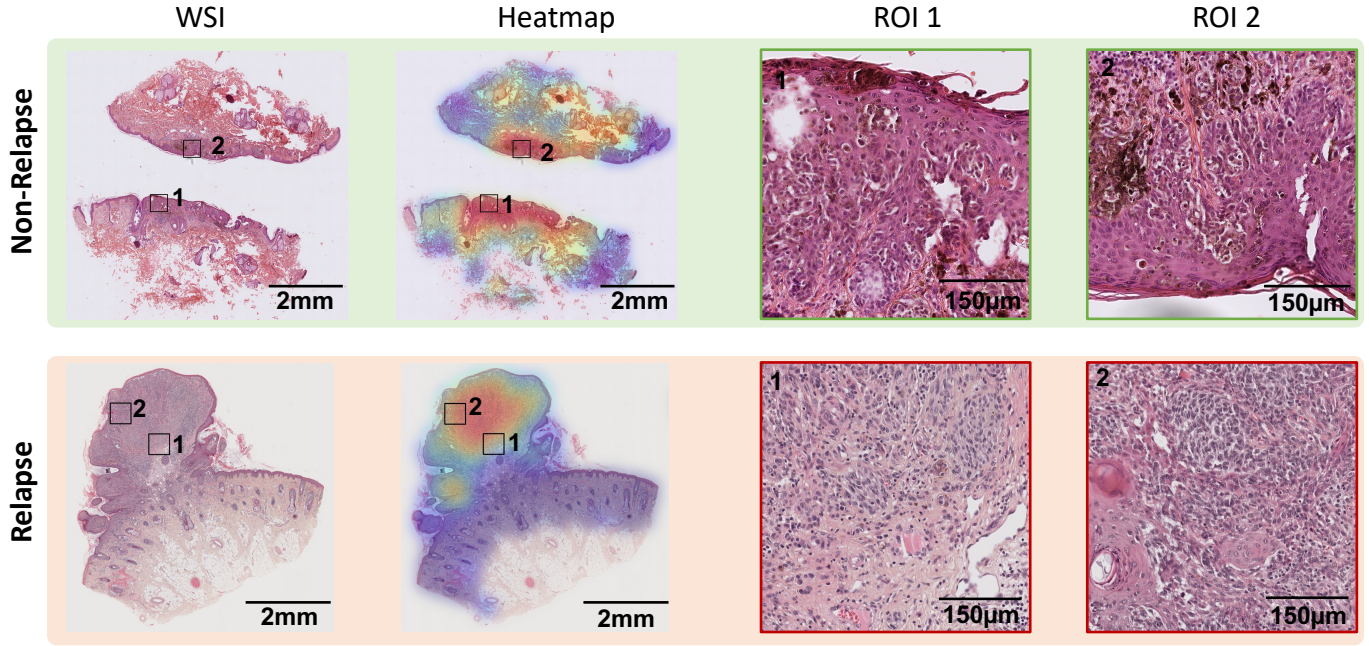
Supplementary Fig. 3. Schematics of four benchmark experiments. **a. Text-to-Image and Image-to-Text Retrieval:** Images and texts are encoded using the visual (MUSK-V) and textual (MUSK-L) encoders of MUSK, respectively, creating two feature databases. A query image or text is encoded to retrieve feature vectors, which are then used to retrieve the nearest cross-modality features in the established databases. **b. Visual Question Answering:** An input image and its corresponding questions are encoded by MUSK-V and MUSK-L to derive image and text representations, respectively. A multilayer perceptron (MLP) fuses and processes these representations to generate the answers (classification results). **c. Zero-shot Image Classification:** This task involves categorizing an image without training. A textual feature database is created by designing four text prompts for each class label and aggregating their embeddings through mean pooling to form a class embedding. The image is encoded by MUSK-V, and its class is determined by computing the cosine similarity between its embedding and the class embeddings. The highest similarity score dictates the classification label. **d. Supervised Image Classification:** Images are initially encoded by MUSK-V, followed by processing through an MLP. This task requires fine-tuning the MLP to achieve accurate classification results.



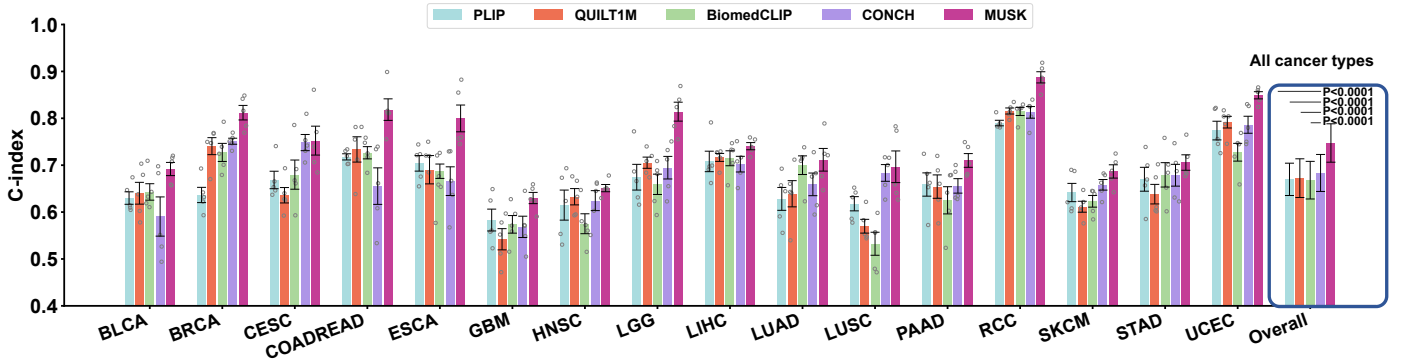
Supplementary Fig. 4. Cross-modal retrieval. Visual examples are shown for the top 4 retrieved candidates. The ground-truth images and text are marked with green boxes, and the image-text and image-text similarity scores are provided.



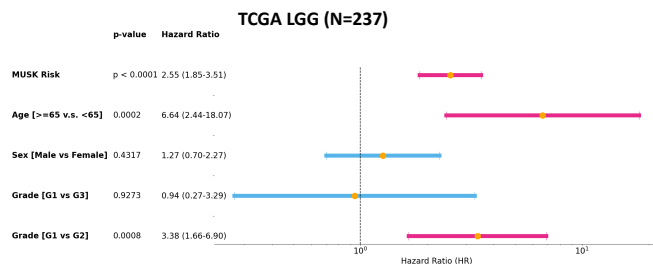
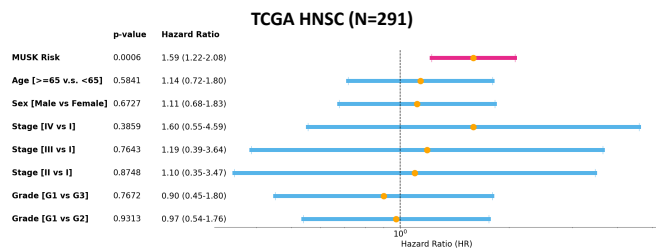
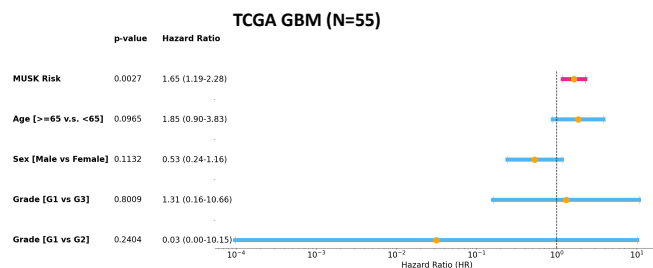
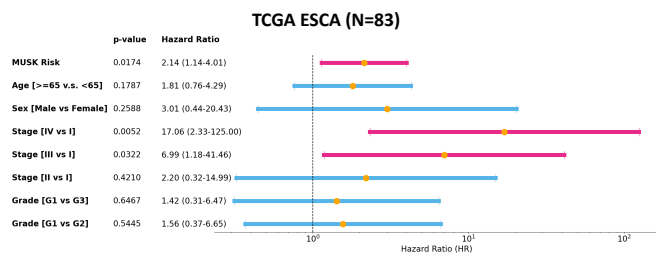
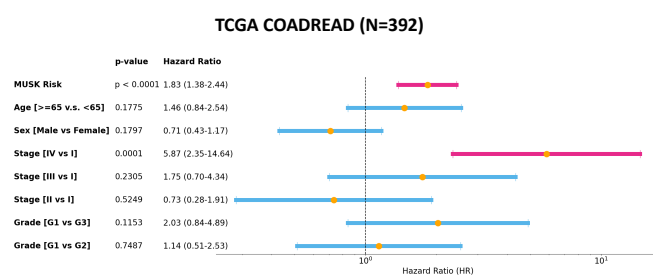
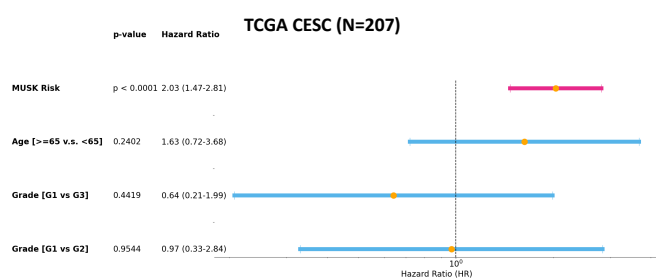
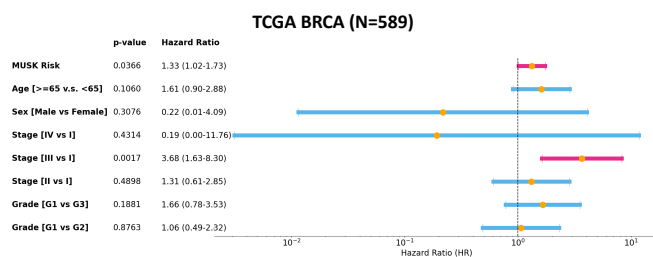
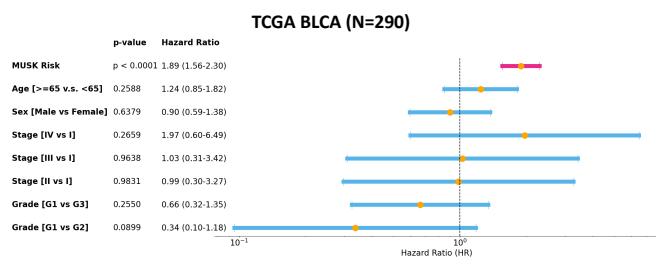
Supplementary Fig. 5. Visualization of image embeddings by the CLIP, PLIP, and MUSK models across three datasets. We employed the UMAP method³⁵ to create the scatter plots. The UniToPatho dataset, known for its complexity, comprises six colorectal polyp types. In this dataset, CLIP and PLIP models failed to form any distinct clusters in the UMAP. Conversely, MUSK identified discernible clusters, particularly for HP, NORM, TA.HG, and TA.LG. The WSSS4LUAD dataset, which is focused on binary classification, presented challenges for the CLIP model in differentiating between benign and malignant tissue characteristics. In contrast, PLIP and MUSK effectively separated these tissue types. Notably, MUSK exhibited more pronounced decision boundary margins for the two categories than PLIP. The NCT-CRC dataset includes nine colorectal cancer tissue classes. Here, MUSK outperformed PLIP in visual clarity of class clusters. For instance, the classes MUS, NORM, and STR are distinctly separated by MUSK, whereas PLIP's results showed overlapping or mixed clusters, thus degrading its classification accuracy.



Supplementary Fig. 6. Visualization of melanoma relapse prediction. Two examples of melanoma cases with and without relapse. In each panel, the left image shows the original WSI, while the middle image displays the corresponding heatmaps that highlights the regions model focused on within the WSIs. The right images provide zoomed-in views of the regions receiving the most model attention.



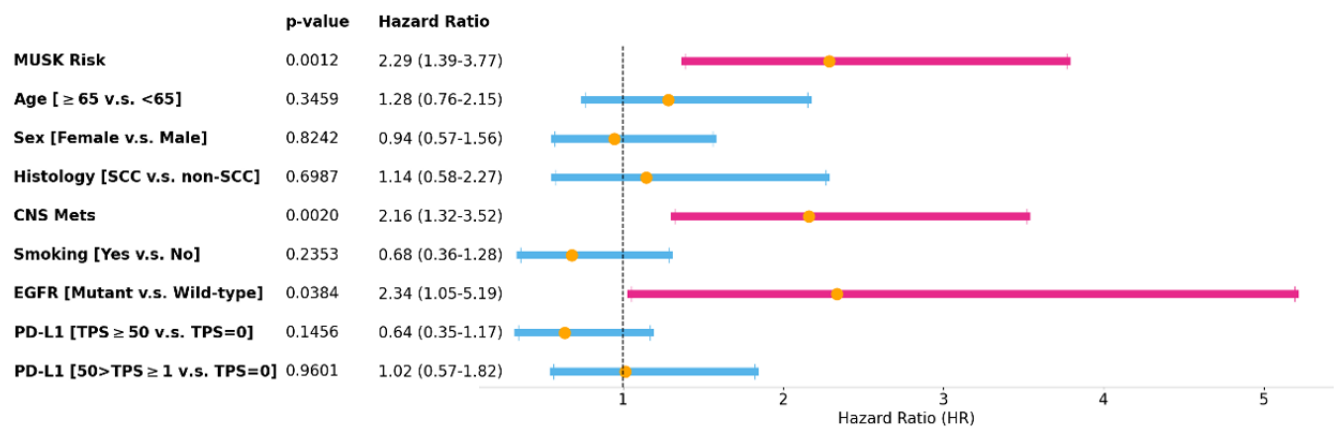
Supplementary Fig. 7. Multimodal prognosis prediction across 16 cancer types. MUSK consistently outperforms other foundation models for prognosis prediction when evaluated on 7,927 WSIs from 6,602 patients in TCGA. The overall bars show that the average c-index for MUSK is 0.746 vs. 0.667 – 0.684 for other models, showing statistical significance with p -value < 0.0001 . The overall bars represent the average performance across 16 projects. Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), Colorectal Adenocarcinoma Rectal Adenocarcinoma (COADREAD), Esophageal Carcinoma (ESCA), Glioblastoma Multiforme (GBM), Head and Neck Squamous Cell Carcinoma (HNSC), Brain Lower Grade Glioma (LGG), Liver Hepatocellular Carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Pancreatic Adenocarcinoma (PAAD), Renal Cell Carcinoma (RCC), Skin Cutaneous Melanoma (SKCM), Stomach Adenocarcinoma (STAD) and Uterine Corpus Endometrial Carcinoma (UCEC). Data is represented as mean with standard deviation calculated using 5-fold cross-validation experiments. The two-sided Mann-Whitney U test is used to assess the statistical significance between MUSK and the comparison methods.



(continued on next page)

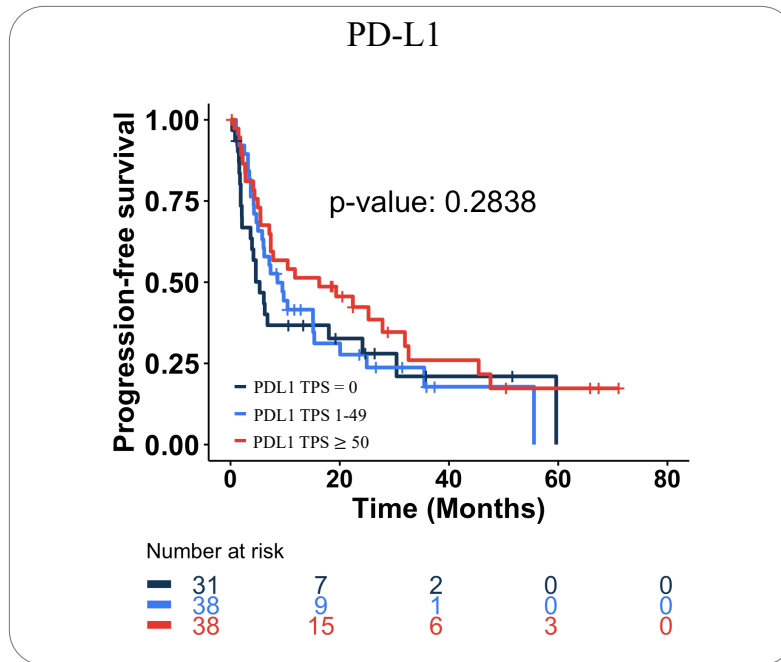


Supplementary Fig. 8. Prognosis multivariate analysis across 16 cancer types. We combined MUSK predicted risk scores with four clinicopathologic factors, including sex, age, cancer stage, and grade, to perform the multivariate Cox regression analysis. Note that the number of samples used for multivariate analysis is lower than the total number of patients in each project due to incomplete clinical data. The corresponding hazard ratios, along with their 95% confidence intervals (CIs) and p-values, quantify their impact on outcomes. Our predicted risk scores emerged as a significant predictive factor across all cancer types. Curated staging information was unavailable for CESC, GBM, and LGG, so we did not include stage in the multivariate analysis for these cancer types. We computed the P-values using the two-sided Wald test. We present the hazard ratio (HR) means with error bars representing the 95% CI.

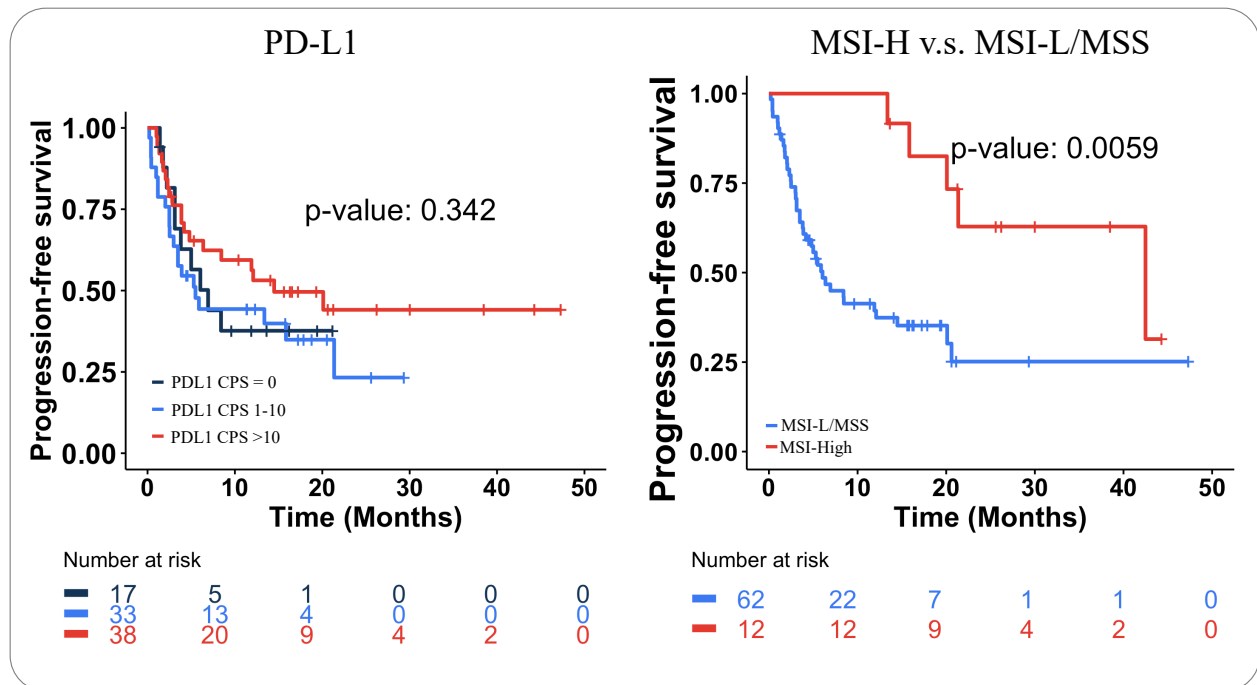


Supplementary Fig. 9. Lung cancer immunotherapy response prediction. Multivariate Cox regression analysis on $n = 118$ NSCLC patients shows MUSK is the most significant predictor of progression-free survival after incorporating clinicopathologic variables. We computed the P-values using the two-sided Wald test, and we presented the hazard ratio (HR) means with 95% confidence intervals.

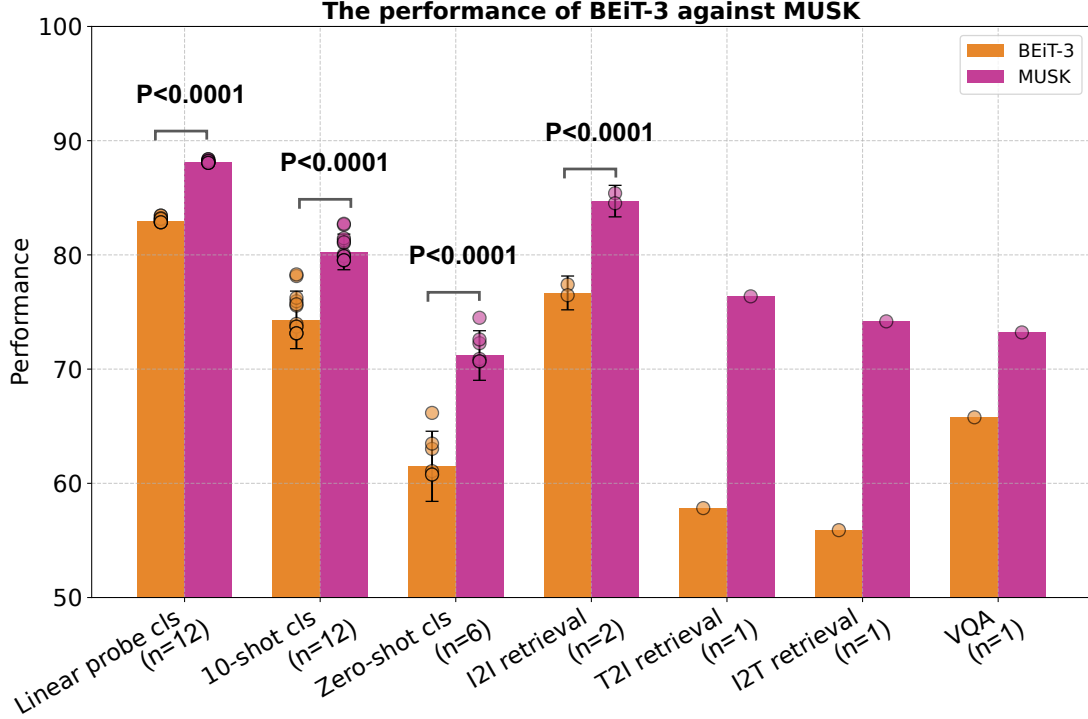
a. Lung cancer



b. Gastro-esophageal cancer



Supplementary Fig. 10. Kaplan-Meier analysis using clinical factors to stratify patients into low- and high-risk groups. PD-L1 expression failed to stratify patients for PFS in both lung and gastro-esophageal cancers. On the other hand, MSI status did stratify patients with gastro-esophageal cancer. The two-sided log-rank test is used to compare the survival differences between the high-risk and low-risk groups.



Supplementary Fig. 11. Ablation study of methodology adaptations. We conducted an ablation study to evaluate the impact of methodology adaptations on model performance. For this purpose, we trained BEiT-3 without any of the proposed adaptations, using the same dataset and training protocol, and compared its performance against that of MUSK. The error bars represent the average of the standard deviations within the datasets. Error bars are not provided for the I2T retrieval, T2I retrieval, and VQA tasks because they are evaluated on a single dataset. The evaluation metrics used are balanced accuracy for the linear probe classification, 10-shot classification, and zero-shot classification tasks; accuracy for the VQA task; *Recall@50* for the I2T and T2I retrievals; and *mMV@5* for the I2I retrieval. T2I: Text-to-Image; I2T: Image-to-Text; I2I: Image-to-Image; VQA: Visual Question Answering; cls: classification. The two-sided Mann-Whitney U test is used to measure the statistical significance between BEiT-3 and MUSK. n denotes the number of datasets corresponding to each task.

Supplementary Table 1: Text-to-image retrieval results on the BookSet and PathMMU datasets. The BookSet contains 4,265 image-text pairs and the PathMMU includes 7,774 image-text pairs. These results are shown with 95% confidence interval. The **bold** numbers indicate the best values, while the underlined numbers represent the second-best.

		<i>Recall@1</i>	<i>Recall@10</i>	<i>Recall@50</i>
BookSet	CLIP (ViT-b16)	0.20 _(0.12,0.30)	1.46 _(1.13,1.81)	4.49 _(3.85,4.95)
	CLIP (ViT-b32)	0.11 _(0.05,0.19)	1.26 _(0.96,1.57)	4.59 _(3.96,5.16)
	PLIP	1.21 _(1.01,1.43)	8.03 _(7.34,8.70)	22.41 _(21.06,23.68)
	QUITL1M (ViT-b16)	1.73 _(1.45,1.95)	10.42 _(9.54,11.51)	27.60 _(26.64,28.96)
	QUITL1M (ViT-b32)	2.01 _(1.76,2.30)	11.54 _(10.55,12.29)	27.84 _(26.66,29.10)
	BiomedicalCLIP	4.79 _(4.43,5.23)	21.34 _(20.02,22.42)	43.48 _(41.90,45.04)
	CONCH	<u>12.67</u> _(12.08,13.32)	<u>46.29</u> _(44.95,47.78)	<u>71.08</u> _(69.82,72.29)
	MUSK	16.07 _(15.49,16.66)	48.64 _(47.05,50.14)	75.03 _(73.94,76.12)
PathMMU	CLIP (ViT-b16)	0.34 _(0.26,0.45)	1.58 _(1.36,1.83)	3.84 _(3.37,4.23)
	CLIP (ViT-b32)	0.21 _(0.12,0.28)	1.34 _(1.08,1.58)	3.92 _(3.58,4.30)
	PLIP	0.31 _(0.23,0.41)	2.94 _(2.57,3.31)	10.17 _(9.57,10.84)
	QUITL1M (ViT-b16)	0.64 _(0.51,0.77)	3.72 _(3.41,4.04)	10.85 _(10.39,11.41)
	QUITL1M (ViT-b32)	0.46 _(0.36,0.55)	3.08 _(2.75,3.45)	10.68 _(10.16,11.33)
	BiomedicalCLIP	1.09 _(0.93,1.25)	7.02 _(6.50,7.59)	18.85 _(18.03,19.69)
	CONCH	<u>1.71</u> _(1.52,1.93)	<u>10.72</u> _(10.07,11.45)	<u>26.35</u> _(25.50,27.22)
	MUSK	2.75 _(2.48,2.97)	15.46 _(14.75,16.26)	33.87 _(32.90,34.80)

Supplementary Table 2: Image-to-text retrieval results on the BookSet and PathMMU datasets. The BookSet dataset contains 4,265 image-text pairs and the PathMMU dataset includes 7,774 image-text pairs. These results are shown with 95% confidence interval. The **bold** numbers indicate the best values, while the underlined numbers represent the second-best.

		<i>Recall@1</i>	<i>Recall@10</i>	<i>Recall@50</i>
BookSet	CLIP (ViT-b16)	0.10 _(0.05,0.19)	1.10 _(0.82,1.38)	3.97 _(3.54,4.45)
	CLIP (ViT-b32)	0.09 _(0.02,0.16)	1.07 _(0.82,1.36)	4.18 _(3.68,4.69)
	PLIP	0.67 _(0.52,0.82)	4.30 _(3.77,4.99)	14.26 _(13.13,15.36)
	QUITL1M (ViT-b16)	1.34 _(1.13,1.59)	8.31 _(7.57,8.89)	23.38 _(22.20,24.55)
	QUITL1M (ViT-b32)	1.60 _(1.38,1.92)	10.64 _(9.80,11.58)	26.38 _(25.28,27.50)
	BiomedicalCLIP	4.69 _(4.31,5.04)	20.18 _(19.06,21.24)	37.84 _(36.67,39.20)
	CONCH	<u>12.24</u> _(11.68,12.85)	<u>45.24</u> _(43.92,46.49)	<u>71.29</u> _(70.04,72.59)
	MUSK	13.51 _(12.89,14.13)	47.62 _(46.32,49.07)	74.84 _(73.59,75.94)
PathMMU	CLIP (ViT-b16)	0.30 _(0.23,0.40)	1.34 _(1.14,1.54)	3.52 _(3.14,3.92)
	CLIP (ViT-b32)	0.29 _(0.21,0.37)	1.31 _(1.11,1.51)	3.70 _(3.31,4.06)
	PLIP	0.14 _(0.09,0.19)	1.12 _(0.94,1.32)	5.00 _(4.54,5.48)
	QUITL1M (ViT-b16)	0.64 _(0.53,0.76)	3.83 _(3.46,4.23)	10.64 _(10.03,11.15)
	QUITL1M (ViT-b32)	0.49 _(0.36,0.63)	3.19 _(2.92,3.51)	9.15 _(8.68,9.63)
	BiomedicalCLIP	1.35 _(1.18,1.54)	8.22 _(7.56,8.77)	20.82 _(20.04,21.62)
	CONCH	<u>1.98</u> _(1.75,2.17)	<u>12.01</u> _(11.41,12.73)	<u>27.34</u> _(26.40,28.25)
	MUSK	3.45 _(3.19,3.67)	17.37 _(16.71,18.38)	34.43 _(33.37,35.46)

Supplementary Table 3: Zero-Shot image-to-image retrieval on UniToPatho and BRACS (6-class). These results are shown with 95% confidence interval. The **bold** numbers indicate the best values, while the underlined numbers represent the second-best.

	Model	Recall@1	Recall@3	Recall@5	mMv@5
BRACS	CLIP (ViT-b16)	56.66 _(54.18,58.65)	72.18 _(70.41,74.16)	80.83 _(79.41,82.29)	57.04 _(54.86,59.08)
	CLIP (ViT-b32)	54.21 _(52.41,56.27)	69.82 _(67.89,71.75)	78.83 _(77.23,80.61)	54.50 _(52.70,56.68)
	PLIP	68.06 _(66.26,69.69)	80.56 _(79.25,82.15)	87.11 _(85.69,88.22)	68.23 _(66.42,70.14)
	QUITL1M (ViT-b16)	67.08 _(65.21,69.30)	79.42 _(77.88,81.18)	86.05 _(84.76,87.44)	67.52 _(65.67,69.50)
	QUITL1M (ViT-b32)	70.84 _(69.10,72.77)	82.50 _(80.77,84.22)	88.55 _(87.38,89.58)	70.34 _(68.05,72.12)
	BiomedCLIP	65.37 _(63.56,67.28)	78.58 _(77.25,80.59)	85.44 _(84.29,87.15)	66.50 _(64.60,68.42)
	CONCH	<u>74.60</u> _(73.23,76.18)	<u>84.85</u> _(83.36,86.26)	<u>89.77</u> _(88.51,90.76)	<u>74.29</u> _(72.77,75.79)
	MUSK	77.13 _(75.59,78.52)	86.64 _(85.20,87.87)	91.43 _(90.30,92.57)	76.82 _(75.18,78.61)
UnitoPatho	CLIP (ViT-b16)	62.52 _(61.32,63.76)	77.33 _(76.14,78.45)	84.93 _(83.87,85.97)	61.18 _(59.43,62.78)
	CLIP (ViT-b32)	57.23 _(55.66,58.81)	74.06 _(72.71,75.27)	82.64 _(81.51,83.65)	57.44 _(56.16,59.03)
	PLIP	81.22 _(80.26,82.21)	89.91 _(89.18,90.83)	93.76 _(93.14,94.34)	79.76 _(78.68,80.74)
	QUITL1M (ViT-b16)	71.94 _(70.84,72.96)	84.25 _(83.34,85.17)	89.97 _(89.30,90.81)	70.36 _(68.99,71.53)
	QUITL1M (ViT-b32)	<u>84.56</u> _(83.76,85.40)	<u>91.53</u> _(90.78,92.19)	<u>94.60</u> _(94.05,95.02)	<u>82.51</u> _(81.57,83.52)
	BiomedCLIP	66.65 _(65.45,67.85)	80.31 _(79.32,81.30)	87.06 _(86.13,87.75)	65.87 _(64.53,67.03)
	CONCH	80.91 _(79.79,82.06)	89.25 _(88.44,90.04)	93.10 _(92.47,93.79)	78.82 _(77.79,79.95)
	MUSK	88.70 _(87.85,89.43)	93.91 _(93.25,94.51)	96.09 _(95.58,96.48)	86.30 _(85.50,87.20)

Supplementary Table 4: Zero-shot image classification results (balanced accuracy). These results are shown with 95% confidence interval. The **bold** numbers indicate the best values, while the underlined numbers represent the second-best.

	PatchCamelyon	SkinCancer	PanNuke	UniToPatho
CLIP (ViT-b16)	55.79 _(55.23,56.31)	14.06 _(13.17,14.97)	65.69 _(63.68,67.70)	16.91 _(15.33,18.60)
CLIP (ViT-b32)	53.17 _(52.64,53.69)	11.34 _(10.87,11.80)	60.43 _(58.95,61.80)	17.56 _(16.97,18.15)
PLIP	58.92 _(58.62,59.22)	26.43 _(25.58,27.34)	63.55 _(61.80,65.25)	24.23 _(22.66,25.76)
QUITL1M (ViT-b16)	<u>67.98</u> _(67.52,68.45)	19.89 _(19.30,20.55)	55.25 _(54.28,56.26)	23.03 _(21.07,25.29)
QUITL1M (ViT-b32)	64.63 _(64.10,65.14)	<u>30.59</u> _(29.77,31.43)	53.92 _(52.79,55.13)	26.28 _(24.43,28.21)
BiomedCLIP	62.84 _(62.29,63.38)	16.99 _(16.49,17.54)	<u>74.88</u> _(73.06,76.93)	22.58 _(20.66,24.51)
CONCH	52.88 _(52.68,53.08)	22.04 _(21.16,22.91)	67.15 _(65.39,68.89)	<u>26.95</u> _(25.24,28.66)
MUSK	78.45 _(78.00,78.89)	58.04 _(57.15,58.93)	82.17 _(80.33,83.89)	37.06 _(34.98,39.02)

Supplementary Table 5: 10-shot classification performance using balanced accuracy across 12 benchmarks. The results represent the mean and standard deviation from 10 independent experiments in supervised training. The **bold** numbers indicate the best values, while the underlined numbers represent the second-best.

	ViT-b32			ViT-b16				ViT-L
	CLIP	PLIP	QUITL1M	CLIP	BiomedCLIP	QUITL1M	CONCH	MUSK
NCT-CRC	76.45(2.40)	88.18(1.75)	90.23(1.21)	71.52(2.98)	87.38(1.07)	80.69(2.68)	87.70(2.72)	90.60(1.76)
PatchCamelyon	63.13(3.97)	76.20(4.60)	<u>76.57(4.02)</u>	65.85(3.45)	73.05(3.28)	72.99(5.00)	77.64(4.55)	83.54(2.02)
SkinCancer	60.50(2.37)	77.62(1.28)	75.83(1.22)	62.71(1.57)	72.61(1.59)	71.59(1.36)	<u>87.89(1.82)</u>	89.66(0.90)
SICAPv2	40.47(2.73)	63.16(3.02)	63.51(5.67)	42.44(4.13)	58.44(3.43)	63.12(4.55)	60.36(6.01)	70.08(4.38)
PanNuke	66.92(4.53)	78.64(2.48)	<u>78.23(2.52)</u>	68.98(6.43)	77.29(2.55)	73.45(7.54)	81.40(3.36)	83.47(3.54)
UniToPatho	33.24(2.90)	44.31(2.75)	45.59(2.15)	34.02(3.16)	39.24(1.59)	39.99(2.92)	<u>45.80(2.59)</u>	55.63(2.17)
WSSS4LUAD	70.29(3.91)	88.78(1.61)	89.86(1.39)	67.58(7.00)	89.09(1.58)	87.30(0.95)	<u>89.93(3.38)</u>	93.47(1.01)
Osteo	77.78(3.04)	88.18(3.07)	85.31(2.71)	78.13(3.62)	84.39(1.49)	86.80(3.19)	<u>89.74(1.81)</u>	92.77(1.86)
LC25000	87.85(1.32)	94.20(1.19)	93.97(1.56)	85.22(2.40)	89.93(1.00)	89.07(0.86)	93.10(1.62)	97.25(0.64)
RenalCell	68.44(2.25)	<u>76.00(1.71)</u>	75.89(1.39)	70.65(1.71)	73.09(1.57)	69.52(1.66)	70.20(3.68)	80.57(1.42)
BRCAS (6-class)	33.80(2.81)	38.88(2.36)	41.52(2.51)	33.69(2.64)	43.98(1.67)	41.52(2.00)	<u>49.47(2.72)</u>	53.72(3.32)
BRCAS (3-class)	48.73(2.62)	52.24(3.54)	58.09(5.25)	50.56(4.20)	57.73(3.22)	58.42(3.47)	65.18(4.57)	63.56(4.31)

Supplementary Table 6: Linear probe classification performance using balanced accuracy across 12 benchmarks. The results represent the mean and standard deviation from 10 independent experiments in supervised training. The **bold** numbers indicate the best values, while the underlined numbers represent the second-best.

	ViT-b32			ViT-b16				ViT-L
	CLIP	PLIP	QUILT1M	CLIP	BiomedCLIP	QUILT1M	CONCH	MUSK
NCT-CRC	90.59(0.07)	90.62(0.06)	<u>94.06(0.05)</u>	89.27(0.04)	91.50(0.05)	92.24(0.04)	93.83(0.24)	94.11(0.20)
PatchCamelyon	81.98(0.03)	86.52(0.07)	<u>86.53(0.03)</u>	84.04(0.03)	83.39(0.03)	86.80(0.03)	87.92(0.14)	89.82(0.04)
SkinCancer	82.99(0.03)	89.29(0.05)	89.22(0.07)	82.91(0.08)	83.44(0.03)	85.87(0.06)	<u>93.40(0.13)</u>	94.44(0.14)
SICAPv2	51.43(0.33)	69.26(0.15)	71.49(0.28)	49.99(0.15)	67.49(0.27)	64.86(0.18)	<u>78.46(0.61)</u>	78.86(0.30)
PanNuke	82.25(0.52)	89.61(0.17)	89.18(0.32)	83.45(0.59)	86.72(0.30)	86.27(0.21)	<u>94.15(0.17)</u>	96.42(0.11)
UniToPatho	37.64(0.28)	46.52(0.41)	50.76(0.31)	37.54(0.34)	44.66(0.16)	43.32(0.45)	<u>75.47(0.24)</u>	78.78(0.32)
WSSS4LUAD	86.72(0.51)	93.77(0.28)	94.47(0.28)	84.90(0.58)	93.05(0.35)	91.57(0.31)	<u>94.84(0.13)</u>	97.78(0.05)
Osteo	79.34(0.26)	91.08(0.32)	88.75(0.69)	78.05(0.62)	86.95(0.27)	87.20(1.04)	<u>93.06(0.67)</u>	96.92(0.34)
LC25000	79.08(0.73)	<u>95.83(0.03)</u>	93.63(0.46)	73.70(0.96)	89.84(0.38)	85.42(0.72)	94.99(0.89)	98.61(0.09)
RenalCell	78.10(0.16)	<u>84.60(0.15)</u>	83.07(0.10)	78.69(0.19)	81.51(0.12)	79.09(0.12)	<u>88.12(0.20)</u>	91.00(0.12)
BRCAS (3-class)	59.29(0.56)	63.33(0.21)	67.13(0.44)	62.22(0.36)	66.78(0.21)	66.75(0.36)	<u>73.94(0.45)</u>	76.63(0.31)
BRCAS (6-class)	43.26(0.55)	48.34(0.65)	50.05(0.29)	43.86(0.54)	50.78(0.35)	49.06(0.48)	<u>63.76(0.96)</u>	64.88(0.37)

Supplementary Table 7: Slide-level benchmark performance using AUC or c-index. These slide-level tasks include biomarker prediction tasks for *ER*, *PR*, *HER2*, and IDH status prediction in BCNB and MUV-IDH datasets, prognosis prediction across 16 TCGA cohorts, and immunotherapy response prediction in lung and gastro-esophageal cancers. MUSK is compared with state-of-the-art pathology foundation models: PLIP, UNI, GigaPath, Virchow, and CONCH. Data are represented with mean and 95% confidence intervals.

Task	Dataset	Methods					
		PLIP	UNI	GigaPath	Virchow	CONCH	MUSK
Biomarker prediction (AUC)	BCNB-ER	0.813 (0.796, 0.829)	0.869 (0.859, 0.88)	0.873 (0.866, 0.881)	0.866 (0.849, 0.883)	0.875 (0.855, 0.895)	0.886 (0.878, 0.894)
	BCNB-PR	0.782 (0.767, 0.797)	0.817 (0.801, 0.833)	0.825 (0.811, 0.839)	0.824 (0.808, 0.839)	0.815 (0.805, 0.826)	0.85 (0.836, 0.864)
	BCNB-HER2	0.706 (0.666, 0.747)	0.769 (0.756, 0.782)	0.786 (0.756, 0.817)	0.77 (0.741, 0.799)	0.771 (0.745, 0.796)	0.826 (0.813, 0.839)
	MUV-IDH	0.947 (0.931, 0.964)	0.962 (0.959, 0.966)	0.959 (0.95, 0.969)	0.95 (0.931, 0.97)	0.953 (0.941, 0.966)	0.978 (0.97, 0.987)
Prognosis prediction (c-index)	BLCA	0.595 (0.546, 0.644)	0.622 (0.509, 0.735)	0.595 (0.494, 0.697)	0.603 (0.501, 0.704)	0.604 (0.561, 0.646)	0.681 (0.646, 0.715)
	BRCA	0.636 (0.591, 0.682)	0.755 (0.686, 0.825)	0.746 (0.670, 0.823)	0.749 (0.660, 0.838)	0.751 (0.733, 0.769)	0.812 (0.769, 0.855)
	CESC	0.668 (0.617, 0.720)	0.715 (0.595, 0.835)	0.685 (0.606, 0.765)	0.669 (0.563, 0.775)	0.748 (0.700, 0.796)	0.752 (0.667, 0.838)
	COADREAD	0.718 (0.699, 0.736)	0.739 (0.692, 0.785)	0.709 (0.662, 0.755)	0.735 (0.667, 0.803)	0.655 (0.547, 0.764)	0.819 (0.755, 0.882)
	ESCA	0.704 (0.658, 0.750)	0.704 (0.608, 0.800)	0.704 (0.638, 0.770)	0.676 (0.530, 0.822)	0.666 (0.581, 0.751)	0.800 (0.720, 0.879)
	GBM	0.583 (0.518, 0.648)	0.603 (0.522, 0.684)	0.604 (0.543, 0.664)	0.582 (0.524, 0.640)	0.568 (0.505, 0.632)	0.630 (0.597, 0.663)
	HNSC	0.615 (0.526, 0.704)	0.607 (0.531, 0.682)	0.604 (0.556, 0.652)	0.576 (0.467, 0.686)	0.624 (0.565, 0.684)	0.651 (0.630, 0.673)
	LGG	0.674 (0.598, 0.751)	0.690 (0.572, 0.808)	0.763 (0.681, 0.846)	0.697 (0.516, 0.879)	0.695 (0.628, 0.762)	0.814 (0.758, 0.870)
	LIHC	0.708 (0.647, 0.769)	0.735 (0.681, 0.789)	0.718 (0.650, 0.786)	0.753 (0.694, 0.812)	0.702 (0.656, 0.749)	0.741 (0.719, 0.762)
	LUAD	0.628 (0.560, 0.696)	0.654 (0.619, 0.690)	0.638 (0.594, 0.683)	0.664 (0.616, 0.712)	0.659 (0.592, 0.726)	0.712 (0.644, 0.779)
	LUSC	0.618 (0.576, 0.660)	0.659 (0.593, 0.725)	0.679 (0.612, 0.745)	0.652 (0.581, 0.724)	0.683 (0.634, 0.733)	0.697 (0.603, 0.790)
	PAAD	0.659 (0.590, 0.728)	0.662 (0.630, 0.694)	0.681 (0.606, 0.755)	0.638 (0.558, 0.718)	0.656 (0.613, 0.699)	0.710 (0.669, 0.751)
	RCC	0.790 (0.773, 0.806)	0.776 (0.727, 0.825)	0.780 (0.721, 0.840)	0.773 (0.720, 0.826)	0.813 (0.778, 0.847)	0.887 (0.854, 0.920)
	SKCM	0.642 (0.588, 0.696)	0.615 (0.538, 0.692)	0.616 (0.560, 0.673)	0.623 (0.542, 0.704)	0.658 (0.627, 0.690)	0.687 (0.647, 0.726)
Treatment response prediction	STAD	0.670 (0.599, 0.741)	0.659 (0.578, 0.741)	0.651 (0.569, 0.733)	0.628 (0.558, 0.698)	0.679 (0.614, 0.743)	0.706 (0.660, 0.751)
	UCEC	0.774 (0.719, 0.829)	0.706 (0.606, 0.805)	0.727 (0.611, 0.842)	0.727 (0.637, 0.817)	0.786 (0.736, 0.837)	0.849 (0.828, 0.870)
	Lung (AUC)	0.636 (0.579, 0.692)	0.636 (0.514, 0.758)	0.647 (0.542, 0.752)	0.643 (0.574, 0.712)	0.692 (0.664, 0.721)	0.768 (0.724, 0.812)
	Lung (c-index)	0.627 (0.57, 0.683)	0.58 (0.509, 0.65)	0.598 (0.543, 0.653)	0.599 (0.545, 0.653)	0.64 (0.615, 0.665)	0.705 (0.677, 0.732)
	Gastro-eso.(AUC)	0.654 (0.635, 0.674)	0.651 (0.557, 0.746)	0.644 (0.561, 0.727)	0.645 (0.551, 0.739)	0.693 (0.658, 0.728)	0.762 (0.718, 0.805)
	Gastro-eso.(c-index)	0.642 (0.616, 0.667)	0.638 (0.581, 0.694)	0.633 (0.59, 0.675)	0.641 (0.584, 0.698)	0.674 (0.642, 0.705)	0.717 (0.678, 0.756)

Supplementary Table 8: Standardization of TCGA clinical reports with expert-designed GPT-4 prompts

Expert-Designed Prompts	Structured Reports
• What is the tumor size? Describe it with concrete numbers if provided and explain it. • What is the differentiation of the lesion? • Is there any evidence of the lesion invading adjacent tissues or organs? • Any evidence of lymph node metastasis. This indicates whether the cancer has spread to the lymph nodes. • Are the margins of the excised tissue clear of disease?	TCGA-78-7167: Tumor size: The tumor in the left upper lobe of the lung has a most significant dimension of 3.5 cm. Differentiation of the lesion: Moderately to Poorly differentiated, with a poorly differentiated component representing less than 50% of the tumor; Evidence of lymph Node Metastasis: Metastatic carcinoma was found in 3 out of 18 lymph nodes; Evidence of invasion into adjacent tissues or organs: Primary tumor shows no direct extension or invasion of adjacent tissues; Margins of the excised tissue: Margins are uninvolved by invasive carcinoma, indicating a negative surgical margin status (R0).

Supplementary Table 9: Standardization of lung cancer clinical reports with expert-designed GPT-4 prompts

Expert-Designed Prompts	Structured Reports
• What is the differentiation of the lesion? (e.g., well-differentiated, moderately differentiated, etc.) • Is there any indication of lymphovascular invasion by the lesion? And explain the reasons. • What are the immunohistology testing results of PD-L? Explain it briefly. • What is the tumor mutational burden result? • What is the cancer diagnosis stage? (e.g., early stage, advanced stage, etc.) • Any evidence of central nervous system metastases? • Did the patient receive other treatments concurrently? (e.g., radiation, chemotherapy)	One example: The lesion is poorly differentiated; There is evidence of the lesion invading the L3 vertebral body; The PD-L1 result shows < 1% positivity, indicating no PD-L1 expression by immunohistochemical staining; The Tumor Mutational Burden (TMB) is reported as 4.8 mutations per megabase, with mutations detected in NF1 and TP53 genes. advanced stage of cancer with the tumor has spread to the opposite lung or distant organs like the brain, liver, or bones; The patient has central nervous system metastases, indicating an advanced stage of cancer with a poor prognosis; the patient received chemotherapy concurrently with immunotherapy;

Supplementary Table 10: Standardization of gastric cancer clinical reports with expert-designed prompts.

Expert-Designed Prompts	Structured Reports
• What is the differentiation of the lesion? (e.g., well-differentiated, moderately differentiated, etc.) • Is there any indication of lymphovascular invasion by the lesion? And explain the reasons. • Is there any evidence of the lesion invading adjacent tissues or organs? If yes, what are they / is it? • What are the immunohistology testing results of PD-L1 and the CPS score? Explain it. • Are the margins of the excised tissue clear of disease? • What's the MSI status? • What's the HER2 result?	One example: Differentiation of the lesion: Poorly differentiated adenocarcinoma; Lymphovascular invasion: Unknown; Invasion of adjacent tissues or organs: Invasion of the GE junction; Immunohistology testing of PD-L1 and CPS score: PD-L1 expression (CPS = 20), which indicates positive PD-L1 expression; Clearance of disease at margins: The margins of the excised tissue are not mentioned; The Microsatellite Instability (MSI) status is that MLH1, MSH2, MSH6, and PMS2 demonstrate intact expression for all four of these enzymes in the neoplastic tissue; The HER2 status is Negative (+1).

Supplementary Table 11: Clinical characteristics of the lung cancer dataset.

Lung Cancer		
Age: Mean(IQR)		67.5 (62.6-75.5)
Gender		
	Male	67 (56.8%)
	Female	51 (43.2%)
Smoking		
	Yes	83 (70.3%)
	No	35 (29.7%)
CNS Metastases		
	Yes	47 (39.8%)
	No	59 (50.0%)
	Unknown	12 (10.2%)
Histology		
	Adenocarcinoma	96 (81.4%)
	Squamous Cell Carcinoma	16 (13.6%)
	Others	6 (5.0%)
Tumor differentiation		
	Well	7 (5.9%)
	Moderately	46 (39.0%)
	Poorly / Undifferentiated	53 (44.9%)
	Unknown	12 (10.2%)
EGFR mutation		
	Yes	16 (13.6%)
	No	95 (80.5%)
	Unknown	7 (5.9%)
PD-L1 TPS		
	0%	31 (26.2%)
	1-49%	37 (31.4%)
	≥50%	37 (31.4%)
	Unknown	13 (11.0%)
Line of therapy		
	1L	75 (63.6%)
	2L	31 (26.3%)
	3L+	12 (10.1%)
Concurrent chemotherapy		
	Yes	72 (61.0%)
	No	46 (39.0%)

Supplementary Table 12: Clinical characteristics of the Gastro-esophageal cancer dataset.

Gastro-esophageal cancers		
Age: Mean(IQR)		63.1 (54.0-73.0)
Gender		
	Male	35 (34.6%)
	Female	66 (65.3%)
Primary tumor site		
	Esophagus	25 (24.8%)
	GE Junction	18 (17.7%)
	Stomach	55 (54.5%)
	Unknown	3 (3.0%)
Histology		
	Adenocarcinoma	81 (80.2%)
	Squamous Cell Carcinoma	13 (12.8%)
	Others	7 (7.0%)
Tumor differentiation		
	Well	1 (1.0%)
	Moderately	24 (23.7%)
	Poorly / Undifferentiated	53 (52.4%)
	Unknown	23 (22.7%)
MSI status		
	MSI-H	12 (11.8%)
	MSS/MSI-L	62 (61.3%)
	Unknown	27 (26.7%)
HER2 status		
	Negative	54 (53.5%)
	Positive	11 (10.9%)
	Equivocal	7 (6.9%)
	Unknown	29 (28.7%)
PD-L1 CPS		
	0%	17 (16.8%)
	1-9%	33 (32.7%)
	≥10%	38 (37.6%)
	Unknown	13 (12.9%)
Line of therapy		
	1L	50 (49.5%)
	2L	29 (28.7%)
	3L+	22 (21.7%)
Concurrent chemotherapy		
	Yes	52 (51.4%)
	No	49 (48.5%)

Supplementary Table 13: Ablation studies for methodology adaptations. BEiT-3 is trained as the baseline without any proposed methodology adaptations. We report the average performance across multiple datasets, along with the average standard deviation within datasets for tasks involving more than one dataset. Standard deviations are not provided for the I2T retrieval, T2I retrieval, and VQA tasks because they are evaluated on a single dataset. The evaluation metrics used are balanced accuracy for the linear probe classification, 10-shot classification, and zero-shot classification tasks; accuracy for the VQA task; *Recall*@50 for the I2T and T2I retrievals; and *mMV*@5 for the I2I retrieval. T2I: Text-to-Image; I2T: Image-to-Text; I2I: Image-to-Image; VQA: Visual Question Answering; cls: classification.

Downstream tasks	Number of datasets	BEiT-3	MUSK	Increase
Linear probe cls	12	82.99 _(0.29)	86.03 _(0.17)	+3.04
10-shot cls	12	74.31 _(2.52)	77.28 _(1.35)	+2.97
Zero-shot cls	6	61.49 _(3.07)	68.01 _(2.56)	+6.52
I2I retrieval	2	76.67 _(1.48)	78.64 _(1.39)	+1.97
T2I retrieval	1	57.83	67.31	+9.48
I2T retrieval	1	55.90	65.13	+9.23
VQA	1	65.78	70.82	+5.04

Supplementary Table 14: Ablation studies for methodology adaptations. We report the average performance across multiple datasets, along with the average standard deviation within datasets for tasks involving more than one dataset. Standard deviations are not provided for the I2T retrieval, T2I retrieval, and VQA tasks because they are evaluated on a single dataset. The evaluation metrics used are balanced accuracy for the linear probe classification, 10-shot classification, and zero-shot classification tasks; accuracy for the VQA task; *Recall@50* for the I2T and T2I retrievals; and *mMV@5* for the I2I retrieval. T2I: Text-to-Image; I2T: Image-to-Text; I2I: Image-to-Image; VQA: Visual Question Answering; cls: classification; Aug: augmentation.

		Number of datasets	Standard Aug.	Pathology Aug.	Increase
Ablation 1	Linear probe cls	12	83.39 _(0.23)	86.03 _(0.17)	+2.64
	10-shot cls	12	75.25 _(2.01)	77.28 _(1.35)	+2.03
	Zero-shot cls	6	66.70 _(2.12)	68.01 _(2.56)	+1.31
	I2I retrieval	2	76.96 _(1.64)	78.64 _(1.39)	+1.68
	T2I retrieval	1	66.15	67.31	+1.16
	I2T retrieval	1	64.11	65.13	+1.02
	VQA	1	69.09	70.82	+1.73
			ImageNet tokenizer	Pathology tokenizer	Increase
Ablation 2	Linear probe cls	12	84.34 _(0.29)	86.03 _(0.21)	+1.69
	10-shot cls	12	75.62 _(1.97)	77.28 _(1.56)	+1.66
	Zero-shot cls	6	66.86 _(2.11)	68.01 _(2.69)	+1.15
	I2I retrieval	2	77.63 _(1.42)	78.64 _(1.79)	+1.01
	T2I retrieval	1	66.32	67.31	+0.99
	I2T retrieval	1	64.51	65.13	+0.62
	VQA	1	70.01	70.82	+0.81
			w/o multimodal decoder	w/ multimodal decoder	Increase
Ablation 3	Linear probe cls	12	86.02 _(0.17)	86.03 _(0.16)	+0.01
	10-shot cls	12	77.20 _(2.25)	77.28 _(1.83)	+0.08
	Zero-shot cls	6	66.88 _(2.87)	68.01 _(2.54)	+1.13
	I2I retrieval	2	78.51 _(1.36)	78.64 _(1.46)	+0.13
	T2I retrieval	1	65.02	67.31	+2.29
	I2T retrieval	1	62.19	65.13	+2.94
	VQA	1	69.23	70.82	+1.59
			Original Quilt1M	Bootstrap Quilt1M	Increase
Ablation 4	Linear probe cls	12	85.63 _(0.24)	86.03 _(0.14)	+0.40
	10-shot cls	12	77.05 _(1.67)	77.28 _(2.21)	+0.23
	Zero-shot cls	6	65.81 _(2.58)	68.01 _(2.81)	+2.20
	I2I retrieval	2	78.58 _(1.18)	78.64 _(1.07)	+0.06
	T2I retrieval	1	60.98	67.31	+6.33
	I2T retrieval	1	60.25	65.13	+4.88
	VQA	1	68.01	70.82	+2.81

Supplementary Table 15: Ablation studies of the first stage mask pretraining. We report the average performance across multiple datasets, along with the average standard deviation within datasets for tasks involving more than one dataset. Standard deviations are not provided for the I2T retrieval, T2I retrieval, and VQA tasks because they are evaluated on a single dataset. The ablation study includes five models: `model-1`, `model-2`, `model-3`, `model-4`, and `model-5`. These models were pretrained on different datasets during the first unified masked pretraining phase, but all were trained on the same Quilt1M dataset during the second contrastive learning pretraining phase. Regarding the pretraining data used in the first phase, `model-1` had no additional masked pretraining and relied on randomly initialized weights. `model-2` was trained on 15 million natural images from ImageNet-21k and 500 million natural language tokens. `model-3` used the Quilt1M for unified masked pretraining. `model-4` employed 15 million pathology images from TCGA and 500 million pathology text tokens. Finally, `model-5` was trained on 50 million pathology images from TCGA and 1 billion pathology text tokens, which is referred to as the MUSK (base version). The evaluation metrics used are balanced accuracy for the linear probe classification, 10-shot classification, and zero-shot classification tasks; accuracy for the VQA task; *Recall*@50 for the I2T and T2I retrievals; and *mMV*@5 for the I2I retrieval.

Downstream tasks	Number of downstream datasets	Models and related pretraining data				
		<code>model-1</code> (None)	<code>model-2</code> (ImageNet-21k 15M)	<code>model-3</code> (Quilt1M)	<code>model-4</code> (TCGA 15M)	<code>model-5</code> (TCGA 50M)
Linear probe cls	12	73.44 _(0.15)	84.47 _(0.23)	83.05 _(0.27)	85.34 _(0.18)	88.10 _(0.13)
10-shot cls	12	60.31 _(1.98)	75.65 _(2.11)	75.01 _(1.89)	77.62 _(1.53)	79.72 _(2.07)
Zero-shot cls	6	37.12 _(2.94)	62.18 _(2.18)	55.58 _(2.59)	66.59 _(2.84)	69.16 _(2.07)
I2I retrieval	2	58.92 _(1.61)	77.11 _(1.48)	75.28 _(1.34)	79.60 _(1.22)	82.10 _(1.75)
T2I retrieval	1	7.59	57.67	25.01	74.19	75.57
I2T retrieval	1	6.54	57.18	22.22	71.91	73.23
VQA	1	55.11	68.09	65.74	70.88	73.01

Supplementary Table 16: Model configuration and training recipe of MUSK.

Hyperparameters		Small	Base	Large
Model Configuration	Layers	6	12	24
	Hidden size	512	768	1024
	FFN hidden size	2048	3072	4096
	Attention heads	8	12	16
	Attention head size	64		
	Image size	384×384		
	Patch size	16×16		
	Parameters	72M	222M	675M
Mask pretraining	Training steps	1M		
	Batch size	2048		
	Adam ϵ	1e-8		
	Adam β	(0.9, 0.95)		
	Peak learning rate	1.5e-3		
	Minimal Learning rate	1e-5		
	Learning rate schedule	Cosine		
	Warmup steps	10k		
	Gradient clipping	3.0	3.0	1.0
	Stoch. depth	0.1		
	Weight decay	0.05		
Image-text Contrastive learning	Training epochs	20		
	Batch size	3072		
	Adam ϵ	1e-8		
	Adam β	(0.9, 0.95)		
	Peak learning rate	1.0e-4		
	Minimal Learning rate	1e-5		
	Learning rate schedule	Cosine		
	Warmup epochs	2		
	Gradient clipping	3.0	3.0	1.0
	Stoch. depth	0.1		
	Weight decay	0.05		

Supplementary Table 17: Model performance of MUSK variants. We report the average performance across multiple datasets, along with the average standard deviation within datasets for tasks involving more than one dataset. Standard deviations are not provided for the I2T retrieval, T2I retrieval, and VQA tasks because they are evaluated on a single dataset. FLOPS is measured using a 384×384 input image and 100 text tokens. The evaluation metrics used are balanced accuracy for the linear probe classification, 10-shot classification, and zero-shot classification tasks; accuracy for the VQA task; *Recall*@50 for the I2T and T2I retrievals; and *mMV*@5 for the I2I retrieval.

Downstream tasks	Number of datasets	Models (parameters/FLOPS)		
		MUSK-small (75M/15G)	MUSK-base (222M/64G)	MUSK-large (675M/222G)
Linear probe cls	12	82.24 _(0.19)	88.10 _(0.13)	88.13 _(0.16)
10-shot cls	12	73.19 _(1.78)	79.72 _(2.07)	80.26 _(1.56)
Zero-shot cls	6	60.92 _(2.45)	69.71 _(2.81)	71.19 _(2.17)
I2I retrieval	2	79.17 _(1.43)	82.10 _(1.75)	84.71 _(1.38)
T2I retrieval	1	53.86	75.57	76.37
I2T retrieval	1	54.31	73.23	74.18
VQA	1	66.86	73.01	73.21

4 Supplementary References

1. Ramesh, A. *et al.* Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831 (Pmlr, 2021).
2. Peng, Z., Dong, L., Bao, H., Ye, Q. & Wei, F. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366* (2022).
3. Shen, Y., Luo, Y., Shen, D. & Ke, J. RandStainNA: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 212–221 (Springer, 2022).
4. Kang, M., Song, H., Park, S., Yoo, D. & Pereira, S. Benchmarking self-supervised learning on diverse pathology datasets. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 3344–3354 (2022).
5. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
6. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics* (2019).
7. Wu, Y. *et al.* Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).
8. Kim, J.-H., Jun, J. & Zhang, B.-T. Bilinear attention networks. *Advances in neural information processing systems* **31** (2018).
9. Nguyen, B. D. *et al.* Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV* 22, 522–530 (Springer, 2019).
10. Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J. & Chang, K.-W. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).
11. Naseem, U., Khushi, M., Dunn, A. G. & Kim, J. K-PathVQA: Knowledge-aware multimodal representation for pathology visual question answering. *IEEE Journal of Biomedical and Health Informatics* **28**, 1886–1895 (2024).
12. He, X., Zhang, Y., Mou, L., Xing, E. & Xie, P. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286* (2020).
13. Barbano, C. A. *et al.* Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. In *2021 IEEE International Conference on Image Processing (ICIP)*, 76–80 (IEEE, 2021).
14. Brancati, N. *et al.* BRACS: A dataset for breast carcinoma subtyping in H&E histology images. *Database* **2022**, baac093 (2022).
15. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
16. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine* **29**, 2307–2316 (2023).
17. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**, 850–862 (2024).

18. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
19. Vorontsov, E. *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine* **30**, 2924–2935 (2024).
20. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874 (2024).
21. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136 (PMLR, 2018).
22. Gamper, J. & Rajpoot, N. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16549–16559 (2021).
23. Sun, Y. *et al.* PathMMU: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. *arXiv preprint arXiv:2401.16355* (2024).
24. Veeling, B. S., Linmans, J., Winkens, J., Cohen, T. & Welling, M. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11*, 210–218 (Springer, 2018).
25. Kather, J. N. *et al.* Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine* **16**, e1002730 (2019).
26. Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R. & Naranjo, V. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine* **195**, 105637 (2020).
27. Arunachalam, H. B. *et al.* Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PloS one* **14**, e0210706 (2019).
28. Brummer, O., Pölönen, P., Mustjoki, S. & Brück, O. Integrative analysis of histological textures and lymphocyte infiltration in renal cell carcinoma using deep learning. *bioRxiv* 2022–08 (2022).
29. Kriegsmann, K. *et al.* Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology* **12**, 1022967 (2022).
30. Borkowski, A. A. *et al.* Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142* (2019).
31. Kumar, N. *et al.* A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging* **39**, 1380–1391 (2019).
32. Han, C. *et al.* Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis* **80**, 102487 (2022).
33. Xu, F. *et al.* Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology* **11**, 759007 (2021).
34. Roetzer-Pejrimovsky, T. *et al.* The digital brain tumour atlas, an open histopathology resource. *Scientific Data* **9**, 55 (2022).
35. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).