Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images

Richard J. Chen\textsuperscript{1,2,3,4}, Ming Y. Lu\textsuperscript{1,3,4}, Wei-Hung Weng\textsuperscript{5}, Tiffany Y. Chen\textsuperscript{1,3,4}, Drew FK. Williamson\textsuperscript{1,3,4}, Trevor Manz\textsuperscript{1,2}, Maha Shady\textsuperscript{1,2,3,4}, Faisal Mahmood\textsuperscript{1,3,4}

\textsuperscript{1}Department of Pathology, Brigham and Women’s Hospital
\textsuperscript{2}Department of Biomedical Informatics, Harvard Medical School
\textsuperscript{3}Cancer Program, Broad Institute of Harvard and MIT
\textsuperscript{4}Cancer Data Science Program, Dana-Farber Cancer Institute
\textsuperscript{5}Computer Science and Artificial Intelligence Laboratory, MIT

richardchen@g.harvard.edu, faisalmahmood@bwh.harvard.edu

Abstract

Survival outcome prediction is a challenging weakly-supervised and ordinal regression task in computational pathology that involves modeling complex interactions within the tumor microenvironment in gigapixel whole slide images (WSIs). Despite recent progress in formulating WSIs as bags for multiple instance learning (MIL), representation learning of entire WSIs remains an open and challenging problem, especially in overcoming: 1) the computational complexity of feature aggregation in large bags, and 2) the data heterogeneity gap in incorporating biological priors such as genomic measurements. In this work, we present a Multimodal Co-Attention Transformer (MCAT) framework that learns an interpretable, dense co-attention mapping between WSIs and genomic features formulated in an embedding space. Inspired by approaches in Visual Question Answering (VQA) that can attribute how word embeddings attend to salient objects in an image when answering a question, MCAT learns how histology patches attend to genes when predicting patient survival. In addition to visualizing multimodal interactions, our co-attention transformation also reduces the space complexity of WSI bags, which enables the adaptation of Transformer layers as a general encoder backbone in MIL. We apply our proposed method on five different cancer datasets (4,730 WSIs, 67 million patches). Our experimental results demonstrate that the proposed method consistently achieves superior performance compared to the state-of-the-art methods.

1. Introduction

Though deep learning has revolutionized computer vision in many disciplines, gigapixel whole-slide imaging (WSI) in computational pathology remains a complex computer vision domain with barriers that render current approaches infeasible for supervised learning tasks such as cancer prognosis. In image classification of natural images, the goal is usually to assign an image-level label to an image with approximate size 256 $\times$ 256 pixels, with the label clearly visible and well-represented in the image. In pathology, WSIs break these assumptions as images exhibit enormous heterogeneity and can be as large as 150,000 $\times$ 150,000 pixels. Depending on the problem, labels for slide-level classification may be: 1) localized in a small pixel region that occupies a tiny proportion of the total image (i.e. - a needle-in-a-haystack problem such as differentiating normal tissue vs micro-metastases) \cite{4, 3, 5, 45}, or 2) spanning the entire composition of a WSI and dependent on the interactions of its components (i.e. - a fine-grained visual recognition problem such as one that involves understanding the complex milieu of stroma, tumor aggregates, immune cells and other visual concepts) \cite{68, 7, 47, 22, 40, 21, 6, 39, 15}.

Due to the enormous gigapixel resolutions of WSIs, many approaches adopt a two-stage multiple instance learning-based (MIL) approach for tractable representation learning of WSIs, in which: 1) instance-level feature representations are extracted from randomly sampled image patches in the WSI, and then 2) global aggregation schemes are applied to the bag of instances to obtain a WSI-level representation for subsequent supervision \cite{23, 59, 11, 38, 69}. Though unable to model complex interactions between instances, MIL is able to solve many needle-in-a-haystack problems in pathology, as the classification of normal tissue vs micro-metastases depends on discriminating only binary instance-level visual concepts \cite{3, 35}. Survival outcome prediction, however, is a challenging ordinal regression task.
that aims to predict relative risk of cancer death, and fits into the latter class of fine-grained visual recognition problems [70]. In contrast to needle-in-a-haystack problems, survival outcome prediction requires modeling a heterogeneous spectrum of visual concepts in the tumor microenvironment that are indiscernable by conventional MIL approaches, e.g. - the co-localization of tumor cells with lymphocyte infiltrates that is associated with favorable prognosis, which would require modeling mid-to-long range interactions between instances in the WSI [48, 25, 1].

Though often approached as a weakly-supervised task using only gigapixel WSIs, survival outcome prediction is traditionally framed as a multimodal learning task in which genomic information can be used as an additional modality for supervision or integration. In the current state-of-the-art, the manual assessment of histology and genomics by pathologists is the gold standard for patient triage, risk assessment, stratification into treatment groups [33]. In further extending weakly-supervised learning with multimodal fusion mechanisms, survival prediction faces an additional challenge due to the large data heterogeneity gap between WSIs and genomics: WSIs represented as bags containing tens of thousands of image patches as instances, while genomic features are often represented as 1 × 1 tabular attributes. As a result, many approaches use late fusion mechanisms for feature integration, which prevents learning important multimodal interactions [42, 9, 10]. Overall, cancer prognostication using WSIs is both a difficult weakly-supervised learning and multimodal learning problem, and is a grand challenge in the characterization of disease progression of many cancer subtypes.

To address these challenges, we propose an interpretable, weakly-supervised, multimodal learning framework called MCAT (Multimodal Co-Attention Transformer) that learns a dense co-attention mapping between WSIs and genomics for interpretable survival outcome prediction. Inspired by deep learning approaches in Visual Question Answering (VQA) that learn relationships attributing how word embeddings attend to salient objects in an image when answering a question [34, 29, 64, 28, 44, 63], in our framework, we learn how instance-level histology features attend to genes when predicting patient survival. One of the key contributions of our work is that we use a cross-modality attention (or co-attention) called genomic-guided co-attention (GCA) as an early fusion strategy for identifying informative instances from a large permutation-invariant set / MIL bag using genomic features as queries (formulated in an embedding space). This yields two advantages for survival outcome prediction:

1. In comparison to late fusion-based architectures that concatenate the WSI-level bag representation with genomic features, our GCA layer captures multimodal interactions that relate histology-based visual concepts to gene embeddings similar to VQA, visualized as a WSI-level attention heatmap for each genomic embedding.

2. We demonstrate how the GCA layer also reduces the effective “sequence length” of WSI bags from \( M \) instance-level patch features to \( N \) gene-guided visual concepts, where \( N \) is the effective sequence length of set of gene embeddings (and \( M \gg N \)). This allows us to develop more sophisticated feature aggregation strategies using self-attention and Transformers for supervision using entire WSIs, which has previously been impossible. In § 3.3, we draw connections between MIL that operate on set-based data structures (bags), and Transformers.

Results in Table 1 shows that MCAT outperforms existing state-of-the-art weakly-supervised methods for survival outcome prediction using gigapixel WSIs, as well as multimodal networks (augmented from existing methods) that would conventionally integrate WSI with genomics via late fusion. We conduct our ablation study on five large publicly-available cancer datasets, and demonstrate that MCAT consistently improves on all prior approaches by 3.0% – 6.87%. Lastly, we visualize the gene-guided visual concepts as heatmaps to analyze feature interactions between WSIs and genomics, shown in Figures 2 and 3, and assess patterns that emerge from how morphological features attend to each gene. Our code is made available at: https://github.com/mahmoodlab/MCAT.

2. Related Work

2.1. Weak Supervision in Gigapixel Images

Recent work has demonstrated remarkable progress in using multiple Instance Learning (MIL) and other set-based deep learning approaches for learning tasks in gigapixel images [23, 5, 52, 51, 62, 69, 37, 36]. Edwards and Storkey [17] and Zaheer et al. [67] proposed one of the first neural network architectures for supervised learning on sets, followed by Ilse et al. [24] later extending set-based deep learning as a general framework for MIL, with applications to pathology. Xu et al. [60] proposed an MIL-based label enrichment approach for tissue semantic segmentation without pixel-annotations. Lu et al. [38], Zhu et al. [70], Yao et al. [61, 62], Zhao et al. [69] explored different strategies for global pooling over patch-based instances. Though demonstrating impressive results in cancer classification, MIL-based approaches in pathology have generally focused only on instance-level feature extraction, and have not yet explored modeling global, long-range interactions through permutation-equivariant feature aggregation techniques such as attention mechanisms.
Figure 1: Overview of the Multimodal Co-Attention Transformer (MCAT) architecture. From gigapixel WSIs and genomic features, we formulate both modalities as bags representations, from which we use: 1) Genomic-Guided Co-Attention to capture multimodal interactions, and 2) set-based MIL Transformers as feature aggregation for survival outcome prediction.

2.2. Attention in Set-Based Deep Learning

Since the seminal work by Vaswani et al. [53], attention mechanisms have since seen widespread adoption across many different domains outside of neural machine translation such as language model pretraining [14, 55, 56], visual recognition [46, 16], visual question answering [34, 29, 64, 28, 44, 63], graph neural networks [54], and point clouds [31, 18]. Outside of language modeling, Lee et al. [31] developed the Set Transformer framework, which extends the original language Transformer to general set-structured data structures such as point clouds and counting problems similar to [67]. Dosovitskiy et al. [16] proposed using Transformer architectures for vision pretraining in natural images, in which 224 × 224 images were formulated as a sequence flattened 16 × 16 image patches. Recently, Kalra et al. [26] used Set Transformers for lung cancer subtyping using bags of 100 randomly sampled histology patches. For large-scale representation learning of entire WSIs, though WSIs can be naturally formulated as a sequence / bag of histology patches, in comparison to word embeddings that have a max sequence length of at most 512, the average size of a WSI contains approximately 15,000 256 × 256 image patches at 20 × magnification, with a max sequence length of 200,000 patches. Due to the large space complexity of WSI bags, using Transformers and other stacked self-attention network architectures in MIL-related tasks is computationally infeasible.

2.3. Multimodal Deep Learning

Learning joint representations via multimodal deep learning is a challenging task due to the heterogeneous statistical properties and noise levels across modalities [43, 2]. To learn shared representations, fusion operators such as vector concatenation, element-wise sum, element-wise multiplication (Hadamard Product), bilinear pooling (Kronecker Product), and co-attention mechanisms are often used in many multimodal learning tasks such as VQA [19, 29, 28, 20], sentiment analysis [65], survival analysis [42, 9, 10], and other tasks in medicine [8, 41]. In pathology, Mobadersany et al. [42] uses vector concatenation to integrate histology and genomic features for survival outcome prediction. Later, Chen et al. [9] uses Kronecker Product fusion to integrate image, graph, and genomic-based features. Though multimodal, many of these approaches are late fusion-based, in that features are only fused towards the penultimate network layers and provide limited interpretability of multimodal interactions. Moreover, in contrast with multimodal fusion approaches in VQA that can relate image features to word embeddings using co-attention learning, current multimodal work using WSIs do not have similar interpretability mechanisms that can relate histology features in WSIs to genomics.

3. Method

In this section, we present our overall framework, the Multimodal Co-Attention Transformer (MCAT), for weakly-supervised and multimodal learning using WSIs and genomics for survival outcome prediction, illustrated in Figure 1. In § 3.1, we present our formulation of WSI and genomic representations as bags using instance-level feature extraction. In § 3.2, we present our core method, the Genomic-Guided Co-Attention (GCA) layer, which learns a dense co-attention mapping between bag representations of WSIs and genomics that can visualize multimodal interactions (Figure 2). We also demonstrate how the GCA layer is able to reduce the space complexity of WSI bags, from which in § 3.3, we adapt set-based Transformers for MIL in survival outcome prediction. In § 3.4, we discuss
implementation details, with further information about our survival loss function described in the Supplementary Materials.

### 3.1. WSI and Genomic Bag Construction

**Problem Formulation:** Multiple Instance Learning (MIL) is a weakly-supervised learning task and framework that operates on set-based data structures. These set-based data structures are also known as “bags”, in which each bag is an unordered (permutation-invariant) set of instances that can be of varying size with incomplete instance-level labels [67]. For single-label classification, given a bag \( X = \{x_1, \ldots, x_M\} \in \mathbb{R}^{M \times d_w} \) containing \( d_w \)-dimensional instances with label \( Y \), the goal is to learn a permutation-invariant function \( F \) that predicts the bag label without detailed knowledge of the instances, and has the general form:

\[
F(X) = \zeta(\rho(\{\phi(x_i) : x_i \in X\}))
\] (1)

where \( \phi : \mathbb{R}^{d_w} \rightarrow \mathbb{R}^{d_{out}} \) is an instance-level function that processes features for each instance independently, \( \rho : \mathbb{R}^{m \times d_{out}} \rightarrow \mathbb{R}^{d_{out}} \) is a symmetric, permutation-invariant aggregation function that pools the extracted features to a single bag-level feature embedding, and \( \zeta : \mathbb{R}^{d_{out}} \rightarrow \mathbb{R}^{d_{class}} \) is usually a bag-level classifier that further processes the bag-level features, which we use to estimate the hazard function in survival analysis.

In our task, let \( X \) represent patient data, \( t_{os} \in \mathbb{R}^+ \) be overall survival time (in months), \( c \in \{0, 1\} \) be right censorship status (death observed) in a single triplet observation in a dataset \( \{X_i, t_{i,os}, c_i\}_{i=1} \). In addition, let \( \{W_{ij}\}_{j=1}^{K_i} \) be the set of \( K_i \) gigapixel WSIs collected for \( X_i \), and \( g_{i,\text{attr}} \) be a vector of genomic attributes matched with \( X_i \). For ease of notation, we drop \( i \) in referring to the \( i \)th observation. Our goal is to develop a set-based neural network architecture \( F \) that integrates \( \{W_{ij}\}_{j=1}^{K_i} \) and \( g_{i,\text{attr}} \) to estimate the hazard function \( h_{\text{hazard}}(T = t | T \geq t, X) \in [0, 1] \), which measure the probability of the patient surviving after time point \( t \) (typically implemented as a Sigmoid activation after the last layer of \( F \)) [12, 27, 66]. Instead of estimating \( t_{os} \) directly, survival models output a ordinal risk value obtained via the cumulative distribution function \( f_{\text{surv}}(T \geq t, X) = \prod_{i=1}^{N}(1 - h_{\text{hazard}}(T = t | T \geq t, X)) \). We include detailed preliminaries on survival analysis in the Supplementary Materials.

**Instance Level Feature Extraction:** To represent \( \{W_j\}_{j=1}^{K_i} \) as a single bag data structure, we follow bag construction methods in conventional MIL approaches, in which instance-level feature representations are extracted from small image patches in the WSI. In contrast with previous approaches that sample image ROIs [59, 42, 11, 61, 62, 9], we construct our bag using all available tissue information across multiple WSIs for large-scale training. For all WSIs \( W_j \), we patch the tissue-containing image regions into a set of non-overlapping \( 256 \times 256 \) patches, which we use as input into an instance-level function \( \phi_{ij} \), implemented as a ResNet-50 CNN + FC layer (pretrained on ImageNet) that extracts \( d_k \)-dim feature embeddings \( h_i \in \mathbb{R}^{d_k} \times 1 \). For \( M \) total histology patches across all \( W_j \), we pack the extracted patch embeddings into a bag \( H_{bag} \in \mathbb{R}^{M \times d_k} \). In utilizing the entire tissue microenvironment across multiple WSIs, the average bag size during training and inference contains approximately \( M = 15, 231 \) instances, with some bags having up to 17 gigapixel WSIs and 230,000 instances. In conventional MIL approaches, from here, global aggregation techniques such as \( \text{SUM}(\cdot) \) can be applied to form \( h_{\text{final}} \), followed by concatenation or bilinear pooling with genomic feature vector \( g_{i,\text{attr}} \) as late fusion.

**Formulating Genes in an Embedding Space:** Genomic features such as gene mutation status, copy number variation, and bulk RNA-Seq abundance are typically quantified as \( 1 \times 1 \) measurements, or attributes, which alone do not incorporate any semantic information that would describe the functional impact of a gene in a biological system. To obtain more expressive, embedding-like feature representations similar to word embeddings in NLP, we categorize genes into \( N \) different sets with similar biological functional impact (e.g. - oncogenesis or cell differentiation). Let \( \{B_n\}_{n=1}^N \) indicate unique functional categories obtained from [49, 32]. For each genomic attribute \( \text{att}_i \in g_{i,\text{attr}} \), we assign \( \text{att}_i \) to gene set \( g_n \) if \( \text{att}_i \in B_n \), which we use as input to a genomics-based instance-level function \( \phi_g \) parameterized using a FC layer. In applying \( \phi_g \) instance-wise over all categorized gene sets, we obtain genomic embeddings \( \{g_n \in \mathbb{R}^{d_k \times 1}\}_{n=1}^N \), which we pack into a bag data structure \( G_{bag} \in \mathbb{R}^{N \times d_k} \). In our implementation, we use \( N = 6 \) functional categories obtained from [32] to define the following genomic embeddings: 1) Tumor Suppression, 2) Oncogenesis, 3) Protein Kinases, 4) Cellular Differentiation, 5) Transcription, and 6) Cytokines and Growth.

### 3.2. Genomic-Guided Co-Attention Layer

Due to the data heterogeneity gap between gigapixel WSI and genomic features, current multimodal approaches in pathology are limited to only incorporating late fusion, which does not capture interpretable genotype-phenotype interactions that exist in the tumor microenvironment. In reformulating both WSIs and genomic features as bag representations \( H_{bag} \) and \( G_{bag} \), we can develop more complex feature aggregation strategies that directly model pairwise interactions between instance-level feature embeddings in \( H_{bag} \) and genomic embeddings in \( G_{bag} \). In this section, we introduce Genomic-Guided Co-Attention (GCA), analogous to the standard Transformer attention that relate image-grid and word embeddings in VQA [53] (Fig. 2). GCA uses \( G_{bag} \in \mathbb{R}^{N \times d_k} \) to guide the feature aggregation of \( H_{bag} \in \mathbb{R}^{M \times d_k} \) into a clustered set of gene-guided visual
concepts $\hat{H}_{bag} \in \mathbb{R}^{N \times d_k}$, using the following mapping:

$$\text{CoAttn}_{G \rightarrow H}(G, H) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right)$$

$$= \text{softmax} \left( \frac{W_q G H^T W_k}{\sqrt{d_k}} \right) W_v H \rightarrow A_{coatt} W_v H \rightarrow \hat{H}$$

where $W_q, W_k, W_v \in \mathbb{R}^{d_k \times d_k}$ are trainable weight matrices multiplied to the queries $G_{bag}$ and key-value pair ($H_{bag}, H_{bag}$), and $A_{coatt} \in \mathbb{R}^{N \times M}$ is the co-attention matrix for computing the weighted average of $H_{bag}$. Distinct from VQA is the complexity of gigapixel WSI and the disparate bag sizes, in which $M = 15, 231$ and $N = 6$. For the task of multimodal survival outcome prediction, we find that: 1) the interpretability of GCA is able to scale up to hundreds of thousands of patches, providing and 2) we can use genomic embeddings in GCA to reduce the complexity of WSI bags.

**Interpretation:** Intuitively, for a single genomic embedding $g_n \in G$, the GCA layer scores the pairwise similarity for how much $h_m$ attends to $g_n$ for all $h_m \in H_{bag}$, written as a row vector $[a_{n1}, a_{n2}, \ldots, a_{nm}] \in A_{coatt}$. These attention weights are then applied element-wise to $H_{bag}$, which constructs a new WSI-level feature embedding $\tilde{h}_n \in \mathbb{R}^{N \times 1}$ that reflects the biological function of $g_n$. For example, if $g_n$ is a genomic embedding that expresses the underlying biological pathways responsible for tumor formation, $A_{coatt}$ computed by the GCA layer would saliently localize image patches containing tumor cells as high attention, which then aggregates $\tilde{h}_n$ as a WSI-level representation primarily containing tumor cells. We describe the set of high-attention image patches that attend to a single genomic embedding $g_n$ as a “gene-guided visual concept”, in which patches that are similar in feature space to $g_n$ would share similar phenotypic information. For $N$ genomic embeddings in $G_{bag}$, the GCA layer captures up to $N$ different gene-guided visual concepts, which we visualize as attention heatmaps in Figures 2 and 3.

**Space Complexity:** An important detail of GCA is that we set up $Q, K, V$ such that the bag size of $Q$ is much smaller than $K, V$. As a result, the query $G_{bag}$ aggregates $H_{bag} \in \mathbb{R}^{M \times d_k}$ containing $M$ instance-level patch embeddings as $\hat{H}_{coatt} \in \mathbb{R}^{N \times d_v}$ containing $N$ WSI-level embeddings, which makes the cost of applying subsequent self-attention layers have asymptotic complexity $O\left(N^2 d_v + N^2 d_v\right)$, which is quadratic with respect to $N$ instead of $M$.  

### 3.3. Set-based MIL Transformers with Survival Prediction

Following the observation in Zaheer et al. [67] that set-based network architectures remain permutation-invariant even if the encoder is a stack of permutation-equivariant layers, we can extend the original MIL framework using a set-based MIL Transformer, written as:

$$E^{(l)} \left( H^{(l)} \right) = \zeta^{(l)} \left( \psi^{(l)} \left( \{ \phi^{(l)} (x_i) : h_i^{(l)} \in H^{(l)} \} \right) \right)$$

$$\mathcal{F}^{(L)} \left( H^{(L)} \right) = \zeta^{(L)} \left( \rho^{(L)} \left( \{ \phi^{(L)} (x_i) : h_i^{(L)} \in H^{(L)} \} \right) \right)$$

$$\mathcal{T} (X) = \mathcal{F}^{(L)} \left( \epsilon^{(L-1)} (\ldots \epsilon^{(1)} (\{ \{ x_i : x_i \in X \} ) ) \right)$$

in which $h_i^{(l)}$ is an arbitrary embedding in the set input $H^{(l)}$ at hidden layer $l$, $E^{(l)}$ is a stackable encoder block that replaces $\phi$ in Equation 1 with a permutation-equivariant set function $\psi, \phi$ and $\zeta$ are permutation-invariant functions applied to feature embeddings (either instance-level or bag-level), $\mathcal{F}^{(L)}$ is the original MIL network but now applied
as a global pooling function as the last layer $L$, and $T$ is the set-based MIL Transformer [53, 31] that uses stacked permutation-equivariant layers followed by a permutation-invariant pooling function. To show that $E^{(l)}$ is the encoder block in Transformers, let $\zeta^{(l)}$ be a position-wise FC layer, $\psi^{(l)}$ be the self-attention layer in [53], and note that the position-wise residual mapping and LayerNorm operations retains permutation-invariance. We can also write $\psi^{(l)}$ more explicitly as the permutation-equivariant set function:

$$
\psi^{(l)}\left(\{h_i^{(l)}\}_{i=1}^M\right) = \left\{ \frac{\exp[h_i^{(l)}h_j^{(l)T}]}{\sum_{i=1}^M \exp[h_i^{(l)}h_j^{(l)T}]} \cdot h_i^{(l)} \rightarrow h_i^{(l+1)} \right\}
$$

(4)

in which permuting the set $\{h_i^{(l)}\}$ permutes the update $\{h_i^{(l+1)}\}$ from the output of $\psi^{(l)}$ in the same order. From this formulation, we can observe that Transformers are a generalization of the shallow set-based data structure commonly used in Equation 1, in which we can compose arbitrary hidden layers with permutation-equivariant pooling functions before global pooling. Using $\tilde{H}_{\text{coatin}} \in \mathbb{R}^{N \times d_x}$ and $G_{\text{bag}} \in \mathbb{R}^{N \times d_x}$ as inputs, we construct two MIL Transformers $T_H, T_G$ to aggregate feature embeddings in $\tilde{H}_{\text{coatin}} \otimes G_{\text{bag}}$. In the process for aggregating features in $\tilde{H}_{\text{coatin}}$, $\psi^{(l)}$ is used to model complex, long-range feature interactions between genomic-guided visual concepts that would otherwise be intractable using the original WSI bag with large $M$.

To implement $\rho_H, \rho_G$, following [24], we use the global attention pooling function $F_{\text{attpool}}(\cdot)$ to adaptively compute a weighted sum of all embeddings within each respective set to finally construct bag-level features $h^{(L)}, g^{(L)}$.

$$
\phi^{(L)}(h_i^{(l)}) = W_{\phi}h_i^{(l)}
$$

$$
\rho^{(L)}\left(\{h_i^{(L)}\}_{i=1}^M\right) = \sum_{i=1}^M a_i \phi^{(L)}(h_i^{(L)}) \rightarrow h^{(l)}
$$

where $W_{\phi}, W_{\rho}, V_{\phi}, V_{\rho}, U_{\phi}, U_{\rho}, W_{\zeta} \in \mathbb{R}^{d_L \times d_x}$ are trainable weight matrices, $\phi^{(L)}$ and $\rho^{(L)}$ are instance-level and bag-level FC layers respectively, $\rho^{(L)}$ is the global attention pooling operator, and $a_i$ scores how much to weigh embedding $h_i^{(L)}$ in the bag-level features $h^{(L)}$.

As a final step, we integrate bag-level features from the output of $T_G, T_H$ using simple vector concatenation of $s_h^{(L)}(h^{(L)}), \zeta_g^{(L)}(g^{(L)})$, which we process using several $\tilde{F}_C$ layers to obtain the final shared representation $h_{\text{final}}$.

### 3.4. Implementation Details

MCAT is implemented in PyTorch and trained on a commercial workstation with 4 NVIDIA GTX 2080Ti GPUs. Functional categories used to define the genomic embeddings were obtained from [32], which categorizes genes into the aforementioned $N = 6$ categories based on similar biological functional impact. During training, we used Adam optimization with a learning rate of $2 \times 10^{-4}$, weight decay of $1 \times 10^{-5}$. Due to samples having varying bag sizes, we use a batch size of 1, with 32 gradient accumulation steps.

### 4. Experiments

#### 4.1. Datasets & Evaluation Metrics

To validate our proposed method, we used the five largest cancer datasets from The Cancer Genome Atlas (TCGA), a public cancer data consortium that contains matched diagnostic WSIs and genomic data with labeled survival times and censorship statuses\(^\dagger\). For this study, we used the following cancer types: Bladder Urothelial Carcinoma (BLCA) ($n = 437$), Breast Invasive Carcinoma (BRCA) ($n = 1022$), Glioblastoma & Lower Grade Glioma (GBMLGG) ($n = 1011$), Lung Adenocarcinoma (LUAD) ($n = 515$), and Uterine Corpus Endometrial Carcinoma (UCEC) ($n = 538$). For each patient sample, we collected all diagnostic WSIs used for primary diagnosis, which resulted in 4,370 WSIs collected with an average bag size of 15,231 256×256 patches per image (approx 5 TB of gigapixel images, 67 million patches). For each cancer dataset, we trained our proposed method in a 5-fold cross-validation, and used the cross-validated concordance index (c-Index) to measure the predictive performance of correctly ranking the predicted patient risk scores with respect to overall survival.

#### 4.2. Comparisons with State-of-the-Art

Using the same 5-fold cross-validation splits for evaluating MCAT, we implemented and evaluated several state-of-the-art methods used for survival outcome prediction in computational pathology, training in total 275 models. For all methods, we use the same instance-level feature extraction pipeline for bag construction of WSIs, as well as identical training hyperparameters and loss function for supervision. Table 1 shows the results of all methods on all five cancer dataset benchmarks.

1. **SNN** [30]: As a unimodal baseline for only genomic features, we trained a feedforward network using the Self-Normalizing Network (SNN) architecture from Klambauer et al. [30], which has been used previously for survival outcome prediction in the TCGA [9].

2. **Deep Sets** [67]: One of the first neural network architectures for set-based deep learning, which proposes sum pooling over instance-level features.

3. **Attention MIL** [24]: A set-based neural network architecture that replaces sum pooling in Deep Sets with

\(^\dagger\)https://gdc.cancer.gov
global attention pooling, in which instances are adaptively weighted using a Softmax function [60, 62].

4. **DeepAttnMISL** [62]: The current state-of-the-art for unimodal survival outcome prediction using WSIs. DeepAttnMISL first applies K-Means clustering to instance-level features, followed by processing each cluster using Siamese networks and then aggregating the cluster features using global attention pooling.

5. **Multimodal Comparisons**: As multimodal comparisons to MCAT, we augment the previous set-based network architectures with two common late fusion mechanisms to integrate bag-level WSI features and genomic features: 1) concatenation [42], and 2) bilinear pooling [19, 65, 57, 9].

<table>
<thead>
<tr>
<th>Model</th>
<th>BLCA</th>
<th>BRCA</th>
<th>GBMLGG</th>
<th>LUAD</th>
<th>UCEC</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNN (Genomic Only) [30]</td>
<td>0.541 ± 0.016</td>
<td>0.466 ± 0.058</td>
<td>0.598 ± 0.054</td>
<td>0.539 ± 0.069</td>
<td>0.493 ± 0.096</td>
<td>0.527</td>
</tr>
<tr>
<td>Deep Sets (WSI Only) [67]</td>
<td>0.500 ± 0.000</td>
<td>0.500 ± 0.000</td>
<td>0.498 ± 0.014</td>
<td>0.496 ± 0.008</td>
<td>0.500 ± 0.000</td>
<td>0.499</td>
</tr>
<tr>
<td>Deep Sets (Concat)</td>
<td>0.604 ± 0.042</td>
<td>0.521 ± 0.079</td>
<td>0.803 ± 0.046</td>
<td>0.616 ± 0.027</td>
<td>0.598 ± 0.077</td>
<td>0.629</td>
</tr>
<tr>
<td>Deep Sets (Bilinear Pooling)</td>
<td>0.589 ± 0.050</td>
<td>0.522 ± 0.029</td>
<td>0.809 ± 0.027</td>
<td>0.558 ± 0.038</td>
<td>0.593 ± 0.055</td>
<td>0.614</td>
</tr>
<tr>
<td>Attention MIL (WSI Only) [24]</td>
<td>0.536 ± 0.038</td>
<td>0.564 ± 0.050</td>
<td>0.787 ± 0.028</td>
<td>0.559 ± 0.060</td>
<td>0.625 ± 0.057</td>
<td>0.614</td>
</tr>
<tr>
<td>Attention MIL (Concat)</td>
<td>0.605 ± 0.045</td>
<td>0.551 ± 0.077</td>
<td>0.816 ± 0.011</td>
<td>0.563 ± 0.050</td>
<td>0.614 ± 0.052</td>
<td>0.630</td>
</tr>
<tr>
<td>Attention MIL (Bilinear Pooling)</td>
<td>0.567 ± 0.034</td>
<td>0.536 ± 0.074</td>
<td>0.812 ± 0.005</td>
<td>0.578 ± 0.036</td>
<td>0.562 ± 0.058</td>
<td>0.611</td>
</tr>
<tr>
<td>DeepAttnMISL (WSI Only) [62]</td>
<td>0.504 ± 0.042</td>
<td>0.524 ± 0.043</td>
<td>0.734 ± 0.029</td>
<td>0.548 ± 0.050</td>
<td>0.597 ± 0.059</td>
<td>0.581</td>
</tr>
<tr>
<td>DeepAttnMISL (Concat)</td>
<td>0.611 ± 0.049</td>
<td>0.545 ± 0.071</td>
<td>0.805 ± 0.014</td>
<td>0.595 ± 0.061</td>
<td>0.615 ± 0.020</td>
<td>0.634</td>
</tr>
<tr>
<td>DeepAttnMISL (Bilinear Pooling)</td>
<td>0.575 ± 0.032</td>
<td>0.577 ± 0.063</td>
<td>0.813 ± 0.022</td>
<td>0.551 ± 0.038</td>
<td>0.586 ± 0.036</td>
<td>0.621</td>
</tr>
</tbody>
</table>

**MCAT (Ours)** | **0.624 ± 0.034** | **0.580 ± 0.069** | **0.817 ± 0.021** | **0.620 ± 0.032** | **0.622 ± 0.019** | **0.653** |

Table 1: Ablation study results assessing c-Index performance of MCAT against several state-of-the-art deep learning-based methods across 5 different cancer datasets.

To visualize the genomic-guided WSI embeddings used as input in our set-based MIL Transformer, we overlay the co-attention weights computed for each histology patch \( h_m \), attending to each genomic embedding \( g_n \) with visual assessment from two pathologists, shown in Figure 3 for both a low and high risk case in the BRCA dataset. In addition, we also used Integrated Gradients [50] to visualize the top-10 genes in each embedding with the highest absolute attribution value.

Overall, we observe that the genomic embeddings used in guided co-attention were able to reflect many known genotype-phenotype relationships in cancer pathology. In BRCA, genomic-guided WSI embeddings for tumor suppression, protein kinases and cellular differentiation generally reflected normal stroma, glands, and adipocytes. For

4.4. **Attention Visualization**
all cases, high attention in the cellular differentiation embedding focused on tumor-associated stroma, while both the tumor suppression and protein kinases embeddings had greater high attention given to stroma adjacent to adipocytes and glandular structures. These findings corroborate the attribution of genes such as FUT3 synthesized by epithelial cells, and TGFB1 used in regulating cell growth. In the oncogenesis and transcription embedding, high attention weights localized image regions such as invasive, high-grade tumor morphology such as dense tumor cellularity and tumor-infiltrated stroma (opposite of tumor suppression). In the cytokines embedding, we observe that high attention regions were focused on immune cells infiltrating normal stroma and tumor cells. Additional visualizations can be found in the Supplementary Materials.

5. Conclusion

In this work, we present the Multimodal Co-Attention Transformer (MCAT) for survival outcome prediction in pathology. Our method formulates both gigapixel WSIs and genomic features as permutation-invariant sets, from which we develop more sophisticated feature aggregation strategies in MIL via transformer attention. A limitation in our current study is that we used a previously-curated gene set with potentially overlapping biological functional impact. Future work would focus on investigating early fusion of WSIs with more fine-grained, distinct biological gene sets, and further quantification of phenotype-genotype correspondences.

6. Acknowledgements

We thank Felix Yu and Quinn Zhen for their insightful feedback. This work was supported in part by internal funds from BWH Pathology, Nvidia GPU Grant Program, and NIGMS R35GM138216 (F.M.). R.J.C. and T.M. were supported by the NSF Graduate Fellowship.
References


