GTP-4o: Modality-prompted Heterogeneous Graph Learning for Omni-modal Biomedical Representation

Chenxin Li[1], Xinyu Liu[1⋆], Cheng Wang[1⋆], Yifan Liu[1], Weihao Yu[1], Jing Shao[2], and Yixuan Yuan[1]

[1] The Chinese University of Hong Kong
[2] Shanghai AI Laboratory

Abstract. Recent advances in learning multi-modal representation have witnessed the success in biomedical domains. While established techniques enable handling multi-modal information, the challenges are posed when extended to various clinical modalities and practical modality-missing setting due to the inherent modality gaps. To tackle these, we propose an innovative Modality-prompted Heterogeneous Graph for Omni-modal Learning (GTP-4o), which embeds the numerous disparate clinical modalities into a unified representation, completes the deficient embedding of missing modality and reformulates the cross-modal learning with a graph-based aggregation. Specially, we establish a heterogeneous graph embedding to explicitly capture the diverse semantic properties on both the modality-specific features (nodes) and the cross-modal relations (edges). Then, we design a modality-prompted completion that enables completing the inadequate graph representation of missing modality through a graph prompting mechanism, which generates hallucination graphic topologies to steer the missing embedding towards the intact representation. Through the completed graph, we meticulously develop a knowledge-guided hierarchical cross-modal aggregation consisting of a global meta-path neighbouring to uncover the potential heterogeneous neighbors along the pathways driven by domain knowledge, and a local multi-relation aggregation module for the comprehensive cross-modal interaction across various heterogeneous relations. We assess the efficacy of our methodology on rigorous benchmarking experiments against prior state-of-the-arts. In a nutshell, GTP-4o presents an initial foray into the intriguing realm of embedding, relating and perceiving the heterogeneous patterns from various clinical modalities holistically via a graph theory. Project page: https://gtp-4-o.github.io/.

Keywords: Biomedical Data · Multimodal Learning · Graph Networks

1 Introduction

Each modality has its own perspective to reflect the specific data characteristics [30, 33, 35, 75, 87]. Integrating multi-modal data empowers the models with

[⋆] Equal second-author contribution

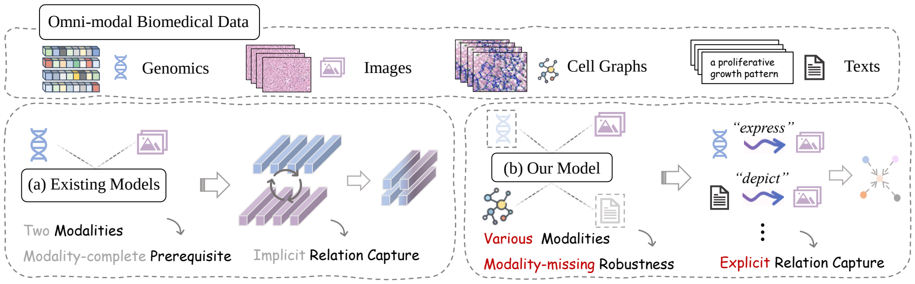

Omni-modal Biomedical Data
Genomics
Images
Cell Graphs
a proliferative growth pattern
Texts
(a) Existing Models
Two Modalities
Modality-complete Prerequisite
Implicit Relation Capture
(b) Our Model
Various Modalities
Modality-missing Robustness
“express”
“depict”
Explicit Relation Capture

Fig. 1: Methodology Comparison. Unlike **(a)** prior methods, **(b)** our framework enables learning unified omni-modal representation from various clinical modalities with modality missing and explicit capture of the cross-modal relations through the established heterogeneous graph representation.

various insights into the conditions of subjects at the macroscopic, microscopic, and molecular levels, and allows for an accurate and comprehensive disease diagnosis [40,44,52,68,69,77]. For instance, multimodal fusion of various imaging techniques has significantly improved gastrointestinal lesion detection and characterization in endoscopic scenes [28,32,49,56]. Similarly, incorporating genomic information with pathological images can improve the prediction accuracy of cancer grading [5,6,16,72,76]. A relevant task, survival prediction, which aims to predict the time interval to a significant event such as death or disease relapse, can also benefit from such multi-modal inclusion [7]. Besides, the cell graphs constructed by the cell nuclei segmentation of pathological images, are shown to provide more fine-grained microscopic information [71]. Recent advances in visual language models also sparks the works in learning from biomedical images and texts [80], whereby the diagnostic texts usually encapsulates abstract semantic information [11,13]. These progress presents potential for extending the capacity boundary of biomedical multi-modal models to omni-modal representation to handle a broader range of clinical modalities.

Established multimodal methods typically follow the principle that first extracts uni-modal features, then learns cross-modal relations in paired multimodal data [42,60,73,80]. Early researches design meticulous fusion techniques for multimodal information and wish to maximize the benefits of each modality [19,41]. Due to the imbalanced learning process with inherent modal disparity and heterogeneity [22,81,82], recent efforts pivot towards improving collaborative learning of multiple modalities by balancing and adjusting learning of each modality [15,59,74]. Through deriving the modality-relevant weighting factors, these methods dynamically modulates the learning and fusing of multimodal information on features [74], gradients [59], attentions [15,50], etc.

Despite the success in alleviating the modality gap, the challenge remains severe when applied to (especially a broad range of) biomedical modalities, primarily due to the two featured challenges. The first challenge lies in the large semantic heterogeneity exhibiting on biological modalities. A straightforward example is that an *"dog"* in natural images shares similar object-related semantics its sound, while the semantic relation and local correspondence between genomic profiles and pathological images is highly ambiguous [5, 6, 43, 53, 72]. Through prior methods employ optimal transport [72] or cross-modal attention [7] to capture fine-grained correlation across genes and images, they still overlook the heterogeneity in a high-order space, i.e., relations across modalities. Every two modalities have their own relation with specific semantics and attributes. As shown in Fig. 1, the relation across images and genomics is semantically related to *"express"*, while that across images and texts could be abstracted as *"depict"*. Therefore, these observations inspire us to introduce a unified non-Euclidean representation that explicitly captures the heterogeneous attributes on both modal features and cross-modal relations.

Secondly, in clinical practice, it is common to encounter partial absence in some modalities due to privacy and ethical considerations. The limitations in data collection technology and the concerns surrounding bioinformatics security make it more challenging to access all the data modalities. However, most multimodal methods have a common assumption on the data completeness [60,73,80]. Once a modality is missing regardless of training or testing, the multimodal fusion becomes unreachable, which leads to sub-optimal performance [26]. Therefore, we are committed to designing algorithms to adaptively complete the feature space messed up by the missing of modality such that all the representation from the missing and existing modality could be handled in a unified fashion.

To address the aforementioned challenges, we propose a modality-prompted heterogeneous graph framework for omni-modal learning (GTP-4o) that allows unifying representations under various biomedical modalities with potential modality missing. Specially, we establish a heterogeneous graph embedding [38,39,51] to explicitly capture the heterogeneous attributes on both modal features and cross-modal relations. Then, we design a modality-prompted completion that completes the deficient graph embedding of missing modality through a novel graph prompting module, which generates hallucination nodes to steer the embedding towards the original complete space. Through the completed graph, we meticulously develop a knowledge-guided hierarchical aggregation that includes a knowledge-derived global meta-path neighbouring to capture the potential heterogeneous neighbors, and a local multi-relation aggregation for the comprehensive interaction of modal information across various heterogeneous relations. GTP-4o presents the first exploration in learning unified representations from various heterogeneous clinical modalities including genomics, pathological images, cell graphs, and diagnostic texts. Our contributions are as follows:

- This paper introduces the new problem of learning unified multimodal representations from various diverse clinical modalities, and presents the first effort to embed and relate heterogeneous multimodal features through a graph representation and aggregation.
- We propose a modality-prompted completion module to complete the corrupted graph embedding of the missing modality by a graph prompting

strategy, which generates hallucination nodes to steer the missing embedding towards the complete representation.

- We present a knowledge-guided hierarchical cross-modal aggregation, employing a global meta-path neighbouring to capture heterogeneous neighbors, and a local multi-relation aggregation module for information interaction across various heterogeneous relations.
- Extensive experiments on comprehensive benchmarks of disease diagnosis including pathological glioma grading and survival outcome prediction exhibits the efficacy of our method against prior state-of-the-arts.

2 Related Work

Biomedical Multimodal Learning. Utilizing multimodal data has gained significant attention for accurate and comprehensive imaging analysis [9,36,55,87] and diagnosis [2,10,64,76]. For instance, the comprehensive features from pathological images [57,89], genomics [46,48,78], are employed in joint for an accurate cancer-related diagnosis, e.g. glioma grading [14] and survival analysis [7]. Meanwhile, with the eGTP-4once of Visual Language Models (VLMs), much efforts have been devoted to enhancing the recognition and analysis ability of vision models by further incorporating the textual information from clinical text reports [80,83]. Inspired by the trends, this paper introduces the new problem of learning unified features from various disparate clinical modalities, including genomics [7,46], pathological images [57], cell graphs [71] and text descriptions [83].

Handling the modality heterogeneity is critical when integrating multimodal information [47]. Early works focus on studying early or late fusion methods [5,19,21,61,66], which integrates the predictions from individually separated models for the final decision. However, these methods suffer from either neglecting intra-modality dynamics [47] or failing to fully relating cross-modal information. Recent progress in intermediate fusion [6,25] has shown promise, which learns uni-modal features and capture cross-modal interactions at the same time, by leveraging the power of cross-modal attention [6,54,72,86]. However, they try to model all potential cross-modal relations with the learned attentions. Different from them, we present to explicitly capture the heterogeneity of modal features and cross-modal relations resorting to a heterogeneous graph space.

Graph Representation in Pathology. Graph representation has shown its promise in the field of pathology analysis [3, 4, 67]. Following previous efforts of Multiple Instance Learning (MIL) that split the high-resolution whole slide pathological images (aka., WSIs) into a bag of instances and pre-define the connective local areas in the Euclidean space, recent graph-based methods [3, 4, 17, 85] models the interactions among instances flexibility via the graphs topology. For instance, PatchGCN [4] models pathological images with homogeneous graphs, and regress survival data with a graph convolutional neural network (GCN) [23]. GTNMIL is designed as a graph-based MIL using graph transformer networks [85]. Recent methods [3,17] extend the prior practice to handling WSIs with heterogeneous graphs, introducing heterogeneity in each patch by different

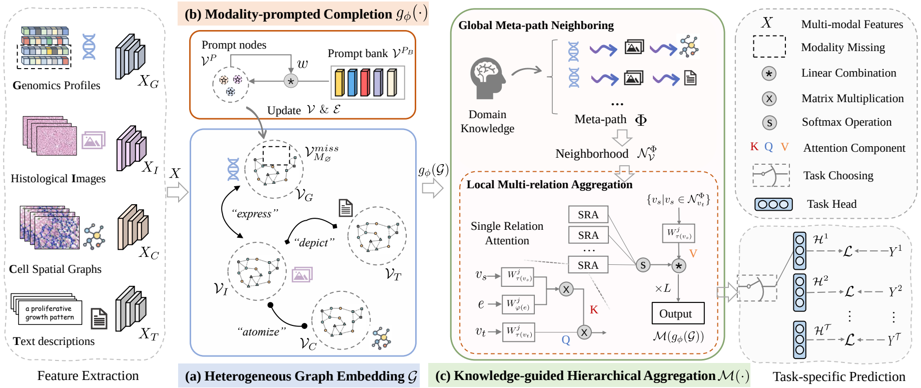


Fig. 2: Pipeline Overview of GTP-4o. We instantiate the omni-modal biomedical features (Sec. 3.1), and embed them onto (a) the heterogeneous graph space (Sec. 3.2). Then, we introduce (b) the modality-prompted completion via graph prompting to complete the missing embedding (Sec. 3.3). After that, we design (c) the knowledge-guided hierarchical aggregation from a global meta-neighbouring to uncover the heterogeneous neighbourhoods and a local multi-relation aggregation to interact features across various heterogeneous relations (Sec. 3.4).

resolution levels [17], or semantic representations via pretext tasks [3]. However, these methods only considers heterogeneity in the image modality, while the more challenging multimodal scenario is left to study. To fill this gap, this paper explores the graph representation in a way more complex setting, i.e., learning from various disparate clinical modalities with significant heterogeneity.

3 Method

Overview. After performing data processing and feature extraction (Sec. 3.1), the omni-modal embedding for a patient subject could be represent by a 4-tuples of four modalities, including genomics (G), pathological images (I), cell spatial graphs (C) and diagnostic texts (T), $X = \{X_G, X_I, X_C, X_T\}$, with different number of instances in each modality, while the common dimension d. Then, we establish the heterogeneous graph representation $\mathcal{G}$ by transforming modal features to the graph space (Sec. 3.2). After that, the modality-prompted completion is performed, which employs a graph prompting $g_\phi(\cdot)$ to transform the incomplete graphic embedding to a prompted and completed representation $g_\phi(\mathcal{G})$ (Sec. 3.3). Afterwards, we conduct a knowledge-guided hierarchical aggregation that is parameterized by $\mathcal{M}$, including a global neighbouring via knowledge-derived meta-paths Φ , and a local multi-relation aggregation along various heterogeneous relations (Sec. 3.4). The final aggregated features $\mathcal{M} \circ g_\phi(\mathcal{G})$ end up with forwarding a task-specific head $\mathcal{H}^\mathcal{T}$ to obtain the specific prediction for task $\mathcal{T}$, based on which we we optimize the network parameters $\mathcal{M}, \mathcal{H}^\mathcal{T}$ and prompt parameters g_ϕ *w.r.t.* the task loss $\mathcal{L}$. (Sec. 3.5).

3.1 Data Processing and Feature Extraction

Genomic Profiles. Following [5], the genomic profiles that we use includes Copy Number Variation (CNV), bulk RNA-Seq expressions and mutation status [5]. We merge the mutation status and CNV and feed it and RNA-Seq as separate groups of gnomic data into Self-Normalizing Neural Network (SNN) [24] to get the embedding $X_G \in \mathbb{R}^{N_G \times d}$, where N_G equals to the number of genomic groups.
Pathological Images. Following [6, 63], we divide WSIs into a series of non-overlapping patches, employ a ImageNet pretrained ResNet-50 to extract the features from each patch, and feed them into a projection layer to obtain $X_I \in \mathbb{R}^{N_I \times d}$, where N_I is the number of patches.
Cell Spatial Graphs. Cell graph representations explicitly capture selected fine-grained features of cells [58]. We utilize the procedure in PathomicFusion [5] to segment cells for each slide, curate graph topology. and use a graph convolutional network (GCN) [23] backbone to obtain the aggregated graph embedding, as $X_C \in \mathbb{R}^{N_c \times d}$, where N_C equals to the number of curated cell graphs.
Text Descriptions. As no actual medical reports are provided in the used materials, we employ an open-source multimodal Large Language Model (LLM), MiniGPT-4 [88] to generate the customized text descriptions for each pathological image[3]. The prompt is customized as follows.
Prompt: *"This is a pathology slide with glioma cells. Write a caption for this slide based on the following properties:* ❶ *size and shape of cells,* ❷ *color of cells,* ❸ *growth pattern and cellularity of cells,* ❹ *uclear atypia and pleomorphism of cells,* ❺ *necrosis of cells,* ❻ *microvascular proliferation of cells,* ❼ *mitotic activity of cells."*
We obtain several sentences for each slide, like: *"Some cells show signs of necrosis, with dark spots in the cytoplasm"*. We take each sentence as a modal instance, and employs a MedBERT [62] to obtain the embedding $X_T \in \mathbb{R}^{N_T \times d}$, where N_T equals to the number of curated sentences.

3.2 Heterogeneous Graph Embedding

With the obtained modal features $X = \{X_G, X_I, X_C, X_T\}$ with the same dimension d, the heterogeneous graph embedding could be established by a feature to graph transformation. Formally, a heterogeneous graph space is formulated by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}\}$, where $\mathcal{V}$ and $\mathcal{E}$ represents the set of entities (i.e., vertices or nodes) and relations (i.e., edges) that has been established in the theory of a classic directed graph. The further introduced $\mathcal{A}$ and $\mathcal{R}$ represent the attribute set of nodes and edges, respectively, by which we can explicitly define the heterogeneous properties for the features of modal entities and cross-modal relations. A function $\tau(v) = a \in \mathcal{A}$, is defined to map each node v to an attribute in the set $\mathcal{A}$, according to its modality, As a result, the attribute set of nodes can be formulated as $\mathcal{A} = \{G, I, C, T\}$. Furthermore, the edges $e \in \mathcal{E}$ in the heterogeneous embedding represent the relations from the source nodes $v_s \in \mathcal{V}_s$ to the

[3] We find that another foundation VLM, BLIP-2 [37] showing effective on curating captions for 3D representation [27,31,73] does not work well for pathological images.

target nodes $v_t \in \mathcal{V}_t$[4], therefore the attribute of an edge $e : v_s \rightarrow v_t$ is determined is determined by the attribute of the source node v_s and target node v_t as well as their actual semantic relations. Thus, a function $\varphi(e) = r \in \mathcal{R}$ that maps each edge $e \in \mathcal{E}$ to a specific attribute $r \in \mathcal{R}$ is introduced. We formulate this attribute set of relations $\mathcal{R}$ by the prior knowledge of biomedical modalities, $\mathcal{R} = \{$*"express"*, *"depict"*, *"atomize"*, *"intra-modal"*, $\}$. This set represents semantic relations of *"express"* between genomics and images, *"depict"* between images and texts, and *"atomize"* between images and cell graphs. We also model all relations between *"intra-modal"* instances. To obtain initial input graph embedding $\mathcal{V}$, we perform a non-linear projection on the modal features X. Edge embeddings are computed as cosine correlations between head and tail nodes.

3.3 Modality-prompted Completion

The modality-prompted completion aims to adapt the deficient embedding of missing modality by updating its with some prompted entities that could be learnt. Formally, we introduce a graph prompt operation, which could be parameterized by g_ϕ, transforming a input graph representation $\mathcal{G}$ into $g_\phi(\mathcal{G})$. Hopefully, it is learnt to transform the missing graph embedding back to its original complete status.

General Prompting. Given the missing modality $M_\varnothing \in \{G, I, C, T\}$, some specific subjects have all the instances of that modality missed, such that the representation of them at $M_\varnothing$ is ruined, as $\mathcal{V}_{M_\varnothing} = \{\varnothing\}$. There are also some patient subjects not affected by the missing, still with the complete data and the representation $\mathcal{V}_{M_\varnothing}$ maintained. We sample hallucination nodes $v^P \in \mathcal{V}^P$, where $\mathcal{V}^P \in \mathbb{R}^{N_P \times d}$, as a basic prompt scheme for graph completion. We extract the representation prior of the missing modality $M_\varnothing$ by collecting the modality-specific feature from all subjects, except for the subjects with the incomplete data at the missing modality, $\mathcal{V}_{M_\varnothing}$ The subjects with the incomplete data at the missing modality, i.e., $\mathcal{V}_{M_\varnothing} = \{\varnothing\}$ Then we initialize the features of the N_P prompt entities by a Gaussian sampling from the extract modality prior, motivated by the intuition that the same modality among different subjects share a basically similar distribution. After initialized, the set of prompt nodes $\mathcal{V}^P$, could be optimized effectively along with the model training.

Entity-dependent Prompting. The introduced prompted entities are agnostic to the context, bringing the risk of yielding sub-optimal results. To encode the entity-dependent contextual information, we further introduce a prompt bank that contains a set of prompt components, $\mathcal{V}^{P_B} \in \mathbb{R}^{N_B \times d}$, where N_B is the number of prompt components. We take the these components as a series of base prompt, the weights $\boldsymbol{w}$ of which could be obtained in an entity-independent fashion. That is, we pass each input node v^P through a channel-downscaling linear layer to obtain a compact feature vector, followed by a softmax operation, thus yielding the weights $w \in \mathbb{R}^{N_B}$,

$$w = \texttt{Softmax}\left(W_{d \times N_B}(v^P)\right), \tag{1}$$

[4] For simplicity, index s and t is omitted when cross-modal relations are not involved.

where $W_{d\times N_P}$ denotes the linear projection layer that transforms the feature dimension from d to N_P. Then we use these weights to modulate the prompt components for each query entity v^P, and sum the general prompts and the entity-dependent prompts that perceives the graphic context,

$$v^P \leftarrow v^P + \sum_{i=1}^{N_B} w_i \cdot v_i^{P_B}, \tag{2}$$

where $v_i^{P_B}$ denotes i-th component in the prompt bank. By doing so, the prompted graph embedding could be described with the formulation of graph prompt function $g_\phi(\cdot)$, by which the nodes $\mathcal{V}$ and edges $\mathcal{E}$ of a graph $\mathcal{G}$ to prompt would be transformed as,

$$\mathcal{V} = \{\mathcal{V}_{/M_\varnothing}, \mathcal{V}^P\},\ \mathcal{E} = \{\mathcal{E}_{/M_\varnothing}, \underset{\forall\varphi(e)\in\mathcal{R}}{\texttt{EdgeUpdate}}(\mathcal{V}_{/\mathcal{M}_\varnothing}, \mathcal{V}^{\mathcal{P}})\} \tag{3}$$

where $\mathcal{V}_{/M_\varnothing}$ denotes the node embedding at all the modalities except for $M_\varnothing$ and $\mathcal{E}_{/M_\varnothing}$ denotes the edge space when removing all the nodes at the modality $M_\varnothing$. $\underset{\forall\varphi(e)\in\mathcal{R}}{\texttt{EdgeUpdate}}(\cdot,\cdot)$ defines the operation that updates the features of edge e between two sets of nodes if the relation can be retrieved in the attribute space of edge $\varphi(e) \in \mathcal{R}$. Effectively, the graph embedding of the missing modality $M_\varnothing$ are adapted through inserted with the prompt nodes as well as uncovered with some ruined relations.

3.4 Knowledge-guided Hierarchical Aggregation

With the completed graph $g_\phi(\mathcal{G})$, the knowledge-guided hierarchical aggregation module effectively embeds the knowledge prior into a series of meta-paths, thereby we can search for the global cross-modal heterogeneous neighboring. With the found neighbouring, the local multi-relation aggregation module is performed across various heterogeneous edges, and the overall hierarchical aggregation [34,84] module can be parameterized by a network function $\mathcal{M}(\cdot)$.

Global Meta-path Neighbouring. The aggregation of graph information highly depends on the established neighboring rules [65,79], and we design novel meta-paths as global information pathways, allowing for interaction of two heterogeneous entities. Given the entities $\mathcal{V}$ and meta-paths $\varPhi$ in the heterogeneous graph, the neighbors derived from meta-paths for all the entities are uniquely identified [79], as $\mathcal{N}_{\mathcal{V}}^{\varPhi}$. Hence, our insight is to embed the domain knowledge into the formulation of $\varPhi$ by considering the semantic relations across the clinical modalities. Recall that the edge attribute space $\mathcal{R}$ is explicitly defined by the biological relations among modalities, we derive that, with the exception of *"intra-modal"* relations, i.e., the entities at the same modality, all heterogeneous nodes can only interact with the nodes whose attributes are semantically related to themselves in a single-hop propagation [45, 79]. Following this principle, we adopt a random walking strategy [12] to search for the optimal meta-paths from all potential candidates. Specially, we randomly start from an entity of one modality, then iterate over all the heterogeneous nodes with valid semantic relations in the attribute space, i.e., $\varphi(e_c) \in \mathcal{R}$. Repeating the iterations, we show that an appropriate customization of meta-paths could be $\Phi = \{G \xrightarrow{\text{“express”}} I \xrightarrow{\text{“atomize”}} C,\quad G \xrightarrow{\text{“express”}} I \xrightarrow{\text{“depict”}} T,\quad C \xrightarrow{\text{“atomize”}} I \xrightarrow{\text{“express”}} G,\quad T \xrightarrow{\text{“depict”}} I \xrightarrow{\text{“express”}} G\}$. Following [79], all meta-paths Φ are formulated with the maximum lengths within two hops. In practical usage, when querying the neighbourhood $\mathcal{N}_v^{\Phi}$ for an entity v, we first project it onto the entity attribute space $\mathcal{A}$ by $\tau(v)$, and then iterate over all meta-paths in Φ with a given number of hops H. After that, all reached entities attributes are collected, and the entities of those collected attributes are taken as the neighbours along the meta-paths,

$$\mathcal{N}_v^{\Phi} = \left\{ v' | \tau(v') \in \{ \underset{i\in[1,|\Phi|]}{||} \underset{H}{\texttt{Reach}}(\Phi_i)\} \right\}, \tag{4}$$

where $\underset{H}{\texttt{Reach}}$ denotes the operation that collects all the reached attributes by walking along a meta-path Φ_i with H hops. $||_{i\in[1,|\Phi|]}$ is the concatenation operator for all the resulted elements.

Local Multi-Relation Aggregation. With the derived entity-wise $\mathcal{N}_v^{\Phi}$ neighbours, we perform the information propagation for each target node $v_t \in \mathcal{V}_t$ as a local feature aggregation from all its neighbored source nodes $\mathcal{V}_s$. To model node-wise interaction [3,18], we introduce a Multi-Head Attention (MHA) mechanism that models the target node features as **Q**uery and source node features as **K**ey and **V**alue. We embed the target node v_t and source node v_s by different linear projection layers $W^j_{\tau(v_s)}$ and $W^j_{\tau(v_s)}$, with each attention head j,

$$\begin{aligned} v_s^{K,j} = W^j_{\tau(v_s)} \cdot v_s^{(l-1)}, \quad v_t^{Q,j} = W^j_{\tau(v_t)} \cdot v_t^{(l-1)}, \\ v_s^{V,j} = W^j_{\tau(v_s)} \cdot v_s^{(l-1)}, \end{aligned} \tag{5}$$

where $v_*^{(l-1)}$ represents the input node feature for node $v \in \mathcal{V}$ from the $(l-1)$-th layer. The projection layers are capable of mapping node features from different node attributes to an embedding space that is invariant across node attributes. The features of edges from the $(l-1)$-th layer $e^{(l-1)}_{v_s\to v_t}$ are also projected by a linear projection layer $W_{\varphi(e)}$, serving as a **K**ey feature, $e^K_{v_s\to v_t} = W_{\varphi(e)} \cdot e^{(l-1)}_{v_s\to v_t}$. Once node embeddings are projected, we calculate the dot-product between the query and key vectors. Besides, we multiply the linearly transformed edge embedding with the similarity score to integrate the edge features into graph $\mathcal{G}$,

$$\texttt{SHA}(e,j) = \left(v_s^{K,j} \cdot e^K_{v_s\to v_t} \cdot v_t^{Q,j} \right) / \sqrt{d}, \tag{6}$$

where d denotes the dimension of node embeddings, $\texttt{SHA}(e,j)$ represents the attention score of edge e by the Single-Head Attention at j-head. We concatenate the scores obtained from each head and apply a softmax to them,

$$\texttt{SRA}(e) = \underset{\forall v_s \in \mathcal{N}^{\Phi}_{v_t}}{\texttt{Softmax}} \left(\underset{j\in[1,h]}{||} \texttt{SHA}(e,j) \right), \tag{7}$$

where $\mathtt{SRA}(e)$ represents Single-Relation Attention, providing the final attention score of the edges aggregating all the heads $j \in [1, h]$. $\mathcal{N}_{v_t}^{\Phi}$ is the set of the neighbours to the target node v_t. Then, we perform target-specific aggregation to update the feature of each target node by averaging its neighboring node features. For each target node v_t, we conduct a softmax operation on all the attention vectors from its neighboring nodes and then aggregate the information of all neighboring source nodes of v_t together. The updated node features $v_t^{(l)}$ for $\mathcal{G}^{(l)}$ can be represented as,

$$v_t^{(l)} = \bigoplus_{\forall v_s \in \mathcal{N}_{v_t}^{\Phi}} \left(\mathop{||}_{j \in [1,h]} \left(v_s^{V,j} \cdot \mathtt{SRA}(e) \right) \right), \tag{8}$$

where $\oplus$ is an aggregation operator, e.g., mean aggregation. The updated graph $\mathcal{G}^{(l)}$ is returned as the output of the l-th layer. Such operation is scalable by using L layers of aggregation. We further introduce modality-specific pooling for all nodes within the modality to obtain the prototype features for all modalities. Then the graph-level feature can be determined by a mean readout layer [3,79].

3.5 Overall Optimization

The aggregated multimodal representation could be obtained by $\mathcal{M} \circ g_{\phi}(\mathcal{G})$ from the heterogeneous graph $\mathcal{G}(\cdot)$, modality-wise graph prompting $g_{\phi}(\cdot)$ and knowledge-guided hierarchical aggregation $\mathcal{M}(\cdot)$. While the diverse diagnostic tasks including the glioma grading (a classification task) and the survival prediction (a integration prediction task) that may differ in the formulation, it is shown that they could be transformed to a uniform supervised learning fashion after some manipulations of task head [5]. Formally, the task-specific task head $\mathcal{H}^{\mathcal{T}}$ for the task $\mathcal{T}$ is introduced, with the task label denoted by $y^{\mathcal{T}}$,

$$\min_{\mathcal{M},\mathcal{H},\phi} \mathbb{E}_{\mathcal{G}} \; \mathcal{L}\left(\mathcal{H}^{\mathcal{T}} \circ \mathcal{M} \circ \phi(\mathcal{G}), y^{\mathcal{T}}\right). \tag{9}$$

where $\mathcal{L}$ denotes the loss function, which could be implemented by a NLL (negative log-likelihood) loss.

4 Experiments

4.1 Datasets and Settings

Datasets. We evaluate our method using data from The Cancer Genome Atlas (TCGA) [1] , a public database that includes genomic and clinical data from thousands of cancer patients. We select the datasets of Glioblastoma & Lower Grade Glioma (GBMLGG) and Kidney Renal Clear Cell Carcinoma (KIRC). For TCGA-GBMLGG, following [5], we use ROIs from diagnostic slides and apply sparse stain normalization [5] to match all images to a standard H&E histology image, creating a total of 1505 images for 769 patients, with WHO grading labels from G2 to G4. We curate 80 CNA and 240 RNA-Seq genomic features for each patient. Note that there are 40% of the patients with inherently actual missing RNA-Seq data. For the KIRC dataset, we use manually extracted 512×512 ROIs from diagnostic whole slide images for 417 patients in CCRCC, yielding 1251 images total that are similarly normalized with stain normalization. We pair these images with 117 CNV and 240 RNA-Seq genomic features. There are grading labels by Fuhrman Grading from G1 to G4.

Evaluation. For each cancer dataset, we perform 5-fold cross-validation and report the average test performance. Different metrics are leveraged for specific evaluation tasks, pathological glioma grading with Area Under the Curve (AUC) and Accuracy (ACC), and survival outcome prediction with concordance index (C-Index). Due to inherent missing of partly genetic modality in GBMLGG, we deploy the framework directly without modifying the data.While for KIRC with the complete modality data, we simulate the same situations in GBMLGG, randomly dropping RNA-Seq data with 40% subjects over the whole dataset. To ensure the consistent missing cases at training and test, we constrain the proportion of incomplete subjects in the training and test splits to be equal when producing the five-fold validation. In order to explore the missing issues under more modalities and missing ratios, we also perform experiments under simulated missing settings in Sec. 4.3.

Implementation. The framework is optimized by the Adam optimizer, with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-5} for the graph aggregation $\mathcal{M}$ and task head $\mathcal{H}$, over 150 epochs with early stopping. We adopt a smaller learning rate of 2×10^{-4} specially for optimizing prompt nodes $\mathcal{V}_P$ and the prompt bank components $\mathcal{V}^{P_B}$ in graph prompt function $g_\phi(\cdot)$. All the multimodal instances of a patient subject are jointly fed to the networks to get the final Data augmentations are performed on the training graphs, which involve randomly dropping edges and nodes, and adding Gaussian noise to the node and edge features [29,70]. The dropout ratio of each dropout layer is selected as 0.2. Regarding the hyper-parameters, we have the number of prompt nodes and prompt bank components in Eq. 2 with $N_P = 5$ and $N_B = 5$, and the dimension of input graph representation as $d = 512$.

4.2 Comparison with State-of-the-arts

We compare the proposed method against several SOTA methods in Tab. 1. For fair comparison, we apply identical settings for all experiments, and use the official code of compared works to deploy on our tasks when necessary.

Unimodal Models. Existing methods to analyze genomic data and pathological images are introduced. For genomic data, we employ SNN [24] for survival outcome prediction in the TCGA [5,6], and SNNTrans [24,63] that incorporates SNN as the feature extractor and TransMIL [63] for a global aggregation. For pathological images, we report the results of the SOTA MIL methods including the transformer-based models: AttnMIL [20], TransMIL [63], and the graph-based models PatchGCN [4], GTNMIL [85], HEAT [3]. It appears that using multimodal data consistently improves the performance under various metrics.

Table 1: Performance Comparison. We report results on four TCGA benchmarks, using various modality combinations of Genomics, Images, Cell graphs and Texts.

Methods	Modality				Glioma Grading (AUC/ACC)				Survival Pred. (C-Idx)	
	G	I	C	T	GBMLGG		KIRC		GBMLGG	KIRC
SNN [24]	✓	✗	✗	✗	0.8527	0.6583	0.8100	0.7790	0.7974	0.6639
SNNTrans [63]	✓	✗	✗	✗	0.8678	0.6725	0.8084	0.7755	0.7970	0.6671
AttMIL [20]	✗	✓	✗	✗	0.9063	0.7533	0.8252	0.7803	0.7908	0.6850
TransMIL [63]	✗	✓	✗	✗	0.9149	0.7683	0.8295	0.7899	0.8017	0.6876
PatchGCN [4]	✗	✓	✗	✗	0.8802	0.7429	0.8288	0.7896	0.7806	0.6795
GTNMIL [85]	✗	✓	✗	✗	0.9225	0.7966	0.8323	0.7980	0.8162	0.6953
HEAT [3]	✗	✓	✗	✗	0.9289	0.8057	0.8300	0.7961	0.8223	0.7059
Pathomic [5]	✓	✓	✗	✗	0.9172	0.7618	0.8295	0.7899	0.8101	0.7152
Porpoise [8]	✓	✓	✗	✗	0.9199	0.7789	0.8278	0.7800	0.8179	0.7179
MCAT [6]	✓	✓	✗	✗	0.9288	0.7929	0.8352	0.7957	0.8274	0.7235
TransFusion [86]	✓	✓	✗	✗	0.9209	0.7815	0.8299	0.7910	0.8251	0.7230
GTP-4o (Ours)	✓	✓	✗	✗	0.9256	0.8036	0.8349	0.7985	0.8296	0.7273
Pathomic [5]	✓	✓	✓	✗	0.9195	0.7674	0.8280	0.7889	0.8199	0.7211
TransFusion [86]	✓	✓	✓	✗	0.9225	0.7952	0.8318	0.7973	0.8283	0.7260
GTP-4o (Ours)	✓	✓	✓	✗	0.9336	0.8068	0.8331	0.8021	0.8329	0.7315
TransFusion [86]	✓	✓	✓	✓	0.9245	0.7986	0.8325	0.7990	0.8296	0.7289
GTP-4o (Ours)	✓	✓	✓	✓	0.9389	0.8126	0.8416	0.8068	0.8351	0.7336

Multimodal Models. We compare the SOTA multimodal methods including Pathomic [5], Porpoise [8] and MCAT [6], which only focus on extracting complementary multimodal information from the genomics and pathological images. It appears that there is a gain in using multimodal complementary information for various diagnostic tasks. Furthermore, as extending to more modalities is still unexplored, we compare our GTP-4o with other baselines by extend existing work Pathomic [5] with the cell graph modality, and also compare with a simple baseline TransFusion [86] which concentrates the intra-modal representations learned by uni-modal models [63]. From the table, the proposed GTP-4o exhibits the obvious improvement under most of different biomedical modalities.

4.3 Further Results

Ablation Studies. The effect of removing each component of GTP-4o is presented in Tab. 2. *No Heterogeneous Embedding* removes all the heterogeneous properties in the embedding such that it degrades to a simple homogeneous graph. *No Heterogeneous Relation* removes the heterogeneous properties of edges while maintaining the diverse attributes among the node features. *No Completion (Zero-init Missing)* handles the missing modality without using the proposed graph prompt completion while sets the features of the missing modality to zero values. *No Completion (Drop Missing)* directly drops the modality data in all patients if it occurs missing for some patients. *No Aggregation (Plain Mean)*

Table 2: Ablation of GTP-4o Variants. Results on TCGA-GBMLGG benchmarks over tasks of Glioma Grading (GG.) and Survival Prediction (SP.) are reported.

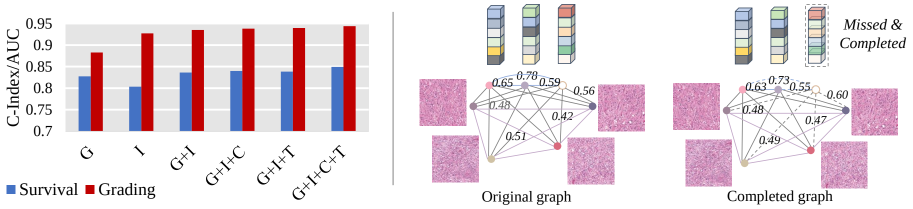
Components	Variants	GG. (AUC)	GG. (AUC)	SP. (C-Idx)
Graph Representation	No *Heterogeneous Embedding*	0.9232	0.8030	0.8168
	No *Heterogeneous Relation*	0.9259	0.8048	0.8201
Modality Completion	No *Completion (Zero-init Missing)*	0.9087	0.7875	0.7946
	No *Completion (Drop Missing)*	0.9288	0.8061	0.8233
	No *Prompt Bank*	0.9275	0.8081	0.8280
Hierarchical Aggregation	No *Aggregation (Plain Mean)*	0.9329	0.8067	0.8311
	No *Knowledge Guidance*	0.9350	0.8071	0.8342
Full Model	The Proposed GTP-4o	0.9389	0.8126	0.8416

removes the knowledge-guided aggregation while performs the plain mean aggregation among the k-NN heterogeneous neighbours ($k = 15$). *No Knowledge Guidance* removes the knowledge guidance for aggregation while uses the random meta-paths. Our ablation study results confirm the pivotal roles of our designs in the overall performance and effectiveness of the model.

Impact of Modality Usage. Fig. 3(**a**) shows the impact of using various combinations of modalities by GTP-4o. For the case of single modality, it is observed that each modality has its advantage for a specific task, as the relative performance of using only genes and only images is opposite for the tasks of glioma grading and survival analysis. We can also see that when more modalities like cell graphs and text descriptions are introduced, the performance of the model is improved on both two tasks. This suggests that GTP-4o is not only capable of generalizing to the various combinations of medical modality usage, but also able to deliver superior performance in terms of AUC (for glioma grading) and C-Index (for survival prediction).

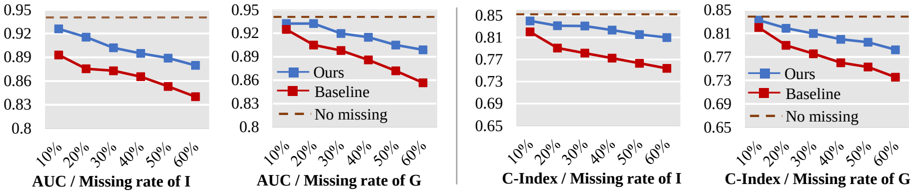
Impact of Modality Missing and Completion. To validate graph prompting's effectiveness, we compared graphs built from original full instances and completed graphs with arbitrary missing instances for a non-missing case (TCGA-02-0006) in Fig. 3(b). The completed graph shows similar relation patterns to the real one, suggesting biological validity of our proposed completion method. We further explored various missing settings by simulating missing data in image and non-RNA genomics modalities on TCGA-GBMLGG benchmarks. Fig. 4 illustrates GTP-4o's performance compared to the baseline without graph-prompted completion under different missing ratios. Results confirm the proposed completion method's effectiveness across various modality missing scenarios.

Limitation and Future Works. The current deployment is limited by the fact that no real-world clinical text reports are available for the datasets, thus we have to generate synthetic text descriptions by LLMs, probably bringing some data noise. Another limitation is that some additional modalities such as tabular data, are not considered in this paper, which could serve as future works.



(a) Ablation on modality usage. **(b)** Visualization on modality-prompted completion.

Fig. 3: (a) Analysis of Modality Usage. We provide the results of GTP-4o by using either Genes, Images, Cell graphs, Texts, or their combinations, on benchmarks of survival prediction (C-Index) and glioma grading (AUC). **(b) Analysis of Modality-prompted Completion**. We compare the relation pattern (similarity) in the original graph and the graph that is first removed specific instances then completed.



(a) Glioma grading w/ missing Image and Gene. **(b)** Survival pred. w/ missing Images and Gene.

Fig. 4: Analysis of Modality Missing. We study the results of **(a)** glioma grading and **(b)** survival prediction with the various missing ratios of Images and Genes. We compare the full framework of Ours and the version without our completion (baseline).

5 Conclusion

Increasing biomedical multimodal data provides not only opportunities for accurate and comprehensive diagnosis but also challenges for learning against the modality heterogeneity as well as the missingness issues. This study presents GTP-4o, which signifies a pioneering exploration into learning unified representations from various clinical modalities via the graph theory, exhibiting the robustness to heterogeneous modalities. Unlike prior methods, GTP-4o explores capturing explicit relations via a heterogeneous graph embedding. A novel graph prompting is proposed to complete deficient graph representations of missing modalities, and a hierarchical multimodal aggregation employs a global meta-path prior to guide the local aggregation across various heterogeneous relations. Extensive experiments demonstrate the efficacy of GTP-4o on disease diagnosis.

Acknowledgments. This work was supported by Research Grants Council (RGC) General Research Fund 14204321.

References

1. https://gdc.cancer.gov
2. Ali, S., Li, J., Pei, Y., Khurram, R., Rehman, K.U., Mahmood, T.: A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal mr image. Archives of computational methods in engineering **29**(7), 4871–4896 (2022)
3. Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: Histopathology whole slide image analysis with heterogeneous graph representation learning. In: CVPR. pp. 15661–15670 (2023)
4. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 339–349. Springer (2021)
5. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F.K., Rodig, S.J., Lindeman, N.I., Mahmood, F.: Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging **41**(4), 757–770 (2022). https://doi.org/10.1109/TMI.2020.3021387
6. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4025 (2021)
7. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4025 (2021)
8. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878 (2022)
9. Chen, Y., Liu, C., Huang, W., Cheng, S., Arcucci, R., Xiong, Z.: Generative text-guided 3d vision-language pretraining for unified medical image segmentation. arXiv preprint arXiv:2306.04811 (2023)
10. Chen, Y., Liu, C., Liu, X., Arcucci, R., Xiong, Z.: Bimcv-r: A landmark dataset for 3d ct text-image retrieval. arXiv preprint arXiv:2403.15992 (2024)
11. Chen, Z., Li, W., Xing, X., Yuan, Y.: Medical federated learning with joint graph purification for noisy label learning. MIA (2023)
12. Codling, E.A., Plank, M.J., Benhamou, S.: Random walk models in biology. Journal of the Royal society interface **5**(25), 813–834 (2008)
13. Ding, Z., Dong, Q., Xu, H., Li, C., Ding, X., Huang, Y.: Unsupervised anomaly segmentation for brain lesions using dual semantic-manifold reconstruction. In: International Conference on Neural Information Processing. pp. 133–144. Springer (2022)
14. Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., Tomaszeweski, J.: Automated grading of prostate cancer using architectural and textural image features. In: 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 1284–1287. IEEE (2007)

15. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C., Wang, X., Li, H.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: CVPR. pp. 6639–6648 (2019)
16. He, Z., Li, W., Zhang, T., Yuan, Y.: H 2 gm: A hierarchical hypergraph matching framework for brain landmark alignment. In: MICCAI. pp. 548–558 (2023)
17. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: Hˆ 2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 933–941 (2022)
18. Hu, Z., Dong, Y., Wang, K., Sun, Y.: Heterogeneous graph transformer. In: Proceedings of The Web Conference 2020. pp. 2704–2710 (2020)
19. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine **3**(1), 136 (2020)
20. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
21. Joo, S., Ko, E.S., Kwon, S., Jeon, E., Jung, H., Kim, J.Y., Chung, M.J., Im, Y.H.: Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. Scientific reports **11**(1), 18800 (2021)
22. Kim, S., Lee, N., Lee, J., Hyun, D., Park, C.: Heterogeneous graph learning for multi-modal medical data analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 5141–5150 (2023)
23. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
24. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
25. Kumar, A., Fulham, M., Feng, D., Kim, J.: Co-learning feature fusion maps from pet-ct images of lung cancer. IEEE Transactions on Medical Imaging **39**(1), 204–217 (2019)
26. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: CVPR. pp. 14943–14952 (2023)
27. Li, C., Feng, B.Y., Fan, Z., Pan, P., Wang, Z.: Steganerf: Embedding invisible information within neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 441–453 (2023)
28. Li, C., Feng, B.Y., Liu, Y., Liu, H., Wang, C., Yu, W., Yuan, Y.: Endosparse: Real-time sparse view synthesis of endoscopic scenes using gaussian splatting. arXiv preprint arXiv:2407.01029 (2024)
29. Li, C., Lin, M., Ding, Z., Lin, N., Zhuang, Y., Huang, Y., Ding, X., Cao, L.: Knowledge condensation distillation. In: European Conference on Computer Vision. pp. 19–35. Springer Nature Switzerland Cham (2022)
30. Li, C., Lin, X., Mao, Y., Lin, W., Qi, Q., Ding, X., Huang, Y., Liang, D., Yu, Y.: Domain generalization on medical imaging classification using episodic training with task augmentation. Computers in biology and medicine **141**, 105144 (2022)
31. Li, C., Liu, H., Fan, Z., Li, W., Liu, Y., Pan, P., Yuan, Y.: Gaussianstego: A generalizable stenography pipeline for generative 3d gaussians splatting. arXiv preprint arXiv:2407.01301 (2024)
32. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. arXiv preprint arXiv:2403.11050 (2024)

33. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. arXiv preprint arXiv:2406.02918 (2024)
34. Li, C., Ma, W., Sun, L., Ding, X., Huang, Y., Wang, G., Yu, Y.: Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. Neural Computing and Applications pp. 1–14 (2022)
35. Li, C., Zhang, Y., Li, J., Huang, Y., Ding, X.: Unsupervised anomaly segmentation using image-semantic cycle translation. arXiv preprint arXiv:2103.09094 (2021)
36. Li, C., Zhang, Y., Liang, Z., Ma, W., Huang, Y., Ding, X.: Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 61–65. IEEE (2021)
37. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
38. Li, W., Chen, Z., Li, B., Zhang, D., Yuan, Y.: Htd: Heterogeneous task decoupling for two-stage object detection. TIP (2021)
39. Li, W., Guo, X., Yuan, Y.: Novel scenes & classes: Towards adaptive open-set object detection. In: ICCV. pp. 15780–15790 (2023)
40. Li, W., Liu, J., Han, B., Yuan, Y.: Adjustment and alignment for unbiased open set domain adaptation. In: CVPR. pp. 24110–24119 (2023)
41. Li, W., Liu, X., Yao, X., Yuan, Y.: Scan: Cross domain object detection with semantic conditioned adaptation. In: AAAI. pp. 1421–1428 (2022)
42. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: CVPR (2022)
43. Li, W., Liu, X., Yuan, Y.: Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. TPAMI (2023)
44. Li, X., Jia, M., Islam, M.T., Yu, L., Xing, L.: Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. IEEE Transactions on Medical Imaging **39**(12), 4023–4033 (2020)
45. Liang, Z., Rong, Y., Li, C., Zhang, Y., Huang, Y., Xu, T., Ding, X., Huang, J.: Unsupervised large-scale social network alignment via cross network embedding. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 1008–1017 (2021)
46. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P.: The molecular signatures database hallmark gene set collection. Cell systems **1**(6), 417–425 (2015)
47. Lipkova, J., Chen, R.J., Chen, B., Lu, M.Y., Barbieri, M., Shao, D., Vaidya, A.J., Chen, C., Zhuang, L., Williamson, D.F., et al.: Artificial intelligence for multimodal data integration in oncology. Cancer Cell **40**(10), 1095–1110 (2022)
48. Liu, D., Yang, X., Wu, X.: Tumor immune microenvironment characterization identifies prognosis and immunotherapy-related gene signatures in melanoma. Frontiers in immunology **12**, 663495 (2021)
49. Liu, H., Liu, Y., Li, C., Li, W., Yuan, Y.: Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction. arXiv preprint arXiv:2406.16073 (2024)
50. Liu, X., Li, W., Yamaguchi, T., Geng, Z., Tanaka, T., Tsai, D.P., Chen, M.K.: Stereo vision meta-lens-assisted driving vision. ACS Photonics (2024)
51. Liu, X., Li, W., Yang, Q., Li, B., Yuan, Y.: Towards robust adaptive object detection under noisy annotations. In: CVPR. pp. 14207–14216 (2022)
52. Liu, X., Li, W., Yuan, Y.: Intervention & interaction federated abnormality detection with noisy clients. In: MICCAI. pp. 309–319 (2022)

53. Liu, X., Li, W., Yuan, Y.: Decoupled unbiased teacher for source-free domain adaptive medical object detection. IEEE Trans. Neural Netw. Learn. Syst. (2023)
54. Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., Yuan, Y.: Efficientvit: Memory efficient vision transformer with cascaded group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14430 (2023)
55. Liu, X., Yuan, Y.: A source-free domain adaptive polyp detection framework with style diversification flow. IEEE Transactions on Medical Imaging **41**(7), 1897–1908 (2022)
56. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. arXiv preprint arXiv:2401.12561 (2024)
57. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
58. Marusyk, A., Almendro, V., Polyak, K.: Intra-tumour heterogeneity: a looking glass for cancer? Nature reviews cancer **12**(5), 323–334 (2012)
59. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: CVPR. pp. 8238–8247 (2022)
60. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
61. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine **34**(6), 96–108 (2017)
62. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 86 (2021)
63. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
64. Sun, L., Li, C., Ding, X., Huang, Y., Chen, Z., Wang, G., Yu, Y., Paisley, J.: Few-shot medical image segmentation using a global correlation network with discriminative embedding. Computers in biology and medicine **140**, 105067 (2022)
65. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. VLDB **4**(11), 992–1003 (2011)
66. Wang, Z., Li, R., Wang, M., Li, A.: Gpdbn: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. Bioinformatics **37**(18), 2963–2970 (2021)
67. Wang, Z., Wen, R., Chen, X., Cao, S., Huang, S.L., Qian, B., Zheng, Y.: Online disease diagnosis with inductive heterogeneous graph convolutional networks. In: Proceedings of the Web Conference 2021. pp. 3349–3358 (2021)
68. Wuyang, L., Chen, Y., Jie, L., Xinyu, L., Xiaoqing, G., Yixuan, Y.: Joint polyp detection and segmentation with heterogeneous endoscopic data. In: 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021): co-located with the 17th IEEE International Symposium on Biomedical Imaging (ISBI 2021). pp. 69–79. CEUR-WS Team (2021)
69. Xu, H., Li, C., Zhang, L., Ding, Z., Lu, T., Hu, H.: Immunotherapy efficacy prediction through a feature re-calibrated 2.5 d neural network. Computer Methods and Programs in Biomedicine **249**, 108135 (2024)

70. Xu, H., Zhang, Y., Sun, L., Li, C., Huang, Y., Ding, X.: Afsc: Adaptive fourier space compression for anomaly detection. arXiv preprint arXiv:2204.07963 (2022)
71. Xu, R., Li, Y., Wang, C., Xu, S., Meng, W., Zhang, X.: Instance segmentation of biological images using graph convolutional network. Engineering Applications of Artificial Intelligence **110**, 104739 (2022)
72. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21241–21251 (October 2023)
73. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: CVPR. pp. 1179–1189 (2023)
74. Xue, Z., Marculescu, R.: Dynamic multimodal fusion. In: CVPR. pp. 2574–2583 (2023)
75. Yang, Q., Guo, X., Chen, Z., Woo, P.Y., Yuan, Y.: D2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. IEEE Transactions on Medical Imaging **41**(10), 2953–2964 (2022)
76. Yang, Q., Li, W., Li, B., Yuan, Y.: Mrm: Masked relation modeling for medical image pre-training with genetics. In: ICCV. pp. 21452–21462 (2023)
77. Yang, Q., Yuan, Y.: Learning dynamic convolutions for multi-modal 3d mri brain tumor segmentation. In: MICCAI Workshop. pp. 441–451. Springer (2021)
78. Zeng, Y., Zeng, Y., Yin, H., Chen, F., Wang, Q., Yu, X., Zhou, Y.: Exploration of the immune cell infiltration-related gene signature in the prognosis of melanoma. Aging (albany NY) **13**(3), 3459 (2021)
79. Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N.V.: Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 793–803 (2019)
80. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)
81. Zhang, Y., Yang, J., Tian, J., Shi, Z., Zhong, C., Zhang, Y., He, Z.: Modality-aware mutual learning for multi-modal medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 589–599. Springer (2021)
82. Zhang, Y., Fang, Q., Qian, S., Xu, C.: Multi-modal multi-relational feature aggregation network for medical knowledge representation learning. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3956–3965 (2020)
83. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
84. Zhang, Y., Li, C., Lin, X., Sun, L., Zhuang, Y., Huang, Y., Ding, X., Liu, X., Yu, Y.: Generator versus segmentor: Pseudo-healthy synthesis. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 150–160. Springer International Publishing (2021)
85. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: A graph-transformer for whole slide image classification. IEEE transactions on medical imaging **41**(11), 3003–3015 (2022)

86. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21485–21494 (2023)
87. Zhou, T., Ruan, S., Canu, S.: A review: Deep learning for medical image segmentation using multi-modality fusion. Array **3**, 100004 (2019)
88. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
89. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: Making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7234–7242 (2017)