# From Representation Space to Prognostic Insights: Whole Slide Image Generation with Hierarchical Diffusion Model for Survival Prediction

**Zhihao Tang**[1], **Xi Zhang**[1*], **Chaozhuo Li**[2]

[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE), Beijing University of Posts and Telecommunications, Beijing 100876, China

[2]Beijing University of Aeronautics and Astronautics, Beijing 100191, China

innerone@bupt.edu.cn, zhangx@bupt.edu.cn, lichaozhuo1991@gmail.com

## Abstract

Deep learning has significantly enhanced survival prediction using whole slide images (WSIs) by adopting a two-stage learning paradigm: WSI preparation and patient-level prediction. While existing research generally concentrates on developing advanced patient-level prediction modules, the critical importance of WSI preparation has been largely overlooked. In practice, WSI preparation is influenced by numerous factors, including tissue heterogeneity, sampling strategies, and technical considerations. These uncontrollable external factors incur variability in the number of WSIs among patients, introducing significant bias and resulting in inferior performance for patients with few WSIs. To address this challenge, we propose a novel approach named WSI-Diffusion. Unlike existing WSI generation models that produce augmented versions of input WSIs, our method generates entirely new WSIs in representation space to serve as complementary data. WSI-Diffusion employs a two-stage hierarchical diffusion process. Two novel modules, WSI-level and patch-level Diffusers are designed to capture complex correlations between WSIs and patches. The generated WSIs are integrated as supplementary data, and a light patient-level prediction module is then trained for survival prediction. Experimental results across five datasets demonstrate the superiority of our proposal.

## 1 Introduction

Survival prediction has been a long-standing clinical task, focusing on estimating the probability of mortality within a specified timeframe (Bakas et al. 2018). The cornerstone of survival prediction involves the microscopic examination of tissues and cells (Azadi et al. 2023), which can be digitized using high-throughput slide scanners, known as Whole Slide Images (WSIs). With the integration of WSIs into routine clinical workflows and the substantial advancement of deep learning, WSI-based deep survival models have become increasingly prevalent (Tang et al. 2019; Zhu et al. 2017).

Existing WSI-based survival prediction models (Yao et al. 2020a; Shen et al. 2022; Fan et al. 2022) generally follow a two-stage paradigm. (1) **WSI Preparation** involves collecting and transforming WSIs into a format suitable for deep learning. Pathology laboratories initially obtain and preserve each WSI from tissue blocks resected during cancer diagnosis and treatment (Aberle et al. 2011). Due to the computational infeasibility of using an entire WSI with gigapixel spatial resolution as input, WSIs are divided into smaller patches (Zhu et al. 2017), which are then fed into a pre-trained encoder (e.g., ResNet) to extract patch-level features. (2) **Patient-level Prediction** involves training a regression model to predict survival outcomes. Based on the patch-level features, existing models employ clustering (Yao et al. 2020a), graph-building (Liu et al. 2023b), or sequential methods (Huang et al. 2021; Shao et al. 2023a) to model survival-related patch-level interactions. These correlations are captured by various encoders (e.g., Transformer (Vaswani et al. 2017)) to learn patient-level representations, which are subsequently mapped to survival outcomes.

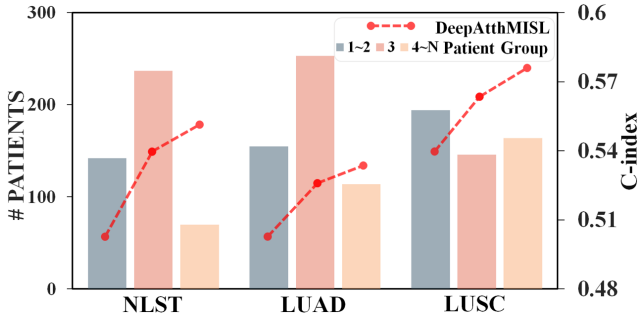


Figure 1: The illustration of the bias in the number of WSIs across patients and its impact on prediction performance.

Significant efforts have been dedicated to improving patient-level prediction through complex patch correlations and advanced encoders (Fan et al. 2022; Jaume et al. 2021; Yao et al. 2020b). However, solely advancing the patient-level prediction might not be the panacea. The WSI preparation stage plays a critical role in establishing a solid data foundation for survival analysis, yet it is often overlooked. Unlike tumor detection, which targets localized regions within WSIs (Liu et al. 2023a), survival prediction requires a comprehensive analysis of histological features across tissues (Zhu et al. 2017). Diverse WSIs offer insights into the patient's health status from multiple vantage points, culminating in a holistic blueprint when these WSIs are collectively analyzed within a unified framework (Fan

*Corresponding author: Xi Zhang.

et al. 2022). In practice, the preparation of WSIs is influenced by numerous factors, including tissue heterogeneity and sampling strategies (Albertina et al. 2016). These uncontrollable external factors result in significant challenges regarding the completeness of WSIs. Consequently, patients may have varying numbers of WSIs, leading to substantial bias and reduced predictive performance.

To investigate the effect of varying WSI numbers across patients, we conduct a preliminary study using a popular deep survival model, DeepAttnMISL (Yao et al. 2020b), on three datasets: NLST (Aberle et al. 2011), TCGA-LUAD, and TCGA-LUSC (Kirk et al. 2016). As shown in Fig. 1, patients are categorized into three groups based on the number of associated WSIs, indicated by different colors. The bars represent the number of patients in each group, while the line chart illustrates the model's performance. DeepAttnMISL is trained and tested on each group. Nearly a third of patients have fewer than three WSIs. However, DeepAttnMISL exhibits degraded performance for patients with fewer WSIs, restricting its practical applicability. The results show that model performance on testing groups with fewer WSIs is consistently inferior to those with more WSIs.

In light of the bias in WSI distribution among patients, our motivation is to forecast "missing" WSIs as supplementary resources. This approach aligns with data augmentation (Shao et al. 2023b; Zaffar et al. 2022), where techniques such as rotation, elastic deformation, and jitter are applied to manipulate patches, as illustrated in Fig. 2(a). However, current augmentation models can only modify existing patches, resulting in augmented WSIs that reflect similar tissue and tumor microenvironments. Therefore, these methods merely expand existing knowledge without predicting new insights as "missing" WSIs, leaving the challenge of incomplete data unresolved. Consequently, developing a WSI generator model capable of generating new WSIs, rather than just manipulating pre-existing ones, becomes essential.

Directly generating WSIs of gigabyte sizes is practically infeasible due to resource limitations and the scarcity of training data. Considering that WSIs are typically encoded into low-dimensional embeddings in deep models, we opt to generate possible representations of "missing" WSIs. A straightforward strategy involves masking one WSI from a patient and attempting to regenerate its representation by generating a series of patches based on the remaining WSIs. However, this naive patch-to-patch generation overlooks crucial WSI-level constraints. (1) **Consistency among Patches**. Patches within a WSI often exhibit strong correlations as they represent different areas of the same tissue block (Fan et al. 2022). This consistency encompasses various attributes, including staining and nuclear morphology. Therefore, when generating a set of patches within a WSI, it is crucial to preserve consistency among them. (2) **Tissue Type Distribution**. Patches within a WSI should adhere to a distribution corresponding to different tissue types, such as neoplastic, necrotic, and inflammatory (Gamper et al. 2019). Variations in tissue type distribution between similar WSIs can convey distinct prognostic information. For instance, a higher proportion of neoplastic patches may indicate a more aggressive tumor and suggest a poorer prognosis (Sanegre et al. 2021). Thus, it is imperative to incorporate this clinically significant constraint during the generation process.

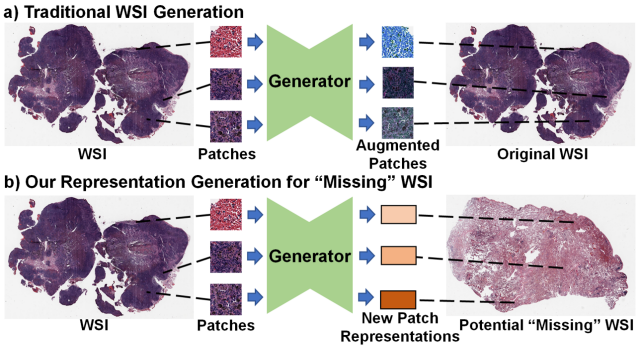


Figure 2: We propose to generate new patch representations to construct a "missing" WSI, as opposed to creating various data augmentations of patches from the original WSI.

Given the challenges mentioned above, we propose a novel model **WSI-Diffusion** to generate comprehensive WSI representations for survival prediction using a hierarchical diffusion approach. The adoption of diffusion model (Ho, Jain, and Abbeel 2020) is justified by its proven effectiveness in scenarios with limited data samples (Giannone, Nielsen, and Winther 2022; Wu et al. 2023). In the WSI preparation stage, WSI-Diffusion employs a two-step hierarchical diffusion process. Initially, it generates a WSI-level feature as a patch consistency constraint, which serves as a condition for generating patch-level representations. This hierarchical structure divides the generation process into two correlated phases, linking each patch with its corresponding WSI to ensure consistency among patches. To maintain tissue type distribution, WSI-Diffusion employs a distribution-aware sampling strategy. During patch consistency constraint generation, it creates a reference tissue distribution by classifying and counting patches by tissue type. Then WSI-Diffusion uses class embeddings sampled from this reference distribution to condition each patch. This sampling strategy ensures diverse generated patches while preserving the distribution of tissue types. Extensive experimental results on five datasets demonstrate its superiority.

The contributions of this work are as follows. First, as far as we know, we are the first to investigate the novel problem of WSI distribution bias across patients and reveal its effect on the survival prediction task. Second, we propose a novel WSI representation generation framework based on a hierarchical diffusion model, which fills the gap between WSI-level and patch-level generation. Third, we demonstrate the effectiveness of our approach by showing its superior performance compared to SOTA baselines.

## 2 Problem Definition

### 2.1 Definition of Survival Data

Given a group of patients $\mathcal{P} = \{p_1, \ldots, p_{|\mathcal{P}|}\}$, each patient $p$ has a set of WSIs, represented as $\mathcal{W}_p = \left\{w_1^{(p)}, \ldots, w_{n_p}^{(p)}\right\}$, along with a follow-up label $(\delta_p, t_p)$. $n_p$ denotes the num-

ber of WSIs associated with patient $p$. The binary indicator $\delta_p$ represents censoring, indicating whether the survival outcome for patient $p$ has been observed before the end of the study. Censoring occurs when patients succumb to other causes or are lost to follow-up. A value of 1 indicates censoring, while a value of 0 indicates that the event of interest has occurred. The variable $t_p$ represents either the patient's survival time/period or a censored time/period.

### 2.2 Definition of Survival Prediction

The WSI-based survival prediction is defined as follows: *Given a set of a patient's WSIs $\mathcal{W}_p$, it aims to predict the most likely survival time $\hat{t_p}$.* This problem is formulated as:

$$\arg\max_{t_i \in \mathcal{R}} Pr\left(\hat{t_p} = t_i \mid \mathcal{W}_p\right). \tag{1}$$

Our motivation, which lies in generating "missing" WSIs, can be presented as: *Given a set of a patient's WSIs $\mathcal{W}_p$, generate the "missing" WSI $\overline{w_p}$ using the generator $f_g()$*:

$$\overline{w_p} = f_g\left(z, \mathcal{W}_p \mid \theta\right), \tag{2}$$

where $\theta$ is the parameter set of $f_g()$, $z$ denotes noise sampled from a specific distribution (e.g., normal distribution). After generating the "missing" WSI $\overline{w_p}$, the definition of survival prediction is revised by combining Eq. (2) into Eq. (1):

$$\arg\max_{t_i \in \mathcal{R}} Pr\left(\hat{t_p} = t_i \mid \mathcal{W}_p \cup f_g\left(z, \mathcal{W}_p \mid \theta\right)\right). \tag{3}$$

As discussed in the introduction, significant effort has been focused on modeling $P_r$. However, the contribution of our approach is centered on how we model the WSI $w$, design the generator $f_g$, and optimize its parameters $\theta$.

## 3 Methodology

The overall framework of WSI-Diffusion is depicted in Fig.3. It consists of two diffusers: one for WSI-level generation and another for patch-level generation. Conditioned on existing WSIs, the WSI-level diffuser generates two key constraints: patch consistency constraint and tissue type distribution, which are then used for generating patch representations in the patch-level diffuser.

### 3.1 Multi-scale WSI Modeling

Multi-scale WSI modeling prepares input data for WSI-Diffusion through two components: Feature Extraction and Patch Classification. Feature Extraction obtains representations at both the WSI and patch levels for hierarchical diffusion. Patch Classification assigns tissue types to each patch, ensuring the preservation of the tissue type distribution.

**Feature Extraction.** In feature extraction, our objective is to first segment WSIs into patches and subsequently map both WSIs and patches into a shared representation space. This approach is informed by previous research (Chen et al. 2022), which supports the modeling of WSIs and patches in a representation space for two key reasons: First, generating WSIs in image space is impractical due to their substantial gigabyte-sized dimensions. Second, generating patches in image space is less direct, as current survival models use patch representations, rather than the raw patches, as input.

For each WSI $w$, a pre-trained feature extractor $f_w()$ (e.g., HIPT (Chen et al. 2022)) is used to map the WSI into a representation vector $\mathbf{z}^w \in \mathbb{R}^{1\times C'}$, where $C'$ denotes the vector dimension. To explore the patch level, non-overlapping and non-background patches are cropped from each WSI $w$ using a sliding window strategy combined with the OTSU thresholding algorithm (Otsu 1979). These patches are then mapped into the representation space using a pre-trained feature encoder $f_e()$ like HIPT, yielding a set of patch representations $\mathcal{E}_w = \left\{e_1^{(w)}, \ldots, e_k^{(w)}, \ldots, e_{n_w}^{(w)}\right\}$, where $n_w$ denotes the number of patches in WSI $w$.

**Patch Classification.** To maintain the distribution of various tissue types within each WSI, the first step involves assigning a tissue type to each patch. This classification is performed using HoverNet (Graham et al. 2019), a model pre-trained on the PanNuke dataset (Gamper et al. 2019). HoverNet classifies nuclei within each patch into predefined types, including neoplastic, dead, inflammatory, non-neoplastic epithelial, connective, and no label (Gamper et al. 2019). Each patch is then categorized based on the most frequently predicted nucleus type through majority voting. Subsequently, each patch is assigned a one-hot tissue type label $y$ with dimensions $1 \times C$, where $C$ represents the number of classes.

### 3.2 Hierarchical WSI Diffusion Model

In this section, we outline the procedure for generating the representation of a potential "missing" WSI $\overline{w}$ using the WSI-Diffusion model, which includes the WSI-level diffuser and the patch-level diffuser.

**3.2.1 WSI-level Diffuser.** Given a WSI $w$, we first extract its feature as described in **Multi-scale WSI Modeling**, resulting in $\mathbf{z}^w \in \mathbb{R}^{1\times C'}$. To derive the corresponding tissue type distribution, we sum the set of patch labels $y^w$ for WSI $w$, where $y^w$ has a size of $n_w \times c$, along the $n_w$ dimension. Here, $n_w$ represents the number of patches in WSI $w$. The resulting sum is then normalized to produce the tissue type distribution $\mathbf{p}^w$ for WSI $w$, with each dimension representing the proportion of each tissue type in WSI $w$.

The WSI-level diffuser is trained to generate the WSI representation $\mathbf{z}^w$ of $\overline{w}$ as patch consistency constraint and its corresponding tissue type distribution $\mathbf{p}^w$, which are used as conditions for the patch-level diffuser. Specifically, our diffuser is based on a conditional denoising diffusion probabilistic model (Ho, Jain, and Abbeel 2020; Zhang et al. 2023, 2024), which involves two processes: forward and reverse.

**Forward process:** Although $\mathbf{z}$ and $\mathbf{p}$ follow different distributions, their independent diffusion processes can be described by a unified formula due to their similar structure:

$$q\left(\mathbf{x}_{1:T}^w \mid \mathbf{x}_0^w\right) = \prod_{t=1}^{T} q\left(\mathbf{x}_t^w \mid \mathbf{x}_{t-1}^w\right), \tag{4}$$

$$q\left(\mathbf{x}_t^w \mid \mathbf{x}_{t-1}^w\right) = \mathcal{N}\left(\mathbf{x}_t^w; \sqrt{1-\beta_t}\mathbf{x}_{t-1}^w, \beta_t\mathbf{I}\right). \tag{5}$$

Here, $\mathbf{z}$ and $\mathbf{p}$ are denoted by $\mathbf{x}$ for unification. $\mathbf{x}_t^w$ represents the noisy patch consistency constraint or tissue type distribution at the $t$-th step, with $\mathbf{x}_0^w = \mathbf{x}^w$ being $\mathbf{z}^w$ or $\mathbf{p}^w$ as their starting point. $\mathcal{N}$ denotes the Gaussian distribution,

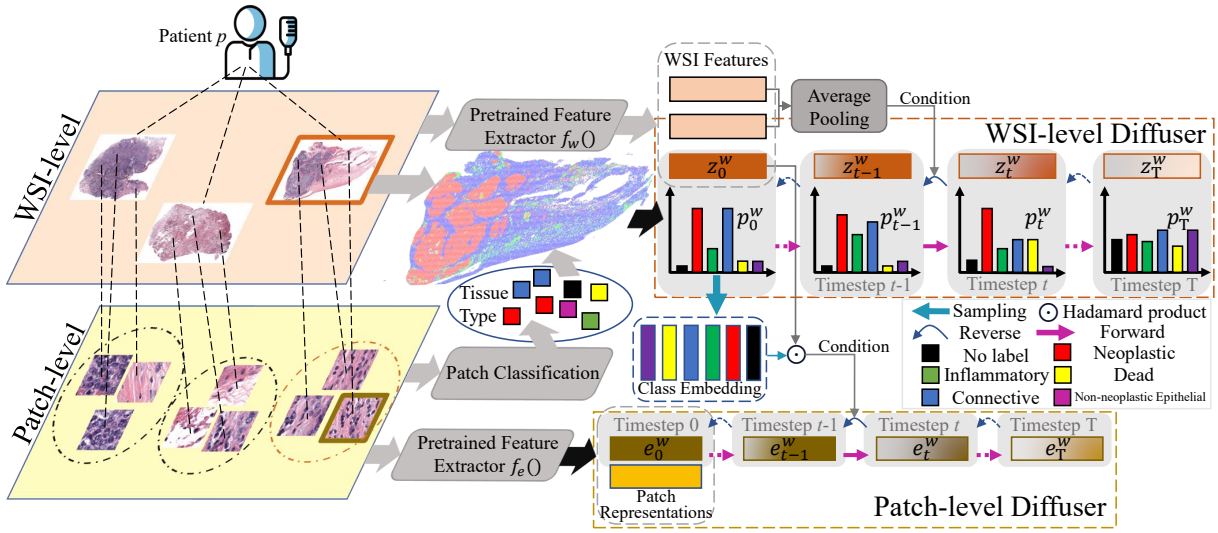


Figure 3: Overview of WSI-Diffusion, which operates at both the WSI and patch levels. The WSI-level diffuser is trained to generate the "missing" WSI feature $z_0^w$ as a patch consistency constraint and its corresponding tissue type distribution $p_0^W$, conditioned on existing WSI features. The patch-level diffuser then generates patch representations $e_0^W$ for different tissue types within the "missing" WSI, conditioned on $z_0^w$ and class embeddings sampled from $p_0^W$. During sampling, the patch consistency constraint and tissue type distribution are generated first, followed by class embeddings to generate the patch representations.

and $\beta_{1:T} \in (0,1)$ is the noise schedule that controls the variance of the noise, which can be set differently for $\mathbf{z}^w$ and $\mathbf{p}^w$. For efficiency, $\mathbf{x}_t^w$ can be derived from $\mathbf{x}_0^w$ using reparameterization techniques (Kingma and Welling 2013):

$$q\left(\mathbf{x}_t^w \mid \mathbf{x}_0^w\right) = \mathcal{N}\left(\mathbf{x}_t^w; \sqrt{\bar{\alpha}_t}\mathbf{x}_0^w, (1-\bar{\alpha}_t)\,\mathbf{I}\right), \tag{6}$$

$$\alpha = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i. \tag{7}$$

**Reverse process:** Given the generated pair of embeddings (patch consistency constraint $\mathbf{z}^w$, tissue type distribution $\mathbf{p}^w$) of $\overline{w}$, and considering their strong interrelation, our goal is to recover them within a single reverse process. We start by sampling noise from a Gaussian distribution with dimension $C + C'$, which is then split into $\mathbf{z}_T^w$ and $\mathbf{p}_T^w$. We use a denoiser $\epsilon_\theta$ to process both inputs simultaneously, predicting the respective noise and enhancing the generation quality of both the patch consistency constraint and the tissue type distribution. Specifically, for a given time step $t$, the reverse process of the WSI-level diffuser is formulated as follows:

$$\begin{aligned} Pr_\theta\left(\mathbf{z}_{t-1}^w \mid (\mathbf{z}_t^w, \mathbf{p}_t^w)\right) &= \mathcal{N}\left(\mathbf{z}_{t-1}^w; \mu_\theta^z\left(\mathbf{z}_t^w, \mathbf{p}_t^w, c^w, t\right)\right), \\ Pr_\theta\left(\mathbf{p}_{t-1}^w \mid (\mathbf{z}_t^w, \mathbf{p}_t^w)\right) &= \mathcal{N}\left(\mathbf{p}_{t-1}^w; \mu_\theta^p\left(\mathbf{z}_t^w, \mathbf{p}_t^w, c^w, t\right)\right), \end{aligned} \tag{8}$$

where $\mu_\theta^z$ and $\mu_\theta^p$ denote the Gaussian mean values predicted by $\theta$, and $c^w$ represents the condition for $\overline{w}$. For a patient $p$ with $n_p$ WSIs, we employ a leave-one-out strategy, selecting one WSI $w$ as $\overline{w}$ for generation, while using the remaining $n_p - 1$ WSIs as the condition. In line with previous works (Ulhaq, Akhtar, and Pogrebna 2022), we average the representations of $n_p - 1$ WSIs to obtain the condition $c \in \mathbb{R}^{1\times C'}$. Finally, the training objective of the WSI-level diffuser is:

$$\begin{aligned} \mathcal{L}_w = \mathbb{E}_{(\epsilon^z,\epsilon^p)\sim\mathcal{N}(0,1),t}[\underbrace{\|\epsilon^z - \epsilon_\theta^z\left(\mathbf{z}^w, \mathbf{p}^w, c^w, t\right)\|_2^2}_{\text{Patch Consistency Constraint}} \\ + \underbrace{\|\epsilon^p - \epsilon_\theta^p\left(\mathbf{z}_t^w, \mathbf{p}^w, c^w, t\right)\|_2^2}_{\text{Tissue Type Distribution}}], \end{aligned} \tag{9}$$

where $t \in [0, T]$; $\epsilon^z \in \mathbb{R}^{1\times C'}$ and $\epsilon^p \in \mathbb{R}^{1\times C}$ are jointly sampled from a Gaussian distribution of dimension $C + C'$.

**3.2.2 Patch-level Diffuser.** The patch-level diffuser generate patch representations that can be used to model the "missing" WSI $\overline{w}$. Given a set of patches cropped from WSI $w$, we extract their representations as described in **Multi-scale WSI Modeling**, resulting in $\mathcal{E}^w \in \mathbb{R}^{n_w\times C'}$. The patch-level diffuser is also based on conditional denoising diffusion probabilistic models (Ho, Jain, and Abbeel 2020), so its forward and reverse processes follow a similar formulation as the WSI-level diffuser. The key distinction lies in the denoiser $\epsilon_\theta$. Unlike the WSI-level denoiser, which is conditioned on the $n_p - 1$ WSIs of the patient, the patch-level diffuser generates a set of patches for the "missing" WSI $\overline{w}$.

To ensure consistency across patches, we impose a constraint on the patch consistency condition $\mathbf{z}^w$ of $\overline{w}$. Additionally, to preserve the tissue type distribution at the WSI level, we apply a condition on the class embedding $s$ for each patch. These constraints are combined into $c^p$ using the Hadamard product, as outlined in previous work (Chen et al. 2024; Zhang, Zhang, and Pan 2022; Zhang et al. 2025; Li et al. 2017; Zhao et al. 2021). To generate patches that conform to $\hat{\mathbf{p}}^w$, we denote the fraction of each tissue class $c$ as $\mathbf{r}_c$, with its corresponding class embedding represented as $s_c$. Given a batch of $N$ noises sampled from a Gaussian

distribution, we randomly select $N\mathbf{r}_c$ of them and use $s_c$ as a condition to generate $N\mathbf{r}_c$ patches. Consequently, we sample $N\mathbf{r}_c$ patches for each class $c$. This sampling strategy ensures the correct distribution of tissue types. Since patches have different class embeddings, we perform denoising on each individual patch $e^w$, which can be expressed as:

$$\mathbf{e}_{t-1}^w = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{e}_t^w - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\mathbf{e}_t^w, c^p, t\right)\right) + \sqrt{1-\alpha_t}\epsilon_t \tag{10}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$. By iteratively sampling $\mathbf{e}_t^w$ using Eq. (10) over $T$ steps, we generate $\hat{\mathbf{e}}^w$, which is then used to reconstruct the "missing" WSI along with other patches.

WSI-Diffusion employs a hierarchical generation process. During sampling, the WSI-level diffuser first generates the "missing" WSI feature $\hat{\mathbf{z}}^{\overline{w}}$ as patch consistency constraint and its corresponding tissue type distribution $\hat{\mathbf{p}}^{\overline{w}}$, conditioned on the $n_p - 1$ existing WSIs. These outputs are then used as conditions for the patch-level diffuser to generate a set of patches $\hat{\mathcal{E}}^{\overline{w}}$, modeling the "missing" WSI $\overline{w}$.

### 3.3 Patient-level Prediction

Utilizing WSI-Diffusion in Section 3.2, we generate a set of patches $\mathcal{E}_{\overline{w}} = \left\{e_1^{\overline{w}}, \ldots, e_k^{\overline{w}}, \ldots, e_{n_w}^{\overline{w}}\right\}$ for the "missing" WSI $\overline{w}$, conditioned on the $n_p - 1$ representations of existing WSIs. Existing models, such as AugDiff (Shao et al. 2023b), typically involve directly integrating $\mathcal{E}_{\overline{w}}$ with $\mathcal{E}_w$, and then aggregating them into a patient-level representation. However, in our approach, we first aggregate the patches into several WSI representations, which are subsequently aggregated into a patient-level representation. We adopt this strategy for two distinct purposes: (1) **Heterogeneity of WSIs**: Each WSI exhibits unique characteristics, making it crucial to process each WSI individually to capture its distinct features. (2) **Utilization of WSI-Level Information**: By introducing two constraints (patch consistency constraint and tissue type distribution), we ensure that the generated set of patches aligns with the entire WSI. Directly aggregating patch representations across WSIs would fail to effectively leverage this WSI-level information, potentially compromising the quality of the final patient-level representation.

Consider a patient denoted as $p$, who possesses $n_p$ WSIs. Each WSI $w$ is associated with a set of patch representations denoted as $\mathcal{E}_w$. These patch representations within $\mathcal{E}_w$ are stacked to form a feature map for $w$, represented as $\mathbf{E}_w \in \mathbb{R}^{n_w \times C}$, where $n_w$ denotes the number of patches within WSI $w$, and $C$ denotes the feature-length determined by a pre-trained feature extractor $f_e()$.

On the obtained feature map $\mathbf{E}_w$ of each WSI, we employ Convolutional Multiple Instance Learning (C-MIL) (Yang et al. 2017) to reduce the dimension, thereby generating WSI representation $E_w \in \mathbb{R}^{n_w \times L}$ separately, where $L = 64$ following the settings in the original paper. Finally, we apply Adaptive Average Pooling to aggregate multiple WSI-level representations $\mathbf{E}_1$,...,$\mathbf{E}_{n_p-1}$,$\mathbf{E}_{\overline{w}}$ into a patient-level representation $\mathbf{E}_p$. This can be formulated as:

$$\mathbf{E}_p = Agg\left(f_C\left(\mathbf{E}_1, ..., \mathbf{E}_k, ..., \mathbf{E}_{n_p-1}, \mathbf{E}_{\overline{w}}\right)\right) \tag{11}$$

| Dataset | # Patient | # WSI | | # WSI of a patient | | Cancer Type |
|---|---|---|---|---|---|---|
| | | Diagnostic | Tissue | Mode | Maximum | |
| NLST | 449 | 1,224 | | | 6 | ADC&SCC |
| LUSC | 504 | 1,100 | 512 | 3 | 13 | SCC |
| LUAD | 514 | 1,067 | 541 | | 14 | ADC |
| BRCA | 1,098 | 1,133 | 1,978 | | 9 | BIC |
| BLCA | 412 | 457 | 469 | 2 | 10 | BUC |

Table 1: The statistics of five datasets.

where $Agg$ denotes Adaptive Average Pooling, and $f_C$ denotes C-MIL. After obtaining the patient-level representation $\mathbf{E}_p$, we use it to generate hazard risks that measure the expected development of cancer. Given patient-level representation $\mathbf{E}_p$, the hazard risk $O_p$ is obtained by passing it through a multilayer perceptron. Following previous work (Zhu, Yao, and Huang 2016), we improve the accuracy of survival prediction by utilizing the Cox loss function:

$$L\left(O_i\right) = \sum_{i:R(t_i)=1}\left(-O_i + \log \sum_{j:t_j \geq t_i} \exp\left(O_j\right)\right) \tag{12}$$

where $O_i$ denotes the hazard risk of patient $p_i$. This Cox loss function optimizes hazard risk predictions by assigning higher risks to patients with shorter survival times and improving the overall effectiveness of survival analysis.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.** Our proposal is evaluated on five large cancer datasets. One is the National Lung Screening Trial (NLST) dataset (Team 2011), which includes adenocarcinoma (ADC) and squamous cell carcinoma (SCC) cases. The other four datasets are from The Cancer Genome Atlas (TCGA), covering lung cancer (LUSC and LUAD), breast cancer (BRCA), and bladder cancer (BLCA). Each dataset contains WSIs stained with hematoxylin and eosin (H&E) and includes clinical information, such as survival data.

**Baselines.** We compare our proposed method with several advanced baselines. WSISA (Zhu et al. 2017) and DeepAttnMisl (Yao et al. 2020a) utilize clustering for patch correlation and CNNs for prediction. HIPT (Chen et al. 2022), SeTransSurv (Huang et al. 2021), ESAT (Shen et al. 2022), LongViT (Wang et al. 2023), and Prov-GigaPath (Xu et al. 2024) employ Transformers to model long-range dependencies at the WSI level. AugDiff (Shao et al. 2023b) leverages a diffusion model to generate augmentations.

**Evaluation Metrics.** Model performance is assessed using the concordance index (C-index) and the STAGE-5 metric, which converts survival prediction into a classification task.

**Implementation Details.** In our experiments, we utilize 5-fold cross-validation and allocated 10% of the training data as validation to facilitate early stopping. We implement WSI-level and patch-level diffusers based on DDPM, with batch sizes of 64 and 128, respectively, and a total of 30 steps. The experiments are conducted using Pytorch 1.9.1 on four NVIDIA V100 GPUs.

| Architecture | Model | Dataset | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NLST | | LUSC | | LUAD | | BRCA | | BLCA | |
| | | C-index | STAGE-5 | C-index | STAGE-5 | C-index | STAGE-5 | C-index | STAGE-5 | C-index | STAGE-5 |
| CNN | WSISA | $0.662_{.033}$ | $0.433_{.021}$ | $0.608_{.048}$ | $0.565_{.028}$ | $0.582_{.011}$ | $0.501_{.030}$ | $0.591_{.035}$ | $0.534_{.022}$ | $0.504_{.041}$ | $0.432_{.036}$ |
| | DeepAttnMISL | $0.630_{.038}$ | $0.427_{.028}$ | $0.670_{.049}$ | $0.569_{.031}$ | $0.563_{.022}$ | $0.522_{.019}$ | $0.603_{.025}$ | $0.531_{.018}$ | $0.517_{.023}$ | $0.420_{.020}$ |
| Transformer | HIPT | $0.619_{.026}$ | $0.450_{.029}$ | $0.655_{.009}$ | $0.571_{.021}$ | $0.552_{.039}$ | $0.516_{.030}$ | $0.589_{.030}$ | $0.549_{.032}$ | $0.552_{.019}$ | $0.451_{.025}$ |
| | SeTranSurv | $0.677_{.032}$ | $0.401_{.021}$ | $0.688_{.047}$ | $0.550_{.033}$ | $0.580_{.009}$ | $0.489_{.018}$ | $0.612_{.020}$ | $0.537_{.019}$ | $0.554_{.030}$ | $0.458_{.021}$ |
| | ESAT | $0.730_{.039}$ | $0.435_{.034}$ | $0.681_{.050}$ | $0.564_{.033}$ | $0.593_{.019}$ | $0.510_{.014}$ | $0.625_{.031}$ | $0.522_{.015}$ | $0.568_{.017}$ | $0.472_{.032}$ |
| | LongViT | $0.677_{.018}$ | $0.440_{.029}$ | $0.665_{.010}$ | $0.562_{.013}$ | $0.618_{.024}$ | $0.531_{.034}$ | $0.628_{.027}$ | $0.545_{.018}$ | $0.556_{.025}$ | $0.477_{.020}$ |
| | Prov-GigaPath | $0.665_{.030}$ | $0.421_{.039}$ | $0.661_{.034}$ | $0.560_{.032}$ | $0.609_{.028}$ | $0.537_{.017}$ | $0.620_{.015}$ | $0.522_{.010}$ | $0.562_{.021}$ | $0.489_{.023}$ |
| DM | AugDiff | $0.653_{.025}$ | $0.441_{.016}$ | $0.678_{.018}$ | $0.562_{.024}$ | $0.580_{.022}$ | $0.507_{.018}$ | $0.619_{.029}$ | $0.538_{.020}$ | $0.532_{.014}$ | $0.462_{.015}$ |
| | **Ours** | $\mathbf{0.751_{.008}}$ | $\mathbf{0.483_{.020}}$ | $\mathbf{0.714_{.012}}$ | $\mathbf{0.596_{.016}}$ | $\mathbf{0.668_{.012}}$ | $\mathbf{0.562_{.013}}$ | $\mathbf{0.660_{.019}}$ | $\mathbf{0.572_{.010}}$ | $\mathbf{0.601_{.010}}$ | $\mathbf{0.512_{.009}}$ |

Table 2: The results achieved by all competing methods on five datasets. The boldface indicates the best result.

| Generation | NLST | | | | LUAD | | | | LUSC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Low | Medium | High | Gap | Low | Medium | High | Gap | Low | Medium | High | Gap |
| w/o Generation | $0.586_{.039}$ | $0.601_{.030}$ | $0.628_{.025}$ | 0.042 | $0.530_{.037}$ | $0.551_{.041}$ | $0.557_{.039}$ | 0.027 | $0.597_{.045}$ | $0.617_{.032}$ | $0.621_{.029}$ | 0.024 |
| WSI Diffusion | $0.630_{.018}$ | $0.639_{.025}$ | $0.644_{.025}$ | 0.014 | $0.572_{.020}$ | $0.581_{.019}$ | $0.588_{.022}$ | 0.016 | $0.629_{.038}$ | $0.641_{.021}$ | $0.642_{.020}$ | 0.013 |

Table 3: C-index performance across different groups of patients with and without generation.

## 4.2 Main Performance Comparison

Table 2 presents the 5-fold cross-validation results, specifically the C-index and STAGE-5 metrics, for all competing methodologies. Our approach achieves superior performance, with an average increase of 3.2% in C-index and 2.6% in STAGE-5. The advantages include its capability to generate "missing" WSIs, which enhances patient-level representation, and its alignment with the aggregation of WSI representations. Transformer-based models generally outperform CNN-based methods. Our model also outperforms WSI classification models such as LongViT and Prov-GigaPath by 3.5% in C-index, underscoring the difference between WSI classification and survival prediction tasks. While WSI classification involves a single WSI with a known label, survival prediction at the patient level requires multiple WSIs with patient-level labels but unknown WSI-level labels. Applying WSI classification models to survival prediction may ignore critical WSI-level information.

## 4.3 Performance on Mitigating Bias

We conduct an experiment to validate our approach to addressing the bias in WSI counts across patients and its impact on the survival prediction task. Specifically, we divide patients into three groups based on their number of WSIs: low (fewer than 3 WSIs), medium (exactly 3 WSIs), and high (more than 3 WSIs). As shown in Table 3, the performance gap between these patient groups is reduced when generated WSIs are utilized, compared to when no generation is applied, confirming that WSI generation effectively mitigates this challenge. Notably, even the high group's performance improves with the use of generated WSI representations, demonstrating the quality of our generated WSIs.

## 4.4 Ablation Study

The effectiveness of various WSI-Diffusion components is evaluated through ablation studies with the following settings: (1) w/o Generation: predictions based on available WSIs only; (2) w/o WSI-level: generation using only the patch-level diffuser; (3) w/o Distribution: generation without tissue type distribution; (4) w/o Consistency: generation without patch consistency constraints; and (5) Complete: the complete WSI-Diffusion. As shown in Table 4, setting (2) does not outperform setting (1), indicating that randomly generated patches without WSI-level constraints may introduce noise rather than improve predictions. Comparing settings (2), (3), and (4) highlights the importance of WSI-level constraints. Jointly generating patch consistency constraints and tissue type distributions significantly boosts performance compared to individual generations.

| Settings | *C-index* ↑ | | *STAGE-5* ↑ | |
|---|---|---|---|---|
| | **NLST** | **LUSC** | **NLST** | **LUSC** |
| (1) w/o Generation | $0.698_{.030}$ | $0.668_{.019}$ | $0.441_{.012}$ | $0.545_{.027}$ |
| (2) w/o WSI-level | $0.685_{.029}$ | $0.663_{.021}$ | $0.440_{.018}$ | $0.552_{.013}$ |
| (3) w/o Distribution | $0.709_{.022}$ | $0.690_{.023}$ | $0.468_{.019}$ | $0.568_{.019}$ |
| (4) w/o Consistency | $0.725_{.013}$ | $0.700_{.023}$ | $0.465_{.011}$ | $0.571_{.015}$ |
| (5) **Complete** | $\mathbf{0.751_{.008}}$ | $\mathbf{0.714_{.012}}$ | $\mathbf{0.483_{.020}}$ | $\mathbf{0.596_{.016}}$ |

Table 4: The results of the ablation experiments on two datasets. The best results are highlighted in bold.

| Dataset | Time step T | | | | |
|---|---|---|---|---|---|
| | 10 | 20 | 30 | 40 | 50 |
| NLST | $0.726_{.012}$ | $0.737_{.019}$ | $\mathbf{0.751_{.008}}$ | $0.750_{.011}$ | $0.750_{.009}$ |
| LUSC | $0.682_{.030}$ | $0.695_{.013}$ | $\mathbf{0.714_{.012}}$ | $0.710_{.017}$ | $0.712_{.020}$ |
| LUAD | $0.632_{.015}$ | $0.652_{.019}$ | $\mathbf{0.668_{.012}}$ | $0.667_{.019}$ | $0.665_{.011}$ |

Table 5: C-index score with varying time steps $T$.

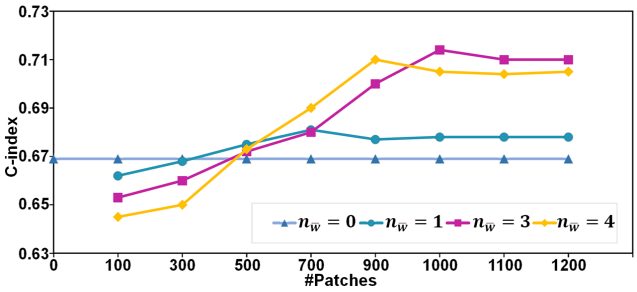


Figure 4: C-index results achieved by the proposed model with varying numbers of generated WSIs and different numbers of patches per generated WSI on the LUSC dataset.

| | NLST | LUSC | LUAD |
|---|---|---|---|
| Wholly | $0.751_{.008}$ | $0.714_{.012}$ | $0.668_{.012}$ |
| Partially | $0.732_{.015}$ | $0.682_{.030}$ | $0.632_{.020}$ |
| Unequally | $0.689_{.042}$ | $0.657_{.033}$ | $0.601_{.030}$ |

Table 6: C-index score of different generation strategies.

### 4.5 Hyperparameter Sensitivity Analysis

**Influence of the number of generated WSIs and patches in each WSI:** The number of generated WSIs is directly correlated with the number of patches within each WSI. In Fig. 4, $n_{\overline{w}}$ denotes the number of generated WSIs. As $n_{\overline{w}}$ increases, performance gradually improves before reaching a plateau, as depicted in Fig. 4. When the number of patches in a generated WSI is limited, the representation may be insufficient. However, as the number of patches increases, performance stabilizes. Since the patches are sufficiently diverse to mimic the WSI sampling process, adding more samples doesn't enhance WSI-level information.

**Influence of the number of time steps $T$:** Since our WSI-level and patch-level representations utilize a common feature extractor (i.e., HIPT (Chen et al. 2022)), the experiments are conducted solely at the patch level by varying $T$ from 10 to 50. As shown in Table 5, the model's performance improves with increasing $T$ and eventually reaches a plateau. This plateau may occur because excessively long time steps can introduce significant noise, thereby reducing the quality of the generated samples (Wallace et al. 2023; Tevet et al. 2022; Zhang et al. 2023).

### 4.6 Analysis of Different Generation Strategies

To address the issue of varying numbers of WSIs across patients, we explore different strategies for determining how many WSIs to generate for each patient. We design three strategies: (1) Wholly: generate the same number of WSIs for all patients as in the current approach; (2) Partially: generate WSIs only for patients with fewer than three WSIs; and (3) Unequally: ensure that the total number of generated WSIs combined with the original WSIs is equal across patients. Table 6 shows that the partially generated WSIs are less effective than the wholly generated WSIs. This may be because even patients with more WSIs benefit from additional generated WSIs, as capturing a comprehensive prognosis with a limited number of WSIs is challenging. Furthermore, generating WSIs unequally further reduces performance, likely due to excessive generation diluting the real WSIs in patients with fewer WSIs, thus introducing noise.

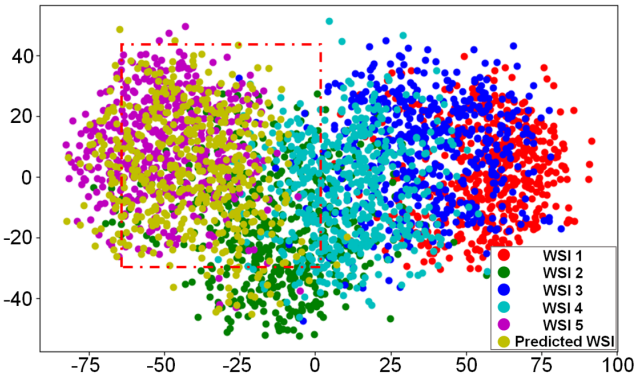


Figure 5: An illustration of the generated patch embeddings.

### 4.7 Case Study

To assess the quality of our generated representations, a case study is conducted on a patient with five WSIs using t-SNE (Van der Maaten and Hinton 2008) for visualization. Our method is trained to predict a set of generated patch representations for WSI-5 using the remaining four WSIs. Each dot represents a patch, with different colors denoting different WSIs. As shown in Fig. 5, the generated WSI closely aligns with the original one, demonstrating the effectiveness of our method in maintaining consistency across patches. The substantial overlap may be attributed to the careful consideration of tissue distribution in the generated patches.

## 5 Conclusions and Limitations

In this paper, we address the issue of varying numbers of WSIs in survival prediction scenarios, which can introduce substantial bias and reduce predictive performance. To mitigate this problem, we propose WSI-Diffusion, a method designed to generate "missing" WSIs for improved survival prediction. This method employs a hierarchical approach with two types of diffusers: WSI-level diffusers establish constraints at the WSI level, while patch-level diffusers generate patches to model the "missing" WSI based on these constraints. Both diffusers are easily trainable simultaneously and enhance survival prediction accuracy. Our approach demonstrates superior performance in predicting survival outcomes across five large cancer datasets. In future work, we will refine tissue type distribution by expanding patch classification. We plan to use clustering to create pseudo-labels to address the absence of subtype labels, although this may impact explainability.

## Acknowledgments

This work was supported by the Natural Science Foundation of China (No. 62372057). This work was partially sponsored by the CAAI-Huawei MindSpore Open Fund. Compute services from Hebei Artificial Intelligence Computing Center.

## References

Aberle, D. R.; Berg, C. D.; Black, W. C.; Church, T. R.; Fagerstrom, R.; Galen, B. A.; Gareen, I. F.; Gatsonis, C.; Goldin, J. G.; Gohagan, J. K.; Hillman, B. J.; Jaffe, C. C.; Kramer, B. S.; Lynch, D. A.; Marcus, P. M.; Schnall, M. D.; Sullivan, D. C.; Sullivan, D.; and Zylak, C. J. 2011. The National Lung Screening Trial: overview and study design. *Radiology*, 258 1: 243–53.

Albertina, B.; Watson, M.; Holback, C.; Jarosz, R.; Kirk, S.; Lee, Y.; Rieger-Christ, K.; and Lemmerman, J. 2016. The Cancer Genome Atlas Lung Adenocarcinoma Collection (TCGA-LUAD)(Version 4)[Data Set]. *The Cancer Imaging Archive*.

Azadi, P.; Suderman, J.; Nakhli, R.; Rich, K.; Asadi, M.; Kung, S.; Oo, H.; Keyes, M.; Farahani, H.; MacAulay, C.; et al. 2023. ALL-IN: AL ocal GL obal Graph-Based DI stillatio N Model for Representation Learning of Gigapixel Histopathology Images With Application In Cancer Risk Assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 765–775. Springer.

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; Prastawa, M.; Alberts, E.; Lipková, J.; Freymann, J. B.; Kirby, J. S.; Bilello, M.; Fathallah-Shaykh, H. M.; Wiest, R.; Kirschke, J. S.; Wiestler, B.; Colen, R. R.; Kotrotsou, A.; LaMontagne, P. J.; Marcus, D. S.; Milchenko, M.; Nazeri, A.; Weber, M.-A.; Mahajan, A.; Baid, U.; Kwon, D.; Agarwal, M.; Alam, M.; Albiol, A.; Albiol, A.; Varghese, A.; Tuan, T. A.; Arbel, T.; Avery, A.; Pranjal, B.; Banerjee, S.; Batchelder, T.; Batmanghelich, N. K.; Battistella, E.; Bendszus, M.; Benson, E.; Bernal, J.; Biros, G.; Cabezas, M.; Chandra, S.; Chang, Y.-J.; and et al. 2018. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *ArXiv*, abs/1811.02629.

Chen, J.; Zhang, R.; Yu, T.; Sharma, R.; Xu, Z.; Sun, T.; and Chen, C. 2024. Label-retrieval-augmented diffusion models for learning from noisy labels. *Advances in Neural Information Processing Systems*, 36.

Chen, R. J.; Chen, C.; Li, Y.; Chen, T. Y.; Trister, A. D.; Krishnan, R. G.; and Mahmood, F. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16144–16155.

Fan, L.; Sowmya, A.; Meijering, E. H. W.; and Song, Y. 2022. Cancer Survival Prediction From Whole Slide Images With Self-Supervised Learning and Slide Consistency. *IEEE Transactions on Medical Imaging*, 42: 1401–1412.

Gamper, J.; Alemi Koohbanani, N.; Benet, K.; Khuram, A.; and Rajpoot, N. 2019. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, 11–19. Springer.

Giannone, G.; Nielsen, D.; and Winther, O. 2022. Few-Shot Diffusion Models. *ArXiv*, abs/2205.15463.

Graham, S.; Vu, Q. D.; Raza, S. E. A.; Azam, A.; Tsang, Y. W.; Kwak, J. T.; and Rajpoot, N. 2019. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58: 101563.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, Z.; Chai, H.; Wang, R.; Wang, H.; Yang, Y.; and Wu, H. 2021. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 561–570. Springer.

Jaume, G.; Pati, P.; Anklin, V.; Foncubierta, A.; and Gabrani, M. 2021. HistoCartography: A Toolkit for Graph Analytics in Digital Pathology. *ArXiv*, abs/2107.10073.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirk, S.; Lee, Y.; Kumar, P.; et al. 2016. The cancer genome atlas lung squamous cell carcinoma collection (TCGA-LUSC), version 4 [Dataset]. *The Cancer Imaging Archive*.

Li, C.; Wang, S.; Yang, D.; Li, Z.; Yang, Y.; Zhang, X.; and Zhou, J. 2017. PPNE: property preserving network embedding. In *Database Systems for Advanced Applications: 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I 22*, 163–179. Springer.

Liu, J.; Zhang, Y.; Chen, J.; Xiao, J.; Lu, Y.; Landman, B. A.; Yuan, Y.; Yuille, A. L.; Tang, Y.; and Zhou, Z. 2023a. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 21095–21107.

Liu, P.; Ji, L.; Ye, F.; and Fu, B. 2023b. GraphLSurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images. *Computer methods and programs in biomedicine*, 231: 107433.

Otsu, N. 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.*, 9: 62–66.

Sanegre, S.; Eritja, N.; de Andrea, C. E.; Díaz-Martín, J. J.; Díaz-Lagares, Á.; Jácome, M. A.; Salguero-Aranda, C.; Ros, D. G.; Davidson, B.; Lopez, R.; Melero, I.; Navarro, S.; y. Cajal, S. R.; de Álava, E.; Matías-Guiu, X.; and Noguera, R. 2021. Characterizing the Invasive Tumor Front of Aggressive Uterine Adenocarcinoma and Leiomyosarcoma. *Frontiers in Cell and Developmental Biology*, 9.

Shao, Z.; Chen, Y.; Bian, H.; Zhang, J.; Liu, G.; and Zhang, Y. 2023a. Hvtsurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2, 2209–2217.

Shao, Z.; Dai, L.; Wang, Y.; Wang, H.; and Zhang, Y. 2023b. Augdiff: Diffusion based feature augmentation for multiple instance learning in whole slide image. *arXiv preprint arXiv:2303.06371*.

Shen, Y.; Liu, L.; Tang, Z.; Chen, Z.; Ma, G.; Dong, J.; Zhang, X.; Yang, L.; and Zheng, Q. 2022. Explainable survival analysis with convolution-involved vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2207–2215.

Tang, B.; Li, A.; Li, B.; and Wang, M. 2019. CapSurv: Capsule Network for Survival Analysis With Whole Slide Pathological Images. *IEEE Access*, 7: 26022–26030.

Team, N. L. S. T. R. 2011. The national lung screening trial: overview and study design. *Radiology*, 258(1): 243–253.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human Motion Diffusion Model. *ArXiv*, abs/2209.14916.

Ulhaq, A.; Akhtar, N.; and Pogrebna, G. 2022. Efficient Diffusion Models for Vision: A Survey. *ArXiv*, abs/2210.09292.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *ArXiv*, abs/1706.03762.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S. R.; and Naik, N. 2023. Diffusion Model Alignment Using Direct Preference Optimization. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8228–8238.

Wang, W.; Ma, S.; Xu, H.; Usuyama, N.; Ding, J.; Poon, H.; and Wei, F. 2023. When an Image is Worth 1, 024 x 1, 024 Words: A Case Study in Computational Pathology. *ArXiv*, abs/2312.03558.

Wu, T.; Ye, S.; Chen, S.; Peng, Q.; and You, X. 2023. Detail Reinforcement Diffusion Model: Augmentation Fine-Grained Visual Categorization in Few-Shot Conditions. *ArXiv*, abs/2309.08097.

Xu, H.; Usuyama, N.; Bagga, J.; Zhang, S.; Rao, R.; Naumann, T.; Wong, C.; Gero, Z.; González, J.; Gu, Y.; Xu, Y.; Wei, M.-H.; Wang, W.; Ma, S.; Wei, F.; Yang, J.; yue Li, C.; Gao, J.; Rosemon, J.; Bower, T.; Lee, S.; Weerasinghe, R. K.; Wright, B.; Robicsek, A.; Piening, B.; Bifulco, C.; Wang, S.; and Poon, H. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630: 181 – 188.

Yang, H.; Zhou, J. T.; Cai, J.; and Ong, Y. 2017. MIML-FCN+: Multi-Instance Multi-Label Learning via Fully Convolutional Networks with Privileged Information. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5996–6004.

Yao, J.; Zhu, X.; Jonnagaddala, J.; Hawkins, N.; and Huang, J. 2020a. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65: 101789.

Yao, J.; Zhu, X.; Jonnagaddala, J.; Hawkins, N.; and Huang, J. 2020b. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis*, 65: 101789.

Zaffar, I.; Jaume, G.; Rajpoot, N. M.; and Mahmood, F. 2022. Embedding Space Augmentation for Weakly Supervised Learning in Whole-Slide Images. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–4.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I.-S. 2023. Text-to-image Diffusion Models in Generative AI: A Survey. *ArXiv*, abs/2303.07909.

Zhang, L.; Zhang, X.; and Pan, J. 2022. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11676–11684.

Zhang, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Wang, S.; Philip, S. Y.; and Li, C. 2024. Early Detection of Multimodal Fake News via Reinforced Propagation Path Generation. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Yu, P. S.; and Li, C. 2025. Knowledge-aware multimodal pre-training for fake news detection. *Information Fusion*, 114: 102715.

Zhao, J.; Li, C.; Wen, Q.; Wang, Y.; Liu, Y.; Sun, H.; Xie, X.; and Ye, Y. 2021. Gophormer: Ego-graph transformer for node classification. *arXiv preprint arXiv:2110.13094*.

Zhu, X.; Yao, J.; and Huang, J. 2016. Deep convolutional neural network for survival analysis with pathological images. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 544–547.

Zhu, X.; Yao, J.; Zhu, F.; and Huang, J. 2017. WSISA: Making Survival Prediction from Whole Slide Histopathological Images. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6855–6863.