Deep Association Multimodal Learning for Zero-shot Spatial Transcriptomics Prediction

Yijing Zhou[1], Yadong Lu[1], Qingli Li[1], Xinxing Li[2], and Yan Wang[1(✉)]

[1] Shanghai Key Laboratory of Multidimensional Information Processing,
East China Normal University, Shanghai, China
[2] Department of Gastrointestinal Surgery,
Tongji Hospital Medical College of Tongji University
51275904063@stu.ecnu.edu.cn, yadonglu@stu.ecnu.edu.cn,
qlli@cs.ecnu.edu.cn, ahtxxxx2015@163.com, ywang@cee.ecnu.edu.cn

Abstract. Spatial transcriptomics enables localized gene expression profiling within histological regions. Current supervised methods struggle to infer patterns for novel gene types beyond their training scope, while existing zero-shot frameworks partially address this by incorporating gene semantics, the "independent learning" paradigms hamper their usage in zero-shot gene expression prediction. Specifically, they learn tissue morphology and gene semantics (inter-modality) independently, and treat gene functions (intra-modality) as independent entities. In this paper, we present a deep association multimodal framework which bridges pathological image with gene functionality semantics for zero-shot expression prediction. Concretely, our framework achieves generalized expression prediction by integrating nuclei-aware spatial modeling that preserves tissue microarchitecture, cross-modal alignment of pathological features with gene functionality semantics via iterative vision-language prompt learning, and gene interaction modeling that dynamically captures relationships across gene descriptions. On standard benchmark datasets, we demonstrate competitive zero-shot performance compared to other competitors (*e.g.*, outperforms 16.3% in mean Pearson Correlation Coefficient on cSCC dataset), and we show clinical interpretability of our method. Codes is publicly available at https://github.com/DeepMed-Lab-ECNU/ALIGN-ST.

Keywords: Spatial transcriptomics · Gene expression prediction · Computational pathology · Zero-shot learning.

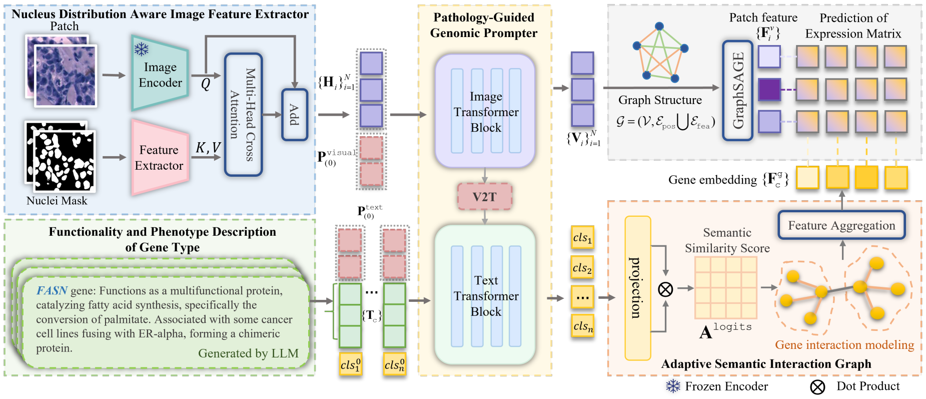
1 Introduction

Spatial transcriptomics (ST) has emerged as a transformative technology for mapping gene expression within histological tissue regions [14], providing critical insights into disease progression and cellular microenvironments [18]. By correlating localized gene activity with morphological patterns in tissue slides, ST enables biomarker discovery and mechanistic studies of pathological processes [15,12,10]. However, widespread adoption of ST remains constrained by the high cost and technical complexity of sequencing-based gene expression profiling [2], creating a critical bottleneck for training robust deep learning models to predict gene expression directly from pathological images.

Traditional supervised methods [3,6,13,9,16] have been employed to predict gene expression from whole slide images. While these models have demonstrated promising performance, they are fundamentally constrained by their training data, as they can only predict gene types seen during training. For example, when trained on the HER2+ dataset, these models are predominantly optimized for highly expressed gene sequences, which restricts their ability to generalize to unseen gene types. This limitation necessitates costly data collection and retraining for each new set of gene types, making supervised approaches impractical for large-scale biomedical applications.

SGN [17] is the only existing zero-shot gene expression prediction framework, pioneering the integration of functional semantics of genes derived from large language models (LLMs) [7]. By generating textual descriptions of gene functions and phenotypes, SGN establishes preliminary associations between pathological patterns and molecular mechanisms, enabling prediction of unseen gene types without prior expression data. While SGN represents a significant step forward, its "independent learning" paradigm still hampers its usage in zero-shot gene expression prediction. Independent learning paradigm includes: (1) tissue morphology and gene semantics are learned independently, and (2) gene functions are treated as independent entities. Thus, the gene type feature extractor is only optimized for the training (seen) gene types, which are sensitive to class shift. Besides, it lacks the ability in explicitly elevating the biological connections and co-expression patterns among gene types, which are crucial for generalization.

To enhance its ability in predicting unseen gene types, we introduce a "deep association learning" paradigm, including (1) dynamic mapping between tissue morphology and gene semantics, which enables gene type features to be better optimized to describe each tissue slide image (instead of being overfitting to the training gene types), and (2) gene interaction modeling, which further models gene-gene dependencies based on semantic relevance, enabling context-aware embeddings. Furthermore, gene spatial expression is tied to the geometric information within histological images, including tissue structure and cell distribution [4]. Thus, to extract more gene-related image features, we further supplement image features with tissue-wide structure and fine-grained nuclei distributions. By integrating these components, our framework establishes a biologically meaningful, dynamically adaptive approach to zero-shot gene expression prediction. Experimental results on benchmark datasets show that our method outperforms SGN, while maintaining competitive performance with supervised methods and extending prediction capabilities to previously unseen gene types.


Nucleus Distribution Aware Image Feature Extractor
Patch
Image Encoder
Q
Multi-Head Cross Attention
Add
Feature Extractor
K, V
Nuclei Mask
$\{\mathbf{H}_i\}_{i=1}^N$
$\mathbf{P}_{(0)}^{\text{visual}}$
$\mathbf{P}_{(0)}^{\text{text}}$
Functionality and Phenotype Description of Gene Type

FASN gene: Functions as a multifunctional protein, catalyzing fatty acid synthesis, specifically the conversion of palmitate. Associated with some cancer cell lines fusing with ER-alpha, forming a chimeric protein.

Generated by LLM

$\{\mathbf{T}_c\}$
cls_1^0
cls_n^0
Pathology-Guided Genomic Prompter
Image Transformer Block
V2T
Text Transformer Block
$\{\mathbf{V}_i\}_{i=1}^N$
cls_1
cls_2
...
cls_n
Graph Structure
$\mathcal{G} = (\mathcal{V}, \mathcal{E}_{\text{pos}} \bigcup \mathcal{E}_{\text{fea}})$
GraphSAGE
Patch feature $\{\mathbf{F}_i^v\}$
Prediction of Expression Matrix
Gene embedding $\{\mathbf{F}_c^g\}$
Feature Aggregation
projection
Semantic Similarity Score
$\mathbf{A}^{\text{logits}}$
Gene interaction modeling
Adaptive Semantic Interaction Graph
Frozen Encoder
⊗ Dot Product

Fig. 1. Given a slide image containing N patches $\{\mathbf{X}_i\}_{i=1}^N$, our framework predicts gene expression for both seen ($\mathtt{C}^{\mathtt{s}}$) and unseen ($\mathtt{C}^{\mathtt{u}}$) gene types. We have four stages: 1)Nuclei Distribution Aware Image Feature Extractor fuses each patch's global tissue semantics and nuclei spatial distributions; 2)Pathology-Guided Genomic Prompter aligns LLM-generated gene descriptions $\{\mathbf{T}_{\mathtt{c}}\}$ with image features through dual stream prompt learning. 3)Adaptive Semantic Interaction Graph models gene dependencies via top$k$ semantic neighbor selection and weighted feature aggregation. 4)Gene Expression Prediction refines patch features via GraphSAGE network and computes expression values through dot products between enhanced features and gene embeddings.

2 Method

Given a WSI containing N patches $\{\mathbf{X}_i\}_{i=1}^N$, where $\mathbf{X}_i \in \mathbb{R}^{H \times W \times 3}$, the goal is to predict the gene expression values $\{y_{i,\mathtt{c}}\}_{i=1}^N$ for all patches, where $\mathtt{c} \in \mathtt{C}^{\mathtt{s}} \bigcup \mathtt{C}^{\mathtt{u}}$ is the gene type of interest (both seen type $\mathtt{C}^{\mathtt{s}}$ and unseen type $\mathtt{C}^{\mathtt{u}}$). For each gene type $\mathtt{c}$, following SGN [17], we utilize a pre-trained LLM [7] to generate the description, using the last hidden state of the LLM as our description $\mathbf{T}_{\mathtt{c}} \in \mathbb{R}^{L \times D^{\mathtt{T}}}$, where L is the length of the description and $D^{\mathtt{T}}$ is the feature dimension. As shown in Fig. 1, our framework extracts nucleus-aware features by integrating tissue morphology with nuclei distributions. These features align with gene descriptions $\{\mathbf{T}_{\mathtt{c}}\}$ via cross-modal prompt fusion, dynamically adapting descriptions to pathological context. Subsequently, an adaptive semantic graph models gene relationships via top-k semantic neighbor selection and weighted feature aggregation. Gene expression prediction is derived from dot product between image feature and gene embedding.

2.1 Nuclei Distribution Aware Image Feature Extractor

Previous gene expression prediction methods [6,9,13,16,17] rely solely on global patch-level features, overlooking fine-grained cellular spatial information essential for modeling localized gene expression patterns. To address this limitation, we propose **N**uclei **D**istribution **A**ware **I**mage **F**eature **E**xtractor (NDA-IFE) to

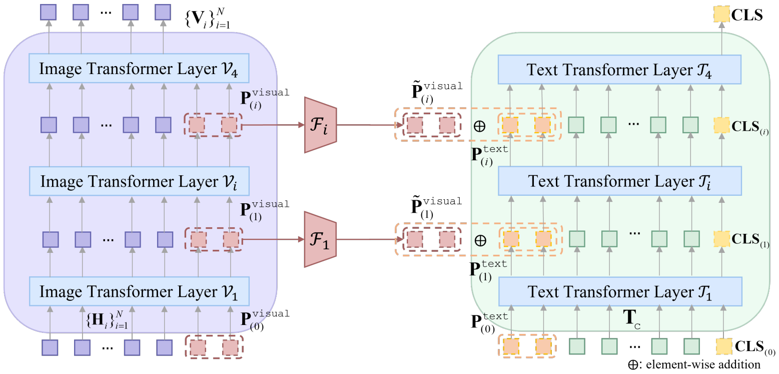

$\{\mathbf{V}_i\}_{i=1}^N$
CLS
Image Transformer Layer $\mathcal{V}_4$
Text Transformer Layer $\mathcal{T}_4$
$\mathbf{P}^{\text{visual}}_{(i)}$
$\mathcal{F}_i$
$\tilde{\mathbf{P}}^{\text{visual}}_{(i)}$
⊕
$\text{CLS}_{(i)}$
$\mathbf{P}^{\text{text}}_{(i)}$
Image Transformer Layer $\mathcal{V}_i$
Text Transformer Layer $\mathcal{T}_i$
$\mathbf{P}^{\text{visual}}_{(1)}$
$\mathcal{F}_1$
$\tilde{\mathbf{P}}^{\text{visual}}_{(1)}$
⊕
$\text{CLS}_{(1)}$
$\mathbf{P}^{\text{text}}_{(1)}$
Image Transformer Layer $\mathcal{V}_1$
Text Transformer Layer $\mathcal{T}_1$
$\{\mathbf{H}_i\}_{i=1}^N$
$\mathbf{P}^{\text{visual}}_{(0)}$
$\mathbf{P}^{\text{text}}_{(0)}$
$\mathbf{T}_c$
$\text{CLS}_{(0)}$
⊕: element-wise addition

Fig. 2. Illustration of Pathology-Guided Genomic Prompter module.

integrate global context modeling and nuclei-aware localization. Given an image patch $\mathbf{X}_i$, we extract global semantic features $\mathbf{h}_i^{\mathsf{g}} \in \mathbb{R}^{D^{\mathsf{h}}}$ using ResNet-18. Meanwhile, Hover-net [5] generates a segmentation mask $\mathbf{S}_i \in \{0,1\}^{H\times W}$, followed by a lightweight convolutional module to encode cellular spatial distributions:

$$\mathbf{h}_i^{\mathsf{s}} = \mathcal{F}_{\text{conv}}(\mathbf{S}_i), \quad \mathbf{h}_i^{\mathsf{s}} \in \mathbb{R}^{D^{\mathsf{h}}} , \tag{1}$$

where $\mathcal{F}_{\text{conv}}(\cdot)$ comprises five 3×3 convolutions with BatchNorm and ReLU.

A cross-attention layer with residual connection dynamically aligns global tissue semantics ($\mathbf{h}_i^{\mathsf{g}}$ as query) with cellular architectures ($\mathbf{h}_i^{\mathsf{s}}$ as key/value), while preserving the original contextual information, thus effectively supplementing image features with fine-grained nuclei distributions:

$$\mathbf{H}_i = \texttt{Attention}(Q = \mathbf{h}_i^{\mathsf{g}},\, K = \mathbf{h}_i^{\mathsf{s}},\, V = \mathbf{h}_i^{\mathsf{s}}) + \mathbf{h}_i^{\mathsf{g}} \ , \quad \mathbf{H}_i \in \mathbb{R}^{D^{\mathsf{h}}} \ . \tag{2}$$

2.2 Pathology-Guided Genomic Prompter

Previous works [17] computes gene expression through a static dot product between patch-level image features and gene type descriptions, neglecting associations between pathological patterns and gene functions. We propose **P**athology-**G**uided **G**enomic **P**rompter (PGGP) module (Fig. 2), a dual-stream architecture that establishes deep associations via dynamic cross-modal alignment, iteratively refining gene semantics via pathology-guided prompt interactions.

Dual-Stream Feature Encoding. For pathological image feature encoding, given patch features $\{\mathbf{H}_i\}_{i=1}^N$, we enhance global contextual modeling through a 4-layer Image Transformer Encoder. The output feature of the l-th image encoder is calculated as:

$$\mathbf{E}_{(l)}^{\texttt{img}} = \texttt{ImageTransformer}_{(l)}\left(\mathbf{E}_{(l-1)}^{\texttt{img}}\right) , \mathbf{E}_{(0)}^{\texttt{img}} = [\mathbf{P}_{(0)}^{\texttt{visual}}; \mathbf{H}] \ , \tag{3}$$

where $\mathbf{E}^{\texttt{img}}_{(l)} \in \mathbb{R}^{(N+2)\times D^{\text{h}}}$ contains two learnable visual prompt tokens $\mathbf{P}^{\texttt{visual}}_{(l)} \in \mathbb{R}^{2\times D^{\text{h}}}$ for the l-th layer.

For genomic text encoding, gene type description $\mathbf{T}_{\texttt{c}} \in \mathbb{R}^{L\times D^{\text{T}}}$ are projected to $\tilde{\mathbf{T}}_{\texttt{c}} \in \mathbb{R}^{L\times D^{\text{t}}}$ via a learnable linear layer. Then, a 4-layer Text Transformer Encoder generates hierarchical semantic representations:

$$\mathbf{E}^{\texttt{text}}_{(l)} = \texttt{TextTransformer}_{(l)}\left(\mathbf{E}^{\texttt{text}}_{(l-1)}\right),\mathbf{E}^{\texttt{text}}_{(0)} = [\mathbf{CLS}_{(0)};\mathbf{P}^{\texttt{text}};\tilde{\mathbf{T}}_{\texttt{c}}]\ , \tag{4}$$

where $\mathbf{E}^{\texttt{text}}_{(l)} \in \mathbb{R}^{(L+3)\times D^{\text{t}}}$ incorporates a learnable class token $[\mathbf{CLS}_{(l)}]$, two text prompts $\mathbf{P}^{\texttt{text}}_{(l)} \in \mathbb{R}^{2\times D^{\text{t}}}$ and text features.

Cross-Modal Prompt Interaction. To enable cross-modal alignment between histology features and gene semantics, as shown in Fig. 2, at the l-th layer, visual prompts $\mathbf{P}^{\texttt{visual}}_{(l)}$ are projected to the text feature space via $f_{\text{proj}} : \mathbb{R}^{D^{\text{h}}} \rightarrow \mathbb{R}^{D^{\text{t}}}$, resulting in $\tilde{\mathbf{P}}^{\texttt{visual}}_{(l)}$. Text prompts are updated through gated fusion:

$$\mathbf{P}^{\texttt{text}}_{(l)} = \alpha\tilde{\mathbf{P}}^{\texttt{visual}}_{(l)} + (1-\alpha)\mathbf{P}^{\texttt{text}}_{(l-1)}, \tag{5}$$

where $\alpha \in [0,1]$ is a hyperparameter that controls cross-modal fusion rate.

The refined image features $\mathbf{V} \in \mathbb{R}^{N\times D^{\text{h}}}$ are derived from the last image encoder layer, while refined gene type embedding $\mathbf{G}_{\texttt{c}} \in \mathbb{R}^{D^{\text{t}}}$ is obtained from the last text encoder layer's class token.

2.3 Adaptive Semantic Interaction Graph

Treating gene functions as independent entities neglects crucial functional dependencies between genes [17]. We propose an **A**daptive **S**emantic **I**nteraction **G**raph (ASIG) module that explicitly constructs gene-gene interaction graphs based on semantic relevance learned from textual descriptions, mining deep associations between independent gene embeddings.

Adaptive Graph Construction. Given gene type embeddings $\{\mathbf{G}_{\texttt{c}}\}^{N^{\text{g}}}$ from Section 2.2, where N^{g} denotes the number of genes. We first project embeddings into an interaction space $\mathbf{Z} \in \mathbb{R}^{N^{\text{g}}\times D^{\text{e}}}$ via a learnable linear layer. The semantic affinity matrix $\mathbf{A}^{\texttt{logits}}$ is computed as:

$$\mathbf{A}^{\texttt{logits}} = \texttt{Softmax}\left(\frac{\mathbf{Z}\mathbf{Z}^{\top}}{\sqrt{D^{\text{e}}}}\right) \in \mathbb{R}^{N^{\text{g}}\times N^{\text{g}}}\ , \tag{6}$$

where each element $\mathbf{A}^{\texttt{logits}}[i,j]$ indicates the semantic similarity between gene $\texttt{C}_i$ and $\texttt{C}_j$. For each gene node, we adaptively select top-k semantically relevant neighbors based on $\mathbf{A}^{\texttt{logits}}$ to construct a adjacency matrix $\mathbf{A} \in \{0,1\}^{N^{\text{g}}\times N^{\text{g}}}$.

Semantic-Aware Gating Aggregation. To enhance feature aggregation, inspired by knowledge-aware graph attention mechanisms [11], we propose a gated attention mechanism that adaptively modulates neighbor influence. For target gene $\texttt{C}_i$ and its neighbors $\texttt{C}_j \in \mathcal{N}^{\text{gene}}_i$, the attention weight is computed as:

$$\beta_{ij} = \frac{\exp(\mathbf{Z}_j \odot \tanh(\mathbf{Z}_i + \mathbf{Z}_j))}{\sum_{\texttt{C}_k\in\mathcal{N}_i}\exp(\mathbf{Z}_k \odot \tanh(\mathbf{Z}_i + \mathbf{Z}_k))}\ , \tag{7}$$

Table 1. Performance comparison of different models under traditional supervised (✗) and zero-shot (✓) learning modes.

Model	ZS	HER2+			cSCC		
		MSE↓	PCC@M↑	PCC@H↑	MSE↓	PCC@M↑	PCC@H↑
ST-Net [6]	✗	0.066	0.314	0.446	0.055	0.409	0.513
HistoGene [13]	✗	0.058	0.291	0.422	0.052	0.428	0.514
THItoGene [9]	✗	0.051	0.261	0.396	0.075	0.374	0.465
BLEEP [16]	✗	0.058	0.323	0.421	0.057	0.331	0.461
TRIPLEX [3]	✗	0.044	0.314	0.484	0.060	0.484	0.596
SGN [17]	✗	0.089	0.314	0.424	0.055	0.384	0.456
Ours	✗	0.053	0.355	0.497	0.051	0.491	0.584
SGN [17]	✓	0.137	0.305	0.417	0.096	0.398	0.477
Ours	✓	0.125	0.329	0.466	0.062	0.463	0.542

where $\tanh(\cdot)$ serves as a non-linear activation function, $\odot$ denotes element-wise multiplication. This formulation enables each gene to dynamically adjust the influence of its neighbors based on their semantic relevance. Neighbor feature of target gene $\mathtt{C}_i$ is aggregated as $\mathbf{Z}_i^{\mathcal{N}} = \sum_{\mathtt{C}_j \in \mathcal{N}_i^{\text{gene}}} \beta_{ij} \mathbf{Z}_j$.

The final gene type embedding $\mathbf{F}^{\mathtt{g}}_{\mathtt{C}_i}$ for gene $\mathtt{C}_i$ is obtained by concatenating $\mathbf{Z}_i$ (original feature) and $\mathbf{Z}_i^{\mathcal{N}}$ (aggregated neighbor feature), followed by a linear projection:

$$\mathbf{F}^{\mathtt{g}}_{\mathtt{C}_i} = \mathbf{W}_f [\mathbf{Z}_i \copyright \mathbf{Z}_i^{\mathcal{N}}] \,, \tag{8}$$

where $\copyright$ denotes concatenation and $\mathbf{W}_f \in \mathbb{R}^{2D^{\mathtt{e}} \times D^{\mathtt{e}}}$ is a learnable weight matric.

2.4 Gene Expression Prediction

With refined image patch features $\{\mathbf{V}_i\}_{i=1}^N$, following SGN [17], we construct a graph based on spatial position and feature similarity, followed by feature refinement using a GraphSAGE network to obtain the final patch features $\{\mathbf{F}_i^{\mathtt{v}}\}_{i=1}^N$. Then, with gene type embedding $\mathbf{F}^{\mathtt{g}}_{\mathtt{c}}$, zero-shot expression prediction of gene type $\mathtt{c}$ is performed as:

$$\{\hat{\mathbf{y}}_{i,\mathtt{c}}\}_{i=1}^N = \{\mathbf{F}_i^{\mathtt{v}} \cdot \mathbf{F}_{\mathtt{c}}^{\mathtt{g}\top}\}_{i=1}^N . \tag{9}$$

We train the model with a combined loss function that includes mean squared error (MSE) and batch-wise Pearson correction coefficient (PCC) losses. The $\mathcal{L}_{\text{mse}}$ measures the difference between predicted $\{\hat{\mathbf{y}}_{i,\mathtt{c}}\}_{i=1}^N$ and true gene expression values $\{\mathbf{y}_{i,\mathtt{c}}\}_{i=1}^N$, while the $\mathcal{L}_{\text{pcc}}$ encourages their correlation. The total loss is thus defined as: $\mathcal{L} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{pcc}}$.

3 Experiment

3.1 Experimental Setup

Datasets. We experiment with two datasets: human HER2 positive breast cancer (HER2+) dataset [1] with 36 tissue sections from 8 patients and human cutaneous squamous cell carcinoma (cSCC) dataset [8] with 12 tissue sections

Table 2. Ablation study of our model components on cSCC dataset. ZS denotes learning modes, where "✓" indicates zero-shot learning and "✗" indicates traditional supervised learning. "`Trans`" means replacing ASIG with a transformer layer.

ZS	NDA-IFE	PGGP		ASIG	MSE$_{\times 10^2}$ ↓	PCC@M↑	PCC@H↑
		prompt	V2T				
✓	✗	✓	✓	✓	8.70	0.407	0.480
✓	✓	✗	✗	✓	9.21	0.431	0.510
✓	✓	✓	✗	✓	9.52	0.443	0.522
✗	✓	✓	✗	✓	5.45	0.456	0.541
✓	✓	✓	✓	✗	7.50	0.433	0.514
✓	✓	✓	✓	`Trans.`	10.20	0.445	0.528
✗	✓	✓	✓	✓	**5.08**	**0.491**	**0.584**
✓	✓	✓	✓	✓	**6.23**	**0.463**	**0.542**

from 4 patients. We follow the dataset pre-processing settings of [17] and cross-fold validation settings of [3]. Past works selected the top 250 highly expressed genes per dataset for prediction. To compare with them, in the zero-shot setting, we use their unselected gene types in training as seen gene types and their selected gene types in testing as unseen gene types.

Evaluation Metrics. Our method is evaluated with mean squared error (MSE), mean PCC for all genes (PCC@M) and for the top 50 highly predictive genes (PCC@H).

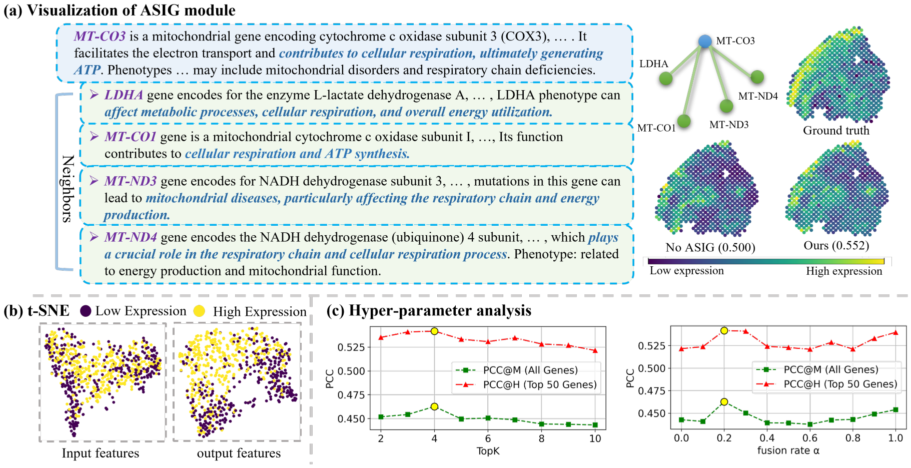
Implementation Details. We train our model respectively for 80 epochs and 100 epochs on the HER2+ dataset and cSCC dataset with batch size 1, where a whole slide image contains up to hundreds of patches in two datasets. We employ a learning rate of 5×10^{-4} and weight decay of 1×10^{-4}. All experiments are conducted on a single NVIDIA GeForce RTX 3090.

3.2 Cross-validation performance on ST datasets

We compare with state-of-the-art methods on the HER2+ and cSCC datasets. As shown in Table 1, our framework significantly advances the zero-shot gene expression prediction paradigm SGN [17], while maintaining competitive performance in supervised learning. In the zero-shot setting, our method comprehensively outperforms SGN, demonstrating strong generalization to unseen gene types through dynamic cross-modal alignment and gene interaction modeling. Notably, our zero-shot predictions significantly narrow the gap with supervised approaches. On cSCC, the PCC@H difference between our zero-shot (0.542) and supervised TRIPLEX (0.596) is only 9%, compared to SGN's 29% gap (0.477 vs. 0.596). Compared to traditional supervised methods, our framework exhibits competitive performance. Our supervised results PCC@M rank first among all methods on two datasets.

3.3 Ablation Study

To quantitatively evaluate the effectiveness of our proposed components, we conduct ablation studies on cSCC dataset, as illustrated in Table 2. Removing



(a) Visualization of ASIG module

MT-CO3 is a mitochondrial gene encoding cytochrome c oxidase subunit 3 (COX3), … . It facilitates the electron transport and ***contributes to cellular respiration, ultimately generating ATP***. Phenotypes … may include mitochondrial disorders and respiratory chain deficiencies.

Neighbors

- ➢ *LDHA* gene encodes for the enzyme L-lactate dehydrogenase A, … , LDHA phenotype can ***affect metabolic processes, cellular respiration, and overall energy utilization.***
- ➢ *MT-CO1* gene is a mitochondrial cytochrome c oxidase subunit I, …, Its function contributes to ***cellular respiration and ATP synthesis.***
- ➢ *MT-ND3* gene encodes for NADH dehydrogenase subunit 3, … , mutations in this gene can lead to ***mitochondrial diseases, particularly affecting the respiratory chain and energy production.***
- ➢ *MT-ND4* gene encodes the NADH dehydrogenase (ubiquinone) 4 subunit, … , which ***plays a crucial role in the respiratory chain and cellular respiration process***. Phenotype: related to energy production and mitochondrial function.

Fig. 3. Visualization of ablation study. (a) Visualization of the ASIG module's impact on gene expression prediction, using MT-CO3 and its neighbor genes (found by ASIG) as an example. (b) t-SNE visualization of input and output patches features of NDA-IFE module. (c) Hyper-parameter analysis of fusion ratio α in PGGP module and number of neighbors top-k in ASIG module.

NDA-IFE results in performance decrease in both PCC and MSE, highlighting the necessity of fine-grained cellular spatial information for accurate gene expression prediction. As shown in Fig. 3 (b), As shown in Fig. 3 (b), using the COL1A1 gene as an example, we find that patch features processed by the NDA-IFE module show a clearer separation of high and low gene expression levels. The absence of cross-modal prompt interaction in PGGP module decreases performance in both learning paradigms, indicating that dynamic cross-modal alignment prevents overfitting by adaptively aligning gene features with histological contexts and capturing generalizable pathology-gene associations in zero-shot settings. Both removing ASIG module and replacing it with a transformer layer degrade performance, indicating that explicitly models gene-gene dependencies based on semantic relevance outperforms uniform attention mechanisms. As shown in Fig. 3 (a), we observe that MT-CO3 and its neighboring genes are related to cellular respiration. The baseline model without ASIG shows a decrease in prediction accuracy for MT-CO3.

As shown in Fig. 3 (c), Hyper-parameter analysis reveals: Optimal cross-modal fusion at $\alpha = 0.2$ balances pathological context and gene-specific semantics, and the optimal ASIG neighbor count is $k = 4$ based on PCC metrics.

4 Conclusion

In this paper, we propose a deep association multimodal framework for zero-shot gene expression prediction, which includes supplementing image features with tissue-wide structure and fine-grained nuclei distributions, dynamic mapping between tissue morphology and gene semantics via iterative vision-language prompt learning, and adaptive gene interaction modeling based on semantic relevance. Experimental results on benchmark datasets demonstrate that our method outperforms SGN, while maintaining competitive performance with supervised approaches.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 62471182), Shanghai Rising-Star Program (Grant No. 24QA2702100), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ2229004).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andersson, A., Larsson, L., Stenbeck, L., Salmén, F., Ehinger, A., Wu, S.Z., Al-Eryani, G., Roden, D., Swarbrick, A., Borg, Å., et al.: Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. Nature communications **12**(1), 6012 (2021)
2. Choe, K., Pak, U., Pang, Y., Hao, W., Yang, X.: Advances and challenges in spatial transcriptomics for developmental biology. Biomolecules **13**(1), 156 (2023)
3. Chung, Y., Ha, J.H., Im, K.C., Lee, J.S.: Accurate spatial gene expression prediction by integrating multi-resolution features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11591–11600 (2024)
4. Gao, R., Yuan, X., Ma, Y., Wei, T., Johnston, L., Shao, Y., Lv, W., Zhu, T., Zhang, Y., Zheng, J., et al.: Harnessing tme depicted by histological images to improve cancer prognosis through a deep learning system. Cell Reports Medicine **5**(5) (2024)
5. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis **58**, 101563 (2019)
6. He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, A., Maaskola, J., Lundeberg, J., Zou, J.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nature Biomedical Engineering **4**, 1–8 (08 2020). https://doi.org/10.1038/s41551-020-0578-x
7. Intel: Neural-chat-v3-1. https://huggingface.co/Intel/neural-chat-7b-v3-1 (2023)
8. Ji, A.L., Rubin, A.J., Thrane, K., Jiang, S., Reynolds, D.L., Meyers, R.M., Guo, M.G., George, B.M., Mollbrink, A., Bergenstråhle, J., et al.: Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. cell **182**(2), 497–514 (2020)
9. Jia, Y., Liu, J., Chen, L., Zhao, T., Wang, Y.: Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. Briefings in Bioinformatics **25**(1), bbad464 (2024)
10. Jiang, J., Liu, Y., Qin, J., Chen, J., Wu, J., Pizzi, M.P., Lazcano, R., Yamashita, K., Xu, Z., Pei, G., et al.: Meti: deep profiling of tumor ecosystems by integrating cell morphology and spatial transcriptomics. Nature communications **15**(1), 7312 (2024)

11. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11323–11332 (2024)
12. Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J.C., Baron, M., Hajdu, C.H., Simeone, D.M., Yanai, I.: Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nature biotechnology **38**(3), 333–342 (2020)
13. Pang, M., Su, K., Li, M.: Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. BioRxiv pp. 2021–11 (2021)
14. Rao, A., Barkley, D., França, G.S., Yanai, I.: Exploring tissue architecture using spatial transcriptomics. Nature **596**(7871), 211–220 (2021)
15. Williams, C.G., Lee, H.J., Asatsuma, T., Vento-Tormo, R., Haque, A.: An introduction to spatial transcriptomics for biomedical research. Genome medicine **14**(1), 68 (2022)
16. Xie, R., Pang, K., Chung, S., Perciani, C., MacParland, S., Wang, B., Bader, G.: Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 70626–70637. Curran Associates, Inc. (2023)
17. Yang, Y., Hossain, M.Z., Li, X., Rahman, S., Stone, E.: Spatial transcriptomics analysis of zero-shot gene expression prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 492–502. Springer (2024)
18. Zhang, L., Chen, D., Song, D., Liu, X., Zhang, Y., Xu, X., Wang, X.: Clinical and translational values of spatial transcriptomics. Signal Transduction and Targeted Therapy **7**(1), 111 (2022)