Contrastive and Adaptive Multi-modal Masked Autoencoder for Spatial Transcriptomics

Joohyeok Kim[1], Taejin Jeong[2], Jinyeong Kim[3], and Seong Jae Hwang[†1]

[1] Department of Artificial Intelligence, Yonsei University, Seoul, Republic of Korea
[2] Department of Computer Science, Yonsei University, Seoul, Republic of Korea
[3] Yonsei University College of Medicine, Seoul, Republic of Korea
{kyyle2114, starforest, jinyeong1324, seongjae}@yonsei.ac.kr

Abstract. The high cost of spatial transcriptomics (ST) has driven extensive studies into predicting gene expression directly from H&E histology images. However, this prediction task faces an inherent limitation, as tissue morphology alone provides insufficient information to fully resolve underlying gene expression. To address this limitation, a recent study leverages partial gene expression to guide the prediction process alongside histology images. Building on this paradigm, we approach the prediction task as a spatial imputation problem, employing a Masked Autoencoder (MAE) to utilize a small fraction of gene expression as genetic anchors for inferring whole-slide gene expression profiles. Specifically, we propose a bio-saliency score and a learning-to-rank strategy to adaptively identify the most informative spots within the tissue. Based on these identified spots, our framework selects contiguous regions as genetic anchors to ensure suitability for real-world ST profiling hardware. To effectively leverage these anchors, we design a cross-modal joint encoder that integrates visual and genetic modalities. By aligning the selected anchors with their corresponding visual features via contrastive learning, the encoder generates robust joint representations to accurately predict gene expression across the whole slide. Notably, our framework consistently surpasses existing methods in both histology-only prediction and spatial imputation, achieving superior accuracy even without genetic anchors and further excelling with as little as 10% transcriptomic coverage. Our code is available at https://github.com/Kyyle2114/CAMMST.

Keywords: Spatial Transcriptomics · Gene Expression Prediction.

1 Introduction

Spatial Transcriptomics (ST) has revolutionized tissue analysis by preserving the spatial context of gene expression patterns, thereby enabling the precise profiling of complex biological phenomena [25]. Despite this potential, the routine clinical application of ST remains constrained by high costs, the need for domain expertise, and labor-intensive workflows [13]. On the other hand, Hematoxylin and

[†] Corresponding Author

Eosin (H&E)-stained whole slide images (WSIs) are routinely acquired in clinical pathology, providing an abundant and widely accessible resource. Accordingly, numerous approaches [8, 27, 28, 23, 16, 6, 20, 14] have attempted to predict gene expression solely from these WSIs.

However, a recent framework [4] indicates that the predictive performance of existing morphology-only methods remains moderate, rendering them far from clinically transferable. This predictive bottleneck is fundamentally attributed to a lack of sufficient shared information between tissue morphology and gene expression. Considering biological phenomena such as time uncoupling [19, 4], where transcriptomic changes precede observable morphological changes, current approaches relying entirely on histological features exhibit clear limitations.

As a practical alternative, a recent study [22] has proposed a spatial imputation method that utilizes gene expression profiles from a small fraction of tissue spots as anchors to guide the prediction process, demonstrating substantial performance improvements. Nevertheless, the strategy for selecting these anchors remains to be further refined. Specifically, relying on non-adaptive sampling strategies, such as random selection, fails to guarantee optimal anchor placement and risks choosing uninformative regions that do not aid prediction. Furthermore, real-world ST profiling hardware (*e.g.*, 10× Visium, GeoMx) is optimized for capturing contiguous ROIs rather than spatially fragmented points [21]. Given these considerations, a practical sampling strategy needs to integrate adaptive, information-driven selection with the requirement of spatial continuity.

To satisfy these prerequisites, a truly practical framework should identify the most informative contiguous regions within the tissue and leverage their measured gene expression as genetic anchors to accurately predict the transcriptomic profiles of the remaining tissue. Achieving this objective requires an adaptive sampler that identifies the optimal anchor regions, alongside a cross-modal architecture capable of effectively integrating genetic information and morphological context. Notably, predicting unmeasured regions from partial measurements naturally aligns with the learning paradigm of Masked Autoencoder (MAE) [11], which infers the complete structure from a visible fraction. In this context, MAE emerges as an optimal framework for spatial imputation, effectively leveraging measured gene expression and morphological context to achieve accurate transcriptomic prediction.

To this end, we propose CAMMST (**C**ontrastive and **A**daptive **M**ulti-modal **M**AE for **ST**). Specifically, we introduce a sampler network that adaptively selects contiguous regions to serve as genetic anchors. We define a bio-saliency score that quantifies local transcriptional deviation, enabling the sampler to identify the most informative regions within the tissue. Additionally, we employ a learning-to-rank strategy to train the sampler network, ensuring robustness against cross-slide batch effects by modeling the relative relationships between tissue regions. Furthermore, we design a cross-modal joint encoder that processes histological images and genetic anchors. This encoder aligns measured gene expression with its corresponding histological features via contrastive learning. Under the MAE paradigm, these aligned representations are then leveraged to

predict whole-slide gene expression profiles. Experimental results demonstrate that CAMMST consistently surpasses existing morphology-only baselines and achieves higher accuracy by incorporating as little as 10% of the total genetic data, offering a practical and high-precision solution for clinical workflows.

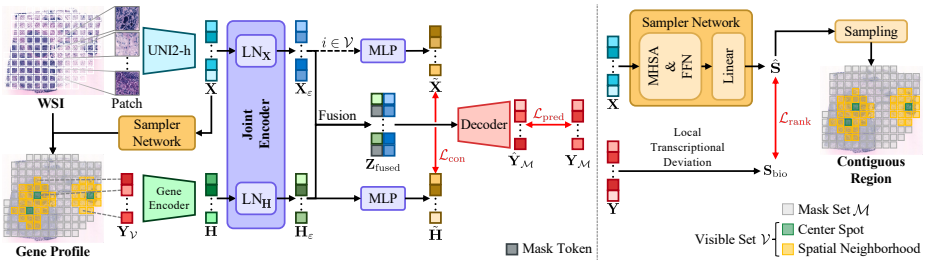

Fig. 1. CAMMST adaptively samples contiguous regions as genetic anchors via a bio-saliency guided sampler. A cross-modal joint encoder then integrates visual and genetic modalities via contrastive alignment to predict whole-slide gene expression profiles.

2 Method

The overall architecture of CAMMST is illustrated in Fig. 1. Formally, ST data is represented as a collection of N spots, where each spot i consists of a gene expression vector $\mathbf{y}_i \in \mathbb{R}^G$ (G denotes the number of genes) and an image patch $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$. These spot-level expression vectors form the complete gene expression set $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, while the corresponding image feature set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is obtained by extracting a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ from each $\mathbf{I}_i$ using UNI2-h [5]. We define the visible set $\mathcal{V}$ as the set of genetic anchors, representing a small subset of spots where gene expression is measured. Following the MAE paradigm, the model predicts the gene expression profiles $\hat{\mathbf{Y}}_{\mathcal{M}} = \{\hat{\mathbf{y}}_i\}_{i\in\mathcal{M}}$ for the masked set $\mathcal{M} = \{1, \ldots, N\} \setminus \mathcal{V}$ by leveraging the measured gene expression $\mathbf{Y}_{\mathcal{V}} = \{\mathbf{y}_i\}_{i\in\mathcal{V}}$ and the morphological context $\mathbf{X}$.

For consistency, we denote the transformer block [26] as $f(\mathbf{Z};\theta)$, where $\mathbf{Z}$ represents the input sequence and $\theta = \{\theta_{\text{attn}}, \theta_{\text{ffn}}\}$ denotes the model parameters. We employ the Attention with Linear Biases (ALiBi) [24, 7], which injects a distance-based bias into the attention operation. This allows the model to encode both local and global contexts based on spot distance, without requiring explicit positional embeddings:

$$\mathbf{Z}' = \mathbf{Z} + \text{MHSA}_{\text{ALiBi}}(\text{LN}(\mathbf{Z}); \theta_{\text{attn}}), \quad f(\mathbf{Z};\theta) = \mathbf{Z}' + \text{FFN}(\text{LN}(\mathbf{Z}'); \theta_{\text{ffn}}). \quad (1)$$

2.1 Bio-Saliency Region Sampling

Bio-Saliency Score. To identify the most informative regions to serve as genetic anchors, we first require a quantitative metric to evaluate the informativeness across the tissue. Since our framework utilizes gene expression from only a small fraction of the tissue, these selected regions should encapsulate as much diverse information as possible to guide the prediction. To this end, we quantify this informativeness based on local transcriptional deviation. By targeting areas that sharply differ from their spatial surroundings, we ensure the selected regions capture a rich transcriptomic profile. Notably, this strategy conceptually aligns with SEPAL [20], which posits that biological variability holds physiological significance. Accordingly, we define the bio-saliency score for spot i as $S_{\text{bio}}^{(i)} = D_i T_i$, combining the local deviation D_i with a noise-suppressing gating term T_i:

$$D_i = \left\| \tilde{\mathbf{y}}_i - \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \tilde{\mathbf{y}}_j \right\|_2, \quad T_i = \frac{1}{1 + \exp(-0.05 \cdot (C_i - \tau_{\text{C}}))}. \tag{2}$$

Here, $\tilde{\mathbf{y}}_i$ denotes the normalized gene expression vector, and $\mathcal{N}(i)$ represents the set of neighbor spots for spot i. The gating term T_i utilizes the total expression count $C_i = \sum_{g=1}^{G} y_{i,g}$ and a threshold τ_{C} to filter out low-signal spots. Such spots often exhibit unreliable signals dominated by technical noise; including them as genetic anchors could adversely affect the training process. To suppress these artifacts, we set τ_{C} to the 20th percentile of the total expression counts.

Contiguous Region Selection. To identify these highly informative anchors solely from histological images during inference, we design a sampler network [2] trained to predict S_{bio}. The sampler network estimates importance scores $\hat{\mathbf{s}} \in \mathbb{R}^N$ from the morphological context $\mathbf{X}$ to select contiguous regions. Formally, for a given input $\mathbf{X}$, the sampler network predicts $\hat{\mathbf{s}}$ as:

$$\mathbf{Z}_{\text{samp}} = f_{\text{samp}}(\mathbf{X}; \theta_{\text{samp}}), \quad \hat{\mathbf{s}} = \text{Linear}(\mathbf{Z}_{\text{samp}}). \tag{3}$$

Using these scores, the sampler identifies K center spots to form regions; K spots are sampled without replacement from a categorical distribution $\text{Cat}(\text{softmax}(\hat{\mathbf{s}}))$ during training, while the top-K spots with the highest scores are selected at inference. Each center spot is expanded to include its spatial neighbor spots, forming a contiguous region that adheres to the physical constraints of ST hardware. The visible set $\mathcal{V}$ is constructed as the union of these selected regions.

Scale-Aware Ranking Loss. A significant challenge in training the sampler network is the inherent batch effects across slides; the absolute scale and distribution of S_{bio} substantially differ among samples due to variations in staining intensity and experimental conditions. Consequently, directly regressing these absolute scores may bias the model toward slides with dominant intensity ranges or hinder stable convergence. To address this, we adopt a pairwise ranking strategy inspired by learning-to-rank methodologies [3]. Specifically, we introduce a scale-aware ranking loss that prioritizes the relative ordering of S_{bio} between spots over fitting their absolute magnitudes:

$$\mathcal{L}_{\text{rank}} = \sum_{(i,j) \in \mathcal{P}} |S_{\text{bio}}^{(i)} - S_{\text{bio}}^{(j)}|^{\beta} \log\left(1 + \exp\left(-(\hat{\mathbf{s}}_i - \hat{\mathbf{s}}_j)\,\text{sgn}(S_{\text{bio}}^{(i)} - S_{\text{bio}}^{(j)})\right)\right), \tag{4}$$

where $\mathcal{P}$ denotes the set of all possible spot pairs within a single slide, and β is a hyperparameter controlling the sensitivity to score differences.

2.2 Contrastive Cross-Modal Fusion

Cross-Modal MAE. To effectively integrate the genetic information and morphological context, we design a multi-stream joint encoder building upon CAV-MAE [9]. First, for the visible set $\mathcal{V}$ identified by the sampler network, the measured gene expression $\mathbf{Y}_{\mathcal{V}}$ is projected into gene features $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}|\times d}$ via an MLP-based gene encoder. These features, alongside the image features $\mathbf{X}$, are processed through independent network streams. For computational efficiency, both streams share the transformer parameters θ_{enc} to align disparate modalities into a shared latent space, while utilizing independent layer normalization to accommodate their distinct distributions:

$$\mathbf{X}_{\varepsilon} = f_{\text{img}}(\mathbf{X}; \theta_{\text{enc}}), \quad \mathbf{H}_{\varepsilon} = f_{\text{gene}}(\mathbf{H}; \theta_{\text{enc}}). \tag{5}$$

The fused sequence $\mathbf{Z}_{\text{fused}}$ is then constructed via a linear layer $\mathbf{W}_{\text{fuse}} \in \mathbb{R}^{2d\times d}$. For visible spots ($i \in \mathcal{V}$), the concatenated gene and image features are projected into a joint representation through $\mathbf{W}_{\text{fuse}}$. Similarly, for masked spots ($i \in \mathcal{M}$), a learnable mask token $\mathbf{m}_{\text{token}}$ and the corresponding image features are integrated to guide the prediction process:

$$\mathbf{z}_{\text{fused}}^{(i)} = \begin{cases} \mathbf{W}_{\text{fuse}}(\mathbf{H}_{\varepsilon,i} \oplus \mathbf{X}_{\varepsilon,i}) & \text{if } i \in \mathcal{V} \\ \mathbf{W}_{\text{fuse}}(\mathbf{m}_{\text{token}} \oplus \mathbf{X}_{\varepsilon,i}) & \text{if } i \in \mathcal{M} \end{cases}. \tag{6}$$

Finally, $\mathbf{Z}_{\text{fused}}$ is fed into the transformer decoder and an MLP head to predict the gene expression profiles $\hat{\mathbf{Y}}_{\mathcal{M}}$ for the masked set:

$$\mathbf{Z}_{\text{dec}} = f_{\text{dec}}(\mathbf{Z}_{\text{fused}}; \theta_{\text{dec}}), \quad \hat{\mathbf{y}}_i = \text{MLP}(\mathbf{z}_{\text{dec}}^{(i)}) \text{ for } i \in \mathcal{M}. \tag{7}$$

Contrastive Feature Alignment. Standard contrastive objectives based on binary supervision often introduce false conflicts by treating biologically similar semantic neighbors as negatives. Since our sampler selects contiguous regions, adjacent spots within $\mathcal{V}$ are highly likely to share similar morphological and transcriptomic profiles, which significantly exacerbates the risk of these false conflicts. To mitigate this risk, we adopt the soft-target contrastive loss from BLEEP [27]. We first project the encoded features $\mathbf{X}_{\varepsilon}$ and $\mathbf{H}_{\varepsilon}$ for the visible spots ($i \in \mathcal{V}$) into $\tilde{\mathbf{X}}, \tilde{\mathbf{H}} \in \mathbb{R}^{|\mathcal{V}|\times c}$ via an MLP head. Then, the soft target $\boldsymbol{\mathcal{T}}$ and the contrastive objective $\mathcal{L}_{\text{con}}$ are defined as:

$$\boldsymbol{\mathcal{T}} = \text{softmax}\left(\frac{\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^T}{2\tau}\right), \ \mathcal{L}_{\text{con}} = \sum_{i\in\mathcal{V}} \frac{[\text{CE}(\boldsymbol{\mathcal{T}}_i, \mathbf{p}_i) + \text{CE}(\boldsymbol{\mathcal{T}}_i, \mathbf{q}_i)]}{2|\mathcal{V}|}, \tag{8}$$

where $\mathbf{p}_i = \tilde{\mathbf{X}}_i\tilde{\mathbf{H}}^T/\tau$ and $\mathbf{q}_i = \tilde{\mathbf{H}}_i\tilde{\mathbf{X}}^T/\tau$ represent the cross-modal similarity logits for the i-th spot, and CE denotes the cross-entropy function. This soft-alignment prevents biologically similar spots from being incorrectly penalized as negatives, yielding more biologically consistent joint representations.

Table 1. Performance comparison under prediction and imputation settings. Best and second-best (prediction only) results are **bolded** and underlined, respectively. The proposed framework is highlighted in gray. $r_{\mathcal{V}}$: ratio of measured genetic anchors.

Method	$r_{\mathcal{V}}$	ST-Net			Her2ST			cSCC		
		MSE ↓	MAE ↓	PCC ↑	MSE ↓	MAE ↓	PCC ↑	MSE ↓	MAE ↓	PCC ↑
BLEEP [27]	-	0.3756	0.4736	0.0784	0.7426	0.6591	0.2747	0.6079	0.6013	0.4176
HisToGene [23]	-	0.2954	0.4269	0.1150	0.7005	0.6628	0.4284	0.6561	0.6677	0.2627
Hist2ST [28]	-	0.3811	0.4822	0.1525	0.7843	0.7286	0.2479	1.0190	0.7639	0.3003
THItoGene [16]	-	0.2925	0.4111	0.3666	0.8436	0.7069	0.3445	0.6798	0.6442	0.3897
TRIPLEX [6]	-	0.1472	0.2943	0.2320	0.8982	0.6946	0.3927	0.4891	0.5356	0.5416
MERGE [8]	-	0.1347	0.2834	0.6795	0.6422	0.6255	0.5037	0.5353	0.5838	0.5512
PH2ST [22]	0.0	0.1599	0.3063	0.6479	0.7520	0.6798	0.3701	0.6421	0.6563	0.3198
CAMMST	0.0	**0.1254**	**0.2740**	**0.6954**	**0.5769**	**0.5865**	**0.5534**	**0.4654**	**0.5350**	**0.5622**
PH2ST	0.1	0.1441	0.2759	0.6960	0.6791	0.6133	0.4735	0.5772	0.5909	0.4330
CAMMST	0.1	**0.1119**	**0.2457**	**0.7337**	**0.5098**	**0.5224**	**0.6236**	**0.4137**	**0.4763**	**0.6273**
PH2ST	0.3	0.1120	0.2147	0.7839	0.5247	0.4760	0.6375	0.4504	0.4602	0.5992
CAMMST	0.3	**0.0865**	**0.1905**	**0.8010**	**0.3876**	**0.4007**	**0.7315**	**0.3202**	**0.3686**	**0.7262**

Objective Functions. The overall training objective of CAMMST is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{9}$$

where λ_{rank} and λ_{con} are hyperparameters balancing the contribution of each loss. The prediction loss $\mathcal{L}_{\text{pred}}$ is defined as $\mathcal{L}_{\text{mse}} + \lambda_{\text{pcc}}\mathcal{L}_{\text{pcc}}$, which represents the sum of mean squared error ($\mathcal{L}_{\text{mse}}$) and Pearson correlation coefficient loss ($\mathcal{L}_{\text{pcc}}$) weighted by λ_{pcc}, both computed based on the predicted gene expression profiles for the masked spots $\hat{\mathbf{Y}}_{\mathcal{M}}$.

3 Experiments

3.1 Experimental Setup

Datasets. Experiments are conducted on three datasets: ST-Net [10], Her2ST [1], and cSCC [15]. To ensure a fair comparison, we strictly follow the experimental protocols established by MERGE [8], utilizing 8-fold cross-validation with SPCS smoothing [18], and targeting the top-250 highly expressed genes for prediction.
Baselines and Evaluation Metrics. We benchmark CAMMST against several state-of-the-art baselines, including image-only prediction models and the spatial imputation framework, PH2ST [22]. For a fair comparison, PH2ST is implemented using the same UNI2-h backbone as our framework. Performance is evaluated using three standard metrics: Pearson Correlation Coefficient (PCC), Mean Squared Error (MSE), and Mean Absolute Error (MAE).
Evaluation Settings. Performance is evaluated under two scenarios representing different levels of data availability: prediction and imputation. While the prediction setting relies solely on WSIs, the imputation setting incorporates a

fraction of gene expression data as spatial anchors. For the latter, predicted values in measured regions are replaced with ground truth to calculate slide-wide metrics, reflecting the overall prediction performance across the entire tissue.

Implementation Details. Our framework is trained using AdamW optimizer (lr $= 2e-4$, weight decay$=5e-2$). During training, 10% of the total spots are selected as visible spots by identifying $K = 3$ regions, with each region containing an approximately equal number of spots. Hyperparameters are assigned as $\beta = 1.5$, $\lambda_{\text{rank}} = 0.001$, $\lambda_{\text{con}} = 0.05$, $\lambda_{\text{pcc}} = 0.5$, and $\tau = 1.0$. Experiments are conducted on a single NVIDIA RTX A6000 GPU using the PyTorch framework.

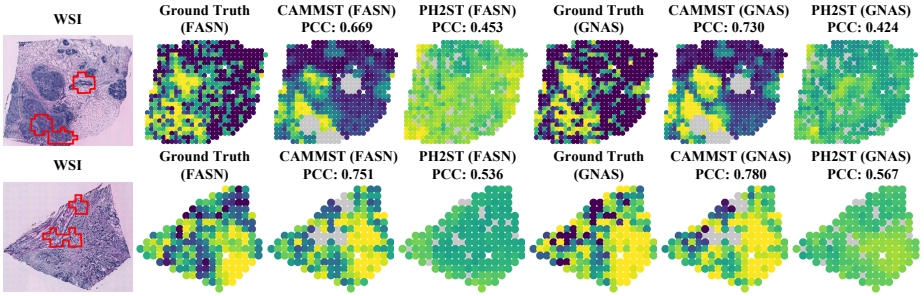

Fig. 2. Qualitative evaluation of gene expression prediction and sampling strategy. Red boundaries on the WSI indicate the contiguous regions selected by our sampler network. Spots in gray denote the corresponding visible spots.

3.2 Main Results

Quantitative Results. Table 1 presents the comparative performance of our framework against state-of-the-art methods in both prediction and imputation settings. In the imputation setting, our framework outperforms the performance of PH2ST; for instance, on the cSCC dataset at $r_{\mathcal{V}} = 0.1$, CAMMST achieves a PCC of 0.6273 compared to 0.4330 for the baseline. These gains indicate that our framework effectively leverages contiguous genetic anchors to predict precise transcriptomic profiles. Notably, even in the prediction setting, CAMMST consistently surpasses all image-only baselines, establishing a robust morphology-to-gene mapping even without the aid of genetic anchors.

Qualitative Results. Fig. 2 presents a qualitative comparison of expression predictions for the tumor biomarker FASN [12] and the breast cancer biomarker GNAS [17]. While PH2ST produces over-smoothed patterns that fail to capture local variations, CAMMST successfully recovers fine-grained gene expression that closely resembles the ground truth. The WSI overlay illustrates how our sampler targets contiguous regions (red boundaries). By utilizing these visible spots (gray spots), our framework effectively guides the prediction process.

Table 2. Comparison of computational efficiency on the Her2ST dataset. Both models utilize pre-computed histological features extracted from the UNI2-h backbone.

Metric	PH2ST	CAMMST (Ours)
Model Parameters (M)	290.73	44.59
GFLOPs (per slide)	2386.21 ± 996.39	33.45 ± 13.94
Training Time (per fold, hours)	7.91 ± 0.80	0.19 ± 0.03
Training GPU Memory (GB)	10.54 ± 0.59	1.04 ± 0.01
Inference Time (per slide, sec)	0.946 ± 0.443	0.021 ± 0.001
Inference GPU Memory (GB)	5.93 ± 2.14	0.21 ± 0.02

Table 3. Ablation study of each component on the Her2ST dataset.

Method	MSE ↓	MAE ↓	PCC ↑
Bio-Saliency Region Sampling			
w/ Random Sampling	0.5205±0.1067	0.5261±0.0647	0.6106±0.0537
w/ Regression Loss (w/o $\mathcal{L}_{\text{rank}}$)	0.5117±0.1035	0.5225±0.0641	0.6197±0.0525
w/ Scattered Sampling	0.5204±0.1098	0.5251±0.0705	0.6119±0.0515
Contrastive Cross-Modal Fusion			
w/o Contrastive Learning	0.5150±0.1075	0.5246±0.0642	0.6171±0.0523
w/ Standard Contrastive Loss	0.5118±0.1057	0.5223±0.0697	0.6212±0.0521
Ours	0.5098±0.1040	0.5224±0.0703	0.6236±0.0508

Computational Efficiency. Table 2 presents a comparison of computational efficiency. Compared to the baseline, our framework substantially reduces resource consumption across both training and inference. This advantage stems from the MAE-based architecture, which provides an optimal framework for the spatial imputation task. Consequently, CAMMST achieves superior prediction performance and operates effectively without high-end hardware, providing a practical and high-precision solution for resource-constrained clinical environments.

3.3 Ablation Study

Table 3 presents the ablation study results evaluating the impact of each module. The sampler network, optimized with the proposed scale-aware ranking loss ($\mathcal{L}_{\text{rank}}$), yields superior performance compared to random selection or regression-based S_{bio} prediction. This improvement demonstrates that the ranking-based strategy identifies more effective genetic anchors while ensuring robustness against batch effects. A scattered sampling variant optimized identically but without regional constraints results in a performance decline, demonstrating that identifying informative points alone is insufficient. This degradation suggests that the value of informational diversity is inherently tied to the surrounding spatial context. By selecting contiguous regions, our framework captures the essential signals for accurate transcriptomic prediction. For cross-modal alignment, the soft-target contrastive loss outperforms standard or absent objectives. These re-

sults demonstrate that our approach effectively integrates multi-modal features by mitigating false conflicts between biologically related spots.

4 Conclusion

In this study, we proposed CAMMST, a novel framework for gene expression prediction. Built upon the MAE paradigm, our model adaptively selects the most informative regions within the tissue and leverages measured gene expression to accurately predict whole-slide transcriptomic profiles. Beyond predictive performance, our framework is explicitly designed for practical efficiency, accommodating the physical constraints of real-world ST hardware and limited computing resources. Experimental results demonstrate that CAMMST consistently outperforms all existing baselines, providing a highly efficient and accurate solution that establishes a practical foundation for routine clinical workflows.

Acknowledgments. This work was supported in part by the IITP RS-2024-00457882 (AI Research Hub Project), IITP 2020-II201361, NRF RS-2024-00345806, NRF RS-2023-002620, and RQT-25-120390.

Disclosure of Interests. The authors have no competing interests.

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. Bandara, W.G.C., Patel, N., Gholami, A., Nikkhah, M., Agrawal, M., Patel, V.M.: Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14507–14517 (2023)
3. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on Machine learning. pp. 89–96 (2005)
4. Chelebian, E., Avenel, C., Wählby, C.: Combining spatial transcriptomics with tissue morphology. Nature Communications **16**(1), 4452 (2025)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
6. 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)
7. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: A multimodal whole-slide foundation model for pathology. Nature medicine pp. 1–13 (2025)

8. Ganguly, A., Chatterjee, D., Huang, W., Zhang, J., Yurovsky, A., Johnson, T.S., Chen, C.: Merge: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15611–15620 (2025)
9. Gong, Y., Rouditchenko, A., Liu, A.H., Harwath, D., Karlinsky, L., Kuehne, H., Glass, J.R.: Contrastive audio-visual masked autoencoder. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=QPtMRyk5rb
10. He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., Zou, J.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nature biomedical engineering **4**(8), 827–834 (2020)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
12. Hunt, D.A., Lane, H.M., Zygmont, M.E., Dervan, P.A., Hennigar, R.A.: Mrna stability and overexpression of fatty acid synthase in human breast cancer cell lines. Anticancer research **27**(1A), 27–34 (2007)
13. Jain, S., Eadon, M.T.: Spatial transcriptomics in health and disease. Nature reviews nephrology **20**(10), 659–671 (2024)
14. Jeong, T., Kim, J., Kim, J., Kim, C., Hwang, S.J.: Feast: Fully connected expressive attention for spatial transcriptomics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26793–26802 (2026)
15. 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)
16. 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) (2023)
17. Jin, X., Zhu, L., Cui, Z., Tang, J., Xie, M., Ren, G.: Elevated expression of gnas promotes breast cancer cell proliferation and migration via the pi3k/akt/snail1/e-cadherin axis. Clinical and Translational Oncology **21**(9), 1207–1219 (2019)
18. Liu, Y., Wang, T., Duggan, B., Sharpnack, M., Huang, K., Zhang, J., Ye, X., Johnson, T.S.: Spcs: a spatial and pattern combined smoothing method for spatial transcriptomic expression. Briefings in Bioinformatics **23**(3) (2022)
19. Lomakin, A., Svedlund, J., Strell, C., Gataric, M., Shmatko, A., Rukhovich, G., Park, J.S., Ju, Y.S., Dentro, S., Kleshchevnikov, V., et al.: Spatial genomics maps the structure, nature and evolution of cancer clones. Nature **611**(7936), 594–602 (2022)
20. Mejia, G., Cárdenas, P., Ruiz, D., Castillo, A., Arbeláez, P.: Sepal: spatial gene expression prediction from local graphs. In: Proceedings of the IEEE/CVF International Conference on computer vision. pp. 2294–2303 (2023)
21. Moffitt, J.R., Lundberg, E., Heyn, H.: The emerging landscape of spatial profiling technologies. Nature Reviews Genetics **23**(12), 741–759 (2022)
22. Niu, Y., Liu, J., Zhan, Y., Shi, J., Zhang, D., Reinius, M., Machado, I., Crispin-Ortuzar, M., Wu, J., Li, C., et al.: Ph2st: Prompt-guided hypergraph learning for spatial transcriptomics prediction in whole slide images. Medical Image Analysis p. 104008 (2026)

23. 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)
24. Press, O., Smith, N., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=R8sQPpGCv0
25. Rao, A., Barkley, D., França, G.S., Yanai, I.: Exploring tissue architecture using spatial transcriptomics. Nature **596**(7871), 211–220 (2021)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
27. Xie, R., Pang, K., Chung, S., Perciani, C., MacParland, S., Wang, B., Bader, G.: Spatially resolved gene expression prediction from histology images via bimodal contrastive learning. Advances in Neural Information Processing Systems **36**, 70626–70637 (2023)
28. Zeng, Y., Wei, Z., Yu, W., Yin, R., Yuan, Y., Li, B., Tang, Z., Lu, Y., Yang, Y.: Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. Briefings in Bioinformatics **23**(5) (2022)