GASP: GENE-AWARE SPATIAL PREDICTOR FOR HISTOLOGY-BASED TRANSCRIPTOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics (ST) provides unprecedented insights into tissue architecture but remains constrained by high costs and complex protocols. Predicting ST directly from ubiquitous Hematoxylin and Eosin (H&E) stained images offers a scalable alternative. However, existing computational methods predominantly adopt a "vision-centric" paradigm, projecting passively extracted visual features into the gene space via a shared decoder. This approach overlooks the inherent heterogeneity of genes, failing to capture their unique morphological dependencies. In this paper, we propose the Gene-Aware Spatial Predictor (GASP), a novel framework that introduces a paradigm shift towards "gene-centric" modeling. GASP employs learnable Gene Queries that actively probe a multi-scale visual context pool—comprising local, neighborhood, and global features—via a Vision-Gene Cross-Attention mechanism. This allows each gene to adaptively retrieve its preferred morphological cues. Furthermore, a Gene-Specific Residual Decoding strategy is introduced to balance global expression trends with gene-specific variations. Extensive experiments on benchmark datasets demonstrate that GASP outperforms state-of-the-art methods, offering superior predictive accuracy for spatial gene expression modeling.

1 INTRODUCTION

Spatial transcriptomics (ST) has emerged as a transformative technology that bridges the gap between gene expression and tissue morphology, offering near-cellular resolution of molecular activity within its native spatial context (Stahl et al., 2016; Burgess, 2019; Asp et al., 2020). By mapping transcriptomes to precise histological coordinates, ST enables researchers to unravel spatial heterogeneity, complex tissue architecture, and intricate cell-cell interactions across diverse biological systems and diseases (Ji et al., 2020; Maniatis et al., 2018; Asp et al., 2019; Wu et al., 2021). However, despite its immense potential, the widespread clinical adoption of ST is significantly hindered by practical constraints. The acquisition of ST data requires specialized instrumentation, complex experimental protocols, and incurs prohibitive costs (Zhu et al., 2025; Ruiz et al., 2025; Wang et al., 2018). Consequently, large-scale diagnostic or population-level applications remain scarce.

In stark contrast, Hematoxylin and Eosin (H&E)-stained whole-slide images (WSIs) are routinely acquired in clinical pathology workflows, are highly cost-effective, and encapsulate a wealth of morphological features that are closely associated with underlying gene expression patterns (Campanella et al., 2019; Lu et al., 2021; Ash et al., 2021). This complementary relationship has motivated a rapidly growing body of research focused on developing computational models, particularly deep learning-based approaches, to infer spatially resolved transcriptomic profiles directly from H&E WSIs (He et al., 2020; Pang et al., 2021; Jia et al., 2023).

Existing methods typically formulate this challenge as a multi-output regression task, mapping spot-aligned image patches to their corresponding gene expression vectors. Early pioneering works, such as ST-Net (He et al., 2020), utilized Convolutional Neural Networks (CNNs) to extract features from isolated spot images. Subsequent advancements have sought to incorporate broader spatial context and long-range dependencies. For instance, HisToGene (Pang et al., 2021) employs Vision Transformers (ViTs) to model inter-spot correlations, while Hist2ST (Zeng et al., 2022) combines ConvMixer and Graph Neural Networks (GNNs) to aggregate neighborhood information. Other approaches, such as EGN (Yang et al., 2023) and BLEEP (Xie et al., 2023), leverage exemplar-guided

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

learning and bi-modal contrastive embeddings, respectively. More recently, models like TRIPLEX (Chung et al., 2024) and HiFusion (Weng et al., 2026) have demonstrated the importance of integrating multi-resolution features—capturing cellular morphology at individual spots, local regional context, and global tissue organization—to enhance prediction accuracy. Furthermore, ASIGN (Zhu et al., 2025) extends this paradigm to 3D ST imputation by aligning adjacent tissue sections.

Despite these significant strides, current methodologies share a fundamental limitation: they predominantly adopt a "vision-centric" paradigm. In these frameworks, visual features are passively extracted from the image and subsequently projected into the gene expression space via a shared, global decoder. This approach treats all genes uniformly, failing to account for the inherent biological heterogeneity and complex relationships among different genes (Taylor et al., 2023; Wolf et al., 2022). Specifically, existing models overlook the fact that different genes exhibit unique morphological preferences; the expression of one gene might be strongly dictated by local nuclear texture, while another might be driven by the broader tissue microenvironment or immune infiltration patterns (Chen et al., 2024a; Zhao et al., 2024). By uniformly fusing multi-scale image features for all genes, current methods dilute these gene-specific signals, leading to suboptimal predictions. There is a critical lack of a mechanism that allows individual genes to actively probe the visual features for their specific morphological cues.

To address these critical bottlenecks, we propose the **Gene-Aware Spatial Predictor (GASP)**, a novel deep learning framework that introduces a paradigm shift from passive visual feature extraction to active, "gene-centric" modeling. GASP explicitly acknowledges gene heterogeneity by introducing learnable Gene Queries. Instead of relying on a shared decoder, each gene query actively interacts with a multi-scale visual context pool (comprising target, neighbor, and global features) via a Vision-Gene Cross-Attention mechanism. This allows each gene to adaptively aggregate the specific spatial and morphological scales most relevant to its expression. Furthermore, we design a Gene-Specific Residual Decoding module that balances global expression trends with gene-specific variations, ensuring robust and highly accurate predictions. This active querying mechanism enables GASP to effectively disentangle the complex, heterogeneous relationships between tissue morphology and spatial gene expression.

Our key contributions can be summarized as follows:

- We introduce a novel "gene-centric" paradigm for spatial transcriptomics prediction, shifting the focus from traditional passive visual feature extraction to active, gene-driven morphological probing.
- We propose the Gene-Aware Spatial Predictor (GASP), which utilizes learnable Gene Queries and a Vision-Gene Cross-Attention mechanism, underpinned by a Cross-View Fusion Encoder that mutually refines multi-scale visual streams to provide a context-enriched retrieval pool. This design empowers each gene to adaptively retrieve its preferred morphological cues from multi-scale visual features, effectively addressing the challenge of gene heterogeneity.
- We develop a Gene-Specific Residual Decoding strategy that preserves the unique expression distribution of individual genes while maintaining the stability of global predictions, significantly improving the modeling of complex spatial patterns.
- Extensive experiments on multiple benchmark datasets demonstrate that GASP achieves state-of-the-art performance, outperforming existing vision-centric methods and providing more biologically meaningful insights into the spatial organization of tissues.

2 Related Work

2.1 Spatial Transcriptomics Prediction from Histology

The prediction of spatial gene expression directly from H&E-stained WSIs has garnered significant attention due to its potential to democratize spatial transcriptomics (Zhang et al., 2022; Chen & Huang, 2025). Early approaches primarily relied on CNNs to extract features from localized image patches. ST-Net (He et al., 2020) pioneered this direction by fine-tuning a pre-trained DenseNet (Huang et al., 2016) to predict expression levels from isolated spot images. To capture the spatial dependencies inherent in tissue architecture, subsequent methods integrated advanced architectural

designs. HisToGene (Pang et al., 2021) utilized Vision Transformers (Dosovitskiy et al., 2020) to model long-range correlations across spots within a WSI. Hist2ST (Zeng et al., 2022) further advanced this by combining ConvMixer (Trockman & Kolter, 2022) for local feature extraction with Graph Convolutional Networks (Kipf & Welling, 2016) to aggregate neighborhood context.

Recognizing the complex hierarchical structures of histopathological images, recent models have increasingly adopted multi-resolution and context-aware strategies. TRIPLEX (Chung et al., 2024) demonstrated that integrating features from three distinct resolutions—the target spot, the local neighbor view, and the global WSI view—significantly outperforms single-resolution models. Similarly, HiFusion (Weng et al., 2026) proposed a hierarchical intra-spot modeling approach combined with context-aware cross-scale fusion. Beyond 2D context, ASIGN (Zhu et al., 2025) introduced a novel framework for 3D ST imputation by aligning adjacent tissue sections. Alternative strategies have explored cross-modal alignment. EGN (Yang et al., 2023) proposed an exemplar-guided network that dynamically selects similar reference spots, while BLEEP (Xie et al., 2023) and ST-Align (Lin et al., 2024) introduced bi-modal embedding frameworks to align image and gene expression modalities via contrastive learning. Despite these advancements, these methods universally employ a shared decoding mechanism that maps visual features to the gene space, thereby neglecting the specific morphological dependencies of individual genes.

2.2 Pathology Foundation Models and Representation Learning

The success of deep learning in computational pathology heavily relies on the quality of visual representations. Historically, models utilized weights pre-trained on natural images (e.g., ImageNet), which often exhibit a domain gap when applied to histology (Campanella et al., 2019). Recently, the paradigm has shifted towards domain-specific self-supervised learning (SSL) on massive cohorts of unannotated histopathological images. For instance, Ciga et al. (Ciga et al., 2022) demonstrated that contrastive learning on large-scale histology patches yields robust representations that are highly transferable to various downstream tasks. Building upon this, pathology foundation models like UNI (Chen et al., 2024b) have scaled up SSL using Vision Transformers (ViT-L) trained on millions of histology patches, capturing rich patch-level pathological representations transferable to diverse downstream tasks.

While these pre-trained models (e.g., CIGAR ResNet-18 and UNI) provide powerful, multi-scale visual features, how to effectively query and fuse them for specific downstream targets remains an open challenge. Existing ST prediction models typically concatenate or average these features uniformly for all genes. In contrast, our GASP framework leverages these robust foundation representations but introduces a "gene-centric" querying mechanism. By assigning a learnable query to each target gene, GASP explicitly models gene heterogeneity, enabling a fine-grained, gene-specific cross-modal alignment that actively retrieves relevant morphological features from the pre-trained visual context pool.

3 Methodology

3.1 Problem Formulation and Assumptions

Problem Formulation. The objective of spatial transcriptomics (ST) prediction is to map a high-resolution Hematoxylin and Eosin (H&E) stained whole-slide image (WSI) to its corresponding spatially resolved gene expression profile. Let $\mathcal{I}$ denote an H&E WSI and $\mathcal{S} = \{\mathbf{s}_i\}_{i=1}^{N}$ be the spatial coordinates of N sequencing spots, where $\mathbf{s}_i \in \mathbb{R}^2$. Based on $\mathcal{S}$, we extract a set of image patches $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ centered at each spot.

Let $\mathbf{Y} \in \mathbb{R}^{N \times G}$ denote the ground truth gene expression matrix, where G is the number of target genes. The i-th row $\mathbf{y}_i \in \mathbb{R}^G$ represents the expression profile of spot i. Our goal is to learn a mapping function $\mathcal{F} : (\mathcal{I}, \mathcal{S}) \rightarrow \hat{\mathbf{Y}}$, where $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times G}$ is the predicted expression matrix.

Overview of GASP. As shown in Figure 1, GASP consists of four components: (1) Histological Feature Extraction; (2) Gene-Aware Contextual Representation; (3) Gene-Specific Residual Decoding; and (4) Joint Optimization.

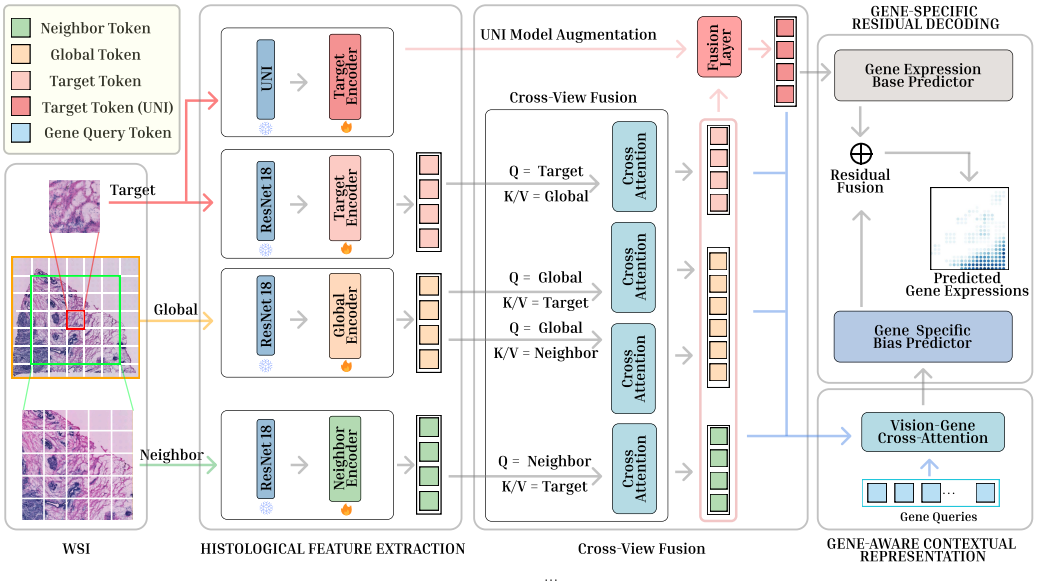

Figure 1: Overview of GASP.

3.2 Histological Feature Extraction

GASP extracts three complementary visual representations for each spot i using a shared frozen ResNet-18 backbone pre-trained on histology images via self-supervised contrastive learning (Ciga et al., 2022).

(1) Target Patch Tokens. The 224×224 target patch $\mathbf{x}_i$ is processed online through the frozen ResNet-18, and its 7×7 spatial feature map is flattened and projected to yield $\mathbf{V}^{(i)}_{\text{target}} \in \mathbb{R}^{49\times D}$, preserving fine-grained local cellular morphology. Online processing enables data augmentation during training.

(2) Neighbor Context Tokens. The patches of the $K{=}25$ spatially adjacent spots ($5{\times}5$ grid) are pre-encoded offline with the same frozen backbone. A trainable 3-layer Transformer encoder aggregates these embeddings into $\mathbf{V}^{(i)}_{\text{neighbor}} \in \mathbb{R}^{25\times D}$, capturing local tissue microenvironment context. Invalid positions at slide boundaries are handled via a padding mask.

(3) Global Context Vector. All spot patches within a WSI are pre-encoded offline. A trainable 3-layer Transformer encoder with an Atypical Position Encoding Generator (APEG) (Chung et al., 2024) processes these slide-wide embeddings, producing a per-spot global context vector $\mathbf{v}^{(i)}_{\text{global}} \in \mathbb{R}^D$ that encodes macro-level tissue architecture.

Cross-View Fusion. A trainable Cross-View Fusion Encoder refines all three representations jointly via bidirectional cross-attention over 3 layers: global features are injected into target and neighbor tokens, while target and neighbor features are fed back to update the global representation. The three streams are then merged via a learned gate into a single fused vector $\mathbf{v}^{(i)}_{\text{fusion}} \in \mathbb{R}^D$. Concretely, a linear layer maps the mean-pooled token of each stream to a scalar logit; the three logits are normalized by softmax to produce weights $(\alpha_t, \alpha_n, \alpha_g)$, and the fused vector is their weighted sum:

$$\mathbf{v}^{(i)}_{\text{fusion}} = \alpha_t \cdot \bar{\mathbf{V}}^{(i)}_{\text{target}} + \alpha_n \cdot \bar{\mathbf{V}}^{(i)}_{\text{neighbor}} + \alpha_g \cdot \tilde{\mathbf{v}}^{(i)}_{\text{global}}, \quad (\alpha_t, \alpha_n, \alpha_g) = \text{softmax}\left(\mathbf{W}_g \begin{bmatrix} \bar{\mathbf{V}}^{(i)}_{\text{target}} \\ \bar{\mathbf{V}}^{(i)}_{\text{neighbor}} \\ \tilde{\mathbf{v}}^{(i)}_{\text{global}} \end{bmatrix} \right) \quad (1)$$

where $\bar{\mathbf{V}}$ denotes mean-pooling over the token sequence and $\mathbf{W}_g \in \mathbb{R}^{3\times D}$ is a learned projection. Crucially, this module outputs not only $\mathbf{v}^{(i)}_{\text{fusion}}$ but also the *updated* token sequences $\tilde{\mathbf{V}}^{(i)}_{\text{target}}$, $\tilde{\mathbf{V}}^{(i)}_{\text{neighbor}}$, and $\tilde{\mathbf{v}}^{(i)}_{\text{global}}$, which are later used as the context pool for gene queries.

Foundation Model Augmentation. To further enrich the representation, the target patch $\mathbf{x}_i$ is additionally encoded offline by the frozen UNI foundation model (Chen et al., 2024b) (ViT-L), yielding a per-spot CLS token $\mathbf{u}^{(i)} \in \mathbb{R}^{1024}$ with richer pathological semantics. This is projected to $\mathbb{R}^D$ and added to the fused vector with a learnable scalar weight w:

$$\tilde{\mathbf{v}}^{(i)}_{\text{fusion}} = \mathbf{v}^{(i)}_{\text{fusion}} + w \cdot \text{MLP}(\mathbf{u}^{(i)}) \tag{2}$$

The resulting $\tilde{\mathbf{v}}^{(i)}_{\text{fusion}} \in \mathbb{R}^D$ consolidates CIGAR and UNI features into a single representation and is used for both base prediction and as part of the gene query context pool.

3.3 Gene-Aware Contextual Representation

Learnable Gene Queries. The core innovation of GASP lies in shifting from passive feature reception to active, gene-driven feature extraction. We introduce a learnable gene query matrix $\mathbf{Q} \in \mathbb{R}^{G\times D}$, where the g-th row $\mathbf{q}_g \in \mathbb{R}^D$ corresponds to the specific query for gene g. These queries are optimized during training to encapsulate the unique morphological preferences of each gene.

Vision-Gene Cross-Attention. To enable each gene to extract its desired morphological cues, we construct a visual context pool $\mathbf{C}^{(i)} \in \mathbb{R}^{L\times D}$ for spot i by concatenating the multi-scale visual tokens:

$$\mathbf{C}^{(i)} = \left[\tilde{\mathbf{V}}^{(i)}_{\text{target}} \parallel \tilde{\mathbf{V}}^{(i)}_{\text{neighbor}} \parallel (\tilde{\mathbf{v}}^{(i)}_{\text{global}})^T \parallel (\tilde{\mathbf{v}}^{(i)}_{\text{fusion}})^T\right] \tag{3}$$

where $\parallel$ denotes concatenation along the sequence dimension, and $L = 49 + 25 + 1 + 1 = 76$. Note that all four sources are the *fusion-updated* representations output by the Cross-View Fusion Encoder, ensuring that each gene query attends to mutually-informed, context-enriched features.

The gene queries $\mathbf{Q}$ act as the query matrix, while $\mathbf{C}^{(i)}$ serves as both the key and value matrices in a multi-head cross-attention block:

$$\mathbf{H}^{(i)} = \text{Softmax}\left(\frac{\mathbf{Q}(\mathbf{C}^{(i)})^T}{\sqrt{D}}\right)\mathbf{C}^{(i)} \tag{4}$$

where $\mathbf{H}^{(i)} \in \mathbb{R}^{G\times D}$ represents the gene-aware contextual features for spot i. The g-th row $\mathbf{h}^{(i)}_g \in \mathbb{R}^D$ is the specific visual feature actively retrieved by gene g.

3.4 Gene-Specific Residual Decoding

Instead of relying solely on a shared linear layer to decode visual features, GASP employs a residual decoding strategy to balance global expression trends with gene-specific variations.

Base Prediction. A shared base decoder generates a preliminary expression profile $\mathbf{y}^{(i)}_{\text{base}} \in \mathbb{R}^G$ directly from the fused visual vector:

$$\mathbf{y}^{(i)}_{\text{base}} = \mathbf{W}_{\text{base}}\tilde{\mathbf{v}}^{(i)}_{\text{fusion}} \tag{5}$$

where $\mathbf{W}_{\text{base}} \in \mathbb{R}^{G\times D}$ is a learnable weight matrix.

Gene-Specific Bias. Simultaneously, each gene's contextual feature $\mathbf{h}^{(i)}_g$ is projected into a scalar bias $b^{(i)}_g$ via an independent readout layer: $b^{(i)}_g = \mathbf{w}^T_{\text{readout}}\mathbf{h}^{(i)}_g$, forming a bias vector $\mathbf{b}^{(i)} = [b^{(i)}_1, \ldots, b^{(i)}_G]^T \in \mathbb{R}^G$.

Residual Fusion. The final predicted expression $\hat{\mathbf{y}}_i \in \mathbb{R}^G$ for spot i is obtained by adding the gene-specific bias to the base prediction:

$$\hat{\mathbf{y}}_i = \mathbf{y}^{(i)}_{\text{base}} + \lambda\mathbf{b}^{(i)} \tag{6}$$

where λ is a scaling hyperparameter (set to 0.15 in our implementation). This residual design ensures stable training while granting each gene the flexibility to fine-tune its expression based on its unique morphological queries.

3.5 Objective Function and Optimization

To optimize GASP, we employ a joint loss function that simultaneously constrains the absolute expression values and the spatial expression patterns across a batch of N spots. The primary loss $\mathcal{L}_{\text{pred}}$ combines the gene-wise Pearson Correlation Coefficient (PCC) loss across spots and the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{PCC}}(\hat{\mathbf{Y}}, \mathbf{Y}) + \gamma \mathcal{L}_{\text{MSE}}(\hat{\mathbf{Y}}, \mathbf{Y}) \tag{7}$$

where γ balances the two terms (set to 0.15). The PCC loss explicitly encourages the model to capture the correct spatial distribution of each gene across the tissue. It is defined as:

$$\mathcal{L}_{\text{PCC}}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{G} \sum_{g=1}^{G} \left(1 - \frac{\sum_{i=1}^{N} (\hat{Y}_{i,g} - \bar{\hat{Y}}_{\cdot,g})(Y_{i,g} - \bar{Y}_{\cdot,g})}{\sqrt{\sum_{i=1}^{N} (\hat{Y}_{i,g} - \bar{\hat{Y}}_{\cdot,g})^2} \sqrt{\sum_{i=1}^{N} (Y_{i,g} - \bar{Y}_{\cdot,g})^2}} \right) \tag{8}$$

where $\hat{Y}_{i,g}$ and $Y_{i,g}$ are the predicted and true expressions of gene g at spot i, and $\bar{\hat{Y}}_{\cdot,g}$ and $\bar{Y}_{\cdot,g}$ are their respective means across the N spots.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate GASP on three publicly available spatial transcriptomics benchmark datasets: two breast cancer datasets (BC1 and BC2) and one squamous cell carcinoma dataset (SCC).

- **BC1** comprises 36 early-stage HER2+ breast cancer tissue sections from 8 patients, totaling 13,620 spots. We predict the expression of 250 highly variable gene symbols.
- **BC2** includes 68 breast cancer tissue sections from 23 patients, containing 68,050 spots, targeting 250 Ensembl IDs.
- **SCC** consists of 12 squamous cell carcinoma sections from 4 patients, with 23,205 spots and 250 gene symbols.

To rigorously assess generalization to unseen patients, we employ patient-level cross-validation across all datasets. For BC1 (8 patients) and SCC (4 patients), we apply strict leave-one-patient-out cross-validation (LOPOCV), yielding 8 and 4 folds respectively. For BC2, which contains 23 patients, we apply 8-fold patient-grouped cross-validation, where patients are partitioned into 8 groups with no patient appearing in both training and test sets within any fold, ensuring the same patient-level generalization guarantee.

Evaluation Metrics. Following previous works (He et al., 2020; Chung et al., 2024), we use the mean Pearson Correlation Coefficient across all target genes, denoted as **PCC(M)**, as our primary evaluation metric. We also report **PCC(H)**, the mean PCC for the top-50 highly variable genes selected per fold based on expression variance in the training set. Additionally, we measure the Mean Squared Error (**MSE**) and Mean Absolute Error (**MAE**) to evaluate the absolute prediction accuracy.

Implementation details are provided in Appendix A.

4.2 Motivation Analysis: Why Gene-Centric Modeling?

To empirically justify the design of GASP, we conduct an analysis of gene expression behavior on the BC1 dataset. As shown in Figure 2, two key observations motivate our architecture.

Gene expression is highly heterogeneous. As shown in Table 1, expression variance across the 250 target genes varies substantially (mean 2.417, std 0.618), and Moran's I ranges from 0.012 to 0.231, confirming that genes exhibit vastly different spatial structures. We also observe a positive Spearman correlation ($\rho = +0.368$) between gene variance and patient specificity, suggesting that highly variable genes are more sensitive to individual differences. This motivates our **Gene-Specific Residual Decoding** module, which assigns a unique bias term to each gene.

Table 1: Gene expression variance statistics on BC1 (250 genes, log-normalized).

Statistic	Value
Mean variance across 250 genes	2.417
Std of variance across genes	0.618
High-variance genes (>mean+1σ)	22 genes (8.8%)
Low-variance genes (<mean-1σ)	26 genes (10.4%)
Mean Moran's I (spatial autocorrelation)	0.087 ± 0.041
Moran's I range	[0.012, 0.231]
ρ(gene variance, patient specificity)	+0.368

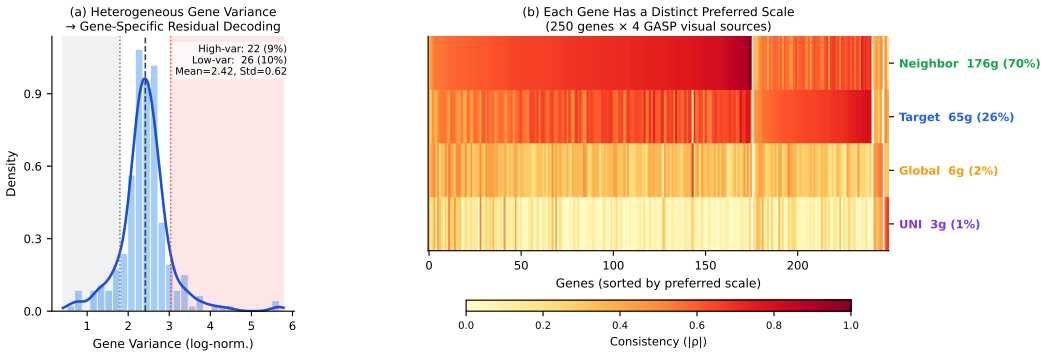

Figure 2: Empirical motivation for GASP's gene-centric design on BC1 (250 genes). **(a)** Gene expression variance is highly heterogeneous; high-variance genes correlate with patient specificity ($\rho = +0.368$), motivating Gene-Specific Residual Decoding. **(b)** Genes exhibit heterogeneous visual scale preferences across local, neighborhood, and global features, motivating Vision-Gene Cross-Attention for adaptive, per-gene scale retrieval.

Table 2: Cross-validation results on three ST benchmark datasets. Bold indicates the best performance.

Model	BC1				BC2				SCC			
	PCC(M)↑	PCC(H)↑	MSE↓	MAE↓	PCC(M)↑	PCC(H)↑	MSE↓	MAE↓	PCC(M)↑	PCC(H)↑	MSE↓	MAE↓
ST-Net	0.1581	0.2912	1.5847	0.9804	0.0995	0.1943	1.1744	0.8340	0.1288	0.2180	2.2904	1.1970
HisToGene	0.1872	0.3288	1.3308	0.8846	0.1110	0.2106	1.2950	0.9061	0.1652	0.2398	1.5525	0.9909
BLEEP	0.0063	0.0970	2.5810	1.2861	-0.0360	0.0317	2.0802	1.1561	-0.0063	0.0652	3.0223	1.4677
TRIPLEX	0.2174	0.4145	1.4299	0.9122	0.1522	0.3679	1.2879	0.8736	0.1822	0.3430	1.6819	1.0098
HiFusion	0.2032	0.4292	1.4736	0.9322	0.1078	0.3218	1.2896	0.8700	0.1882	0.3072	1.6920	1.0124
GASP (Ours)	**0.2946**	**0.5425**	**1.2509**	**0.8692**	**0.1612**	**0.3835**	**1.1712**	**0.8311**	**0.2489**	**0.4049**	**1.5443**	**0.9852**

Genes differ in their preferred visual scale. We find that genes vary in how much they rely on local patch features versus broader neighborhood or global context. This heterogeneity in scale preference suggests that a fixed-weight fusion strategy, as used in prior works, may not be optimal for all genes, motivating our **Vision-Gene Cross-Attention** mechanism for adaptive, per-gene feature retrieval.

4.3 Spatial Gene Expression Prediction

We evaluate the predictive performance of GASP against five state-of-the-art vision-centric ST prediction models: ST-Net (He et al., 2020), HisToGene (Pang et al., 2021), BLEEP (Xie et al., 2023), TRIPLEX (Chung et al., 2024), and HiFusion (Weng et al., 2026).

As shown in Table 2, GASP consistently and significantly outperforms all baseline methods across all three datasets and evaluation metrics. On the BC1 dataset, GASP achieves a PCC(M) of 0.2946, representing a substantial relative improvement of 35.5% over the strongest baseline, TRIPLEX (0.2174), and 45.1% over HiFusion. This superiority is similarly observed in the PCC(H) metric, where GASP reaches 0.5425, indicating its exceptional capability in predicting highly variable and biologically significant genes.

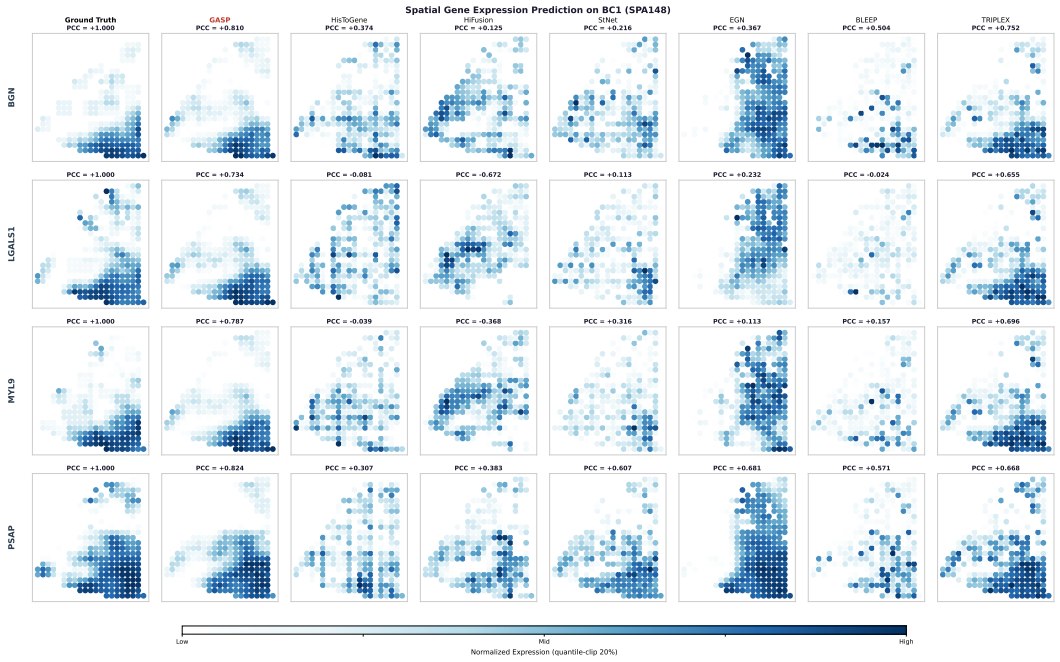

Figure 3: Spatial gene expression visualization of selected marker genes (BGN, LGALS1, MYL9, PSAP) on a representative BC1 sample (SPA148). GASP captures complex spatial patterns more accurately than baseline methods.

Table 3: Per-gene PCC on BC1 sample SPA148 for selected marker genes.

Gene	Biological Role	GASP PCC	TRIPLEX PCC	Δ Advantage
BGN	Stromal remodeling, tumor boundary	**0.810**	0.752	+0.058
LGALS1	Tumor immune evasion	**0.734**	0.655	+0.079
MYL9	Smooth muscle / stromal	**0.787**	0.696	+0.091
PSAP	Lysosomal metabolism	**0.824**	0.668	+0.156

On the larger BC2 dataset and the SCC dataset, GASP maintains its leading position, outperforming TRIPLEX by 5.9% and 36.6% in PCC(M), respectively. The consistent reduction in MSE and MAE further confirms that GASP not only captures the spatial distribution trends (as measured by PCC) but also accurately estimates the absolute expression levels. The significant performance gap between GASP and multi-scale vision-centric models (like TRIPLEX and HiFusion) strongly validates our hypothesis: passively fusing multi-scale features is insufficient. Active, gene-specific querying is essential to fully unlock the potential of complex histological representations.

4.4 Spatial Gene Expression Visualization

To qualitatively assess the predictive power of GASP, we visualize the spatial expression patterns of several key marker genes on a representative BC1 sample (SPA148). As detailed in Table 3, GASP demonstrates a profound advantage in predicting genes associated with complex biological roles.

As shown in Table 3, GASP consistently outperforms TRIPLEX across all four marker genes: immune evasion gene **LGALS1** (Δ+0.079), stromal markers **BGN** (Δ+0.058) and **MYL9** (Δ+0.091), and lysosomal marker **PSAP** (Δ+0.156). These gains demonstrate that gene-specific queries effectively capture biologically relevant spatial patterns that global feature fusion misses.

Table 4: Core component ablation on BC1 (8-fold average). All five components contribute positively; removing the Cross-View Fusion Encoder yields the largest performance drop.

Variant	PCC(M)	ΔPCC(M)	PCC(H)	MSE	MAE
GASP (Full)	**0.2946**	—	**0.5425**	**1.2509**	**0.8692**
w/o Gene Queries	0.2675	−0.0271	0.5023	1.3818	0.9006
w/o Cross-View Fusion	0.2432	−0.0514	0.4990	1.3210	0.8729
w/o Residual Decoding	0.2602	−0.0344	0.5126	1.3881	0.9190
w/o UNI Features	0.2627	−0.0319	0.5203	1.4054	0.9002
w/o Pearson Loss	0.2635	−0.0311	0.5113	1.2956	0.8710

Table 5: Sensitivity of the two core hyperparameters on BC1 (8-fold average).

λ	PCC(M)	γ	PCC(M)
0.05	0.2935	0.05	0.2912
0.15 (default)	**0.2946**	**0.15 (default)**	**0.2946**
0.30	0.2893	0.30	0.2901

4.5 Ablation Study

We conduct extensive ablation studies on the BC1 dataset (8-fold cross-validation) to evaluate the contribution of each component in GASP.

4.5.1 Module Ablation

Table 4 presents the results of removing key modules from the full GASP architecture.

- **w/o Cross-view Attention:** Removing the Cross-View Fusion Encoder causes the largest drop (ΔPCC(M) = -0.0514), as the context pool degrades to unrefined independent features, depriving gene queries of mutually-informed representations.
- **w/o Gene Queries:** Replacing learnable gene queries with a static representation reduces PCC(M) by 0.0271, highlighting the importance of allowing the model to learn gene-specific probing vectors.
- **w/o UNI Features:** Removing the UNI foundation model augmentation leads to a significant decrease (ΔPCC(M) = -0.0319), confirming that the richer patch-level pathological semantics from UNI are crucial for accurate expression prediction.
- **w/o Residual Decoding:** Setting the gene-specific bias weight to zero (λ=0) reduces PCC(M) by 0.0344, demonstrating that gene-specific residual corrections are essential for modeling high-variance genes.
- **w/o Pearson Loss:** Training solely with MSE loss drops the PCC(M) by 0.0311, demonstrating that the joint PCC loss is vital for capturing the correct spatial distribution patterns.

4.5.2 Key Hyperparameter Sensitivity

We analyze the two core model-specific hyperparameters: the residual scaling factor λ and the MSE loss weight γ. As shown in Table 5, both are robust within a reasonable range. The residual scale λ varies by less than 0.006 PCC(M) across $[0.05, 0.30]$. Sensitivity analyses for training hyperparameters (lr, wd) and architecture choices (heads, query dimension D) are provided in Appendix C.

5 Conclusion

We presented GASP, a gene-centric framework for spatial transcriptomics prediction that leverages learnable Gene Queries, Vision-Gene Cross-Attention, and Gene-Specific Residual Decoding to move beyond vision-centric baselines. GASP achieves up to 35.5% relative improvement in PCC(M) across three benchmarks, opening promising directions for biologically-aware ST prediction.

AI USE STATEMENT

Large language models (LLMs) were used during the preparation of this work for assistance with writing and editing of the manuscript text. The authors take full responsibility for all scientific content, experimental results, and conclusions presented in this paper.

ETHICS STATEMENT

This work uses publicly available spatial transcriptomics datasets (BC1, BC2, SCC) that do not contain personally identifiable information. No new human subjects research was conducted. The proposed method aims to reduce the cost of spatial transcriptomics profiling, which may broaden access to this technology in under-resourced settings. We foresee no direct ethical concerns with this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide full implementation details in Appendix A, including model architecture, hyperparameters, training procedure, and hardware specifications. Hyperparameter sensitivity analyses are reported in Appendix C. All datasets used in this work (BC1, BC2, SCC) are publicly available. We will release the source code upon acceptance.

REFERENCES

Jordan T. Ash, Gregory Darnell, Daniel Munro, and Barbara E. Engelhardt. Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nature Communications*, 12(1), 2021. doi: 10.1038/s41467-021-21727-x.

Michaela Asp, Stefania Giacomello, Ludvig Larsson, Chenglin Wu, Daniel Furth, Xiaoyan Qian, Eva Wardell, Joaquin Custodio, Johan Reimegard, Fredrik Salmen, Cecilia Osterholm, Patrik L. Stahl, Erik Sundstrom, Elisabet Akesson, Olaf Bergmann, Magda Bienko, Agneta Mansson-Broberg, Mats Nilsson, Christer Sylven, and Joakim Lundeberg. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660.e19, 2019. doi: 10.1016/j.cell.2019.11.025.

Michaela Asp, Joseph Bergenstrahle, and Joakim Lundeberg. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10), 2020. doi: 10.1002/bies.201900221.

Darren J. Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(6):317–317, 2019. doi: 10.1038/s41576-019-0129-z.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. Stimage-1k4m: A histopathology image-gene expression dataset for spatial transcriptomics. In *Advances in Neural Information Processing Systems 37*, pp. 35796–35823. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024a. doi: 10.52202/079017-1129.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024b. doi: 10.1038/s41591-024-02857-3.

Xulin Chen and Junzhou Huang. Delst: Dual entailment learning for hyperbolic image-gene pretraining in spatial transcriptomics, 2025. URL http://arxiv.org/abs/2503.00804v1.

Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11591–11600. IEEE, 2024. doi: 10.1109/cvpr52733.2024.01101.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. doi: 10.1016/j.mlwa.2021.100198.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL http://arxiv.org/abs/2010.11929v2.

Bryan He, Ludvig Bergenstrahle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Ake Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8):827–834, 2020. doi: 10.1038/s41551-020-0578-x.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016. URL https://arxiv.org/abs/1608.06993.

Andrew L. Ji, Adam J. Rubin, Kim Thrane, Sizun Jiang, David L. Reynolds, Robin M. Meyers, Margaret G. Guo, Benson M. George, Annelie Mollbrink, Joseph Bergenstrahle, Ludvig Larsson, Yunhao Bai, Bokai Zhu, Aparna Bhaduri, Jordan M. Meyers, Xavier Rovira-Clave, S. Tyler Hollmig, Sumaira Z. Aasi, Garry P. Nolan, Joakim Lundeberg, and Paul A. Khavari. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514.e22, 2020. doi: 10.1016/j.cell.2020.05.039.

Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao, and Yadong Wang. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1), 2023. doi: 10.1093/bib/bbad464.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016. URL http://arxiv.org/abs/1609.02907v4.

Yuxiang Lin, Ling Luo, Ying Chen, Xushi Zhang, Zihui Wang, Wenxian Yang, Mengsha Tong, and Rongshan Yu. St-align: A multimodal foundation model for image-gene alignment in spatial transcriptomics, 2024. URL http://arxiv.org/abs/2411.16793v1.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Silas Maniatis, Tarmo Aijo, Sanja Vickovic, Catherine Braine, Kristy Kang, Annelie Mollbrink, Delphine Fagegaltier, Zaneta Andrusivova, Sami Saarenpaa, Gonzalo Saiz-Castro, Miguel Cuevas, Aaron Watters, Joakim Lundeberg, Richard Bonneau, and Hemali Phatnani. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis, 2018.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors, 2021.

Daniela Ruiz, Paula Cardenas, Leonardo Manrique, Daniela Vega, Gabriel M. Mejia, and Pablo Arbelaez. Completing spatial transcriptomics data for gene expression prediction benchmarking. *Medical Image Analysis*, 106:103754, 2025. doi: 10.1016/j.media.2025.103754.

Patrik L. Stahl, Fredrik Salmen, Sanja Vickovic, Anna Lundmark, Jose Fernandez Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Ake Borg, Fredrik Ponten, Paul Igor Costea, Pelin Sahlen, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisen. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. doi: 10.1126/science.aaf2403.

Dylan J. Taylor, Surya B. Chhetri, Michael G. Tassia, Arjun Biddanda, Alexis Battle, and Rajiv C. McCoy. Sources of gene expression variation in a globally diverse human cohort, 2023.

Asher Trockman and J. Zico Kolter. Patches are all you need?, 2022. URL http://arxiv.org/abs/2201.09792v1.

Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), 2018. doi: 10.1126/science.aat5691.

Ziqiao Weng, Yaoyu Fang, Jiahe Qian, Xinkun Wang, Lee AD Cooper, Weidong Cai, and Bo Zhou. Hifusion: Hierarchical intra-spot alignment and regional context fusion for spatial gene expression prediction from histopathology, 2026.

Scott Wolf, Diogo Melo, Kristina M. Garske, Luisa F. Pallares, Amanda J. Lea, and Julien F. Ayroles. Characterizing the landscape of gene expression variance in humans, 2022.

Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9):1334–1347, 2021. doi: 10.1038/s41588-021-00911-1.

Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya Macparland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In *Advances in Neural Information Processing Systems 36*, pp. 70626–70637. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-3095.

Yan Yang, Md Zakir Hossain, Eric A Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5028–5037. IEEE, 2023. doi: 10.1109/wacv56688.2023.00501.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks, 2022.

Linlin Zhang, Dongsheng Chen, Dongli Song, Xiaoxia Liu, Yanan Zhang, Xun Xu, and Xiangdong Wang. Clinical and translational values of spatial transcriptomics. *Signal Transduction and Targeted Therapy*, 7(1), 2022. doi: 10.1038/s41392-022-00960-w.

Chongyue Zhao, Zhongli Xu, Xinjun Wang, Shiyue Tao, William A MacDonald, Kun He, Amanda C Poholek, Kong Chen, Heng Huang, and Wei Chen. Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Briefings in Bioinformatics*, 25(2), 2024. doi: 10.1093/bib/bbae052.

Junchao Zhu, Ruining Deng, Tianyuan Yao, Juming Xiong, Chongyu Qu, Junlin Guo, Siqi Lu, Mengmeng Yin, Yu Wang, Shilin Zhao, Haichun Yang, and Yuankai Huo. Asign: An anatomy-aware spatial imputation graphic network for 3d spatial transcriptomics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30829–30838. IEEE, 2025. doi: 10.1109/cvpr52734.2025.02871.

A Implementation Details

Table 6 summarizes the key implementation settings of GASP. The visual backbone consists of three complementary encoders: a frozen ResNet-18 (CIGAR) for target patch encoding, a 3-layer Transformer over a 5×5 neighbor grid for local context, and a 3-layer Transformer with APEG for global spatial context. A frozen UNI ViT-L foundation model provides rich patch-level pathological semantics via offline CLS token extraction. The gene-centric module uses $G{=}250$ learnable queries of dimension $D{=}512$, interacting with multi-scale visual features through 8-head Vision-Gene Cross-Attention. The residual scale $\lambda{=}0.15$ and MSE loss weight $\gamma{=}0.15$ balance gene-specific corrections and absolute expression accuracy respectively. All models are trained with AdamW for 30 epochs on a single NVIDIA A100 40 GB GPU, with early stopping patience of 15 epochs.

Table 6: GASP implementation details.

Component	**Setting**
Target patch encoder	Frozen ResNet-18 (CIGAR, self-supervised (Ciga et al., 2022)), 224×224 input
Neighbor context encoder	3-layer Transformer, $K{=}25$ adjacent spots (5×5 grid)
Global context encoder	3-layer Transformer + APEG (Chung et al., 2024)
Foundation model	Frozen UNI ViT-L (Chen et al., 2024b), offline CLS token extraction
Gene queries	$G{=}250$ learnable queries, dimension $D{=}512$
Cross-View Fusion Encoder	3 layers, 8 heads, bidirectional cross-attention over target/neighbor/global streams
Vision-Gene Cross-Attention	8 heads
Residual scale λ	0.15
MSE loss weight γ	0.15
Optimizer	AdamW, lr $= 5\times10^{-4}$, weight decay $= 1\times10^{-4}$
Training	30 epochs, batch size 1024, early stopping patience 15
Hardware	Single NVIDIA A100 40 GB GPU

B Evaluation Metric Definitions

We use four evaluation metrics throughout this paper. Let $\hat{Y}_{i,g}$ and $Y_{i,g}$ denote the predicted and ground-truth expression of gene g at spot i, with N spots and G genes.

PCC(M) — Mean Pearson Correlation Coefficient across all genes:

$$\text{PCC(M)} = \frac{1}{G}\sum_{g=1}^{G}\frac{\sum_{i=1}^{N}(\hat{Y}_{i,g}-\bar{\hat{Y}}_{\cdot,g})(Y_{i,g}-\bar{Y}_{\cdot,g})}{\sqrt{\sum_i(\hat{Y}_{i,g}-\bar{\hat{Y}}_{\cdot,g})^2}\sqrt{\sum_i(Y_{i,g}-\bar{Y}_{\cdot,g})^2}} \tag{9}$$

PCC(H) — Mean PCC for the top-50 highly variable genes (selected per fold based on expression variance in the training set):

$$\text{PCC(H)} = \frac{1}{50}\sum_{g\in\mathcal{G}_{\text{top50}}}\frac{\sum_i(\hat{Y}_{i,g}-\bar{\hat{Y}}_{\cdot,g})(Y_{i,g}-\bar{Y}_{\cdot,g})}{\sqrt{\sum_i(\hat{Y}_{i,g}-\bar{\hat{Y}}_{\cdot,g})^2}\sqrt{\sum_i(Y_{i,g}-\bar{Y}_{\cdot,g})^2}} \tag{10}$$

MSE — Mean Squared Error:

$$\text{MSE} = \frac{1}{NG}\sum_{i=1}^{N}\sum_{g=1}^{G}(\hat{Y}_{i,g}-Y_{i,g})^2 \tag{11}$$

MAE — Mean Absolute Error:

$$\text{MAE} = \frac{1}{NG}\sum_{i=1}^{N}\sum_{g=1}^{G}|\hat{Y}_{i,g}-Y_{i,g}| \tag{12}$$

Table 7: Training hyperparameter sensitivity on BC1 (8-fold average).

Param	Value	PCC(M)	ΔPCC(M)
Learning rate	1×10^{-4}	0.2923	−0.002
	$\mathbf{5 \times 10^{-4}}$	**0.2946**	—
	1×10^{-3}	0.2838	−0.011
Weight decay	0	0.2912	−0.003
	$\mathbf{1 \times 10^{-4}}$	**0.2946**	—
	2×10^{-3}	0.2945	≈0

Table 8: Architecture hyperparameter sensitivity on BC1 (8-fold average).

Param	Value	PCC(M)	ΔPCC(M)
Attention heads	4	0.2919	−0.003
	8	**0.2946**	—
	16	0.2784	−0.016
Query dim D	256	0.2874	−0.007
	512	**0.2946**	—
	1024	0.2918	−0.003

C Additional Hyperparameter Sensitivity

Training Hyperparameters. Table 7 reports sensitivity to learning rate and weight decay. The default settings (lr=5×10^{-4}, wd=1×10^{-4}) yield the best or near-best performance across both axes.

Architecture Hyperparameters. Table 8 reports sensitivity to the number of cross-attention heads and the gene query dimension D.