Preserving DEG Rankings for Gene Discovery in Histology-Based Spatial Gene Expression Prediction

Kaito Shiku[1] **Kazuya Nishimura**[2] **Yasuhiro Kojima**[3] **Ryoma Bise**[1]
[1] Kyushu University, Japan [2] The University of Osaka, Japan [3] National Cancer Center, Japan
kaito.shiku@human.ait.kyushu-u.ac.jp

Abstract

Predicting spatial gene expression from histology images could scale spatial transcriptomics (ST) to image-only cohorts, but conventional histology-based ST prediction is trained and evaluated mainly by per-gene spatial-profile reconstruction. This objective is misaligned with a key downstream use of ST: differentially expressed gene (DEG) discovery, where genes are ranked for a biological or morphology-defined contrast by evidence of between-group expression differences. We formulate image-based differential expression ranking (IDER), which asks whether predicted expression profiles preserve the contrast-specific ranked gene list obtained from measured profiles. IDER compares gene rankings induced by differential-expression statistics, rather than raw expression magnitudes or per-gene spatial correlations. We further introduce a differentiable IDER objective that aligns these statistics across genes and can be trained with morphology-derived proxy contrasts without predefined biological group labels. Experiments on public ST datasets show improved DEG-ranking agreement and pathway-enrichment overlap over conventional reconstruction objectives, including morphology-derived and pathologist-annotated tissue-region evaluations. The code is available at https://github.com/Shiku-Kaito/IDER.

1 Introduction

Predicting spatial gene expression from routine histopathology images is a promising way to extend molecular analysis to large image-only cohorts [8, 35, 27, 37, 36]. Histopathology images are widely collected in clinical and research settings, but they do not directly provide the gene expression measurements needed to interpret tissue biology. Recent methods address this gap by learning from paired histology images and spatial transcriptomics (ST) measurements, where ST provides spatially resolved gene expression while preserving tissue organization [23]. If reliable, such models could enable molecular analysis in cohorts where sequencing-based ST assays are unavailable or costly.

Conventional objectives and evaluation protocols for histology-based ST prediction primarily assess whether spatial expression profiles are accurately reconstructed. Pearson correlation coefficient (PCC)-based losses and evaluation metrics evaluate, for each gene independently, whether the predicted expression pattern across spots is consistent with the measured pattern, and then average this agreement across genes [4, 37, 36, 30]. While this is a natural criterion for expression reconstruction, it emphasizes agreement along the spatial axis within each gene rather than identifying which genes should be prioritized for downstream biological discovery. Figure 1 (a) illustrates this conventional setup, where each target gene is evaluated separately based on spatial-profile agreement.

However, a central downstream goal of ST analysis is gene discovery: identifying genes whose expression differences are associated with biologically meaningful tissue contrasts, such as those across tissue regions, disease-associated phenotypes, or morphology-defined groups [3, 32]. This perspective goes beyond expression reconstruction by asking which genes should be prioritized for

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

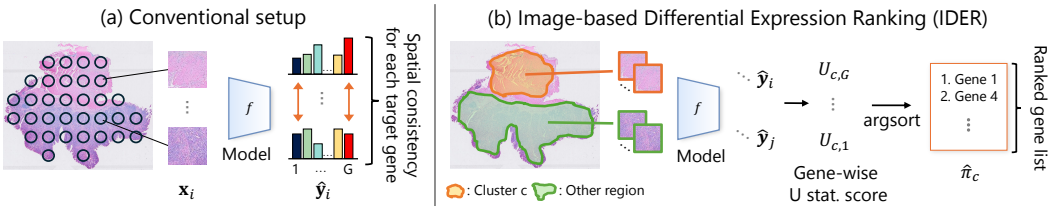

Figure 1: (a) Conventional evaluation measures spatial-profile agreement within each gene across spots. (b) Image-based Differential Expression Ranking (IDER) ranks genes by contrast statistics for a biological or morphology-defined comparison and evaluates whether predicted profiles preserve this ranked list of DEG candidates.

biological interpretation. In differential expression analysis [29, 28, 3], researchers specify such biological contrasts and seek to identify, from thousands of measured genes, those with strong statistical evidence of between-group expression differences. To this end, differential expression analysis assigns each gene a group-contrast statistic and ranks genes by the strength of this evidence. The top-ranked genes are reported as candidate differentially expressed genes (DEGs) or marker genes that help explain the tissue difference, support pathway interpretation, and guide downstream biological validation. Thus, a key output of differential expression analysis is a contrast-specific ranked gene list.[1]

This creates a direct mismatch between conventional reconstruction criteria and image-based DEG discovery. Conventional criteria compare predicted and measured expression along the spatial axis within each gene, and then summarize these per-gene agreements. In differential expression analysis, however, the relevant output is defined along the gene axis: genes are ranked against one another by their contrast scores for a biological or morphology-defined comparison. The missing evaluation axis is therefore whether predicted expression profiles induce the same contrast-specific ranked gene list as measured profiles. Errors in this ordering can change which genes are reported as candidate DEGs or marker genes, and consequently, which pathways and biological hypotheses are carried forward.

To address this limitation, we formulate image-based differential expression ranking (IDER), a downstream-aware task for histology-based spatial gene expression prediction. At evaluation time, for a specified biological or morphology-defined contrast, we compute gene-level differential-expression statistics from both measured and predicted expression profiles and evaluate whether the resulting contrast-specific ranked gene lists agree, as illustrated in Figure 1 (b). This criterion directly assesses whether predicted profiles preserve the gene prioritization used in DEG-based discovery, rather than only evaluating each gene's spatial pattern in isolation. Importantly, genes are not compared by raw expression magnitudes across genes; they are compared through their differential statistics, as in standard differential expression analysis.

Beyond evaluation, we introduce a training objective specifically designed for IDER. During training, the objective uses morphology-derived proxy contrasts obtained by clustering histology features, without requiring predefined biological group labels such as tissue-region or phenotype annotations. For each proxy contrast, we compute differentiable differential-expression statistics from predicted profiles and align them with the corresponding statistics computed from measured profiles. This encourages the model to capture genes whose expression differences are associated with morphology-defined image variation, rather than only reconstructing individual spatial expression patterns.

We evaluate the proposed framework on public ST datasets under both morphology-derived contrasts and pathologist-annotated tissue-region contrasts. Experiments show that the proposed objective improves DEG-ranking agreement and pathway-enrichment overlap compared with conventional reconstruction objectives. These results suggest that image-based differential expression ranking provides a useful downstream-aware formulation for connecting histology-based ST prediction with morphology-associated gene discovery.

Our contributions are summarized as follows:

[1]While standard DEG analysis often considers both up- and down-regulated genes, our study focuses on identifying genes that are more highly expressed in a particular tissue morphology than in other tissue groups.

- We identify a mismatch between conventional objectives and evaluation protocols for histology-based ST prediction and downstream gene discovery from ST profiles. Conventional criteria primarily measure spatial-profile agreement within each gene, whereas differential expression analysis depends on contrast-specific ranked gene lists that prioritize genes by between-group expression differences.
- We formulate image-based differential expression ranking (IDER), a downstream-aware task that evaluates whether predicted expression profiles preserve the contrast-specific ranked gene lists induced by measured expression profiles. For a specified biological or morphology-defined contrast, IDER compares gene rankings derived from differential-expression statistics computed from predicted and measured profiles.
- We introduce a differentiable IDER objective that aligns differential-expression statistics across genes using morphology-derived proxy contrasts from clustered histology features, without requiring predefined biological group labels during training. Experiments on public ST datasets show improved DEG-ranking agreement and pathway-enrichment overlap over conventional reconstruction objectives, including evaluations with morphology-derived groups and pathologist-annotated tissue regions.

2 Related Work

Spatial Gene Expression Estimation from Histopathological Images. Predicting gene expression from histology images has emerged as a promising direction for bridging morphological and molecular information. A pioneering work in this area is ST-Net [8], which estimates gene expression from image patches using convolutional neural networks. Subsequent studies have explored architectures that capture both global and local context. In particular, Transformer-based models [27] and graph convolutional networks (GCNs) [6, 15, 38] have been introduced to model long-range spatial dependencies. Multi-scale modeling strategies have also been proposed to effectively combine local and global contexts [33, 4]. Retrieval-based approaches [35, 37, 5] estimate gene expression by leveraging similarities between images and transcriptomic profiles in a joint embedding space, rather than directly regressing expression values. Methods based on diffusion models [39] and flow matching [10] have also been proposed, formulating the task as generating image-conditioned gene expression distributions.

Ranking-based Objectives for Gene Expression Estimation. Recent studies have introduced ranking-based objectives for gene expression prediction. RankByGene [11] uses spatial ranking-based alignment to improve image and gene expression correspondence, while STRank [25] models relative expression ordering across spatial spots. Both methods mainly focus on spatial-level relativity, i.e., expression rankings across spots.

While prior studies have advanced histology-based gene expression estimation, they mainly aim to reconstruct spatial trends across target genes and do not explicitly model the relative biological importance of genes. In contrast, we focus on gene-wise rankings derived from differential expression analysis, which are useful for discovering genes associated with image-derived features.

Utilizing Biological Signals in Image and Gene Expression Data. Recent studies have increasingly emphasized higher-level biological signals, such as pathways, rather than relying solely on individual gene expression profiles. PEaRL [20] incorporates pathway priors into histology-to-transcriptomics learning, whereas SurvPath [14] and PathCLAST [26] integrate pathway graphs to encode known biological relationships during model estimation. HistoPrism [9] further advocates pathway-level benchmarking to evaluate whether predicted transcriptomic profiles preserve functionally coherent biological programs.

Motivated by recent studies, we extend the ST estimation task toward downstream biological analysis to better capture biologically meaningful signals.

3 Preliminary: Spatial Gene Expression Prediction from Histology Images

The objective of conventional gene expression estimation is to reconstruct spot-wise gene expression values for all target genes, while preserving spatial expression patterns across spots [8, 27, 4, 30].

Formally, given a set of spot images $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ and corresponding gene expression vectors $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^N$, where $\mathbf{y}_i \in \mathbb{R}^G$, N denotes the number of spots and G the number of genes, existing methods learn a mapping

$$f : \mathbf{x}_i \mapsto \hat{\mathbf{y}}_i,$$

by optimizing reconstruction-based objectives such as Mean Squared Error (MSE) [8, 27, 4] or correlation-based metrics such as the Pearson Correlation Coefficient (PCC) [30, 37, 36]. PCC-based objectives measure, for each gene separately, the correlation between predicted and measured expression profiles across spots.

These objectives are natural for reconstructing spatial expression profiles for individual genes. However, they do not explicitly assess whether the predicted profiles preserve the contrast-specific ranked gene lists used in DEG-based gene discovery. This motivates the downstream-aware formulation introduced in the following section.

4 Task Definition: Image-based Differential Expression Ranking

We define image-based differential expression ranking (IDER) as the task of making contrast-specific ranked gene lists induced by predicted expression profiles agree with those induced by measured expression profiles. Given group labels for spots, differential expression analysis assigns each gene a contrast statistic and ranks genes by differential-expression evidence. The output of interest is therefore not the statistic for each gene in isolation, but the ordering of genes induced by these statistics.

Formally, let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ denote pathology images, $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^N$ the corresponding measured expression profiles, and $\mathcal{L} = \{l_i\}_{i=1}^N$ the group labels, where $\mathbf{y}_i \in \mathbb{R}^G$ and $l_i \in \{1, \ldots, K\}$. For each group c, we consider a one-versus-rest contrast:

$$A_c = \{i \mid l_i = c\}, \quad B_c = \{j \mid l_j \neq c\}.$$

For each contrast c and gene g, we use the one-sided Mann–Whitney U statistic [22, 31] as a rank-based differential-expression score. Intuitively, $U_{c,g}$ measures how often a spot in the target group A_c has higher expression of gene g than a spot in the reference group B_c; larger values indicate stronger evidence that gene g is more highly expressed in A_c relative to B_c. We compute this score as:

$$U_{c,g} = \sum_{i \in A_c} \sum_{j \in B_c} \left[\mathbf{1}[y_{i,g} > y_{j,g}] + \frac{1}{2} \mathbf{1}[y_{i,g} = y_{j,g}] \right],$$

where $\mathbf{1}[\cdot]$ denotes the indicator function, and ties are assigned half weight.

The reference DEG output for contrast c is the ranked gene list induced by these scores:

$$\pi_c = \operatorname{argsort}_g^{\downarrow} (U_{c,g}),$$

where $\operatorname{argsort}^{\downarrow}$ returns gene indices in descending order of $U_{c,g}$. Thus, π_c prioritizes genes with stronger one-sided differential-expression evidence in group A_c relative to group B_c. Given predicted expression profiles $\hat{\mathcal{Y}} = \{\hat{\mathbf{y}}_i\}_{i=1}^N$, we analogously compute $\hat{U}_{c,g}$ by replacing $y_{i,g}$ with $\hat{y}_{i,g}$ and obtain $\hat{\pi}_c = \operatorname{argsort}_g^{\downarrow}(\hat{U}_{c,g})$. The goal of IDER is to make the predicted ranked gene list $\hat{\pi}_c$ agree with the reference ranked gene list π_c for each contrast.

5 Method: Differentiable Optimization for IDER

Based on the IDER task, our goal is to learn an image-to-expression predictor whose outputs preserve contrast-specific DEG rankings. Since the ranked gene list $\hat{\pi}_c$ is induced by pairwise comparisons and sorting, directly optimizing this ranking is not differentiable. We therefore optimize a differentiable surrogate that aligns gene-level differential-expression statistics computed from predicted expression profiles with those computed from measured expression profiles.

Biological group-label-free training with morphology-derived proxy contrasts. During training, our objective does not require predefined biological group labels such as tissue-region or phenotype

annotations. Instead, we obtain morphology-derived proxy groups by clustering pathology foundation-model [19] embeddings with K-means, and construct one-versus-rest contrasts by varying the target group during optimization. This exposes the model to diverse morphology-associated expression differences and encourages it to preserve cross-gene differential-expression ordering across contrasts.

Differentiable approximation of the U statistic. For each training contrast c, with groups A_c and B_c defined as in the task formulation, we approximate the one-sided Mann–Whitney comparison by replacing the hard indicator $\mathbf{1}[\hat{y}_{i,g} > \hat{y}_{j,g}]$ with a temperature-scaled sigmoid function. The differentiable statistic for gene g is defined as:

$$\hat{U}_{c,g} = \sum_{i \in A_c} \sum_{j \in B_c} \sigma\left(\frac{\hat{y}_{i,g} - \hat{y}_{j,g}}{\tau}\right), \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\tau > 0$ is a temperature parameter. This statistic smoothly relaxes the one-sided U statistic: as $\tau \to 0$, it approaches hard pairwise comparison, while larger τ yields smoother gradients for stable optimization. Because network outputs are continuous, exact ties are rare, so we omit the tie term used in the discrete statistic.

DEG-ranking surrogate objective. For each contrast c, let $\mathbf{U}_c = (U_{c,1}, \ldots, U_{c,G})$ be the reference U-statistic vector and $\hat{\mathbf{U}}_c = (\hat{U}_{c,1}, \ldots, \hat{U}_{c,G})$ be its differentiable counterpart computed from predicted profiles. The DEG ranking is determined by the relative magnitudes of these entries across genes. We therefore optimize a correlation-based surrogate over the gene axis:

$$\mathcal{L}_{\mathrm{DEG}} = 1 - \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{Corr}_g\left(\hat{\mathbf{U}}_c, \mathbf{U}_c\right), \tag{2}$$

where Corr_g denotes the Pearson correlation computed across genes, and $\mathcal{C}$ denotes the set of group labels. This differs from conventional PCC losses, which compute correlation across spots for each fixed gene. Here, each gene is summarized by a contrast-specific differential-expression score, and the loss encourages genes with high reference U statistics to also receive high predicted U statistics, serving as a differentiable surrogate for preserving the contrast-specific DEG ranking.

Auxiliary expression-consistency regularization. The DEG-ranking objective aligns differential-expression evidence across genes, but it does not explicitly constrain the spatial expression profile of each individual gene. We therefore add a lightweight expression-consistency regularizer $\mathcal{L}_{\mathrm{const}}$, defined as a conventional PCC-based loss that correlates predicted and measured expression across spots for each gene and averages the result over genes. This term is used to stabilize training and preserve standard spatial-profile agreement, while the main objective remains DEG-ranking reproducibility. The overall loss is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{DEG}} + \lambda \mathcal{L}_{\mathrm{const}}, \tag{3}$$

where λ controls the strength of the auxiliary regularization.

6 Evaluation Metrics for IDER

We evaluate whether predicted profiles preserve the contrast-specific DEG output using two complementary ranking metrics. For each contrast c, DEG-SCC measures global agreement of differential-expression evidence by computing Spearman's rank correlation across genes $\mathrm{Spearman}_g(\cdot)$ between the reference and predicted U-statistic vectors:

$$\mathrm{SCC}_c^{\mathrm{DEG}} = \mathrm{Spearman}_g\left(\mathbf{U}_c, \hat{\mathbf{U}}_c\right). \tag{4}$$

To emphasize recovery of top-ranked DEG candidates, we additionally use binary DEG-nDCG@k [12]. Let $\pi_c(r)$ and $\hat{\pi}_c(r)$ denote the genes ranked at position r in the reference and predicted DEG lists, respectively. We define the binary relevance of gene g, indicating whether the gene is included in the top-k genes of the reference DEG list, as $\mathrm{rel}_c(g) = \mathbf{1}\left[g \in \{\pi_c(1), \ldots, \pi_c(k)\}\right]$, where $\mathbf{1}[\cdot]$ is the indicator function. Then, binary nDCG@k is defined as

$$\mathrm{nDCG}_c^{\mathrm{DEG}}@k = \frac{\sum_{r=1}^{k} \frac{\mathrm{rel}_c(\hat{\pi}_c(r))}{\log_2(r+1)}}{\sum_{r=1}^{k} \frac{1}{\log_2(r+1)}}. \tag{5}$$

$\mathrm{SCC}_c^{\mathrm{DEG}}$ evaluates agreement across all genes, whereas $\mathrm{nDCG}_c^{\mathrm{DEG}}@k$ focuses on whether genes highly ranked in the reference DEG list also appear near the top of the predicted DEG list. We report both metrics averaged across the considered contrasts.

7 Experiments

We evaluate the proposed method on spatial transcriptomics datasets to assess its ability to estimate DEG rankings from inter-group comparisons. We consider two settings: (i) a **cluster-based DEG setting** (Sec. 7.1), where groups are defined via clustering of pathology foundation model features, and (ii) a **real annotation-based DEG setting** (Sec. 7.2), where groups correspond to pathologist-provided tissue regions. Across both settings, we further analyze downstream pathway recovery, compare against conventional objectives, and conduct ablation and plug-in studies (Sec. 7.3).

Implementation Details. All experiments are implemented in PyTorch [1] and optimized using Adam [18] with a learning rate of 3×10^{-4}. We use CONCH [19] as the feature extractor for the gene expression prediction network f, with its parameters frozen during training. Only a three-layer MLP following the feature extractor is trained. Models are trained for 1,000 epochs, and we report the best validation checkpoint. The auxiliary regularization weight and temperature parameter are set to $\lambda = 1.0$ and $\tau = 0.01$, respectively.

During training, groups are defined as regions sharing similar cellular morphology, obtained via K-means clustering on image features with the number of clusters set to 10. We adopt a one-vs-rest scheme, where a target group is treated as positive and the others as negative. The positive group is randomly sampled at each iteration, encouraging the model to learn DEG rankings across all groups.

Evaluation Metrics. We use evaluation metrics based on group-wise comparison that assess the ranking of DEGs obtained by the proposed framework. Evaluation is conducted using $\mathrm{SCC}^{\mathrm{DEG}}$, which reflects the global agreement of DEG rankings, and $\mathrm{nDCG}^{\mathrm{DEG}}@k$, which evaluates the recovery of highly ranked genes. We report $\mathrm{nDCG}^{\mathrm{DEG}}$@50, @100, and @200. Similar to training, we construct a one-vs-rest setting in which one group is treated as the positive class and the remaining groups as negative classes, and compute the ranking of DEGs for evaluation. This procedure is performed for all groups, and the final performance is obtained by averaging the results across groups.

Baseline Objective Functions. To demonstrate the effectiveness of the proposed DEG-based ranking objective, we compare our method with six conventional baselines that optimize only spatial expression independently for each gene. 1) to 3) "MSE" [8, 27, 4], "PCC" [30], and "MSE & PCC" [37, 36] are the most commonly used objective functions in state-of-the-art ST estimation methods, where the model is trained using losses computed independently for each gene, such as MSE loss, PCC loss, or their combination. 4) "Poisson" and 5) "NB" [24] are likelihood-based objectives for modeling gene expression counts at the individual sample level, where Poisson assumes a Poisson distribution and NB additionally models overdispersion using a dispersion parameter. 6) "STRank" [25] is a state-of-the-art objective function that learns relative spatial expression patterns through a ranking-based loss robust to stochastic noise in ST measurements.

7.1 Cluster-based DEG Experiments

Datasets. We conducted experiments on the Bowel, Ovary, and Lymph Node samples from the public HEST-1K dataset [13]. For each organ, we used a single slide to enable a controlled comparison of different training objectives without the influence of cross-patient batch effects. Each dataset consists of paired patch images and gene expression profiles, with 4,096, 4,617, and 4,840 samples for Bowel, Ovary, and Lymph Node, respectively. All patch images are of size 224×224. ST data were obtained using Visium technology [34]. After preprocessing, as detailed in the Supplementary Material (Sec. A.7), the numbers of genes were 18,066, 18,054, and 17,564 for Bowel, Ovary, and Lymph Node, respectively. While conventional studies typically select only a small subset of highly variable genes (HVGs) for training and evaluation [13, 25], our goal is to evaluate DEG rankings across all genes; therefore, we use all genes in our experiments. Following [30], we split each dataset into training, validation, and test sets with a ratio of 3:1:1.

These datasets do not include annotations for tissue regions such as diseased or normal areas. Therefore, we construct groups of tissues with similar morphological characteristics by clustering features extracted using the pretrained pathology foundation model CONCH [19].

Table 1: **Comparison of DEG ranking performance in the cluster-based DEG setting** against methods optimized with conventional objective functions.

Objective	Ovary				Lymph Node				Bowel			
	SCC^{DEG}	nDCG^{DEG}			SCC^{DEG}	nDCG^{DEG}			SCC^{DEG}	nDCG^{DEG}		
		@50	@100	@200		@50	@100	@200		@50	@100	@200
MSE [4]	0.587	0.415	0.446	0.434	0.624	0.407	0.434	0.477	0.709	0.364	0.432	0.468
PCC [30]	0.584	0.425	0.451	0.451	0.629	0.439	0.478	0.521	0.710	0.384	0.461	0.494
MSE & PCC [37]	0.609	0.437	0.465	0.453	0.650	0.477	0.511	0.548	0.714	0.383	0.459	0.496
Poisson	0.460	0.363	0.401	0.422	0.234	0.448	0.493	0.501	0.694	0.440	0.476	0.492
NB [24]	0.245	0.404	0.415	0.417	0.090	0.541	0.548	0.534	0.601	**0.494**	0.533	0.546
STRank [25]	0.457	0.359	0.399	0.416	0.300	0.375	0.425	0.446	0.687	0.435	0.470	0.498
Ours	**0.686**	**0.486**	**0.515**	**0.531**	**0.676**	**0.601**	**0.641**	**0.643**	**0.731**	0.481	**0.565**	**0.595**

The clustering procedure is performed completely independently on the training data to ensure that no data leakage occurs. This approach allows us to generate groups at arbitrary granularity simply by varying the number of clusters, without requiring costly annotation procedures. Thus, this setting is not only useful for evaluating machine learning methods, but also has the potential to facilitate efficient differential expression analysis in real-world gene expression studies by enabling annotation-free grouping.

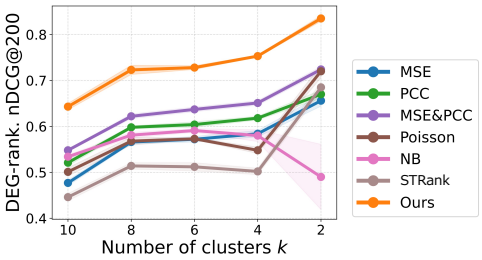

Figure 2: **Generalization Performance Across Various Grouping Granularities.**

Comparison of DEG Ranking Performance. Table 1 reports DEG ranking results under the cluster-based DEG setting. The proposed method achieves the best performance on nearly all metrics, demonstrating substantial improvements over conventional objectives. These results suggest that conventional methods, which optimize the reproducibility of gene-wise spatial expression patterns, do not necessarily improve DEG ranking estimation between morphologically distinct tissue groups. The strong performance of our method further highlights the importance of explicitly optimizing DEG rankings derived from morphology-based proxy groups.

Evaluation of Generalization Performance Across Various Grouping Granularities. In practical applications, the granularity of group partitioning varies depending on the objective of differential expression analysis, making it important to generalize DEG estimation across different grouping granularities. Therefore, in this experiment, we evaluate the generalization capability of the proposed method under varying grouping granularities using the Lymph Node dataset.

Figure 2 shows the DEG ranking performance of each method, evaluated by nDCG^{DEG}@200, under the cluster-based DEG setting with varying numbers of clusters $k = 10, 8, 6, 4, 2$. The proposed method consistently achieves the best performance across different grouping granularities. These results indicate that, although trained with a specific grouping granularity, the proposed method generalizes well to DEG ranking estimation under different grouping granularities, demonstrating strong potential for practical differential expression analyses at various levels of granularity.

7.2 Real Annotation-based DEG Experiments

Datasets. We conduct experiments using the publicly available breast cancer ST dataset Her2st [2], which includes expert-provided region annotations. Her2st contains histopathology images and ST data acquired from 8 patients using Spatial Transcriptomics technology. For each patient, 3 to 6 histopathology images are available, resulting in a total of 36 images. After applying the preprocessing procedure detailed in the Supplementary Material (Sec. A.7), we obtain 13,620 pairs of 224×224 image patches and gene expression profiles with 11,871 genes. As in the previous setting, all genes are used for both training and evaluation. We conduct patient-level cross-validation following [4] by splitting the dataset at the patient level with a ratio of 2:1:1.

In this dataset, one WSI per patient is annotated with tissue-type region labels among the multiple WSIs available for each patient. The annotations include adipose tissue, cancer in situ, connective

Table 3: **Comparison of DEG ranking performance in the real annotation-based DEG setting** against methods optimized with conventional objective functions, in terms of $\mathrm{SCC}^{\mathrm{DEG}}$ and $\mathrm{nDCG}^{\mathrm{DEG}}$@200. Due to space limitations, $\mathrm{nDCG}^{\mathrm{DEG}}$@200 is abbreviated as $\mathrm{nDCG}^{\mathrm{DEG}}$.

Objective	Average		adipose tissue		connective tissue		immune infiltrate		invasive cancer	
	$\mathrm{SCC}^{\mathrm{DEG}}$	$\mathrm{nDCG}^{\mathrm{DEG}}$	$\mathrm{SCC}^{\mathrm{DEG}}$	$\mathrm{nDCG}^{\mathrm{DEG}}$	$\mathrm{SCC}^{\mathrm{DEG}}$	$\mathrm{nDCG}^{\mathrm{DEG}}$	$\mathrm{SCC}^{\mathrm{DEG}}$	$\mathrm{nDCG}^{\mathrm{DEG}}$	$\mathrm{SCC}^{\mathrm{DEG}}$	$\mathrm{nDCG}^{\mathrm{DEG}}$
MSE [4]	0.239	0.084	0.393	0.069	0.206	0.080	0.197	0.176	0.272	0.082
PCC [30]	0.181	0.073	0.297	0.060	0.143	0.066	0.132	0.199	0.182	0.035
MSE & PCC [37]	0.195	0.073	0.317	0.076	0.208	0.048	0.134	0.239	0.216	0.043
Poisson	0.116	0.101	0.123	0.057	0.105	0.109	0.191	0.228	0.114	0.081
NB [24]	0.137	0.098	0.145	0.038	0.127	0.097	0.190	0.269	0.120	0.073
STRank [25]	0.101	0.109	0.053	0.083	0.114	0.121	0.189	0.227	0.096	0.077
Ours	**0.353**	**0.247**	**0.475**	**0.100**	**0.313**	**0.164**	**0.263**	**0.446**	**0.357**	**0.301**

tissue, immune infiltrate, invasive cancer, breast glands, and undetermined. For evaluation, we use annotated WSIs from the test patients and treat each tissue type as a distinct group. We exclude the undetermined category, as well as cancer in situ and breast glands, which are present in only a small subset of patients. In addition, classes with fewer than 10 spots per patient are excluded from evaluation, as reliable statistical testing is not feasible. Importantly, we use the annotations only for evaluation and not for training at all. Instead, training is performed using groups obtained through the automatic group construction method proposed in our framework.

Comparison of DEG Ranking Performance. Table 3 presents the DEG ranking performance for all subtypes as well as for each individual subtype under the real annotation-based DEG setting. Due to space limitations, for $\mathrm{nDCG}^{\mathrm{DEG}}@k$, we report only the results for @200. The proposed method consistently outperformed baseline methods across all tissue types and evaluation metrics. Similar to the cluster-based setting, it improved both the global agreement of DEG rankings measured by $\mathrm{SCC}^{\mathrm{DEG}}$ and the identification of top-ranked genes measured by $\mathrm{nDCG}^{\mathrm{DEG}}$@200. Notably, although the model was trained without tissue-type annotations and relied only on clusters derived from image features, it generalized well to real pathological annotations at test time. These results suggest that the proposed objective encourages the model to capture gene signals more strongly associated with histopathological tissue patterns. Therefore, compared with existing approaches, the proposed method may provide a more effective framework for discovering genes associated with image-derived pathological findings.

7.3 Further Analysis and Validation

Pathway Enrichment Overlap Analysis for Estimated ST. To assess whether predicted profiles preserve downstream functional interpretation, we perform pathway enrichment analysis using the top-ranked genes from the predicted and measured DEG lists. For each contrast, we first select the top 200 genes from the predicted and measured DEG-ranked lists, respectively. We then conduct enrichment analysis for each gene set using the Reactome_2022 and KEGG_2021_Human pathway panels [7, 16], and retain the top 200 enriched pathways according to the enrichment scores. Subsequently, following previous studies [21, 17], we compute the Jaccard index between the predicted and measured pathway sets to quantify their overlap. This analysis evaluates whether the proposed method preserves pathway-level interpretations derived from contrast-specific gene prioritization, rather than only matching individual gene expression profiles. We conduct this analysis on all datasets under both morphology-derived group and pathologist-annotated tissue-region settings.

Table 2: **Analysis of pathway enrichment overlap.**

Objective	Ovary	Lymph Node	Bowel	Breast
MSE [4]	0.326	0.382	0.336	0.144
PCC [30]	0.341	0.414	0.345	0.127
MSE & PCC [37]	0.332	0.433	0.314	0.125
Poisson	0.307	0.413	0.362	0.136
NB [24]	0.289	0.443	0.385	0.143
STRank [25]	0.302	0.383	0.360	0.147
Ours	**0.381**	**0.571**	**0.405**	**0.235**

Table 2 shows that the proposed method achieved the highest Jaccard scores across all datasets, indicating that it better recovered biologically relevant pathways compared with baseline methods. These results suggest that the proposed method better preserves biological functional structures in ST and more accurately estimates biologically meaningful signals from pathology images.

Table 4: **Ablation experiments of the proposed method**.

Objective	U stat.	DEG rank.	Ovary		Lymph Node		Bowel		Her2st	
			SCC^{DEG}	nDCG^{DEG}	SCC^{DEG}	nDCG^{DEG}	SCC^{DEG}	nDCG^{DEG}	SCC^{DEG}	nDCG^{DEG}
Baseline	✗	✗	0.584	0.451	0.629	0.521	0.710	0.494	0.181	0.073
Ours w/o DEG rank.	✓	✗	0.587	0.454	0.617	0.521	0.710	0.491	0.178	0.077
Ours	✓	✓	**0.686**	**0.531**	**0.676**	**0.643**	**0.731**	**0.595**	**0.353**	**0.247**

Table 5: **Evaluation on per-gene spatial-profile PCC.** L. N. indicates Lymph Node.

Objective	Ovary	L. N.	Bowel	Breast
MSE [4]	**0.235**	0.171	**0.345**	0.113
PCC [30]	0.233	**0.172**	**0.345**	**0.145**
MSE & PCC [37]	**0.235**	**0.172**	0.341	0.141
Poisson	0.192	0.126	0.300	0.052
NB [24]	0.195	0.130	0.303	0.056
STRank [25]	0.190	0.130	0.300	0.065
Ours	0.216	0.148	0.322	0.097

Table 6: **Plug-in experiments** for existing methods on the Her2st dataset.

Objective	SCC^{DEG}	nDCG^{DEG}		
		@50	@100	@200
ST-Net [8]	0.181	0.060	0.062	0.073
ST-Net + **Ours**	**0.353**	**0.212**	**0.230**	**0.247**
TRIPLEX [4]	0.299	0.079	0.097	0.115
TRIPLEX+ **Ours**	**0.457**	**0.190**	**0.212**	**0.242**

Ablation Experiments. We conduct ablation experiments to analyze the contribution of each component in the proposed method. We compare the following two ablation methods with our proposed method: (i) "Baseline," a baseline model trained with a PCC loss that learns spatial expression patterns for each gene independently; and (ii) "Ours w/o DEG ranking," which additionally introduces a gene-independent MSE loss on the DEG-based statistic U without considering ranking relationships between DEGs. These experiments were conducted on all datasets, and performance was evaluated using SCC^{DEG} and nDCG^{DEG}@200.

Table 4 shows the ablation results, where nDCG^{DEG} denotes nDCG^{DEG}@200 due to space limitations. Baseline achieved reasonable SCC^{DEG} performance but showed limited ability to identify top-ranked genes, as reflected by the lower nDCG^{DEG}@200 scores. Introducing MSE-based supervision on U in "Ours w/o DEG ranking" partially improved nDCG^{DEG}@200, suggesting the effectiveness of incorporating DEG signals. However, the improvements were inconsistent, indicating that gene-independent supervision on U alone is insufficient to capture inter-gene ranking relationships. In contrast, the proposed method, which optimizes inter-gene ranking relationships on U, achieved substantial improvements across all metrics. These results highlight the importance of optimizing inter-gene ranking relationships for DEG ranking estimation.

Performance in Estimating Per-Gene Spatial Expression Patterns. We also evaluate all methods using conventional per-gene PCC, which measures spatial-profile agreement across spots. As shown in Table 5, MSE- and PCC-based objectives generally achieve the highest scores, as expected because this metric matches their training targets. Although our method is not best under this reconstruction-oriented metric, it remains in an intermediate range, suggesting that the auxiliary expression-consistency term preserves reasonable spatial-profile agreement.

More importantly, together with the DEG-ranking and pathway-overlap results, this analysis highlights the gap between conventional reconstruction evaluation and downstream gene-discovery performance. Methods with higher per-gene spatial PCC do not necessarily achieve better DEG-ranking agreement or pathway-enrichment overlap, whereas our method improves these downstream metrics despite not being optimized primarily for conventional PCC. These results support the need to evaluate histology-based ST prediction not only by spatial-profile reconstruction, but also by whether predicted profiles preserve contrast-specific gene prioritization for downstream analysis.

Plug-in to Existing Methods. The proposed method is model-architecture-agnostic and can therefore be integrated into existing ST prediction architectures. Here, we applied the proposed method as a plug-in to two representative ST prediction methods, ST-Net [8] and TRIPLEX [4], and evaluated them on the Her2st dataset. The results shown in Table 6 demonstrate that DEG ranking performance was consistently and substantially improved for both methods. This indicates that the proposed method can be easily applied in a plug-and-play manner and has strong generalizability.

8 Conclusion

In this paper, we highlighted the mismatch between conventional histology-based ST prediction based on per-gene spatial-profile reconstruction and differential expression analysis, an important downstream biological analysis that ranks genes by between-group expression differences. To address this issue, we proposed Image-based Differential Expression Ranking (IDER), a framework that evaluates whether predicted expression profiles preserve contrast-specific DEG rankings. We further introduced a differentiable objective that aligns differential-expression statistics across genes using morphology-derived proxy groups without predefined biological annotations. Experiments on public ST datasets demonstrated that the proposed method improved DEG-ranking agreement and pathway-enrichment overlap compared with conventional reconstruction objectives.

Acknowledgments and Disclosure of Funding

This work was supported by JST BOOST Grant Number JPMJBS2406, SIP Grant Number JPJ012425, JST ASPIRE Grant Number JPMJAP2403, JSPS KAKENHI Grant Numbers JP26K21391 and JP24KJ2205, and MEXT Supporting Pioneering Research through AI for 1,000 Discovery Challenges Program, SPReAD, Japan Grant Number JPMXP1726260906.

References

[1] Paszke Adam, Gross Sam, Massa Francisco, Lerer Adam, Bradbury James, Chanan Gregory, Killeen Trevor, Lin Zeming, Gimelshein Natalia, Antiga Luca, Desmaison Alban, Köpf Andreas, Yang Edward, DeVito Zach, Raison Martin, Tejani Alykhan, Chilamkurthy Sasank, Steiner Benoit, Fang Lu, Bai Junjie, and Chintala Soumith. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Neural Information Processing Systems*, pages 8026 – 8037, 2019.

[2] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial Deconvolution of HER2-positive Breast Tumors Reveals Novel Intercellular Relationships. *BioRxiv*, pages 2020–07, 2020.

[3] Dylan M Cable, Evan Murray, Vignesh Shanmugam, Simon Zhang, Luli S Zou, Michael Diao, Haiqi Chen, Evan Z Macosko, Rafael A Irizarry, and Fei Chen. Cell Type-Specific Inference of Differential Expression in Spatial Transcriptomics. *Nature Methods*, 19(9):1076–1087, 2022.

[4] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate Spatial Gene Expression Prediction by Integrating Multi-resolution Features. In *Computer Vision and Pattern Recognition*, pages 11591–11600, 2024.

[5] Thao M Dang, Haiqing Li, Yuzhi Guo, Hehuan Ma, Feng Jiang, Yuwei Miao, Qifeng Zhou, Jean Gao, and Junzhou Huang. HAGE: Hierarchical Alignment Gene-Enhanced Pathology Representation Learning with Spatial Transcriptomics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 228–238, 2025.

[6] Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. Merge: Multi-faceted Hierarchical Graph-based GNN for Gene Expression Prediction from Whole Slide Histopathology Images. In *Computer Vision and Pattern Recognition Conference*, pages 15611–15620, 2025.

[7] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Alexandra Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The Reactome Pathway Knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2022.

[8] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke 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.

[9] Susu Hu, Qinghe Zeng, Nithya Bhasker, Jakob Nikolas Kather, and Stefanie Speidel. HistoPrism: Unlocking Functional Pathway Analysis from Pan-Cancer Histology via Gene Expression Prediction. In *International Conference on Learning Representations*, 2026.

[10] Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable Generation of Spatial Transcriptomics from Histology Images via Whole-Slide Flow Matching. In *International Conference on Machine Learning*, 2025.

[11] Wentao Huang, Meilong Xu, Xiaoling Hu, Shahira Abousamra, Aniruddha Ganguly, Saarthak Kapse, Alisa Yurovsky, Prateek Prasanna, Tahsin Kurc, Joel Saltz, Michael L. Miller, and Chao Chen. RankByGene: Gene-Guided Histopathology Representation Learning Through Cross-Modal Ranking Consistency. *Transactions on Medical Imaging*, pages 1–1, 2026.

[12] Kalervo Järvelin and Jaana Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. In *ACM Transactions on Information Systems*, pages 422–446, 2002.

[13] Guillaume Jaume, Paul Doucet, Andrew Song, Ming Yang Lu, Cristina Almagro Pérez, Sophia Wagner, Anurag Vaidya, Richard Chen, Drew Williamson, Ahrong Kim, et al. Hest-1k: A dataset for Spatial Transcriptomics and Histology Image Analysis. *Neural Information Processing Systems*, 37:53798–53833, 2024.

[14] Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction. In *Computer Vision and Pattern Recognition*, pages 11579–11590, 2024.

[15] 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):bbad464, 2024.

[16] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Mayu Ishiguro-Watanabe, and Mao Tanabe. KEGG: Integrating Viruses and Cellular Organisms. *Nucleic Acids Research*, 49(D1):D545–D551, 2021.

[17] Ieva Kerseviciute and Juozas Gordevicius. aPEAR: an R Package for Autonomous Visualization of Pathway Enrichment Networks. *Bioinformatics*, 39(11):btad672, 2023.

[18] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *ArXiv preprint arXiv:1412.6980*, 2014.

[19] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A Visual-Language Foundation Model for Computational Pathology. *Nature medicine*, pages 863–874, 2024.

[20] Sejuti Majumder, Saarthak Kapse, Moinak Bhattacharya, Xuan Xu, Alisa Yurovsky, and Prateek Prasanna. PEaRL: Pathway-Enhanced Representation Learning for Gene and Pathway Expression Prediction from Histology. In *Winter Conference on Applications of Computer Vision*, pages 8052–8062, 2026.

[21] Farhad Maleki, Katie L. Ovens, Elham babazadeh Rezaei, Alan M. Rosenberg, and Anthony J. Kusalik. Method Choice in Gene Set Analysis Has Important Consequences for Analysis Outcome. In *Bioinformatics*, 2019.

[22] Henry B Mann and Donald R Whitney. On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

[23] Vivien Marx. Method of the Year: Spatially Resolved Transcriptomics. *Nature Methods*, 18(1):9–14, 2021.

[24] Kazuya Nishimura, Ryoma Bise, Shinnosuke Matsuo, Haruka Hirose, and Yasuhiro Kojima. Cell-Type Prototype-informed Neural Network for Gene Expression Estimation from Pathology Images. *arXiv preprint arXiv:2603.18461*, 2026.

[25] Kazuya Nishimura, Haruka Hirose, Ryoma Bise, Kaito Shiku, and Yasuhiro Kojima. Learning Relative Gene Expression Trends from Pathology Images in Spatial Transcriptomics. In *Neural Information Processing Systems*, 2025.

[26] Minho Noh, Sungkyung Lee, Sunghyun Kim, and Sangsoo Lim. PathCLAST: Pathway-augmented Contrastive Learning with Attention for Interpretable Spatial Transcriptomics. *Briefings in Bioinformatics*, 27(1):bbag029, 2026.

[27] Minxing Pang, Kenong Su, and Mingyao Li. Leveraging Information in Spatial Transcriptomics to Predict Super-resolution Gene Expression from Histology Images in Tumors. *BioRxiv*, pages 2021–11, 2021.

[28] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies. *Nucleic Acids Research*, pages e47–e47, 2015.

[29] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression data. *Bioinformatics*, pages 139–140, 2010.

[30] Kaito Shiku, Kazuya Nishimura, Shinnosuke Matsuo, Yasuhiro Kojima, and Ryoma Bise. Auxiliary Gene Learning: Spatial Gene Expression Estimation by Auxiliary Gene Selection. In *Association for the Advancement of Artificial Intelligence*, volume 40, pages 9015–9023, 2026.

[31] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-cell Data. *cell*, 177(7):1888–1902, 2019.

[32] Benjamin L Walker, Zixuan Cang, Honglei Ren, Eric Bourgain-Chang, and Qing Nie. Deciphering Tissue Structure and Function Using Spatial Transcriptomics. *Communications Biology*, 5(1):220, 2022.

[33] Hongyi Wang, Xiuju Du, Jing Liu, Shuyi Ouyang, Yen wei Chen, and Lanfen Lin. M2ORT: Many-To-One Regression Transformer for Spatial Transcriptomics Prediction from Histopathology Images. *Association for the Advancement of Artificial Intelligence*, 2025.

[34] Cameron G Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An Introduction to Spatial Transcriptomics for Biomedical Research. *Genome Medicine*, 14(1):68, 2022.

[35] 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. *Neural Information Processing Systems*, 36:70626–70637, 2023.

[36] Yan Yang, Md Zakir Hossain, Eric Stone, and Shafin Rahman. Spatial Transcriptomics Analysis of Gene Expression Prediction Using Exemplar Guided Graph Neural Network. *Pattern Recognition*, 145:109966, 2024.

[37] 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 *Winter Conference on Applications of Computer Vision*, pages 5039–5048, 2023.

[38] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial Transcriptomics Prediction from Histology Jointly Through Transformer and Graph Neural Networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022.

[39] Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion Generative Modeling for Spatially Resolved Gene Expression Inference from Histology Images. In *International Conference on Learning Representations*, 2025.

A Supplementary Materials

A.1 Limitations and Future Work

Our results demonstrate the effectiveness of explicitly optimizing DEG rankings compared with conventional objectives that do not consider gene rankings. However, our current objective optimizes the overall ranking agreement across all genes. Developing ranking objectives that place greater emphasis on top-ranked DEGs, which are particularly important for downstream biological analysis, is a promising direction for future work.

Additionally, the absolute DEG-ranking performance on pathologist-annotated regions remains relatively low. This may partly reflect the inherent difficulty of histology-based gene expression prediction on challenging datasets such as Her2st, indicating room for further improvement.

A.2 Evaluation and Biological Interpretation of Significantly Enriched Pathways

In the main paper, we conducted a Pathway Enrichment Overlap Analysis by calculating the Jaccard index over the top 200 enriched pathways. We further evaluated pathway enrichment using only significantly enriched pathways on the Her2st dataset. Specifically, we restricted the analysis to pathways with FDR-adjusted $p < 0.05$ and calculated the Jaccard index. Furthermore, we ranked pathways according to their enrichment combined scores and used SCC to quantify the agreement between the ground-truth and predicted rankings. We also report PCC to evaluate the agreement between the ground-truth and predicted combined scores.

As shown in Table 1, IDER achieves the best performance across all three metrics. In particular, the improvement in SCC demonstrates that IDER better preserves the ranking of enriched pathways, highlighting its effectiveness in capturing pathway-level ranking agreement beyond pathway-set overlap.

The recovered pathways are also consistent with known biology. For example, invasive cancer regions exhibit highly ranked ERBB2 signaling pathways associated with HER2-positive breast cancer, while immune infiltrate regions exhibit highly ranked tumor-immune pathways related to antigen presentation, T-cell activation, and PD-1 signaling. These findings suggest that the improved pathway-ranking agreement reflects biologically relevant signals rather than merely increased overlap of enriched pathways.

Table 1: **Comparison of pathway enrichment performance across different training objectives on the Her2st dataset.** Evaluation is restricted to significantly enriched pathways (FDR-adjusted $p < 0.05$). Jaccard measures the overlap of significant pathway sets, while PCC and SCC measure the agreement of enrichment combined scores and pathway rankings, respectively, between ground-truth and predicted gene expression.

Objective	Jaccard	PCC	SCC
MSE	0.071	0.157	0.010
PCC	0.058	0.161	0.020
MSE & PCC	0.062	0.222	-0.139
Poisson	0.073	0.374	0.088
NB	0.064	0.196	-0.045
STRank	0.086	0.169	-0.025
Ours	**0.227**	**0.480**	**0.271**

A.3 Generalization Across Multiple Pathology Foundation Models

To evaluate whether our method generalizes beyond the CONCH-specific morphological representations used for group construction during training, we conducted additional evaluations using tissue morphology groups constructed with different pathology foundation models. Specifically, we evaluated our method on the Ovary dataset using four state-of-the-art pathology foundation models: UNIv2 [2], Virchow2 [5], OmniCLIP [3], and GigaPath [4].

As shown in Table 2, our method consistently achieves the best DEG-ranking performance across morphology groups constructed using all four foundation models. These results demonstrate that

IDER, although trained using CONCH-derived morphology groups, is not strongly dependent on CONCH-specific group definitions and generalizes to morphological contrasts captured by different pathology foundation models.

Table 2: **Comparison of performance across different pathology foundation models** against methods optimized with conventional objective functions.

Objective	UNIv2	Virchow2	OmniCLIP	GigaPath
MSE	0.595	0.550	0.607	0.598
PCC	0.593	0.551	0.604	0.592
MSE & PCC	0.612	0.569	0.627	0.614
Poisson	0.506	0.471	0.510	0.473
NB	0.323	0.288	0.302	0.281
STRank	0.502	0.470	0.503	0.468
Ours	**0.674**	**0.633**	**0.692**	**0.672**

A.4 DEG Ranking Performance on Highly Variable Genes

IDER allows DEG-ranking performance to be evaluated on any subset of genes at inference time. To assess its effectiveness on highly variable genes (HVGs), we additionally evaluated IDER on the Her2st dataset using the top 2,000 and 5,000 HVGs, which are commonly used in transcriptomic analyses.

As shown in Table 3, IDER consistently achieves the best performance in terms of both SCC and nDCG@200 for the top 2,000 and 5,000 HVGs. These results demonstrate that IDER remains effective when focusing on HVGs, indicating that its improvements are not solely attributable to the inclusion of all genes.

Table 3: **Comparison of DEG ranking performance across different numbers of highly variable genes (HVGs)** against methods optimized with conventional objective functions, in terms of $\mathrm{SCC}^{\mathrm{DEG}}$ and $\mathrm{nDCG}^{\mathrm{DEG}}$@200. Due to space limitations, $\mathrm{nDCG}^{\mathrm{DEG}}$@200 is abbreviated as $\mathrm{nDCG}^{\mathrm{DEG}}$.

Objective	**HVG2000**		**HVG5000**	
	$\mathrm{SCC}^{\mathrm{DEG}}$	$\mathrm{nDCG}^{\mathrm{DEG}}$	$\mathrm{SCC}^{\mathrm{DEG}}$	$\mathrm{nDCG}^{\mathrm{DEG}}$
MSE	0.205	0.231	0.221	0.145
PCC	0.168	0.203	0.178	0.126
MSE & PCC	0.193	0.212	0.188	0.123
Poisson	0.149	0.215	0.143	0.145
NB	0.177	0.216	0.173	0.153
STRank	0.161	0.217	0.143	0.154
Ours	**0.351**	**0.343**	**0.362**	**0.286**

A.5 Comparative experiments reporting mean and standard deviation

Due to space limitations, only the mean values from cross-validation were reported in the main text for the comparative experiments under the cluster-based DEG setting and the real annotation-based DEG setting. In this section, we additionally present the results with standard deviations, where the results for the cluster-based DEG setting are shown in Table 4 and those for the real annotation-based DEG setting are shown in Table 5. Furthermore, for the real annotation-based DEG setting, we also report the $\mathrm{nDCG}^{\mathrm{DEG}}$@50 and @100 metrics, which were omitted in the main text.

Table 4: **Comparison of DEG ranking performance in the cluster-based DEG setting.**

Objective	SCC^{DEG}	nDCG^{DEG}		
		@50	@100	@200
Ovary				
MSE	0.587±0.009	0.415±0.003	0.446±0.004	0.434±0.005
PCC	0.584±0.011	0.425±0.006	0.451±0.003	0.451±0.006
MSE & PCC	0.609±0.014	0.437±0.011	0.465±0.010	0.453±0.006
Poisson	0.460±0.009	0.363±0.004	0.401±0.004	0.422±0.008
NB	0.245±0.016	0.404±0.012	0.415±0.012	0.417±0.009
STRank	0.457±0.018	0.359±0.010	0.399±0.006	0.416±0.005
Ours	**0.686**±0.004	**0.486**±0.026	**0.515**±0.011	**0.531**±0.011
Lymph Node				
MSE	0.624±0.013	0.407±0.008	0.434±0.008	0.477±0.006
PCC	0.629±0.007	0.439±0.012	0.478±0.010	0.521±0.007
MSE & PCC	0.650±0.002	0.477±0.010	0.511±0.005	0.548±0.002
Poisson	0.234±0.012	0.448±0.020	0.493±0.014	0.501±0.009
NB	0.090±0.018	0.541±0.004	0.548±0.006	0.534±0.004
STRank	0.300±0.009	0.375±0.014	0.425±0.003	0.446±0.008
Ours	**0.676**±0.006	**0.601**±0.011	**0.641**±0.007	**0.643**±0.007
Bowel				
MSE	0.709±0.007	0.364±0.005	0.432±0.006	0.468±0.007
PCC	0.710±0.001	0.384±0.005	0.461±0.004	0.494±0.006
MSE & PCC	0.714±0.001	0.383±0.008	0.459±0.005	0.496±0.004
Poisson	0.694±0.001	0.440±0.009	0.476±0.006	0.492±0.007
NB	0.601±0.003	**0.494**±0.019	0.533±0.013	0.546±0.008
STRank	0.687±0.002	0.435±0.011	0.470±0.008	0.498±0.001
Ours	**0.731**±0.005	0.481±0.017	**0.565**±0.012	**0.595**±0.012

Table 5: **Comparison of DEG ranking performance in the real annotation-based DEG setting**.

Objective	SCC^{DEG}	nDCG^{DEG}		
		@50	@100	@200
Average				
MSE	0.239±0.085	0.053±0.050	0.065±0.054	0.084±0.053
PCC	0.181±0.119	0.060±0.057	0.062±0.055	0.073±0.047
MSE & PCC	0.195±0.117	0.051±0.049	0.061±0.054	0.073±0.051
Poisson	0.116±0.079	0.086±0.060	0.095±0.060	0.101±0.057
NB	0.137±0.138	0.072±0.064	0.086±0.070	0.098±0.066
STRank	0.101±0.104	0.085±0.081	0.100±0.086	0.109±0.076
Ours	**0.353**±0.084	**0.212**±0.104	**0.230**±0.101	**0.247**±0.067
Adipose tissue				
MSE	0.393±0.080	0.083±0.083	0.068±0.052	0.069±0.034
PCC	0.297±0.177	0.071±0.071	0.057±0.053	0.060±0.034
MSE & PCC	0.317±0.102	0.083±0.083	0.069±0.057	0.076±0.043
Poisson	0.123±0.052	0.044±0.035	0.053±0.026	0.057±0.014
NB	0.145±0.196	0.025±0.025	0.036±0.027	0.038±0.018
STRank	0.053±0.006	0.081±0.020	0.085±0.001	0.083±0.014
Ours	**0.475**±0.099	**0.092**±0.029	**0.087**±0.021	**0.100**±0.031
Connective tissue				
MSE	0.206±0.054	0.053±0.039	0.063±0.042	0.080±0.041
PCC	0.143±0.107	0.052±0.058	0.055±0.063	0.066±0.045
MSE & PCC	0.208±0.120	0.027±0.039	0.039±0.038	0.048±0.028
Poisson	0.105±0.092	0.080±0.076	0.099±0.079	0.109±0.072
NB	0.127±0.133	0.062±0.076	0.079±0.093	0.097±0.083
STRank	0.114±0.119	0.102±0.151	0.107±0.137	0.121±0.112
Ours	**0.313**±0.133	**0.210**±0.185	**0.189**±0.166	**0.164**±0.123
Immune infiltrate				
MSE	0.197±0.039	0.080±0.033	0.130±0.037	0.176±0.021
PCC	0.132±0.040	0.152±0.124	0.165±0.096	0.199±0.074
MSE & PCC	0.134±0.112	0.167±0.001	0.214±0.030	0.239±0.014
Poisson	0.191±0.025	0.191±0.059	0.203±0.060	0.228±0.074
NB	0.190±0.010	0.219±0.078	0.256±0.113	0.269±0.089
STRank	0.189±0.001	0.191±0.070	0.223±0.043	0.227±0.043
Ours	**0.263**±0.006	**0.387**±0.002	**0.426**±0.026	**0.446**±0.026
Invasive cancer				
MSE	0.272±0.104	0.056±0.060	0.065±0.076	0.082±0.082
PCC	0.182±0.123	0.029±0.026	0.030±0.027	0.035±0.030
MSE & PCC	0.216±0.125	0.025±0.029	0.031±0.034	0.043±0.042
Poisson	0.114±0.097	0.083±0.062	0.081±0.051	0.081±0.040
NB	0.120±0.185	0.057±0.047	0.062±0.045	0.073±0.041
STRank	0.096±0.127	0.040±0.064	0.065±0.070	0.077±0.063
Ours	**0.357**±0.125	**0.188**±0.071	**0.245**±0.062	**0.301**±0.068

A.6 Ablation Results for All Evaluation Metrics

Due to space limitations, only a subset of evaluation metrics for the ablation experiments was reported in the main text. Here, we additionally report the performance on the omitted metrics, as shown in Table 6.

Table 6: **Ablation experiments across datasets.**

Objective	Setting		SCC^{DEG}	nDCG^{DEG}		
	U stat.	DEG rank.		@50	@100	@200
Ovary						
Baseline	✗	✗	0.584±0.011	0.425±0.006	0.451±0.003	0.451±0.006
Ours w/o DEG rank.	✓	✗	0.587±0.005	0.439±0.009	0.459±0.004	0.454±0.001
Ours	✓	✓	**0.686**±0.004	**0.486**±0.026	**0.515**±0.011	**0.531**±0.011
Lymph Node						
Baseline	✗	✗	0.629±0.007	0.439±0.012	0.478±0.010	0.521±0.007
Ours w/o DEG rank.	✓	✗	0.617±0.004	0.440±0.009	0.475±0.004	0.521±0.003
Ours	✓	✓	**0.676**±0.006	**0.601**±0.011	**0.641**±0.007	**0.643**±0.007
Bowel						
Baseline	✗	✗	0.710±0.001	0.384±0.005	0.461±0.004	0.494±0.006
Ours w/o DEG rank.	✓	✗	0.710±0.004	0.379±0.013	0.464±0.015	0.491±0.010
Ours	✓	✓	**0.731**±0.005	**0.481**±0.017	**0.565**±0.012	**0.595**±0.012
Her2st						
Baseline	✗	✗	0.181±0.119	0.060±0.057	0.062±0.055	0.073±0.047
Ours w/o DEG rank.	✓	✗	0.178±0.113	0.068±0.060	0.066±0.051	0.077±0.046
Ours	✓	✓	**0.353**±0.084	**0.212**±0.104	**0.230**±0.101	**0.247**±0.067

A.7 Processing Details of Dataset

Bowel, Ovary, and Lymph Node Datasets. We removed spots with fewer than 200 detected genes and genes detected in fewer than 3 spots. We then applied spot-wise total-count normalization followed by a log1p transformation. No imputation or spatial smoothing was performed. After preprocessing, 18,066, 18,054, and 17,564 genes remained for Bowel, Ovary, and Lymph Node, respectively. Each dataset was split into training, validation, and test sets at a ratio of 3:1:1.

Her2st Dataset. We first retained genes shared across all patients. We then removed spots with fewer than 200 detected genes and genes detected in fewer than 3 spots. We applied a log1p transformation. No imputation or spatial smoothing was performed. After preprocessing, 11,871 genes remained. We performed four-fold patient-wise cross-validation on patients A–H, splitting them into training, validation, and test sets at a ratio of 2:1:1.

A.8 Computer Resources

All experiments were conducted on a desktop computer with the following specifications:

- CPU: Intel(R) Xeon(R) Gold 5122 CPU (8 cores, 16 threads)
- GPU: NVIDIA RTX A6000 (48 GB GPU memory)
- Memory: 128 GB

A.9 Licenses for Existing Assets

We implemented our method using PyTorch [1], which is released under the BSD 3-Clause License. For the experiments, we used publicly available datasets, including HEST-1k (CC BY-NC-SA 4.0) and Her2st (CC BY 4.0).

References

[1] Paszke Adam, Gross Sam, Massa Francisco, Lerer Adam, Bradbury James, Chanan Gregory, Killeen Trevor, Lin Zeming, Gimelshein Natalia, Antiga Luca, Desmaison Alban, Köpf Andreas,

Yang Edward, DeVito Zach, Raison Martin, Tejani Alykhan, Chilamkurthy Sasank, Steiner Benoit, Fang Lu, Bai Junjie, and Chintala Soumith. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Neural Information Processing Systems*, pages 8026 – 8037, 2019.

[2] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a General-purpose Foundation Model for Computational Pathology. *Nature Medicine*, 30(3):850–862, 2024.

[3] Weiqing Chen, Pengzhi Zhang, Tu N Tran, Yiwei Xiao, Shengyu Li, Vrutant V Shah, Hao Cheng, Kristopher W Brannan, Keith Youker, Li Lai, et al. A Visual–omics Foundation Model to Bridge Histopathology with Spatial Transcriptomics. *Nature Methods*, 22(7):1568–1582, 2025.

[4] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A Whole-slide Foundation Model for Digital Pathology from Real-world Data. *Nature*, 630(8015):181–188, 2024.

[5] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology. *arXiv preprint arXiv:2408.00738*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly stated our claims, and the experiments validated our assumptions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and future directions in Section A.1 of the Supplementary Material.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our result does not contain theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The source code is publicly available at `https://github.com/Shiku-Kaito/IDER`. The details of the experiment conditions were described in the experiment session.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The source code is publicly available at https://github.com/Shiku-Kaito/IDER.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We described the details of the implementation detail in our paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our figure contains error bars, and the variance of the main result is shown in the Supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details of computer resources are described in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We used publicly available datasets, including HEST-1k (CC BY-NC-SA 4.0) and Her2st (CC BY 4.0).

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work was conducted on publicly available, de-identified data and focuses on methodological development. Although the data originate from human subjects, they are anonymized and released under appropriate usage conditions, and no immediate societal impact is expected.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not use such types of data and models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All methods employed in this study are appropriately cited. Additional details, including licensing information, are provided in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The source code is publicly available at `https://github.com/Shiku-Kaito/IDER`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our experiments used only publicly available datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We used LLMs for only writing, editing, and coding.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.