Appendix

1. Method

Algorithm 1 presents an overview of our approach.

Algorithm 1 GSAT-RAG Framework

Input: Image patch V; training gene expressions E; gene identifiers G; E_{db} is all gene expression(E) in the training set.
Function TRAINCONTRASTIVE(V, E, G)
  $V_{\text{proj}} \leftarrow \mathcal{F}_{\text{proj}}^{\text{Contr}}(\mathcal{F}_{\text{image}}(V))$
  $C_{\text{CLS}} \leftarrow \mathcal{F}_{\text{encoder}}(G, E)$
  Compute CLIP loss: $\mathcal{L}_{\text{CLIP}}(V_{\text{proj}}, C_{\text{CLS}})$
Function RETRIEVAL(V, E_{db}, G)
  $V_{\text{proj}} \leftarrow \mathcal{F}_{\text{proj}}^{\text{Contr}}(\mathcal{F}_{\text{image}}(V))$
  $C_{\text{CLS}} \leftarrow \mathcal{F}_{\text{encoder}}(G, E)$
  $d \leftarrow \text{dist}(V_{\text{proj}}, C_{\text{CLS}})$
  $I \leftarrow \text{Top-K}_G(d)$
  Retrieve $\{E_r\}_{r=1}^{K_G}$ using indices I
  $E^{\text{retr}} \leftarrow \frac{1}{K_G} \sum_{r=1}^{K_G} E_r$
  return E^{retr}
Function TRAINGENERATION(V, E_{db}, G, E_y)
  $E^{\text{retr}} \leftarrow \text{RETRIEVAL}(V, E_{db}, G)$
  $V_{\text{proj}} \leftarrow \mathcal{F}_{\text{proj}}^{\text{gen}}(\mathcal{F}_{\text{image}}(V))$
  $C_{\text{CLS}}^{\text{retr}} \leftarrow \mathcal{F}_{\text{encoder}}(G, E^{\text{retr}})$
  $h(E_y \mid V, G) \leftarrow \mathcal{F}_{\text{decoder}}([C_{\text{CLS}}^{\text{retr}} \cdot V_{\text{proj}}], G)$
  $\mathcal{L}_{\text{RAG}} \leftarrow \mathcal{L}_{\text{MSE}}(h(E_y \mid V, G), E_y)$
Function PREDICT(V, G, E_{db}, G^*)
  $E^{\text{retr}} \leftarrow \text{RETRIEVAL}(V, E_{db}, G)$
  $V_{\text{proj}} \leftarrow \mathcal{F}_{\text{proj}}^{\text{gen}}(\mathcal{F}_{\text{image}}(V))$
  $C_{\text{CLS}}^{\text{retr}} \leftarrow \mathcal{F}_{\text{encoder}}(G, E^{\text{retr}})$
  $h(E_y^* \mid V, G^*) \leftarrow \mathcal{F}_{\text{decoder}}([C_{\text{CLS}}^{\text{retr}} \cdot V_{\text{proj}}], G^*)$
  return $h(E_y^* \mid V, G^*)$

2. Experiment

2.1. Time measurement

In Table 1, we show the inference time (on sample CID4465) in comparison between GSAT-RAG and baseline methods.

Table 1. Time measurement in second

Method	GSAT-RAG	Stem	BLEEP	TRIPLEX
Time(Sec)	125	63376	110	153

2.2. Scaling unseen prediction results on the HER2ST dataset

Seen and unseen gene sets have a difference in overall expression (row-wise sums). To make the row-wise gene-expression sums of the predictions match those of the ground truth, each row of the predictions is scaled by:

$$\hat{P}_i = P_i \times \frac{\sum_j T_{i,j}}{\sum_j P_{i,j}}, \tag{1}$$

where T denotes the ground truth, P denotes the predictions, and $\hat{P}$ denotes the scaled predictions whose row-wise sums satisfy

$$\sum_j \hat{P}_{i,j} = \sum_j T_{i,j}. \tag{2}$$

The results are reported in Table 2.

Table 2. Performance comparison with and without scaling

Scaling	MSE	MAE
No	0.9719	1.7143
Yes	0.8278	1.2210

2.3. Additional visualization

In Figure 1, we compare the expression levels of four oncogenes between the ground truth and GSAT-RAG under the seen setting on the HER2ST dataset. As for the oncogene HMGB1 [1], we show seen (Figure 2) and unseen (Figure 3) expression predictions from GSAT-RAG and compare them with the ground truth and baseline methods.

2.4. GSAT-RAG retains gene heterogeneity under seen setting

As mentioned in previous studies [2], recapitulating gene variance is important for preserving the biological heterogeneity of ST data. Therefore, we investigate gene variance in the breast cancer dataset using sample CID4465.

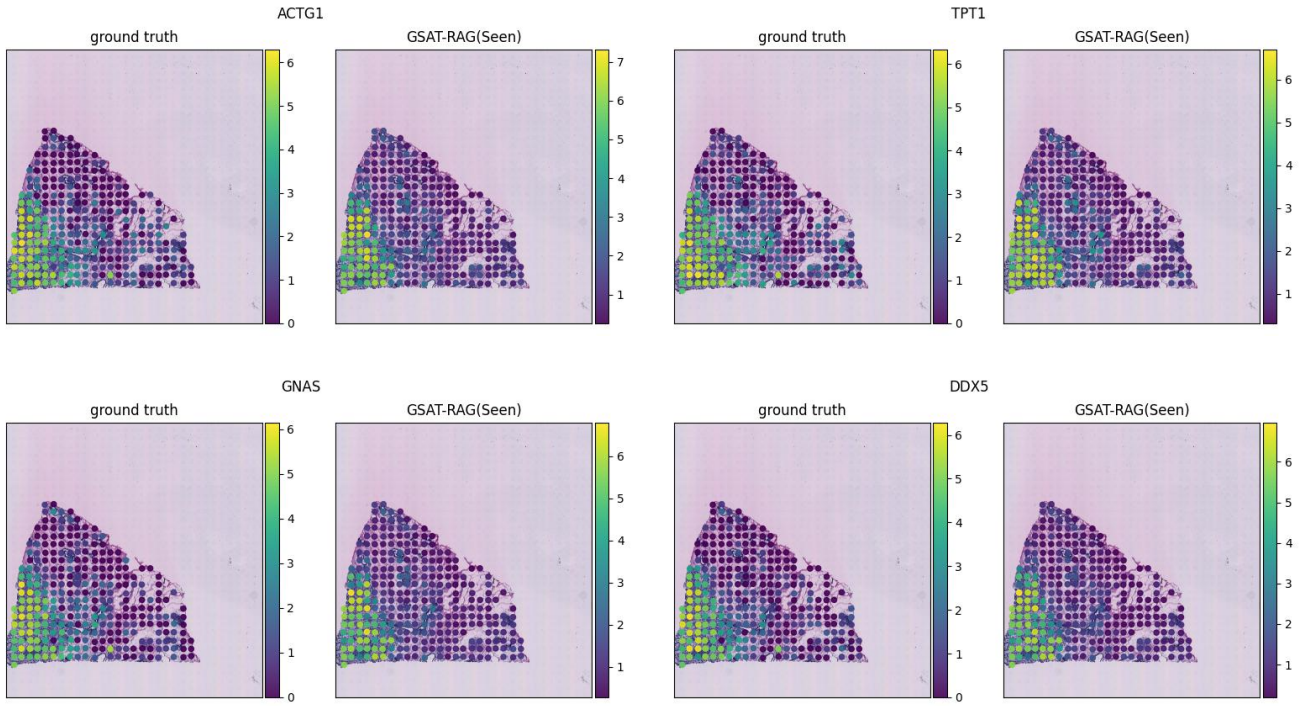

Figure 1. For the HER2ST dataset, we compare the expression of four oncogenes between the ground truth and GSAT-RAG under the seen setting.

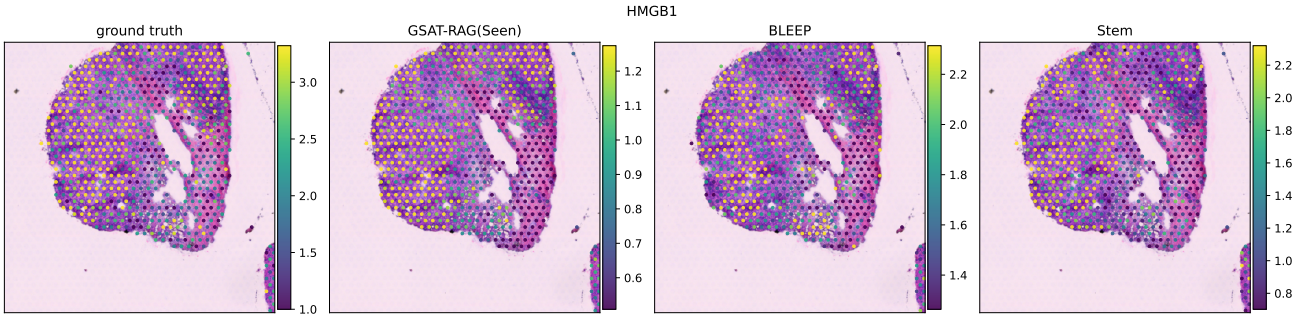

Figure 2. Expression of oncogene HMGB1; all methods use the seen setting.

We compare GSAT-RAG (seen and unseen settings) with baseline methods (BLEEP and Stem). Figure 4 shows normalization using mean expressions in the first row and variances in the second row. Under the seen setting, GSAT-RAG, BLEEP, and Stem closely resemble the original expression profiles. As suggested by previous work [3], we applied the Relative Variation Distance (RVD) to evaluate whether the model captures heterogeneity within the data. As shown in Table 3, under the seen setting, GSAT-RAG significantly outperforms the baseline methods in terms of the RVD metric.

Table 3. Heterogeneity evaluation with RVD

Method	RVD↓
GSAT-RAG (seen)	**0.7850**
BLEEP (seen)	1.313
Stem (seen)	1.114

2.5. Implementation details

GSAT: The hidden dimension d_c is set to 768. The number of attention heads is 12. The number of encoder layers is 12, and the number of decoder layers is 1. The batch size

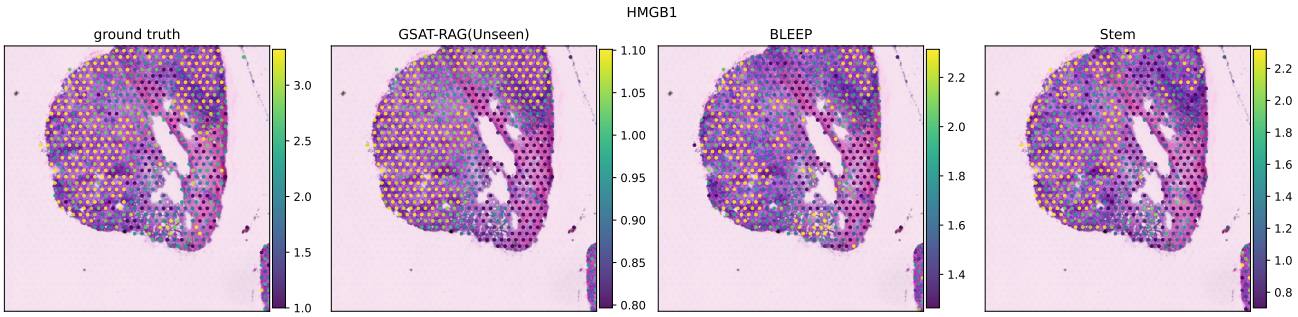

Figure 3. Expressions for oncogene HMGB1. Note that predictions from GSAT-RAG are in a zero-shot setting but those from BLEEP and Stem are in a full-shot setting.

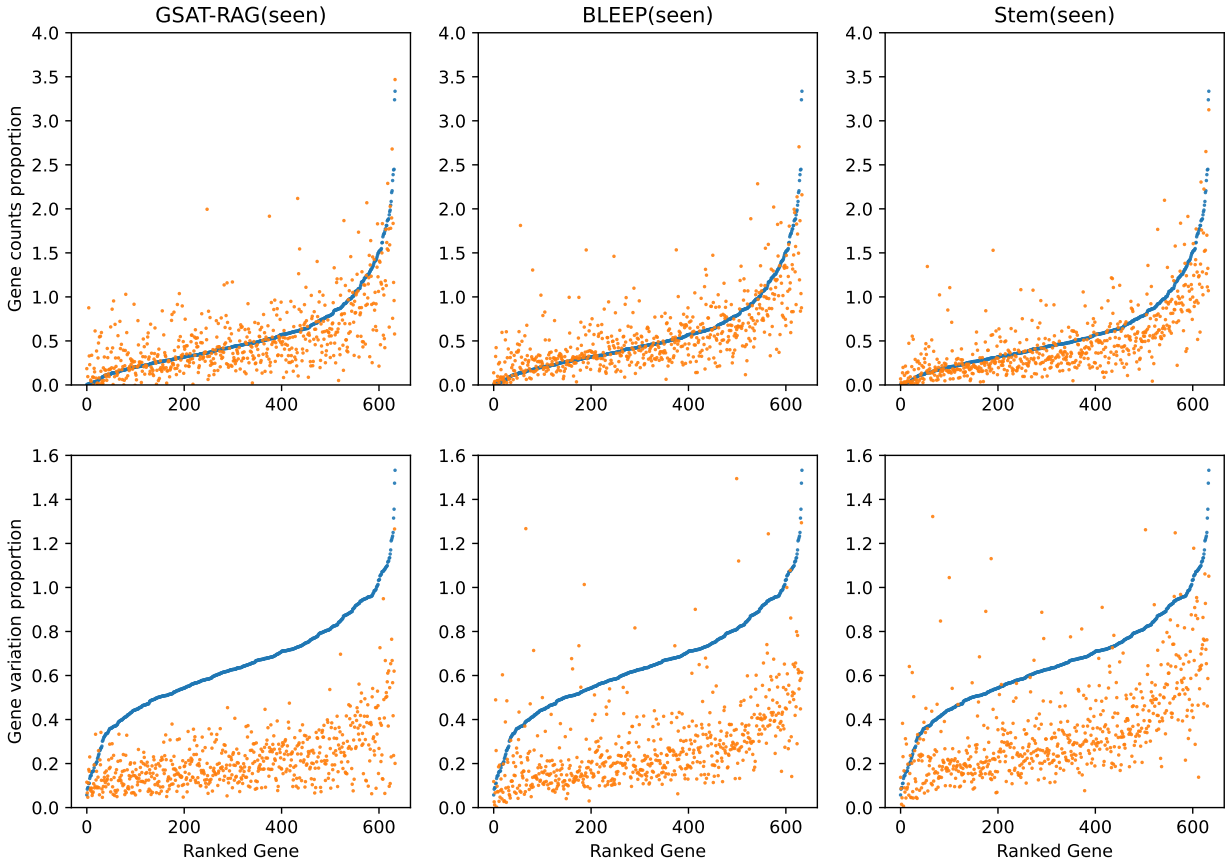

Figure 4. Predicted (Pred) expression profiles are compared with ground truth (GT) expression profiles, normalized either by the average gene count (first row) or by the variance of gene count (second row).

is 128, and the number of pairs (K) is 4. The learning rate is 1×10^{-5}. GSAT is pre-trained on more than 1,700,000 cells.

GSAT-RAG: For image-gene contrastive learning, the learning rate is 4×10^{-3}, and training is stopped based on the validation loss. Top-K_G is set to 50 for generating gene prompts. For generation training, the learning rate is 4×10^{-4}, and training is stopped based on the average PCC score on the validation set. The order of input genes is shuffled in each epoch.

References

[1] Kamolporn Amornsupak, Suyanee Thongchot, Chanida Thinyakul, Carol Box, Somaieh Hedayat, Peti Thuwajit, Suzanne A Eccles, and Chanitra Thuwajit. Hmgb1 mediates invasion and pd-l1 expression through rage-pi3k/akt signaling pathway in mda-mb-231 breast cancer cells. *BMC cancer*, 22 (1):578, 2022. 1

[2] 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. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023. 1

[3] Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *The Thirteenth International Conference on Learning Representations*, 2025. 2