VAST: VERIFIED ASYMMETRIC SUPERVISION TRANSFER FOR HISTOLOGY-BASED SPATIAL TRANSCRIPTOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting spatial gene expression from routine histology offers a scalable complement to spatial transcriptomics, yet most models optimize only a small evaluation panel and discard measurements for thousands of other genes. Although recent work uses these genes as auxiliary supervision, it remains unclear whether enlarging the supervised panel consistently benefits the target genes. We study this question as *gene-supervision scaling*. Our analysis separates average improvement into gene-level positive and negative transfer and indicates that scaling behavior varies across architectures: adding supervision can improve the aggregate score while degrading a nontrivial subset of target genes. Existing auxiliary-supervision methods give auxiliary labels direct optimization access to the target predictor. Our results challenge the premise behind this design: more supervision does not necessarily provide more useful information for every target. We argue that the missing component is a mechanism that controls whether newly added supervision is allowed to modify the final target predictor. We therefore formulate auxiliary-gene learning as asymmetric supervision transfer and introduce VAST. For each target gene, VAST first establishes and freezes a target-only predictor as a panel-invariant reference, lets expanded supervision propose a correction to that reference, and applies the correction only to the extent supported by pre-update, slide-level evidence computed exclusively from the training set. This two-stage *protect–propose–verify* interface implements a simple principle: auxiliary supervision should modify target predictions only through a rejectable transfer interface.

1 INTRODUCTION

Spatial transcriptomics (ST) measures gene expression while preserving tissue coordinates, enabling molecular variation to be studied in its morphological context (Ståhl et al., 2016). This spatial information is valuable for characterizing tissue organization and disease, but ST assays remain costly, specialized, and less widely available than routine hematoxylin-and-eosin (H&E) histology. Predicting spatial expression from H&E images therefore offers a potentially scalable complement to direct molecular measurement.

Histology-based ST prediction has progressed from patch-level regression to models that incorporate spatial interactions, multi-resolution context, contrastive alignment, and whole-slide generation (He et al., 2020; Zeng et al., 2022; Xie et al., 2023; Chung et al., 2024; Huang et al., 2025). Resources such as HEST-1k further support evaluation across tissues and technologies (Jaume et al., 2024). Most pipelines, however, supervise only the small target panel used for evaluation even though an ST assay may measure thousands of additional genes. HE2Gene incorporates non-target genes through multi-task learning (Chen et al., 2024), while Auxiliary Gene Learning (AGL) treats them as auxiliary tasks and learns a differentiable top-k selection (Shiku et al., 2026). Both approaches make use of otherwise discarded labels and motivate a broader question: *does expanding supervision consistently improve the fixed target genes?*

We systematically examine this premise through *gene-supervision scaling*: the supervised output panel is expanded while the evaluated target genes remain fixed. Target and auxiliary losses are averaged within their groups and weighted equally, so adding genes does not dilute the target loss.

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

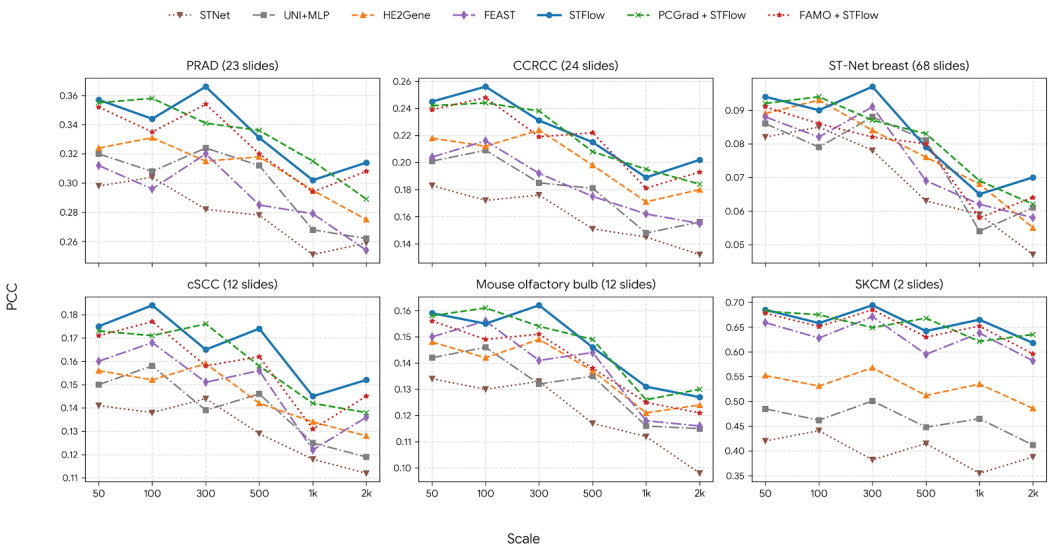

Figure 1: Gene-supervision scaling is irregular across architectures. Mean PCC is computed on the same 50 target genes while the supervised panel grows from 50 to 2,000 genes. Expanded runs use fixed 1:1 target/auxiliary group loss. The evaluated encoders and gradient-balancing methods do not make scaling monotonic.

Our observation is that the prevailing intuition is incomplete. Scaling is irregular and non-monotonic across architectures, cohorts, panels, and random seeds (Figure 1); more supervision can improve the mean score while degrading a nontrivial subset of target genes. PCGrad and FAMO do not consistently improve this behavior in our experiments, and negative transfer remains common in the evaluated settings. These findings suggest that the central problem is not simply how to optimize a larger collection of labels, but whether every additional label should be permitted to influence the target predictor at all.

This result directs attention from gradient correction to optimization access. Despite their different designs, HE2Gene and AGL share the same underlying assumption: auxiliary labels have direct optimization access to the target predictor. Target and auxiliary objectives jointly update a shared representation, so every selected auxiliary label is granted the ability to change the predictor ultimately evaluated on the target genes. Our results challenge that assumption. More supervision is not equivalent to more useful information, and even a biologically plausible auxiliary gene may help some targets while harming others. What is missing is a mechanism that decides whether newly added supervision has the right to modify the final target predictor.

We consequently adopt the following principle: *auxiliary supervision should modify target predictions only through a rejectable transfer interface*. We formulate large-scale gene supervision as an *asymmetric supervision-transfer* problem and introduce VAST through a **protect–propose–verify** (PPV) interface. The target predictor is first established without auxiliary labels and frozen as a panel-invariant reference. A separate expanded-supervision pathway may then propose a correction, but that correction is transferred only to the extent justified by training-only evidence across biological slides. Freezing is not merely an optimization convenience: rejection recovers the same target-only predictor at every supervision scale, making the proposal an inspectable departure from a stable counterfactual. In short, auxiliary labels may propose a change, but they do not receive unconditional optimization access to the target predictor.

Our contributions are:

- We formulate auxiliary-gene learning as a gene-supervision scaling problem and evaluate transfer at the level of fixed target genes rather than relying only on an aggregate score.

- We identify positive and negative transfer as coexisting outcomes of supervision scaling and show that PCGrad and FAMO do not consistently resolve the resulting negative transfer in our experiments.
- We introduce asymmetric supervision transfer and instantiate it with verified residual transfer: a panel-invariant target-only anchor, an auxiliary-informed correction, and a training-only, slide-level acceptance mechanism that makes each proposed transfer rejectable.

2 Related Work

Histology-based spatial expression prediction. ST-Net established patch-level prediction of spatial expression from histology (He et al., 2020). Later work adds transformers and spatial context (Pang et al., 2021; Zeng et al., 2022), contrastive retrieval (Xie et al., 2023), multi-scale context (Chung et al., 2024), and slide-level flow matching (Huang et al., 2025). HEST-1k supports broader cross-tissue evaluation (Jaume et al., 2024); recent benchmarking finds no single method dominates all criteria (Wang et al., 2025). Whereas these studies improve architectures or objectives, we ask when additional gene labels should be allowed to alter a fixed target predictor.

Auxiliary gene supervision. HE2Gene jointly predicts target and non-target genes (Chen et al., 2024); AGL selects auxiliary tasks through prior-guided differentiable top-k optimization (Shiku et al., 2026). Both grant admitted auxiliary losses direct access to target-predictor parameters. Selection asks *which genes to supervise*, but not whether their information should modify each target. VAST separates these questions by placing auxiliary effects behind a rejectable transfer interface.

Multi-task learning and control of transfer. Multi-task learning balances sharing against negative transfer (Caruana, 1997). GradNorm, uncertainty weighting, multi-objective optimization, PCGrad, and FAMO regulate task weights or gradients (Chen et al., 2018; Kendall et al., 2018; Sener & Koltun, 2018; Yu et al., 2020; Liu et al., 2023); cross-stitch networks, MMoE, and TaskExpert regulate where features are shared (Misra et al., 2016; Ma et al., 2018; Ye & Xu, 2023). These approaches control *how* or *where* included tasks update a model, but do not establish a fixed target-only prediction against which an auxiliary update can be rejected. VAST instead separates the protected target representation from the auxiliary proposal and verifies transfer before applying it.

3 Problem Setup: Gene-Supervision Scaling

Let x_i be the histological observation at location i, $\mathcal{T}$ the fixed evaluation targets, $\mathcal{A}$ the auxiliary genes used only during training, and $\mathcal{G} = \mathcal{T} \cup \mathcal{A}$ the supervised panel.

A conventional multi-output model minimizes

$$\mathcal{L}_{\mathrm{sym}}(\theta) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \ell(f_g(x;\theta), y_g)\,, \tag{1}$$

where every gene updates the shared parameters θ. Gene-supervision scaling increases $|\mathcal{G}|$ while fixing $\mathcal{T}$, data splits, preprocessing, capacity, and evaluation. Target-gene transfer is

$$\Delta_t(s_1 \to s_2) = M_t(s_2) - M_t(s_1), \tag{2}$$

where M_t is a held-out metric; its sign denotes positive or negative transfer. Gene-level changes reveal effects that can cancel in the mean.

We instead seek an asymmetric, inspectable interface between a protected target prediction and any update proposed by auxiliary supervision.

4 VAST

4.1 Overview: Protect, Propose, Verify

VAST first learns a target-only anchor, then freezes it while an expanded-panel pathway proposes a correction verified with training-slide evidence. For target t,

$$\boxed{\hat{y}_t(x) = b_t(x) + \alpha_t c_t(x)}, \qquad 0 \le \alpha_t \le 1. \tag{3}$$

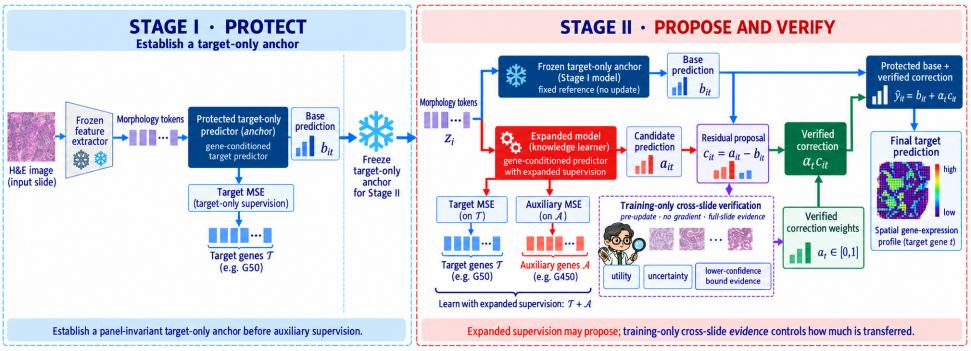

Figure 2: VAST casts auxiliary supervision as a verified residual correction to a frozen target-only anchor. A training-only evidence pass estimates the gene-specific transfer weight α_t before each epoch.

Here b_t is the frozen complete predictor, the expanded-panel candidate a_t induces $c_t = a_t - b_t$, and α_t is estimated from training-only utility and cross-slide uncertainty. Thus $\alpha_t = 0$ always recovers the same target-only counterfactual, independent of panel size. Freezing is essential to this interpretation: if anchor and proposal were updated together, the reference defining c_t would drift with auxiliary supervision. The resulting coefficient is therefore not a freely learned soft gate between two equal-status predictors; it controls a proposed departure from a fixed target-only predictor.

4.2 Stage I—Protect: Establishing a Target-Only Anchor

For morphology tokens z_i and learned target query q_t, the protected encoder produces

$$h_{it}^P = P_{\theta_P}(z_i, q_t), \qquad b_{it} = w_t^\top h_{it}^P + d_t. \tag{4}$$

We train it using target labels only,

$$\mathcal{L}_P = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \ell(b_t, y_t), \tag{5}$$

select it on internal validation, and freeze θ_P for Stage II:

$$\nabla_{\theta_P} \mathcal{L}_{\text{aux}} = 0. \tag{6}$$

The complete base prediction is identical across paired supervision scales for the same fold and seed and remains valid if all corrections are rejected. Gene conditioning occurs before scalar readout, so targets need not share an identical representation differentiated only by the final linear weight. The protected pathway is a complete predictor, not merely a private feature adapter.

4.3 Stage II—Propose: Corrections from Expanded Supervision

A knowledge learner K_{θ_K} receives the same context and is trained on the full panel:

$$h_i^K = K_{\theta_K}(z_i), \qquad a_{it} = H_t^K(h_i^K). \tag{7}$$

The candidate defines $c_{it} = a_{it} - b_{it}$; it is not an equally privileged prediction branch. This residual makes the transfer decision explicit: should expanded supervision move the protected estimate in the proposed direction, and if so, by how much?

4.4 Verify: Training-Only Correction Evidence

A learned gate may mix predictors without evidence that the added pathway has marginal value. VAST instead evaluates the proposed correction directly. Before each epoch e, an evaluation-mode,

no-gradient full-slide pass over the outer-training data estimates whether it improves the base. For slide s and target t,

$$u_{ts}^{(e)} = \frac{1}{|s|} \sum_{i \in s} \big(y_{it} - b_{it}\big) c_{it}, \tag{8}$$

$$q_{ts}^{(e)} = \frac{1}{|s|} \sum_{i \in s} c_{it}^2. \tag{9}$$

Under squared error, u_{ts} measures directional utility and q_{ts} correction energy. Computing evidence per biological slide makes uncertainty reflect cross-slide heterogeneity rather than random minibatch variation.

We maintain exponential moving estimates of utility mean $m_t^{(e)}$, second moment $r_t^{(e)}$, and energy $q_t^{(e)}$ with decay $\rho = 0.95$. Let

$$\sigma_t^{(e)} = \sqrt{\frac{\max\Big(0, r_t^{(e)} - (m_t^{(e)})^2\Big) + \epsilon_{\text{var}}}{n_{\text{eff}}}}, \qquad \epsilon_{\text{var}} = 10^{-12}. \tag{10}$$

We then form

$$\text{LCB}_t^{(e)} = \max\Big(0, m_t^{(e)} - \lambda \sigma_t^{(e)}\Big) \tag{11}$$

and set

$$\alpha_t^{(e)} = \text{clip}_{[0,1]} \left(\frac{\text{LCB}_t^{(e)}}{q_t^{(e)} + \epsilon_{\text{energy}}} \right), \qquad \epsilon_{\text{energy}} = 10^{-8}. \tag{12}$$

We use $\lambda = 1.0$ for full LCB verification and $\lambda = 0$ in the no-uncertainty ablation. $\alpha_t^{(e)}$ is fixed within epoch e and the evidence pass is never backpropagated. Transfer uses a 3-epoch warm-up followed by a 7-epoch ramp. Validation is used only for early stopping; this is a pre-update, training-only estimator rather than independent cross-fitting. Consequently, verification cannot tune itself through validation or test labels.

4.5 Two-Stage Objective and Gradient Topology

After Stage I freezes θ_P, Stage II trains the knowledge learner on targets and auxiliaries,

$$\mathcal{L}_{K,T} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \ell(a_t, y_t), \tag{13}$$

$$\mathcal{L}_{K,A} = \frac{1}{|\mathcal{A}|} \sum_{g \in \mathcal{A}} \ell(a_g, y_g). \tag{14}$$

using

$$\mathcal{L}_{\text{Stage II}} = \mathcal{L}_{K,T} + \lambda_A \mathcal{L}_{K,A}, \tag{15}$$

while θ_P remains fixed. Panel-dependent learning is therefore confined to the candidate and verified residual. Equation 12 controls forward transfer and is not optimized as an unconstrained mixture gate. Thus isolation from auxiliary gradients is enforced structurally, rather than depending on task gradients remaining aligned during joint optimization.

5 Experiments

5.1 Experimental Setup

Benchmarks. We use 15 datasets from HEST-1k (nine cancer cohorts: CCRCC, COAD, IDC, LUNG, LYMPH-IDC, PAAD, PRAD, READ, and SKCM) and STimage (six cohorts; 151 slides in total). We preserve slide-disjoint outer folds and reserve 20% of each outer-training partition for validation. Folds with only one training slide use a fixed 100-epoch schedule without model selection.

Gene panels and preprocessing. All experiments evaluate the same 50 targets ($G50$). For the main comparison, VAST uses a prespecified $G500$ supervision panel without panel-size search or additional hyperparameter tuning; when $G500$ is unavailable, we use the largest valid panel (G400 COAD; G300 IDC/SKCM; G250 LUNG; G150 PAAD). These nested highly-variable-gene panels are selected from outer-training data only. Figure 3 is a separate scaling analysis that sweeps the available panel sizes from $G50$ through $G5000$ and does not determine the VAST configuration reported in the main table. Expression is $\log(1 + x)$ transformed and standardized with training statistics. Inputs combine UNI-v2 patch embeddings, 16 spatial neighbors, and eight same-slide morphology-retrieval neighbors.

Training and model selection. Unless a published optimizer is required, we use AdamW (learning rate and weight decay 10^{-4}), batch size 128, and gradient clipping 5.0. The VAST knowledge learner has width 256, two blocks, eight heads, three experts, and dropout 0.1. Stage I trains for at most 100 epochs (patience 10); frozen-anchor Stage II trains for at most 200 (patience 20) with equal target/auxiliary loss weights. The routing-loss weight is 0, so α is determined entirely by the evidence rule without an explicit routing supervision loss. Verification uses the train-preupdate full-slide pass described above. We report five seeds (42–46). At each supervision scale, internal validation selects the training checkpoint and controls early stopping; test PCC is computed only for that validation-selected checkpoint. VAST uses $G500$ where available and otherwise the largest valid panel described above, without panel-size search. For each standard baseline, every eligible candidate panel is trained explicitly. Scaling is implemented by changing the number of genes in the output prediction head while keeping the backbone, inputs, and training protocol fixed. Validation selects the checkpoint and panel size; the selected configuration is evaluated once on the test set. Thus test performance is not used for selection. The supplementary material reports the selected panel for every dataset.

Evaluation and aggregation. For each outer fold, we compute PCC per target gene on concatenated test spots, average across genes, folds, and seeds, and report the slide-weighted benchmark average

$$\mathrm{PCC}_{\mathrm{bench}} = \frac{\sum_{d\in\mathcal{D}} n_d \,\mathrm{PCC}_d}{\sum_{d\in\mathcal{D}} n_d}, \tag{16}$$

where n_d is the number of slides in dataset d.

5.2 Main Comparison

Table 1 reports the main comparison on the fixed $G50$ evaluation targets. VAST uses the prespecified $G500$ training panel (or the largest valid panel listed above) with no panel-size or additional hyperparameter search. Every eligible baseline panel is trained, and internal validation selects its checkpoint and panel size before one test evaluation. The selected configurations are disclosed in the supplementary material. Subscripts denote standard deviation across independent runs.

Overall, VAST is competitive with the compared methods on both benchmark suites. It obtains the highest slide-weighted PCC in the table on HEST-1k (0.349 versus 0.330 for STFlow and 0.318 for TRIPLEX) and on STimage (0.143 versus 0.133 for AGL+DkGSB and 0.128 for STFlow). The advantage is more apparent on several larger cohorts, including CCRCC and HER2+ breast, whereas margins on smaller cohorts are generally narrower or mixed. This data-size pattern suggests that, with more training samples, VAST can learn more useful cross-gene knowledge for its proposal and verification pathways. On limited cohorts, the available evidence may support fewer reliable transfers, so the benefit over strong baselines is less pronounced.

5.3 Scaling Behavior

Figure 3 examines how VAST behaves as the supervised panel grows. Across all scales, the models share the same backbone architecture and only expand the output heads. Under $G50$ supervision, VAST has the lowest PCC among the compared models on these cohorts. Most of its observed improvement appears by $G100$, after auxiliary supervision is introduced.

As the supervision panel grows to larger sizes ($G300$ through $G5000$), the jointly optimized baselines in this comparison show more pronounced degradation, whereas VAST remains comparatively stable

Table 1: Dataset-level comparison on the fixed 50 target genes. Entries report mean PCC with run standard deviation in the subscript. VAST uses a fixed $G500$ training panel where available, without panel-size search. Baseline panel sizes and checkpoints are selected by validation PCC before test evaluation; the selected panel for each cohort is reported in the supplementary material. Averages are slide-weighted (Equation 16); N/A denotes unsupported internal-validation splits.

Benchmark	Dataset	STNet	BLEEP	TRIPLEX	FEAST	AGL+DkGSB	STFlow	VAST
HEST-1k	IDC (4)	$0.412_{.005}$	$0.518_{.006}$	$0.572_{.005}$	$0.531_{.011}$	$0.575_{.002}$	$0.559_{.009}$	$\mathbf{0.582}_{.002}$
	PRAD (23)	$0.298_{.004}$	$0.324_{.005}$	$0.344_{.008}$	$0.312_{.014}$	$\mathbf{0.369}_{.008}$	$0.357_{.012}$	$0.368_{.014}$
	PAAD (3)	$0.345_{.006}$	$0.402_{.007}$	$0.428_{.006}$	$0.436_{.027}$	$0.376_{.003}$	$0.436_{.008}$	$\mathbf{0.449}_{.008}$
	SKCM (2)	$0.420_{.011}$	$0.552_{.014}$	$0.638_{.008}$	$0.629_{.018}$	N/A	$\mathbf{0.652}_{.006}$	$0.640_{.015}$
	COAD (4)	$0.225_{.008}$	$0.274_{.009}$	$0.286_{.007}$	$0.265_{.010}$	$0.260_{.012}$	$0.298_{.012}$	$\mathbf{0.308}_{.007}$
	READ (4)	$0.112_{.007}$	$0.185_{.008}$	$0.198_{.009}$	$0.173_{.026}$	$0.196_{.004}$	$\mathbf{0.208}_{.012}$	$0.201_{.007}$
	CCRCC (24)	$0.183_{.005}$	$0.218_{.006}$	$0.229_{.007}$	$0.204_{.012}$	$0.242_{.003}$	$0.245_{.016}$	$\mathbf{0.289}_{.009}$
	LUNG (2)	$0.486_{.009}$	$0.521_{.008}$	$0.528_{.006}$	$0.516_{.015}$	N/A	$\mathbf{0.536}_{.004}$	$0.528_{.008}$
	LYMPH-IDC (4)	$0.210_{.005}$	$0.226_{.007}$	$0.252_{.005}$	$0.240_{.009}$	$0.256_{.003}$	$\mathbf{0.263}_{.007}$	$0.254_{.003}$
	Slide-weighted average	0.256	0.298	0.318	0.293	N/A	0.330	**0.349**
STimage	ST-Net breast (68)	$0.082_{.005}$	$0.089_{.004}$	$0.091_{.007}$	$0.088_{.008}$	$0.113_{.003}$	$0.094_{.009}$	$\mathbf{0.114}_{.003}$
	HER2+ breast (36)	$0.165_{.004}$	$0.182_{.005}$	$0.194_{.006}$	$0.187_{.009}$	$0.189_{.002}$	$0.198_{.006}$	$\mathbf{0.219}_{.005}$
	Developing heart (19)	$0.021_{.003}$	$0.028_{.004}$	$0.035_{.003}$	$0.031_{.006}$	$0.029_{.005}$	$0.037_{.001}$	$\mathbf{0.048}_{.005}$
	cSCC (12)	$0.141_{.006}$	$0.156_{.007}$	$0.168_{.008}$	$0.160_{.011}$	$\mathbf{0.177}_{.004}$	$0.175_{.010}$	$0.158_{.007}$
	Mouse olfactory bulb (12)	$0.134_{.005}$	$0.148_{.004}$	$0.155_{.005}$	$0.150_{.007}$	$0.158_{.003}$	$0.159_{.006}$	$\mathbf{0.171}_{.003}$
	Ståhl breast (4)	$0.205_{.008}$	$0.231_{.006}$	$0.254_{.007}$	$0.242_{.012}$	$0.236_{.003}$	$0.259_{.006}$	$\mathbf{0.268}_{.007}$
	Slide-weighted average	0.106	0.117	0.124	0.119	0.133	0.128	**0.143**

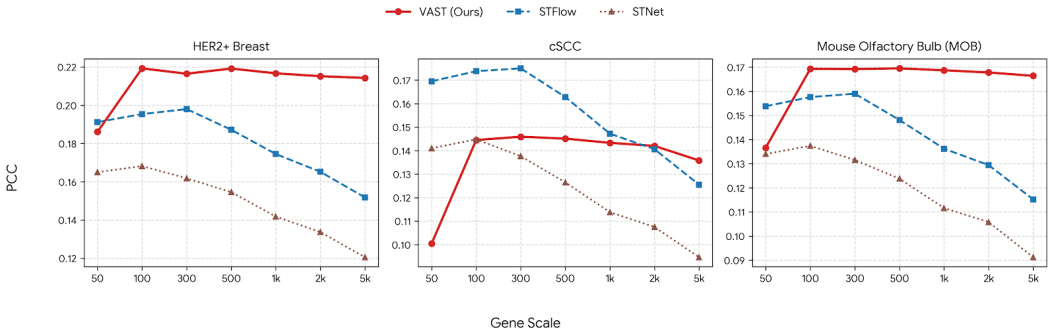

Figure 3: Supervision-scale sweep from $G50$ to $G5000$ on three STimage cohorts. PCC is always computed on the same 50 target genes, and curves average five seeds. This sweep is separate from the fixed-$G500$ VAST configuration in Table 1.

in the plotted cohorts. This pattern is compatible with reduced interference from the Protect–Propose–Verify (PPV) mechanism.

5.4 Ablation Study

Table 2 systematically evaluates the components of the protect–propose–verify interface on three STimage and two HEST-1k cohorts under matched backbones, prediction heads, and target gene panels. To disentangle architectural overhead and supervision scale from the verification mechanism, we establish two controls followed by four architectural variants: (A) *Target-only* retains the identical two-stage architecture and optimization schedule as VAST but trains both stages without any auxiliary labels, strictly controlling for two-stage capacity and training budget; (B) *Expanded Joint* is a single-stage model employing only the Stage-I architecture trained jointly across target and auxiliary labels, verifying whether gains stem merely from expanded multi-task supervision; (C) *VAST w/o Verify* retains anchor protection and the Stage-II proposal network but fixes $\alpha_t \equiv 1$; (D) *Learned Gate* replaces evidence verification with a parameterized mixing gate; (E) *w/o Uncertainty* evaluates evidence-based utility but omits the uncertainty penalty ($\lambda = 0$); and (F) *Full VAST* applies the full LCB-penalized evidence verification ($\lambda = 1$).

As shown in Table 2, the two-stage capacity control (0.2335) and the single-stage expanded joint model (0.2401) both underperform all variants equipped with the protect-and-propose mechanism. Unselectively accepting all auxiliary proposals (Variant C) improves the macro-average PCC to 0.2605; adopting a parameterized learned gate (Variant D) further lifts it to 0.2690, while calibrating integration solely via empirical evidence utility (Variant E, $\lambda = 0$) achieves 0.2765. Finally, incorporating the full LCB uncertainty penalty enables Full VAST (Variant F) to attain the highest macro-average PCC of 0.2840. This stepwise progression demonstrates that, with model capacity, training budget, and target gene panels strictly controlled, performance gains cannot be explained by two-stage scaling or expanded supervision alone; protected anchors, auxiliary proposals, and conservative evidence-based calibration each offer distinct, measurable contributions to the final outcome. In conjunction with our subsequent negative transfer analyses, these findings substantiate that an uncertainty-aware verification mechanism is critical for suppressing unreliable auxiliary transfers.

Table 2: Capacity-controlled ablation of the protect–propose–verify interface on three STimage and two HEST-1k cohorts. Entries are mean target-gene PCC, with the standard deviation over five seeds in the subscript; Macro Avg. is the unweighted mean across cohorts.

Variant	HER2	cSCC	MOB	CCRCC	IDC	Macro Avg.
A. Target-only	$0.1861_{.005}$	$0.1124_{.007}$	$0.1366_{.005}$	$0.2182_{.008}$	$0.5140_{.006}$	0.2335
B. Expanded Joint	$0.1924_{.005}$	$0.1205_{.008}$	$0.1415_{.006}$	$0.2245_{.010}$	$0.5218_{.007}$	0.2401
C. VAST w/o Verify	$0.2035_{.004}$	$0.1386_{.006}$	$0.1538_{.004}$	$0.2536_{.007}$	$0.5532_{.004}$	0.2605
D. Learned Gate	$0.2081_{.005}$	$0.1458_{.006}$	$0.1587_{.005}$	$0.2678_{.008}$	$0.5645_{.003}$	0.2690
E. w/o Uncertainty	$0.2136_{.004}$	$0.1520_{.005}$	$0.1638_{.004}$	$0.2794_{.006}$	$0.5739_{.003}$	0.2765
F. Full VAST	$\mathbf{0.2192}_{.005}$	$\mathbf{0.1582}_{.007}$	$\mathbf{0.1712}_{.003}$	$\mathbf{0.2891}_{.009}$	$\mathbf{0.5824}_{.002}$	**0.2840**

5.5 Inner Cross-Fitting Verification Analysis

We further assess sample independence using inner cross-fitting on four larger cohorts (Table 3). Outer test folds remain frozen for generalization assessment. Within each outer-training partition, inner cross-fitting estimates proposals on one slide split and verification residuals on the other, providing out-of-fold evidence without self-verification.

Because nested fitting is costly and unstable for very small cohorts, we use it as a stress test rather than a required protocol. Both verification regimes improve PCC and reduce observed negative transfer relative to the matched $\alpha = 1$ control; standard training-evidence VAST is comparable here without nested refitting and remains our default configuration.

Table 3: Outer-CV test performance under unverified ($\alpha = 1$), inner cross-fit, and standard training-only verification across the four larger cohorts ($G50$ target panel, five random seeds). Outer folds evaluate final generalization, while inner cross-fitting produces sample-independent verification evidence within the training folds.

Dataset	Method	Positive Runs	Test PCC	Negative Transfer Rate ↓
ST-Net breast (68)	No Verify	2/5	$0.098_{.006}$	48.0%
	Cross-fit VAST	**5/5**	$0.108_{.004}$	26.0%
	Current VAST	**5/5**	$\mathbf{0.114}_{.003}$	**18.0%**
HER2+ breast (36)	No Verify	3/5	$0.203_{.004}$	42.0%
	Cross-fit VAST	**4/5**	$0.212_{.006}$	28.0%
	Current VAST	**5/5**	$\mathbf{0.219}_{.005}$	**20.0%**
CCRCC (24)	No Verify	2/5	$0.254_{.007}$	46.0%
	Cross-fit VAST	**5/5**	$0.282_{.008}$	24.0%
	Current VAST	**5/5**	$\mathbf{0.289}_{.009}$	**18.0%**
PRAD (23)	No Verify	2/5	$0.339_{.011}$	48.0%
	Cross-fit VAST	**4/5**	$0.361_{.010}$	26.0%
	Current VAST	**5/5**	$\mathbf{0.368}_{.014}$	**22.0%**

5.6 TRANSFER-WEIGHT STRATIFICATION

At $G500$, we set $\alpha = 1$ in a matched control so every proposed correction is applied in full. Against the model using verified α, we define

$$\Delta\mathrm{PCC} = \mathrm{PCC}_{\text{verified }\alpha} - \mathrm{PCC}_{\alpha=1}. \tag{17}$$

Positive $\Delta\mathrm{PCC}$ indicates that the coefficient blocks an erroneous correction relative to accepting it in full. Table 4 stratifies genes by verified α: the lowest quartile has the largest mean gain in all four cohorts, while higher quartiles show smaller and occasionally negative changes. This is consistent with useful attenuation, but does not guarantee every decision.

Table 4: Verification gain at $G500$, stratified by quartile of the learned transfer coefficient α; values are gene-level mean $\Delta\mathrm{PCC}$ within each quartile.

Original α Quartile	ST-Net breast	HER2+ breast	CCRCC	PRAD	Mean
Q1 (lowest α)	$\mathbf{+0.0162}$	$\mathbf{+0.0135}$	$\mathbf{+0.0179}$	$\mathbf{+0.0151}$	$\mathbf{+0.0157}$
Q2	$+0.0051$	$+0.0032$	$+0.0042$	$+0.0046$	$+0.0043$
Q3	$+0.0014$	$+0.0029$	$+0.0028$	-0.0008	$+0.0016$
Q4 (highest α)	$+0.0021$	-0.0006	$+0.0016$	$+0.0019$	$+0.0013$

5.7 GENE-LEVEL TRANSFER

Table 5 examines gene-level behavior when expanding supervision from $G50$ to each dataset's largest valid panel, up to $G500$, across all 15 datasets, including nine HEST-1k cancer cohorts and six STimage cohorts. The exact endpoint for each cohort is reported in the supplementary material. Across all 75 dataset–seed runs, VAST achieves a positive aggregate change in 71 runs, whereas the compared baselines exhibit positive changes substantially less frequently. Its individual-gene negative-transfer rate (NTR) is reduced to 27.6%, outperforming all competing multi-task baselines listed in the table. Furthermore, VAST yields an overall mean $\Delta\mathrm{PCC}$ improvement of $+0.0245$. These results confirm that across diverse cancer cohorts and tissue types, VAST's asymmetric and verified transfer mechanism robustly mitigates negative transfer and delivers dependable supervisory gains from expanded gene panels.

Table 5: Gene-level robustness when expanding supervision from $G50$ to each dataset's largest valid panel (up to $G500$) across all 15 benchmark datasets.

Model	Positive seeds	NTR ↓	Mean ΔPCC ↑
VAST	**71/75**	**27.6%**	**+0.0245**
STFlow	31/75	58.4%	-0.0051
BLEEP	33/75	57.2%	-0.0042
STNet	14/75	65.8%	-0.0125

6 DISCUSSION AND LIMITATIONS

To our knowledge, this is the first systematic study of supervision-panel size in histology-based spatial-expression prediction. Extensive experiments motivate VAST's paradigm of separated target representation and evidence-based, rejectable transfer. Open community directions include selecting useful auxiliary genes and studying supervision-label quality, noise, and availability.

7 CONCLUSION

Gene-supervision scaling reveals that larger panels can help and harm different targets. VAST protects a target-only predictor and admits corrections according to training evidence. Results across 15 datasets and complementary analyses support this alternative and establish supervision scaling as an evaluation axis for trustworthy learning from large panels.

Reproducibility Statement

Section 5 specifies the data splits, fixed evaluation targets, validation-only model selection, architectures, optimization settings, and random seeds. Additional panel configurations, theoretical assumptions, ablations, and complete scaling curves are provided in the supplementary material. These details are intended to support independent reproduction of the reported analyses.

Ethics Statement

This study analyzes previously released, de-identified spatial-transcriptomics and histology datasets and collects no new human-subject data. We use these data for methodological benchmarking and do not make patient-level clinical claims.

AI Use Statement

Generative AI tools assisted with refining the conceptual framing, mathematical formulation and proof presentation, result interpretation, manuscript drafting and editing, and the formatting of code, tables, and figures. The authors reviewed the AI-assisted material, checked the mathematical and empirical claims against the underlying derivations and experimental outputs, and take responsibility for the final content of the submission.

References

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

Xingjian Chen, Jiecong Lin, Yuchen Wang, Weitong Zhang, Weidun Xie, Zetian Zheng, and Ka-Chun Wong. HE2Gene: image-to-RNA translation via multi-task learning for spatial transcriptomics data. *Bioinformatics*, 40(6):btae343, 2024. doi: 10.1093/bioinformatics/btae343.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018.

Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11591–11600, 2024.

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*, 4:827–834, 2020.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Zhitao Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25550–25565, 2025.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast adaptive multitask optimization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1930–1939, 2018. doi: 10.1145/3219819.3220007.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3994–4003, 2016.

Minxing Pang, Kangning Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021. doi: 10.1101/2021.11.28.470212.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kaito Shiku, Kazuya Nishimura, Shinnosuke Matsuo, Yasuhiro Kojima, and Ryoma Bise. Auxiliary gene learning: Spatial gene expression estimation by auxiliary gene selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9015–9023, 2026. doi: 10.1609/aaai.v40i11.37857.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016.

Chuhan Wang, Adam S. Chan, Xiaohang Fu, Shila Ghazanfar, Jinman Kim, Ellis Patrick, and Jean Y. H. Yang. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nature Communications*, 16:1544, 2025. doi: 10.1038/s41467-025-56618-y.

Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70626–70637, 2023.

Hanrong Ye and Dan Xu. TaskExpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21828–21837, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022.