SPARSE RNA PROMPTS CALIBRATE SPATIAL GENE EXPRESSION FROM HISTOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Histology-based models of spatial gene expression may capture within-slide variation while remaining miscalibrated across donors. We evaluated whether sparse same-slide RNA measurements reduce this cross-donor error and whether any improvement reflects local spatial adaptation or slide-level calibration. A frozen UNI2-h–Ridge predictor was evaluated on 70 slides from 47 donors across nine HEST cohorts. For each slide, 20 deterministic spatial partitions defined immutable target fields of view (FOVs) and disjoint prompt FOVs. We compared a no-prompt baseline with scalar, gene-wise, and low-rank empirical-Bayes offsets. At requested budget $B = 5\%$ (mean achieved fraction, 5.94%; range, 4.97–14.48%), the prespecified low-rank comparison reduced donor-macro variance-normalized log-MSE from 1.441 to 1.239 (absolute difference, -0.202; paired bootstrap 95% interval, $[-0.500, -0.048]$), with improvement for 40 of 47 donors. Gene-wise offsets performed best, reducing normalized log-MSE to 1.200 (-16.7% relative to the no-prompt baseline) and mean Poisson deviance from 31.90 to 26.33 (-17.5%), with improvement for all 47 donors. A scalar offset achieved normalized log-MSE of 1.206 and therefore recovered 97.3% of the gene-wise reduction relative to the no-prompt baseline; the low-rank estimator had 2.7% higher normalized log-MSE than the scalar estimator. Within-slide gene and program correlations were unchanged because each estimator adds a spot-invariant vector. These findings support same-slide calibration of expression levels, but not recovery of donor-specific spatial structure or identification of a uniquely technical source of the residual offset.

1 INTRODUCTION

Spatial transcriptomics (ST) measures gene expression while preserving tissue coordinates, but comprehensive molecular profiling remains costly and assay-dependent (Ståhl et al., 2016; Rao et al., 2021). Histology is routinely acquired at substantially lower marginal cost, motivating models that predict spatial expression from H&E images (He et al., 2020; Schmauch et al., 2020; Pang et al., 2021). Pathology foundation models have improved image-based expression prediction across tissues; nevertheless, a predictor trained on other donors may remain poorly calibrated with respect to slide-level expression baselines that are not identifiable from morphology alone (Chen et al., 2024; Jaume et al., 2024).

A natural extension is sparse molecular acquisition: measure a small set of regions, estimate a correction, and potentially use the corrected model to select subsequent regions. Four preliminary methodological evaluations considered stochastic count models, nuisance-invariant contrasts, moment-based objectives, and spatial residual updates. In each case, a stronger control constrained the supported claim: Ridge regression outperformed the stochastic model, theoretical identifiability of contrasts did not translate into real-data prediction, an apparent acquisition advantage disappeared relative to a zero-prompt baseline, and leakage-controlled spatial updates increased error. Table 1 summarizes these comparisons, and Appendix A provides the corresponding formulations and results. Collectively, they motivate a more fundamental question: *Do sparse RNA measurements improve prediction when prompt and evaluation regions are rigorously separated?*

We address this question using a hierarchy of slide-level calibration estimators. A frozen Ridge model predicts $\log(1 + y)$ from UNI2-h embeddings. Sparse same-slide prompt spots then estimate

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.

Table 1: **Summary of preliminary methodological evaluations motivating the present study.** Each row reports the principal comparison used to determine whether the corresponding formulation warranted further evaluation. AUBC denotes area under the budget curve. Lower AUBC and assignment loss indicate better performance, whereas higher Pearson r indicates better performance.

Formulation	Research Question	Principal Comparison	Implication
Stochastic count model	Can stochastic count transformations outperform a linear image baseline?	Multiscale $r = 0.142$ vs. Ridge $r = 0.414$.	Conditioning, model capacity, and pooling require explicit controls.
Nuisance-invariant contrasts	Can identifiable expression ratios support transfer across gene panels?	Synthetic support result held; real-data contrast loss was 0.703 vs. 0.693 for a constant predictor.	Identifiability does not by itself establish predictive utility.
Moment-based selection	Can unbiased query moments train a region selector?	Learned AUBC was $1.500\,\mathrm{M}$ vs. $1.349\,\mathrm{M}$ with no acquisition.	A zero-prompt baseline is required for acquisition claims.
Spatial residual acquisition	Do leakage-controlled spatial updates improve fixed-target prediction?	Analytic AUBC was 18.56 and random-policy AUBC was 19.42, vs. 14.80 with no acquisition.	Prompt value should be established before optimizing selection.
Slide-level calibration	Do sparse prompts estimate a stable residual baseline?	Gene-wise normalized log-MSE changed by -16.7%, with improvement for 47/47 donors.	This comparison defines the supported claim of the present study.

either a single scalar offset, a gene-specific offset vector, or a low-rank gene-space offset. Each correction is estimated exclusively from prompt spots and applied to a fixed, disjoint target set. This design isolates the value of same-slide molecular information from the separate problems of learning spatially varying residuals and optimizing region selection.

The resulting evidence is positive but narrower than an active-learning or spatial-adaptation claim. At requested budget $B = 5\%$, gene-wise empirical-Bayes calibration reduced donor-macro normalized log-MSE by 16.7% and improved the donor-level mean for all 47 donors. A scalar correction captured nearly all of the mean reduction, whereas the low-rank correction performed worse than both the scalar and gene-wise estimators. Because all three corrections are constant across spots, they cannot alter within-slide spatial rankings. The study therefore supports calibration of absolute expression levels, not recovery of local spatial structure. Moreover, a slide-constant residual may reflect assay capture, tissue composition, donor-wide biology, or a combination of these factors; the retrospective design does not identify a unique mechanism.

Our contributions are:

- We define a fixed-target evaluation protocol in which target and prompt FOVs are disjoint, cross-validation folds are grouped by biological donor, and all calibration estimators and placement policies share identical target masks within each slide-partition replicate.
- Across nine HEST cohorts comprising 70 slides and 47 donors, we show that sparse same-slide calibration reduces variance-normalized log-MSE and mean Poisson deviance. At requested budget $B = 5\%$, the gene-wise estimator improves the donor-level mean for all 47 donors.
- We characterize the boundary of the supported contribution. A scalar shift recovers 97.3% of the best normalized log-MSE reduction, the low-rank estimator provides no additional benefit, and within-slide correlation metrics are mathematically invariant to all evaluated corrections.
- We distinguish requested from achieved acquisition fractions, retain all 56,000 evaluation records, and restrict inference to retrospective HEST development data.

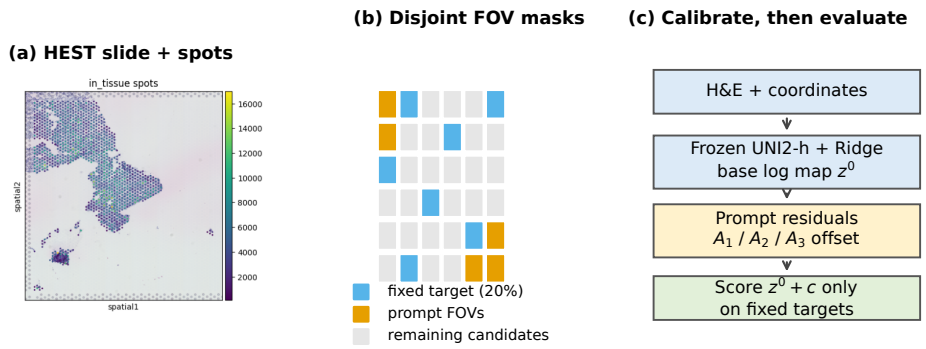

Figure 1: **Dataset example and fixed-target evaluation design.** (a) Representative HEST clear-cell renal cell carcinoma (CCRCC) slide INT1 with spatial transcriptomics spots overlaid on H&E histology. (b) For each of 20 deterministic partitions per slide, target FOVs are selected before prompt FOVs and removed from the prompt candidate set; all estimators and placement policies use the same target mask for a given slide-partition pair. (c) Residuals observed in the prompt FOVs calibrate the frozen histology-based prediction in log space, and all metrics are computed only on the immutable target FOVs. The full evaluation comprises 70 slides from 47 donors; the mask illustration is schematic, and achieved prompt fractions are computed from the selected complete FOVs.

2 Related Work

Histology-to-expression prediction. ST-Net, HE2RNA, and HisToGene established image-conditioned expression prediction (He et al., 2020; Schmauch et al., 2020; Pang et al., 2021). Subsequent systems incorporated bimodal alignment, spatial graphs, and pathology foundation models (Xie et al., 2023; Zeng et al., 2022; Chen et al., 2024). HEST-1k standardizes paired histology, expression matrices, patient-level splits, and foundation-model benchmarks across tissues (Jaume et al., 2024). We retain a frozen image predictor and quantify the incremental information provided by a small number of molecular observations from an otherwise unseen donor.

Count normalization and empirical Bayes. RNA-sequencing workflows model library size, dispersion, and gene-specific uncertainty because between-sample scale differences may reflect both technical and biological variation (Robinson et al., 2010; Love et al., 2014). Empirical-Bayes shrinkage stabilizes high-dimensional estimates when the number of observations per feature is limited (Smyth, 2004; Efron, 2012). The calibration estimators studied here apply the same principle to prediction: outer-training donors determine prior and observation variances, and prompt spots estimate the residual baseline of a new slide.

Spatial experimental design. Bayesian experimental design selects measurements according to expected utility (Chaloner & Verdinelli, 1995; Rainforth et al., 2024). Recent spatial-genomics studies examine tissue-section design, learned FOV selection, single-cell-guided acquisition, and histology-based region prioritization (Jones et al., 2024; Li et al., 2026; Zhu et al., 2026a; Xu et al., 2026). A contemporaneous active-spot benchmark reports unstable policy rankings at small acquisition budgets (Zhu et al., 2026b). The present study evaluates a necessary condition for such methods: a prompt-conditioned predictor should outperform its no-prompt counterpart on a policy-independent target set. We test this condition without fitting a region selector.

3 Problem and Protocol

3.1 Data, predictor, and donor folds

For slide s, let $X_s \in \mathbb{R}^{N_s \times d}$ denote precomputed UNI2-h spot embeddings, $Y_s \in \mathbb{R}_+^{N_s \times G}$ the raw counts for the cohort-specific 50-gene panel, and $S_s \in [0,1]^{N_s \times 2}$ the normalized spatial coordinates.

The present analysis uses the full UNI2-h representation, with $d = 1536$. Within each outer training fold, embedding dimensions are standardized using training data only, and Ridge regression with $\alpha = 100$ is fitted to the log-transformed target $z_{ig} = \log(1 + y_{ig})$. No principal-component projection is applied in this analysis; in particular, the embeddings are not reduced to the 256-dimensional representation used in earlier preliminary evaluations.

Slides are grouped by biological donor. Cohorts with at least four donors use four outer folds, whereas cohorts with two or three donors use leave-one-donor-out folds. All slides from a donor remain in the same fold. The resulting evaluation set contains 70 slides, 47 donors, 232,299 spots, and nine cohorts.

3.2 Fixed targets and sparse prompts

Figure 1 summarizes the retrospective prediction task. Each slide is tessellated into non-overlapping rectangular FOVs containing approximately 50 spots. For each slide and each of 20 deterministic partition indices, a spatially stratified target set is selected from FOV centroids in four quadrants, and the selected target FOVs are removed from the prompt candidate set. All calibration estimators, acquisition budgets, and placement policies use the same target set for a given slide-partition pair. Cryptographic identifiers of the target and candidate index sets are retained to verify equality across comparisons, and any prompt–target overlap invalidates the corresponding evaluation replicate.

Requested budgets are $B \in \{0, 1, 2, 5, 10\}\%$ of all spots on a slide. Because acquisition proceeds in complete FOVs until the requested number of spots is reached, the achieved fraction may exceed the request, particularly for small slides. Mean achieved fractions are 2.85%, 3.41%, 5.94%, and 11.20% for requested budgets of 1%, 2%, 5%, and 10%, respectively. In the primary requested 5% condition, achieved fractions range from 4.97% to 14.48%. We therefore use *requested budget* to denote the experimental condition and report achieved fractions separately. Prompt FOVs are selected either uniformly at random from the candidate set or by deterministic farthest-first coverage of the normalized spatial coordinates.

Target expression is excluded from base-model fitting, prompt placement, and calibration. Histology-derived features at target spots remain available to the frozen predictor, consistent with the intended prediction setting.

4 Calibration Ladder

Let z^0_{ig} denote the frozen Ridge prediction in log space. For prompt set P, define the residual

$$r_{ig} = \log(1 + y_{ig}) - z^0_{ig}, \qquad i \in P. \tag{1}$$

All prior parameters and variance components described below are estimated exclusively from outer-training donors. Calibration is performed in log space; $\exp(\hat{z}) - 1$ is used only for count-scale evaluation.

A_0: no prompt. The reference prediction is $\hat{z}^{(0)}_{ig} = z^0_{ig}$.

A_1: scalar empirical-Bayes offset. Let $\bar{r} = (|P|G)^{-1} \sum_{i \in P} \sum_{g=1}^{G} r_{ig}$. Outer-training donors provide estimates of the between-donor variance τ_0^2 and within-donor variance σ_0^2. The scalar correction is

$$\begin{aligned} w_0 &= \frac{\tau_0^2}{\tau_0^2 + \sigma_0^2/(|P|G)}, \\ \hat{z}^{(1)}_{ig} &= z^0_{ig} + w_0 \bar{r}. \end{aligned} \tag{2}$$

A_2: gene-wise empirical-Bayes offsets. For $\bar{r}_g = |P|^{-1} \sum_{i \in P} r_{ig}$ and outer-training variance estimates τ_g^2 and σ_g^2, the gene-specific correction is

$$w_g = \frac{\tau_g^2}{\tau_g^2 + \sigma_g^2/|P|}, \qquad \hat{z}_{ig}^{(2)} = z_{ig}^0 + w_g \bar{r}_g. \tag{3}$$

A_3: low-rank empirical-Bayes offsets. Let $\bar{r} = (\bar{r}_1, \ldots, \bar{r}_G)^\top$, and let $V_q \in \mathbb{R}^{G\times q}$ and $\mu \in \mathbb{R}^G$ denote the principal-component basis and mean estimated from outer-training residuals. The effective dimension is $q = \min(r, G, N_{\text{train}} - 1, D_{\text{train}} - 1)$, where N_{train} and D_{train} denote the sample and donor counts represented in the training residual matrix. Requested ranks $r \in \{4, 8, 16\}$ are compared using an inner donor split, and the smallest rank within one standard error of the best validation performance is selected. With component-wise shrinkage matrix W_q, the correction is

$$c^{(3)} = V_q W_q V_q^\top (\bar{r} - \mu), \qquad \hat{z}_{ig}^{(3)} = z_{ig}^0 + c_g^{(3)}. \tag{4}$$

Thus, A_3 is low-rank in gene space but remains constant across target locations.

4.1 What the ladder can and cannot measure

Proposition 1 (Spatial-ranking invariance)**.** *For any slide-constant vector $c \in \mathbb{R}^G$, replacing predictions z_i^0 by $z_i^0 + c$ leaves within-slide gene-wise Pearson and Spearman correlations unchanged whenever those correlations are defined. It also leaves the Pearson correlation of every fixed linear gene-program score unchanged.*

Proof. For gene g, the correction adds the same scalar c_g to every spot, and correlation is invariant to translation. For a fixed program vector v, $(z_i^0 + c)^\top v = (z_i^0)^\top v + c^\top v$, which is again a constant translation across spots. The rank order is therefore unchanged as well. □

Proposition 1 determines the interpretation of the corresponding secondary metrics. Equality of within-slide correlations across estimators is a mathematical consequence of the model class and serves as an implementation check; it is not evidence that the low-rank estimator attempted and failed to recover local spatial structure. The calibration ladder compares increasingly expressive *slide-level* offset vectors. Evaluation of a spatially varying estimator would require a separate model class and a leakage-controlled validation protocol.

5 Evaluation

The primary endpoint is variance-normalized log-MSE on the fixed target spots,

$$\text{NMSE} = \frac{1}{|T|G} \sum_{i\in T} \sum_{g=1}^{G} \frac{[\log(1 + y_{ig}) - \hat{z}_{ig}]^2}{\widehat{\text{Var}}_{\text{train}}[\log(1 + Y_g)]}. \tag{5}$$

The denominator is estimated separately for each gene using outer-training data. Secondary endpoints are mean Poisson deviance on the count scale, within-slide gene-wise Pearson and Spearman correlations, Pearson correlation for five fixed linear gene programs derived from training data, and Moran's-I error. All endpoints are evaluated on the immutable target spots.

For each endpoint, procedural partitions and slides are averaged within donor, after which each of the 47 donors receives equal weight. Paired 95% bootstrap intervals are obtained by resampling the 47 donor-level paired effects with replacement 1,000 times using seed 42. These intervals quantify uncertainty within the retrospective development dataset and are not external confirmatory intervals. Cohort-specific results for cohorts with fewer than four donors are treated as descriptive, as are the secondary endpoints and placement-policy comparisons.

Table 2: **Calibration ladder at requested budget $B = 5\%$ with spatial-coverage placement.** The requested 5% condition has a mean achieved fraction of 5.94% (range, 4.97–14.48%). NMSE and mean Poisson deviance are averaged over 20 partitions and all slides within donor and then averaged equally over 47 donors (70 slides). Relative changes are computed against A_0; lower values are better. Wins denote donors with lower mean NMSE than under A_0. Uncertainty intervals are not displayed in the table; paired donor-bootstrap intervals are reported in the text.

Estimator	Correction	NMSE	Δ	Deviance	Δ	Wins
A_0	none	1.441	–	31.90	–	–
A_1	scalar	1.206	-16.3%	28.28	-11.3%	28/47
A_2	gene-wise	**1.200**	-16.7%	**26.33**	-17.5%	**47/47**
A_3	low-rank	1.239	-14.0%	29.14	-8.7%	40/47

The prespecified primary success criterion evaluates A_3 with spatial-coverage placement at requested budget $B = 5\%$. Success requires at least a 10% reduction in donor-macro NMSE relative to A_0, a paired 95% interval entirely below zero, improvement for at least 70% of donors, and favorable point estimates in multiple cohorts. A secondary classification criterion would designate the result as spatial adaptation only if A_3 improved upon A_1 by at least 3% and improved a spatial-structure metric by at least 0.01. Because Proposition 1 makes the structure-metric condition unattainable for the implemented spot-invariant estimators, this criterion is retained only as a record of the prespecified analysis and is not interpreted as a valid test of local adaptation.

6 Results

6.1 Protocol integrity and the prespecified primary comparison

The evaluation comprises 56,000 records: 70 slides $\times$ 20 partitions $\times$ 5 requested budgets $\times$ 2 placement policies $\times$ 4 calibration estimators. Within every outer fold, training and test donors are disjoint. Stored set identifiers confirm that all estimators and placement policies use the same target indices for a given slide-partition pair, and automated integrity checks detected no target–candidate or target–prompt overlap in the analyzed records.

The prespecified A_3 comparison satisfies all primary success criteria. At requested budget $B = 5\%$ with spatial-coverage placement, donor-macro NMSE decreases from 1.441 for A_0 to 1.239 for A_3, an absolute difference of -0.202 and a relative reduction of 14.0%. The paired 95% bootstrap interval for the absolute difference is $[-0.500, -0.048]$, and 40 of 47 donors have lower mean NMSE under A_3. All nine cohort-level point estimates are favorable, although estimates from cohorts with two or three donors remain descriptive.

6.2 Gene-wise calibration is the strongest and most consistent estimator

Table 2 and Figure 2(a) compare the full calibration ladder. A_2 attains the lowest donor-macro NMSE (1.200; -16.7% relative to A_0) and mean Poisson deviance (26.33; -17.5%), with lower donor-level NMSE for all 47 donors. Its paired NMSE difference relative to A_0 has a 95% bootstrap interval of $[-0.414, -0.140]$. A_1 yields a similar mean NMSE of 1.206 (-16.3%), corresponding to 97.3% of the reduction achieved by A_2, but improves 28 of 47 donor means. The NMSE of A_3 is 2.7% higher than that of A_1 and 3.3% higher than that of A_2.

The contrast between the mean effect and the donor win count indicates substantial heterogeneity for the scalar estimator. Specifically, A_1 achieves a 16.3% reduction in donor-macro NMSE despite improving only 28 of 47 donor means, implying that its average benefit is concentrated among a subset of donors with larger baseline errors. By comparison, the gene-wise estimator combines a slightly lower donor-macro mean with a uniform direction of effect. These observations support A_2 as the preferred calibration estimator among the evaluated alternatives, while the performance of A_1 shows that a large fraction of the aggregate error is attributable to a shared slide-level offset.

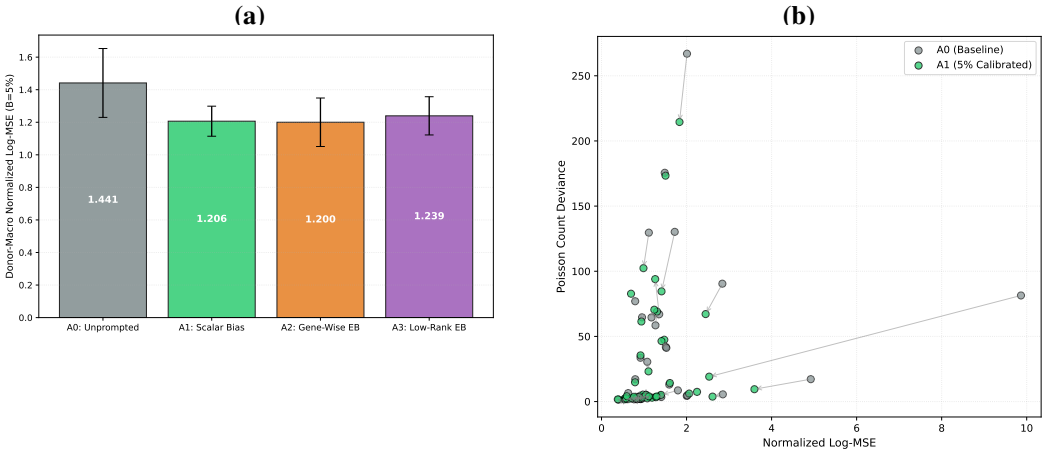

Figure 2: **Calibration magnitude and donor heterogeneity at requested budget $B = 5\%$.** Results use spatial-coverage placement and 47 donor-level means derived from 70 slides and 20 partitions per slide. (a) Donor-macro normalized log-MSE for the four estimators; error bars show standard errors across donor means. A_2 has the lowest mean, whereas A_1 recovers 97.3% of the reduction from A_0 to A_2. (b) Each arrow connects one donor's A_0 and A_1 mean NMSE. The wide range of baseline errors motivates paired donor-level inference rather than comparison of unpaired aggregate distributions.

6.3 The first complete FOV provides most scalar calibration, but gene-wise estimates continue to improve

Figure 3 shows the budget response for A_1. Under spatial-coverage placement, requested budget $B = 1\%$ reduces donor-macro NMSE from 1.441 to 1.214; requested budgets of 5% and 10% yield NMSE values of 1.206 and 1.200, respectively. The first nonzero condition therefore captures 94% of the scalar reduction observed at requested budget 10%. At requested budget 1%, spatial coverage has 2.8% lower NMSE than random placement (1.214 versus 1.249). The reported results do not show a persistent advantage for spatial coverage at larger requested budgets.

The gene-wise estimator exhibits a larger response to additional prompt spots: under spatial-coverage placement, NMSE is 1.255, 1.241, 1.200, and 1.112 at requested budgets of 1%, 2%, 5%, and 10%, respectively. This pattern is consistent with reduced sampling variance when estimating 50 gene-specific residual means, although that mechanism was not separately tested. The result therefore does not support a general claim that calibration saturates at requested budgets of 1–2%.

6.4 Gene-wise calibration improves mean error in every cohort

Table 3 reports the no-prompt A_0 and gene-wise A_2 results by cohort. Donor-mean NMSE is lower under A_2 in all nine cohorts, with relative reductions ranging from 2.2% in PRAD to 25.1% in CCRCC. Mean Poisson deviance is also lower in every cohort, with the largest relative reduction in CCRCC (34.8%). These comparisons use biological donors as the unit of aggregation; consequently, the 23 PRAD slides contribute two donor-level observations rather than 23 independent observations.

6.5 The experiment demonstrates level calibration, not spatial adaptation

Donor-macro gene-wise Pearson correlation is 0.282 for all four estimators to the displayed precision, and program-level correlation is 0.361. The maximum reported difference among estimators is below 3×10^{-9}. Moran's-I error is likewise unchanged. As predicted by Proposition 1, a spot-invariant correction cannot alter these within-slide structure metrics. The reductions in NMSE and mean Poisson deviance therefore arise from improved absolute expression levels rather than changed spatial rankings.

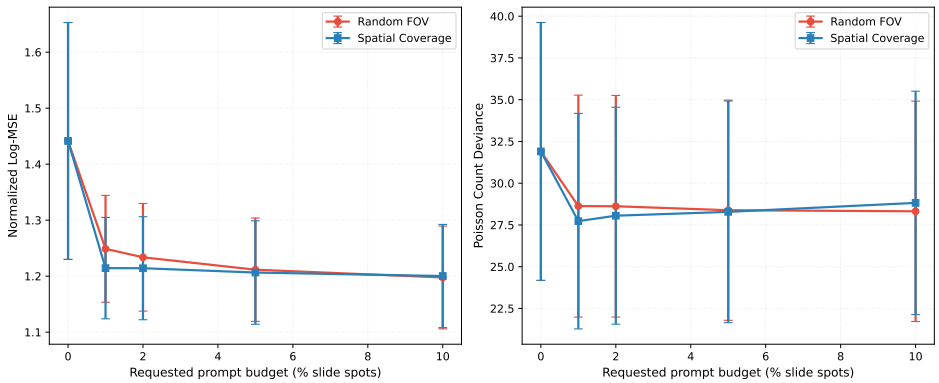

Figure 3: **Budget response of scalar calibration.** Curves report donor-macro mean normalized log-MSE for A_1 across 47 donors, after averaging 20 partitions and all slides within donor; error bars show standard errors across donor means. The horizontal coordinate is the requested fraction of slide spots. Complete-FOV selection yields mean achieved fractions of 2.85%, 3.41%, 5.94%, and 11.20% for requested budgets of 1%, 2%, 5%, and 10%, respectively.

Table 3: **Cohort-level gene-wise calibration at requested budget $B = 5\%$ with spatial-coverage placement.** N denotes slides and D denotes biological donors. NMSE values are donor means within cohort, and relative changes compare A_2 with the no-prompt baseline A_0; lower values are better. No uncertainty intervals are shown. Cohorts with fewer than four donors are interpreted descriptively.

Cohort	N	D	A_0 NMSE	A_2 NMSE	Δ NMSE	Δ dev.
CCRCC	24	24	1.279	0.958	−25.1%	−34.8%
COAD	4	4	1.032	0.975	−5.5%	−9.7%
IDC	4	4	1.456	1.217	−16.4%	−24.4%
LUNG	2	2	2.178	1.984	−8.9%	−15.7%
LYMPH-IDC	4	4	2.508	2.141	−14.6%	−27.5%
PAAD	3	3	1.131	1.081	−4.4%	−10.2%
PRAD	23	2	1.110	1.086	−2.2%	−6.2%
READ	4	2	2.129	1.999	−6.1%	−20.6%
SKCM	2	2	1.422	1.349	−5.1%	−6.2%
Donor macro	70	47	1.441	1.200	−16.7%	−17.5%

The direct empirical conclusion is that same-slide RNA measurements identify a residual expression baseline not recovered by the histology-only predictor. The scalar result is compatible with a slide-wide scale or batch component, and the greater consistency of A_2 is compatible with additional gene-specific slide effects. These explanations remain hypotheses: without technical replicates, controlled library-size perturbations, or an independent assay-depth intervention, the experiment cannot determine whether the offsets are technical, biological, or mixed.

7 Discussion

Why the simpler model succeeds. The previously evaluated spatial approach estimated local residual fields from sparse prompt observations and performed worse than the no-prompt baseline. The present estimators instead infer at most 50 slide-level quantities, shrink them using outer-training donors, and apply them uniformly to target spots. This lower-dimensional parameterization is designed to reduce estimation variance and is deterministic conditional on the prompt set. The comparison between A_2 and A_3 further shows that additional structural constraints need not improve calibration: projecting residuals onto a learned low-rank subspace increases NMSE relative to unrestricted gene-wise offsets. This result is consistent with information loss under the low-rank projection, although the experiment does not isolate the responsible mechanism.

What the result means for active acquisition. Demonstrating prompt value is necessary but not sufficient to justify active region selection. Spatial coverage improves upon random placement at the smallest requested budget, but the reported advantage is not sustained as more complete FOVs are acquired. A future active-selection study should therefore establish, under independent molecular observations, that candidate FOVs have reproducibly different downstream value. If candidate utilities are approximately exchangeable, random or space-filling sampling may be sufficient and a learned selector may add unnecessary variance.

What the result means for histology models. The positive within-slide correlations of the no-prompt predictor (0.282 gene-wise and 0.361 at the program level) indicate that predicted variation covaries with observed spatial variation, whereas the calibration gains show that absolute cross-slide levels remain uncertain. Correlation metrics alone cannot reveal a translation error because they are invariant to additive offsets. Histology-to-expression benchmarks should therefore report both within-slide ranking metrics and level-sensitive metrics on a log or count scale, and should distinguish prediction on an unseen donor from calibration using molecular observations from the same slide.

8 LIMITATIONS

First, all nine cohorts belong to HEST development data used during the broader methodological investigation; no sealed external cohort was reserved for confirmatory evaluation. The reported intervals therefore quantify uncertainty within this dataset and do not establish prospective generalization. Second, the study retrospectively masks regions of complete Visium slides. It does not establish that arbitrary FOVs can be acquired sequentially, at the assumed cost, or without affecting tissue available for subsequent assays. Third, complete-FOV acquisition overshoots small requested budgets, reaching as much as 14.48% on the smallest slides under nominal 1%, 2%, and 5% conditions. The results must therefore be interpreted in terms of requested conditions and reported achieved fractions rather than exact assayed-area percentages. Fourth, the evaluation is restricted to cohort-specific panels of 50 genes and does not establish transcriptome-wide performance.

Fifth, all evaluated corrections are spatially constant. They cannot improve local boundaries, within-slide rankings, or centered spatial-autocorrelation measures, and the study does not compare calibration with a valid spatially varying estimator. Sixth, residual offsets conflate technical capture, batch effects, tissue-wide composition, and donor biology. Technical replicates or explicit assay-depth interventions are required to identify their sources. Seventh, five cohorts contain only two or three donors, so their cohort-specific estimates are descriptive. Eighth, the A_3 rank-selection procedure reuses the outer-fold Ridge predictor while scoring inner calibration splits rather than cross-fitting the base predictor within each inner split. This dependence may favor A_3, although A_3 nevertheless performs worse than the untuned A_1 and A_2 estimators on held-out outer donors. Finally, the implemented A_1 estimator divides a per-spot gene-averaged residual variance by $|P|G$ rather than the conventional $|P|$, thereby weakening shrinkage toward zero. We report the evaluated estimator exactly and provide the conventional alternative in Appendix B. Future evaluations of spatial adaptation should cross-fit both the base predictor and calibration estimator within the inner loop.

9 CONCLUSION

In this retrospective HEST development-set evaluation, sparse same-slide RNA measurements reduce level-sensitive error in cross-donor histology-based expression prediction. At requested budget $B = 5\%$ (mean achieved fraction, 5.94%), gene-wise empirical-Bayes offsets reduce donor-macro normalized log-MSE by 16.7%, reduce mean Poisson deviance by 17.5%, and improve the donor-level mean for all 47 donors across nine cohorts. A scalar offset captures nearly all of the mean NMSE reduction, whereas low-rank calibration performs worse. The supported contribution is therefore a slide-level calibration baseline: measure disjoint prompt FOVs, estimate shrunk offsets, and preserve the spatial pattern supplied by the histology predictor. Claims regarding local biological adaptation, optimal region selection, or technical capture depth as the unique mechanism require a different experimental design.

Reproducibility Statement

The accompanying repository contains the calibration code, base-model evaluation code, automated protocol-integrity checks, 56,000 row-level evaluation records, configuration manifests, prespecified decision criteria, and deterministic figure-generation code. Because the CCRCC cohort was evaluated before the remaining cohorts, separate immutable manifests are retained for the initial and extended analyses to preserve provenance. External HEST matrices and precomputed UNI2-h embeddings are not redistributed and must be obtained from their original sources under the applicable access terms.

Ethics Statement

This retrospective computational study uses de-identified profiles distributed through HEST-1k under its stated access terms (Jaume et al., 2024). The method has not been validated for diagnosis, treatment selection, or prospective tissue acquisition. Calibration may reduce aggregate error while preserving gene-specific, cohort-specific, and demographic biases inherited from the base predictor. Prospective use would require independent validation, assay-specific oversight, and explicit accounting for tissue consumed by prompt measurements.

References

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939.

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. doi: 10.1038/s41591-024-02857-3.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2012.

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. doi: 10.1038/s41551-020-0578-x.

Guillaume Jaume, Paul Doucet, Andrew H Song, Ming Y Lu, Cristina Almagro-Pérez, Sophia J Wagner, Anurag J Vaidya, Richard J Chen, Drew FK Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:53798–53833, 2024.

Andrew Jones, Diana Cai, Didong Li, and Barbara E Engelhardt. Optimizing the design of spatial genomic studies. *Nature Communications*, 15(1):4987, 2024. doi: 10.1038/s41467-024-49174-4.

Zhuo Li, Weiran Wu, Chuangyi Han, Yan Cui, Tian Lu, Rongqin Ke, Jian Sun, and Zhiyuan Yuan. SOFisher: Reinforcement learning-guided experiment designs for spatial omics. *Nature Communications*, 17(1):6830, 2026. doi: 10.1038/s41467-026-73404-6.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. doi: 10.1186/s13059-014-0550-8.

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

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. doi: 10.1214/23-STS915.

Ananya Rao, Dalia Barkley, Gabriel S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature Reviews Genetics*, 22(8):515–530, 2021. doi: 10.1038/s41576-021-00370-8.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010. doi: 10.1093/bioinformatics/btp616.

Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pierre Maillé, Julien Calderaro, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications*, 11(1):3877, 2020. doi: 10.1038/s41467-020-17678-4.

Gordon K Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1): Article 3, 2004. doi: 10.2202/1544-6115.1027.

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. doi: 10.1126/science.aaf2403.

Ronald Xie, Kuan Pang, Sai W 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 (NeurIPS)*, volume 36, pp. 70832–70845, 2023.

Jonathan Xu, Michelle Jiang, Shunsuke Koga, Nancy Zhang, and Zhi Huang. SpatialFinder: A human-in-the-loop vision-language framework for prioritizing high-value regions in spatial transcriptomics. *Frontiers in Bioinformatics*, 6:1746714, 2026. doi: 10.3389/fbinf.2026.1746714.

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. doi: 10.1093/bib/bbac297.

Junchao Zhu, Ruining Deng, Junlin Guo, Tianyuan Yao, Chongyu Qu, Juming Xiong, Siqi Lu, Zhengyi Lu, Yanfan Zhu, Marilyn Lionts, Yuechen Yang, Yalin Zheng, Yu Wang, Shilin Zhao, Haichun Yang, and Yuankai Huo. SCR^2-ST: Combine single cell with spatial transcriptomics for efficient active sampling via reinforcement learning. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 488–509. PMLR, 2026a. URL https://proceedings.mlr.press/v315/zhu26a.html.

Zheyu Zhu, Junchao Zhu, Fengbei Liu, Tianyuan Yao, Gelei Xu, John Cannon, Haichun Yang, Yuankai Huo, Mert R Sabuncu, and Ruining Deng. Benchmarking active spot selection for cost-efficient spatial transcriptomics. *arXiv preprint arXiv:2609.27208*, 2026b. URL https://arxiv.org/abs/2609.27208v1.