FROM HISTOLOGY TO CELL SENTENCES: VISUAL SOFT-PROMPTS FOR LONG-HORIZON RANKED GENE EXPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics (ST) measures gene expression in tissue but remains costly and sparse, whereas histology images (H&E) are abundant. Existing virtual-ST methods typically regress a small highly variable gene (HVG) panel or retrieve expression from a reference atlas. We instead cast H&E→expression as ranked sequence generation in the spirit of LLaVA-style visual conditioning of an LLM, adapted to a domain where that recipe does not directly transfer. A lightweight adapter couples two domain-specific foundation models trained on massive corpora, a pathology expert and a single-cell language model, mapping frozen visual features into *Visual Soft-Prompts* that elicit *cell sentences*: gene symbols ranked by relative expression over a ∼27k-gene vocabulary.

Prediction length is a central challenge: short-context models give the strongest early-rank fidelity but emit only a handful of genes, while naively extending the autoregressive horizon with a frozen decoder yields longer lists that eventually derail into duplicate loops and invalid tokens. A failure mode analysis shows that standard decoding fixes such as nucleus sampling or constrained decoding do not repair this. With attention LoRA and a larger training budget, derailing is largely prevented, producing near-complete cell sentences while retaining early-rank precision.

Scored against the full ground-truth cell sentence at every depth, dense panel regressors and averaging retrieval methods (ST-Net, BLEEP) lead briefly but collapse once they exhaust the panel they were built and evaluated on: they cannot, by construction, name genes outside it, whereas our model continues informatively over the full expressed set. On a frozen 1k HVG panel used only as a secondary calibration, a mid-length LoRA model (2048 tokens) approaches ST-Net and BLEEP on panel correlation (within ∼0.03 PCC). The result is a generative path from H&E to ranked cell sentences that needs no pre-specified gene panel and can compose with external predictions rather than only replace them.

1 INTRODUCTION

Histopathology, most often Hematoxylin and Eosin (H&E) staining, is the clinical gold standard for high-resolution maps of tissue architecture, but it provides no direct measurement of the molecular state. Single-cell RNA sequencing (scRNA-seq) quantifies gene expression at cellular resolution across a large transcriptome, yet dissociating tissue into a suspension destroys spatial context. Spatial transcriptomics (ST) restores location, but remains constrained by cost, coverage, and resolution trade-offs that limit large clinical cohorts. Virtual ST aims to close this gap by predicting molecular profiles from histology alone.

From a learning perspective, image-to-expression is hard. Spot- or cell-level transcriptomes are high-dimensional, extremely sparse, and zero-inflated: most genes are inactive in a given observation, and technical dropouts further corrupt counts. Absolute expression is therefore a brittle regression target. Relative ordering of detected genes is often more stable and more directly useful as a molecular signature. Our target is accordingly a ranked list of detected genes, not quantitative expression values: the method should be read as predicting relative gene ordering and molecular signatures, not reconstructing absolute expression magnitude,

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.

Prior work largely follows three paradigms, dense panel regression, retrieval, and continuous generative modeling of counts, each with structural limits on what it can name or compose (Sec. 2); we compare against representative panel and retrieval baselines under a shared protocol in Sec. 4.

We cast H&E→expression as ranked sequence generation, conditioning a language model on visual features in the style of LLaVA (Liu et al., 2023). The recipe does not transfer as-is: with no natural-language image captions available, we cannot run a caption-alignment warm-up stage. We instead keep both foundation models frozen and adapt only a lightweight interface, for parameter efficiency and to avoid forgetting either pretrained representation, optionally allowing a low-rank update to the decoder's attention alone when longer cell sentences require it (Sec. 4.3).

We introduce *Visual Soft-Prompts*: GPFM patch embeddings, plus assay metadata, are mapped to k=16 continuous tokens in the Cell2Sentence (Levine et al., 2024) embedding space and spliced after BOS, eliciting *cell sentences*, gene symbols ordered by relative expression over a ∼27k-gene keep-set, with optional LoRA on decoder attention for long horizons (architecture details in Sec. 3).

This transfer raises a systems question that panel regressors do not face: *how long can autoregressive cell-sentence decoding remain faithful, and what does it take to make it so?* We answer this directly below, and test what else the result depends on.

Contributions.

- An analysis of *autoregressive derailing* in long-horizon virtual ST: decode-time fixes (nucleus sampling, vocabulary/uniqueness constraints) restore surface validity but not rank fidelity, and a matched-length, matched-budget ablation shows that continued training *without* attention LoRA does not fix it either. We propose attention LoRA combined with extended training budget as a fix for autoregressive derailing.
- *Visual Soft-Prompts*: a parameter-efficient adapter that conditions a frozen cell-sentence LLM on a frozen pathology foundation model without natural-language captions, adapter-only for short contexts, with an optional low-rank attention update for long horizons (Sec. 4.3).
- A dual evaluation protocol on patient-split HEST: full cell-sentence metrics over the keep-set vocabulary, plus a frozen ∼1k HVG panel for calibration against ST-Net and BLEEP.
- Controls that test whether histology pairing, the C2S prior, and technology-token conditioning are necessary for ranked gene prediction.

2 Related Work

2.1 Image-to-expression from histology

Virtual ST maps H&E morphology to gene expression. **Dense panel regressors** such as ST-Net (He et al., 2020) and HisToGene (Pang et al., 2021) predict a fixed HVG vector with MSE or count likelihoods. In sparse, zero-inflated transcriptomes these objectives often regress toward tissue averages and cannot emit genes outside the trained panel. **Retrieval / contrastive** models such as BLEEP (Xie et al., 2023) and OmiCLIP (Chen et al., 2025) align patch and expression embeddings; at inference they typically require a reference library, which limits composition of novel ranked lists. We compare against BLEEP, whose released setup retrieves from a reference library rather than emitting a rankable list. **Continuous generative** approaches such as STFlow (Huang et al., 2025) model expression as a continuous field (e.g., flow matching) at whole-slide granularity, remaining count-centric and sensitive to dropout noise.

Autoregressive count generation. GenAR (Ouyang et al., 2026) reframes virtual ST as *discrete* multi-scale sequence modeling, predicting raw counts on a curated ∼200-gene inventory conditioned on histology (UNI (Chen et al., 2024)) and spatial coordinates. Both approaches treat cross-gene structure as sequential, but at opposite scales, GenAR curates a small inventory for tractable dense counts, whereas we generate **rank-ordered cell sentences** over an open ∼27k-gene vocabulary via a *frozen* cell LLM (Sec. 4.4), so we compare against classical baselines on the shared 1k HVG panel (Sec. 4) rather than report a head-to-head number against GenAR's narrower, curated protocol.

2.2 PATHOLOGY AND SINGLE-CELL FOUNDATION MODELS

Pathology foundation models (e.g., UNI (Chen et al., 2024), GPFM (Ma et al., 2026), CONCH (Lu et al., 2024)) learn morphological representations from large H&E corpora, often with self-supervised ViT objectives such as DINOv2 (Oquab et al., 2024).

Single-cell foundation models (e.g., scGPT (Cui et al., 2024), scFoundation (Hao et al., 2024)) treat cells and genes as structured sequences. Cell2Sentence (C2S) (Levine et al., 2024) linearizes expression into ranked gene sentences, enabling causal LMs to generate transcriptomic signatures autoregressively. C2S-Scale extends this recipe on Gemma-2 (Team et al., 2024) at multi-billion scale with long context (Rizvi et al., 2026).

2.3 SOFT PROMPTS, MULTIMODAL CONDITIONING, AND PEFT

Soft / continuous prompts adapt frozen LMs by learning embeddings in token space rather than updating all weights (Lester et al., 2021; Li & Liang, 2021). LLaVA-style systems (Liu et al., 2023) map vision features into the LM embedding space, typically with caption supervision and partial LLM updates. Our setting differs: there are no H&E–caption pairs; Visual Soft-Prompts are trained via next-token cross-entropy on gene-rank sequences while keeping both the pathology encoder and the cell LLM frozen. Continuous tokens projected from pathology embeddings into the C2S-Scale-Gemma-2-2B latent space let morphological cues reach the genomic decoder without passing through a natural-language bottleneck. Optional LoRA (Hu et al., 2022) on decoder attention addresses long-horizon generation when freezing the decoder entirely is insufficient.

Relative to that landscape, we focus on Visual Soft-Prompts for a *frozen* cell-sentence LLM (parameter-efficient conditioning rather than GenAR's end-to-end trained decoder), the length–fidelity trade-off of autoregressive virtual ST, and a protocol that separates full-sentence evaluation from shared-panel comparison with classical baselines.

3 METHOD

We condition a frozen cell-sentence language model on a frozen pathology foundation model via *Visual Soft-Prompts*: a lightweight adapter that maps H&E patch embeddings (plus assay metadata) into continuous visual tokens. Optional attention LoRA stabilizes long-horizon decoding. Figure 1 summarizes the stack.

3.1 ARCHITECTURE

Vision encoder. Each H&E spot patch is encoded with GPFM (Ma et al., 2026) into a global embedding $z \in \mathbb{R}^{1024}$. GPFM weights remain frozen; at train time we use precomputed embeddings.

Visual Soft-Prompt adapter. We concatenate z with one-hot encodings of ST technology and organ, then map the result through a residual MLP (4 layers, hidden width 4096) to $k{=}16$ continuous tokens in the LLM embedding space. Adapter outputs are scaled to the text-embedding standard deviation and cast to bf16. These soft tokens are not discrete subword IDs; they are `inputs_embeds` spliced into the decoder sequence. The adapter has $\sim$223M trainable parameters; with attention LoRA enabled it adds a further $\sim$3.2M ($r{=}8$ on `q/k/v/o_proj` in every decoder layer), together under 9% of the $\sim$2.6B frozen C2S-Scale-Gemma-2-2B parameters. This residual MLP ends in a single unrolled linear head that produces all $k{=}16$ tokens at once, rather than a sequence-aware or cross-attention resampler (e.g., Perceiver); a crossed ablation against $k{=}64$ (Fig. 8, Appendix B) shows why this is a reasonable default: a wider token budget gives the frozen decoder more surface to attend to (a real Jaccard/NDCG edge for $k{=}64$), but attention LoRA closes most of that gap, so a more expressive projection buys comparatively little once LoRA is enabled anyway.

Cell-sentence LLM. We use frozen C2S-Scale-Gemma-2-2B (Levine et al., 2024; Rizvi et al., 2026). The input layout is

$$[\text{BOS},\ v_1, \ldots, v_{16},\ g_1, \ldots, g_m,\ \text{EOS}], \tag{1}$$

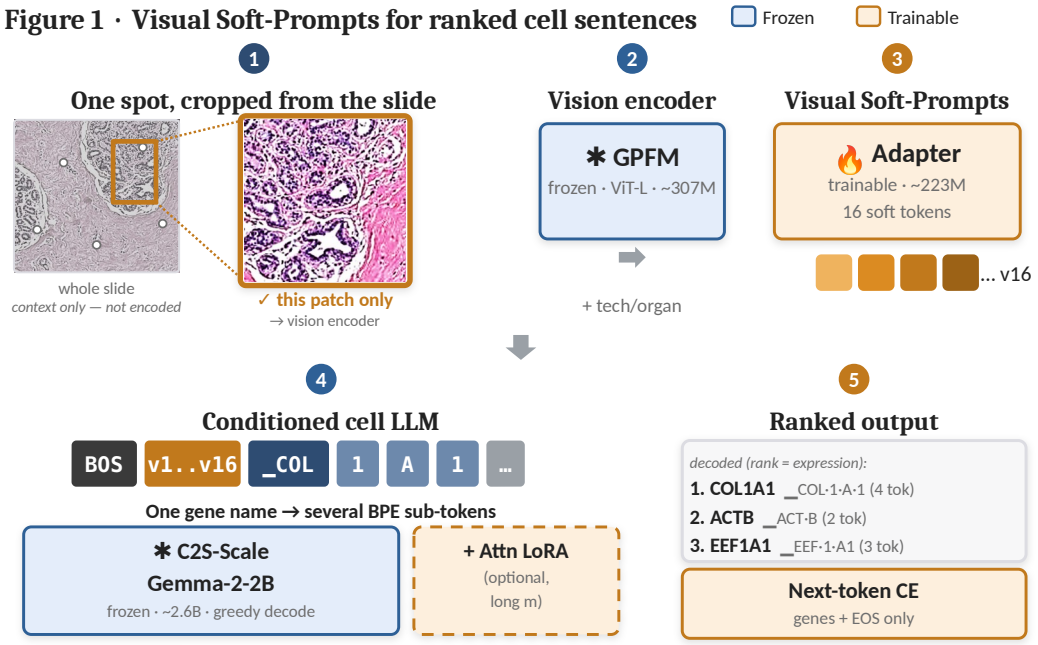


Figure 1: Visual Soft-Prompts for ranked cell sentences. A single H&E patch (one per spot; the rest of the slide is unseen) is mapped, from frozen GPFM embeddings plus assay metadata, via a trainable adapter, to k=16 continuous tokens conditioning a frozen C2S-Scale Gemma-2 decoder, with optional attention LoRA for long-m models. Supervision is next-token CE on gene slots only; each gene name decodes as several SentencePiece sub-tokens, not one token per gene. Compact operating point: adapter-only, short m (early-rank fidelity); long operating point: adapter+LoRA, m=8192 (near-complete sentences). Histology image: Mikael Häggström, M.D. et al., CC BY 4.0, Wikimedia Commons (cropped, illustrative, not a HEST sample).

where v_i are visual soft tokens and g_j are gene-sentence tokens from the frozen embedding table. Cross-entropy is applied only on gene (and EOS) positions; visual slots use ignore labels. C2S-Scale-Gemma-2-2B is pretrained on scRNA-seq cell sentences, each representing one dissociated cell; the ST spots we condition on are not single cells, HEST-1k is dominated by 10x Visium, where a capture spot aggregates ∼10–50 cells, so the frozen decoder's linguistic prior is applied to a coarser unit of observation than it was trained on. Attention LoRA (below) adapts decoder attention to this mismatch when enabled, but the embedding table and LM head, where the gene-cooccurrence prior actually lives, stay frozen and single-cell-derived regardless: the gap only ever closes partially, and not at all for the frozen adapter-only operating points.

Attention LoRA (optional). When enabled, we attach LoRA (Hu et al., 2022) with rank r=8, α=16, and dropout 0.05 to every attention projection (`q_proj`, `k_proj`, `v_proj`, `o_proj`) in all decoder layers. Base attention weights, MLP blocks, embeddings, LayerNorms, and the LM head stay frozen. Short-context models in our study are adapter-only; long-horizon models train adapter + attention LoRA.

3.2 Data and Cell Sentences

Corpus and split. We use HEST paired H&E–ST data (Jaume et al., 2024) across organs and technologies (including Visium, Visium HD, Xenium, and Spatial Transcriptomics). Within-family assay variants (Xenium and Xenium 5k; Visium HD and Visium HD 3′) are collapsed to their parent technology label before one-hot encoding and for all technology-stratified reporting (e.g. Fig. 6), since we found no consistent, model-relevant signal from treating them separately. Slides are partitioned by patient/study `split_group` into 1014 train and 262 test slides (no patient leakage).

Unless noted, test metrics use 1310 spots (5 spots sampled per test slide, seed 42, minimum GT sentence length 200 genes).

Gene symbols and keep-set. HEST count matrices expose heterogeneous gene identifiers across slides; we normalize these to canonical HGNC symbols via a global mapping pipeline (full procedure in the Reproducibility Statement), and mapped nonzero genes become the cell-sentence targets. We then restrict the modeling vocabulary to genes present in at least 0.05% of training sentences ($\sim$27,216 symbols), matching C2S rare-gene filtering. This keep-set bounds training targets and evaluation scope, not the decoder's output space: at inference the model samples from the frozen LM's full token vocabulary, so nothing prevents it from emitting a string outside the keep-set, scored as invalid, which is exactly what the gene-validity diagnostic tracks (Sec. 4.6), and why vocabulary-constrained decoding is tested there as one decoding ablation among several rather than as the default.

Cell-sentence construction. For each spot, genes with zero count are dropped. Remaining genes are sorted by descending expression; ties break by gene name, inherited from Cell2Sentence (Levine et al., 2024), count-tied genes therefore receive an arbitrary order, relevant for rank-sensitive metrics in the sparse tail. The resulting ordered list is the ground-truth cell sentence: rank encodes relative expression, and absent genes are treated as unexpressed or unmeasured. Technology and organ labels are likewise given to the adapter as one-hots, over the collapsed categories above; classical baselines (Sec. 4) do not receive this metadata, an asymmetry we test via shuffle controls (Sec. 4.6).

3.3 Training and inference

We train with next-token cross-entropy on cell sentences conditioned on Visual Soft-Prompts (DeepSpeed ZeRO-3, FlashAttention-2, bf16; hardware, batch sizes, and checkpoint selection in Appendix A). The main hyperparameter is train/decode `max_length` m: we study short contexts (notably m=128) up to m=8192, under matched short budgets (12k optimizer steps) and longer full-train budgets (100k–200k steps). At inference we decode greedily from the BOS + visual soft-token prefix, with `max_new_tokens` matched to the trained m; nucleus sampling and constrained decoding appear only as controls (Sec. 4.6). Three operating points matter for the paper: a **compact** adapter-only short-m model (m=128) for early-signature ranking, a **mid-length** LoRA model (m=2048) for panel comparison, and a **long** attention-LoRA model (m=8192) for near-complete cell sentences.

3.4 Evaluation protocol

We use a dual protocol so generative lists and dense panel regressors are not conflated.

(A) Full cell-sentence evaluation. Predictions are scored against the ground-truth sentence over the keep-set vocabulary with Jaccard at depth, plus diagnostics for derailing (duplicate rate, gene validity vs. position, novelty vs. depth, loop structure). Early-signature HVG analyses that take top genes from a GT prefix (e.g., GT[:100]) are internal to this track.

(B) Shared HVG panel (Table 1). For comparison with ST-Net and BLEEP use a frozen HVG Panel: the top 1000 genes by variance of $1/\mathrm{rank}$ pseudo-expression on train spots only. Predicted and GT lists are restricted to this panel (relative order preserved). We report Jaccard@10, NDCG@50 and PCC(inv): Pearson correlation of length-1000 vectors of $1/\mathrm{rank}$ with absent genes set to 0 on both sides. Baselines regress `log1p` CPM on mapped `st_normalized` counts; generative models never see dense CPM targets.

This separation is intentional: panel metrics answer "how do we compare to dense regressors on a shared gene universe?", while full-sentence metrics answer "how faithful is ranked generation over the expressed transcriptome?"

4 EXPERIMENTS

We evaluate Visual Soft-Prompts as a generative alternative to panel regression and retrieval, across the subsections below, characterizing ranked generation's failure modes and a practical long-horizon recipe rather than claiming HEST leaderboard SOTA.

4.1 SETUP

Data and models. Unless noted, quantitative tables use the full $\mathbf{n{=}1310}$ protocol from Sec. 3.2; a lean protocol (†) uses 1 spot $\times$ 262 test slides ($n{=}262$) for directional checks only, and we do not mix † and $n{=}1310$ rows without a footnote. The three operating points from Sec. 3.3 use, respectively, a 12k-step checkpoint ($m{=}128$), 100k steps ($m{=}2048$; Table 1), and 200k steps ($m{=}8192$; Table 1, global batch 2 on two GPUs vs. batch 1 for $m{=}2048$, so step budgets are not example-matched).

Baselines and metrics. ST-Net, BLEEP, and HisToGene are trained on the training split and evaluated on the same test spots as our models, with mapped log1p-CPM targets on our defined HVG Panel; our models receive organ and technology one-hots, while baselines follow their standard image-only recipes (Sec. 3).

4.2 FINITE SIGNATURES AND AUTOREGRESSIVE DERAILING

Under a matched **12k**-step budget with a frozen decoder, early-rank quality peaks at short m. The $m{=}128$ setting yields the strongest early Jaccard with median lists of only $\sim$30 genes. Lengthening the train/decode horizon produces longer strings that **eventually derail**: novelty collapses with depth, duplicate loops appear, and late gene validity falls. At the longest frozen setting in the matched-budget sweep ($m{=}8192$), median duplication reaches $\sim$0.74, late validity (genes after position 250) falls to $\sim$0.07, and full-sentence J@10 collapses to $\sim$0.04 (Fig. 4, Appendix B; Sec. 4.3). Novelty typically dies by positions $\sim$250–300 even when generation continues, the same frozen-decoder failure mode diagnosed at $m{=}2048$ in Fig. 5 (Appendix B) and, alongside its LoRA fix, at $m{\in}\{2048, 8192\}$ in Fig. 4.

The takeaway is not that longer context is useless, but that with a frozen cell LLM, extra length mostly buys late tokens that are near-noise for early-rank metrics. Extending length without adaptation is therefore not a reliable path to near-complete cell sentences.

We then ask whether derailing is merely a decoding artifact. At frozen $m{=}2048$, vocabulary-constrained decoding drives late validity to 0.999 (from greedy's 0.223) while leaving J@10 essentially unchanged (0.175 vs. 0.174 greedy), the constrained string is syntactically clean, yet no more biologically correct, and duplication stays high (0.776 vs. 0.718 greedy), since forcing a valid gene name does nothing to stop the model repeating one, immediately. The uniqueness constraint reaches the same 0.999 late validity and lowers duplication somewhat (0.553) and unchanged J@10 (0.175). Nucleus sampling ($p{=}0.9$) is not neutral either: it lowers J@10 further (0.174$\rightarrow$0.115), consistent with perturbing the very rank order being scored, even as it also raises late validity (0.789) and lowers duplication (0.097) by slowing the decay. J@10 barely moves under any constraint, and Jaccard at depth tracks the same frozen-decoder derailing curve regardless of arm (Fig. 5, Appendix B): each constraint blocks one escape route for a collapsed conditional distribution and the model routes its mass into another, first into valid but off-manifold genes, then into immediate self-restarts, so cleaning the surface string adds no biological meaning, it only changes which symptom of lost conditional structure is visible.

4.3 ATTENTION LORA RESTORES LONG CELL SENTENCES

Attention LoRA changes the length story, but only in combination with a longer training budget, not on its own. Under a short 12k budget, long-m LoRA still derails (e.g., $m{=}8192$: duplication 0.540, late validity 0.427 at $n{=}1310$). With 100k–200k steps at $m{=}8192$, derailing is largely prevented: median lists reach $\sim$2,100 genes, covering nearly all expressed genes in a typical cell sentence (GT median $\sim$2.1k on the same spots), with full-sentence J@10 within 0.01 of the short model (0.303 vs. 0.314 at $n{=}1310$). Scaling 100k$\rightarrow$200k at this length does not further lift early Jaccard/HVG; the main win is stable long coverage.

This raises the question we answer directly with a matched-length, matched-budget control: is LoRA actually necessary, or would continued training alone eventually fix frozen $m{=}8192$ too? Fig. 4 (Appendix B) answers it: training the *frozen* adapter-only model at $m{=}8192$ for 100k steps does not rescue it, J@10 falls from 0.040 to 0.035, median length *shrinks* from 556 to 246, late validity only creeps from 0.067 to 0.120, and duplication stays high (0.744→0.511), the opposite of what a budget-alone fix would predict. Attention LoRA at the same lengths and budgets moves from 0.255 to 0.320 J@10, median length growing to 2122, late validity from 0.427 to 0.917, and duplication falling to 0.091: fidelity and hygiene improve together with budget, unlike the frozen arm above. LoRA is not a proxy for more training, it is the ingredient that makes long-horizon Visual Soft-Prompt decoding usable at all.

4.4 Long-horizon fidelity against ground truth

With attention LoRA established as the fix for derailing, we compare the resulting model against the full ground-truth cell sentence, every gene detected in the spot, in its true rank order, with no hint about which genes are worth naming (protocol (A), Sec. 3.4). We score all methods with *windowed* NDCG (window $W{=}50$: the NDCG of the genes emitted in $[\mathrm{depth}{-}50, \mathrm{depth}]$ under global relevance $1/\mathrm{GT\text{-}rank}$), judging each depth on local order-consistency rather than cumulative agreement with the head of the list. The long tail is dominated by rank ties, so deep windowed NDCG (Fig. 2A) becomes a proxy for membership rather than fine order, confirmed by the order-free expressed-set precision panel (Fig. 2B), which reproduces the same pattern with earlier crossovers.

ST-Net and BLEEP lead at shallow depth (Fig. 2, consistent with Table 1), then collapse once past the genes they are actually confident about (BLEEP: 0.408→0.271 from $k{=}500$ to 750; ST-Net: 0.374→0.185), only $\sim$28% of their 1k panel is expressed per spot, so their supply of true positives runs out well before the panel's edge. Our long-LoRA model ($m{=}8192$) trails near the top but degrades gracefully instead (0.430, 0.404, 0.375 at $k{=}500/750/1000$, settling near 0.36 to $k{=}2200$), staying informative across the full open-vocabulary sentence rather than only a fixed panel.

This graceful degradation also enables composition: we additionally decode a *hybrid* arm, splicing a hard ST-Net prefix (its own thresholded output, $\tau{=}0.5$, capped at $K{=}100$) into the Visual Soft-Prompt decoder and letting it continue generating. The hybrid tracks ST-Net through the prefix, then keeps high quality where ST-Net alone collapses (0.539, 0.486, 0.466, 0.450, 0.410 at $k{=}250/500/750/1000/2000$; Fig. 2), an improvement over our long-LoRA model alone at each depth beyond the prefix (ΔJ@10 $= +0.081$, ΔJ_full $= +0.036$ at $n{=}1310$). This works because our decoder accepts *any* gene-list prefix and continues it coherently, unlike a dense regressor or retrieval baseline, which emits a fixed-size output in one shot, suggesting this architecture could serve as a general substrate for multimodal composition, accepting and extending predictions from other tools rather than only replacing them.

4.5 Secondary: shared-panel calibration

As a second, narrower comparison, we calibrate against ST-Net and BLEEP on a shared 1k highly-variable-gene (HVG) panel, Sec. 3.4B), itself an oracle, the top 1000 genes by variance of $1/\mathrm{rank}$ pseudo-expression on training-split data. Our method needs no such list, ranking instead over the full $\sim$27k-gene open vocabulary with no hint about which genes matter for a given spot.

ST-Net and BLEEP lead J@10, NDCG@50, and PCC(inv) here, both dense regressors trained to score every gene in a panel selected, by construction, to match the variance structure they were built to fit. HisToGene also emits a full-panel ranking (Recall 1.000) but trails every method on early ranks and PCC(inv) (J@10 0.161, PCC(inv) 0.233).

The two LoRA operating points trade panel coverage for selectivity: the mid-length model ($m{=}2048$) is closest to ST-Net/BLEEP on J@10 and PCC(inv) and posting the best precision in the table (0.734); the long model ($m{=}8192$) trades panel PCC for coverage (Recall 0.61), landing slightly below BLEEP's best F1 (0.581$\pm$0.008 vs. 0.585$\pm$0.010). Both baselines score every panel gene rather than abstaining, so once only $\sim$28% of panel genes are truly expressed most calls are false positives (Precision 0.360 ST-Net, 0.439 BLEEP, vs. 0.636–0.734 across our operating points).

Beyond these spot-wise metrics, we also compare whole-slide marker maps against ground truth (Fig. 3, Appendix B). ST-Net and BLEEP still lead spot-wise, but their whole-slide maps show a

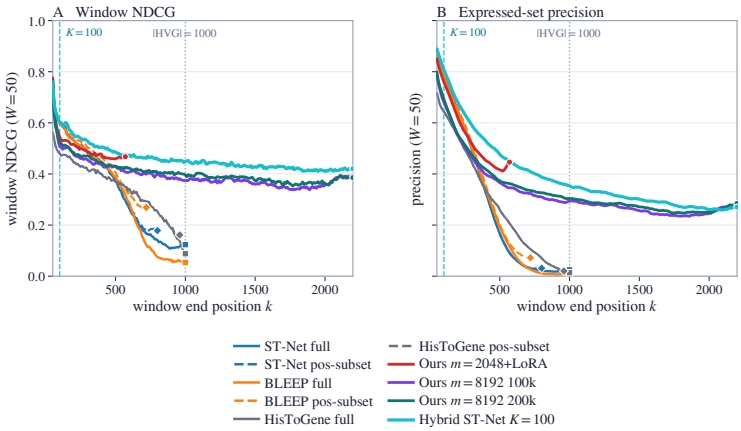


Figure 2: Windowed (W=50) fidelity vs. depth against the full ground-truth cell sentence, n=1310. (A) NDCG (relevance = $1/\mathrm{GT\text{-}rank}$): windowed agreement with the ground-truth ranked list, best read as set-membership consistency once rank ties dominate the tail (Sec. 4.4). (B) Order-free: fraction of each predicted window expressed anywhere in the ground-truth sentence. Both show the same pattern: ST-Net/BLEEP lead briefly then collapse once past their confident range, only ∼28% of their 1k panel is expressed per spot, so their true positives run out well before the panel's edge (markers show each list's end), while our long-LoRA model ($m \in \{2048, 8192\}$) degrades gracefully instead, crossing over earlier under (B). The hybrid (cyan) splices a hard ST-Net prefix (K=100) into our decoder: it inherits ST-Net's early quality, then keeps our long-horizon stability where ST-Net alone collapses.

Table 1: Secondary calibration against fixed-panel baselines on a shared 1k-HVG universe (HEST test spots, n=1310); not a leaderboard comparison (Sec. 4.5). Dense baselines keep only predicted log1p CPM >0 genes, matching cell sentences' zero-omission; Ours is unchanged (already emit-only). Each cell is the mean ± the half-width of a normal 95% CI ($1.96\,s/\sqrt{n}$) over spots. PCC(inv): Pearson of $1/\mathrm{rank}$ over all 1k genes (absent → 0). Prec/F1 on expressed $\mathrm{GT_panel}$ only. NDCG@50: relevance = $1/\mathrm{GT}$ rank (absent → 0); short lists are zero-padded. **Bold** = best mean per column, underline = second best.

Method	J@10	NDCG@50	PCC(inv)	Prec	F1
ST-Net	**0.394**±0.013	**0.659**±0.014	**0.580**±0.017	0.360±0.008	0.508±0.009
BLEEP	0.382±0.013	0.644±0.015	0.568±0.017	0.439±0.010	**0.585**±0.010
HisToGene	0.161±0.008	0.332±0.011	0.233±0.012	0.298±0.007	0.442±0.009
Ours (m=2048+LoRA)	0.355±0.013	0.599±0.014	0.548±0.017	**0.734**±0.013	0.575±0.008
Ours (m=8192+LoRA)	0.314±0.013	0.539±0.014	0.492±0.016	0.636±0.013	0.581±0.008

ST-Net scores ≤ 0 are dropped; BLEEP/HisToGene switch from full-panel ranking to the positive subset (Cov(pred) 1.00→0.65/0.95). Ours lists are generative and already sparse. Recall is omitted: dense baselines sit at 0.983–0.998, so precision and F1 carry the set-overlap comparison. m=128 is omitted (near-zero panel coverage) and discussed in Sec. 4.5 instead.

visible averaging effect, smoothing over spatial structure the ground truth clearly has; our predictions track this spatial structure more closely, consistent with the Moran's I comparison in Table 2.

4.6 Controls: Histology, Prior, and Metadata

Blind controls shuffle GPFM embeddings while holding organ/technology fixed: on the m=512 C2S twin, J@10 drops from 0.262 to 0.082 (∼69% relative drop), so correct histology pairing is required for ranked gene identity even given metadata. The same controls at the long-horizon checkpoint (m=8192+LoRA, 200k steps; Fig. 7, Appendix B) show the gap persists deep into the sentence rather than closing as the frozen C2S prior takes over: shuffle-keep-meta, zero-, and

noise-embedding all stay well below the real-image Jaccard from k=250 to 2000, so the long tail characterized in Sec. 4.4 remains image-conditioned, not just a longer completion of a generic prior.

Replacing C2S-Scale-Gemma-2-2B with vanilla Gemma-2-2B under the same adapter, GPFM features, m=512, and n=1310 protocol collapses J@10 from 0.262 to 0.071, near the histology-shuffle floor (0.082), so the cell-sentence prior is load-bearing too, not just histology pairing.

5 DISCUSSION

What Visual Soft-Prompts buy. The blind-histology and Gemma-2 controls (Sec. 4.6) together argue that performance is not explained by metadata leakage or a generic language prior alone, and Fig. 4 (Appendix B) extends the same logic to training compute: continued optimization of the frozen adapter cannot substitute for attention LoRA at long horizons.

Visual capacity and attention LoRA are not independent levers. A crossed sweep of vision-token width (k=16 vs. k=64) against attention LoRA at m=2048 (Fig. 8, Appendix B) shows the two do not simply add: k=64 helps the frozen decoder but the gap largely closes once LoRA is applied, attention LoRA, not extra visual capacity, is the lever that fixes the long-horizon bottleneck.

Technology-token conditioning is a secondary, spatially uncorrelated modulator. On a held-out lung Xenium slide (NCBI866, n=1124 spots, long-LoRA checkpoint), forcing the technology soft-prompt to Visium rather than the native Xenium token changes predictions less than ordinary spot-to-spot biological variation (same-spot swap-vs-native J@10 0.796 vs. spatial-neighbor-vs-native 0.582). The effect is biologically coherent (Visium shifts mass toward secretory markers, Xenium toward resolved stromal/vascular identity) and is a global slide-level shift with spatially uncorrelated local perturbations, as expected for a one-hot identical at every spot on a slide. Histology remains the dominant, necessary signal; the quantitative token-swap analysis above carries the causal claim about technology conditioning, and Fig. 3 (Appendix B) is illustrative only.

Limitations. We do not compare against a regressor trained on the same GPFM features, so our comparisons do not isolate the decoder's contribution from encoder quality; all results are from single training runs with no seed variance. Metrics score rank order, not absolute expression magnitude, and Visium ground truth is itself a spot-level mixture rather than single cells. Vision conditioning uses one global GPFM embedding with no spatial context, which limits per-marker map quality (Fig. 3, Appendix B), the spatial-structure result in Table 2 does not imply better local accuracy, and adding explicit spatial context is a natural next step. GPFM and C2S-Scale-Gemma-2-2B are fixed choices rather than the strongest available encoder/decoder pair.

6 CONCLUSION

We presented Visual Soft-Prompts: a parameter-efficient interface that conditions a frozen cell-sentence LLM on a frozen pathology foundation model to generate ranked gene lists from H&E. Useful morphology→rank signal concentrates in a finite early signature; lengthening a frozen decoder yields autoregressive derailing that neither decode-time constraints nor continued training of the frozen model repair, a matched-length, matched-budget control shows frozen m=8192 getting *worse*, not better, from 12k to 100k steps. Attention LoRA combined with longer training stabilizes generation to a median length of ∼2.1k genes, approximately matching the ground-truth sentence length on the evaluated spots, while maintaining useful agreement at early ranks. On a shared HVG panel, generative lists calibrate against, but do not dominate, classical panel regressors and retrievers; their advantage is selective naming: they are selective by default, without a tuned abstention threshold, and can continue open-vocabulary generation past the panel's edge. Dual evaluation (full sentence vs. shared panel), histology / prior controls, and a technology-token ablation make this trade-off explicit and and show that histology pairing is necessary for the predictions even given metadata. A practical recipe follows: deploy short-m models for early signatures, and long-m attention LoRA when near-complete ranked transcriptomes are required. The same prefix-continuation interface that powers the hybrid arm (Sec. 4.4) points to downstream uses beyond prediction alone, e.g. extending an existing HVG panel or a dense regressor's output into the unmeasured tail, or composing with other tools' predictions rather than only replacing them.

AI USE STATEMENT

Generative AI tools (a large language model assistant) were used during manuscript preparation to help draft and edit prose, restructure figures and tables, and cross-check numerical claims against the underlying result files; all such output was reviewed line-by-line against the source data before inclusion. No generative AI tool was used to design experiments, run training or evaluation jobs, or generate results. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses only de-identified, publicly available histology–spatial-transcriptomics pairs from the HEST-1k collection; no new human-subjects data were collected. We see the primary societal benefit as reducing the cost of accessing molecular information from archival H&E slides, particularly in settings where spatial transcriptomics is not routinely available. We caution against clinical use of model predictions without prospective validation: reported metrics are rank-fidelity measures on a research benchmark, not diagnostic accuracy, and the model can derail under long free-running decoding (Sec. 4.2) in ways that would not be visually obvious downstream without the diagnostics we report.

REPRODUCIBILITY STATEMENT

Architecture and training details are given in Sec. 3 (adapter dimensions, LoRA configuration, parameter counts); the evaluation protocol, including the patient-level train/test split and the frozen HVG panel construction, is described in Sec. 3.2–3.4. All reported metrics use either the full n=1310 HEST test-spot protocol or an explicitly labeled lean n=262 subset (†); we do not mix the two within a comparison. HEST-1k is a public dataset; GPFM, C2S-Scale-Gemma-2-2B, ST-Net, BLEEP, and HisToGene follow their original public releases. Code and configuration files for training, inference, and all evaluation scripts referenced in the text and appendix will be released alongside the camera-ready version. Full gene-symbol mapping and per-length compute/hardware details are given in Appendix A.

ACKNOWLEDGMENTS

Omitted for anonymous review.

REFERENCES

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3.

Weiqing Chen, Pengzhi Zhang, Tu N. Tran, Yiwei Xiao, Shengyu Li, Vrutant V. Shah, Hao Cheng, Kristopher W. Brannan, Keith Youker, Li Lai, Longhou Fang, Yu Yang, Nhat-Tu Le, Jun-ichi Abe, Shu-Hsia Chen, Qin Ma, Ken Chen, Qianqian Song, John P. Cooke, and Guangyu Wang. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582, July 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02707-1. URL https://www.nature.com/articles/s41592-025-02707-1.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on

single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02305-7. URL https://www.nature.com/articles/s41592-024-02305-7.

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, August 2020. ISSN 2157-846X. doi: 10.1038/s41551-020-0578-x. URL https://www.nature.com/articles/s41551-020-0578-x.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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 *Forty-Second International Conference on Machine Learning*, June 2025. URL https://openreview.net/forum?id=Ossg1IbHDT.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 53798–53833. Curran Associates, Inc., 2024. doi: 10.52202/079017-1704. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/60a899cc31f763be0bde781a75e04458-Paper-Datasets_and_Benchmarks_Track.pdf.

Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Daniel Levine, Syed A. Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique De Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David Van Dijk. Cell2Sentence: Teaching Large Language Models the Language of Biology. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 27299–27325. PMLR, July 2024. URL https://proceedings.mlr.press/v235/levine24a.html.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02856-4. URL https://www.nature.com/articles/s41591-024-02856-4.

Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Jinbang Li, Fang Yan, Yu Cai, Zhengjie Zhu, Cheng Jin, Yi Lin, Xinrui Jiang, Chenglong Zhao, Danyi Li, Anjia Han, Zhenhui Li, Ronald Cheong Kin Chan, Jiguang Wang, Peng Fei, Kwang-Ting Cheng, Shaoting Zhang, Li Liang, and Hao Chen. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *Nature Biomedical Engineering*, 10(3):545–564, March 2026. ISSN 2157-846X. doi: 10.1038/s41551-025-01488-4. URL https://www.nature.com/articles/s41551-025-01488-4.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193.

Jiarui Ouyang, Yihui Wang, Yihang Gao, Yingxue Xu, Shu Yang, and Hao Chen. GenAR: Next-scale autoregressive generation for spatial gene expression prediction, November 2026. ISSN 1361-8415. URL https://www.sciencedirect.com/science/article/pii/S1361841526003014.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv : the preprint server for biology*, 2021. doi: 10.1101/2021.11.28.470212. URL https://www.biorxiv.org/content/early/2021/11/28/2021.11.28.470212.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Curtis Jamison Perry, Ivan Vrkic, Nicole Mayerli Constante, Zirui Fu, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, Chang Li, Emily Sun, David Jeong, Lawrence Zhao, Jennifer Kwan, David Braun, Brian Hafler, Hattie Chung, Rahul M. Dhodapkar, Paul Jaeger, Bryan Perozzi, Jeffrey Ishizuka, Shekoofeh Azizi, and David van Dijk. Scaling Large Language Models for Next-Generation Single-Cell Analysis. *bioRxiv*, pp. 2025.04.14.648850, January 2026. ISSN 2692-8205. doi: 10.1101/2025.04.14.648850. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12632461/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee

Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, October 2024. URL http://arxiv.org/abs/2408.00118.

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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70626–70637. Curran Associates, Inc., 2023. doi: 10.52202/075280-3095. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/df656d6ed77b565e8dcdfbf568aead0a-Paper-Conference.pdf.

A DATA PROCESSING AND COMPUTE DETAILS

Gene-symbol mapping (Sec. 3.2). HEST count matrices expose ∼326k unique raw column names: genome-assembly prefixes (e.g. `mm10___`, `GRCh38_`), Ensembl IDs with version suffixes, Xenium codewords/controls, mitochondrial naming variants, and mixed human/mouse aliases. We normalize every column through a global mapping: drop probe/control/ambiguous artifacts; strip assembly prefixes; resolve Ensembl IDs (and h5ad `gene_ids` hints) to HGNC symbols; canonicalize mitochondrial symbols (e.g. `mt-Co1`/`MT.CO3` → `MT-CO1`/`MT-CO3`); look up remaining aliases in an HGNC reference; and resolve legacy mouse symbols via mygene. Unresolved locus/accession IDs are dropped. On the full inventory this keeps ∼256k columns (∼78%) and drops ∼71k (∼22%; mostly unmapped Ensembl, versioned loci, and accession/clone IDs); many KEEP rows collapse onto shared symbols (∼230k collapses, e.g. `Alb`/`ALB` → `ALB`), with counts summed before ranking.

Compute (Sec. 3.3). Shorter- and mid-context runs fit on single NVIDIA A100 GPUs (40GB or 80GB) at batch 1; the longest m=8192 LoRA trains use two Blackwell RTX PRO 6000 GPUs at global batch 2, because peak memory is dominated by attention over the long gene-sentence context, step counts are therefore not example-matched across lengths (Sec. 5 limitations). Inference for reported metrics is run on A100s. Checkpoint selection uses validation NLL.

B ADDITIONAL FIGURES

Supplementary companions to the figures and tables in the main text, referenced from Sec. 4.2–4.3 and Sec. 5.

Table 2 quantifies this qualitative view: on these two slides, Ours tracks which genes are spatially structured, while the baselines are smooth almost everywhere regardless of the underlying biology.

Table 2: Spatial autocorrelation (Moran's I, k=6 neighbor graph on spot coordinates) of predicted vs. ground-truth expression, pooled over 669 panel genes with nonzero variance in the ground truth and every method, on the two whole-slide cases in Fig. 3 (NCBI866, MISC109; n=2 slides, preliminary). "r with GT" is the correlation between a method's per-gene Moran's I and the ground truth's per-gene Moran's I: whether a method reproduces *which* genes are spatially structured, not merely how smooth its maps are on average. Ground-truth mean Moran's I is itself low (0.120), most genes are not strongly spatially patterned, so a uniformly smooth method inflates its own mean without tracking the genes that actually are; ST-Net, BLEEP, and HisToGene are smooth almost everywhere (mean $I \geq 0.52$) with near-zero or negative r, while Ours matches the ground-truth mean and correlates with it. This does not indicate better local map accuracy: spot-wise correlation with the ground truth on NCBI866 is tied across methods (≈0.31–0.35). HisToGene's whole-slide pass is also outside its training regime (packs of ≤32 spots), likely contributing to its over-smoothing.

Method	Mean Moran's I	r with GT Moran's I
Ground truth	0.120	,
Ours	0.124	0.636
ST-Net	0.543	−0.030
BLEEP	0.526	−0.139
HisToGene	0.656	−0.403

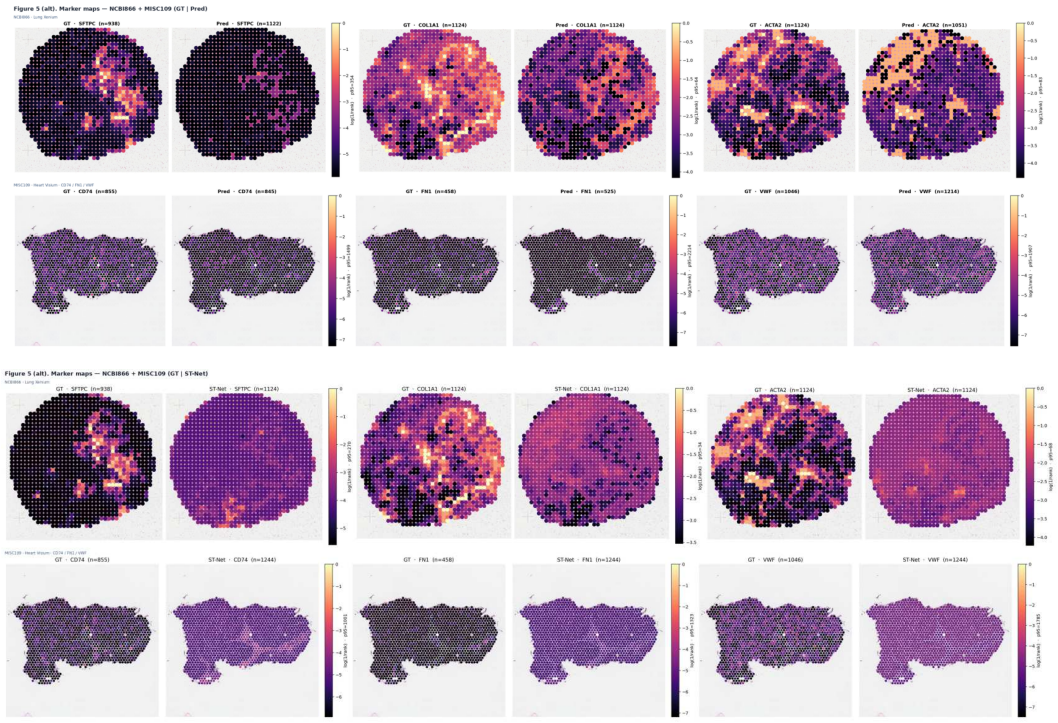


Figure 3: Whole-slide marker maps on held-out test slides (illustrative only; Sec. 5), GT | Ours (top) vs. GT | ST-Net (bottom). Each cell shows GT | Pred ($\log(1/\mathrm{rank})$; joint p95 ceiling; fill-empty). Rows: NCBI866 (lung Xenium; SFTPC, COL1A1, ACTA2) and MISC109 (heart Visium; CD74, FN1, VWF). Checkpoint: long LoRA (m=8192). Train slides (TENX94, MISC30) excluded. MISC109 markers are selected for spatial contrast rather than a priori; see Sec. 5 limitations. ST-Net's maps are visibly smoother and lack the sharp spatial boundaries the ground truth shows, most clearly for SFTPC; Table 2 quantifies this.

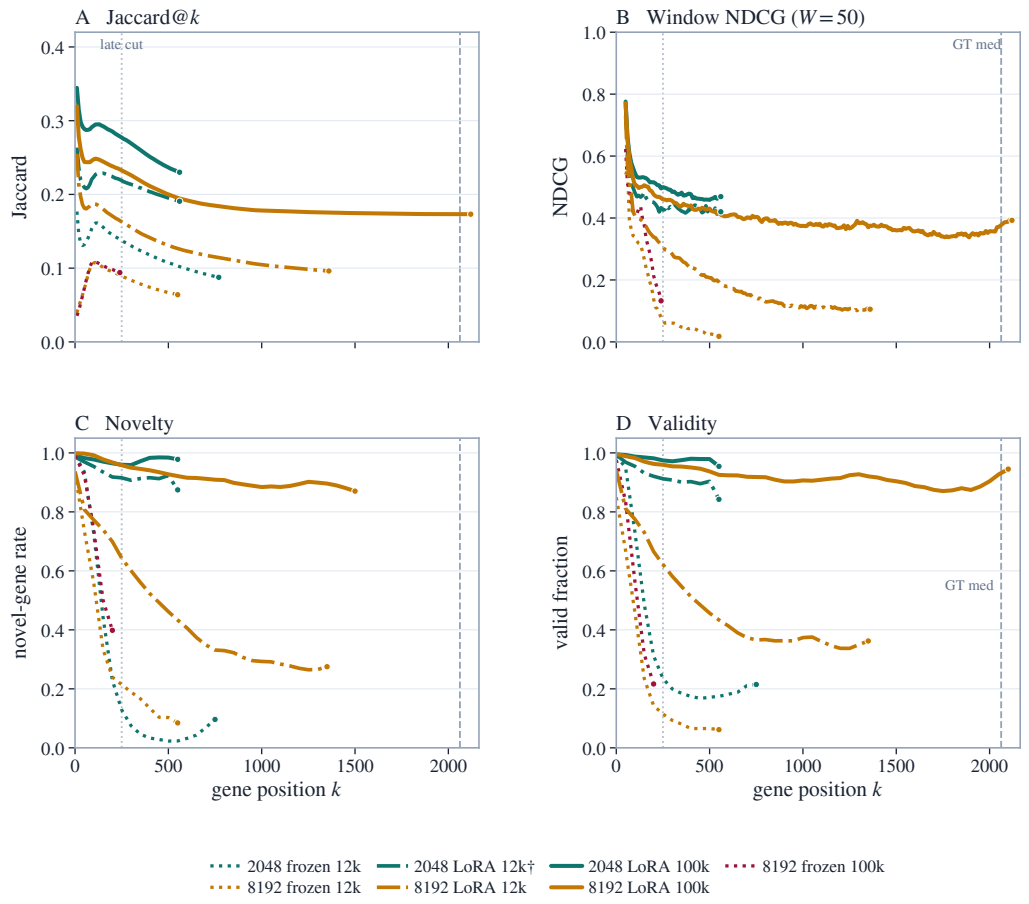


Figure 4: Position-resolved view of the frozen-vs-LoRA budget ablation (numbers in Sec. 4.3): Jaccard@k / windowed NDCG (W=50; relevance = 1/GT-rank) / novelty / validity at $m \in \{2048, 8192\}$ across frozen@12k, LoRA@12k, and LoRA@100k, plus frozen@8192@100k. Frozen m=8192 stays collapsed at both budgets; attention LoRA restores late hygiene and early ranks with budget. † Lean n=262 for m=2048 LoRA@12k only; all other curves (m=2048 frozen@12k, m=8192 frozen/LoRA@12k/100k) use the full Table-1 protocol (n=1310).

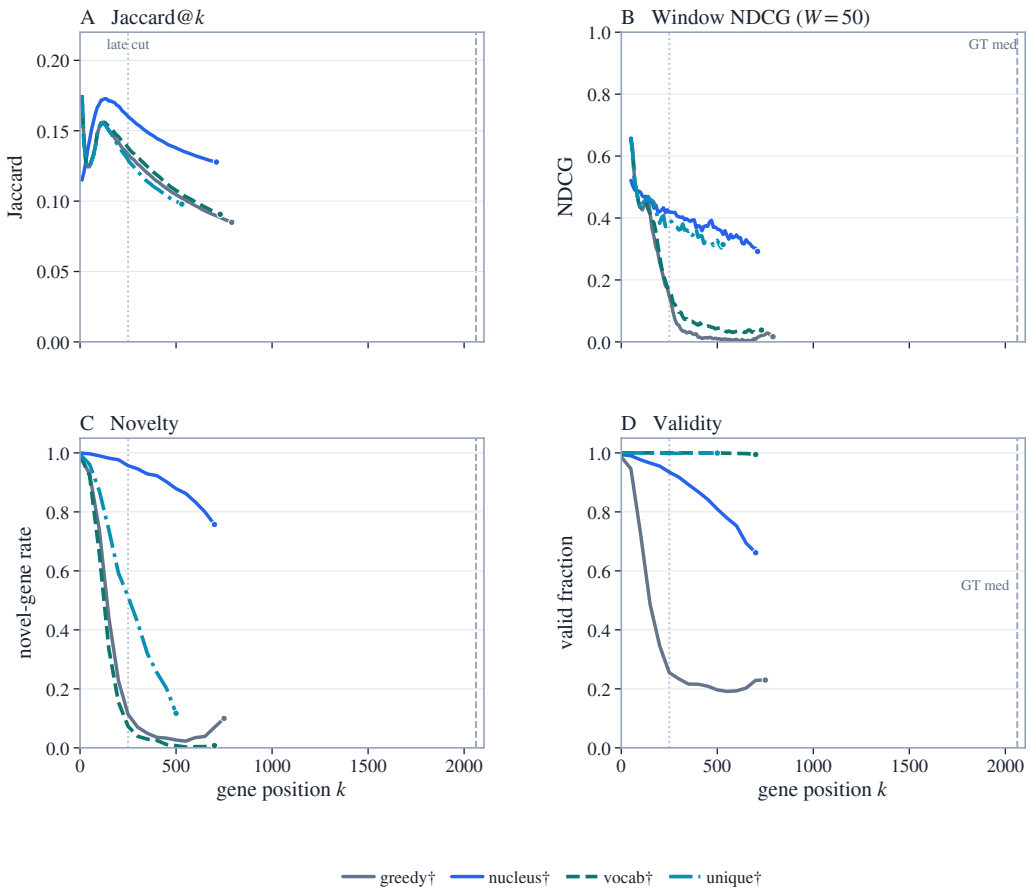


Figure 5: Decode hygiene overview at frozen m=2048 (lean † n=262, 12k): Jaccard@k / windowed NDCG (W=50; relevance = 1/GT-rank) / novel-gene rate / gene validity for greedy, nucleus, vocab, and unique. Vocab/unique hold late validity ≈1 (D) while Jaccard tracks greedy (A) and novelty still collapses by ∼250 (C): valid tokens without biological lift. Nucleus slows novelty/validity decay but hurts the earliest ranks; unique's novelty is not pegged at exactly 1 since its constraint is not fully airtight. Curves stop at each arm's median length (k/bins with <25% support omitted).



Figure 6: Table 1's PCC(inv) broken down by technology × organ, for Ours (m=8192+LoRA, top), ST-Net (middle), and BLEEP (bottom). Panel ranks, absent→0, over the frozen 1k HVG panel, technologies collapsed to a per-model map; all three panels use the full Table-1 protocol (n=1310). All three methods show the same qualitative pattern, strongest correlation on Spinal cord/Heart/Bowel, weakest on Bladder/Cervix/Lymphoid, indicating the aggregate Table 1 gap is not driven by a handful of outlier organs or technologies for any one method.

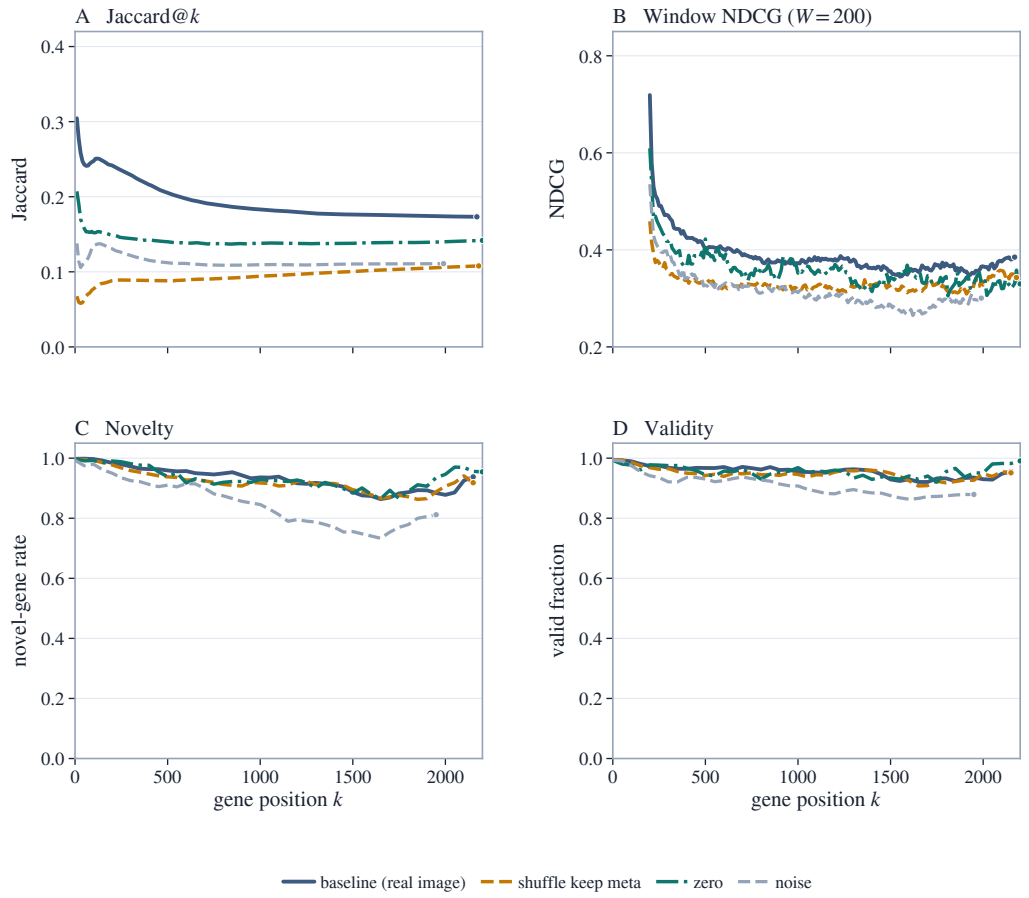


Figure 7: Blind-histology controls at the long-horizon checkpoint (m=8192+LoRA, 200k steps; lean n=262): open-vocabulary Jaccard@k, sliding-window NDCG (W=200), novel-gene rate, and valid-gene fraction for real image, `shuffle_keep_meta`, zero-embedding, and noise-embedding arms (organ/technology tokens kept on every arm). Real image sustains Jaccard ∼0.17–0.24 from k=250 to 2000; all three image-ablation arms drop and stay lower throughout the same range (e.g. shuffle keep-meta 0.089–0.106), confirming the long tail characterized in Sec. 4.4 remains image-conditioned, not only carried by the frozen C2S prior.

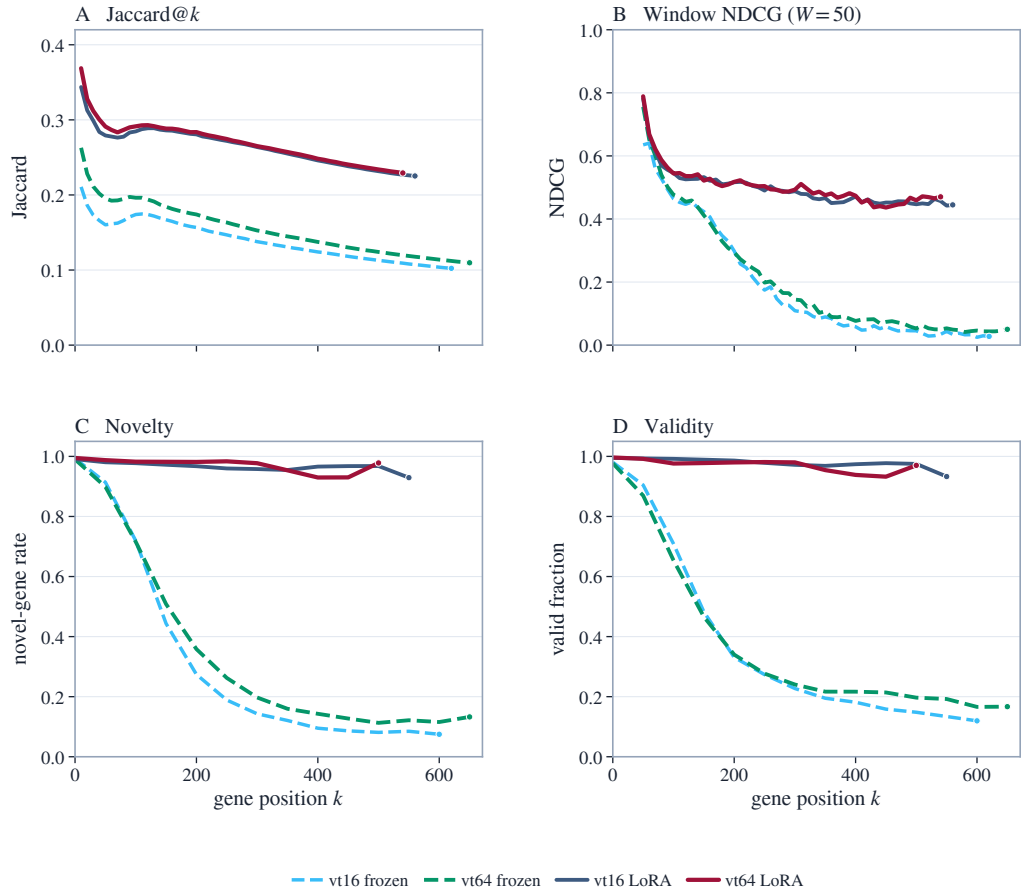


Figure 8: Vision-token width (k=16 vs. k=64) crossed with attention LoRA, m=2048, 100k steps (lean n=262): Jaccard@k, windowed NDCG (W=50), novelty, and validity for all four frozen/LoRA $\times$ k=16/k=64 arms. Interpretation in Sec. 5: LoRA closes most of the frozen-decoder rank-quality gap between k=16 and k=64, while k=16+LoRA is the more length-stable arm on novelty/validity.