SPATIALDATAAGENT: AUTONOMOUS CURATION OF A DECADE OF PUBLIC SPATIAL OMICS AT SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics maps gene activity within tissue sections. When paired with tissue images, these measurements connect molecular activity to visible tissue structures, providing a basis for studying cellular organization and developing image-based molecular analyses. Yet public spatial omics data often remain difficult to reuse—stored across non-standard archives where coordinates, histology images, and molecular counts frequently diverge, use proprietary containers, or are mixed with dissociated single-cell data. Here, we present **SpatialDataAgent**, an autonomous system that converts public deposits into standardized multimodal datasets at scale. SpatialDataAgent is designed to preserve source-deposited measurements through source-grounded coordinate transformations and explicit data-preservation constraints: it grounds coordinate mappings in deposited metadata, uses a catalog of 26 specialized domain tools for spatial formats, and checks candidate alignment through automated visual diagnostics. Evaluated on 400 reference studies, our screening module identifies eligible datasets with 91.75% classification accuracy and a 0.913 macro-F1 score. In a controlled evaluation across 40 diverse GEO studies (331 samples per condition), domain workflow support increases human-reviewed visual alignment from 54.7% to 77.6% (+23.0 percentage points; 95% bootstrap CI: 8.6–45.3), reaching **81.0%** when coupled with visual quality feedback. Using SpatialDataAgent, we construct and release **HESRT (Histology-paired Expression atlas of Spatially Resolved Transcriptomics)**, an open atlas of 4,468 human-reviewed multimodal samples comprising 37.05 million spatial observations across multiple species and technologies. HESRT makes matched images and gene measurements available for studies of tissue organization and image-based molecular modeling.

1 INTRODUCTION

Spatial transcriptomics, introduced in 2016 (Ståhl et al., 2016), measures gene expression at spatially resolved locations within tissue sections. When paired with a histology image from the same sample, it lets researchers relate gene activity to visible structures such as tumor regions and immune-cell-rich areas (Jaume et al., 2024; Chen et al., 2024; Zhang et al., 2025). This pairing can reveal how cellular states vary across tissue and help identify spatial patterns worth testing for clinical relevance (Zhang et al., 2025; Zou et al., 2026). It also supports models that infer molecular features from routine, lower-cost H&E slides (Schroeder et al., 2025; Zou et al., 2026; Liu et al., 2026), potentially extending such analyses to larger collections of tissue images. Over the past decade, researchers have deposited thousands of spatial transcriptomics runs into public repositories such as the NCBI Gene Expression Omnibus (GEO).

In a public study record, a tissue image, the positions where genes were measured, and the gene counts may be stored in different files and formats. Some studies also include dissociated single-cell data with no tissue positions, and the recorded spatial positions may refer to a resized image rather than the original scan. A researcher must work out which files belong to each sample, recover the correct image scale, and check that the positions line up with the tissue. If those files or scales are mismatched, gene activity can appear in the wrong tissue region. Doing this manually for each study slows the reuse of public data. Existing databases make many studies easier to discover (Yuan et al., 2023; Xu et al., 2024; Li et al., 2025; Wang et al., 2024), but many records still need sample-level work before they can be used as reliable image–gene pairs.

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.

Large language model (LLM) agents offer a practical approach to automating these scientific workflows (Xiao et al., 2026; Xu et al., 2026; Bu et al., 2026). Recent frameworks show that autonomous agents can plan, generate code, execute tools, and debug errors within interactive computing environments (Chen et al., 2025; Bragg et al., 2026; Chan et al., 2025). Scientific data curation depends on preserving source measurements during file parsing and coordinate recovery. SpatialDataAgent carries deposited expression values and sample identities through reconstruction, derives coordinate relationships from source records, and records the evidence supporting each mapping.

To address this challenge, we introduce SpatialDataAgent, an autonomous agent framework designed for large-scale curation of public spatial omics data (Figure 1). SpatialDataAgent connects study screening to sample-level reconstruction. Its workflow guides the agent to identify files from the same sample, recover coordinate mappings from deposited metadata, and check the resulting image–expression pairs. Three design principles support this process:

- **Evidence-based screening**: A dedicated screening module reviews study descriptions, protocols, and supplementary file trees, separating paired spatial assays from dissociated single-cell samples before code execution begins.
- **Grounded program synthesis**: The reconstruction agent writes code using an auditable library of 26 domain tools across five functional modules and derives coordinate and scale mappings from documented scale factors, homographies, or array geometries.
- **Integrated diagnostic quality control**: The agent inspects candidate alignments via automated visual checks to detect coordinate misplacements while maintaining an independent evaluation protocol to verify final data fidelity.

Through an evaluation across 40 diverse GEO studies (331 required samples per condition), we show that domain-guided workflow support increases human-reviewed visual alignment from 54.7% to 77.6% (+23.0 percentage points), reaching 81.0% under joint workflow and diagnostic support. Finally, we deploy SpatialDataAgent across public repositories spanning 2016 to 2026 to build and release HESRT (Histology-paired Expression atlas of Spatially Resolved Transcriptomics), an open atlas of 4,468 human-reviewed multimodal samples and 37.05 million spatial observations across diverse species and platforms.

2 Related Work

AI Agents for Scientific Workflows. Language model agents are increasingly used to automate scientific analysis and experimental planning (Miao et al., 2026; Roohani et al., 2025). Recent systems such as CellAgent (Xiao et al., 2026) and MedAgentGym (Xu et al., 2026) show that LLMs can write code for single-cell analyses and medical data science tasks. BioDiscoveryAgent (Roohani et al., 2025) demonstrated that language models can guide closed-loop genetic perturbation experiments. While these works focus on downstream analyses (such as cell clustering or target prediction), SpatialDataAgent addresses the upstream foundation: assembling raw public archives into standardized, analysis-ready datasets.

Agent Evaluation on Grounded Tasks. Evaluating scientific agents requires sandboxed environments and realistic execution checks (Jimenez et al., 2024; Chen et al., 2025; Fa et al., 2026). Benchmarks like ScienceAgentBench (Chen et al., 2025) and BioAgent Bench (Fa et al., 2026) emphasize evaluating executable code alongside the scientific plausibility of outputs. We build on these principles by assessing code execution, coordinate agreement with histology, and preservation of original gene counts.

Public Spatial Transcriptomics Collections. Curated collections combining histology and gene expression, such as HEST-1k (Jaume et al., 2024) and STimage-1K4M (Chen et al., 2024), have provided valuable paired data for multimodal machine learning. Such pairs have supported studies of cellular organization in tumors (Zhang et al., 2025), gene-activity inference from routine images (Zou et al., 2026), and integration of molecular measurements across tissue sections (Liu et al., 2026). Several online portals, including SODB (Yuan et al., 2023), STOmicsDB (Xu et al., 2024), SOAR (Li et al., 2025), and CROST (Wang et al., 2024), organize public spatial studies. However, many public records remain non-standardized or lack full-resolution image pairing. SpatialDataAgent targets these irregular source archives by reconstructing paired images, coordinates, and counts

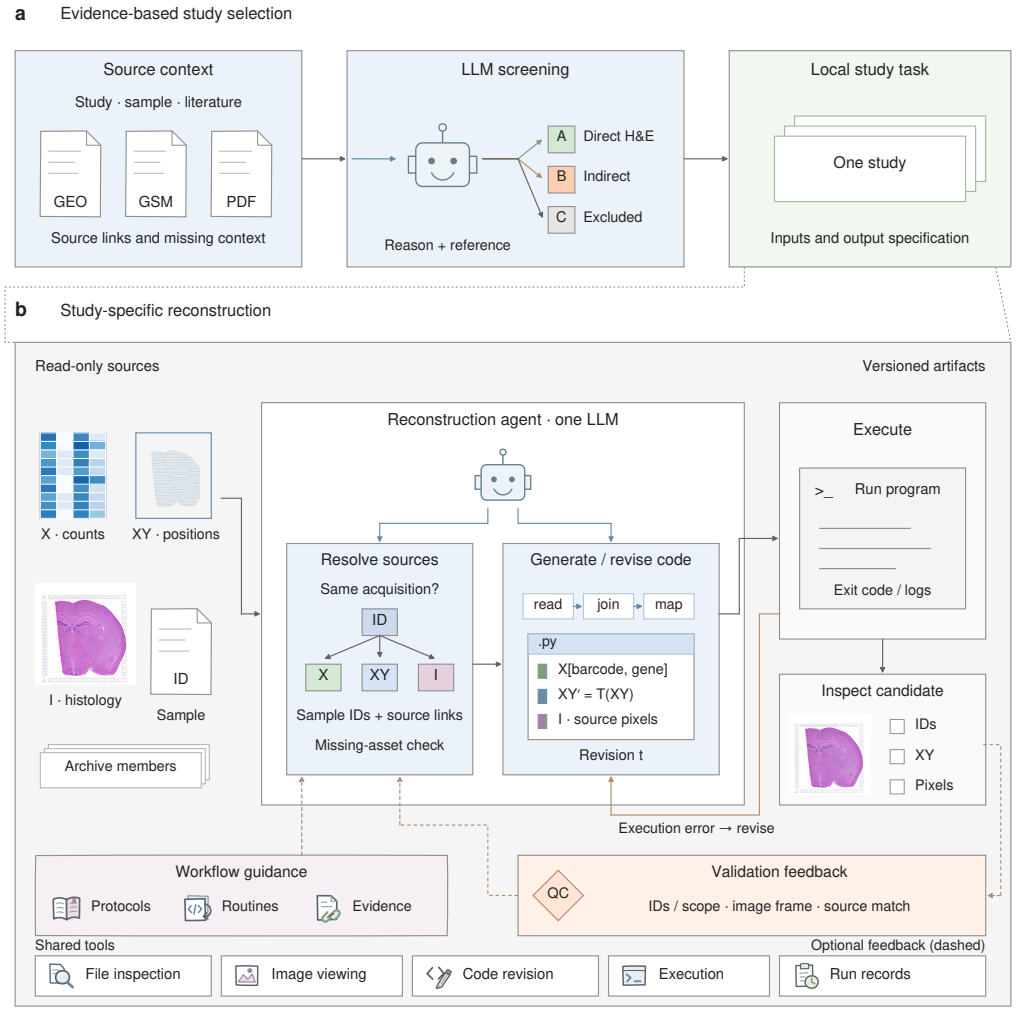


Figure 1: **Overview of SpatialDataAgent.** (a) The screening module reads study and sample records and linked publications, then classifies each study as direct H&E support, indirect or unresolved morphology, or excluded or unsupported. It records the evidence behind the decision before reconstruction begins. (b) For a selected study, the reconstruction agent links deposited gene counts, spatial positions, and the tissue image to the same sample. It writes and runs a program to map the positions onto the image, inspects the result, and revises the code when needed, without changing the deposited source files. Dashed arrows show optional workflow guidance and quality-control feedback. Additional tools and diagnostics are shown in Figure S2.

with source manifests, complementing existing curated collections. SpatialData (Marconato et al., 2025) provides a related standard for representing spatial data.

3 The SpatialDataAgent Framework

3.1 Task Formulation and Data Integrity

For a given study, the inputs consist of the deposited metadata (protocols, sample descriptions, publications) and the raw files. The goal is to generate an executable Python script that processes these raw files into clean, paired samples. Each reconstructed sample consists of four standardized elements:

- **Tissue Image**: The tissue image (e.g., brightfield H&E or fluorescence).
- **Image Coordinates**: The location of each spot, bin, or cell in the image pixel frame.
- **Expression Matrix**: Source-deposited expression values associated with unique observation IDs, with their processing state recorded.
- **Provenance Trail**: A recorded log of file paths, checksums, and transformation steps linking the output directly to the original archive.

SpatialDataAgent parses deposited files, applies coordinate transformations supported by each acquisition, and records file paths, checksums, and transformation steps. Shared observation IDs link expression values to coordinates and the selected histology image; the set of in-tissue observations follows the source definition.

3.2 Screening Candidate Studies

Before generating code, SpatialDataAgent reads the study title, abstract, experimental protocols, supplementary file names, and linked publication text to classify the study into one of three operational categories:

- **Class A (Direct H&E Support)**: Studies with a spatial assay supported by explicit H&E evidence or characteristic Visium/Space Ranger morphology files.
- **Class B (Indirect or Unresolved Morphology)**: Studies with eligible spatial assays but indirect or unresolved morphology linkage, including DAPI or immunofluorescence without explicit H&E evidence.
- **Class C (Excluded or Unsupported)**: Studies with excluded assay types or insufficient evidence for an eligible spatial assay–morphology combination.

For studies containing multiple assay types, classification follows the evidence for the relevant spatial assay. Evidence from dissociated single-cell assays is evaluated separately from spatial-sample evidence.

This initial screen prioritizes records with eligible spatial data and retains structured notes for future data recovery.

3.3 Program Synthesis with Domain Tools

To reconstruct a study, a coding agent works within an isolated Python and bash environment. In our workflow-assisted setting, the agent is provided with documentation and a library of 26 specialized domain tools organized into five practical modules:

- **Visium Coordinate Tools (7 tools)**: Read position tables and named full-resolution pixel coordinates, and recover or apply source-supported mappings from rendered annotations, overlays, or selections.
- **Platform Adapters (6 tools)**: Compute coordinate transformations and offsets for alternative platforms such as Xenium, CosMx, and Stereo-seq.
- **Provenance Tools (4 tools)**: Record source paths, hashes, and supporting evidence for scale-factor mappings, identity mappings, rendered annotations, and embedded coordinate frames.
- **Native Frame Checker (1 tool)**: Checks linked observation IDs, coordinates, expression, and image components within the same native source frame.
- **Cloupe Decoders (8 tools)**: Read `.cloupe` source containers to extract expression matrices, feature names, coordinates, and image tiles, construct AnnData objects, split recorded expression summaries, and verify native image pixels.

The agent uses these tools to read source formats, identify sample components, and recover coordinate mappings from deposited metadata. It then writes and runs a study-specific reconstruction program, inspects the resulting overlay, and revises the code when needed.

3.4 Diagnostic Quality Control

When diagnostic quality control (QC) is enabled, the agent can call an automated checker to evaluate its candidate outputs before finishing:

- **Background Spot Check**: Because active tissue spots should lie on tissue rather than empty glass, the checker identifies the top 10% highest-expression spots and examines their image neighborhood. If more than 3% of these spots fall on uniform blank background, an alignment failure is returned.
- **Tissue Coverage Check**: The checker calculates whether the spatial array covers the segmented tissue area, alerting the agent if large sections of tissue appear unmeasured.

The agent uses these diagnostic reports to adjust its code and correct coordinate alignments during the execution trajectory.

4 Experimental Setup

We report screening accuracy on the 400-GSE reference, reconstruction performance on the 40-GSE benchmark, and resource scale on the HESRT release.

4.1 Screening Benchmark: Reference Set and Protocol

To evaluate study screening, we use the 400-GSE reference (150 Class A, 109 Class B, 141 Class C) accepted as the project gold standard. All records received primary source reads with supplemental AI cross-review and human final verification. Sampling was stratified by historical Muse predictions for retrospective assessment.

We use the same frozen ABC criteria as the screening module; the full specification is provided in Appendix A.

We measure accuracy with study-level agreement and macro-F1: agreement is the share of correct calls, and macro-F1 averages F1 equally over A, B, and C so the large Class C cannot dominate. Per-class precision, recall, and F1 then show where errors fall.

4.2 Reconstruction Benchmark: Cohort and Factorial Design

To evaluate SpatialDataAgent across realistic repository challenges, we selected a stratified benchmark of 40 GEO studies (GSEs) containing 331 required samples (GSMs), with 10 GSEs in each of four archival formats. The cohort is balanced across formats by design:

- Standard Space Ranger directory layouts;
- Non-standard compressed tar/zip archives;
- Containerized files (`.h5ad`, `.rds`, `.cloupe`);
- Non-Visium or alternative high-density spatial arrays.

We evaluated four system configurations on the exact same 40 studies:

- **Base**: The shared Codex-based agent scaffold with Python, bash, image viewing, and execution logs.
- **QC**: Base with automated diagnostic visual checks.
- **Workflow**: Base with workflow guides, 26 domain tools, and dedicated source-image evidence interfaces.
- **Workflow + QC**: The full system combining domain tools, workflow guides, and diagnostic checks.

We ran each of the 40 studies once under each condition, giving 160 agent runs. All conditions used Muse Spark 1.2 via its official API interface with fixed high-reasoning settings and the same source

inputs for each study, with budgets of 3 hours, 5 million tokens, and up to 15 program execution attempts per run. Comparisons involving Workflow assess the combined contribution of its guides, tools, and evidence interfaces.

4.3 Evaluation Metrics

We evaluated all outputs using two main criteria:

- **Human Visual Alignment (Primary Metric)**: Double-blind human review of the final overlay between histological images and spatial coordinates for all 331 required samples per condition. Samples are graded as **PASS** (coordinates accurately match tissue morphology), **FAIL** (detectable misalignment or inverted axes), or **Unavailable** (execution failure or missing output). All 331 samples remain in the denominator.
- **Automated Source Audit (Secondary Metric)**: A strict script checking three criteria: valid file schemas, coordinate accuracy within 1.0 pixel of reference ground truth, and exact preservation of deposited expression values.

We estimated 95% confidence intervals for paired differences in alignment pass rates using 20,000 bootstrap resamples of the 40 studies. Each resample retained all samples and all four condition outcomes within each selected study.

5 Results

5.1 Screening Accurately Identifies Usable Spatial Studies

On 400 reference studies, our screening module achieved 91.8% agreement (367 of 400 studies) and a macro-F1 of 0.913 (Table 1; Figure 2a,b). Disagreements (33 studies) were mostly borderline distinctions between Class A (direct H&E support) and Class B (indirect or unresolved morphology; 24 cases), with only 9 crossing the exclusion boundary (Class C). Applied to a frozen pool of 3,771 candidate GEO studies retrieved by broad keyword screening of 216,510 Series records from 2016–2026 (Appendix A), the module identified 781 Class A, 602 Class B, and 2,388 Class C studies.

Table 1: Screening performance across 400 reference-annotated studies.

Class	Ground truth n	Precision (%)	Recall (%)	F1
Class A (Direct H&E Support)	150	93.2%	90.7%	0.919
Class B (Indirect or Unresolved Morphology)	109	83.9%	86.2%	0.851
Class C (Excluded or Unsupported)	141	96.5%	97.2%	0.968
Overall / Macro-F1	400	—	—	**0.913**

5.2 Domain Tools and Quality Checks Drive Successful Reconstruction

Across 331 required samples from 40 studies, domain workflow support substantially improved human-reviewed visual alignment (Figure 2c; Table 2). The visual alignment pass rate increased from 54.7% (181/331) in the Base agent to 77.6% (257/331) in the Workflow agent—an absolute gain of +23.0 percentage points (95% CI: [8.6, 45.3]). In sample-by-sample comparisons, Workflow brought 79 samples that did not pass under Base to PASS, while regressing on only 3 (Table 2). Gains were largest for standard Visium files and archives: alignment pass rates increased from 60.2% to 98.9% and from 62.2% to 83.2%, respectively (Figure 2d). Improvements were smaller for containers and alternative grids, where Workflow pass rates remained at 52.6% and 48.8%.

Diagnostic QC feedback serves as an automated visual guardrail, complementing the domain workflow tools used for source-grounded coordinate recovery. With QC, visual alignment increased from 54.7% to 57.4% under Base and from 77.6% to 81.0% under Workflow (268/331; Table 2). In highly irregular formats, including complex containers and dense arrays (Figure 2d), visual diagnostics can prompt iterative re-inspection and consume additional execution budget. Workflow + QC used more tokens and reached the budget limit more often than Workflow (Appendix D, Table S5), illustrating a practical trade-off between visual refinement and compute cost.

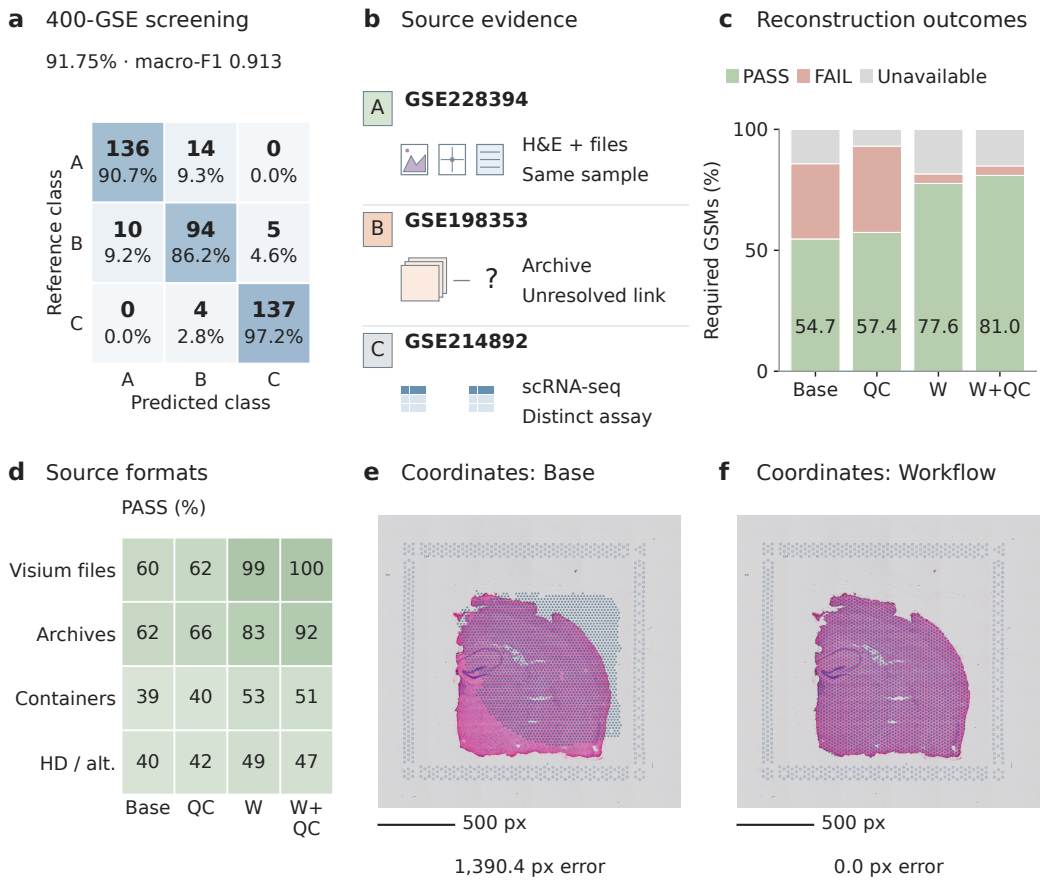


Figure 2: **Screening and reconstruction performance.** (a) The screening module's calls against the 400-GSE reference (150 Class A, 109 Class B, 141 Class C); cell shading encodes row percentage. (b) Illustrative source evidence for Classes A, B, and C; icons denote file types. Expanded source examples are shown in Figure S1b. (c) Complete outcomes for 331 required GSMs per condition; FAIL and unavailable outputs remain in the denominator. W denotes Workflow; W+QC denotes Workflow + QC. (d) PASS percentages in four source-format strata, each containing 10 GSEs; row denominators are 88, 143, 57 and 43 GSMs. Values are descriptive and rounded to whole percentages. (e, f) Base and Workflow overlays for GSE209583/GSM6380080, with 2,380 barcodes, at the same scale. Shown are the unchanged overlays; the Workflow output matches the source-supported coordinates, whereas Base misreads row/column orientations. Errors are maximum source-coordinate discrepancies. Rulers use native image pixels. Screening and reconstruction cohorts are separate.

Table 2: Paired reconstruction comparisons across 40 GEO studies (331 required samples per condition). FAIL and unavailable outputs remain in the denominator. Intervals are GSE-cluster bootstrap 95% CIs.

Comparison	PASS rate: former / latter (%)	Improved / regressed	Difference and 95% CI (pp)
Workflow vs. Base	77.6 / 54.7	79 / 3	**+23.0 [8.6, 45.3]**
Workflow + QC vs. QC	81.0 / 57.4	84 / 6	**+23.6 [9.9, 44.7]**
QC vs. Base	57.4 / 54.7	17 / 8	+2.7 [-3.6, 9.9]
Workflow + QC vs. Workflow	81.0 / 77.6	14 / 3	+3.3 [-1.8, 13.5]

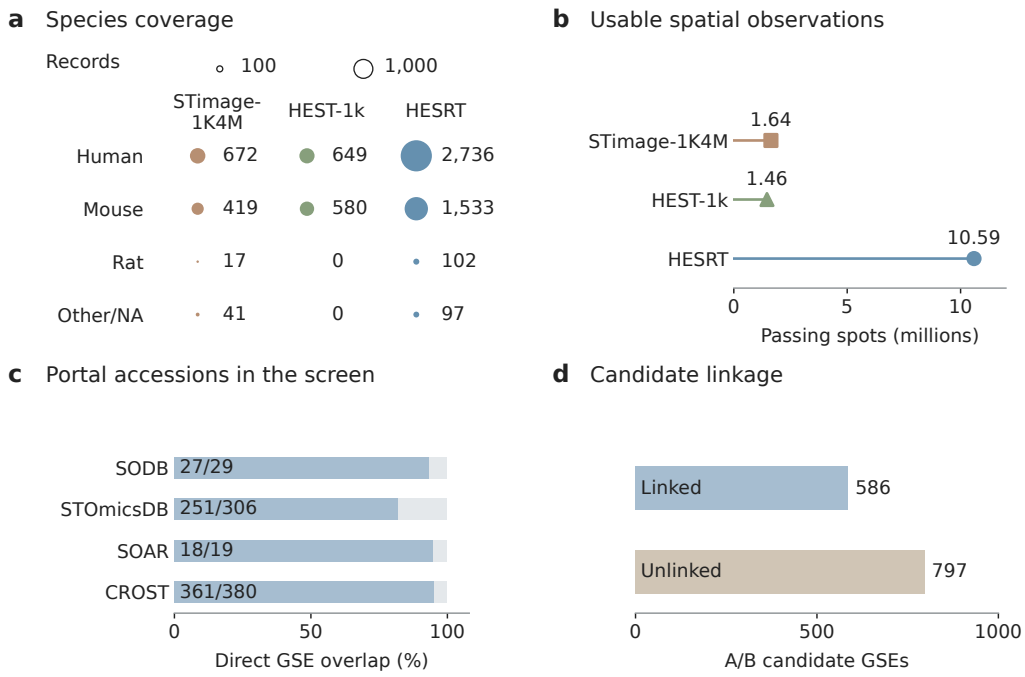


Figure 3: **Resource scale and catalog linkage.** (a) All-platform species composition. Circle area encodes record count; labels give exact counts. Other/NA includes all remaining labels. (b) Human/mouse Visium spots passing four shared geometric checks, for STimage-1K4M, HEST-1k, and HESRT, respectively: 1,636,687; 1,462,322; and 10,592,238. Measurement denominators are in Supplementary Table S8. (c) Direct GSE overlap between each portal snapshot (SODB, STOmicsDB, SOAR, and CROST) and the 3,771-study screening pool; labels show overlap/available GSE identifiers. (d) Among 1,383 operational A/B candidates, 586 are linked and 797 remain unlinked to these four snapshots after accession/publication matching. Panels c,d concern study identifiers; panel a includes all platforms, whereas b uses measured human/mouse Visium artifacts.

5.3 Case Study: Correcting Coordinate Axis Inversion

A clear example of how domain tools help occurs in study GSE209583 (sample GSM6380080), which contains 2,380 spatial barcodes paired with a 2000×1939 histology image of a mouse brain tissue section (Figure 2e,f).

- **Base & QC Agents**: Both agents successfully extracted the count matrix and found the spatial coordinate table. However, because they lacked spatial array parsers, both agents swapped row and column indices ($x \leftrightarrow y$). This transposed the coordinate grid, projecting spots onto empty background glass. Maximum coordinate discrepancies were **1,390.4 pixels for Base** and **1,390.6 pixels for QC**.

- **SpatialDataAgent (Workflow)**: The agent called `load_visium_positions` and read the Space Ranger scale factor file. It recovered the source-supported image coordinates for all 2,380 barcodes, with a maximum coordinate discrepancy of 0.0 pixels.

5.4 The HESRT Resource: Scale and Database Complementarity

Applying SpatialDataAgent across public repositories spanning 2016 to 2026 produced HESRT, an open collection of 4,468 human-reviewed samples across 507 studies, comprising 37,054,005 spatial observations (Figure 3a; Supplementary Table S7).

Paired data at scale. Among the human and mouse Visium samples assessed with the same four coordinate checks, HESRT provides 3,802 passing samples containing 10.59 million spatial spots, compared with 784 samples and 1.64 million spots in STimage-1K4M, and 583 samples and 1.46 million spots in HEST-1k (Figure 3b). These results show the scale at which SpatialDataAgent can turn public deposits into paired image–expression data for reuse.

Coverage of public studies. Our screening pool includes 516 of the 591 unique GEO studies listed across SODB, STOmicsDB, SOAR, and CROST (Figure 3c; Supplementary Table S9). Among the 1,383 studies classified as A or B, 797 remained unlinked to these four portal snapshots after matching accessions and publications (Figure 3d). The screening stage therefore covers many studies represented in existing portals while also identifying a substantial pool of candidates for further reconstruction, complementing catalog-based data discovery.

6 Discussion and Conclusion

SpatialDataAgent shows how domain workflow support can improve autonomous reconstruction of scientific data from heterogeneous public deposits. Using the same model, source inputs, and budget limits, workflow support increased human-reviewed visual alignment from 54.7% to 77.6%, reaching 81.0% with diagnostic feedback. Domain knowledge is supplied through reusable guides and tools, which the agent uses to construct programs adapted to individual studies. This connects flexible code generation with explicit requirements for preserving sample identities, expression measurements, and source-supported coordinate relationships.

More broadly, SpatialDataAgent provides a practical design pattern for agent-based scientific data curation. Experts define the required outputs, provide tools for common domain operations, and specify checks for file structure and consistency with the source data. The agent inspects each input, selects the relevant tools, and writes, executes, and revises a processing program. A common submission and execution interface allows these tasks to be organized into batches, with outputs, logs, and unresolved cases recorded consistently. This design gives experts a way to encode reusable domain knowledge while allowing the agent to handle differences between individual deposits. Independent evaluation remains part of assessing the resulting data.

Applying this workflow to public spatial-omics deposits produced HESRT, an open resource of 4,468 human-reviewed samples containing 37.05 million spatial observations. The resource demonstrates the scale at which agent-assisted reconstruction can make existing measurements available as paired image–expression data. Extending this design to other imaging and omics repositories will require suitable tools, output specifications, and checks for the measurements that must be preserved. Future work can test its transfer across domains and models, and develop targeted expert review for cases that automated processing leaves unresolved.

AI USE STATEMENT

Generative AI tools supported inspection and classification of repository records, scientific data reconstruction, code and figure preparation, literature search and synthesis, manuscript drafting, and methodological and editorial feedback. The experimental agent, its tools and budgets, and the AI-assisted screening reference with human final verification are described in the main text and appendix. Manuscript revisions were reviewed by the author. Frozen experimental outcomes were retained during writing and typesetting. The authors are responsible for the submitted text, claims, citations, code and data artifacts.

REPRODUCIBILITY STATEMENT

Data and code availability. HESRT is available at https://huggingface.co/datasets/Biogod/HESRT. The analyses use release `2026-09-17-v1.3`; the pinned catalog and artifact checksums accompany the supplementary archive. Access to dataset files requires Hugging Face approval and compliance with original-source terms. The anonymous supplementary archive contains the complete paper source, numerical outcomes, metadata catalog, and frozen tool, prompt and evaluator implementations. It supports paper rebuilding and result auditing; repeating agent generation additionally requires the source inputs, a compatible runtime and the historical model service.

ETHICS STATEMENT

The work studies reconstruction of existing public research deposits and does not involve a new clinical intervention or diagnostic deployment. Public accessibility does not itself establish redistribution rights, donor-level consent for every reuse, or freedom from identifying metadata. Any public resource release must preserve source attribution and applicable access conditions. Incorrect pairing or synthetic replacement of measurements can contaminate downstream research; the workflow therefore records provenance, retains uncertainty, and excludes known integrity holds from the prepared collection. Visual review is not a clinical safety certification.

REFERENCES

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefanus Candra, Jason Dunkelberger, Daniel Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Ngoc-Uyen Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel Weld. AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b2ce9568dbb559aefc8c98ca5b5314ce-Paper-Conference.pdf.

Dechao Bu, Jingbo Sun, Kun Li, Zihao He, Wei Huang, Jinlin Hu, Shanshan Zhang, Shuangshuang Lei, Peipei Huo, Zhihao Wang, Sheng Wang, Tao Wang, Kai Gao, Yang Wu, Lianhe Zhao, Kai Wang, Gen Li, Huan Song, Yang Jin, Kang Zhang, Runsheng Chen, and Yi Zhao. Empowering AI data scientists using a multi-agent LLM framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01634-6. URL https://www.nature.com/articles/s41551-026-01634-6.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. In *International Conference on Learning Representations*, 2025.

URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7e3767db483c942b883eb4f8cfb74e31-Paper-Conference.pdf.

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. STimage-1K4M: A histopathology image-gene expression dataset for spatial transcriptomics. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3ef2b740cb22dcce67c20989cb3d3fce-Abstract-Datasets_and_Benchmarks_Track.html.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f12b4df26344f3be803c06b555252efe-Paper-Conference.pdf.

Dionizije Fa, Marko Čuljak, Bruno Pandža, and Mateo Čupić. BioAgent Bench: An AI agent evaluation suite for bioinformatics. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.21800.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/60a899cc31f763be0bde781a75e04458-Abstract-Datasets_and_Benchmarks_Track.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06770.

Yiming Li, Yanyi Ding, Saya Dennis, Meghan R. Hutch, Jiaqi Zhou, Yadi Zhou, Yawei Li, Maalavika Pillai, Sanaz Ghotbaldini, Mario Alberto Garcia, Mia S. Broad, Chengsheng Mao, Parambir S. Dulai, Feixiong Cheng, Zexian Zeng, and Yuan Luo. SOAR elucidates biological insights and empowers drug discovery through spatial transcriptomics. *Science Advances*, 11 (24):eadt7450, 2025. doi: 10.1126/sciadv.adt7450. URL https://doi.org/10.1126/sciadv.adt7450.

Yonghao Liu, Chuyao Wang, Zhikang Wang, Liang Chen, Zhi Li, Jiangning Song, Qi Zou, Rui Gao, Bin-Zhi Qian, Xiaoyue Feng, Renchu Guan, and Zhiyuan Yuan. High-parameter spatial multi-omics through histology-anchored integration. *Nature Methods*, 23(2):373–386, 2026. doi: 10.1038/s41592-025-02926-6. URL https://doi.org/10.1038/s41592-025-02926-6.

Luca Marconato, Giovanni Palla, Kevin A. Yamauchi, Isaac Virshup, Elyas Heidari, Tim Treis, Wouter-Michiel Vierdag, Marcella Toth, Sonja Stockhaus, Rahul B. Shrestha, Benjamin Rombaut, Lotte Pollaris, Laurens Lehner, Harald Vöhringer, Ilia Kats, Yvan Saeys, Sinem K. Saka, Wolfgang Huber, Moritz Gerstung, Josh Moore, Fabian J. Theis, and Oliver Stegle. SpatialData: an open and universal data framework for spatial omics. *Nature Methods*, 22:58–62, 2025. doi: 10.1038/s41592-024-02212-x. URL https://www.nature.com/articles/s41592-024-02212-x. First published online 20 March 2024.

Jiacheng Miao, Joe R. Davis, Yaohui Zhang, Jonathan K. Pritchard, and James Zou. Reimagining research papers as interactive and reliable AI agents. *Nature*, 2026. doi: 10.1038/s41586-026-11044-y. URL https://www.nature.com/articles/s41586-026-11044-y.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4252dc94531833029000f85dc5fac792-Paper-Conference.pdf.

Amelia Schroeder, Melanie L. Loth, Chunyu Luo, et al. Scaling up spatial transcriptomics for large-sized tissues: uncovering cellular-level tissue architecture beyond conventional platforms with iSCALE. *Nature Methods*, 22:1911–1922, 2025. doi: 10.1038/s41592-025-02770-8. URL https://doi.org/10.1038/s41592-025-02770-8.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016. doi: 10.1126/science.aaf2403. URL https://doi.org/10.1126/science.aaf2403.

Guoliang Wang, Song Wu, Zhuang Xiong, Hongzhu Qu, Xiangdong Fang, and Yiming Bao. CROST: a comprehensive repository of spatial transcriptomics. *Nucleic Acids Research*, 52(D1): D882–D890, 2024. doi: 10.1093/nar/gkad782. URL https://doi.org/10.1093/nar/gkad782.

Yihang Xiao, Jinyi Liu, Yan Zheng, Shaoqing Jiao, Jianye Hao, Xiaohan Xie, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Zhen Wang, Xuequn Shang, Zhijie Bao, Changxiao Yang, and Jiajie Peng. CellAgent: LLM-Driven Multi-Agent Framework for Natural Language-Based Single-Cell Analysis. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7c482acecdd1b3386a3a11acc536a22a-Paper-Conference.pdf.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May Dongmei Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Xin Liu, Carl Yang, Yang Xie, and Wenqi Shi. MedAgentGym: A Scalable Agentic Training Environment for Code-Centric Reasoning in Biomedical Data Science. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/911ca386fecaf7ebac35bb64733050fc-Paper-Conference.pdf.

Zhicheng Xu, Weiwen Wang, Tao Yang, Ling Li, Xizheng Ma, Jing Chen, Jieyu Wang, Yan Huang, Joshua Gould, Huifang Lu, Wensi Du, Sunil Kumar Sahu, Fan Yang, Zhiyong Li, Qingjiang Hu, Cong Hua, Shoujie Hu, Yiqun Liu, Jia Cai, Lijin You, Yong Zhang, YuXiang Li, Wenjun Zeng, Ao Chen, Bo Wang, Longqi Liu, Fengzhen Chen, Kailong Ma, Xun Xu, and Xiaofeng Wei. STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. *Nucleic Acids Research*, 52(D1):D1053–D1061, 2024. doi: 10.1093/nar/gkad933. URL https://doi.org/10.1093/nar/gkad933.

Zhiyuan Yuan, Wentao Pan, Xuan Zhao, Fangyuan Zhao, Zhimeng Xu, Xiu Li, Yi Zhao, Michael Q. Zhang, and Jianhua Yao. SODB facilitates comprehensive exploration of spatial omics data. *Nature Methods*, 20(3):387–399, 2023. doi: 10.1038/s41592-023-01773-7. URL https://doi.org/10.1038/s41592-023-01773-7.

Peng Zhang, Chaofei Gao, Zhuoyu Zhang, Zhiyuan Yuan, Qian Zhang, Ping Zhang, Shiyu Du, Weixun Zhou, Yan Li, and Shao Li. Systematic inference of super-resolution cell spatial profiles from histology images. *Nature Communications*, 16:1838, 2025. doi: 10.1038/s41467-025-57072-6. URL https://doi.org/10.1038/s41467-025-57072-6.

Jiawei Zou, Kai Xiao, Zexi Chen, Jiazheng Pei, Jing Xu, Tao Chen, Likun Hou, Chunyan Wu, Yunlang She, Zhiyuan Yuan, and Luonan Chen. The multimodal pretraining framework CarHE predicts spatial transcriptomics in tumors from routine pathology images. *Cancer Research*, 2026. doi: 10.1158/0008-5472.CAN-26-1480. URL https://doi.org/10.1158/0008-5472.CAN-26-1480.

A SCREENING REFERENCE AND SOURCE EVIDENCE

The screening reference, reconstruction experiment, and released resource serve distinct roles (Table S1), with labels and artifact versions defined for each analysis.

Table S1: Cohorts and evaluation roles.

Cohort	Size	Role
Operational screening	3,771 GSEs	Curated A781/B602/C2388 predictions used for candidate selection.
Screening reference	400 GSEs	A150/B109/C141; primary source reads, AI cross-review and human final verification.
Reconstruction experiment	40 GSEs; 331 GSMs per condition	160 runs; 1,324 required sample-condition outcomes across four conditions.
Complete output review	5,035 GSMs; 592 GSEs	4,704 PASS and 331 FAIL across all admitted candidates.
Published HESRT	4,468 GSMs; 507 GSEs	37,054,005 spatial observations; a release selected from human-reviewed outputs.

Candidate retrieval. To identify spatial transcriptomics studies at repository scale, we systematically screened a GEO metadata snapshot of 216,510 Series records from 2016–2026. Search fields included series titles, summaries, organism annotations, available sample titles, and supplementary-file descriptors. Records were retrieved when they contained explicit spatial assay or platform terms (e.g., "spatial transcriptomics", "Visium", "Xenium", "CosMx", or "MERFISH"), or combined "spatial" or "spatially" with a contextual term such as "transcript", "expression", "omics", "atlas", "profiling", or "resolved". This broad keyword retrieval yielded 3,771 unique candidate GSE accessions before histology-pairing review. Their recorded publication dates range from 5 January 2016 to 31 March 2026. All 3,771 candidates were subsequently classified into operational Classes A, B, and C using Muse Spark 1.2 and the evidence criteria described below. The keyword configuration, candidate accessions with recorded hits, historical filter implementation, and retrieval counts are retained in `provenance/retrieval/`.

Reference and criteria. The 400 reference studies were selected across publication years and experimental assay classes to form the benchmark. Each study underwent manual review of primary publications and raw repository files, supplemented by automated cross-validation and senior human adjudication. All 400 curated studies serve as the evaluation denominator.

Class A requires an eligible spatial assay with explicit H&E evidence or characteristic Visium/Space Ranger morphology files. Class B retains indirect or unresolved morphology, including DAPI or immunofluorescence without explicit H&E. Class C covers excluded or insufficiently supported combinations. In mixed-assay studies, classification uses evidence linked to the relevant spatial sample, with dissociated assays evaluated separately. Figure S1 illustrates direct support, unresolved archive pairing, and competing assay evidence. The operational class summary and the 400-GSE reference evaluation use distinct cohorts.

Model and context. Operational screening used Muse Spark 1.2 through its official API interface with pre-defined prompt instructions and a fixed context-recovery procedure. The recorded model identifier and execution settings are retained in the accompanying provenance files. The context-recovery cohort and character-budget records are retained in `data/screening_evaluation.json`; the screening instructions are in `provenance/screening_judging.py.txt`.

Complete agreement and uncertainty. Figure 2a and Table 1 give the matrix and per-class metrics. Across the 400 reference studies, the screening module achieved 91.75% agreement (367/400) and a macro-F1 score of 0.913. Among the 33 classification discrepancies, 24 represented distinctions between Class A and Class B, while nine crossed the exclusion boundary (five reference-B/predicted-C and four reference-C/predicted-B cases). A+B precision is 98.45% and recall is 98.07%. Non-parametric bootstrap resampling (2,000 iterations) yields empirical 95% confidence intervals of 89.0–94.5% for agreement, 0.884–0.941 for macro-F1 and 96.2–99.6% for A+B recall.

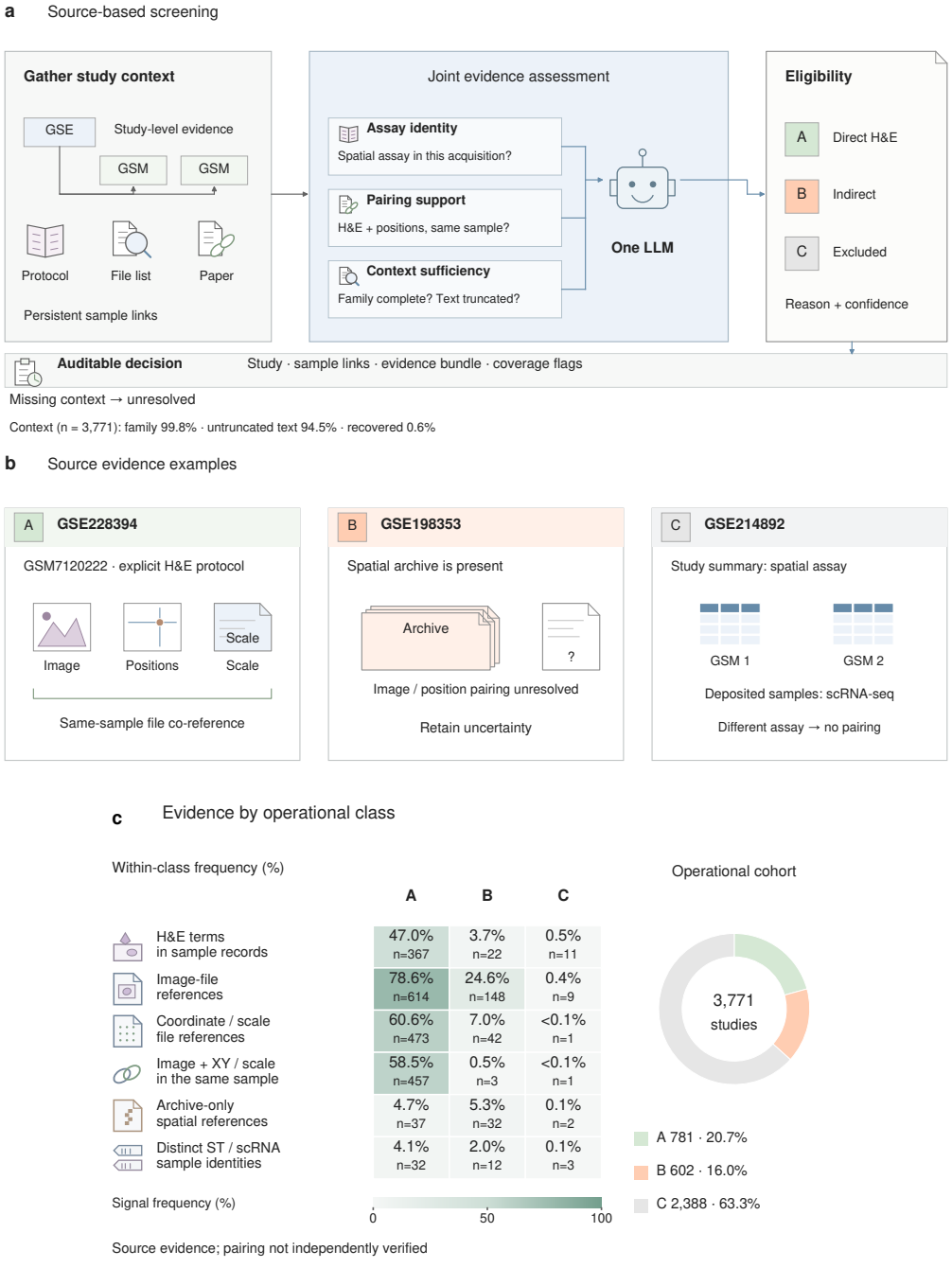


Figure S1: **Detailed screening workflow, cases and evidence coverage.** (a) Evidence integration; (b) direct support, unresolved pairing and competing-assay cases; (c) evidence signals across the 3,771 operational classes. Main Figure 2a reports the separate 400-GSE reference evaluation.

B RECONSTRUCTION REQUIREMENTS, TOOLS AND CHECKS

B.1 RECONSTRUCTION OUTPUTS AND SOURCE-BASED MAPPING

For each study, the agent reconstructs the acquisitions specified by the deposited source files and produces linked expression, histology, coordinates, and provenance outputs. Shared observation and feature identifiers connect the deposited measurements to the reconstructed outputs. Coordinates are expressed in the pixel frame of the selected image, with x denoting columns and y denoting rows. Figure S3 illustrates these links for a shared acquisition.

The agent inspects the source files, develops and executes a study-specific reconstruction program, and uses the resulting outputs and logs to guide revisions within the common budget. The final artifacts are then frozen for independent evaluation of data validity.

Coordinate mapping follows the geometric information deposited with each acquisition, including scale factors, homographies, field-of-view offsets, coordinate units, and embedded image frames. When coordinates are recovered from rendered source marks, a complete correspondence between mark identities and molecular observations establishes the mapping. Selection-based recovery uses the deposited array geometry and a uniquely determined orientation. In both cases, source records establish the transformation, while image diagnostics assess the resulting alignment. Acceptance requires sufficient source evidence for the mapping.

Observation selection follows the in-tissue subset defined by the source files. Image cropping changes the image extent while retaining the observations required by that source definition.

B.2 DOMAIN TOOLS AND EXPERIMENTAL AVAILABILITY

All four conditions share shell/Python execution, image viewing, file writing, execution logs, schema-error feedback, and agent-written self-checks. Workflow and W+QC additionally receive workflow guides, 26 domain helpers, and source-image evidence interfaces. QC and W+QC receive the same black-box diagnostic checker (Table S2). Reference labels are used only in the independent evaluation stage. The Workflow comparison evaluates the guides, helpers, and evidence interfaces as a combined intervention.

The catalog contains seven Visium readers/mappers, six platform adapters, four provenance helpers, one native-frame checker and eight container-decoding functions (Figure S2). Signatures, docstrings, frozen code and task instructions are included in the evidence archive.

Table S2: **Controlled tool visibility.** Checks denote availability, not mandatory use.

Capability	Base	QC	Workflow	W+QC
Shell/Python, actual pixel viewer	✓	✓	✓	✓
Write/patch, manifest, execute, logs	✓	✓	✓	✓
Agent-written self-checks and revision	✓	✓	✓	✓
Domain workflow documents and helpers	–	–	✓	✓
Dedicated source-image evidence tools	–	–	✓	✓
Production checker diagnostics and images	–	✓	–	✓
Independent source evaluator during generation	–	–	–	–

B.3 IMAGE DIAGNOSTICS AND OPERATING THRESHOLDS

The checker jointly rescales the image and coordinates for display-level geometric diagnostics, while the delivered coordinates retain their source resolution. Table S3 summarizes three measurements: visible coordinate coverage, blank-background spots, and tissue outside the measurement region.

Table S3: Diagnostic checks, operating thresholds, and interpretation. Checks with missing prerequisites are reported as unevaluable, with the reason recorded alongside the diagnostic result.

Check	Measurement	Operating rule
Visible coverage	Fraction of all observations with finite centers inside the image.	Below 50% is an alignment failure.
Blank-background spots	Top 10% of finite, visible observations by total expression; test footprint and surrounding annulus for bright, uniform background.	Above 3% blank spots is an alignment failure.
Tissue outside coverage	Fraction of target tissue outside the buffered measurement region. Target tissue includes only connected components touching that region.	Above 80% gives an advisory warning, suppressed if mask polarity was inverted.

Visible coverage. Visible coverage is the fraction of all observations whose coordinates are finite and whose centers fall within the image; values below 50% indicate an alignment failure.

Blank-background spots. The blank-background check selects the top 10% of finite, visible observations by total expression. It tests each circular footprint and its surrounding annulus for bright, uniform background using the frozen contrast representation. The brightness threshold defaults to 0.98 and is adjusted downward for bright image corners; the standard-deviation threshold is 0.01. Footprints have a minimum radius of two display pixels. A blank-background fraction above 3% indicates an alignment failure.

Tissue outside coverage. The tissue-coverage check constructs the convex hull of observations above the fifth expression percentile and buffers it by half the median observation spacing, with a minimum buffer of two display pixels. Tissue segmentation uses RGB distance from the dominant background color, Otsu thresholding, and fixed morphological operations. Target tissue comprises connected components that intersect the buffered measurement region. The fraction of target-tissue pixels outside this region produces an advisory warning when it exceeds 80%.

When low tissue occupancy triggers mask inversion, the checker records the inversion and suppresses the tissue-coverage advisory. The evidence archive contains the frozen implementation, including fallback rules and exception handling. Geometric diagnostics provide alignment feedback during reconstruction; independent source evaluation subsequently assesses expression identity and acquisition completeness.

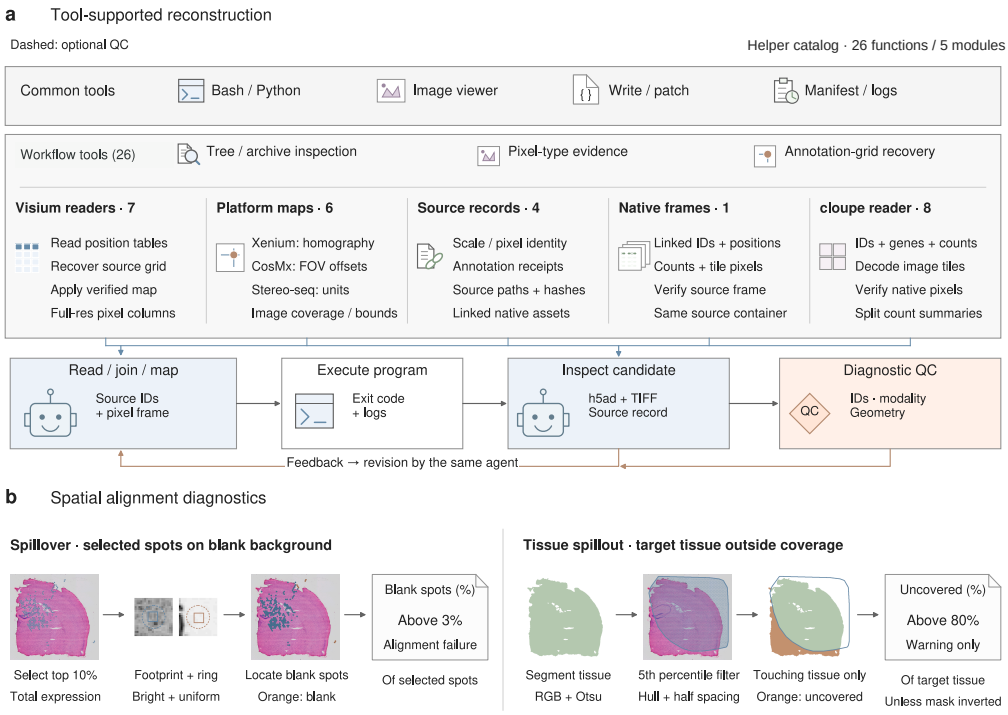


Figure S2: **Full tool catalog and spatial diagnostics.** (a) Common tools, 26 domain helpers in five modules, and the reconstruction loop. (b) Spillover and tissue-spillout operations. Thresholds and exceptions are specified in Appendix B.3; tissue spillout is advisory.

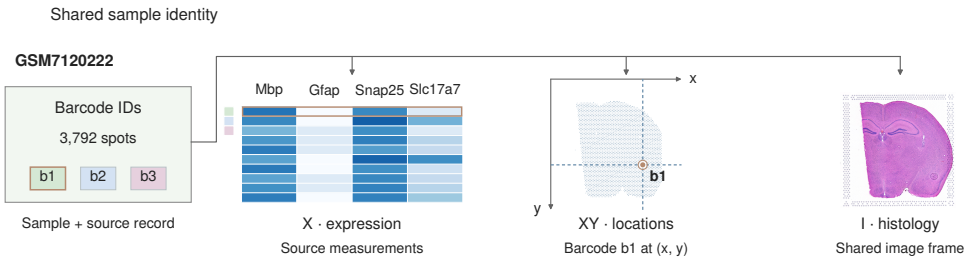


Figure S3: **A shared acquisition links molecular observations to histology.** The GSM7120222 example illustrates how shared observation IDs connect deposited expression values, spatial coordinates and histology. The displayed gene columns illustrate these links.

C EXPERIMENTAL AND REVIEW PROTOCOL

Cohort and controlled execution. The 40-GSE evaluation cohort was randomly selected before generation from receipt-complete operational-A studies, with ten studies per source-format stratum and a fixed random seed. Studies used in earlier exploratory workflow and tool-development pilots were excluded to separate development from benchmark evaluation. The exact selection records accompany the archive. Source inventories define the same 331 required GSMs and 24 source-excluded entries in each condition. Every GSE–condition cell had one fresh generation, no fresh-task retry and at most two transport retries. Model, source inputs and limits were shared: 3 hours, 5 million provider tokens and at most 15 program executions. Sources were materialized and hash-verified before the batch, then exposed immutably in separate workspaces. The experiment therefore includes common archive/container preparation in its input condition.

All 160 runs shared 12 CPU execution slots. Recorded generation time includes transport and resource waits but excludes common preparation. All 160 visibility probes passed, source hashes were unchanged, and logs showed no Workflow-helper calls in Base/QC or dedicated-checker calls in Base/Workflow. Remaining independent evaluation jobs used the same candidates and scoring rules.

Double-blind visual review. The formal visual review evaluated 1,324 required sample-condition entries: 331 required GSMs across four conditions, after 24 source-level exclusions per condition. Each visual verdict was scored against the frozen candidate output under the double-blind review protocol. Artifacts with congruent coordinate–histology overlays were graded as PASS, visible coordinate misalignments or axis flips as FAIL, and execution halts or missing visualizations as UNAVAILABLE. Human visual verification provides the primary reference for morphological and spatial alignment, while automated scripts independently evaluate container schemas, source-supported geometry, and preservation of deposited expression values.

Independent source audit. Sample-local PASS requires structure, source-supported geometry and expression to pass together. Geometry compares identities, source pixels and deposited mappings at a one-pixel tolerance; expression checks deposited values and observation/feature identities. Verified source-supported scale changes are allowed. Fabricated, normalized or aggregated substitutes do not satisfy exact source fidelity. GSE-wide acquisition completeness is recorded separately. FAIL, INCOMPLETE and UNKNOWN remain distinct, with only PASS counted as success among all 331 GSMs. Before generation, 24/40 studies had fully bound numerical references. Source-only derivations for the remaining studies followed the frozen rules, and unresolved references remained UNKNOWN. Under a sensitivity scenario that counted all UNKNOWN outcomes as PASS, the source-PASS bounds were 65.9/70.7/79.5/76.7% for Base/QC/Workflow/W+QC. These bounds are reported separately from confidence intervals; no outcomes were imputed.

Paired effects and uncertainty. Pass rates weight each required GSM equally. Paired differences compare the same GSMs across conditions; unavailable outputs contribute no positive verdict. Paired differences were estimated using 20,000 cluster-bootstrap resamples of the 40 GSE studies, preserving within-study sample correlation structures. Each resample retains all GSMs and all four conditions within each selected study and uses its GSM count as the denominator. Reported marginal 95% confidence intervals reflect the 2.5th and 97.5th percentiles of the paired difference distribution. Source-audit effects use the same resampling rule and are stored separately from human-review effects. The archive records the random seed and exact implementation.

D COMPLETE RESULTS AND RUN COSTS

Table S4 retains every required visual-review outcome. The four paired comparisons and intervals appear in main Table 2 and are not repeated here.

Independent automated verification. Independent automated checks assessed source structure, coordinate geometry and expression preservation across the four conditions. In Base/QC/Workflow/Workflow + QC order, composite PASS counts were 77/77/134/133 of 331 required GSMs. Among the final Workflow + QC artifacts, all 133 samples passing the independent

Table S4: **Double-blind human visual alignment outcomes across all 331 required GSMs per condition.** Unavailable outputs remain in the denominator. Source-level expression preservation is independently evaluated by automated audit.

Condition	PASS	FAIL	Unavailable	PASS / 331 (%)
Base	181	103	47	54.7
QC	190	118	23	57.4
Workflow	257	13	61	77.6
W+QC	268	13	50	81.0

source audit preserved deposited expression values and observation identities without imputation or synthetic alteration. Full outcome counts and unresolved states accompany the source tables.

For Base/QC/Workflow/W+QC, structure passes for 279/308/266/255 GSMs, source geometry for 77/93/134/134 and expression for 124/133/144/145. Components are paired and correlated; their rates cannot be multiplied to reconstruct the composite outcome. Complete component states and source-audit contrasts, including all cluster intervals, are in `data/gsm_source_summary.json` and `data/per_gsm_source_outcomes.csv`.

Table S5: **Operational outcomes and costs for all attempted runs.** Each row contains 40 attempts. Tokens are provider input plus output, including cached/repeated input; M denotes millions. Generation minutes exclude common preparation and are not pure model-compute time.

Condition	Normal exit	Budget end	Runtime error	Mean M	Median M	Median min
Base	36	0	4	1.21	0.94	20.2
QC	32	4	4	2.09	1.37	30.4
Workflow	38	1	1	2.34	2.11	35.2
W+QC	29	10	1	2.98	2.64	39.0

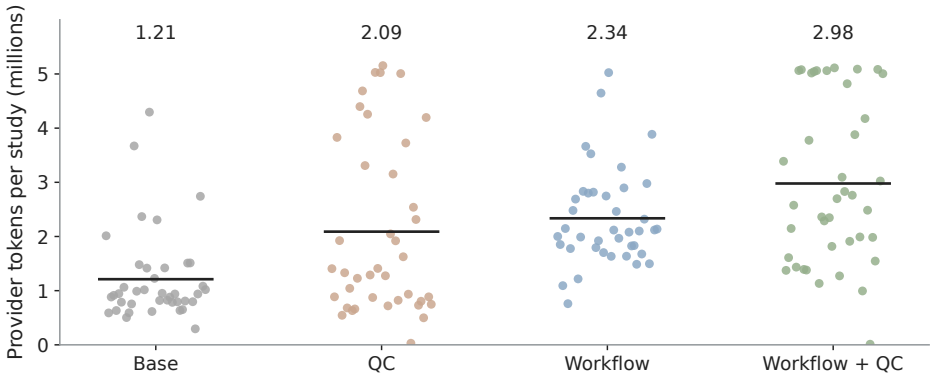


Figure S4: **Complete generation costs for the four experimental conditions.** Each dot is one GSE run: 40 per condition, all 160 attempts retained. Horizontal marks and numeric labels show mean provider-reported input-plus-output tokens, including the recorded cached-token accounting. Dots are offset deterministically for visibility. Complete integer costs, durations and terminal outcomes remain in the accompanying source tables.

All 160 attempts contribute to costs. Total provider tokens are 48,460,566; 83,575,614; 93,418,662; and 119,122,000 for Base/QC/Workflow/W+QC. The token cap is checked after a provider response, so one response may cross it. Normal exit is not scientific success. Tokens include cached/repeated input; durations are observed generation times, not pure model-compute time.

Feedback and iterative refinement. In QC-assisted conditions, diagnostic checks were invoked in 37/40 QC runs (58 calls) and 39/40 W+QC runs (59 calls; Table S6). These diagnostics provide intermediate guidance during code generation, helping the agent detect image-frame misplacements before output freezing. The table summarizes final production-checker outcomes and changes in independent source-verification outcomes across stored executions.

Table S6: Recorded feedback outcomes at the GSE level. Upper: independent source-audit outcomes among final production-checker passes. P/F/I/U denote PASS, FAIL, INCOMPLETE and UNKNOWN. Lower: changes in complete source-verification outcomes between the first and final stored executions; the Missing column identifies runs without a first snapshot. The blocks describe diagnostic agreement and execution trajectories.

Condition	Final production pass	Independent P/F/I/U
QC	22	10 / 4 / 2 / 6
W+QC	26	16 / 1 / 0 / 9

Condition	Pairs	Missing	First P	Final P	Improved	Regressed
Base	37	3	10	10	0	0
QC	38	2	8	11	3	0
Workflow	39	1	18	18	0	0
W+QC	39	1	18	16	0	2

Per-run records in the archive document the source-verification transitions, image-source changes, and execution outcomes associated with Table S6.

For the coordinate case in Figure 2e,f, both displayed conditions preserve the 2,380 identities and expression matrix (32,285 genes). The source image is 2000×1939 pixels. Across these two conditions, maximum coordinate discrepancies are 1,390.4 pixels for Base and 0.0 pixels for Workflow, illustrating the role of source-supported orientation parsing.

E HESRT COMPOSITION AND RESOURCE COMPARISONS

Cohort curation and release selection. A total of 5,035 GSMs across 592 GSEs were evaluated through the human-in-the-loop review pipeline, yielding 4,704 visually reviewed PASS samples and 331 alignment FAIL samples (Figure S5). From this pool, the frozen HESRT release contains 4,468 human-reviewed samples across 507 studies. Samples outside the public release comprise the 331 failed alignments, seven integrity exclusions (five synthetic-expression holds and two output-requirement failures), and 229 later unpublished outputs. Resource-composition and geometric analyses use the fixed released artifacts. Appendix F documents the relationship between reviewed versions, released versions, and source-verification records.

Table S7: Published HESRT composition. Observations are expression-matrix rows: spots, bins or cells, depending on platform.

Platform	Human	Mouse	Rat	Other/NA	Samples	Observations
Visium	2,462	1,420	101	97	4,080	12,272,028
ST	192	77	0	0	269	126,366
Visium HD	43	14	0	0	57	21,770,955
Xenium	6	2	0	0	8	1,668,764
Other ST	15	20	1	0	36	1,144,989
Unspecified	18	0	0	0	18	70,903
Total	2,736	1,533	102	97	4,468	37,054,005

Observations mix spots, bins and cells; they are not independent patients or uniformly defined single cells. Other ST includes Decoder-seq, DynaSpatial, BMKMANU, Array-seq and HDST; Unspecified includes missing/generic platform labels. The release has 4,403 operational-A, 61 B and four unclassified samples, so not every artifact is confirmed H&E. Versioned checksums, source attribution and access conditions accompany the catalog.

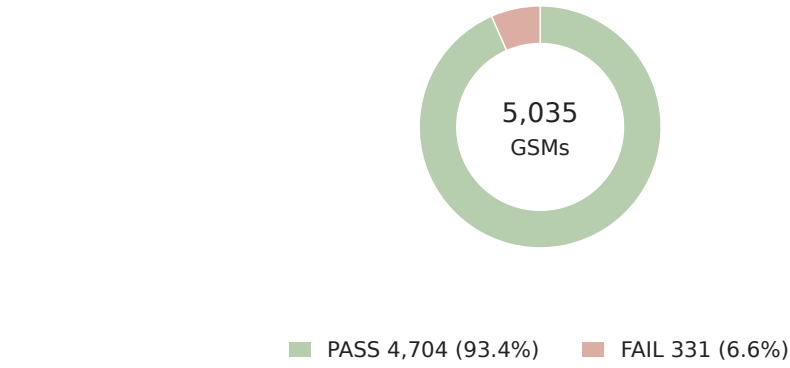


Figure S5: **Human-in-the-loop review outcomes.** The review cohort comprises 5,035 unique GSMs from 592 GSEs, using the latest explicitly reviewed artifact per GSM: 4,704 PASS and 331 FAIL. The public release selects 4,468 human-reviewed samples from this cohort.

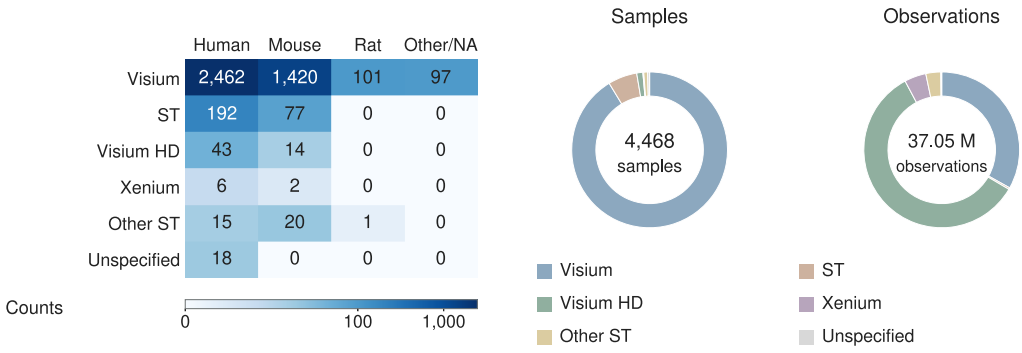


Figure S6: **Complete published resource composition.** Platform-by-species counts and both platform-share rings describe 4,468 GSMs from 507 GSEs and 37,054,005 spatial observations. The heatmap uses $\log_{10}(1 + \text{count})$ color encoding and exact count labels. The two rings use different denominators: samples and observations (spots, bins, or cells). Other/NA includes every remaining species label. Release provenance is recorded in `data/resource_release.json`; the review-to-release relationship is described in Appendix E.

Shared geometric measurements. Main Figure 3a includes all platforms; panel b uses measured human/mouse Visium artifacts. The shared comparison reports four geometric checks: unique observation IDs, finite coordinates, unique centers, and centers inside the associated image (Table S8). It summarizes coordinate geometry; expression fidelity, anatomical registration, and downstream patch quality are separate dimensions. The table lists unmeasured records separately from measured outcomes. Human/mouse/rat/other counts are 672/419/17/41 for STimage-1K4M, 649/580/0/0 for HEST-1k and 2,736/1,533/102/97 for HESRT. Other/NA retains all remaining labels, including mixed-species and unspecified records.

Table S8: **Shared geometric checks and measurement availability.** The table reports record counts, with unmeasured records listed separately. A record passes when observation IDs and centers are unique, coordinates are finite, and all centers fall within the associated image.

Resource	Eligible	Measured	Unmeasured	Passing	Passing spots
STimage-1K4M	936	786	150	784	1,636,687
HEST-1k	602	602	0	583	1,462,322
HESRT	3,882	3,882	0	3,802	10,592,238

Portal coverage and matching. The comparison uses a GEO metadata cache dated 1 April 2026, operational predictions from 9 September, and portal snapshots from 16 September. It summarizes accession overlap across these dated records. Primary matching uses explicit GSE accession/source fields, excluding CROST's paired-project field; sensitivity analyses add shared GSMs and explicit PMID/DOI associations with cached publication-alias resolution. Shared publications identify source associations. Unresolved or non-GEO identifiers are excluded from uncovered-GEO counts.

Table S9: Four-portal coverage. GSE denominators count exposed unique accessions; publication denominators count resolved source-publication components. Neither is a count of successfully reconstructed samples.

Portal	GSE overlap	Publication overlap
SODB	27 / 29	33 / 93
STOmicsDB	251 / 306	168 / 206
SOAR	18 / 19	248 / 331
CROST	361 / 380	291 / 324
Deduplicated union	516 / 591	458 / 610

The 591-GSE union contains 338 A, 117 B, 61 C and 75 accessions absent from the frozen screen; the 516 overlaps therefore include Class C. SOAR lists 19 direct GSE accessions, of which 18 match the screen. Unlinked A candidates decrease from 443 to 410 to 360 as GSM and publication links are added; A+B candidates decrease from 928 to 879 to 797. Of the 797 A+B candidates, 487 have a local publication identifier and 310 do not. They remain candidates for reconstruction; incomplete metadata may account for some unmatched records.

The explicit GSM union contains 8,956 identifiers: 8,228 in the candidate metadata pool and 2,368 in the prepared resource. Study membership across external portals may occasionally include non-spatial assay arms or minor accession discrepancies due to repository release timelines. Full intersections, identifier availability, pagination audits and snapshots are in `resource_data/`; all portals and lower-overlap results are retained there.

F REPRODUCIBILITY AND EVIDENCE INDEX

The accompanying archive contains the paper source, figure PDFs, complete outcomes, and frozen implementation records (Table S10). Per-study tables, checker operations, and version histories provide the detailed evidence underlying the reported results. Upload copies retain public source identifiers and scientific instructions, with operational paths, hostnames, and reviewer identifiers anonymized as documented in the inventory.

Table S10: Locations in the accompanying anonymous archive.

Evidence	Files or directory
400-GSE screening	`data/screening_evaluation.json`; `provenance/screening/`
Human and source outcomes	`data/per_gsm_human_review.csv`; `data/per_gsm_source_outcomes.csv`
All 160 runs and costs	`data/per_run_results.csv`
Review and release records	`provenance/full_review/`; `provenance/published_review/`; `data/hf_catalog.csv`
Tools, prompts and diagnostics	`provenance/implementation/`; `provenance/frozen_workflow_documents/`
Resource-comparison evidence	`resource_data/`
Paper build and integrity	`main.tex`; `build.sh`; `SHA256SUMS`

Pipeline reproduction uses the standardized source inputs, model API endpoints, runtime environments and container isolation controls. The model backbone is Muse Spark 1.2 operated under high-reasoning mode. Execution prompts, tool catalogs and evaluation harnesses are packaged in the accompanying repository for transparent and verifiable evaluation. Analysis used Python 3.11.16, NumPy 2.2.6, SciPy 1.17.1 and Matplotlib 3.10.9.

Artifact versioning and freeze. All HESRT resource analyses and downstream evaluations in this work use the frozen HESRT release (v1.3; 4,468 multimodal samples). To ensure backward reproducibility, repository catalogs maintain cryptographic checksums for all published artifacts. Any subsequent iterative refinements or expanded provenance comparisons are tracked in the versioned repository manifests without altering the frozen benchmark artifacts reported in this paper.

G FROM PUBLIC DEPOSITS TO REUSABLE SPATIAL DATA

Figure S7 brings together source selection, identity-preserving reconstruction, and verification in one study, alongside the screening, reconstruction, and resource results reported in the main text.

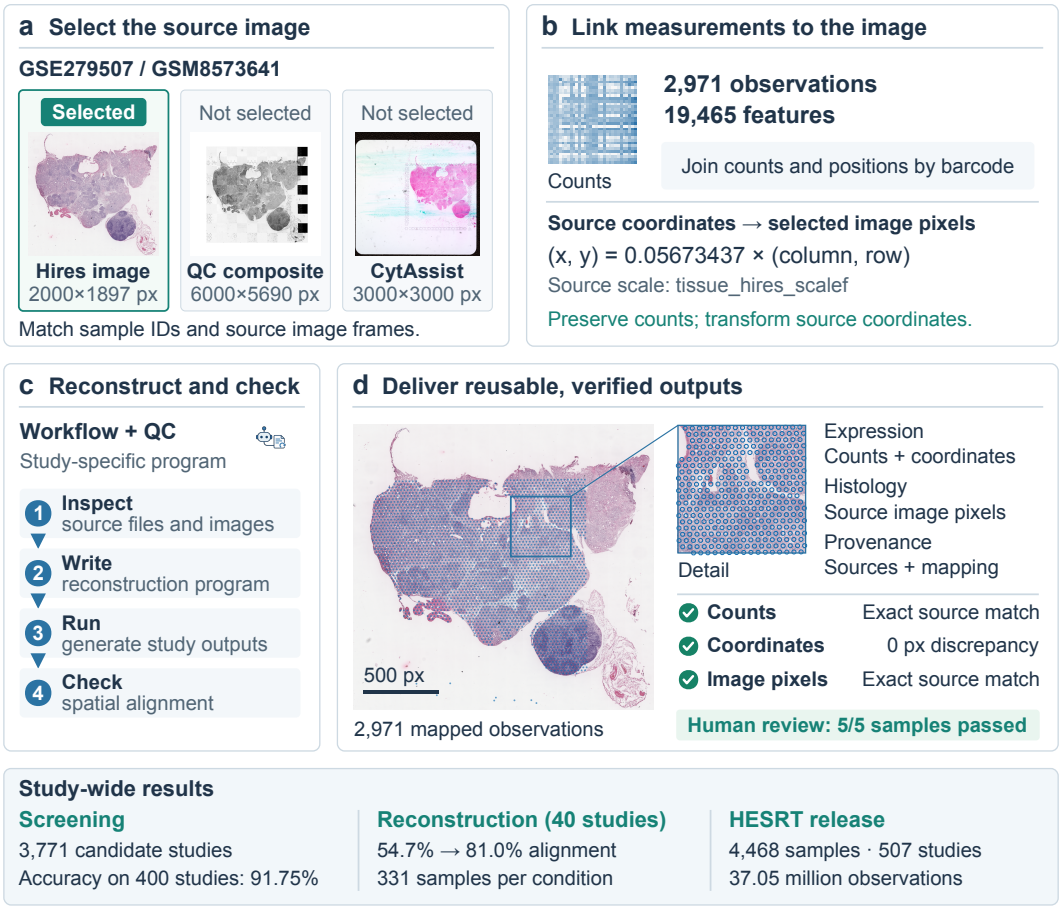


Figure S7: **From heterogeneous source files to reusable spatial measurements.** (a) Source image alternatives for GSE279507/GSM8573641 and the selected sample-linked hires histology image. (b) Counts for 2,971 observations and 19,465 features are joined to source positions by barcode. The dimensionless source scale maps these positions to pixels in the selected image. The count excerpt displays $\log(1 + \text{count})$ for 32 observations and 32 features. (c) Recorded stages of the Workflow+QC execution. (d) All delivered coordinates are overlaid on the image, with the boxed region enlarged. Source comparisons confirm identical counts and image pixels, and zero coordinate discrepancy for this sample. All five samples in the study passed human alignment review; the scale bar uses native image pixels. The bottom strip summarizes distinct screening, reconstruction, and resource cohorts. The alignment change compares Base with Workflow+QC. Figure assets and case-verification records accompany the source archive in `provenance/closing_summary/`.