CHORUS: A MULTI-AGENT FRAMEWORK WITH STATE-CONDITIONED EXPERIENCE ADAPTATION FOR SPATIAL BIOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial biology requires interpreting molecular variation within its tissue context. Although large language model agents have simplified spatial analysis, existing systems typically center on either spatial transcriptomics or histopathology. Even when both are accepted, histopathology images are often used only inside task-specific methods rather than as an independent source of evidence throughout reasoning. We introduce CHORUS, a multi-agent system that coordinates molecular and tissue-structural information for spatial-biology analysis. CHORUS assigns spatial transcriptomics and H&E-stained histopathology images to role-isolated specialists and converts their executed results into provenance-grounded evidence logic trees. A scientific-state gate permits cross-modal interaction only after verifying specimen, coordinate, registration, and region correspondence. Grounded evidence remains immutable, while specialists may revise only their inferences and conclusions or request additional evidence. Unresolved disagreement therefore triggers targeted evidence acquisition rather than forced consensus. Additionally, a state-conditioned experience memory retrieves prior trajectories only under compatible scientific conditions, enabling behavioral evolution without model updates. Controlled experiments demonstrate improved cross-modal reasoning, while state-conditioned memory enhances accuracy and error recovery.

1 INTRODUCTION

Spatial biology links molecular activity to its tissue context. Spatial transcriptomics (ST) measures gene expression together with spatial coordinates, revealing molecular domains, cell states, and local interactions in intact tissue. Pathological image provide complementary information about tissue architecture, cellular morphology, and structural boundaries. Together, these sources can connect molecular variation with tissue organization, but their analysis requires heterogeneous preprocessing, specialized tools, and careful interpretation (Wang et al., 2026). Large language model (LLM) agents offer a natural interface to this workflow by translating scientific questions into executable analyses and grounded reports (Li et al., 2025; Pavilion et al., 2026).

Current agents, however, usually organize analysis around one primary information source. STAgent and SpatialAgent mainly automate ST-centered workflows, whereas NOVA focuses on histopathology analysis (Lin et al., 2025; Vaidya et al., 2025; Wang et al., 2025). Some spatial agents can also invoke image-aware downstream methods. For example, ChatSpatial can call SpaGCN for spatial domain identification (Yang et al., 2026); SpaGCN combines expression, coordinates, and histology pixels within a predefined graph model (Hu et al., 2021). This is a useful multimodal computation, but the pathological image is consumed inside a task-specific method. It does not independently generate tissue-level evidence or continue to interact with molecular findings throughout the analysis. This distinction is important for domain identification, where molecular patterns and tissue structure may agree, reveal complementary organization, or disagree. Using two modalities in a tool is not the same as coordinating two sources of scientific information.

Multi-agent collaboration provides a promising mechanism for such coordination. PathGen separates pathology image–text generation, revision, and summarization among specialized agents (Sun et al., 2025), while MedLA uses logic trees to support premise-level interaction in medical reason-

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.

ing (Ma et al., 2026). Yet the central issue is not simply whether agents communicate, but what they exchange and what they may revise. If tool-derived observations and interpretations are mixed in free-form discussion, grounded findings can be blurred by later messages. Repeated discussion can also amplify early errors or favor agreement over correctness (Wynn et al., 2025). Scientific collaboration therefore needs explicit constraints: executed evidence should remain traceable and immutable, revisions should target only inference or conclusion, and unresolved claims should trigger new evidence rather than forced consensus.

The Cross-modal interaction must also respect the scientific state of the data. An ST assay and a pathological image may come from different specimens, use incompatible coordinates, lack verified registration, or cover different regions of interest. Under these conditions, the two sources are incomparable, not contradictory. The same principle applies to memory. Prior work shows that agents can improve by retaining successful trajectories without updating model parameters (Shinn et al., 2023; Zhao et al., 2024; Jin et al., 2025), but a workflow learned from registered, matched data may be unsafe for an unregistered or mismatched sample. Reusing experience therefore requires scientific-state compatibility.

To address these limitations, we introduce **CHORUS**, a multi-agent framework for coordinating ST and H&E analysis through explicit evidence records and state-dependent interaction. A Coordinator activates the specialists required by the question. For joint tasks, the ST specialist analyzes expression and spatial coordinates, while the H&E specialist analyzes the tissue image; each constructs its initial evidence independently of the other specialist's outputs. Modality-specific questions can use a single specialist. This organization preserves separate molecular and tissue-structural analysis trajectories before their findings are brought into a shared interpretation.

The interaction interface is a scientific evidence logic tree that separates tool-derived evidence, inference, and conclusion, with provenance and spatial support attached to its nodes. A scientific-state gate consults the recorded specimen, coordinate, registration, and region correspondence before classifying a cross-modal relation. Failed or unknown prerequisites yield an incomparable outcome. Admissible comparisons classified as complementary, conflicting, or uncertain trigger targeted discussion of the claim and its supporting evidence from a shared pre-discussion snapshot. Specialists may revise their own inferences or conclusions, retain them, or request further evidence; existing evidence records remain unchanged. A Judgment Agent directs unresolved questions to additional tool execution or retains an unresolved outcome. The resulting preservation property concerns the recorded evidence under these updates; the validity of the underlying measurements still depends on acquisition and validation. CHORUS also stores tool trajectories with their scientific states to support compatibility-conditioned workflow reuse.

Our contributions are summarized as follows:

- We introduce a multi-agent framework that jointly considers molecular and tissue-structural information while preserving independent specialist evidence and supporting molecular-only, histopathology-only, and joint analysis.
- We use provenance-aware logic trees, state-gated alignment, immutable evidence, and restricted revision. Remaining disagreements trigger evidence acquisition rather than unconstrained debate or forced consensus.
- We combine scientific-state validation with outcome-grounded trajectory memory, allowing past experience to guide future analysis only when the underlying scientific conditions are compatible and experimental results demonstrate the superiority of our agent.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

We consider a spatial-biology dataset

$$\mathcal{D} = (\mathbf{X}, \mathbf{S}, \mathbf{H}, \boldsymbol{\xi}), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{N \times G}$ is an expression matrix, $\mathbf{S} \in \mathbb{R}^{N \times 2}$ contains spatial coordinates, $\mathbf{H}$ is a pathological image, and $\boldsymbol{\xi}$ contains platform, sample, scale, and registration metadata. Given a natural-

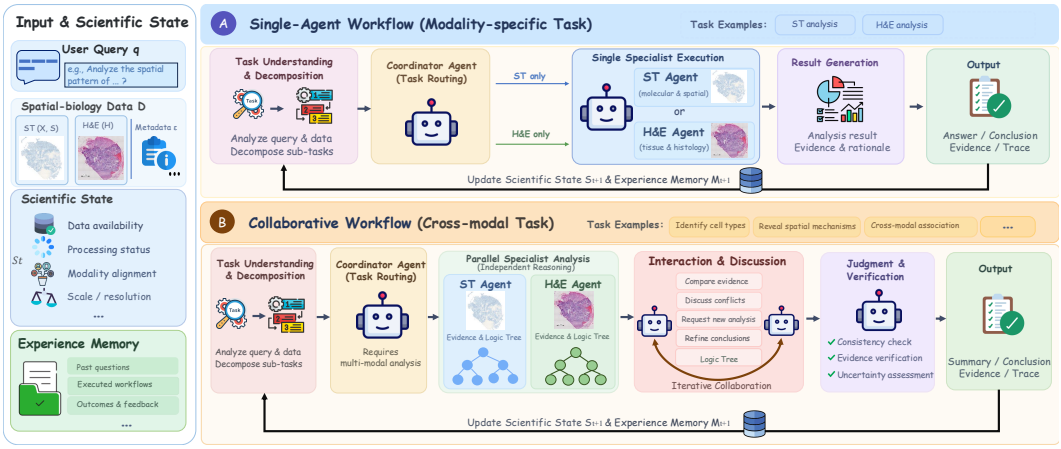


Figure 1: **Overview of CHORUS.** Given a user query, scientific state, and experience memory, the Coordinator routes modality-specific tasks to a single ST or H&E specialist, while joint tasks invoke independent specialists to construct tool-grounded evidence logic trees. State-gated alignment, constrained interaction, and judgment-based verification produce an auditable conclusion and update the scientific state and experience memory.

language scientific question q, the goal is not merely to generate an answer, but to produce

$$\mathcal{Y} = (\hat{y}, J, \mathcal{E}^\star, \Gamma), \tag{2}$$

where $\hat{y}$ is the scientific conclusion, J is its support status, $\mathcal{E}^\star$ is the decisive evidence, and Γ is an auditable interaction trace. The system may analyze molecular data, pathological image, or both, without requiring the user to select a modality. When the available evidence is scientifically incomparable or insufficient, a valid output may abstain from a joint conclusion.

2.2 Scientific Evidence Units and Logic Trees

The minimal reasoning object in *CHORUS* is a scientific evidence unit

$$v = \big(t_v, \ell_v, \Omega_v, \mathrm{par}(v), \mathrm{prov}(v), u_v\big), \tag{3}$$

where $t_v \in \{\textit{Evidence}, \textit{Inference}, \textit{Conclusion}\}$ is the node type, ℓ_v is its statement, Ω_v denotes its spatial or sample-level support, $\mathrm{par}(v)$ contains the supporting nodes, $\mathrm{prov}(v)$ records tool and artifact provenance, and u_v represents uncertainty. Chaining these units yields a scientific evidence logic tree for specialist a,

$$\mathcal{T}^a = (\mathcal{V}^a, \mathcal{E}^a, r^a), \qquad a \in \mathcal{A}_{\mathrm{spec}}, \tag{4}$$

where r^a is the root conclusion and $(v_i, v_j) \in \mathcal{E}^a$ means that v_i supports v_j. This representation exposes which observations support each inference and allows agents to compare individual claims rather than exchange unrestricted free-form reports.

An evidence node is grounded only when it can be traced to a successfully executed tool or a persisted scientific artifact:

$$g(v) = \mathbb{I}[\mathrm{prov}(v) \neq \varnothing], \qquad t_v = \textsc{Evidence} \implies g(v) = 1. \tag{5}$$

LLM explanations, generated code, and agent judgements may guide execution or interpretation, but do not constitute scientific evidence by themselves.

3 Method

3.1 Overview

As shown in Fig 3, we introduce *CHORUS*, a multi-agent framework with state-conditioned experience evolution for spatial biology. *CHORUS* contains a Coordinator Agent, a registry of task- and

modality-specialized agents, and a Judgment Agent. In the current instantiation, the principal specialists are an ST Agent for molecular and spatial evidence and an H&E Agent for tissue-structural evidence.

Given a question, *CHORUS* first profiles the current scientific state and retrieves relevant experience. The Coordinator then activates the smallest set of specialists required by the question. Activated specialists produce independent, tool-grounded logic trees before any cross-agent communication. Their claims are compared only when the scientific state establishes that the underlying samples, coordinate systems, and regions are compatible. Remaining uncertainty or disagreement triggers targeted discussion and, when necessary, new tool execution. The complete process is

$$\begin{aligned}(q, \mathcal{D}, \mathcal{M}_t) \xrightarrow{\text{profile and retrieve}} (\mathbf{s}_t, \mathcal{K}_t) \xrightarrow{\text{orchestrate}} \{\mathcal{T}^{a,0} : a \in \mathcal{U}_t\} \\ \xrightarrow{\text{align, discuss, verify}} \mathcal{Y}_t \xrightarrow{\text{experience update}} \mathcal{M}_{t+1},\end{aligned} \tag{6}$$

where $\mathcal{M}_t$ is the experience memory, $\mathcal{K}_t$ is the retrieved experience set, and $\mathcal{U}_t$ is the activated specialist set.

3.2 State-Conditioned Experience Evolution

Scientific state. Spatial-biology workflows depend on more than data presence. A valid action may require a particular platform, preprocessing result, physical scale, or verified relationship between modalities. *CHORUS* therefore represents the state at interaction t as

$$\mathbf{s}_t = \left(\mathbf{s}_t^{\text{data}}, \mathbf{s}_t^{\text{analysis}}, \mathbf{s}_t^{\text{cross}}\right). \tag{7}$$

Here, $\mathbf{s}_t^{\text{data}}$ records the platform, dataset scale, and available modalities; $\mathbf{s}_t^{\text{analysis}}$ records preprocessing, completed analyses, and reusable artifacts; and $\mathbf{s}_t^{\text{cross}}$ records specimen correspondence, coordinate compatibility, registration readiness, and ROI correspondence. Cross-modal predicates use three-valued semantics $\{\textit{true}, \textit{false}, \textit{unknown}\}$, and *unknown* is never promoted to *true*. The state determines which actions and comparisons are scientifically admissible; it does not determine which specialist is correct.

Evolving experience memory. *CHORUS* stores each completed interaction as an experience episode

$$e_i = (q_i, \mathbf{s}_i, \tau_i, o_i), \qquad \tau_i = \left[(g_{i1}, x_{i1}), \ldots, (g_{iL}, x_{iL})\right], \tag{8}$$

where τ_i is the executed tool trajectory and o_i summarizes execution success, scientific quality, agent verdicts, failures, and cost. For a new question, state compatibility first removes experiences that are unsafe or irrelevant; the remaining episodes are ranked by task similarity and observed outcome:

$$\mathcal{K}_t = operatornameRetrieve(q_t, \mathbf{s}_t; \mathcal{M}_t), \qquad \pi_t = \text{Plan}(q_t, \mathbf{s}_t, \mathcal{K}_t). \tag{9}$$

State therefore constrains the admissible strategy space, while memory ranks strategies within that space. After execution, *CHORUS* updates

$$\mathcal{M}_{t+1} = \mathcal{M}_t \cup \{(q_t, \mathbf{s}_t, \tau_t, o_t)\}. \tag{10}$$

Because new outcomes alter which trajectories are retrieved and preferred in later interactions, *CHORUS* evolves its behavior through experience without updating the underlying model parameters.

3.3 Generalist Multi-Agent Organization

Coordinator Agent. Different questions require different combinations of expertise. The Coordinator Agent converts the question, state, and retrieved experience into a structured orchestration decision

$$d_t = \mathcal{F}_{\text{C}}(q_t, \mathbf{s}_t, \mathcal{K}_t) = (\kappa_t, m_t, \mathcal{U}_t), \tag{11}$$

where κ_t is the task family, $m_t \in \{\textit{ST}, \textit{HE}, \textit{Joint}\}$ is the analysis mode, and $\mathcal{U}_t$ is the activated specialist set. The Coordinator minimizes unnecessary activation: image availability alone does not invoke the H&E Agent unless structural evidence is relevant to the question. If a joint task lacks pathological image, *CHORUS* preserves the joint intent, executes the admissible molecular path, and reports the fallback; an H&E-only request without an image terminates explicitly.

Modality specialists. Each specialist owns a distinct evidential role and a restricted tool scope. The ST Agent analyzes $(\mathbf{X}, \mathbf{S})$ to produce molecular domains, markers, annotations, and spatial statistics. The H&E Agent analyzes $\mathbf{H}$ to produce morphology, tissue-region, boundary, and image-quality evidence. Their common interface is

$$\mathcal{T}^{a,0} = \mathcal{F}_a(\mathcal{D}_a, q_t, \pi_t^a), \qquad a \in \mathcal{U}_t, \tag{12}$$

where $\mathcal{D}_a$ contains only the inputs permitted for specialist a. During the first pass, the ST Agent cannot consume histology or histology-derived features, and the H&E Agent cannot consume ST labels or outputs. The resulting modality-isolated trees preserve distinct scientific evidence before interaction begins.

Judgment Agent. Independently generated molecular and histological conclusions may appear inconsistent even when neither specialist is necessarily incorrect. Such disagreement can arise because the two claims refer to unmatched tissue regions, operate at different biological scales, or are supported by incomplete evidence. To resolve these cases without forcing consensus, *CHORUS* introduces a Judgment Agent that examines the aligned evidence bundles together with the relevant scientific state. It first diagnoses whether the unresolved relation is caused by scientific incomparability, missing evidence, scale complementarity, or a genuine contradiction. When the available evidence is insufficient, the agent converts the unresolved claim into a structured evidence request specifying the scientific question to verify, the required evidence, and the responsible specialist. The selected specialist then plans and executes the appropriate domain tools, after which the newly grounded evidence is returned for re-alignment and final adjudication. By separating disagreement diagnosis from evidence generation, the Judgment Agent guides targeted verification without fabricating or modifying scientific evidence.

3.4 Evidence-Grounded Collaborative Reasoning

Following the agent definitions above, *CHORUS* performs four steps for a joint analysis.

Step 1: Modality-isolated evidence construction. The Coordinator sends the same resolved scientific question to the activated ST and H&E Agents, but neither receives the other's results. Each specialist executes its own tools and converts their outputs into an initial logic tree $\mathcal{T}^{a,0}$. Deterministic adapters preserve real measurements, markers, regions, and quality metrics as evidence nodes. When a required measurement is absent, the adapter records uncertainty instead of inventing a biological label.

Step 2: Scientific-state-gated claim alignment. For each query-relevant conclusion, *CHORUS* extracts the conclusion and its minimal supporting ancestor closure to form a claim bundle B_a. Before classifying the relationship between two bundles, it evaluates

$$G(B_{\mathrm{ST}}, B_{\mathrm{HE}}; \mathbf{s}_t) \in \{\textit{Pass}, \textit{Unknown}, \textit{Fail}\}. \tag{13}$$

Spatial comparison requires confirmed specimen correspondence, compatible coordinates, valid registration, and corresponding ROIs. An unknown or failed gate produces *Incomparable*, not *Conflict*. Only when $G = \textit{Pass}$ does *CHORUS* assign

$$R(B_{\mathrm{ST}}, B_{\mathrm{HE}}) \in \{\textit{Agreement}, \textit{Complementary}, \textit{Conflict}, \textit{Uncertain}\}. \tag{14}$$

A conflict requires an explicit contradiction about the same registered target and biological scale. Different cluster identifiers, different numbers of regions, or different tool choices are not conflicts by themselves. When a registered region mapping is needed, a deterministic correspondence tool maps ST observations through the verified transform and summarizes overlap and mapping coverage. This computation describes how molecular and structural partitions correspond; it does not decide which partition is correct.

Step 3: Logic-tree-constrained discussion. Agreement and incomparability do not require debate. Complementary, conflicting, or uncertain bundles trigger a targeted exchange containing only the focal claim, its supporting ancestors, relevant peer evidence, and the state needed to interpret them. Each specialist chooses

$$a \in \{\textit{Keep}, \textit{Revise}, \textit{RequestEvidence}\}. \tag{15}$$

Both specialists condition on the same pre-discussion snapshot. This eliminates within-round revision leakage; conditional on the returned actions, the two local updates are order-independent. Evidence nodes are immutable. A specialist may revise only its own inference or conclusion, or request new evidence when the existing evidence is insufficient.

Definition 1 (Admissible discussion update)**.** *Let $\mathcal{T}^k$ be the logic tree owned by specialist $k \in \{\mathrm{ST}, \mathrm{HE}\}$. A discussion update is* admissible *if it satisfies one of the following conditions:* KEEP *leaves $\mathcal{T}^k$ unchanged;* REVISE *modifies only the statement and optional confidence of an owned* INFERENCE *or* CONCLUSION *node; and* REQUESTEVIDENCE *emits an evidence request without modifying the current tree. In all cases, evidence nodes, node types, spatial support, parent identifiers, provenance, and existing graph edges remain unchanged. Peer nodes may be included as discussion context but cannot be attached as parents of a local node.*

Theorem 1 (Evidence preservation and within-round independence)**.** *Assume that the initial specialist trees have nonempty evidence sets and are fully grounded, i.e., $\rho(\mathcal{T}^{k,0}) = 1$ for every specialist k. After any finite number R of admissible discussion rounds,*

$$\mathcal{V}_E^{k,R} = \mathcal{V}_E^{k,0}, \qquad \mathcal{E}^{k,R} = \mathcal{E}^{k,0}, \qquad \rho(\mathcal{T}^{k,R}) = 1, \tag{16}$$

and, for every evidence node v,

$$\mathrm{prov}_R(v) = \mathrm{prov}_0(v). \tag{17}$$

In addition, conditional on the actions returned from a shared pre-discussion snapshot, the two specialists' local updates commute:

$$\mathsf{U}_{\mathrm{ST}}^r \mathsf{U}_{\mathrm{HE}}^r = \mathsf{U}_{\mathrm{HE}}^r \mathsf{U}_{\mathrm{ST}}^r. \tag{18}$$

Therefore, discussion alone cannot create scientific evidence or introduce within-round revision leakage. The evidence set can expand only through a subsequent acquisition step whose outputs pass provenance validation. The full proof is provided in Appendix 5.

Step 4: Evidence verification, user clarification, and final adjudication. If uncertainty or disagreement remains after discussion, the Judgment Agent generates a small set of prioritized evidence requests. Each request specifies the unresolved scientific question, required evidence, rationale, related nodes, and responsible specialist. The selected specialist then plans and executes an appropriate domain tool, and successful outputs are appended as new grounded evidence nodes. After all admissible analyses have been completed, *CHORUS* re-aligns the updated logic trees and determines whether the available evidence supports agreement, complementarity, genuine discordance, or a modality-specific conclusion.

Additionally, *CHORUS* stores the scientific state, retrieved experience, specialist trajectories, initial and revised trees, evidence requests, and final adjudication as the new episode e_t. Thus the verified outcomes from the current analysis become experience for future multi-agent reasoning.

4 EXPERIMENTS

We evaluate four questions: whether CHORUS executes joint molecular–structural analysis with traceable evidence; whether it correctly completes multi-step quantitative ST tasks; whether evidence organization affects reasoning over identical spatial outputs; and whether its interaction and retrieval mechanisms enforce their stated constraints. These questions require different endpoints. Tool execution, numerical correctness, agreement with published tissue-region annotations, and implementation invariants are evaluated separately; none is treated as biological correctness of a final interpretation.

4.1 DATASETS AND EVALUATION DESIGN

Our primary collection is GSE208253, comprising 12 oral squamous cell carcinoma (OSCC) sections profiled using 10x Genomics Visium (Denisenko et al., 2024). Inputs include filtered gene-expression counts, spot coordinates, H&E images, and image-scale metadata. Across these sections,

Table 1: Evaluation tasks and assigned units

Task	Sections	Design	Assigned
Joint workflow/context	11	2 contexts × 2 repeats	44 raw runs
Joint workflow/ablation	14	2 models × 4 settings × 2 repeats	224 raw slots
Multi-step ST	4	2 tasks × 2 repeats	16 raw runs
Quantitative evidence	11	2 tasks × 2 formats × 2 protocols	88 answers
Frozen adjudication	11	2 contexts	22 judgments

the retained in-tissue matrices contain 969–3,591 spots and 36,601 Gene Expression feature rows per section. Visium observations are spots, not individual cells. Patient-level independence is not established by the evaluation manifest. Section s9 is reserved for development. Joint analysis uses the other 11 sections, s1–s8 and s10–s12. The multi-step ST panel uses s2, s3, s6, and s11, with 1,836, 969, 3,311, and 2,130 spots, respectively. The numerical panel is retrospective and does not establish generalization to previously unseen biological specimens. For region agreement, nonmissing `pathologist_anno.x` labels from the published section-specific annotation files associated with Figshare record 20304456 are reserved for scoring. Predictions and references are matched by barcode within the correct section. Published annotations provide a partition reference, not an independently collected human workflow comparator. The broader collection additionally contains one Xenium Prime FFPE ovarian-cancer section and two Xenium FFPE breast-cancer sections representing IDC/DCIS and ILC. These sections contribute to the complementary end-to-end comparison described below. Table 1 summarizes the principal assigned evaluations. Repeated executions measure runtime variability and do not increase the number of biological sections. The *Task Construction* and *Implementation and Metrics* can be found in Appendix 5 and 5

4.2 Results

Correspondence Evidence Exposure. We investigate the effect of explicitly exposing executed correspondence evidence to the judgment context. As shown in Table 2, the augmented context increases the visibility of the correspondence matrix from $0/22$ to $16/22$, while preserving $22/22$ process completion and deterministic H&E fallback. This confirms that the intervention successfully delivers existing cross-modal evidence to downstream judgment. Greater evidence visibility, however, does not yield stronger molecular–structural correspondence. The two settings achieve comparable partition agreement, with ARI of 0.146 vs. 0.138, region macro-F1 of 0.255 vs. 0.254, NMI of 0.253 vs. 0.248, and boundary F1 of 0.499 vs. 0.485. Likewise, dual-branch execution and verified correspondence metadata do not increase with the additional context. These results reveal a clear functional distinction: *evidence exposure* improves access to existing correspondence evidence, but cannot create missing upstream evidence or independently improve partition correspondence. Effective cross-modal reasoning therefore depends not on providing more context alone, but on scientifically verified evidence and its controlled use.

Table 2: Joint workflows and context ablation.

Endpoint	Original	Augmented
Process completed	22/22	22/22
Dual-branch execution, ST verified	21/22	19/22
Verified correspondence metadata	18/22	16/22
Matrix visible in added field	0/22	16/22
Deterministic H&E fallback	22/22	22/22
Scored partitions/assigned	18/22	16/22
Scored sections/assigned	10/11	11/11
ARI ↑	0.146 ± 0.095	0.138 ± 0.094
Region macro-F1 ↑	0.255 ± 0.104	0.254 ± 0.099
NMI ↑	0.253 ± 0.088	0.248 ± 0.086
Boundary F1 ↑	0.499 ± 0.200	0.485 ± 0.196

Agreement with published tissue-region annotations. We further assess how well the spatial partitions produced by CHORUS agree with published tissue-region annotations. As shown in Table 3, CHORUS shows consistently higher agreement across all four partition-level metrics. Compared with Leiden, ARI increases from 0.223 to 0.332, region macro-F1 from 0.381 to 0.485, NMI from 0.244 to 0.308, and boundary F1 from 0.470 to 0.537. A similar pattern is observed relative to

Table 3: Agreement with published tissue-region annotations

Method	Runs/sections	ARI ↑	Region macro-F1 ↑	NMI ↑	Boundary F1 ↑
CHORUS	60/12	0.332 ± 0.094	0.485 ± 0.079	0.308 ± 0.100	0.537 ± 0.160
Leiden	60/12	0.223 ± 0.096	0.381 ± 0.097	0.244 ± 0.092	0.470 ± 0.146
Louvain	60/12	0.214 ± 0.108	0.361 ± 0.088	0.234 ± 0.105	0.461 ± 0.147
SpaGCN	60/12	0.178 ± 0.088	0.314 ± 0.065	0.255 ± 0.104	0.483 ± 0.174
spEMO-CONCH adaptation	60/12	0.059 ± 0.074	0.203 ± 0.056	0.071 ± 0.087	0.115 ± 0.130

Louvain and SpaGCN, indicating that the difference is not driven by a single clustering metric, but is reflected across region correspondence, global partition structure, and spatial boundary agreement. The magnitude of the difference, however, varies across metrics. The gains are more pronounced for ARI and region macro-F1 than for boundary F1. For example, relative to SpaGCN, CHORUS improves ARI by 0.154, whereas boundary F1 increases by only 0.054. This suggests that the benefit of joint molecular–structural reasoning is more apparent in organizing coherent tissue regions and their correspondence than in local boundary localization alone. These results provide an external descriptive reference showing that the joint spatial partitions produced by CHORUS are well aligned with published tissue-region annotations.

End-to-end component ablations. We further remove the alignment gate, evidence acquisition, and discussion under two different language-model backbones to examine their effects on the joint workflow. As shown in Table 4, DeepSeek-flash (Xu et al., 2026) exhibits stable execution across configurations. The Full, No Alignment Gate, and No Discussion settings all complete $27/28$ workflows with $27/28$ ST-verified dual-branch executions, while removing acquisition only reduces both quantities to $26/28$. Thus, raw workflow completion is largely insensitive to removing alignment or discussion, with acquisition showing only a modest effect on end-to-end execution robustness. A larger difference arises across backbones. Under the Full configuration, DeepSeek-flash achieves ST-verified dual-branch execution in $27/28$ runs, compared with only $8/28$ for Qwen3-8B (Yang et al., 2025), despite similar process-completion rates of $27/28$ and $26/28$, respectively. This distinction shows that *process completion* is not equivalent to *verified joint execution*, motivating explicit checks of branch execution, input identity, and correspondence provenance in agentic cross-modal workflows. Interestingly, removing components from Qwen3-8B increases the frequency of conclusion text from $6/28$ under Full to $11/28$, $16/28$, and $12/28$ without the alignment gate, acquisition, and discussion, respectively. These increases are not accompanied by consistent improvements in ST-verified dual-branch execution. Producing a conclusion more frequently therefore should not be interpreted as more reliable cross-modal reasoning; rather, the results further distinguish *answer generation* from *evidence-verified reasoning*.

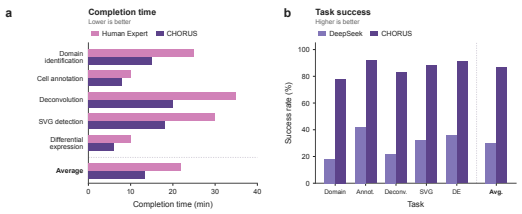


Figure 2: Performance across five downstream spatial biology tasks. (a) Completion time of CHORUS and human experts, with human results aggregated over 10 experts. (b) Success rates of CHORUS and Baseline. Average/Avg. denotes the mean across the five tasks.

4.3 Case Study

We demonstrate modality-specific and cross-modal analysis using Visium and Xenium data. For the ovarian cancer Visium dataset GSE211956, the agent estimates spot-level cell composition and tumor purity, using RCTD-derived tumor-cell enrichment weights from Denisenko et al. Denisenko et al. (2024) as the reference (GT). For Xenium, the agent checks marker availability and performs coarse cell annotation with scType, using reference cell-type maps from the CELLama study of lung adenocarcinoma Park et al. (2026). For cross-modal analysis, ST and H&E agents independently identify molecular and structural boundaries, verify registration, and examine boundary offsets. Within a 20,267-cell region, offset areas exhibit a 7.2-percentage-point higher immune-like

Table 4: End-to-end component ablations

Backbone	Setting	Completed	ST-verified dual branch	Conclusion text
DeepSeek-flash	Full	27/28	27/28	0/28
DeepSeek-flash	No alignment gate	27/28	27/28	0/28
DeepSeek-flash	No acquisition	26/28	26/28	0/28
DeepSeek-flash	No discussion	27/28	27/28	0/28
Qwen3-8B	Full	26/28	8/28	6/28
Qwen3-8B	No alignment gate	26/28	10/28	11/28
Qwen3-8B	No acquisition	24/28	10/28	16/28
Qwen3-8B	No discussion	26/28	13/28	12/28

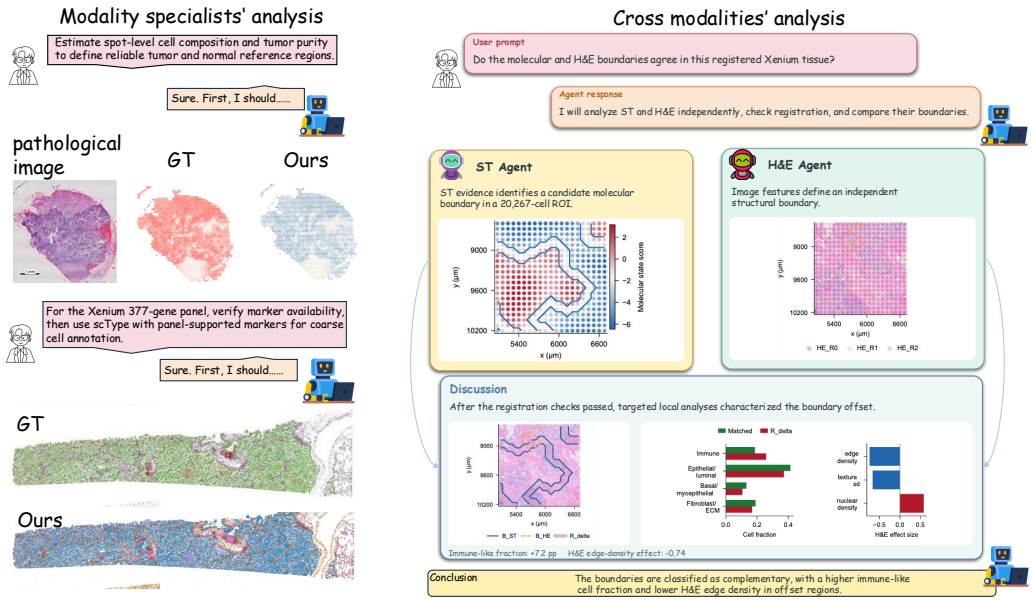


Figure 3: Case studies of modality-specific and joint cross-modal analysis. Left: Visium tumor-cell enrichment estimates and Xenium cell annotations compared with published reference results (GT). Right: ST and H&E agents jointly assess molecular and tissue-structural boundaries, using local cellular composition and morphological features to interpret their complementarity.

cell fraction and lower H&E edge density (effect size: -0.74) than matched areas. These findings support the complementarity of molecular and morphological boundaries, illustrating how agent collaboration integrates cellular and histological evidence to explain cross-modal differences.

User Study. As shown in Fig. 2, CHORUS requires less time than human experts on all five tasks, reducing average completion time from 22.0 to 13.4 minutes (39.1%). CHORUS also achieves higher success rates than DeepSeek on every task, averaging 86.4% versus 30.0%, a gain of 56.4 percentage points. These results demonstrate efficient execution and effective task completion across the evaluated tasks.

5 Conclusion

We introduced CHORUS, a multi-agent framework for spatial biology that coordinates independent molecular and histological specialists through provenance-grounded logic trees and scientific-state gating. Controlled evaluations show that preserving evidence and enforcing state compatibility improve cross-modal decisions, while state-conditioned memory improves accuracy and error recovery without model updates. Together, these mechanisms support auditable analysis and experience-driven adaptation.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, figure preparation, and discussion of mathematical arguments. All AI-assisted content was reviewed and verified by the authors against the relevant sources and experimental artifacts. The authors take full responsibility for the final manuscript.

REFERENCES

Yuheng Chen, Shi Han, Zitong Chao, Yuyao Liu, Fan Zhang, Hao Chen, Jiguang Wang, Jiashun Xiao, and Can Yang. Stat: A multi-agent framework for integrated and interactive spatial transcriptomics analysis. *bioRxiv*, pp. 2026–05, 2026.

Elena Denisenko, Leanne De Kock, Adeline Tan, Aaron B Beasley, Maria Beilin, Matthew E Jones, Rui Hou, Dáithí Ó Muirí, Sanela Bilic, G Raj KA Mohan, et al. Spatial transcriptomics reveals discrete tumour microenvironments and autocrine loops within ovarian cancer subclones. *Nature communications*, 15(1):2860, 2024.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.

Ruofan Jin, Zaixi Zhang, Mengdi Wang, and Le Cong. Stella: Self-evolving llm agent for biomedical research. *arXiv preprint arXiv:2507.02004*, 2025.

Yiyao Li, Nirvi Vakharia, Weixin Liang, Aaron T Mayer, Ruibang Luo, Alexandro E Trevino, and Zhenqin Wu. Omicsnavigator: An auditable scientific partner for scalable hypothesis validation in spatial omics. *bioRxiv*, pp. 2025–07, 2025.

Zuwan Lin, Wenbo Wang, Arnau Marin-Llobet, Qiang Li, Samuel D Pollock, Xin Sui, Almir Aljovic, Jaeyong Lee, Jongmin Baek, Ningyue Liang, et al. Spatial transcriptomics ai agent charts hpsc-pancreas maturation in vivo. *Biorxiv: the Preprint Server for Biology*, 2025.

Xinheng Lyu, Yuci Liang, Wenting Chen, Meidan Ding, Jiaqi Yang, Guolin Huang, Daokun Zhang, Xiangjian He, and Linlin Shen. Wsi-agents: A collaborative multi-agent system for multi-modal whole slide image analysis. *arXiv preprint arXiv:2507.14680*, 2025.

Siqi Ma, Jiajie Huang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. Medla: A logic-driven multi-agent framework for complex medical reasoning with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 845–853, 2026.

Jeongbin Park, Sumin Kim, Jiwon Kim, Dongjoo Lee, Sungwoo Bae, Haenara Shin, Daeseung Lee, and Hongyoon Choi. Cellama: foundation model for single cell and spatial transcriptomics by cell embedding leveraging language model abilities. *Advanced Science*, 13(7):e13210, 2026.

Gianna Pavilion, Cathrine Mbigidde, Rafael Tubelleza, Priya Ramarao-Milne, and Arutha Kulasinghe. Agentic ai for spatial omics. *Immunology and cell biology*, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, Xinheng Lyu, Tao Lin, et al. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 94611–94653, 2025.

Anurag J Vaidya, Felix Meissen, Daniel C Castro, Shruthi Bannur, Tristan Lazard, Drew FK Williamson, Faisal Mahmood, Javier Alvarez-Valle, Stephanie L Hyland, and Kenza Bouzid. Nova: An agentic framework for automated histopathology analysis and discovery. *arXiv preprint arXiv:2511.11324*, 2025.

Hanchen Wang, Yichun He, Paula P Coelho, Matthew Bucci, Abbas Nazir, Bob Chen, Linh T Trinh, Serena Zhang, Ziyu Lu, Kexin Huang, et al. Spatialagent: an autonomous ai agent for spatial biology. *bioRxiv*, pp. 2025–04, 2025.

Haoxiu Wang, Xinwang Yang, Siheng Wang, Zhe Yang, Xiuhui Yang, Yutong Yang, Zirong Li, Yuqi Ren, Qianqian Zhang, Bowen Zhao, et al. Spatial omics in the ai era: Technologies, algorithmic ecosystems, biological applications, and large model perspectives. *iMeta*, pp. e70146, 2026.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chen Yang, Xianyang Zhang, and Jun Chen. Chatspatial: schema-enforced agentic orchestration for reproducible and cross-platform spatial transcriptomics. *bioRxiv*, 2026.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

APPENDIX

RELATED WORK

Agents for spatial transcriptomics. LLM agents are increasingly used to automate spatial-transcriptomics workflows. STAgent combines code generation, visual interpretation, literature retrieval, and report synthesis for end-to-end ST analysis (Lin et al., 2025), while SpatialAgent extends autonomous planning across experimental design, data analysis, and hypothesis generation (Wang et al., 2025). ChatSpatial emphasizes reproducible tool orchestration through schema-constrained interfaces (Yang et al., 2026); STAT organizes ST analysis through multiple specialized agents and persistent sessions (Chen et al., 2026). These systems substantially improve the accessibility and automation of ST analysis, but their reasoning remains centered on molecular measurements and spatial coordinates. When tissue images are used, they are commonly inspected as auxiliary context or consumed within an image-aware backend such as SpaGCN (Hu et al., 2021), rather than analyzed through an independent tissue-evidence trajectory. CHORUS instead coordinates molecular and tissue-structural information as two evidence-producing processes within the same analysis.

Agents for histopathology. Agentic systems have also advanced computational pathology. NOVA translates scientific questions into iterative, executable histopathology pipelines and orchestrates a large collection of domain-specific tools (Vaidya et al., 2025). PathGen uses multiple agents to extract representative whole-slide-image patches and generate, revise, and summarize captions, producing a large-scale pathology image–text dataset (Sun et al., 2025). WSI-Agents further combines task allocation, specialized pathology models, verification, and summarization for multi-task whole-slide-image analysis (Lyu et al., 2025). These approaches establish strong pathology-centered analysis or data-generation workflows. They are not, however, designed to compare morphological claims with paired spatial transcriptomic evidence under verified specimen and coordinate correspondence. CHORUS complements these efforts by retaining an independent histopathology specialist while enabling its evidence to interact with molecular findings when the scientific state permits.

Multi-agent collaboration. Multi-agent frameworks improve complex reasoning by assigning complementary roles and allowing agents to critique or revise one another. MedLA makes this interaction explicit by representing each agent's medical reasoning as a logic tree and using graph-guided discussion to detect premise-level conflicts Ma et al. (2026). However, more discussion does not necessarily yield more reliable conclusions: multi-agent debate can amplify incorrect reasoning and move initially correct agents toward agreement with erroneous peers (Wynn et al., 2025). This risk is greater in tool-using science, where executed observations must not be overwritten by conversational persuasion. CHORUS therefore constrains collaboration at the evidence level. Tool-grounded evidence is immutable; specialists may revise only their own inferences or conclusions; both respond to the same pre-discussion snapshot; and scientific state is checked before a relation is classified. Remaining disagreement triggers targeted evidence acquisition or an unresolved outcome rather than additional debate for its own sake.

TASK CONSTRUCTION

T1: Joint molecular–structural correspondence. Given a section's expression matrix, coordinates, H&E image, and scale metadata, the system must independently produce molecular spatial domains and image-derived structural regions, establish whether local comparison is supported, and execute correspondence when admissible. The report must include actual mapped-observation coverage, a molecular-by-structural contingency summary, supporting artifacts, and unresolved prerequisites. Different cluster identifiers or numbers of regions do not by themselves establish biological conflict. Partial or unresolved answers are allowed when comparison is unsupported. We compare the original judgment context with an augmented context that includes the already executed correspondence matrix and its provenance. The intervention changes evidence delivery, without setting gate values or supplying a desired verdict. Each section receives both conditions with two paired repetition seeds. A complementary end-to-end comparison evaluates the full configuration and removal of the alignment gate, acquisition, or discussion using deepseek-flash and Qwen3-8B on 14 target sections with two repetitions, yielding 224 assigned positions. Its retained execution records are analyzed under their recorded model settings; they are not pooled with the context intervention as a single budget-matched experiment.

T2: Multi-step quantitative ST analysis. We construct AnnData inputs directly from the original in-tissue counts, without normalization, filtering, clustering, or biological labels. The only added observation field is `roi_quadrant`: full-resolution image coordinates x and y are independently divided at their section medians, with values at or below the median assigned L and higher values H. The four ROIs are LL, LH, HL, and HH, with the first letter denoting x. These are geometric regions, not tissue classes. Requested marker lists and computational definitions are provided, while actual feature availability and reference measurements are withheld. The first task uses Epithelial={KRT8, KRT18, EPCAM} and Myeloid={LST1, TYROBP, FCER1G}. The agent must execute native `score_marker_signatures`, report spot counts and both signature means in all four ROIs, select the ROI with the highest epithelial signature, and calculate its myeloid mean minus the mean over all spots in the other three ROIs pooled. It must also distinguish the global spot-weighted myeloid mean from the unweighted mean of the four ROI means. Thus selecting a region and evaluating a different signature require sequential operations, and unequal ROI sizes make the weighting consequential. The second task requests Mixed={CXCL9, CXCL10, CXCL13, MIR155, MIR21}. The agent must identify present and missing features from the input, report the number of available markers, calculate four ROI means and both weighted and unweighted aggregate means, rank ROIs, and return the smoothed extreme ratio $(\mu_{\max} + 1)/(\mu_{\min} + 1)$. Missing markers are excluded from the divisor; assay absence is not biological non-expression. Ties are resolved by lexicographic ROI order. Both tasks require one final JSON answer and the concrete native measurement CSVs. Marker signatures are raw-count summaries, not cell fractions or enrichment tests. For present requested markers G^+ and ROI R, the independent reference is

$$\mu_R = \frac{\sum_{i \in R} \sum_{g \in G^+} X_{ig}}{|R||G^+|}, \qquad \mu_{\text{global}} = \frac{\sum_R |R| \mu_R}{\sum_R |R|}, \qquad \mu_{\text{macro}} = \frac{1}{4} \sum_R \mu_R. \tag{19}$$

References are calculated by reading the original integer counts, without using native tool outputs. The first task's pooled-rest mean uses the same spot weighting after excluding the selected ROI.

T3: Reasoning over identical real spatial outputs. For each of the 11 joint sections, we use one frozen correspondence artifact from an executed upstream analysis. The input retains only its integer ST-by-HE contingency counts C_{ij}, original observation denominator, opaque region identifiers, and a visible artifact ID. Precomputed normalized overlaps, dominant mappings, coverage, relationship labels, and gate decisions are removed. The two conditions receive exactly the same information as either a flat table or a hierarchy of ST domains and HE-count leaves; reconstructing the table from the hierarchy must recover the identical counts. Each format is evaluated under an unaided single-completion protocol and a calculator-assisted protocol. The latter is a diagnostic extension on the same frozen tasks following the unaided evaluation, not a prospectively held-out validation set. Its read-only calculator exposes the supplied counts and identifiers for agent-specified mathematical operations, without reference answers. Thus 22 tasks yield 88 assigned answers across the four conditions. This is a component experiment on evidence organization and computational assistance, not a full CHORUS versus single-agent comparison. Q1 requires mapped-observation coverage, both directional weighted purities, macro-row purity, every ST domain's normalized structural entropy, and the two most mixed ST domains in order. Let $n_i = \sum_j C_{ij}$, $N = \sum_{ij} C_{ij}$, and K be the number of nonempty HE columns. The requested quantities include

$$P_{S \to H} = \frac{\sum_i \max_j C_{ij}}{N}, \quad P_{H \to S} = \frac{\sum_j \max_i C_{ij}}{N}, \quad \overline{P}_S = \frac{1}{|\{i : n_i > 0\}|} \sum_{i : n_i > 0} \frac{\max_j C_{ij}}{n_i}, \tag{20}$$

and $H_i = -\sum_{j : p_{ij} > 0} p_{ij} \log p_{ij} / \log K$, where $p_{ij} = C_{ij}/n_i$. Entropy is zero for $K = 1$ and undefined for empty rows. Ranking uses entropy rounded to 12 decimal places, followed by lexicographic ST ID to resolve ties. Q2 removes the HE column with the largest total, then requires the remaining mapped count, coverage over the original observations, retained fraction of originally mapped observations, both recomputed directional purities, and the set of ST domains whose dominant HE region changes. Dominant-region ties use lexicographic HE ID; a row losing all observations has a null remaining dominant and remains eligible for the change set. Both tasks cite the supplied artifact ID. References are independently recomputed from the integer table, not taken from the original correspondence scorer. Neither spatial mixing nor a changed dominant mapping is assigned a biological conflict or complementarity label.

T4: Interaction, adjudication, and retrieval constraints. We inspect final-judgment stability by freezing 11 executed upstream contexts and their declared correspondence artifacts, then regenerating only the judgment under original and augmented context. The first response and any native schema-repaired response are retained separately. No upstream scientific tools are rerun. We additionally execute 70 enumerated checks of native interaction components, including evidence immutability, ownership of revisions, shared pre-discussion snapshots, explicit state prerequisites, and malformed-response handling. These deterministic checks use controlled backend outputs and require no model calls; they assess implementation properties rather than biological accuracy. Retrieval is evaluated using two actually executed s9 development trajectories, which contain one unique tool sequence, and 16 explicit task-state queries. We compare no memory, native ranking, and a new explicit compatibility filter before ranking, using top-5 retrieval. Whole-procedure reuse is eligible only when no executed step is explicitly forbidden by the query and all required state fields are known and equal; unknown required values are ineligible. Final answers are removed uniformly from the memory records. This small deterministic diagnostic measures contract-compatible retrieval, not whether an LLM follows retrieved advice or improves its scientific answer. The hard filter is an added wrapper and its behavior must not be attributed to the unmodified native retriever.

IMPLEMENTATION AND METRICS

Execution settings. The T1 context intervention and T2 native task panel use `chat.py` through the instrumented execution harness, with local Qwen3-8B served by vLLM 0.26.0 as a text backbone (BF16, tensor parallelism two, two RTX 5090 GPUs, and a 40,960-token context). Image information is processed through scientific tools. T1 conditions share the 17-capability ST and 13-capability H&E catalogs and native fallback behavior; T2 uses the ST catalog. Raw runs have ceilings of 32 model calls, 400,000 total tokens, 2,048 generated tokens per call, and 24 instrumented scientific actions, with four CPU threads. One NOVA code program counts as one action even if it invokes several internal functions; bookkeeping is recorded separately. A 900-s soft timeout is supplemented by external termination after 1,080 s and a 15-s grace period. Both panels disable memory and use temperature 0.2. T1 repetition seeds are 2026091801/02; T2 seeds are 2026091901/02. T2 enforces a unique run-local destination for every native signature CSV at the tool boundary in all runs; this shared persistence support is not a memory intervention.

T3 uses the same Qwen3-8B backbone, temperature 0.2, and seed 2026091901. Its unaided protocol uses one completion capped at 4,096 output tokens, with no tools, retries, or schema repair. The assisted protocol allows at most four model calls and three read-only calculator calls; the frozen questions, references, and final-answer scorer remain unchanged. Because assistance also increases the allowed interaction budget, differences between these protocols cannot isolate calculator access from additional model computation. Formatting comparisons are made within each computation protocol. T4 judgment replay uses temperature 0, seed 20260919, and at most two 2,048-token calls, including one native repair. Actual returned model IDs, token usage, artifacts, failures, and runtime are recorded. Equal limits do not imply equal realized cost; concurrent service also limits interpretation of wall-time comparisons.

The 224-slot ablation panel uses the same 32-call, 400,000-token, 2,048-output-token ceilings and a configured 24-tool-round limit, temperature 0.2 and a 1,800-s wall limit; its two executions do not send a provider sampling seed. It requests deepseek-flash or qwen3-8b and uses the same registered-region-correspondence task across all 14 targets.

Execution and evidence. For T1 we report process completion, execution of both branches with independently verified ST input identity, correspondence count/coordinate/registration-metadata consistency, and final evidence availability. The H&E branch is checked for execution without an independent image-identity audit. Visibility refers specifically to the added final-context field. Metadata consistency and mapped coverage are not measurements of physical registration accuracy. H&E fallback regions are color/texture-derived structural partitions and do not establish pathological tissue identities.

Numerical answers and support. For T2/T3, counts, marker sets, identifiers, and prescribed ranking orders require exact agreement. Continuous values satisfy

$$|\hat{y} - y| \leq \max(10^{-3}, 10^{-3}|y|). \tag{21}$$

We score the final answer only and retain all assigned runs, including parse and execution failures. We report partial field correctness and correctness of all requested answer fields separately. In T2, scalar ROI entries count separately; marker sets and ordering lists count as one field each. Supported task success additionally requires a completed process, the requested final JSON contract, successful native measurements on the correct hashed input, and final-cited run-local CSVs whose values match independent references. Tool intermediates cannot fill missing final answers. In T3, the nine requested answer fields are scored, with nested-map correctness requiring all entries; its support endpoint checks the supplied artifact-ID citation and does not imply new tool execution. Field averages are not pooled across these different scoring granularities.

Partition agreement and comparisons. We match annotated spots by barcode. ARI measures pairwise partition agreement adjusted for chance. We use NMI with arithmetic normalization, $2I(Y;\hat{Y})/(H(Y)+H(\hat{Y}))$, region macro-F1 after Hungarian overlap matching, and boundary F1. Region macro-F1 includes the union of reference and mapped predicted labels, retaining unmatched predicted clusters. A boundary spot has a differently labeled member among its six nearest other scored spots; matching requires the opposite partition's boundary within 55 μm, represented using the section-specific spot diameter. Matches use directional nearest-neighbor distances rather than one-to-one assignment. Availability is reported alongside scores; missing partitions are not silently selected away or interpreted as valid predictions. Repeats are averaged within section before section-level summaries. The domain-method comparison includes Leiden, Louvain, SpaGCN, and a hires spEMO–CONCH adaptation using concatenated features on all 12 OSCC sections. It uses one previously selected CHORUS output and five runs per comparator per section, so it provides descriptive partition context rather than a matched agent-budget ranking.

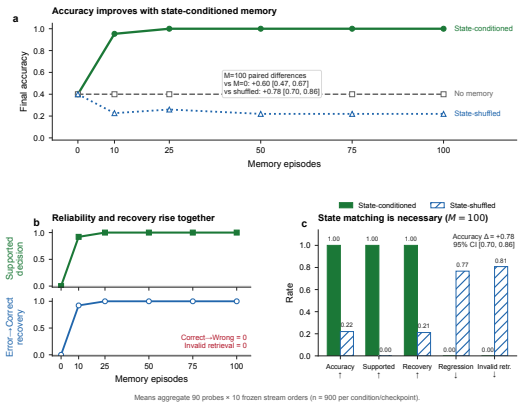


Figure 4: Memory-adaptation results. **(a)** Accuracy at six checkpoints under three memory conditions. **(b)** Two vertically separated panels with a shared horizontal axis show that Supported Decision and Error-to-Correct Recovery increase with Full State-conditioned Memory. **(c)** Full versus State-shuffled Memory at $M = 100$. Confidence intervals use 10,000 paired cluster-bootstrap replicates.

For Leiden and Louvain, we retain genes detected in at least three spots, normalize to 10,000 counts, log-transform, select 2,000 Seurat HVGs, scale with clipping at 10, and use 30 PCs, a 15-neighbor graph and resolution 0.5. SpaGCN additionally removes ERCC/MT genes, omits gene scaling, and uses spatial Louvain initialization with 10 neighbors, up to 30 PCs, resolution 0.5, learning rate 0.05 and 200 epochs. No reference-derived cluster count is supplied. Comparator seeds are 0–4. The spEMO–CONCH entry is a hires feature-concatenation adaptation rather than a claim about the original method under its official protocol.

For T3 format differences, we average Q1 and Q2 within each source case and use 10,000 paired source-case bootstrap resamples with seed 20260919. The resulting 95 percent percentile intervals are descriptive case-level uncertainty, not patient-level generalization intervals.

Constraint endpoints. T4 reports first-response schema validity, verdict preservation after repair, and decisive-evidence references separately. Interaction checks report protected-property violations and the exact tested boundary. Retrieval reports compatible and incompatible returned procedures and the number retained, separating abstention from successful retrieval. Rejection under a declared hard rule is not evidence of improved answer accuracy. Human-expert answers or blinded scientific ratings have not been collected.

ADDITIONAL EXPERIMENTS

Cost and interpretation. Table 5 reports actual inference usage for the controlled panels. The demonstrated scope is tool-mediated workflow execution, independently scored quantitative reporting, regional agreement, and tested implementation constraints. Evidence transport alone does not ensure correct evidence use. The present results do not establish expert-level biological interpretation, a general accuracy benefit from memory, or universal superiority.

Table 5: Observed inference cost in the controlled panels

Panel	Assigned	Model calls	Total tokens
Joint / original	22	493	3821722
Joint / augmented	22	496	3957726
Multi-step ST	16	121	1555574
Evidence / Unaided / flat	22	22	20297
Evidence / Unaided / tree	22	22	24197
Evidence / Assisted / flat	22	76	85545
Evidence / Assisted / tree	22	80	111927
Adjudication replay	22	44	233687

Actual provider usage, excluding development runs, service checks, deterministic tests and the complementary 224-slot panel. Shared-service wall time is not used as an efficiency ranking.

MEMORY ADAPTATION

Figure 4 summarizes how memory growth affects accuracy, reliability, and error recovery. With Full State-conditioned Memory, Final Accuracy increases from 0.400 at $M = 0$ to 0.953 at $M = 10$ and reaches 1.000 for $M \geq 25$. The paired improvement at $M = 100$ over $M = 0$ is 0.600 (95% CI $[0.467, 0.667]$). Supported Decision Rate and Error-to-Correct Recovery Rate likewise increase from 0 to 1.000, while Correct-to-Wrong Regression and Invalid Retrieval remain zero at every checkpoint. At $M = 100$, State-shuffled Memory obtains only 0.220 Final Accuracy, 0 Supported Decision, and 0.211 Error-to-Correct Recovery, while its Correct-to-Wrong Regression and Invalid Retrieval rates rise to 0.767 and 0.808, respectively. Full State-conditioned Memory exceeds this control in accuracy by 0.780 (95% CI $[0.700, 0.860]$). The preregistered claim gate is therefore satisfied, supporting a scoped conclusion: reusable experience improves task accuracy, decision reliability, and recovery from erroneous trajectories only when retrieval is matched to the current scientific state.

PROOF OF DISCUSSION SAFETY

Proof of Theorem 1. We proceed by induction over discussion updates. Before discussion, all evidence nodes are grounded by assumption.

For the induction step, consider the three admissible actions in Definition 1. A KEEP action is the identity operation and therefore preserves every tree component. A REVISE action is defined only for an owned INFERENCE or CONCLUSION node and leaves the evidence-node set, graph edges, parent identifiers, and provenance unchanged. A REQUESTEVIDENCE action emits a request but does not modify the current tree. Hence, none of the three cases can add, delete, modify, or reattribute an evidence node. The evidence set and its provenance are consequently invariant after any finite number of discussion updates. Since every preserved evidence node satisfies $g(v) = 1$, the grounding fraction remains one.

For within-round independence, both actions are computed from the same pre-discussion snapshot. Once these actions have been returned, the ST update acts only on $\mathcal{T}^{\mathrm{ST}}$, whereas the H&E update

acts only on $\mathcal{T}^{\mathrm{HE}}$. The two operators therefore have disjoint supports and commute, so their application order cannot affect the resulting pair of trees. In particular, neither action is conditioned on the peer specialist's current-round revision.

Discussion does not modify the evidence set. If the subsequent evidence-acquisition step appends a set of new nodes only after verifying $g(v) = 1$ for each appended node, then the grounding invariant is preserved after acquisition as well. □