MURALAGENT: LANGUAGE-GUIDED AGENTIC INSPECTION FOR LARGE-SCALE MURAL STITCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale image stitching is essential for the digital preservation of cultural heritage murals, yet stitching artifacts are difficult to identify due to their sparse distribution in images and their visual similarity to intrinsic mural degradations. Existing vision-language models and multimodal agents often rely on fixed or general-purpose visual exploration, making it difficult to efficiently acquire the evidence needed to resolve such ambiguities under a limited inspection budget. We formulate mural stitching inspection as an uncertainty-driven evidence acquisition problem, where an agent must decide not only where to look, but also what evidence to acquire to resolve uncertain observations. We present **MuralAgent**, a multimodal agent that performs closed-loop inspection by maintaining spatial beliefs and competing hypotheses, adaptively selecting visual observations and hypothesis-conditioned forensic evidence. It further performs source-aware attribution to distinguish stitching-induced artifacts from intrinsic mural degradation. Experiments on MuralSI-Bench show that MuralAgent improves attribution F1 by 11.6 percentage points over the strongest multimodal agent baseline while reducing high-resolution inspection area by 44.3%.

1 INTRODUCTION

Large-scale image stitching is an important technique for the digital preservation of cultural heritage, enabling reconstruction of murals whose spatial extent exceeds the field of view of individual captures (Jiang et al., 2026; Li et al., 2025b; Wang & Yang, 2020; Lin et al., 2015). The fidelity of these reconstructions, however, can be compromised by stitching artifacts arising from inaccurate registration, geometric deformation, or imperfect blending, including geometric misalignment, ghosting, visible seams, appearance inconsistency, and duplicated or missing content. Reliable inspection of such artifacts is therefore essential before reconstructed murals can support subsequent preservation, restoration, and analysis. Mural stitching inspection is particularly challenging due to the large spatial scale and complex appearance of historical murals (Mei et al., 2024; Wang et al., 2026; Fang et al., 2026). As illustrated in Fig. 1, stitching artifacts are often sparse and scale-dependent: subtle local errors can occupy only a tiny fraction of the complete mural and become observable only at sufficiently high resolution. Meanwhile, historical murals inherently contain cracks, fading, material loss, and irregular painted structures that can closely resemble stitching artifacts. Consequently, reliable inspection requires more than detecting suspicious regions from a single observation. An inspector must progressively gather visual and source evidence to determine whether an observed anomaly is truly stitching-induced or instead originates from the mural itself.

This observation raises a fundamental question: *given a limited inspection budget, what should an agent observe next, and what evidence should it acquire to resolve the current uncertainty?* Conventional image-quality assessment and anomaly detection methods typically evaluate predefined images or local regions, providing limited support for actively searching extremely large visual spaces or reasoning about the origin of an anomaly. Vision-language models (VLMs) offer stronger semantic reasoning, yet directly processing a downsampled mural can suppress small artifacts, while exhaustive patch-wise inspection incurs substantial computation without adapting visual attention to the current uncertainty. More importantly, visual appearance alone may be insufficient to distinguish stitching artifacts from intrinsic mural degradation. These challenges call for an inspection paradigm that treats visual observations as evidence acquired sequentially, rather than as a fixed input to a one-shot prediction. Recent multimodal agents (Zeng et al., 2026; Zhang et al., 2026b; Xia et al., 2026;

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

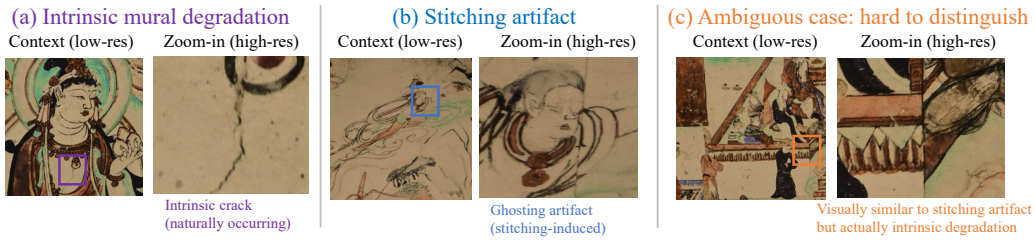

Figure 1: **Motivation for agentic mural stitching inspection.** (a) An intrinsic mural crack. (b) A stitching-induced ghosting artifact. (c) An ambiguous intrinsic degradation that visually resembles a stitching artifact. These examples highlight the need for active multi-scale inspection and evidence-grounded artifact attribution.

He et al., 2026; Lin et al., 2026) provide a promising foundation for such interactive visual reasoning. These methods have demonstrated iterative visual reasoning, tool use, multi-critic verification, reflection, and experience memory in multimodal creation and editing tasks. However, mural inspection presents a different requirement: the agent must actively allocate a limited visual budget according to the uncertainty of its current beliefs (Colace et al., 2025; Shen et al., 2025; Cao et al., 2025). In particular, it needs to determine *where and at what scale* to inspect, maintain alternative explanations for ambiguous observations, and acquire *discriminative evidence* that can distinguish competing explanations. Thus, the key challenge is not simply to reason over visual observations, but to decide which observations and evidence are most informative for the current inspection state.

To address this challenge, we formulate large-scale mural stitching inspection as an *uncertainty-driven evidence acquisition* problem and propose **MuralAgent**, a training-free multimodal agent for active and evidence-grounded inspection. Given a natural-language inspection request, MuralAgent performs closed-loop inspection by maintaining an instruction-conditioned spatial belief and competing hypotheses over suspicious regions. Based on the evolving inspection state, it adaptively selects visual actions to acquire more informative observations at appropriate locations and scales. When visual observations remain insufficient to distinguish competing hypotheses, MuralAgent further acquires hypothesis-conditioned forensic evidence, including semantic, geometric, appearance, and source-level evidence. Source observations enable the agent to attribute detected abnormalities to either the stitching process or intrinsic mural degradation. A reflective experience memory is additionally used to reuse informative inspection strategies across subsequent episodes. We construct **MuralSI-Bench** from real Dunhuang murals in MuralDH, combining intrinsic mural degradations with controllable stitching artifacts and source-level provenance. Extensive experiments demonstrate that MuralAgent consistently outperforms existing methods in artifact detection, localization, and attribution, while requiring substantially fewer VLM calls.

2 Related Work

Mural image stitching and inspection. Image stitching reconstructs wide-field or high-resolution scenes from overlapping observations. Classical methods improve alignment through spatially varying warps (Zaragoza et al., 2013; Chang et al., 2014), while learning-based approaches such as UDIS (Nie et al., 2021) and UDIS++ (Nie et al., 2023) address alignment and parallax with deep models. For cultural heritage, MuralDH (Xu et al., 2024) provides Dunhuang mural images with annotations of intrinsic deterioration. Existing studies primarily focus on image reconstruction or mural degradation analysis. We instead study *post-stitching inspection*, aiming to actively locate sparse stitching artifacts and distinguish them from intrinsic mural degradations.

Multimodal agentic visual reasoning. Recent multimodal agents combine visual reasoning, tool use, verification, and memory for complex visual tasks. MIRA (Zeng et al., 2026) and IntentEdit (Zhang et al., 2026b) explore iterative visual reasoning and reflective execution, while SAGE (Xia et al., 2026) employs multi-critic verification and tool interaction. GEMS (He et al., 2026) and JarvisEvo (Lin et al., 2026) further introduce experience memory and reflection, with related evidence- and tool-driven mechanisms explored in Data2Story (Qinghong Lin et al., 2026), NEWTON (Feng et al., 2026), and WorldClaw (Guo et al., 2026). MuralAgent targets a distinct *evidence-seeking inspection* problem, actively exploring large-scale murals, maintaining competing artifact hypotheses, and using forensic and source evidence for reliable stitching-specific attribution.

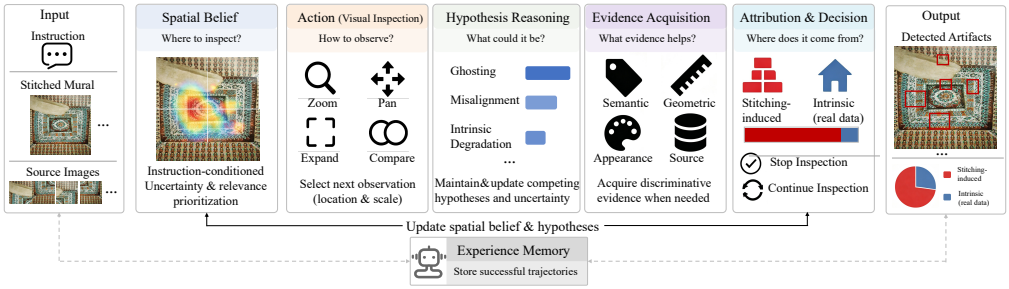

Figure 2: **The architecture of MuralAgent.** MuralAgent follows a closed-loop process of *spatial prioritization, informative observation, hypothesis resolution, and artifact provenance verification*.

3 METHOD

We propose **MuralAgent** as shown in Figure 2, a language-guided multimodal agent for mural stitching artifact inspection. Given a natural-language instruction, MuralAgent sequentially explores the mural, maintains competing artifact hypotheses, acquires targeted forensic evidence, and accumulates reusable inspection experience.

3.1 PROBLEM FORMULATION

Given a large-scale stitched mural image $I \in \mathbb{R}^{H\times W\times 3}$ and a natural-language instruction q, our goal is to identify stitching artifacts relevant to the specified regions or artifact types. We represent the inspection output a $\mathcal{Y} = \{y_i\}_{i=1}^{N}$, $y_i = (b_i, c_i, p_i, d_i, \mathcal{E}_i)$ denotes the artifact location, category, confidence, natural-language explanation, and supporting evidence, respectively. We consider six artifact types (Xu et al., 2026; Wen et al., 2026; Yang et al., 2025): geometric misalignment, ghosting, visible seams, appearance inconsistency, content duplication, and content missing.

3.2 INTERLEAVED LANGUAGE-GUIDED VISUAL INSPECTION

Large-scale murals contain stitching artifacts with diverse spatial extents and visual characteristics (Ren et al., 2024; Liu et al., 2025). We develop an interleaved visual inspection mechanism that jointly maintains instruction-conditioned spatial beliefs and competing artifact hypotheses, and progressively acquires multi-scale observations to resolve them, following the general principle of actively acquiring task-relevant visual evidence under partial observability (Kolner et al., 2025; Zhang et al., 2026a; Li et al., 2025a).

Instruction-conditioned inspection state. Given an instruction q, MuralAgent extracts a compact inspection specification $G_q = \Phi_{\text{ins}}(q) = \{\mathcal{O}_q, \mathcal{C}_q, \mathcal{R}_q\}$. $\mathcal{O}_q$ denotes semantic targets of interest, $\mathcal{C}_q$ specifies candidate stitching artifact types, and $\mathcal{R}_q$ describes inspection constraints such as regions and preferred observation scales. For a general inspection request, $\mathcal{C}_q$ is initialized with the complete artifact taxonomy and progressively specialized during inspection. At step t, the agent maintains an inspection state $S_t = (G_q, R_t, l_t, M_t^{\text{sus}}, M_t^{\text{cov}}, \mathcal{H}_t)$. R_t and l_t denote the current spatial region and observation scale. M_t^{sus} is a spatial suspicion map encoding the current likelihood of instruction-relevant stitching artifacts, M_t^{cov} records inspection coverage across spatial regions and scales, and $\mathcal{H}_t$ maintains competing explanations for the current observations.

Instruction-conditioned spatial belief. We represent the mural using a multi-resolution image pyramid (Burt & Adelson, 1987; Zhu et al., 2024; Wang et al., 2025) $\mathcal{I} = \{I^{(0)}, I^{(1)}, \ldots, I^{(L)}\}$ and initialize M_0^{sus} from coarse-scale observations conditioned on G_q. For each candidate region R, we compute its inspection priority as $U_t(R) = S_t^{\text{vis}}(R) + S_t^{\text{lang}}(R) + S_t^{\text{unc}}(R) - S_t^{\text{cov}}(R)$. Here, $S_t^{\text{vis}}(R)$ measures local visual abnormality, $S_t^{\text{lang}}(R)$ is the cosine similarity (Radford et al., 2021; Cicchetti et al., 2026) between the visual embedding of R and the textual embedding of the instruction-derived targets in G_q, and $S_t^{\text{unc}}(R) = -\sum_k p_t^k(R) \log p_t^k(R) / \log K$ measures the uncertainty of the current hypotheses. $S_t^{\text{cov}}(R)$ is the fraction of spatial-scale cells associated with R that have already been observed. All terms are normalized to $[0, 1]$, yielding a spatial priority that jointly favors visually suspicious, instruction-relevant, uncertain, and insufficiently inspected

regions. After acquiring a new observation, we update the spatial belief as

$$M_{t+1}^{\text{sus}} = \Phi_{\text{belief}}\left(M_t^{\text{sus}}, O_{t+1}, \mathcal{H}_t, G_q\right), \tag{1}$$

where Φ_{belief} recomputes the four scores for the affected regions and updates their priorities, while M_{t+1}^{cov} records the newly observed region and scale. This enables MuralAgent to progressively redistribute its inspection budget as new visual evidence becomes available.

Interleaved multi-hypothesis reasoning. For the selected region, MuralAgent jointly observes global context and local details $O_t = \left\{I^{\text{global}}, I[R_t, l_t], I[R_t, l_t + 1]\right\}$. Based on O_t and G_q, the agent constructs a set of competing hypotheses

$$\mathcal{H}_t = \left\{(h_t^k, p_t^k)\right\}_{k=1}^{K}, \qquad \sum_{k=1}^{K} p_t^k = 1,$$

where h_t^k represents a possible explanation of the observed pattern and p_t^k denotes its current plausibility. The hypothesis space jointly covers stitching artifacts and intrinsic mural structures. For example, a local double contour may instantiate hypotheses corresponding to `ghosting`, `painted double contour`, and `mural crack`. Each newly acquired observation updates the hypothesis state through

$$\mathcal{H}_{t+1} = \Phi_{\text{reason}}\left(\mathcal{H}_t, O_{t+1}, G_q\right), \tag{2}$$

where Φ_{reason} is implemented by the VLM as an evidence-conditioned hypothesis reasoner. Given the previous hypotheses and the newly acquired multi-scale observation, the VLM evaluates each hypothesis against the new evidence and outputs a structured support score $r_{t+1}^k \in [0, 1]$. We combine this evidence with the previous plausibility and normalize across hypotheses $p_{t+1}^k = \frac{p_t^k r_{t+1}^k}{\sum_j p_t^j r_{t+1}^j}$. Hypotheses with negligible support are discarded, while a new hypothesis can be introduced when the observation cannot be sufficiently explained by the existing set; the top-K hypotheses are retained after normalization. The resulting distribution is propagated to the spatial belief in Eq. 1, coupling local hypothesis revision with global inspection.

Uncertainty-guided atomic inspection. MuralAgent interacts with the multi-resolution mural through a set of atomic visual actions $\mathcal{A}^{\text{vis}} = \{\texttt{ZOOM}, \texttt{PAN}, \texttt{EXPAND}, \texttt{COMPARE}, \texttt{BACKTRACK}, \texttt{VERIFY}\}$. These actions control the spatial location, observation scale, and contextual information available to the agent. `ZOOM` acquires finer visual details, `PAN` moves the inspection window to a neighboring region, `EXPAND` enlarges the contextual field of view, `COMPARE` retrieves adjacent or corresponding regions for consistency reasoning, `BACKTRACK` restores a previous inspection state, and `VERIFY` requests forensic evidence for the current hypotheses.

We measure the ambiguity of the current hypotheses using $\mathcal{U}_t = -\sum_{k=1}^{K} p_t^k \log p_t^k$. MuralAgent first selects the unresolved region with the highest spatial priority $U_t(R)$ and then determines the visual action from its current spatial and hypothesis states. Specifically, `ZOOM` is selected when the current resolution is insufficient to resolve a hypothesis; `EXPAND` is used when additional surrounding context is required; `PAN` moves to a neighboring high-priority region; and `COMPARE` is triggered when a hypothesis requires cross-region or cross-scale consistency. When newly acquired observations contradict the current inspection branch, the agent invokes `BACKTRACK`. If visual observations remain ambiguous and the current hypotheses require geometric, appearance, semantic, or source evidence, the agent invokes `VERIFY` and enters the evidence verification stage in Sec. 3.3.

After executing an action, the resulting observation updates the competing hypotheses through Φ_{reason} and subsequently updates the spatial belief M_t^{sus}. The inspection therefore proceeds as

$$\boxed{\text{Observe} \rightarrow \text{Hypothesize} \rightarrow \text{Act} \rightarrow \text{Re-observe} \rightarrow \text{Revise}}.$$

The loop continues until the current region is resolved, transferred to evidence verification, or deprioritized in favor of another unresolved region.

3.3 Evidence Verification and Source-Aware Attribution

Visual observations alone can leave multiple explanations plausible for a suspicious region. MuralAgent therefore performs evidence-driven verification by adaptively acquiring complementary

evidence according to the current hypothesis state, and subsequently attributes the observed anomaly to either the stitching process or the intrinsic mural content.

Hypothesis-conditioned evidence routing. Given the current hypotheses $\mathcal{H}_t$, MuralAgent acquires complementary forensic evidence to verify suspicious observations. We organize the available tools into four evidence families $\mathcal{T} = \{\mathcal{T}^{\text{sem}}, \mathcal{T}^{\text{geo}}, \mathcal{T}^{\text{app}}, \mathcal{T}^{\text{src}}\}$, corresponding to semantic, geometric, appearance, and source evidence. Semantic tools examine structural consistency of mural content; geometric tools measure feature correspondence, contour continuity, and registration residuals; appearance tools evaluate color, gradient, texture, and sharpness consistency; and provenance tools retrieve source tiles, overlap regions, and stitching-boundary information.

We use a predefined hypothesis–evidence mapping to determine the relevant evidence families for each artifact hypothesis. Geometric misalignment and ghosting prioritize geometric and source evidence; visible seams and appearance inconsistencies invoke appearance and source evidence; and duplicated or missing content primarily uses semantic and source evidence. When multiple hypotheses remain active, MuralAgent takes the union of their required evidence families and skips evidence that has already been acquired for the current region. This produces a compact evidence set targeted to the current ambiguity without requiring an additional learned or VLM-based tool-scoring module. At verification step j, the selected tool returns evidence e_j, which is appended to the current evidence set:

$$\mathcal{E}_{j+1} = \mathcal{E}_j \cup \{e_j\}, \qquad \mathcal{H}_{j+1} = \Phi_{\text{verify}}(\mathcal{H}_j, \mathcal{E}_{j+1}). \tag{3}$$

The returned evidence is appended to $\mathcal{E}_j$ and passed to the multi-evidence verifier described below.

Multi-evidence forensic reasoning. The acquired evidence is organized into four complementary groups, $\mathcal{E}_j = \{E_j^{\text{sem}}, E_j^{\text{geo}}, E_j^{\text{app}}, E_j^{\text{prov}}\}$. Each group retains the original tool measurements together with their spatial support, such as feature displacement and registration residuals for geometric evidence, cross-boundary color and gradient differences for appearance evidence, and source-tile correspondence and boundary location for source evidence.

MuralAgent jointly presents the available evidence and competing hypotheses to the evidence-conditioned verifier Φ_{verify}. For each hypothesis, the verifier determines whether each available evidence group is `supporting`, `contradicting`, or `inconclusive`, and revises the hypothesis set accordingly. A hypothesis proceeds to attribution when it is consistently supported by its required evidence families and no acquired evidence provides a direct contradiction. Conflicting or insufficient evidence keeps the hypothesis unresolved and triggers acquisition of the next evidence source specified by the hypothesis–evidence mapping.

Source-aware artifact attribution. For mural inspection, verification further requires determining whether an observed anomaly is introduced during stitching or already exists in the underlying mural. MuralAgent retrieves candidate source observations corresponding to the suspicious region using the original stitching inputs and estimated geometric correspondences. Let $A(\cdot)$ denote the anomaly evidence associated with the verified hypothesis and I_i^{src} denote the retrieved source observations. We define the source contrast as $\Delta A = A(I^{\text{stitch}}) - \text{Agg}_i\, A(I_i^{\text{src}})$, which measures whether the observed abnormality is already present in the source observations or emerges after reconstruction. The final attribution is inferred jointly from the forensic evidence and source contrast:

$$z = \Phi_{\text{attr}}\left(\mathcal{H}, \mathcal{E}, \Delta A\right), \qquad z \in \{\texttt{stitching}, \texttt{intrinsic}, \texttt{uncertain}\}. \tag{4}$$

An anomaly that is absent from the corresponding source observations but emerges together with geometric or appearance inconsistencies after reconstruction supports a `stitching` attribution. Conversely, an anomaly consistently observed in the source images supports an `intrinsic` attribution. Cases with insufficient or conflicting evidence remain `uncertain` and are returned to the inspection loop for additional observation. This verification process progressively refines competing hypotheses using complementary forensic measurements and source observations, producing evidence-grounded attribution without access to ground-truth artifact locations, stitching boundaries, or perturbation metadata.

3.4 Reflective Experience Memory

MuralAgent maintains a hierarchical experience memory that preserves inspection trajectories and distills reusable knowledge from verified inspections, following reflective memory mechanisms in language agents (Shinn et al., 2023; Zhao et al., 2024).

Hierarchical inspection memory. We organize memory into three levels: *factual*, *trajectory*, and *experience* memory. Factual memory stores verified findings, including artifact hypotheses, evidence, and attribution results. Trajectory memory records observations, actions, tool calls, and hypothesis revisions within an inspection episode. After completing a trajectory τ, MuralAgent compresses it into $m^{\mathrm{exp}} = \Phi_{\mathrm{mem}}(\tau)$, which summarizes discriminative cues, effective actions, informative evidence, and failure patterns without retaining the complete trajectory.

Experience-conditioned inspection. After each verified inspection, MuralAgent reflects on the collected observations and evidence, and stores the resulting strategy in experience memory. At each step, relevant experiences are retrieved according to the current state $\mathcal{M}_t^{\mathrm{rel}} = \mathrm{Retrieve}(S_t, \mathcal{M}^{\mathrm{exp}})$.The retrieved experiences guide region exploration, hypothesis revision, and evidence selection. Repeatedly validated experiences are further consolidated into reusable inspection skills, enabling previous inspection outcomes to improve subsequent decisions.

4 Experiments

4.1 Experimental Setup

Datasets. We primarily evaluate on **MuralSI-Bench**, constructed from real Dunhuang murals in MuralDH (Xu et al., 2024). It contains six stitching artifact categories—geometric misalignment, ghosting, visible seam, appearance inconsistency, content duplication, and content missing—at multiple severity levels, while preserving intrinsic degradations such as cracks, fading, and material loss. Each artifact is annotated with its location, category, severity, and source correspondence. Construction details and statistics are provided in Appendix A.1.

Baselines. We compare three groups of methods: visual baselines (DINOv2 (Oquab et al., 2024), WinCLIP (Jeong et al., 2023), and AnomalyCLIP (Zhou et al., 2024)); VLM baselines using direct, sliding-window, and fixed hierarchical inspection; and multimodal agents, including MIRA (Zeng et al., 2026), IntentEdit (Zhang et al., 2026b), SAGE (Xia et al., 2026), GEMS (He et al., 2026), Data2Story (Qinghong Lin et al., 2026), and JarvisEvo (Lin et al., 2026). All agent methods share the same VLM backbone, observations, forensic tools, and inference budget, differing only in their reasoning and control strategies. Adaptation details are provided in Appendix A.2.

Evaluation metrics. We report detection F1 (**Det. F1**), localization AP_{50}, artifact-type F1 (**Type F1**), and attribution F1 (**Attr. F1**). Inspection efficiency is measured by the percentage of mural area accessed at high resolution (**Area**) and the average number of VLM calls. For language-guided inspection, predictions are considered relevant only when they satisfy the region and artifact constraints specified by the instruction. Detailed metric definitions are provided in Appendix A.

Implementation. All agent methods use the same inference budgets and a shared forensic tool pool covering semantic, geometric, appearance, and source evidence. Experience memory is constructed exclusively from the development split and remains fixed during test evaluation. Detailed configurations, tool implementations, prompts, budgets, and baseline adaptations are provided in Appendix A.3. The operators Φ denote training-free VLM prompting or deterministic state updates; their exact prompts and decision rules are given in Appendix C.

4.2 Main Results

Table 1 compares MuralAgent with conventional visual methods, VLM-based inspection strategies, and adapted multimodal agent frameworks on MuralSI-Bench.

Table 1: Main results on MuralSI-Bench.

Method	Det. F1 ↑	AP_{50} ↑	Type F1 ↑	Attr. F1 ↑	Area (%) ↓	Calls ↓
Visual anomaly detection						
DINOv2	0.312	0.247	0.198	–	100	–
WinCLIP	0.358	0.291	0.241	–	100	–
AnomalyCLIP	0.401	0.336	0.279	–	100	–
VLM-based inspection						
Direct VLM	0.487	0.412	0.365	0.298	100	1.0
Sliding-window VLM	0.623	0.558	0.501	0.437	100	48.6
Hierarchical VLM	0.671	0.604	0.552	0.489	67.3	29.4
Multimodal agents						
MIRA	0.714	0.653	0.598	0.541	41.8	22.7
IntentEdit	0.698	0.641	0.583	0.527	38.5	19.3
SAGE	0.731	0.672	0.615	0.558	35.2	17.9
GEMS	0.725	0.661	0.607	0.549	36.9	18.6
Data2Story	0.709	0.648	0.591	0.533	39.7	20.4
JarvisEvo	0.742	0.685	0.628	0.571	33.4	16.2
MuralAgent	**0.813**	**0.764**	**0.721**	**0.687**	**18.6**	**9.4**

Overall inspection performance. MuralAgent consistently achieves the best performance across all inspection tasks, reaching 0.813 Det. F1, 0.764 AP_{50}, 0.721 Type F1, and 0.687 Attr. F1. Compared with JarvisEvo, the strongest agent baseline, MuralAgent improves these metrics by 7.1, 7.9, 9.3, and 11.6 percentage points, respectively. The particularly large gain in attribution indicates the benefit of jointly reasoning over competing hypotheses and source-level evidence for distinguishing stitching artifacts from intrinsic mural degradations.

Comparison with VLM and multimodal agents. Direct VLM inspection performs poorly on large-scale murals (0.487 Det. F1), while sliding-window and hierarchical inspection improve Det. F1 to 0.623 and 0.671 by accessing finer visual details. Multimodal agents further improve performance through iterative perception, tool use, verification, and memory, with JarvisEvo achieving the strongest baseline. MuralAgent further raises Det. F1 from 0.742 to 0.813 and Attr. F1 from 0.571 to 0.687. These results show that spatial belief, competing-hypothesis reasoning, and evidence acquisition provide complementary benefits beyond general-purpose agentic reasoning and tool use.

Inspection efficiency. The performance gains are accompanied by lower high-resolution coverage and fewer VLM calls. MuralAgent accesses only 18.6% of the mural at high resolution, compared with 33.4% for JarvisEvo, 35.2% for SAGE, and 67.3% for hierarchical VLM inspection. It also requires only 9.4 inference calls on average, reducing high-resolution coverage and calls by 44.3% and 42.0%, respectively, relative to JarvisEvo. Compared with exhaustive sliding-window inspection, MuralAgent reduces the average number of calls from 48.6 to 9.4 while improving Det. F1 from 0.623 to 0.813. These results show improved detection with less high-resolution coverage and fewer VLM calls on MuralSI-Bench.

4.3 Agentic Inspection Analysis

We further analyze how MuralAgent actively allocates visual observations and follows language instructions during inspection. We focus on two properties central to our formulation: whether the agent can efficiently discover sparse artifacts under a limited inspection budget, and whether its inspection trajectory changes according to the user's intent.

Active inspection efficiency. We compare MuralAgent with three exploration strategies under increasing high-resolution inspection budgets: *Random*, which uniformly samples regions; *Uniform*, which traverses the mural with a regular grid; and *Saliency*, which prioritizes regions according to their initial visual anomaly scores.

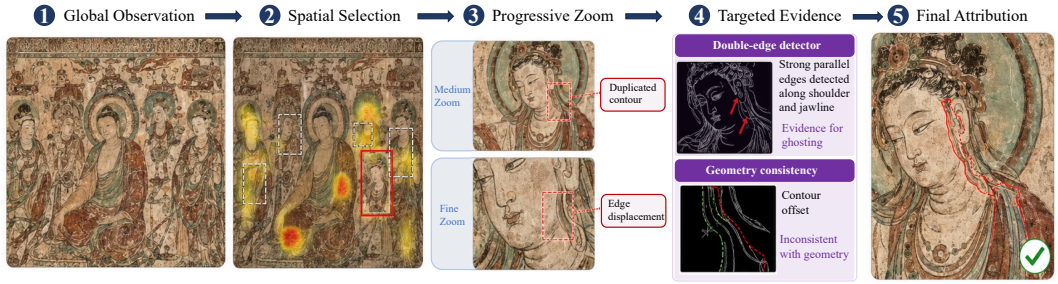

Figure 4: Representative inspection trajectory of MuralAgent. Starting from the instruction *"Inspect human figures for ghosting and geometric misalignment"*, the agent constructs a spatial suspicion map, selects a high-suspicion human figure, and progressively examines it at finer resolutions. Local observations produce competing artifact hypotheses, which are further resolved using double-edge detection and geometry-consistency analysis. The accumulated evidence verifies ghosting along the human-figure contour and supports the ghosting hypothesis.

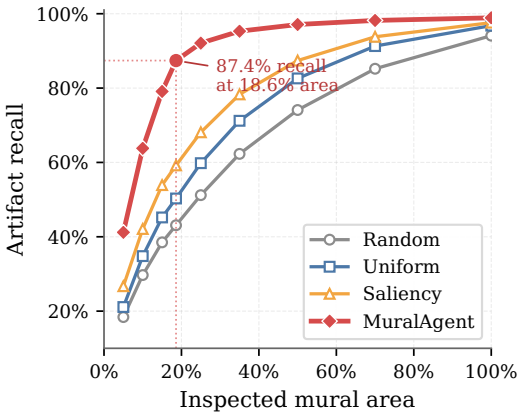

Figure 3: Artifact recall versus high-resolution inspected area.

Figure 3 shows a clear advantage of active inspection, particularly under constrained visual budgets. MuralAgent achieves 63.8% recall with only 10% high-resolution coverage and reaches 87.4% recall at its default operating point of 18.6%. In comparison, Saliency requires 50% coverage to reach the same 87.4% recall, while Uniform remains at 82.6% even at 50% coverage. The gap is particularly pronounced at low budgets; for example, at 15% coverage, MuralAgent achieves 79.1% recall, compared with 53.9%, 45.2%, and 38.5% for Saliency, Uniform, and Random, respectively. As the inspection budget approaches full coverage, the gap gradually narrows, indicating that MuralAgent primarily improves *where limited high-resolution observations are allocated*.

Language-guided inspection. We evaluate whether natural-language instructions effectively control the inspection process using four instruction types: *global*, *artifact-conditioned*, *region-conditioned*, and *compositional*. We report instruction-conditioned artifact recall (**IC-Recall**) over artifacts satisfying the requested constraints and inspection precision (**IP**), defined as the fraction of high-resolution observations allocated to instruction-relevant regions.

Table 2: Language-guided inspection on MuralSI-Bench.

Method	IC-Recall ↑	IP ↑
No Guidance	0.692	0.241
Direct VLM	0.518	–
MIRA	0.734	0.387
IntentEdit	0.721	0.412
GEMS	0.748	0.396
JarvisEvo	0.763	0.435
MuralAgent	**0.857**	**0.682**

As shown in Table 2, MuralAgent achieves an IC-Recall of 0.857, improving over the strongest agent baseline, JarvisEvo (0.763), by 9.4 percentage points. The improvement is more pronounced in inspection precision: MuralAgent reaches 0.682 IP, compared with 0.435 for JarvisEvo and 0.241 without language guidance. This indicates that the instruction not only improves artifact discovery but also substantially changes *where* the agent allocates its high-resolution observations. Removing instruction-specific guidance reduces IC-Recall from 0.857 to 0.692 and IP from 0.682 to 0.241, further confirming the role of language-conditioned spatial belief. More results are provided in Appendix B.4.

Inspection trajectory visualization. Figure 4 presents a representative inspection trajectory to illustrate the closed-loop interaction between spatial exploration, hypothesis refinement, and evidence acquisition. Given the instruction *"Inspect human figures for ghosting and geometric misalign-*

ment", MuralAgent begins with a coarse global observation and constructs a spatial suspicion map over the mural. It then prioritizes a high-suspicion human figure and progressively examines the selected region at medium and fine resolutions. The resulting local observations reveal duplicated contours and edge displacement, yielding competing hypotheses of ghosting, geometric misalignment, and natural deterioration.

Because the local visual evidence alone does not fully distinguish duplicated painting contours from geometric displacement, the remaining uncertainty triggers targeted forensic analysis. The double-edge detector identifies strong parallel boundaries along the shoulder and jawline, while the geometry-consistency tool measures a contour offset beyond the expected tolerance. Incorporating this evidence increases the confidence of the ghosting hypothesis from 0.55 to 0.88, while reducing the probabilities assigned to geometric misalignment and natural deterioration. MuralAgent consequently verifies ghosting along the human-figure contour and attributes the anomaly to registration displacement. This trajectory demonstrates that inspection is not executed as a fixed sequence: each observation updates the current hypotheses, and the residual uncertainty adaptively determines the next visual action or forensic tool call.

4.4 Ablation Studies

We ablate the key components of MuralAgent on MuralSI-Bench while keeping the VLM backbone, tool pool, and inference budget fixed. Results are summarized in Table 3.

Table 3: Ablation study of MuralAgent. "Top-1 Hyp." replaces competing hypotheses with a single hypothesis; "Fixed Routing" uses a predefined tool sequence; and "w/o Source Attr." removes source-level provenance during attribution.

Variant	Det. F1 ↑	AP_{50} ↑	Attr. F1 ↑	Area ↓	Calls ↓
w/o Spatial Belief	0.741	0.628	0.612	31.4	14.8
Top-1 Hyp.	0.762	0.701	0.583	22.7	11.3
Fixed Routing	0.779	0.718	0.641	21.9	13.6
w/o Source Attr.	0.796	0.743	0.492	19.8	10.1
w/o Experience Memory	0.786	0.731	0.524	21.5	12.4
MuralAgent	**0.813**	**0.764**	**0.687**	**18.6**	**9.4**

1) Interleaved visual inspection. Removing spatial belief decreases AP_{50} from 0.764 to 0.628 and increases the inspected area from 18.6% to 31.4%, confirming its importance for efficient localization. Using only the top-1 hypothesis further reduces Det. F1 and Attr. F1 by 5.1 and 10.4 points, respectively, showing the benefit of maintaining competing explanations. **2) Evidence verification and attribution.** Fixed evidence routing reduces Attr. F1 from 0.687 to 0.641 and increases calls from 9.4 to 13.6. Removing source observations causes the largest attribution drop, from 0.687 to 0.492, demonstrating their importance for distinguishing stitching artifacts from intrinsic degradations. **3) Reflective experience memory.** Without experience memory, Det. F1 and Attr. F1 decrease from 0.813/0.687 to 0.786/0.524, respectively, indicating that reusable inspection experience particularly benefits artifact attribution. Further memory analysis is provided in Appendix B.7.

5 Conclusion

We presented MuralAgent, a multimodal agent for large-scale mural stitching inspection. By formulating inspection as uncertainty-driven evidence acquisition, MuralAgent jointly performs spatial prioritization, competing-hypothesis reasoning, adaptive observation, and source-aware verification. Experiments on MuralSI-Bench show improved artifact detection and attribution while substantially reducing visual inspection cost.

AI use statement

In this work, we used generative AI tools to assist with language editing, improving the clarity and organization of the manuscript. The authors take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

Reproducibility statement

We provide detailed descriptions of MuralAgent and its implementation throughout the main paper and appendix. The construction procedure, artifact generation process, and evaluation protocol of MuralSI-Bench are documented in Appendix A.1; model configurations, forensic tools, inference budgets, and baseline adaptations are provided in Appendix A.3; and the exact VLM prompts and deterministic decision rules are provided in Appendix C. The anonymized source code, benchmark construction and evaluation scripts, prompts, and configuration files are included in the supplementary material and are also available at https://anonymous.4open.science/r/MuralAgent-2B9D.

References

Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pp. 671–679. 1987.

Yan Cao, Yi Zhang, Yingjie Lin, and Kaijun Wu. Dunhuang art style transfer via hierarchical vision transformer and color consistency constraints. *IEEE Transactions on Consumer Electronics*, 71 (2):3240–3251, 2025.

Che-Han Chang, Yoichi Sato, and Yung-Yu Chuang. Shape-preserving half-projective warps for image stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3254–3261, 2014.

Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal alignment beyond cosine similarity. *Advances in Neural Information Processing Systems*, 38: 112004–112022, 2026.

Francesco Colace, Rosario Gaeta, Angelo Lorusso, Michele Pellegrino, and Domenico Santaniello. New ai challenges for cultural heritage protection: A general overview. *Journal of Cultural Heritage*, 75:168–193, 2025.

Kaiyi Fang, Zhongyu Min, and Changyu Diao. An sfm system for mural digitization with attention-guided feature matching and robust sparse reconstruction. *npj Heritage Science*, 14(1):166, 2026.

Yuxiang Feng, Juncheng Wang, Chao Xu, Yijie Qian, Huihan Wang, Wenlong Hou, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. Newton: Agentic planning for physically grounded video generation. *arXiv preprint arXiv:2605.18396*, 2026.

Chunchao Guo, Jinpeng Li, Yang Li, and Zilong Huang. Worldclaw: Agentic 3d open-world generation at scale. *arXiv preprint arXiv:2608.05248*, 2026.

Zefeng He, Siyuan Huang, Xiaoye Qu, Yafu Li, Tong Zhu, Yu Cheng, and Yang Yang. Gems: Agent-native multimodal generation with memory and skills. *arXiv preprint arXiv:2603.28088*, 2026.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616. IEEE, 2023.

Zhiying Jiang, Ruhao Yan, Zengxi Zhang, Bowei Zhang, and Jinyuan Liu. Depth-supervised fusion network for seamless-free image stitching. *Advances in Neural Information Processing Systems*, 38:130258–130277, 2026.

Oleh Kolner, Thomas Ortner, Stanisław Woźniak, and Angeliki Pantazi. Mind the gap: Glimpse-based active perception improves generalization and sample efficiency of visual reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 85703–85725, 2025.

Kaican Li, Lewei Yao, Jiannan Wu, Tiezheng Yu, Jierun Chen, Haoli Bai, Lu Hou, Lanqing Hong, Wei Zhang, and Nevin L Zhang. Insight-o3: Empowering multimodal foundation models with generalized visual search. In *The Fourteenth International Conference on Learning Representations*, 2025a.

Shumeng Li, Lei Qi, Qian Yu, Jing Huo, Yinghuan Shi, and Yang Gao. Stitching, fine-tuning, and re-training: a sam-enabled framework for semi-supervised 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(10):3909–3923, 2025b.

Chung-Ching Lin, Sharathchandra U Pankanti, Karthikeyan Natesan Ramamurthy, and Aleksandr Y Aravkin. Adaptive as-natural-as-possible image stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1155–1163, 2015.

Yunlong Lin, Linqing Wang, Kunjie Lin, Zixu Lin, Kaixiong Gong, Wenbo Li, Bin Lin, Zhenxi Li, Shiyi Zhang, Yuyang Peng, et al. Jarvisevo: Towards a self-evolving photo editing agent with synergistic editor-evaluator optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27291–27302, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Zhen Liu, Silu Liu, and Shuo Fan. Research on the virtual restoration of faded dunhuang murals with a global attention mechanism. *npj Heritage Science*, 13(1):35, 2025.

Yuan Mei, Lichun Yang, Mengsi Wang, Tianxiu Yu, and Kaijun Wu. Dunhuangstitch: Unsupervised deep image stitching of dunhuang murals. *IEEE transactions on visualization and computer graphics*, 31(8):4226–4240, 2024.

Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE transactions on image processing*, 30:6184–6197, 2021.

Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Parallax-tolerant unsupervised deep image stitching. In *IEEE/CVF International Conference on Computer Vision*, pp. 7365–7374, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Kevin Qinghong Lin, Batu EI, Yuhong Shi, Pan Lu, Philip Torr, and James Zou. Data journalist agent: Transforming data into verifiable multimodal stories. *arXiv e-prints*, pp. arXiv–2606, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763, 2021.

Hui Ren, Ke Sun, Fanhua Zhao, and Xian Zhu. Dunhuang murals image restoration method based on generative adversarial network. *Heritage Science*, 12(1):39, 2024.

Jiaquan Shen, Ningzhong Liu, Han Sun, Deguang Li, Yongxin Zhang, and Lulu Han. An algorithm based on lightweight semantic features for ancient mural element object detection. *npj Heritage Science*, 13(1):70, 2025.

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.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8918–8927. IEEE, 2021.

Fubo Wang, Mingcong Dang, Zeyu Jia, Wanyi Zhao, Fuxiang Ma, and Shengling Geng. A structure-aware diffusion framework for generalizable mural restoration. *npj Heritage Science*, 2026.

Zhaobin Wang and Zekun Yang. Review on image-stitching techniques. *Multimedia Systems*, 26 (4):413–430, 2020.

Zhaokai Wang, Xizhou Zhu, Xue Yang, Gen Luo, Hao Li, Changyao Tian, Wenhan Dou, Junqi Ge, Lewei Lu, Yu Qiao, et al. Parameter-inverted image pyramid networks for visual perception and multimodal understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

Siwei Wen, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Conghui He, Weijia Li, et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *Advances in Neural Information Processing Systems*, 38:58972–59005, 2026.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, et al. Sage: Scalable agentic 3d scene generation for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22358–22368, 2026.

Zhipei Xu, Xuanyu Zhang, Youmin Xu, Qing Huang, Shen Chen, Taiping Yao, Shouhong Ding, and Jian Zhang. Genshield: Unified detection and artifact correction for ai-generated images. *arXiv preprint arXiv:2605.16122*, 2026.

Zishan Xu, Yuqing Yang, Qianzhen Fang, Wei Chen, Tingting Xu, Jueting Liu, and Zehua Wang. A comprehensive dataset for digital restoration of dunhuang murals. *Scientific Data*, 11(1):955, 2024.

Zhilin Yang, Yong Yin, Haitong Xu, Qianfeng Jing, Zhiying Jiang, Tianli Liao, and Carlos Guedes Soares. Advancements of image and video stitching techniques: A review. *IEEE Sensors Journal*, 25(13):23526–23551, 2025.

Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2339–2346, 2013.

Ziyun Zeng, Hang Hua, and Jiebo Luo. Mira: Multimodal iterative reasoning agent for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9563–9573, 2026.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, et al. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *The Fourteenth International Conference on Learning Representations*, 2026a.

Yuxuan Zhang, Shijia Huang, and Liwei Wang. Intentedit: Multi-agent reasoning for intent-driven complex image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8776–8785, 2026b.

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.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, volume 2024, pp. 49705–49737, 2024.

Xizhou Zhu, Xue Yang, Zhaokai Wang, Hao Li, Wenhan Dou, Junqi Ge, Lewei Lu, Yu Qiao, and Jifeng Dai. Parameter-inverted image pyramid networks. *Advances in Neural Information Processing Systems*, 37:132267–132288, 2024.

A Additional Experimental Details

A.1 Dataset and Evaluation Protocol

MuralSI-Bench. MuralSI-Bench is constructed from MuralDH (Xu et al., 2024). We split the benchmark at the *source-mural level*, ensuring that all source tiles and artifact variants derived from the same mural belong to only one split. This prevents visual-content leakage between development and test sets.

For each source mural, we construct overlapping source tiles and their corresponding stitched reconstruction.MuralSI-Bench includes artifacts observed in mural reconstructions and controlled artifacts induced by perturbing stages of the stitching pipeline. These controlled cases are generated only by perturbing the corresponding stages of the stitching pipeline, including registration, geometric warping, blending, photometric consistency, and source composition, rather than directly corrupting the final stitched image. Together, the benchmark covers six representative artifact categories: geometric misalignment, ghosting, visible seam, appearance inconsistency, content duplication, and content missing. Intrinsic mural degradations, including cracks, fading, and material loss, are preserved throughout construction.

Each benchmark instance contains the stitched mural, corresponding source tiles, artifact annotations, source provenance, and language-guided inspection instructions. The annotations include artifact category, bounding region, severity, and attribution label. Table 4 summarizes the benchmark statistics.

Table 4: Statistics of MuralSI-Bench.

Statistic	Number
Source murals	512
Stitched images	4,256
Clean reconstructions	387
Artifact instances	11,843
Misalignment	2,041
Ghosting	1,967
Visible seam	2,013
Appearance inconsistency	1,988
Content duplication	1,902
Content missing	1,932
Intrinsic degradation regions	7,694
Language instructions	4,256

Language-guided evaluation. We construct four forms of inspection instructions: global, artifact-conditioned, region-conditioned, and compositional. Global instructions request inspection of the complete mural; artifact-conditioned instructions specify one or more artifact categories; region-conditioned instructions specify semantic mural regions; and compositional instructions jointly constrain both region and artifact type. Predictions outside the requested scope are not counted as true positives, allowing the benchmark to measure whether the inspection trajectory follows the language request.

Metrics. For image-level artifact detection, a prediction is considered correct when the predicted artifact category matches a ground-truth category. We report precision, recall, and F1, using F1 as the primary metric. For localization, predicted and ground-truth regions are matched by intersection-over-union (IoU), and we report AP_{50}. Type F1 is computed over matched artifact regions. For attribution, each localized anomaly is classified as either `stitching` or `intrinsic`, and we report macro F1 over the two classes.

Inspection efficiency is measured using

$$\mathrm{Area} = \frac{\left|\bigcup_t R_t^{\mathrm{HR}}\right|}{|I|},$$

where R_t^{HR} denotes regions accessed at high resolution. We also report the total numbers of VLM calls and external forensic-tool calls per mural. Repeated observations of an already inspected high-resolution region are counted once for Area but remain included in the call statistics.

A.2 Adaptation of Multimodal Agent Baselines

Directly applying existing multimodal agents is infeasible because they were originally developed for different tasks such as image editing, 3D generation, or data analysis. We therefore preserve the defining reasoning mechanism of each method while replacing its task-specific environment with our mural inspection environment. All adapted agents receive identical observations, tool APIs, and inference budgets.

- MIRA agent (Zeng et al., 2026). We adapt MIRA's iterative multimodal reasoning paradigm into an observation–reasoning–action loop. At each step, the agent observes the current mural crop and predicts one atomic inspection action before receiving the next visual observation. No explicit spatial belief or multi-hypothesis state is maintained.
- IntentEdit agent (Zhang et al., 2026b). We instantiate a planner–executor–evaluator architecture. The planner decomposes the inspection instruction into staged inspection objectives, the executor performs visual and tool actions, and the reflective evaluator determines whether a stage should be accepted, refined, or replanned.
- SAGE agent (Xia et al., 2026). We retain SAGE's multi-critic iterative refinement strategy. Semantic, geometric, and appearance critics independently evaluate suspicious regions, and their feedback is provided to the agent for subsequent inspection and tool selection.
- GEMS agent (He et al., 2026). We implement an agent loop augmented with trajectory memory and reusable inspection skills. Previous successful and failed inspection trajectories are summarized into memory, while task-relevant skills are retrieved before subsequent decisions.
- Data2Story agent (Qinghong Lin et al., 2026). We adapt its evidence-grounded inspector mechanism. The agent first produces candidate findings, after which an inspector invokes available tools to verify supporting evidence and returns failed or unsupported findings for revision.
- JarvisEvo agent (Lin et al., 2026). We adapt its interleaved multimodal reasoning and reflection strategy. Visual observations and verification results are repeatedly incorporated into subsequent reasoning, and previous inspection outcomes are retained as experience for later decisions.
- Additional agent baselines. We additionally evaluate WorldClaw (Guo et al., 2026), NEWTON (Feng et al., 2026), and Auto-Slides (Guo et al., 2026). WorldClaw is adapted as a hierarchical coarse-to-fine inspection agent; NEWTON is instantiated as a planner that orchestrates inspection tools followed by verifier-driven replanning; and Auto-Slides is adapted as a multi-stage generation–verification loop in which inspection findings are iteratively reviewed and revised.
- Direct VLM input protocol. Direct VLM receives the complete high-resolution mural in a single inference call, without agent-controlled region selection. Its high-resolution coverage is therefore 100%. This coverage measures spatial access rather than inference cost. Backbone-specific image preprocessing and input-resolution configurations are provided in the supplementary code.

Controlled comparison. All adapted agents use the same underlying VLM, mural pyramid, crop interface, and forensic tool pool. They are given the same maximum VLM-call and tool-call budgets. We preserve only the reasoning, memory, planning, reflection, and tool-control mechanisms that characterize each framework. This protocol isolates differences in agentic inspection strategy from differences in perception models or external tools.

A.3 Implementation Details

VLM backbone. We use **Qwen2.5-VL-72B-Instruct** as the default multimodal backbone for MuralAgent and all agent-based baselines. The input configuration, decoding parameters, temperature,

and maximum output length are kept identical across methods. We additionally evaluate **GPT-4o** and **InternVL2.5-78B** to examine backbone generalization.

Multi-scale observation. Each mural is represented by a four-level image pyramid at $\{1/8, 1/4, 1/2, 1\}$ of the original resolution. Inspection starts from the coarsest level for global context and progressively accesses higher-resolution regions through visual actions. Each local observation corresponds to a 512×512 crop at the selected pyramid level and is directly provided to the VLM without additional spatial downsampling. Neighboring regions use a 25% overlap when contextual expansion is requested.

Forensic tools. All agent-based methods share the same training-free forensic tool pool, covering semantic, geometric, appearance, and source evidence. Semantic grounding is performed using **Grounding DINO** (Liu et al., 2023). Geometric consistency is measured using **LoFTR** (Sun et al., 2021) correspondences followed by RANSAC-based homography estimation; correspondence consistency and reprojection residuals are used as geometric evidence. Contour continuity is additionally measured from Canny edges across candidate regions. For appearance evidence, we compute CIE Lab color difference, Sobel gradient discontinuity, local texture similarity, and Laplacian-variance sharpness. Source evidence is obtained by retrieving candidate source observations overlapping the queried region using the original stitching inputs and estimated geometric correspondence. Importantly, benchmark annotations, injected artifact locations, perturbation parameters, and ground-truth stitching boundaries are used only for evaluation and are never exposed to the agent. The complete tool pool is summarized in Table 5.

Table 5: Forensic tools shared by all agent-based methods.

Family	Implementation	Evidence
Semantic	Grounding DINO	Region grounding
Geometry	LoFTR	Feature correspondence
Geometry	RANSAC Homography	Registration residual
Geometry	Canny	Contour continuity
Appearance	CIE Lab + Sobel	Color/gradient consistency
Appearance	Texture + Laplacian	Texture/sharpness consistency
Source	Source-tile retrieval	Source observation

Action selection and stopping. Visual actions are selected from the current spatial and hypothesis states following the state-conditioned rules described in Sec. 3.2. MuralAgent first prioritizes unresolved regions according to $U_t(R)$ and then selects `ZOOM`, `PAN`, `EXPAND`, or `COMPARE` according to the required observation scale and context. When visual observations remain insufficient to resolve the current hypotheses, `VERIFY` triggers the hypothesis-conditioned forensic tools defined in Sec. 3.3. No learned or VLM-predicted utility score is used for action or tool selection. Inspection of the current region terminates when the hypothesis is verified or rejected by the acquired evidence, while the complete episode terminates when no unresolved high-priority region remains, the agent issues `STOP`, or the inference budget is reached.

Experience memory. The factual, trajectory, and experience memories are constructed exclusively from completed inspection episodes in the development split. Each trajectory is compressed into a compact experience containing the encountered ambiguity, informative observations, effective evidence sources, final attribution, and reflection. The memory is frozen during evaluation to prevent information leakage from test murals.

Reproducibility. All methods are evaluated using identical dataset splits and inspection environments. For stochastic VLM inference, we repeat each experiment 5 times and report the mean. Prompts, benchmark generation code, tool interfaces, baseline adaptations, and evaluation scripts are also available at https://anonymous.4open.science/r/MuralAgent-2B9D.

Table 6: Artifact-wise F1 scores on MuralSI-Bench.

Method	Misalign.	Ghosting	Seam	Appearance	Duplication	Missing	Avg.
Direct VLM	0.541	0.428	0.563	0.479	0.412	0.397	0.470
Sliding-window VLM	0.682	0.571	0.694	0.613	0.548	0.531	0.607
MIRA	0.763	0.672	0.781	0.698	0.641	0.629	0.697
SAGE	0.781	0.694	0.798	0.715	0.658	0.647	0.716
GEMS	0.774	0.683	0.791	0.708	0.652	0.639	0.708
JarvisEvo	0.798	0.711	0.812	0.734	0.679	0.668	0.734
MuralAgent	**0.864**	**0.791**	**0.873**	**0.802**	**0.768**	**0.751**	**0.808**

Table 7: Detection F1 across artifact severity levels.

Method	Easy	Medium	Hard	Avg.
Direct VLM	0.612	0.478	0.341	0.477
Sliding-window VLM	0.741	0.618	0.473	0.611
Hierarchical VLM	0.783	0.667	0.528	0.659
MIRA	0.821	0.714	0.582	0.706
SAGE	0.836	0.732	0.601	0.723
JarvisEvo	0.854	0.751	0.624	0.743
MuralAgent	**0.907**	**0.821**	**0.718**	**0.815**

B Additional Main Results

B.1 Results across Artifact Categories

Table 6 reports detection performance for the six stitching artifact categories in MuralSI-Bench. This evaluation characterizes whether improvements are concentrated on a particular artifact or generalize across geometric, appearance, and content-level stitching failures.

Geometric misalignment and visible seams can often be identified from strong local discontinuities, whereas subtle ghosting and content-level inconsistencies require additional contextual or source evidence. MuralAgent shows consistent improvements across these categories, with larger gains on artifacts that require multi-scale inspection or cross-source verification.

B.2 Results across Artifact Severity

We further evaluate robustness under the three artifact severity levels defined in MuralSI-Bench. Table 7 reports detection F1 for easy, medium, and hard subsets.

The performance gap becomes larger as artifact severity decreases. Hard examples often contain weak local visual cues and stronger ambiguity with intrinsic mural structures, making them particularly dependent on active high-resolution observation and complementary forensic evidence.

B.3 Extended Multimodal Agent Comparison

Table 8 extends the agent comparison to additional multimodal frameworks. These methods are adapted according to the controlled protocol described in Appendix A.2.

The additional results show that hierarchical exploration, verifier-driven replanning, and multi-stage review are individually useful for mural inspection. MuralAgent combines these general agentic capabilities with an inspection-specific state and evidence-seeking strategy, yielding stronger artifact localization and attribution under the same inference budget.

B.4 Analysis by Instruction Type

Table 9 reports language-guided inspection performance separately for the four instruction types in MuralSI-Bench.

Table 8: Extended comparison with multimodal agent frameworks.

Method	Det. F1	AP_{50}	Attr. F1	Area ↓	Calls ↓
WorldClaw	0.728	0.661	0.552	37.2	18.9
NEWTON	0.751	0.693	0.579	34.6	17.4
Auto-Slides	0.736	0.674	0.561	36.1	19.2
MuralAgent	**0.813**	**0.764**	**0.687**	**18.6**	**9.4**

Table 9: Results across different language instruction types.

Instruction	IC-Recall ↑	IP ↑	Area ↓
Global	0.891	1	21.4
Artifact-conditioned	0.864	0.583	17.9
Region-conditioned	0.842	0.721	14.6
Compositional	0.813	0.768	12.3

Artifact-conditioned instructions primarily alter which anomaly hypotheses are prioritized, whereas region-conditioned instructions directly modify spatial exploration. Compositional instructions require both behaviors simultaneously and therefore provide the strongest test of instruction-conditioned inspection.

B.5 Effect of Inspection Strategy

We further isolate the region-selection mechanism by replacing the inspection policy of MuralAgent with alternative exploration strategies while keeping the same VLM, hypothesis reasoning, and verification modules.

The Spatial Belief variant selects regions using the instruction-conditioned spatial belief but removes hypothesis-uncertainty-based action selection. Comparing it with the full MuralAgent isolates the contribution of uncertainty-guided information acquisition beyond spatial anomaly prioritization.

B.6 Backbone Generalization

To examine whether the inspection gains depend on a particular multimodal backbone, we instantiate Direct VLM and MuralAgent with three VLM families while keeping the inspection environment unchanged. As demonstrated in Figure 11, performance improvements are evident across all three evaluated backbones.

B.7 Experience Memory Analysis

We further examine whether the gain from experience memory arises from retrieving semantically useful inspection knowledge or simply from providing additional context to the VLM. All memory variants use the same number of entries, prompt format, and approximately matched token budget.

Random Memory retrieves entries uniformly from the memory bank, while *Irrelevant Memory* retrieves entries that are dissimilar to the current visual state and artifact hypotheses. *Relevant Memory* uses the retrieval mechanism of MuralAgent, and *Oracle Memory* retrieves experiences using the ground-truth artifact category solely to estimate an upper bound. The comparison between Relevant and Random/Irrelevant Memory isolates the semantic utility of retrieved experience from the effect of additional context.

Memory size scaling. We further vary the size of the available experience memory while keeping the number of retrieved entries fixed to three for every query. This ensures that the inference-time context length remains approximately unchanged as the memory bank grows.

Performance improves as the memory bank contains more inspection experience and gradually saturates with increasing memory size. Since every setting retrieves the same number of entries with a

Table 10: Comparison of visual exploration strategies under the same high-resolution inspection budget.

Strategy	Det. F1 ↑	Recall ↑	Area ↓
Random	0.521	0.431	18.6
Uniform Grid	0.583	0.503	18.6
Visual Saliency	0.647	0.592	18.6
Spatial Belief	0.742	0.781	18.6
MuralAgent	**0.813**	**0.874**	**18.6**

Table 11: Generalization across multimodal backbones. Δ denotes the improvement of MuralAgent over direct inspection using the same backbone.

Backbone	Method	Det. F1	Attr. F1	Area
Qwen2.5-VL-72B	Direct VLM	0.487	0.298	100
	MuralAgent	0.813	0.687	18.6
GPT-4o	Direct VLM	0.512	0.321	100
	MuralAgent	0.829	0.704	17.9
InternVL2.5-78B	Direct VLM	0.469	0.284	100
	MuralAgent	0.796	0.671	19.4

matched context budget, the improvement cannot be attributed to providing the VLM with a longer prompt. Instead, a larger memory bank increases the probability of retrieving experience that is relevant to the current visual ambiguity and verification process.

C Prompts and Decision Rules

MuralAgent is training-free and does not introduce additional learned modules beyond the underlying VLM and off-the-shelf forensic tools. The operators used in the main text denote either structured VLM prompting or deterministic state updates. Table 14 summarizes their implementations.

C.1 Structured VLM Prompts

Inspection specification (Φ_{ins}). The VLM receives the user instruction and the predefined artifact taxonomy and returns:

```
Target regions:
Target artifact types:
Inspection constraints:
```

Hypothesis reasoning (Φ_{reason}). Given the instruction, current multi-scale observations, and previous hypotheses, the VLM is prompted as:

```
You are inspecting a stitched mural image.
Given the global context, local observation, and
previous hypotheses, identify plausible explanations
for the observed pattern.

Consider both:
(1) stitching artifacts, and
(2) intrinsic mural structures or degradation.

For each hypothesis, return:
- hypothesis
- supporting visual observations
- contradicting observations
```

Table 12: Effect of experience retrieval. All variants use matched memory size and context budget.

Memory	Det. F1 ↑	Attr. F1 ↑	Calls ↓
No Memory	0.786	0.524	12.4
Random Memory	0.791	0.541	12.1
Irrelevant Memory	0.783	0.518	12.6
Relevant Memory	**0.813**	**0.687**	9.4
Oracle Memory	0.841	0.729	8.7

Table 13: Effect of experience-memory size. The retrieval size is fixed to three entries across all settings.

Memory Size	Det. F1 ↑	Attr. F1 ↑	Calls ↓
0	0.786	0.524	12.4
5	0.794	0.561	11.7
10	0.801	0.608	10.9
20	0.808	0.649	10.1
50	0.811	0.674	9.7
Full	0.813	0.687	9.4

```
Retain at most K hypotheses.
```

Evidence verification (Φ_{verify}). The verifier receives the current hypotheses together with the structured measurements returned by the forensic tools:

```
For each hypothesis, evaluate the available forensic
evidence.

Return one of:
SUPPORTING / CONTRADICTING / INCONCLUSIVE

Use only the provided measurements and observations.
Do not infer unavailable stitching metadata.

Return:
- hypothesis
- evidence status
- short justification
```

Source-aware attribution (Φ_{attr}). The attribution prompt compares the verified observation with retrieved source observations:

```
Determine whether the verified abnormality is:
STITCHING / INTRINSIC / UNCERTAIN.

STITCHING:
the abnormality is absent from the corresponding
source observations and emerges after reconstruction.

INTRINSIC:
the same abnormality is consistently present in
the source observations.

UNCERTAIN:
the available evidence is insufficient or conflicting.

Return the attribution and supporting evidence.
```

Table 14: Implementation of the operators used in MuralAgent.

Operator	Implementation	Output
Φ_{ins}	VLM prompt	Inspection specification
Φ_{reason}	VLM prompt	Competing hypotheses
Φ_{verify}	VLM prompt	Evidence-supported hypotheses
Φ_{attr}	VLM prompt	Artifact attribution
Φ_{mem}	VLM prompt	Experience summary
Spatial belief update	Deterministic	Region priority
Action selection	Rule-based	Visual action
Evidence routing	Rule-based	Forensic tools

Experience summarization (Φ_{mem}). After a completed inspection trajectory, the VLM summarizes reusable experience:

```
Summarize the completed inspection trajectory into
reusable inspection knowledge.

Include only:
- visual ambiguity encountered
- useful inspection actions
- informative forensic evidence
- failed or misleading cues
- final verification strategy

Do not store mural identity, artifact location,
ground-truth labels, or benchmark generation metadata.
```

C.2 Deterministic Decision Rules

Spatial belief update. For each candidate region, we compute the normalized visual, language, uncertainty, and coverage terms defined in Sec. 3.2. Their parameter-free average determines the region priority $U_t(R)$. After each observation, only affected regions are recomputed and the corresponding entries in M_t^{sus} and M_t^{cov} are updated.

Visual action selection. After selecting the highest-priority unresolved region, the following rules determine the visual action:

- `ZOOM`: finer spatial detail is required;
- `EXPAND`: surrounding context is required;
- `PAN`: another neighboring region has higher priority;
- `COMPARE`: the current hypothesis requires cross-region or cross-scale comparison;
- `BACKTRACK`: new evidence contradicts the current inspection branch;
- `VERIFY`: visual observations remain ambiguous and forensic evidence is required.

Hypothesis-conditioned evidence routing. The evidence families are selected according to the current artifact hypotheses:

Hypothesis	Evidence
Misalignment	Geometry + Source
Ghosting	Geometry + Source
Visible seam	Appearance + Source
Appearance inconsistency	Appearance + Source
Content duplication	Semantic + Source
Content missing	Semantic + Source

For multiple active hypotheses, we take the union of the required evidence families and skip evidence already acquired for the current region.