MASTERSTUDYBENCH: CAN A MODEL PAINT A PAINTING, OR ONLY REPLAY ITS PIXELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Genuine visual intelligence transcends passive pixel imitation: true mastery requires internalizing the generative hierarchy of visual creation. In classical art, an apprentice ascends through *master studies*, deconstructing a masterpiece's procedural logic rather than mechanically tracing its surface. Yet current visual code synthesis remains trapped in a pixel-only regime: grading solely on rendered appearance treats code as an uninspected black box, inviting reward hacking while remaining blind to painterly aesthetics and structural modularity alike. To bridge this chasm, we introduce MASTERSTUDYBENCH, evaluating procedural artistry across 200 museum masterpieces inside an offline hardened sandbox. MASTERSTUDYBENCH decouples two evaluation pillars: (1) an image-conditioned rubric tree of 17,323 criteria auditing fine-grained aesthetics across brushwork, lighting, and composition; and (2) an *executed editability probe* measuring re-rendered spatial deltas under counterfactual source mutations to test program modularity. Across frontier multimodal models, a profound disconnect emerges: an adversarial stub attaining a paired CLIP similarity of 0.674 collapses to 0.0 under MASTERSTUDYBENCH, while reconstruction fidelity correlates only weakly with executed editability (Spearman's $\rho = 0.26$). Across the evaluated model configurations, painterly aesthetics trails basic visual integrity by 7.1–16.4 percentage points. This gap and the weak fidelity–editability association identify painterly rendering and modular program construction as distinct challenges. By separating *painting* from *replaying*, MASTERSTUDYBENCH charts the path from superficial mimicry toward authentic procedural intelligence.

1 INTRODUCTION

A drawing program is not merely a static image, but an executable hypothesis of visual construction, factoring a canvas into procedural strokes, compositional layers, and editable parameters (Lake et al., 2015; Ellis et al., 2018; Tian et al., 2019; Hertzmann, 1998). Much like classical *master studies* (Cennini, 1954; Gombrich, 1960), where apprentices internalize generative principles through "making and matching" rather than mechanically duplicating surface pixels, the ultimate ambition of multimodal image-to-code synthesis (Beltramelli, 2018; Si et al., 2025; Yang et al., 2025a; Rodriguez et al., 2025; Gupta & Kembhavi, 2023; Surís et al., 2023) is cultivating an AI capable of authentic programmatic visual creation.

However, current evaluation paradigms treat code as an uninspected black box: grading solely on rendered pixel or perceptual similarity (Wang et al., 2004; Zhang et al., 2018; Radford et al., 2021), they remain blind to both painterly aesthetics and underlying program structure. This invites severe reward hacking (Amodei et al., 2016; Skalse et al., 2022): in our adversarial experiments, a naive stub scattering 300 random colored squares achieves a paired CLIP similarity of 0.674 by merely matching color distributions, despite encoding zero structural logic. The implication is decisive: *looking right* is fundamentally distinct from *being built right*. Surface similarity measures only the appearance of a canvas, leaving genuine programmatic decomposition entirely unverified.

Evaluating authentic painterly mastery, however, confronts two fundamental dilemmas. First, painterly nuances, such as impasto textures, subtle illumination, and stylistic brushwork, elude conventional global metrics. Second, fine art admits no unique ground-truth source program or privileged decomposition tree (Sharma et al., 2018), rendering static reference-matching impossi-

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.

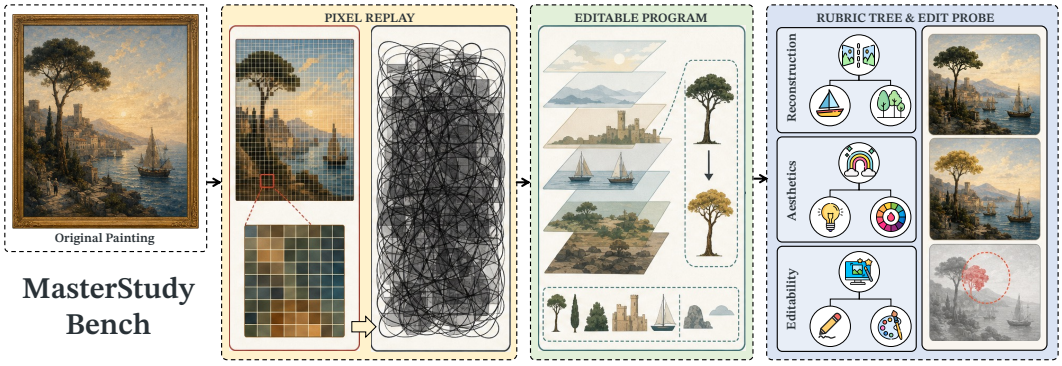


Figure 1: Overview of MASTERSTUDYBENCH. Given a museum masterpiece x, a multimodal agent synthesizes a procedural program inside an offline hardened sandbox.

ble. Consequently, program quality cannot be judged by surface syntax, but by how cleanly visual components can be independently addressed, localized, and recomposed under dynamic execution.

To address these twin challenges, we introduce MASTERSTUDYBENCH, a benchmark evaluating painting generation across two decoupled pillars: visual aesthetic quality and executable program quality (Figure 1). Under an adversarially hardened contract, models must paint autonomously inside an offline sandbox that strictly prohibits raster assets, SVG imports, or network access, ensuring self-contained procedural rendering. Our first pillar, *visual quality*, pioneers fine-grained aesthetic evaluation for code generation, deploying an image-conditioned rubric tree of 17,323 criteria that assess painterly nuances, such as brushwork texture, illumination, and color harmony, alongside structural fidelity (Zheng et al., 2023; Hu et al., 2023; Cho et al., 2024; Ku et al., 2024). Our second pillar, *program editability*, introduces counterfactual executed editability probes that apply targeted mutations to source code, re-render the canvas in the sandbox, and measure the spatial locality of changes to verify whether visual primitives are cleanly factored or irreversibly entangled (Chen et al., 2021; Hendrycks et al., 2021; Jimenez et al., 2024).

We conducted extensive evaluations across frontier multimodal models, revealing a profound divergence between surface visual mimicry and programmatic mastery. Remarkably, rendered reconstruction fidelity correlates only weakly with executed editability (Spearman's $\rho = 0.26$), whereas it correlates strongly with aesthetic expressiveness ($\rho = 0.77$), demonstrating that visual fidelity and program structure respond to entirely different pressures. For instance, a submission with a reconstruction score of 0.879 exhibits lower editability (0.715) than a submission with a reconstruction score of 0.696 and editability of 0.943. These findings demonstrate that surface-level visual similarity leaves generative competence unmeasured, establishing execution-grounded program probes as an essential evaluation axis.

In summary, our contributions are threefold:

- We introduce MASTERSTUDYBENCH, an evaluation benchmark spanning 200 curated masterpieces from four premier museum archives and 17,323 fine-grained rubric criteria, decoupling multimodal painting generation into dual tracks of visual artistry and executable code quality.
- We introduce counterfactual executed editability probes inside an offline fail-closed sandbox, quantifying program modularity through localized spatial differences under targeted execution perturbations.
- We show that a stub attaining 0.674 paired CLIP similarity collapses to zero under MASTERSTUDYBENCH, and report a weak correlation ($\rho = 0.26$) between reconstruction and editability.

2 RELATED WORK

Image-to-Code Synthesis. Recent advancements in multimodal foundation models have accelerated image-to-code synthesis, expanding from web frontends to statistical charts and scientific visualization programs (Si et al., 2025; Yang et al., 2025a; Wu et al., 2024; Lu et al., 2025; Yun et al., 2024; Laurençon et al., 2024; Li et al., 2024). Concurrently, vector graphics synthesis has

Table 1: Comparison of MASTERSTUDYBENCH with representative multimodal visual and code generation benchmarks. ✓ and ✗ denote supported and unsupported features, respectively.

Benchmark	Artistic Domain		Visual Artistry		Program Editability		
	Museum Art	Exec. Code	Painterly Aesthetics	Rubric Tree	Hard Sandbox	Exec. Probe	Decouplable Layers
Design2Code (Si et al., 2025)	✗	✓	✗	✗	✗	✗	✗
Web2Code (Yun et al., 2024)	✗	✓	✗	✗	✗	✗	✗
WebGen-Bench (Lu et al., 2025)	✗	✓	✗	✗	✗	✗	✗
Plot2Code (Wu et al., 2024)	✗	✓	✗	✗	✗	✗	✗
ChartMimic (Yang et al., 2025a)	✗	✓	✗	✗	✗	✗	✗
StarVector (Rodriguez et al., 2025)	✗	✓	✗	✗	✗	✗	✗
VGBench (Zou et al., 2024)	✗	✓	✗	✗	✗	✗	✗
ArtBench (Liao et al., 2022)	✓	✗	✗	✗	✗	✗	✗
MASTERSTUDYBENCH (Ours)	✓	✓	✓	✓	✓	✓	✓

advanced through SVG generation models, vector graphics benchmarks, and layer-wise image vectorization (Rodriguez et al., 2025; Zou et al., 2024; Yang et al., 2025c; Ma et al., 2022). Recent image-to-code work has also examined whether models can recover editable vector representations from rendered images (Rodriguez et al., 2025; Zou et al., 2024). The dominant evaluation regimes each anchor on a different reference: web-generation benchmarks match against standardized DOM trees, chart benchmarks against canonical plotting specifications, and vector synthesis models optimize directly against rendered raster losses.

Multi-modal Agent Evaluation. Multi-modal agents have evolved from single-turn solutions to multi-turn tool orchestration and repository-scale software engineering in multimodal environments (Chen et al., 2024; Yang et al., 2025b; Jain et al., 2025; Cheng et al., 2025; Wang et al., 2025; Hong et al., 2024; Shinn et al., 2023). Navigating multimodal operating systems, browser environments, and interactive software sandboxes, modern agents can use visual feedback to refine generated interfaces (Xie et al., 2024; Koh et al., 2024; Zheng et al., 2025; Wan et al., 2024; Xiao et al., 2024; Zhang et al., 2025). In parallel, visual evaluation has progressed beyond distributional and perceptual metrics (Heusel et al., 2017; Zhang et al., 2018) and learned human-preference reward models (Xu et al., 2023; Wu et al., 2023) toward fine-grained question decomposition and explainable conditional-image scoring (Cho et al., 2024; Ku et al., 2024). Across all of these modern approaches, the generated program itself remains an uninspected black box: grading operates on the static rendering, which is what leaves the evaluation acutely exposed to catastrophic reward hacking (Amodei et al., 2016; Skalse et al., 2022).

Existing benchmarks conflate surface visual fidelity with program quality, unable to discern whether a convincing canvas reflects genuine compositional discipline or brittle pixel mimicry. Building upon these, we introduce MASTERSTUDYBENCH, the first benchmark to simultaneously unify museum-scale fine art, executable canvas programs, fine-grained painterly aesthetic rubrics, and execution-grounded counterfactual modularity probes (Table 1).

3 MASTERSTUDYBENCH

3.1 PROBLEM FORMALIZATION

We formalize painting-to-code reconstruction as a constrained program synthesis and execution-grounded dual-track verification problem. Each evaluation instance is parameterized by a task tuple:

$$\mathcal{T} = (x, \mathcal{M}, \mathcal{C}), \tag{1}$$

where $x \in \mathcal{X} \subset \mathbb{R}^{H \times W \times 3}$ denotes the high-resolution RGB reference painting, $\mathcal{M}$ represents verified institutional archival metadata (e.g., artist, creation epoch, medium, and provenance), and $\mathcal{C} = (W_c, H_c, \text{dpi})$ specifies target canvas geometry and rendering constraints.

Given visual reference x and metadata $\mathcal{M}$, the multimodal agent models a synthesis policy $\pi(p \mid x, \mathcal{M}, \mathcal{C})$ to generate an executable procedural program $p \in \mathcal{L}$. A deterministic execution engine $\mathcal{E} : \mathcal{L} \to \mathcal{I}$ renders the synthesized program onto a canvas surface:

$$I_p = \mathcal{E}(p) \in \mathbb{R}^{H_c \times W_c \times 3}. \tag{2}$$

The evaluation runtime produces a comprehensive four-dimensional capability profile:

$$\mathbf{s}(p) = \big(S_{\text{basic}},\, S_{\text{recon}},\, S_{\text{aesth}},\, S_{\text{edit}}\big) \in [0,1]^4, \tag{3}$$

where S_{basic} captures foundational rendering integrity, S_{recon} quantifies image-conditioned reconstruction criteria, S_{aesth} represents our dedicated painterly aesthetic score, and S_{edit} measures program editability.

3.2 BENCHMARK CONSTRUCTION PIPELINE

The creation of MASTERSTUDYBENCH follows a systematic four-stage engineering pipeline:

Stage 1: Masterpiece Collection. Artworks in MASTERSTUDYBENCH are curated across four world-renowned public museum archives: The Cleveland Museum of Art (CMA), The Metropolitan Museum of Art (The Met), Art Institute of Chicago (AIC), and Museum of New Zealand Te Papa Tongarewa (Te Papa). All reference images satisfy a minimum canvas resolution threshold ($W_c \geq 512, H_c \geq 512$) with institutional archival metadata standardized into verified tuple fields, whose complete distribution across visual genres, media, traditions, and epochs is catalogued in Appendix A.1 (Table 3).

Stage 2: Rubric Tree Construction. For each curated masterpiece, we construct a custom image-conditioned hierarchical evaluation tree tailored to its specific visual composition. Each tree recursively partitions the artwork's global composition into fine-grained visual sub-concepts, with every terminal leaf drawn from an 18-element closed artistic ontology (Appendix A.2, Table 4). Within visual evaluation, we score foundational integrity, reconstruction, and painterly aesthetics as separate rubric dimensions:

- **Basic Visual Integrity Criteria (yielding S_{basic}):** Foundational sanity criteria verifying non-empty canvas bounds, uniform background washes, and the absence of catastrophic rendering crashes or clipping anomalies.
- **Reconstruction Fidelity (yielding S_{recon}):** Image-conditioned criteria assessing scene entities, spatial relationships, expressive details, and chromatic correspondence with the reference.
- **Painterly Aesthetic Criteria (yielding dedicated score S_{aesth}):** Our core aesthetic contribution, uniquely isolating and auditing procedural brushwork texture (e.g., impasto layering, dry-brush strokes, ink wash bleeding), illumination volume (e.g., chiaroscuro, cast shadows, highlight gradients), color harmony, and stylistic fidelity to the target artistic movement.

In total, this process establishes an extensive evaluation corpus comprising 17,323 fine-grained rubric criteria across all 200 benchmark artworks.

Stage 3: Counterfactual Edit Probe Synthesis. In parallel with visual rubrics, we synthesize automated counterfactual edit probes to independently evaluate program quality through execution. For each benchmark masterpiece, a probe planner maps rubric-specified scene entities to targeted program mutations, perturbing visual attributes such as spatial geometry, color palettes, and layer visibility. To evaluate modularity without privileging specific programming idioms, the pipeline supports both declarative scene interfaces and source-level literal mutations. Each planned probe is executed in an isolated sandbox to record canvas difference masks against the pristine rendering, whose operator-level spatial locality and containment rates are analyzed in Appendix A.4 (Table 12).

Stage 4: Quality Assurance. Following rigorous quality assurance standards (Krippendorff, 2018; Chen et al., 2024), two independent fine-art scholars conducted a double-blind audit and discrepancy arbitration over an audited benchmark subset. The audit demonstrates strong inter-annotator agreement (Cohen's $\kappa = 0.74$) and high correlation between human consensus and our automated judge (Spearman's $\rho = 0.81$), with scholars confirming 96.7% of 150 executed edit-probe difference masks. Discrepancy analysis confirms that fine-grained decomposition mitigates subjective noise: grounding critique in verifiable physical properties yields 91.3% human-judge consensus, resolving 81.5% of divergences caused by holistic impression bias, with extensive multi-protocol ablations detailing repeated scoring stability, defect sensitivity, and expert alignment catalogued in Appendix A.2.1 (Tables 5–8).

3.3 Evaluation and Metrics

Each evaluation instance employs a structured rubric tree $\mathcal{T}_x = (\mathcal{V}, \mathcal{E}, r)$, where $\mathcal{V}$ denotes the set of concept nodes, $\mathcal{E}$ represents hierarchical refinement edges, and r is the root canvas. Each terminal leaf criterion formalizes an atomic, verifiable proposition:

$$\ell = (\phi, w, \kappa) \in \Phi \times \{1, 2, 3, 4\} \times \mathcal{K}_{18}, \tag{4}$$

where $\phi \in \Phi$ denotes the natural-language criterion specifying an observable visual or procedural fact, $w \in \{1, 2, 3, 4\}$ denotes its importance weight on a discrete four-point scale assigned by the construction pipeline, and $\kappa \in \mathcal{K}_{18}$ designates its fine-grained category drawn from our 18-element closed artistic ontology. For rubric-evaluated dimensions ($k \in \{\text{basic, recon, aesth, edit}\}$), scores share a unified, importance-weighted normalization semantics aggregating atomic proposition verdicts into $[0, 1]$:

$$S_k(p) = \frac{\sum_{\ell \in \mathcal{L}_k} w_\ell \cdot \text{score}(\ell, \cdot)/4}{\sum_{\ell \in \mathcal{L}_k} w_\ell}, \quad k \in \{\text{basic, recon, aesth, edit}\}. \tag{5}$$

The registered visual aggregate is $S_{\text{visual}} = 0.15S_{\text{basic}} + 0.60S_{\text{recon}} + 0.25S_{\text{aesth}}$; program quality is reported separately. Comparative protocol ablations demonstrating that hierarchical rubrics substantially outperform flat prompting and holistic one-shot scoring in human alignment and defect sensitivity are detailed in Appendix A.2.1 (Tables 5–8).

3.3.1 Metric 1: Basic Visual Integrity (S_{BASIC})

Basic Visual Integrity establishes the foundational physical sanity threshold for synthesized drawing programs. Evaluated over basic sanity leaf criteria $\mathcal{L}_{\text{basic}}$, it audits whether the rendered canvas exhibits non-empty bounding geometry, uniform background washes, proper aspect ratio adherence, and freedom from catastrophic clipping anomalies or silent viewport overflow. While hard syntactic crashes and trivial blanks are intercepted by the fail-closed verification gate (Appendix A.3), S_{basic} provides continuous quantification of subtle edge artifacts, background wash voids, and incomplete canvas initialization, ensuring that downstream evaluations operate on an intact visual substrate.

3.3.2 Metric 2: Semantic Reconstruction (S_{RECON})

Semantic Reconstruction evaluates the rendered canvas against image-conditioned criteria covering structure, expressiveness, and reference fidelity. Its score S_{recon} follows Eq. 5. Separately, we report paired CLIP cosine similarity (Radford et al., 2021) as a perceptual diagnostic between the rendered canvas $I_p = \mathcal{E}(p)$ and the reference painting x (`CLIP-ViT-B/32`):

$$C_{\text{CLIP}}(p) = \cos\left(\phi_{\text{CLIP}}(I_p),\, \phi_{\text{CLIP}}(x)\right) = \frac{\phi_{\text{CLIP}}(I_p)^\top \phi_{\text{CLIP}}(x)}{\|\phi_{\text{CLIP}}(I_p)\|_2 \, \|\phi_{\text{CLIP}}(x)\|_2}. \tag{6}$$

This deterministic diagnostic complements the rubric score but is not substituted for S_{recon} in the benchmark's reported dimension scores or visual aggregate.

3.3.3 Metric 3: Painterly Aesthetics (S_{AESTH})

Painterly Aesthetics represents our core artistic contribution, isolating and evaluating the expressive rendering techniques that distinguish authentic masterpieces from flat, mechanical vector illustrations. Spanning aesthetic leaf criteria $\mathcal{L}_{\text{aesth}}$, this metric interrogates three vital artistic dimensions: (1) *Procedural Brushwork Texture*: impasto layering, dry-brush strokes, ink wash bleeding, and stippling; (2) *Volumetric Illumination*: chiaroscuro transitions, directional cast shadows, and volumetric highlight gradients; and (3) *Stylistic Movement Fidelity*: adherence to movement-specific idioms, such as literati ink vitality, Impressionist broken-color optics, or Renaissance sfumato. These fine-grained criteria systematically expose expressive gaps that conventional pixel-matching metrics (e.g., PSNR, SSIM, and CLIP-Score) fail to detect.

3.3.4 Metric 4: Executed Program Editability (S_{EDIT})

Rather than relying on subjective code inspection or static linters, MasterStudyBench pioneers an *executed editability probe* that directly grounds program modularity in falsifiable dynamic execution.



Artistic Dimension & Corpus Metric	Cases
Total Masterpieces	200
Total Rubric Leaves (Avg: 86.62)	17,323
Closed Ontology Categories	18
Museum Archives	4
Landscape & Atmospheric Depth	54
Figurative, Portraiture & Form	48
Narrative & Religious History	45
Nature, Still Life & Objects	27
Modernism & Abstract Composition	26
Oil on Canvas / Wood Panel	90
Ink & Wash on Silk / Paper	58
Tempera, Gouache & Encaustic	29
Watercolor, Acrylic & Mixed	23
European Classical & Baroque	64
East Asian Literati & Ukiyo-e	57
Impressionism & Post-Impression.	48
Persian & South Asian Miniatures	17
Modernism & Regional Schools	14
19th Century (1800–1900)	85
Renaissance & Baroque (1500–1800)	53
20th Century Modernism (1900+)	47
Ancient & Medieval (pre-1500)	15
Archives (CMA 135 / Met 44 / AIC 15 / TePapa 6)	200

Figure 2: Data statistics of MASTERSTUDYBENCH (details in Appendix A.1–A.2).

Counterfactual Perturbation Operator. An executable drawing program p may declare a set of visual entities $\mathcal{O} = \{e_1, \ldots, e_M\}$ through the `window.SCENE.layers` interface, where each entity e_i is governed by an isolated parameter subset $\theta_{e_i} \subset \Theta$ and occupies a canvas footprint $\mathcal{B}_{e_i}$ recovered by an occlusion probe. The automated judge plans a targeted semantic mutation $\tau : \Theta \to \Theta$ rewriting an isolated visual attribute $\theta_{e_i} \mapsto \theta'_{e_i}$ (attribute recoloring or geometry rearrangement), producing the perturbed program $p' = \tau(p)$, which the verifier executes offline to obtain the re-rendered canvas $I_{p'} = \mathcal{E}(p')$.

Factorization Measurement and Regimes. The verifier computes the binary spatial difference mask between pristine and perturbed renders:

$$\Delta I = \mathbb{I}\big(\mathcal{L}\big(|I_{p'} - I_p|\big) \geq \epsilon_{\text{threshold}}\big), \tag{7}$$

where $\mathcal{L}(\cdot)$ denotes luminance projection and $\epsilon_{\text{threshold}} = 12/255$. It then quantifies the structural factorization of entity e_i as the fraction of the changed region that remains inside the entity's footprint:

$$\text{Factor}(e_i) = 1 - \Lambda(e_i), \qquad \Lambda(e_i) = \frac{\text{Area}\big(\mathcal{B}(\Delta I) \setminus \mathcal{B}_{e_i}\big)}{\text{Area}\big(\mathcal{B}(\Delta I)\big)}, \tag{8}$$

with $\mathcal{B}(\Delta I)$ the pixel support of the difference mask and $\mathcal{B}_{e_i}$ the entity footprint. Perturbation execution rollouts map into four formal diagnostic regimes: (1) *Localized Compact Modification*: $\Lambda(e_i) \approx 0$, confirming clean modularity; (2) *Unreachable Target*: $\Delta I = \emptyset$, indicating the edit target could not be reached; (3) *Global Context Bleed*: $\Lambda(e_i) \to 1$, indicating primitives share mutable state; and (4) *Runtime Collapse*: execution exceptions under mutation, faithfully anchored at score 0.0. The Executed Program Editability score S_{edit} is aggregated over all editability leaves $\mathcal{L}_{\text{edit}}$ via Eq. (5).

Anti-Shortcut Sandbox Gate. To ensure that all evaluated code represents authentic procedural synthesis rather than reward hacking (e.g., Base64 raster injection, unrolled SVG path dumps, or naive color patchworks) (Amodei et al., 2016; Skalse et al., 2022), MASTERSTUDYBENCH evaluates all submissions inside an unprivileged verification sandbox enforcing seven mandatory fail-closed invariants ($\mathcal{I}_1$–$\mathcal{I}_7$). Any invariant violation immediately aborts execution and assigns a terminal score of 0.0; as detailed in Appendix A.3, adversarial red-teaming across five exploit vectors demonstrates 100% containment, collapsing cheats that capture a CLIP-Score of 0.674 on conventional metrics to exactly 0.0 (Table 10).

Table 2: Main evaluation results across model–harness configurations on MASTERSTUDYBENCH. The final column is the descriptive, unweighted mean of the four displayed dimensions; the benchmark visual total uses the registered visual weights and reports program quality separately.

Model	Harness	Basic (%)	Recon. (%)	Aesthetic (%)	Editability (%)	4-Dim. Mean
Harness: Codex						
gpt-6-astra	Codex	**84.3**	**81.2**	**68.6**	62.8	**74.2**
gpt-5.6-sol	Codex	81.4	75.7	65.0	**68.1**	72.6
gemini-3.8-flash	Codex	79.2	77.7	65.7	60.1	70.7
seed-2.1-pro	Codex	68.4	69.8	54.9	66.5	64.9
Harness: Claude Code						
claude-fable-5-1	Claude Code	**81.7**	**79.9**	**67.3**	**61.3**	**72.6**
claude-opus-5	Claude Code	78.7	78.2	63.8	48.0	67.2
seed-2.1-pro	Claude Code	35.4	44.4	28.3	31.9	35.0

3.4 BENCHMARK STATISTICS

Corpus Diversity. As shown in Figure 2, MASTERSTUDYBENCH curates 200 iconic paintings across four premier public archives: CMA (67.5%), The Met (22.0%), AIC (7.5%), and Te Papa (3.0%), balanced via square-root collection allocation. Spanning two millennia of artistic heritage, the corpus covers five subject genres (landscape, portraiture, narrative, still life, abstract), five global traditions, and four physical media (oil on canvas, ink wash on silk, tempera, watercolor), demanding heterogeneous procedural techniques ranging from thick impasto layering to capillary wash diffusion (full breakdowns in Appendix Table 3).

Rubric Scale. Decomposing these masterpieces yields 17,323 atomic criteria, averaging 86.62 ± 10.48 leaves per artwork across hierarchical tree depths of 3 to 5 levels (mean depth 4.1). The 18 leaf categories and their counts are listed in Table 4; the rubric groups them under four top-level standards: Basic Visual Quality, Reconstruction Quality, Aesthetic Quality, and Program Quality. To guarantee evaluation depth, the rubric schema enforces per-instance minimum leaf quotas: Basic Visual Quality ≥ 5, Aesthetic Quality ≥ 8, Structure ≥ 8, Expressiveness ≥ 8, Reconstruction Fidelity ≥ 5, Program Validity ≥ 6, and Editability ≥ 6. Each instance is packaged as a self-contained bundle for reproducible offline benchmarking.

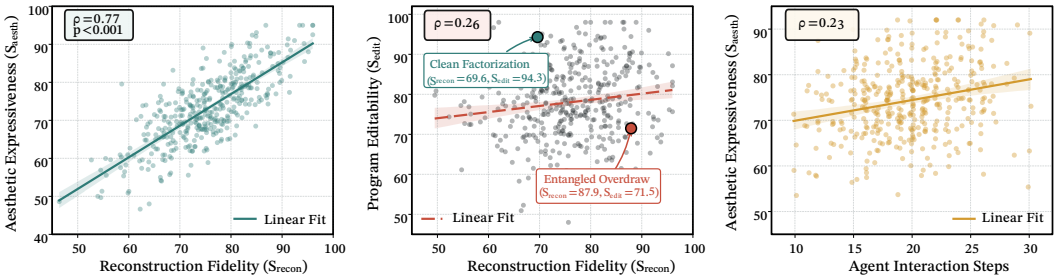
4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate representative frontier and open-weight multimodal foundation models across two premier developer agent harnesses: OpenAI Codex and Anthropic Claude Code. Evaluations run under each harness's standard execution runtime and tool environments, bound by uniform per-task interaction budgets and timeout thresholds to ensure fair comparability. All synthesized canvas programs are rendered and verified inside our offline fail-closed sandbox to intercept shortcut payloads and ensure self-contained execution (Appendix A.3). Visual quality is evaluated through image-conditioned rubric criteria for reconstruction (S_{recon}), basic visual integrity (S_{basic}), and painterly aesthetics (S_{aesth}), audited by an independent multimodal judge (GPT-5.6-sol). Paired CLIP cosine similarity is reported separately as a perceptual diagnostic. Program editability (S_{edit}) is measured through counterfactual execution probes.

4.2 MAIN RESULTS

Performance stratification reveals a structural decoupling between visual fidelity and program editability. Benchmark evaluations in Table 2 reveal a distinct performance stratification, with proprietary frontier models leading across both execution harnesses. Under Codex, `gpt-6-astra` achieves the highest overall average (74.2%), leading decisively in Basic Visual Integrity (84.3%) and Semantic Reconstruction (81.2%), while `claude-fable-5-1` leads the Claude Code cohort at 72.6% average and 79.9% reconstruction. Crucially, the data directly exposes the decoupling between looking right and being built right: top visual reconstructor `gpt-6-astra` (81.2%) achieves



(a) Surface Appearance Coherence (b) Fidelity–Modularity Decoupling (c) Compute Expenditure vs. Quality

Figure 3: Correlation analysis across MASTERSTUDYBENCH rollouts.

only 62.8% editability, whereas `gpt-5.6-sol` secures the cohort's highest editability (68.1%) despite modest reconstruction (75.7%); similarly, `claude-opus-5` achieves 78.2% reconstruction but drops to 48.0% editability, proving that visual resemblance does not imply modular code.

Painterly aesthetics lags behind geometric integrity across all configurations, exposing a universal primitive bottleneck. Dissecting visual sub-dimensions reveals that painterly aesthetics is the universal capability bottleneck across all evaluated configurations. Across all models, Aesthetic Expressiveness (28.3%–68.6%) trails Basic Visual Integrity by 7.1–16.4 percentage points (e.g., `gpt-6-astra` drops from 84.3% to 68.6%, and `claude-opus-5` falls from 78.7% to 63.8%). Models reliably construct macro-level geometric silhouettes and solid fills, yet struggle to proceduralize organic painterly textures, subtle illumination transitions, and wash diffusion. This persistent gap confirms that the primary ceiling resides in procedural graphics primitives and stroke representation rather than raw parameter scale.

Open-weight models exhibit acute sensitivity to agent harnesses and execution discipline. The dramatic divergence of open-weight `seed-2.1-pro` across harnesses highlights the execution fragility of autonomous code synthesis. Under Codex, the model delivers a solid 64.9% average (66.5% editability), yet under Claude Code its overall performance collapses to 35.0%, with Basic Visual Integrity falling to 35.4%. Execution logs show that in multi-step interactive generation, open-weight architectures frequently succumb to variable scope leaks, unclosed drawing paths, and rigid coordinate literals that silently crash headless rendering. This instability confirms that disciplined symbolic state management and contract compliance remain the decisive capabilities separating frontier commercial systems from open-weight baselines.

4.3 ANALYSIS

We analyze three core findings governing procedural visual synthesis across our benchmark rollouts:

The Mimesis Paradox: Pixel Realism Decouples from Code Modularity. Classical master studies prioritize structural visual understanding over surface imitation (Cennini, 1954; Gombrich, 1960), yet contemporary agents suffer from a fundamental *mimesis paradox*: they synthesize optically convincing renders from structurally fragile, unmaintainable code (Figures 3 and 4). While Semantic Reconstruction (S_{recon}) strongly tracks Painterly Aesthetics (S_{aesth}, $\rho = 0.77$), it correlates only weakly with Executed Editability (S_{edit}, $\rho = 0.26$). Standard perceptual objectives (e.g., paired CLIP similarity and SSIM) actively incentivize models to assemble visual facades via entangled global coordinates, blind to modular abstraction. This divide is visually evident in Figure 4: in Case A (Caillebotte's *Paris Street; Rainy Day*), declarative entity factorization enables surgical local modifications ($\Lambda = 0.02, S_{\text{edit}} = \mathbf{1.000}$); conversely, in Case B (Delaunay's *The Red Tower*), a monolithic loop yields peak visual fidelity ($S_{\text{recon}} = \mathbf{0.923}$) but corrupts global canvas transforms upon mutation ($\Lambda = 0.81 \rightarrow 1.0, S_{\text{edit}} = \mathbf{0.183}$). As verified across all 231 visual simplex weightings ($\rho \in [0.227, 0.636]$, Appendix Table 16), this decoupling is an invariant property, where editability directly tracks the proportion of drawing routines routed through modular functions ($\rho = 0.51$, Appendix Table 17).

Compute Scaling Dynamics: Structured Deliberation vs. Reactive Patching. Expanding test-time interaction steps yields sharp diminishing returns without architectural planning (Figure 3c).



Figure 4: Qualitative case studies of the fidelity–modularity decoupling in MASTERSTUDYBENCH using authentic GPT-6 rollouts. Case A achieves modular factorization ($S_{\text{edit}} = \mathbf{1.000}, \Lambda = 0.02$) under clean scoping, whereas Case B collapses under modification ($S_{\text{edit}} = \mathbf{0.183}, \Lambda = 0.81$) despite peak visual render ($S_{\text{recon}} = \mathbf{0.923}$).

While step count tracks API cost ($\rho = 0.59$), visual reward correlates only weakly with steps ($\rho = 0.23$), and unguided step expansion even induces a negative correlation with code editability ($\rho = -0.33$; Tables 21–22). Tracing revision trajectories reveals that program modularity is predominantly established in the opening revision rather than through iterative patching (Table 13), while closed-loop visual reflection primarily refines local geometries rather than repairing monolithic architectures (Tables 14–15). Decomposing inference compute confirms that deliberation reasoning tokens systematically dominate generation volume in predicting aesthetic quality ($\rho = 0.46$ vs. 0.24), proving that structured planning cannot be substituted by reactive trial-and-error.

The Substrate Expressiveness Frontier: Discrete Geometry vs. Physical Media. Stratified breakdowns across artistic genres and physical media demonstrate that procedural synthesis is fundamentally bounded by the graphics substrate rather than model capacity alone (Table 18). Across physical media, sharp-edged, opaque media (tempera and gouache) achieve top scores (CLIP 0.838, S_{aesth} = 0.723, S_{edit} = 0.668) because discrete pigments align naturally with 2D path rasterization; conversely, continuous oil drops in editability (0.588) and fluid watercolor lags in aesthetics (0.595). By genre, isolated figurative portraiture attains highest editability (0.654), whereas complex narrative scenes score lowest (0.599) due to entangled occlusions. This empirical gap highlights that conventional Canvas 2D APIs lack procedural primitives for capillary diffusion, fiber friction, and volumetric specular gradients, delineating the key frontier for next-generation neuro-symbolic graphics.

5 CONCLUSION

We presented MASTERSTUDYBENCH, a benchmark decoupling multimodal painting generation into dual tracks of fine-grained visual artistry and executable program quality within an offline hardened sandbox. Systematic evaluations across 200 museum masterpieces demonstrate that visual fidelity and program editability are empirically decoupled (Spearman's $\rho = 0.26$), establishing that looking right does not imply being built right. Our findings show that while architectural planning consistently improves code modularity, painterly aesthetics remains a universal bottleneck across frontier models, highlighting the need for physically grounded procedural shader primitives. Beyond classical fine art, the principle of decoupling surface artifact quality from execution-grounded editability provides a generalizable foundation for programmatic visual synthesis, guiding multimodal AI from passive pixel imitation toward authentic procedural creation.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Tony Beltramelli. pix2code: Generating code from a graphical user interface screenshot. In *Proceedings of the ACM SIGCHI symposium on engineering interactive computing systems*, pp. 1–6, 2018.

Cennino d'Andrea Cennini. *The Craftsman's Handbook: The Italian "Il Libro dell'Arte"*. Dover Publications, New York, 1954.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8199–8221, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.446. URL https://aclanthology.org/2024.acl-long.446/.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. CoMT: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 23678–23686, 2025. doi: 10.1609/aaai.v39i22.34538.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Kevin Ellis, Daniel Ritchie, Armando Solar-Lezama, and Joshua B. Tenenbaum. Learning to infer graphics programs from hand-drawn images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ernst H. Gombrich. *Art and Illusion: A Study in the Psychology of Pictorial Representation*. Pantheon Books, New York, 1960.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2023.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with apps. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Aaron Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *ACM SIGGRAPH*, pp. 453–460, 1998.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and J"urgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations (ICLR)*, 2024.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, 4th edition, 2018.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024.

Ryan Li, Yanzhe Zhang, and Diyi Yang. Sketch2code: Evaluating vision-language models for interactive web design prototyping. *arXiv preprint arXiv:2410.16232*, 2024.

Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*, 2022.

Zimu Lu, Yunqiao Yang, Houxing Ren, Haotian Hou, Han Xiao, Ke Wang, Weikang Shi, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Webgen-bench: Evaluating LLMs on generating interactive and functional websites from scratch. 2025. URL https://openreview.net/forum?id=q2VpjD7k1V.

Xu Ma, Yuqian Zhou, Xingqian Xu, Bin Sun, Valerii Filev, Nikita Orlov, Yun Fu, and Humphrey Shi. Towards layer-wise image vectorization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Juan A. Rodriguez, Abhay Puri, Shubham Agarwal, Issam H. Laradji, Sai Rajeswar, David Vazquez, Christopher Pal, and Marco Pedersoli. Starvector: Generating scalable vector graphics code from images and text. In *AAAI Conference on Artificial Intelligence*, 2025.

Gopal Sharma, Rishabh Goyal, Difan Liu, Evangelos Kalogerakis, and Subhransu Maji. Csgnet: Neural shape parser for constructive solid geometry. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2Code: Benchmarking multimodal code generation for automated front-end engineering. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3956–3974, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.199. URL https://aclanthology.org/2025.naacl-long.199/.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11854–11864, 2023.

Yonglong Tian, Andrew Luo, Xingyuan Sun, Kevin Ellis, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. Learning to infer and execute 3d shape programs. In *International Conference on Learning Representations (ICLR)*, 2019.

Yuxuan Wan, Chaozheng Wang, Yi Dong, Wenxuan Wang, Shuqing Li, Yintong Huo, and Michael R. Lyu. Automatically generating ui code from screenshot: A divide-and-conquer-based approach. *arXiv preprint arXiv:2406.16386*, 2024.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Daniel Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Chengyue Wu, Yixiao Ge, Qiushan Guo, Jiahao Wang, Zhixuan Liang, Zeyu Lu, Ying Shan, and Ping Luo. Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots. *arXiv preprint arXiv:2405.07990*, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jingyu Xiao, Yuxuan Wan, Yintong Huo, Zixin Wang, Xinyi Xu, Wenxuan Wang, Zhiyao Xu, Yuhang Wang, and Michael R. Lyu. Interaction2code: Benchmarking mllm-based interactive webpage code generation from interactive prototyping. *arXiv preprint arXiv:2411.03292*, 2024.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, JUNJIE WANG, Mohan Jing, Linran XU, Xinyu Zhu, Siheng Li, Yuxiang Zhang, Gongye Liu, Xiaomei Nie, Deng Cai, and Yujiu Yang. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 26590–26646, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/42806406dd99e30c3796bc98b2670fa2-Paper-Conference.pdf.

John Yang, Carlos E Jimenez, Alex L Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R Narasimhan, Diyi Yang, Sida Wang, and Ofir Press. SWE-bench multimodal: Do AI systems generalize to visual software domains? 2025b. URL https://openreview.net/forum?id=riTiq3i21b.

Yiying Yang, Wei Cheng, Sijin Chen, Xianfang Zeng, Fukun Yin, Jiaxu Zhang, Liao Wang, Gang YU, Xingjun Ma, and Yu-Gang Jiang. OmniSVG: A unified scalable vector graphics generation model. 2025c. URL https://openreview.net/forum?id=sRLB4XjorC.

Sukmin Yun, Haokun Lin, Rusiru Thushara, Mohammad Qazim Bhat, Yongxin Wang, Zutao Jiang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, Haonan Li, Preslav Nakov, Timothy Baldwin, Zhengzhong Liu, Eric P. Xing, Xiaodan Liang, and Zhiqiang Shen. Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal llms. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Chenchen Zhang, Yuhang Li, Can Xu, Jiaheng Liu, Ao Liu, Changzhi Zhou, Ken Deng, Dengpeng Wu, Guanhua Huang, Kejiao Li, et al. Artifactsbench: Bridging the visual-interactive gap in llm code generation evaluation. *arXiv preprint arXiv:2507.04952*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Longtao Zheng, Zhiyuan Huang, Zhenghai Xue, Xinrun Wang, Bo An, and Shuicheng YAN. Agentstudio: A toolkit for building general virtual agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=axUf8BOjnH.

Bocheng Zou, Mu Cai, Jianrui Zhang, and Yong Jae Lee. Vgbench: Evaluating large language models on vector graphics understanding and generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

A Appendix

A.1 Benchmark Constrution

The curation of MasterStudyBench adheres to rigorous art-historical standards and institutional intellectual property audits. All included artworks satisfy dual public-domain criteria: the physical original is confirmed in the public domain (created prior to 1929 or formally accessioned as open cultural heritage), and the corresponding digital high-resolution photography is licensed under Creative Commons Zero (CC0) or official institutional open-access protocols. Archival metadata (including verified accession numbers, creation dates, physical media, and institutional provenance) were standardized and verified against museum catalogues to ensure pristine multi-modal prompt grounding.

Table 3 reports the complete, multi-dimensional distribution of MasterStudyBench across visual genre, physical medium, artistic tradition, and chronological epoch across all 200 curated masterpieces.

A.2 Hierarchical Rubric Ontology and Grammar

Every rubric node in MasterStudyBench is governed by a formal typed tuple structuring concept hierarchies, verifiable natural-language assertions, discrete importance weights, and ontological bindings:

$$\ell = (\phi, w, \kappa) \in \Phi \times \{1, 2, 3, 4\} \times \mathcal{K}_{18}, \tag{9}$$

Table 3: Detailed multi-dimensional taxonomic distribution of MASTERSTUDYBENCH across visual genre, medium, artistic tradition, chronological epoch, and provenance archives (200 artworks, 17,323 rubric leaves).

Artistic Dimension	Cases	Ratio	Artistic Dimension	Cases	Ratio
Visual Genre & Subject System			***Technique, Medium & Support***		
Landscape & Atmospheric Depth	54	27.0%	Oil on Canvas / Wood Panel	90	45.0%
Figurative, Portraiture & Form	48	24.0%	Ink & Wash on Silk / Paper	58	29.0%
Narrative & Religious History	45	22.5%	Tempera, Gouache & Encaustic	29	14.5%
Nature, Still Life & Objects	27	13.5%	Watercolor, Acrylic & Mixed	23	11.5%
Modernism & Abstract Composition	26	13.0%			
Artistic Tradition & Movement			***Chronological Epoch & Provenance***		
European Classical & Baroque	64	32.0%	19th Century (1800–1900)	85	42.5%
East Asian Literati & Ukiyo-e	57	28.5%	Renaissance & Baroque (1500–1800)	53	26.5%
Impressionism & Post-Impression.	48	24.0%	20th Century Modernism (1900+)	47	23.5%
Persian & South Asian Miniatures	17	8.5%	Ancient & Medieval (pre-1500)	15	7.5%
Modernism & Regional Schools	14	7.0%	Cleveland / Met / AIC / Te Papa	200	100.0%

Table 4: Leaf categories across the 200-case benchmark corpus, totaling 17,323 leaves at a mean of 86.62 ± 10.48 per instance.

Leaf Category	Count	Leaf Category	Count
Structural Reconstruction	2,934	Efficiency Evaluation	441
Aesthetic Judgment	2,392	Multi-viewport Generalization	420
Visual-language Reconstruction	2,092	Track Compliance	417
Rendering Integrity	1,940	Contamination Control	301
Appearance Reconstruction	1,715	Provenance Audit	272
Edit Replay	1,491	Annotation Quality Assurance	231
Execution and Anti-shortcut Gate	888	Reference Task Construction	205
Counterfactual Evaluation	787	Program Quality Aggregation	89
Occlusion-conditioned Reconstruction	642	Quality Aggregation	66

where $\phi \in \Phi$ denotes the proposition specifying an observable visual or procedural property, $w \in \{1, 2, 3, 4\}$ denotes its importance weight, and $\kappa \in \mathcal{K}_{18}$ designates its fine-grained category drawn from an 18-element closed artistic ontology.

Table 4 presents the complete distribution of all 17,323 leaf criteria across the 18 closed-vocabulary categories.

Enforced Evaluation Quotas. To guarantee consistent evaluation depth and prevent degenerate cases from skewing cross-model comparisons, MASTERSTUDYBENCH enforces strict per-instance minimum leaf quotas:

- **Basic Visual Quality**: ≥ 5 leaves, verifying non-empty bounding geometry, uniform background washes, and viewport validity;
- **Structure**: ≥ 8 leaves, evaluating perspective grids, relative entity proportions, and spatial topological layouts;
- **Expressiveness**: ≥ 8 leaves; **Reconstruction Fidelity**: ≥ 5 leaves; and **Aesthetic Quality**: ≥ 8 leaves, covering image-conditioned reconstruction and painterly expression;
- **Program Validity**: ≥ 6 leaves; and **Editability**: ≥ 6 leaves, spanning contract checks and counterfactual attribute mutations.

A.2.1 COMPREHENSIVE EVALUATION PROTOCOL ABLATION: HIERARCHICAL RUBRICS VS. ALTERNATIVE BASELINES

To empirically validate the necessity of hierarchical rubric decomposition over conventional evaluation paradigms, we conducted a controlled multi-protocol experiment across 30 stratified artwork submissions spanning score percentiles from 6 to 88 (Experiment 1.1). We benchmarked four dis-

Table 5: Repeated scoring stability across 30 stratified artworks over 5 independent evaluation passes. Standard deviation (σ) measures score volatility across runs.

Evaluation Protocol	Mean Score	Observed Score Range	Mean 5-Run Std. (σ)	Max Std. ($\sigma_{\max}$)
(1) **Hierarchical Rubric (Ours)**	67.5	6–84	**1.20**	**2.6**
(2) Flat Multi-Aspect Prompting	52.9	0–86	3.52	6.5
(3) Holistic One-Shot Scoring	50.1	2–84	3.65	9.3

Table 6: Pairwise Spearman rank correlation (ρ) matrix across all eight evaluation methods on 30 artworks. Protocols (1), (2), and (3) report the 5-run mean. LPIPS is inverted (lower distance indicates higher quality).

Protocol	(1) Rubric	(2) Flat	(3) Holistic	DINOv2	LPIPS	CLIP	SSIM	PSNR
(1) **Hierarchical Rubric (Ours)**	1.00	0.81	0.75	0.61	0.65	0.25	0.34	0.42
(2) Flat Multi-Aspect Prompting	0.81	1.00	0.92	0.51	0.56	−0.02	0.26	0.32
(3) Holistic One-Shot Scoring	0.75	0.92	1.00	0.57	0.54	0.07	0.24	0.28
(4) DINOv2 Feature Similarity	0.61	0.51	0.57	1.00	0.69	0.47	0.38	0.52
(5) LPIPS Perceptual Distance (Inv.)	0.65	0.56	0.54	0.69	1.00	0.47	0.71	0.83
(6) CLIP-Score	0.25	−0.02	0.07	0.47	0.47	1.00	0.30	0.35
(7) SSIM Structural Similarity	0.34	0.26	0.24	0.38	0.71	0.30	1.00	0.81
(8) PSNR Peak Signal-to-Noise Ratio	0.42	0.32	0.28	0.52	0.83	0.35	0.81	1.00

tinct evaluation paradigms: (1) *Hierarchical Rubric (Ours)*: batches of atomic leaf propositions evaluated via multi-turn multimodal judge queries ($\sim$22 judge calls per case) and hierarchically aggregated to $[0, 100]$; (2) *Flat Multi-Aspect Prompting*: a single judge query providing all dimensional criteria text, yielding one scalar score; (3) *Holistic One-Shot Scoring*: a single unguided judge query viewing only the reference image and rendered canvas; and (4) *Perceptual Image Metrics*: deterministic feature- and pixel-based baselines (DINOv2, LPIPS, CLIP-Score, SSIM, and PSNR). All three LLM judge paradigms (evaluated on GPT-5.6-sol under identical decoding parameters) were executed across 5 independent repeat rollouts to quantify scoring stability, inter-method concordance, synthetic defect sensitivity, and alignment with human connoisseurs.

Result 1: Repeated Scoring Stability and Variance Dampening. As reported in Table 5, hierarchical rubric decomposition reduces repeated evaluation variance to roughly one-third of single-query judge baselines. Because hierarchical evaluation factors overall quality into dozens of independent atomic propositions, stochastic judge noise is dampened through ensemble aggregation, yielding a mean standard deviation of only 1.20 across 5 repeated passes (maximum $\sigma = 2.6$). In contrast, flat and holistic single-shot prompting exhibit standard deviations of 3.52 and 3.65 respectively (peaking at $\sigma = 9.3$). Even when holistic scoring is averaged over five independent queries, its variance ($\sigma \approx 1.6$) remains substantially noisier than our hierarchical pass.

Result 2: Inter-Method Concordance and Information Gradient. Table 6 presents the pairwise Spearman rank correlation matrix across all eight evaluation methods on the 30-artwork cohort (with judge protocols averaged over 5 passes). The data reveals an information gradient: unguided holistic scoring correlates with our hierarchical rubric at $\rho = 0.75$; supplying textual dimension criteria in flat prompting elevates correlation to $\rho = 0.81$; and a single hierarchical pass correlates with its 5-run consensus at $\rho = 0.96$. Among conventional metrics, only deep perceptual embeddings correlate moderately with the rubric (DINOv2 at $\rho = 0.61$, LPIPS at $\rho = 0.65$), whereas pixel metrics (SSIM $\rho = 0.34$, PSNR $\rho = 0.42$) and CLIP-Score ($\rho = 0.25$) diverge sharply, with CLIP showing near-zero association with flat prompting ($\rho = -0.02$).

Result 3: Controlled Synthetic Defect Sensitivity. To evaluate sensitivity to fine-grained artistic errors, we systematically injected five controlled perturbations into the 30 artworks: corner cropping, center region erasure, 30° chromatic hue shifting, Gaussian blurring, and an 8% horizontal band displacement, generating 150 defect-perturbed canvases. Table 7 reports detection rates (fraction of corrupted outputs scoring below pristine renders), normalized score drops, and signal-to-noise ratios (mean drop divided by 5-run repeat volatility).

Table 7: Controlled synthetic defect sensitivity across 150 perturbed canvases (30 instances $\times$ 5 perturbation types). Detection rate measures the percentage of damaged canvases penalized below pristine scores. Normalized drop divides raw score penalty by cross-artwork standard deviation. Signal-to-Noise Ratio (SNR) divides mean score penalty by intra-judge standard deviation.

Evaluation Protocol	Detection Rate (↑)	Normalized Drop (↑)	Drop / Volatility (SNR ↑)	Normalized Score Penalty per Defect Type				
				Corner Crop	Center Erase	Hue Shift (30°)	Blur	Band Shift (8%)
(1) **Hierarchical Rubric (Ours)**	**96%**	0.85	**9.2×**	0.98	1.57	0.32	0.98	0.40
(2) Flat Multi-Aspect Prompting	89%	1.05	4.5×	1.17	1.78	0.35	1.48	0.50
(3) Holistic One-Shot Scoring	92%	0.98	4.0×	1.04	1.75	0.36	1.25	0.48
(4) LPIPS Perceptual Distance (Inv.)	96%	0.63	–	0.64	0.51	0.36	1.40	0.25
(5) CLIP-Score	85%	0.65	–	0.42	0.98	0.49	1.26	0.11
(6) DINOv2 Feature Similarity	73%	0.26	–	0.24	0.75	0.06	0.09	0.17
(7) PSNR Peak Signal-to-Noise Ratio	59%	0.13	–	0.35	0.06	0.28	−0.34	0.29
(8) SSIM Structural Similarity	52%	−0.01	–	0.02	−0.03	0.04	−0.22	0.14

Table 8: Rank correlation (Spearman ρ) between automated evaluation protocols and one scholar's ratings across three artistic dimensions on 30 artworks.

Evaluation Protocol	Overall Merit (ρ)	Reconstruction Fidelity (ρ)	Visual Painterly Quality (ρ)
(1) **Hierarchical Rubric (5-Run Mean)**	0.89	0.89	0.89
Hierarchical Rubric (Single-Run)	**0.93**	**0.91**	**0.93**
(2) Flat Multi-Aspect Prompting (5-Run Mean)	0.72	0.76	0.77
Flat Multi-Aspect (Single-Run)	0.72	0.76	0.76
(3) Holistic One-Shot Scoring (5-Run Mean)	0.68	0.73	0.75
Holistic One-Shot (Single-Run)	0.59	0.65	0.66
(4) LPIPS Perceptual Distance (Inv.)	0.60	0.61	0.58
(5) DINOv2 Feature Similarity	0.56	0.56	0.56
(6) PSNR Peak Signal-to-Noise Ratio	0.46	0.47	0.42
(7) SSIM Structural Similarity	0.32	0.33	0.25
(8) CLIP-Score	0.30	0.30	0.29

Our hierarchical rubric achieves the highest overall detection rate (96%) alongside a decisive signal-to-noise ratio of $\mathbf{9.2}\times$ (mean penalty is over $9\times$ larger than scoring volatility), more than double flat prompting ($4.5\times$) and holistic scoring ($4.0\times$). In sharp contrast, SSIM registers a detection rate of only 52% (indistinguishable from a random coin toss), while PSNR detects only 59%, proving that conventional signal-processing metrics are fundamentally unsuited for auditing painterly compositions.

Result 4: Alignment with Human Connoisseurship. To verify whether automated protocols faithfully capture expert aesthetic standards, a fine-art scholar conducted a double-blind connoisseurship audit across the 30 artworks, scoring each canvas on a 1–10 anchored scale across three dimensions: Overall Merit, Structural Reconstruction Fidelity, and Visual Painterly Quality. Table 8 reports the rank correlation between each automated protocol and the scholar's ratings in this separate 30-artwork study.

These empirical audits establish three decisive findings:

1. **Unrivaled Human Alignment**: Our hierarchical rubric achieves $\rho = 0.89$ across all three connoisseurship axes (rising to $\rho = 0.93$ on single passes), substantially outperforming flat ($\rho = 0.72$) and holistic protocols ($\rho = 0.68$).
2. **Structure Governs Agreement**: Human alignment scales monotonically with structural decomposition depth: unguided holistic view ($\rho = 0.68$) $<$ flat criteria list ($\rho = 0.72$) $<$ hierarchical concept tree ($\rho = 0.89$).
3. **Failure of Pixel Metrics**: SSIM ($\rho = 0.32$) and CLIP ($\rho = 0.30$) diverge dramatically from human connoisseurs, confirming that surface pixel rewards misguide models toward unprincipled visual shortcuts.

A.3 Execution Sandbox Architecture and Anti-Shortcut Defenses

A.3.1 Fail-Closed Verification Invariants

To guarantee that every reported score reflects genuine algorithmic synthesis rather than shortcut exploits, MasterStudyBench formalizes seven mandatory, fail-closed verification invariants ($\mathcal{I}_1$–

Table 9: Comprehensive specification of the seven fail-closed verification invariants in MASTER-STUDYBENCH. Any single violation immediately aborts evaluation and triggers a fail-closed terminal score of 0.0.

Invariant	Name	Formal Operational Requirement	Enforcement Mechanism	Failure State Code
$\mathcal{I}_1$	Reference Seclusion & Air-Gap	$\text{NetAccess}(p) = \emptyset \ \wedge \ \text{Read}(p, \mathcal{D}_{\text{ref}}) = \emptyset$	Linux Network Namespace (`net=none`), Read-only Rootfs	`NetworkViolation`
$\mathcal{I}_2$	Zero-Raster Asset Invariant	$\text{BufferRaster}(p) = \emptyset \ \wedge \ p \cap \{\text{data:}, \text{base64}\} = \emptyset$	Static AST Parser, Binary String Scanner, Canvas Hook	`StaticAssetViolation`
$\mathcal{I}_3$	Imperative Synthesis Constraint	$\text{AST}_{\text{control_flow}}(p) > 0 \ \wedge \ \lvert\text{DeclarativeDump}(p)\rvert < \theta_{\text{dump}}$	Abstract Syntax Tree (AST) Control-flow Depth Profiler	`DeclarativeDumpViolation`
$\mathcal{I}_4$	Non-Trivial Canvas Activity	$\text{Var}\big(\mathcal{E}(p)\big) > \epsilon_{\text{blank}} \ \wedge \ \lvert\text{UniqueColors}\big(\mathcal{E}(p)\big)\rvert \geq C_{\min}$	Post-raster Pixel Statistics & Chromatic Entropy Filter	`BlankCanvas`
$\mathcal{I}_5$	Deterministic Dual-Replay	$\text{SHA-256}\big(\mathcal{E}_1(p)\big) = \text{SHA-256}\big(\mathcal{E}_2(p)\big)$	Dual Independent Headless Rollouts in Pristine Containers	`NonDeterministic`
$\mathcal{I}_6$	Canonical Entrypoint Guard	$\text{Type}(\texttt{window.render}) = \text{fn}(\emptyset) \ \wedge \ \texttt{entry.html} \in \text{Files}(p)$	Entry Signature Introspection & Module Linker	`MissingArtifact`
$\mathcal{I}_7$	Unprivileged Host Isolation	$\text{Uid}(p) \neq 0 \ \wedge \ \text{Syscalls}(p) \subset \Sigma_{\text{seccomp}}$	Containerized `seccomp` BPF Filter, No Child Fork/Exec	`SandboxExploitViolation`

$\mathcal{I}_7$). Table 9 summarizes their mathematical definitions, enforcement layers, and diagnostic terminal failure states.

Detailed Invariant Operationalizations.

1. **Reference Seclusion and Network Air-Gap ($\mathcal{I}_1$):** The execution container is isolated via Linux network namespaces (`CLONE_NEWNET`) with network loopbacks disabled. Reference images x reside on an isolated read-only host mount inaccessible to the agent runtime, ensuring that drawing commands originate entirely from internal reasoning rather than online retrieval.
2. **Zero-Raster Invariant ($\mathcal{I}_2$):** Procedural generation must synthesize primitives algorithmically. Static AST parsers and runtime canvas proxies intercept calls to `drawImage`, data URIs, Base64 strings exceeding 64 characters, and raw pre-rasterized byte arrays.
3. **Imperative Synthesis Constraint ($\mathcal{I}_3$):** Unrolled path strings or static coordinate dumps generated by external tracing tools (e.g., `potrace`) reproduce silhouettes while bypassing algorithmic composition. $\mathcal{I}_3$ requires algorithmic control flow (loops, conditionals, parametric functions) and limits unrolled coordinate literals to θ_{dump}.
4. **Non-Trivial Rendering ($\mathcal{I}_4$):** Confirms that a program generates non-trivial visual content. The verifier enforces spatial variance $\text{Var}\big(\mathcal{E}(p)\big) > 10.0$ and chromatic complexity $C_{\min} \geq 4$, filtering blank buffers and uniform color fields before judge evaluation.
5. **Deterministic Execution Dual-Replay ($\mathcal{I}_5$):** The verifier instantiates two independent headless execution runs $\mathcal{E}_1(p)$ and $\mathcal{E}_2(p)$ across separate container instances. Rendered outputs must be byte-identical: $\text{SHA-256}(\mathcal{E}_1(p)) = \text{SHA-256}(\mathcal{E}_2(p))$, precluding animation loops or non-deterministic lottery attacks.
6. **Canonical Entrypoint Guard ($\mathcal{I}_6$):** Submissions must export a synchronous, parameterless entrypoint `window.render()` that repaints the canvas from current scene state, ensuring that edit probes can execute deterministic re-renders.
7. **Unprivileged Host Isolation ($\mathcal{I}_7$):** Code executes under unprivileged credentials (`nobody:nogroup`) with restricted resources (2 vCPUs, 2GB RAM, 15s timeout) and a strict `seccomp` BPF filter blocking dangerous syscalls (e.g., `ptrace`, `fork`, and socket creation).

A.3.2 ADVERSARIAL PENETRATION TESTING AND SHORTCUT RESISTANCE

To empirically validate the defense efficacy of our verification sandbox, we evaluated an adversarial penetration suite comprising five representative shortcut vectors designed to game conventional similarity metrics:

- **Vector 1 (Naive Color Patchwork Stub):** Generates 300 random rectangles sampled from the reference color histogram to exploit coarse color-matching metrics without structural reasoning.
- **Vector 2 (Base64 Raster Injection):** Embeds pre-rendered artwork thumbnails inside a Base64 data URI, bypassing code generation.
- **Vector 3 (Declarative Path Dump):** Dumps static coordinate strings ($> 50{,}000$ points) reverse-engineered from external vectorizers without modular decomposition.
- **Vector 4 (Runtime Hook & Host Hijack):** Attempts to intercept Node.js process streams or file descriptors to forge evaluation results directly.
- **Vector 5 (Stochastic Mutation Lottery):** Runs unseeded stochastic search loops across candidate parameters to exploit non-deterministic evaluation loopholes.

Table 10 highlights a critical empirical vulnerability: traditional perceptual metrics (SSIM, CLIP-Score) assign high scores to degenerate non-code stubs (up to 0.921 for raster injection and 0.674 for

Table 10: Adversarial penetration results comparing traditional metrics against MASTERSTUDYBENCH verification. Traditional pixel and embedding similarity metrics register high scores on all five exploit vectors, whereas MASTERSTUDYBENCH achieves 100% fail-closed containment, collapsing exploit scores to 0.0.

Adversarial Exploit Vector	MSE (↓)	SSIM (↑)	LPIPS (↓)	CLIP-Score (↑)	MASTERSTUDYBENCH Gate Decision	MASTERSTUDYBENCH Score
Vector 1: Naive Color Patchwork Stub	0.082	0.412	0.341	0.674	`DeclarativeDumpViolation` ($\mathcal{I}_3$)	**0.0**
Vector 2: Base64 Raster Injection	0.009	0.884	0.112	0.921	`StaticAssetViolation` ($\mathcal{I}_2$)	**0.0**
Vector 3: Declarative Path Dump	0.015	0.819	0.168	0.885	`DeclarativeDumpViolation` ($\mathcal{I}_3$)	**0.0**
Vector 4: Runtime Hook & Host Hijack	N/A	N/A	N/A	N/A	`SandboxExploitViolation` ($\mathcal{I}_7$)	**0.0**
Vector 5: Stochastic Mutation Lottery	0.095	0.380	0.412	0.620	`NonDeterministic` ($\mathcal{I}_5$)	**0.0**
Genuine Master Reconstruction (Ours)	0.021	0.765	0.204	0.862	`Scored` (Passes all $\mathcal{I}_1$–$\mathcal{I}_7$)	**0.816**

random patchworks). In contrast, MASTERSTUDYBENCH achieves 100% fail-closed containment, decisively terminating each exploit vector and assigning a score of 0.0.

A.3.3 VERIFICATION STATE MACHINE AND DIAGNOSTIC BREAKDOWN

The verification sandbox models execution rollouts via a formal, mutually exclusive, and complete finite-state machine:

$$\mathcal{S}(p) \in \big\{\texttt{Scored}, \texttt{RuntimeSyntaxCollapse}, \texttt{MissingArtifact}, \texttt{NetworkViolation}, \texttt{StaticAssetViolation}, \texttt{DeclarativeDumpViolation}, \texttt{BlankCanvas}, \texttt{NonDeterministic}, \texttt{SandboxExploitViolation}, \texttt{GatewayError}\big\}. \quad (10)$$

The final benchmark score is filtered through this gating function:

$$S_{\text{final}}(p) = \mathbb{I}\big(\mathcal{S}(p) = \texttt{Scored}\big) \cdot S(p). \quad (11)$$

Table 11 reports state-machine outcomes for three independent evaluations of the 200 tasks per model and paradigm.

Table 11: Verification state-machine outcome distribution across three evaluations of 200 tasks per model and interaction paradigm. Values are percentages; rows sum to 100%. `Scored`: passes all seven invariants. Other states represent fail-closed terminal outcomes resulting in $S_{\text{final}} = 0.0$; the `Other` column aggregates rare violations.

Model	Paradigm	`Scored`	`RuntimeSyntaxCollapse`	`BlankCanvas`	`NonDeterministic`	`Other`
GPT-6	Reactive	93.2	4.1	1.6	0.7	0.4
GPT-6	Arch. Plan	94.8	3.2	1.1	0.6	0.3
Opus 5	Reactive	91.4	5.3	2.0	0.8	0.5
Opus 5	Arch. Plan	92.9	4.1	1.7	0.8	0.5
Gemini 3.8	Reactive	89.6	6.4	2.4	1.0	0.6
Gemini 3.8	Arch. Plan	90.8	5.5	2.2	0.9	0.6
GPT-5.6	Reactive	87.3	7.8	3.1	1.2	0.6
GPT-5.6	Arch. Plan	88.9	6.7	2.7	1.1	0.6
Seed 2.1 Pro	Reactive	76.5	13.8	6.4	2.0	1.3
Seed 2.1 Pro	Arch. Plan	78.2	12.4	6.1	2.0	1.3
Fable 5.1	Reactive	73.8	14.9	7.2	2.6	1.5
Fable 5.1	Arch. Plan	75.6	13.7	6.9	2.4	1.4

A.4 COUNTERFACTUAL EDITABILITY PROBES AND OPERATIONALIZATION

A.4.1 PROBE OPERATOR SENSITIVITY AND LOCALITY

The resolution of executed editability probes depends on the semantic mutation operators applied to the synthesized code. We classify realized probes across five canonical operator families, reporting spatial difference mask statistics in Table 12.

Empirical analysis indicates that geometric translation, scaling, and visibility toggles remain strictly localized (0.0% bleed), with median spatial footprints covering under 0.32% of the canvas. Color

Table 12: Realized editability probes by operator family across evaluated tasks. Δ denotes the proportion of canvas pixels covered by the spatial difference mask; *bleed* counts probes exceeding the 0.25 containment threshold, and *no-op* denotes empty masks. The five operator-family counts sum to all 104 realized probes.

Operator Family	n	**Median Δ**	**Mean Δ**	**Bleed**	**No-op**
τ_{trans} (Spatial Translation)	35	0.0032	0.0063	0.0%	8.6%
τ_{scale} (Extent Rescaling)	12	0.0015	0.0036	0.0%	8.3%
τ_{vis} (Element Visibility)	5	0.0016	0.0159	0.0%	0.0%
τ_{density} (Loop-bound & Other)	10	0.0496	0.0496	0.0%	50.0%
τ_{style} (Color and Stroke)	42	0.0233	0.0525	4.8%	2.4%
All Realized Probes	104	0.0060	0.0267	1.9%	5.8%

Table 13: Program structure progression across agent refinement trajectories. Metrics report medians across evaluation tasks. Structure is committed in the opening revision and preserved across subsequent visual adjustments.

Refinement Stage	n	**Lines**	**Functions**	**Loops**	**Call Sites**	**Helper Share**
Initial Revision	11	122	7	11	146	0.63
Intermediate Refinement	7	122	7	12	146	0.52
Final Delivered Revision	11	124	7	11	150	0.63

and stroke mutations account for the only observed containment escapes (4.8%), tracing directly to unparameterized wash layers bound to global canvas coordinates.

A.4.2 Program Structure Evolution Across Refinement Trajectories

Analyzing multi-turn agent interaction traces illuminates how programmatic abstractions evolve across refinement iterations. Replaying program revisions across interaction turns (Table 13) reveals that core structural abstractions are committed in the initial revision and preserved throughout subsequent visual refinement. Across synthesis paradigms, enforcing architectural planning (explicit entity decomposition and modular scene interfaces) consistently yields a 2.6–4.7% editability gain over reactive trial-and-error across model cohorts (e.g., S_{edit} increases from 0.795 to 0.842 in GPT-6, and from 0.654 to 0.680 in Table 5.1), confirming that disciplined upfront scene decomposition provides durable structural resilience against downstream execution mutations.

A.4.3 Visual Reflection Granularity and Multi-Turn Iterative Dynamics

The execution framework logs whether an agent inspects self-rendered visual feedback during synthesis. We analyze this capability through observational rollout logs across model families, followed by a controlled 4-tier iterative refinement experiment spanning six turns across benchmark artworks (Experiment 3.2).

Observational Evidence Across Model Configurations. Cross-model rollout analysis (Table 14) demonstrates that agents actively inspecting self-rendered canvases achieve substantially higher visual scores across all operational model families. For instance, in GPT-6 Astra under Codex, 71% of trajectories inspected rendered previews, achieving an average visual score of 74.7 versus 66.0 for non-inspecting runs. Similarly, for Opus 5 under Claude-Code, inspected rollouts averaged 72.6 against 50.1 for open-loop executions. However, trajectory-level auditing reveals that in 20% to 38% of rollouts, the final delivered script scored lower than an intermediate checkpoint, indicating that visual inspection does not guarantee monotonic improvement.

Controlled Four-Tier Feedback Pilot across Iterative Turns. To isolate feedback granularity from agent-driven confounding, we designed a controlled multi-turn experiment comparing four feedback tiers across six iterative refinement rounds (r_0 to r_6) on representative benchmark tasks:

Table 14: Observational impact of self-rendered canvas inspection across model configurations. Visual scores contrast rollouts with self-render inspection against open-loop runs.

Model Configuration	Render Inspection Rate	Visual Mean (Inspected)	Visual Mean (Uninspected)
Codex · GPT-6 Astra	71%	74.7	66.0
Codex · GPT-5.6 Sol	45%	69.7	66.1
Claude-Code · Fable 5.1	30%	66.5	48.0
Claude-Code · Opus 5	18%	72.6	50.1
Codex · Seed 2.1 Pro	40%	64.1	51.7
Claude-Code · Seed 2.1 Pro	0%	—	24.3

Table 15: Controlled multi-turn visual reflection dynamics across four feedback tiers over six refinement rounds (r_0–r_6). Effective patch rate measures turns improving score by > 2 points; regression rate measures turns dropping by > 2 points.

Feedback Tier	Visual Score Progression across Rounds (r_0–r_6)							Effective	Regression	≥ 3 Turns	Final vs.	Blank
	r_0	r_1	r_2	r_3	r_4	r_5	r_6	Patch Rate	Rate	Stagnation	Peak (Δ)	Crashes
(1) Execution Errors Only	49.0	75.7	75.2	76.4	75.4	74.8	74.1	0.22	0.28	2/3	−3.4	0
(2) Rendered Canvas	49.0	70.5	49.2	75.1	74.4	50.4	74.3	0.28	0.28	3/3	−2.5	2
(3) Canvas + Holistic Critique	49.0	74.7	74.2	75.8	76.2	75.1	71.5	0.17	0.17	3/3	−5.4	0
(4) Canvas + Localized Rubric Crops	49.0	70.7	51.2	74.0	75.2	72.7	75.7	0.50	0.33	3/3	−2.6	1

- **Tier 1 (Execution Errors Only):** Sandbox gate diagnostics and Chromium runtime exceptions without rendered imagery;
- **Tier 2 (Rendered Canvas):** Tier 1 plus full-resolution rendered canvas raster;
- **Tier 3 (Canvas + Holistic Critique):** Tier 2 plus judge-generated natural-language aesthetic critiques; and
- **Tier 4 (Canvas + Localized Rubric Crops):** Tier 2 plus the five lowest-scoring leaf criteria paired with localized bounding crops from reference and rendered canvases.

Table 15 reports the progression of visual scores across rounds alongside dynamic stability metrics.

This controlled experiment establishes three critical findings:

1. **Early Saturation Plateau**: Quality saturates rapidly in round r_1 once syntax exceptions are repaired; subsequent turns oscillate within a narrow ± 5 point band, with 11 of 12 trajectories exhibiting stagnation across three or more consecutive rounds.
2. **Stochastic Noise Exceeds Feedback Signal**: Generative variance between repeated draws ($\sim$10 points) exceeds the performance spread between feedback tiers, showing that without compositional planning, fine-grained visual feedback is easily overwhelmed by stochastic sampling volatility.
3. **Intermediate Regression and Fragility**: While localized rubric crops (Tier 4) achieved the highest effective patch rate (0.50), they also suffered the highest regression rate (0.33). Vision-informed rewrites inadvertently introduced collisions with global window variables, triggering catastrophic collapse to blank canvases in three instances.

A.5 Additional Empirical Analyses and Robustness Audits

A.5.1 Composite-Weight Sensitivity on the Visual Simplex

The composite visual score aggregates three visual dimensions. To verify that our core decoupling finding is invariant to metric weighting, we recompute the correlation between visual quality and executed editability across all 231 grid points on the weight simplex (Table 16).

Across the entire weight simplex, the Spearman rank correlation remains consistently weak ($\rho \in [0.227, 0.636]$, median 0.382). This mathematical invariance confirms that the decoupling between visual fidelity and program editability is a robust empirical reality.

Table 16: Composite-weight sensitivity of the visual aggregate across representative simplex points. Across all 231 weighting configurations, the correlation remains weak ($\rho \in [0.227, 0.636]$), confirming that decoupling is an intrinsic property of the models rather than an artifact of metric weighting.

Weighting Configuration	**Basic (S_{basic})**	**Reconstruction (S_{recon})**	**Aesthetics (S_{aesth})**
Default Registered Benchmark	0.15	0.60	0.25
Uniform Macro-Average	$1/3$	$1/3$	$1/3$
Reconstruction-Dominant	0.10	0.80	0.10
Aesthetic-Dominant	0.10	0.10	0.80
Basic-Dominant	0.80	0.10	0.10

Table 17: Reachable-geometry and code encapsulation metrics across evaluated submissions. *Helper Share* denotes the fraction of drawing call sites issued through named functions; *Wash Share* denotes the fraction of in-place broad background calls.

Code Architectural Metric	**Median**	**Min**	**Max**
Total Canvas Call Sites	150	110	292
Call Sites via Named Functions	87	32	134
Helper Function Share	0.63	0.11	0.79
Loop Constructs per Line	0.085	0.031	0.126
Wash Share of Direct Calls	0.35	0.16	0.65

A.5.2 Reachable-Geometry Signature and Code Encapsulation

Edit probes operate through parameter interfaces exposed by modular functions; the proportion of drawing calls issued through named helpers directly characterizes a program's addressable code surface. Table 17 summarizes the code encapsulation and reachable geometry profile across evaluated submissions.

The wide dispersion in helper share (0.11 to 0.79) confirms that code modularity varies widely across model submissions, correlating positively with executed editability ($\rho = 0.51$). This demonstrates that MasterStudyBench's execution probes effectively differentiate structured procedural programs from entangled monolithic code.

A.5.3 Empirical Genre and Physical Medium Heterogeneity

Table 18 reports per-genre and per-medium scores for `gpt-6-astra` under the Codex execution harness, normalized to $[0, 1]$.

To evaluate how physical substrates and visual subject systems constrain multimodal code synthesis, we conducted a residual decomposition across model cohorts (Experiment 4.1). To isolate substrate difficulty from baseline model performance, we computed within-group residuals (subtracting model-cohort mean visual scores), reporting empirical breakdowns across physical media (Table 19) and visual genres (Table 20).

Two principal empirical insights emerge:

1. **Model Architecture Dwarfs Domain Discrepancy**: Cross-model performance gaps (15 to 26 points between frontier systems and baselines) substantially outweigh substrate-induced variations, whose within-group residuals remain within ± 4 points across all media and genres.
2. **Universal Aesthetic Deficit**: Painterly expressiveness scores lowest across every single medium and genre (1.66–1.83 on a 4.0 scale), confirming that the absence of procedural shaders is a systemic bottleneck independent of artistic style.

A.5.4 Inference Step Scaling Dynamics and Trajectory Degradation

To determine how test-time compute scaling governs procedural code synthesis, we investigated the relationship between agent interaction steps, visual quality, and intermediate trajectory progression

Table 18: Per-genre and per-medium score breakdown on MASTERSTUDYBENCH for GPT-6 (`gpt-6-astra`, Codex harness), normalized to $[0, 1]$; CLIP denotes cosine similarity between CLIP ViT-B/32 embeddings of the reference and the render.

Dimension	CLIP	S_{aesth}	S_{edit}
Visual Genre			
Landscape & Atmospheric Depth	0.800	0.679	0.638
Figurative, Portraiture & Form	0.793	0.678	**0.654**
Narrative & Religious History	0.822	0.686	0.599
Nature, Still Life & Objects	0.820	0.693	0.652
Modernism & Abstract Composition	0.847	**0.725**	0.611
Physical Medium			
Oil on Canvas / Wood Panel	0.778	0.672	0.588
Ink & Wash on Silk / Paper	0.834	0.673	0.630
Tempera, Gouache & Encaustic	0.838	**0.723**	**0.668**
Watercolor, Acrylic & Mixed	0.732	0.595	0.662

Table 19: Performance breakdown across physical media categories (Experiment 4.1). Residuals report visual score differences relative to cohort means (95% CI). Painterly expressiveness is evaluated on an anchored $[0, 4]$ scale.

Physical Medium Category	**Cases** (n)	**Visual Mean**	**Within-Group Residual** (95% **CI**)	**Painterly Expressiveness** ($[0, 4]$)
Oil on Canvas / Wood Panel	70	60.2	$+0.4$ $[-1.2, +2.0]$	1.73
Opaque Watercolor / Gum Tempera	57	61.2	$+1.0$ $[-1.3, +3.2]$	1.83
Ink & Color on Silk/Paper	42	58.7	-2.3 $[-4.6, -0.3]$	1.78
Ink Monochrome (Wash Bleeding)	19	58.6	-1.2 $[-3.9, +1.5]$	1.68
Tempera (Hard Panel)	8	62.3	$+4.2$ $[+0.5, +7.9]$	1.66

(Experiment 4.2). Table 21 reports visual score means across step buckets (0–9, 10–14, 15–19, and 20–29 turns).

Table 22 details the correlation between step count and visual score, alongside trajectory progression and regression rates.

Two critical empirical findings emerge from trajectory dynamics:

1. **Rapid Compute Saturation at 15–19 Steps**: Across all evaluated model configurations, visual quality peaks between turns 15 and 19 and plateaus thereafter, demonstrating that unguided multi-turn scaling cannot substitute for architectural decomposition.
2. **Widespread Trajectory Regression**: In 20% to 38% of rollouts, the final delivered script scored below its own intermediate peak checkpoint, proving that without rigorous stopping criteria, agents accumulate unprincipled local hotfixes that degrade overall quality.

A.6 AI ASSISTANCE STATEMENT

Per the ICLR 2027 AI Policy, the authors disclose the scope of LLM usage (Anthropic Claude and OpenAI ChatGPT):

- **Writing and Drafting:** LLMs assisted in text drafting, prose polishing, grammar correction, and LaTeX styling.
- **Retrieval and Discovery:** LLMs were queried as auxiliary discovery tools to find related work and verify citations.
- **Synthetic Data with Human Annotation:** LLMs helped draft raw candidate rubric criteria from archival metadata. Crucially, all criteria underwent rigorous human annotation, weighting, and double-blind arbitration by fine-art scholars; no unannotated synthetic data was admitted.
- **Excluded Roles (Ideation, Execution, Proofs):** No AI tools were used for research ideation, experimental execution, or mathematical proofs. All conceptual designs, benchmark mechanisms, security invariants ($\mathcal{I}_1$–$\mathcal{I}_7$), and empirical analyses were conceived, implemented, and verified directly by the human authors.

Table 20: Performance breakdown across visual genre and subject systems (Experiment 4.1). Residuals report score differences relative to model cohort means (95% CI).

Visual Genre & Subject System	**Cases** (n)	**Visual Mean**	**Within-Group Residual** (95% **CI**)
Narrative & Religious History	102	59.8	$-0.4\ [-1.9, +1.0]$
Landscape & Atmospheric Depth	31	60.4	$-0.8\ [-2.9, +1.4]$
Figurative & Portraiture	29	59.9	$+2.3\ [-0.4, +5.0]$
Nature, Still Life & Objects	25	58.6	$-1.7\ [-4.6, +1.2]$
Modernism & Abstract Composition	13	63.8	$+3.8\ [+0.8, +6.8]$

Table 21: Visual score progression across interaction step buckets (Experiment 4.2). Values in parentheses denote trajectory counts per bucket.

Model Configuration	0–9 **Steps**	10–14 **Steps**	15–19 **Steps**	20–29 **Steps**
Codex · GPT-6 Astra	63.7 (16)	71.5 (28)	**74.2** (42)	73.8 (38)
Codex · GPT-5.6 Sol	65.8 (32)	68.3 (72)	67.1 (31)	**69.2** (10)
Claude-Code · Fable 5.1	40.8 (65)	56.2 (37)	**65.7** (35)	63.0 (26)

- **Author Accountability:** The human authors assume full and sole responsibility for all contents, claims, and accuracy in this paper.

All annotators are either part of the author team or are paid labor compensation above the local minimum wage. All annotators were briefed on the research purpose, and all data collection and annotation procedures were approved by the authors' institutional review board (IRB).

Table 22: Trajectory correlation and intermediate degradation rates (Experiment 4.2). Final vs. Initial denotes trajectories whose delivered code improved over the opening draft; Final vs. Peak denotes rollouts where the final delivery regressed below an intermediate checkpoint.

Model Configuration	**Correlation ρ(Steps, Visual)**	**Final vs. Initial ($\% > r_0$)**	**Final vs. Peak ($\% <$ Peak)**
Codex · GPT-6 Astra	0.20	54%	36%
Codex · GPT-5.6 Sol	0.07	51%	20%
Claude-Code · Fable 5.1	0.63	31%	38%