VEvo-Harness: Causally-Guarded Self-Evolving Harness for Agentic Image Generation

Anonymous authors
Paper under double-blind review

Abstract

The central bottleneck in visual generation and editing has migrated from achieving pixel-level fidelity to satisfying densely constrained user specifications through increasingly agentic workflows. Prevailing agentic generation pipelines remain shallow generate-verify-rewrite loops treating VLM judgments as coarse visual feedback, either discarding repair experience or admitting ungated updates that poison skill library. With memorize and govern periods left open, they degrade into ungoverned iterators of repeated trial-and-error rather than continual self-evolving harness. Towards this issue, we present **VEvo-Harness**, the first causally-guarded continual self-evolving harness system for agentic visual generation. VEvo-Harness introduces four coupled innovations: (1) *a visual traceback* that densifies sparse judgments into root-cause attributions with a hierarchical verification tree; (2) *a visual continual governance* that admits only causally valid skill patches through counterfactual dual-arm image generation replay; (3) *a lifecycle-managed harness state* that hierarchically distills root-cause-anchored episodes across working, experience, and pattern tiers into a persistent skill library of routable and causally verified capabilities; (4) *a heterogeneous sandbox policy* that orchestrates a diverse action space of repair primitives, dynamically routing cost-aware interventions based on root causes to select the most effective strategy. VEvo-Harness surpasses state-of-the-art agentic methods and rivals proprietary generators (*e.g.*, GPT-Image, Nano-Pro) while maintaining the continual harness evolution and generalization.

1 Introduction

Recently, multimodal generation (Song et al., 2026; Hu et al., 2026; MiniMax-AI, 2026; Seedance et al., 2026) has evolved from research capability into general-purpose utility. Leading proprietary systems such as GPT-Image (OpenAI, 2026) and Nano Banana (Pichai, 2024), alongside prominent open-source frameworks like Qwen-Image (Wu et al., 2025) and Z-Image (Cai et al., 2025) achieve unprecedented levels of fidelity. As these models become ubiquitous creative tools, research attention shifts toward increasingly complex prompts. Modern generation requests steadily outgrow a single inference pass, as a growing fraction hinge on exact object counts (Kamath et al., 2025), attribute binding (Zhou et al., 2026), external knowledge (Pan et al., 2026a; Wang et al., 2026b), or multi-step reasoning (Wang et al., 2026c) beyond aesthetic description. The central bottleneck has thus shifted from pixel-level fidelity to satisfying multi-faceted specifications under densely coupled constraints.

To counter this escalating complexity, recent work shifts scaling from model weights to test-time computation, *i.e.*, prompt densification (Wang et al., 2024; Yang et al., 2026a) and semantic rewriting (Zhan et al., 2024; Yang et al., 2024). However, prompt-only interventions lack fine-grained attribution and pixel-grounded feedback, while newer multi-agent systems (Feng et al., 2026; Chen et al., 2026b;a; He et al., 2026b) surround generator with planning, web search, skill invocation, and refinement loops. Despite this progress, their harnesses function as read-only control-flow orchestrators. GEMS and Gen-Searcher augment generation via iterative verification or external retrieval, yet their stateless or statically curated harnesses discard repair trajectories and freeze capability maps at deployment. GenRouter evolves only bucket-level routing utilities while leaving the routed repair loops and their admission criteria untouched, so none of them possesses a governed write path through which verified experience become persistent capability. Moreover, sparse visual feedback makes such naive correlation-based accumulation self-deceiving, inducing harness-level hallucination. We thus ask *how the scaffolding of a frozen visual generator can consolidate cross-task repair experience into persistent state under a causally governed write path, without collapsing into such hallucination*?

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: *From stateless refinement to continual capability growth.* VEvo-Harness causally verifies and reuses visual experience, enabling frozen generators to become progressively stronger with usage.

The emerging agent harness (Kang et al., 2026; Karten et al., 2026b) treats the scaffold around models as a site of continual adaptation to accumulate capability. Extending this to visual synthesis, however, demands an explicit causal guarantee. In Fig. 1, we present **VEvo-Harness**, to our knowledge the first continual causally-guarded self-evolving harness for agentic image generation with four principles:

(1) **Visual traceback.** Unlike coding harnesses (DeepSeek-AI, 2026; OpenClaw, 2026) with line-level tracebacks, visual critics provide only coarse binary feedback (*e.g.*, pass/fail), because visual failures are continuous and multi-scaled. We construct a hierarchical verification tree that dynamically expands failed judgments into fine-grained failure modes (*e.g.*, attribute binding), yielding structured root-cause signals that anchor experience memory and causally verified skill updates. (2) **Visual continual governance**. To prevent harness hallucination, we introduce a visual continual evaluator that estimates each candidate patch's causal effect via counterfactual dual-arm replay under pinned skill routing, and admits only patches with positive estimated effects, thereby keeping the skill library causally validated. (3) **A lifecycle-managed harness state.** We organize persistent memory into working, instance, and pattern tiers, distilling root-cause-anchored repair episodes into reusable failure patterns that are causally verified before promotion into routable skills. The lifecycle enables the harness to accumulate causally verified experience and self-evolve monotonically across tasks. (4) **A heterogeneous sandbox policy.** Rather than hard-wiring the harness to a single prompt-rewriting tool, we expand its action registry into a heterogeneous sandbox, and route failures with a cost-aware policy based on visual traceback and persistent harness state. Our approach contributes threefold:

- We propose **VEvo-Harness**, the first causally-guarded self-evolving harness for agentic image generation. It recasts the scaffold around a frozen generator into a persistent, causally verified state space, establishing a new paradigm of harness-level capability accumulation.
- VEvo-Harness integrates four innovations, *i.e.*, counterfactual dual-arm replay for visual continual governance, hierarchical verification for precise failure attribution, lifecycle-managed state for experience accumulation, and sandbox routing policy for adaptive intervention.
- Extensive experiments across holistic generation and editing benchmarks show that VEvo-Harness continuously self-evolves via accumulated experience, yielding reset-free monotonic capability growth and zero-shot generalization to highly complex visual instructions.

2 Related Works

Agentic Image Generation. Recent work reframes image generation from text-to-pixel decoding as an agentic runtime, replacing prompt expansion and heuristic rewriting with planning, retrieval, verification, and iterative repair. Qwen-Image-Agent (Zhang et al., 2026b) casts generation as context construction, while Mind-Brush (He et al., 2026a), Gen-Searcher (Feng et al., 2026), and RS-Gen (Bian et al., 2026) use search and reasoning to ground knowledge-intensive prompts. CRAFT (Kovalev et al., 2025) and SCOPE (Ren et al., 2026) move toward structured control by maintaining explicit

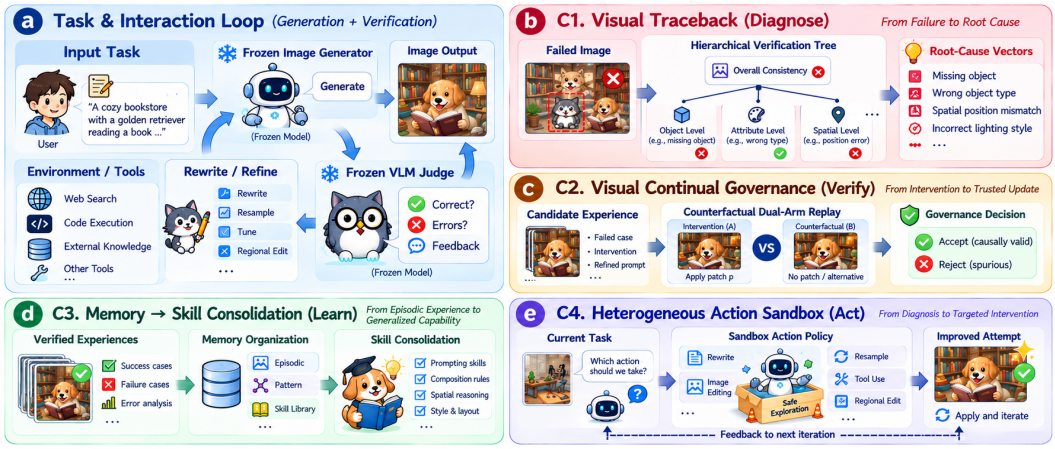


Figure 2: Overview of our proposed VEvo-Harness pipeline. A self-evolving scaffold is built around (a) a Interaction Loop for generation and verification, comprising (b) Visual Traceback, (c) Visual Continual Governance, (d) Lifecycle-Managed Harness State, and (e) Heterogeneous Action Sandbox.

visual constraints and invoking targeted repair only when violations arise. GenClaw (Ye et al., 2026) adds executable visual code as an intermediate canvas for layout control, whereas AR-RAG (Qi et al., 2026) performs autoregressive patch-level retrieval during decoding. GEMS (He et al., 2026b) and GenRouter (Chen et al., 2026a) extend this toward agent-native runtimes with persistent memory and adaptive workflow routing. They remain tied to fixed workflow families, leaving fine-grained visual attribution and harness-level governance of heterogeneous visual failures insufficiently addressed. Our VEvo-Harness, a unified visual generation scaffold, fill this gap via causally guarded self-evolution.

Agentic Harnesses and Scaffolding. Recent work treats harness as runtime infrastructure rather than prompt wrapper. Ning et al. (2026) identify tool orchestration, state management, and execution verification as core harness functions, and Karten et al. (2026a) emphasizes long-horizon performance depends on persistent substrates, *i.e.*, continual memory, durable state, reusable skills, and runbook control. Lou et al. (2026); Zhang et al. (2026a); Lee et al. (2026); Pan et al. (2026b) show substantial gains come from synthesizing, revising, or optimizing harness itself through environment feedback, regression-tested edits, or interface-side adaptation. Thangarajah et al. (2026) further clarifies that scaffold-specific planning conventions can hinder cross-scaffold transfer. Further extensions cover behavior localization, lifecycle governance, runtime mutation, and structured policy-environment interfaces Wang et al. (2026a); Garralda-Barrio (2026); Hou et al. (2026); Gan et al. (2026). Prior work focuses on coding, search, and web agents, leaving visual generation underexplored, where sparse feedback and stochastic execution demand causal tracing and persistent adaptation.

3 Method

3.1 Preliminary and Framework Overview

Formulation of Agentic Visual Generation. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the multimodal input conditions and generated image spaces. Agent-based visual generation expands standard single-step inference by introducing a foundational primitive space and orchestrating structured workflows. The generation process is formulated as $y = M(x, v_{\text{ref}}; \mathcal{H})$, where $v_{\text{ref}} \in \mathcal{Y}$ represents optional reference images, M is a frozen generation model, and $\mathcal{H}$ denotes the agent harness, formalized as a composite system $\mathcal{H} = (\mathcal{A}, \mathcal{M}, \mathcal{G}, \pi)$. Here, $\mathcal{A}$ is the execution sandbox, $\mathcal{G}$ is the governance mechanism, π is the routing policy, and $\mathcal{M} = (\mathcal{E}, \mathcal{S})$ denotes the lifecycle-managed state, comprising the episodic memory $\mathcal{E}$ and the persistent skill library $\mathcal{S}$. To handle complex visual synthesis, the harness orchestrates a directed workflow $w \in \mathcal{W}$ over a primitive set $\mathcal{O}$. In VEvo-Harness, accumulated experience continually updates $\mathcal{H}_t$, forming a self-evolving agent $\mathcal{A}_t = (M, \mathcal{H}_t, w)$ that generates image $y = \mathcal{A}_t(x, v_{\text{ref}})$.

Framework Overview. As illustrated in Fig. 2(a) and Algo. 1, VEvo-Harness forms a bidirectional closed loop between a frozen generation model M and an evolving harness $\mathcal{H}_t$, which integrates external tools (*e.g.*, web search, code execution) to drive iterative generation via targeted interventions. In the forward flow, a user prompt is first parsed into verifiable visual constraints. The routing policy π then queries the **lifecycle-managed harness state** (Sec. 3.3) $\mathcal{M}$ to retrieve failure-aware episodic traces from $\mathcal{E}$ and applicable heuristics from $\mathcal{S}$, which guide workflow construction and prompt enhancement. In the backward flow, failures of generated images are localized by **visual traceback**

(Sec. 3.2) and addressed via the **heterogeneous sandbox policy** (Sec. 3.5). The repaired trajectories are assessed by **visual continual governance** (Sec. 3.4), which promotes only causally validated gains into the persistent skill library and retains the others as episodic experience. With generate–verify–intervene–memorize loop, the harness continually evolves through recursive self-improvement.

3.2 VISUAL TRACEBACK

Agentic visual generation relies on a judge VLM to evaluate a set of compositional visual constraints c—spanning object counting, attribute binding, and spatial relations—against a synthesized image y_{cand}, yielding a sparse Boolean verdict vector $\mathbf{v} \in \{0,1\}^m$, where $v_j = \text{VLM}(y_{\text{cand}}, c_j)$. Crucially, these constraints are cached per prompt to ensure strict criterion alignment across all arms of the downstream counterfactual protocol. While sufficient for accept/reject decisions, binary signals obscure visual failure attributions (*e.g.*, spatial distortion or attribute misbinding), offering no actionable, pixel-aligned guidance for targeted intervention when $v_j = 0$. As shown in Fig. 2(b), we introduce the *Visual Traceback* mechanism, which densifies these non-differentiable global verdicts into structured, semantically grounded root causes, enabling precise diagnostic feedback for downstream repair.

Visual traceback densifies each failed boolean verdict into localized root causes via a per-constraint two-level verification tree $\mathcal{T}_j$, whose root encodes the coarse requirement c_j and expands lazily only when $v_j = 0$. Upon failure, the VLM is prompted to generate a bounded set of finer-grained sub-questions that disentangle the failure mechanism. This generation is explicitly guided to span orthogonal diagnostic dimensions—encompassing object existence, fine-grained attributes, and stylistic fidelity—while semantically redundant sub-questions are dynamically filtered. Unlike a predefined dimension checklist, this *failure-conditioned adaptive decomposition* tailors the diagnostic structure to the specific semantic nature of c_j (*e.g.*, relational constraints expand into spatial and existence checks). All sub-questions of all failed constraints are evaluated concurrently, bounding the diagnostic overhead to a single parallelized VLM round. The visual traceback process is defined as:

$$\mathbf{e} = \text{Traceback}(\mathbf{v}, y_{\text{cand}}, c) = \bigcup\nolimits_{j:v_j=0} \text{Leaves}^{-}(\mathcal{T}_j), \tag{1}$$

where $\text{Leaves}^{-}(\mathcal{T}_j)$ collects false leaf sub-questions, or c_j if none exist. Thus, $\mathbf{e}$ converts a scalar verdict into re-verifiable, pixel-grounded visual predicates actionable by subsequent sandbox policy.

3.3 LIFECYCLE-MANAGED HARNESS STATE

Conventional harnesses suffer from limited transfer in flat episodic buffers or degradation in append-only skill registries. As illustrated in Fig. 2(d), we thus architect a lifecycle-managed harness state $\mathcal{M} = \{\mathcal{M}_{\text{work}}, \mathcal{E}, \mathcal{P}, \mathcal{S}\}$ with a three-tier distillation pipeline. Working memory $\mathcal{M}_{\text{work}}$ buffers per-iteration prompts, pixel-level verifications, and structured root-cause vectors of the current task, flushing at task completion to prevent context leakage. Experience memory $\mathcal{E}$ archives trajectories that generate images passing compositional verification, enabling retrieval of refined prompts that previously resolved similar failures. Pattern memory $\mathcal{P}$ distills recurring visual failure modes into preventive strategies injected before rendering, shifting the cost from post-hoc repair to pre-emptive prevention. $\mathcal{S}$ consumes causally verified knowledge from this pipeline as routable, executable skills.

Each entry $\tau_j = (\mathbf{e}_j, \xi_j, p_j^*, \Delta_{\text{cf}}^{(j)})$ in $\mathcal{E}$ explicitly anchors a verified repair to its structured root-cause vector $\mathbf{e}_j$ rather than to raw prompt-image pairs, with ξ_j, p_j^*, and Δ_{cf} denoting the refinement sequence, verified prompt, and counterfactual causal utility (Sec. 3.4). Two-stage retrieval first recalls relevant candidates by root-cause similarity $r(\mathbf{e}_{\text{curr}}, \tau_j) = \text{sim}(\mathbf{e}_{\text{curr}}, \mathbf{e}_j)$ to match current failure mode, then reranks them by descending Δ_{cf} to return intervention with the highest validated causal gain.

As repair trajectories accumulate in $\mathcal{E}$, an induction operator distills recurring visual faults into pattern memory $\mathcal{P}$. Each pattern $\rho_k = (\mu_k, \phi_k, \sigma_k, \kappa_k)$ abstracts a cluster of trajectories sharing similar root-cause vectors. μ_k denotes the generalized failure mode, ϕ_k the preventive strategy, σ_k the empirical support count, and κ_k the dynamic confidence score. The confidence κ_k is modulated by usage feedback, which tracks the empirical reliability of the preventive strategy ϕ_k (*details in Appendix*). Ineffective patterns are progressively pruned, leaving only high-signal regularities in $\mathcal{P}$. Promotion to the skill library $\mathcal{S}$ is strictly gated by counterfactual verification. Post-admission, verified patches undergo conflict-aware consolidation via Exponential Moving Average (EMA), which permits state updates only when a new patch $\mathbf{s}_{\text{new}}$ satisfies semantic and causal criteria:

$$\text{Merge}(\mathbf{s}_{\text{new}}) \iff \text{Sim}(\phi_{\mathbf{s}_{\text{new}}}, \phi_{\mathbf{s}_{\text{exist}}}) > \theta_{\text{sem}} \wedge \Delta_{\text{cf}}(\mathbf{s}_{\text{new}}) > \Delta_{\text{cf}}(\mathbf{s}_{\text{exist}}), \tag{2}$$

where $\phi_{\mathbf{s}}$ is the preventive strategy's semantic embedding, which prevents redundancy while ensuring causal superiority over rendered images. When merged, the existing strategy is updated via $\phi_{\text{exist}} \leftarrow \alpha\phi_{\text{exist}} + (1-\alpha)\phi_{\text{new}}$ to balance historical stability with new evidence. The usage feedback reinforces or penalizes retrieved patterns based on pixel-level verifications. To prevent memory bloat, a capacity-



Figure 3: Qualitative results of VEvo-Harness with baselines on generation (*left*) and editing (*right*).

aware eviction policy prunes elements via a composite score of confidence decay and idleness:

$$\text{Evict}(x_k) = \arg\min_{x \in \mathcal{P} \cup \mathcal{S}} [\lambda_1 \cdot \kappa_x + \lambda_2 \cdot \text{idle}(x) + \lambda_3 \cdot \text{fail}(x)], \tag{3}$$

where x_k represents either a pattern $\rho_k \in \mathcal{P}$ or a skill $\mathbf{s}_k \in \mathcal{S}$, $\text{idle}(x)$ is the time since last reinforcement, and $\text{fail}(x)$ is the recent failure count. Elements falling below a threshold θ_{prune} are purged. Success preservation protects validated rules from later overwrites, while π jointly retrieves skills and root-cause evidence from $\mathcal{E}, \mathcal{P}$ to condition generation with verified instructions and episodic context.

3.4 Visual Continual Governance

Traditional harness updates based on text-level success rates, risking harness-level hallucination where textually plausible edits degrade visual fidelity. To address this, we propose Visual Continual Governance in Fig. 2(c), a causal admission protocol that certifies every candidate skill patch by counterfactually replaying it against the unpatched skill on rendered pixels so that only interventions with positive visual causal utility enter the skill library. Let $\mathcal{G}(\cdot)$ be the frozen generator and $\mathcal{R}_{\text{vis}}(\cdot)$ the reward verifying the image y against visual contract $\mathcal{C}_x$. The causal effect of a candidate patch ϕ on skill s_k is intervention $s_k \oplus \phi$, whose counterfactual utility is expected visual reward shift over z:

$$\Delta_{\text{cf}}(\phi \mid s_k, x) = \mathbb{E}_{z \sim p(\mathcal{Z})} \Big[\mathcal{R}_{\text{vis}}\big(\mathcal{G}(x, s_k \oplus \phi; z), \mathcal{C}_x\big) - \mathcal{R}_{\text{vis}}\big(\mathcal{G}(x, s_k; z), \mathcal{C}_x\big) \Big]. \tag{4}$$

After computing the causal utility Δ_{cf}, we apply a three-valued admission gate: verified ($\Delta_{\text{cf}} > \epsilon$), detrimental ($\Delta_{\text{cf}} < 0$), or uncertain ($0 < \Delta_{\text{cf}} < \epsilon$). Only verified patches are incorporated into the skill library. To avoid repeatedly attempting rejected edit directions, we introduce a denied-change buffer $\mathcal{D}$ that archives detrimental and uncertain patches. Proposals exhibiting semantic equivalence to any denied edits, quantified by $\max_{\phi' \in \mathcal{D}} \mathcal{J}(\phi, \phi') > \theta_{\mathcal{J}}$, are pre-emptively suppressed prior to the verification. It precludes redundant verification of rejected failure patterns. Post-admission, patches incorporated into skill library $\mathcal{S}$ undergo continuous monitoring to prevent the catastrophic forgetting. For each skill $s_k \in \mathcal{S}$, we track per-sample records indexed by (s_k, x) and compute pass ratio shifts $\delta_r(x) = r_t(x) - r_{t-1}(x)$ at epoch boundaries. When a run satisfies visual contract, correctly rendered instruction fragments are stored as protected rules ψ^+ in s_k, preventing later failure-driven updates from degrading verified content while repairing other defects. As patches accumulate, the skill library risks unbounded growth and conflicting visual priors. We regulate these updates via semantic conflict resolution and edit-budget scheduling to prevent visual degradation (**see Appendix A.2 for details**).

3.5 Heterogeneous Sandbox Policy

Existing pipelines (He et al., 2026b; Feng et al., 2026) adopt a monolithic intervention paradigm that attributes all visual failures to prompt misalignment, disregarding heterogeneous causal origins. To this issue, as shown in Fig. 2(e), we construct a heterogeneous action sandbox $\mathcal{A} = \{a_{\text{rewrite}}, a_{\text{resample}}, a_{\text{tune}}\}$ comprising distinct intervention types, each action intervening on a specific component of the image generation pipeline. Given the structured root-cause vector $\mathbf{e}$ from visual traceback (Sec. 3.2), the routing policy π selects actions via a softmax distribution parameterized by risk-sensitive Q-values:

$$\pi(a \mid \mathbf{e}) = \exp\Big(\hat{Q}(a, \mathbf{e})/\tau\Big) \Big/ \sum_{a' \in \mathcal{A}} \exp\Big(\hat{Q}(a', \mathbf{e})/\tau\Big), \quad \hat{Q}(a, \mathbf{e}) = \hat{\mu}(a, \mathbf{e}) - \lambda \cdot \hat{\sigma}^2(a, \mathbf{e}), \tag{5}$$

Table 1: Main results on general image generation benchmarks. All inference-time scaling methods are evaluated on the same frozen backends, where **bold** and underlined mark the best and second-best.

	Model	GenEval	GenEval2	DPG-Bench	OneIG-EN	OneIG-ZH	WISE	Avg.
	Nano-Banana Comanici et al. (2025)	0.75	44.6	85.23	0.550	0.405	–	–
	GPT-Image-1 Hurst et al. (2024)	0.84	–	85.15	0.533	0.474	0.80	–
	Seedream-4 Seedream (2025)	0.84	–	88.25	0.576	0.553	0.78	–
	Bagel Deng et al. (2025)	0.82	23.1	85.10	0.361	0.370	0.70	59.35
	Ovis-U1 Wang et al. (2025)	0.89	21.7	83.72	0.382	0.395	0.64	56.02
	Echo-4o Ye et al. (2025)	0.89	67.9	86.07	0.467	0.488	0.74	68.75
	Z-Image-Turbo Cai et al. (2025)	0.77	31.0	85.08	0.526	0.501	0.57	60.29
	Qwen-Image Wu et al. (2025)	0.87	33.8	88.32	0.539	0.548	0.62	65.09
	Inference-Time Scaling							
Z-Image-Turbo	Original (Cai et al., 2025)	0.77 ↑0.00	31.0 ↑0.0	85.08 ↑0.00	0.526 ↑0.000	0.501 ↑0.000	0.57 ↑0.00	60.29
Z-Image-Turbo	Rewrite (He et al., 2026b)	0.77 ↑0.00	39.7 ↑8.7	84.48 ↓0.60	0.520 ↓0.006	0.469 ↓0.032	0.84 ↑0.27	66.93
Z-Image-Turbo	Promptist (Hao et al., 2023)	0.76 ↓0.01	32.3 ↑1.3	65.69 ↓19.39	0.333 ↓0.193	0.327 ↓0.174	0.54 ↓0.03	52.20
Z-Image-Turbo	Search (Ma et al., 2025)	0.81 ↑0.04	43.5 ↑12.5	85.47 ↑0.39	0.530 ↑0.004	0.506 ↑0.005	0.61 ↑0.04	64.55
Z-Image-Turbo	Maestro (Wan et al., 2025)	0.82 ↑0.05	44.6 ↑13.6	85.29 ↑0.21	0.548 ↑0.022	0.519 ↑0.018	0.85 ↑0.28	70.05
Z-Image-Turbo	CRAFT (Kovalev et al., 2025)	0.80 ↑0.03	62.4 ↑31.4	85.29 ↑0.21	0.582 ↑0.056	0.542 ↑0.041	0.78 ↑0.21	72.38
Z-Image-Turbo	GEMS (He et al., 2026b)	0.86 ↑0.09	63.5 ↑32.5	86.01 ↑0.93	0.569 ↑0.043	0.552 ↑0.051	0.81 ↑0.24	74.51
Z-Image-Turbo	GenRouter (Chen et al., 2026a)	——	57.6 ↑26.6	85.49 ↑0.41	0.570 ↑0.044	0.545 ↑0.044	0.86 ↑0.29	68.12
Z-Image-Turbo	VEvo-Harness (Ours)	**0.89** ↑0.12	**67.8** ↑36.8	**86.72** ↑1.64	**0.594** ↑0.068	**0.563** ↑0.062	**0.88** ↑0.31	**76.28**
Qwen-Image-2512	Original (Wu et al., 2025)	0.66 ↑0.00	29.0 ↑0.0	84.69 ↑0.00	0.487 ↑0.000	0.489 ↑0.000	0.59 ↑0.00	57.50
Qwen-Image-2512	Rewrite (He et al., 2026b)	0.68 ↑0.02	39.8 ↑10.8	83.49 ↓1.20	0.465 ↓0.022	0.435 ↓0.054	0.81 ↑0.22	63.46
Qwen-Image-2512	Promptist (Hao et al., 2023)	0.66 ↑0.00	27.4 ↓1.6	64.02 ↓20.67	0.288 ↓0.199	0.307 ↓0.182	0.57 ↓0.02	48.83
Qwen-Image-2512	Search (Ma et al., 2025)	0.74 ↑0.08	42.1 ↑13.1	86.17 ↑1.48	0.502 ↑0.015	0.497 ↑0.008	0.67 ↑0.08	63.84
Qwen-Image-2512	Maestro (Wan et al., 2025)	0.76 ↑0.10	48.0 ↑19.0	84.07 ↓0.62	0.494 ↑0.007	0.489 ↑0.000	0.84 ↑0.25	68.24
Qwen-Image-2512	CRAFT (Kovalev et al., 2025)	0.79 ↑0.13	66.3 ↑37.3	85.87 ↑1.18	0.533 ↑0.046	0.518 ↑0.029	0.79 ↑0.20	72.54
Qwen-Image-2512	GEMS (He et al., 2026b)	0.79 ↑0.13	70.4 ↑41.4	85.59 ↑0.90	0.542 ↑0.055	0.532 ↑0.043	0.80 ↑0.21	73.74
Qwen-Image-2512	GenRouter (Chen et al., 2026a)	——	71.6 ↑42.6	87.39 ↑2.70	0.553 ↑0.066	0.541 ↑0.052	**0.88** ↑0.29	71.28
Qwen-Image-2512	VEvo-Harness (Ours)	**0.90** ↑0.24	**86.1** ↑57.1	**88.14** ↑3.45	**0.572** ↑0.085	**0.581** ↑0.092	0.86 ↑0.27	**77.59**

where $\hat{\mu}(a,\mathbf{e})$ and $\hat{\sigma}^2(a,\mathbf{e})$ are the empirical mean and variance of $\Delta_{\text{cf}}(a)$ accumulated over historical rollouts, τ is a temperature parameter controlling the exploration-exploitation tradeoff, and $\lambda \geq 0$ penalizes high-variance actions. The Q-value estimates are updated via EMA after each rollout:

$$\hat{\mu}(a,\mathbf{e}) \leftarrow (1-\alpha)\hat{\mu}(a,\mathbf{e}) + \alpha \cdot \Delta_{\text{cf}}(a), \quad \hat{\sigma}^2(a,\mathbf{e}) \leftarrow (1-\alpha)\hat{\sigma}^2(a,\mathbf{e}) + \alpha\left(\Delta_{\text{cf}}(a) - \hat{\mu}(a,\mathbf{e})\right)^2, \quad (6)$$

By matching each failure mode to its structurally appropriate intervention, the heterogeneous sandbox policy achieves targeted repair at the true causal origin, broader failure coverage across generation pipeline levels, and adaptive convergence toward consistently effective intervention strategies.

4 Experiments[1]

4.1 Experimental Setup

Implementation Details. We evaluate VEvo-Harness on Z-Image-Turbo and Qwen-Image-2512, using Qwen3.5-27B for the core reasoning backend, with at most 5 harness iterations. For visual traceback, each failed constraint expands up to 4 sub-questions judged by Gemini-3.7-Flash. For visual continual governance, Δ_{cf} uses $K = 4$ seed-fixed rollouts per arm; the denied buffer holds 50 edits, suppresses proposals with $\mathcal{J} > 0.6$, and refines verified patches within $\text{sim} = 0.45$. The edit budget cosine-decays from $\eta_0 = 4$ to $\eta_{\min} = 1$ over 20 steps, with slow updates every 25 runs. For sandbox policy, $(\tau, \lambda, \alpha, W) = (1.0, 0.5, 0.3, 5)$. Lifecycle consolidation starts after 10 episodes and 5 runs, distilling up to 5 patterns from the latest 30 episodes; patterns merge at Jaccard > 0.6, update confidence by ± 0.05, decay by 0.02, and are pruned below 0.15. Retrieval uses the top-3 patterns and top-2 instances; skill merging triggers at $|\Phi_k| > 8$ with $\theta_{\text{sem}} = 0.6$, and EMA coefficient $\alpha = 0.3$.

Benchmarks and Baselines. Following the protocol of GEMS (He et al., 2026b), we assess the VEvo-Harness on five general image generation benchmarks—GenEval (Ghosh et al., 2023), GenEval2 (Kamath et al., 2025), DPG-Bench (Hu et al., 2024), OneIG (Chang et al., 2026), and WISE (Niu et al., 2025)—and extend the analysis to downstream tasks (*e.g.*, creativity, spatial reasoning) on LongText-Bench (Geng et al., 2025), SpatialGenEval (Wang et al., 2026d), CREA (Venkatesh et al., 2026), and ArtiMuse (Cao et al., 2026). We also incorporate four editing benchmarks: GEdit-Bench (Liu et al., 2025), RISE-Bench (Zhao et al., 2026), KRIS-Bench (Wu et al., 2026), WiseEdit (Pan et al., 2026a).

4.2 Main Results on Image Generation and Editing Benchmarks

Performance on Image Generation Benchmarks. Tab. 1 shows that VEvo-Harness achieves the best average performance on both backends, reaching 76.28% on Z-Image-Turbo (Cai et al., 2025) and 77.59% on Qwen-Image-2512 (Wu et al., 2025). Gains are strongest on compositionally demanding benchmarks: on GenEval2, VEvo-Harness reaches 86.1% with Qwen-Image-2512, outperforming

[1]Cost-efficiency trade-offs, framework details, holistic ablations, and visualizations are in Append. A.2, A.3.

Table 2: Generalization results to downstream tasks comparison with inference-time scaling methods.

	Model	LongText-EN	LongText-ZH	SpatialGenEval	CREA	ArtiMuse	Avg.
Z-Image-Turbo	Original (Cai et al., 2025)	0.912 ↑0.000	0.932 ↑0.000	58.7 ↑0.0	11.84 ↑0.00	43.27 ↑0.00	58.41
	Rewrite (He et al., 2026b)	0.571 ↓0.341	0.495 ↓0.437	59.6 ↑0.9	16.75 ↑4.91	57.86 ↑14.59	56.65
	Promptist (Hao et al., 2023)	0.034 ↓0.878	0.070 ↓0.862	48.3 ↓10.4	14.00 ↑2.16	43.74 ↑0.47	35.98
	Search (Ma et al., 2025)	0.918 ↑0.006	0.937 ↑0.005	59.3 ↑0.6	12.98 ↑1.14	44.21 ↑0.94	59.88
	Maestro (Wan et al., 2025)	0.877 ↓0.035	0.807 ↓0.125	60.3 ↑1.6	15.81 ↑3.97	56.86 ↑13.59	63.52
	CRAFT (Kovalev et al., 2025)	0.951 ↑0.039	0.760 ↓0.172	60.6 ↑1.9	13.63 ↑1.79	54.95 ↑11.68	61.63
	GEMS (He et al., 2026b)	0.952 ↑0.040	0.940 ↑0.008	61.4 ↑2.7	22.55 ↑10.71	58.58 ↑15.31	72.44
	VEvo-Harness (Ours)	**0.961** ↑0.049	**0.948** ↑0.016	**62.8** ↑4.1	**25.73** ↑13.89	**60.42** ↑17.15	**75.62**
Qwen-Image-2512	Original (Cai et al., 2025)	0.854 ↑0.000	0.892 ↑0.000	60.0 ↑0.0	17.45 ↑0.00	59.94 ↑0.00	66.35
	Rewrite (He et al., 2026b)	0.563 ↓0.291	0.468 ↓0.424	60.3 ↑0.3	18.12 ↑0.67	62.23 ↑2.29	58.62
	Promptist (Hao et al., 2023)	0.030 ↓0.824	0.058 ↓0.834	48.7 ↓11.3	18.35 ↑0.90	60.65 ↑0.71	43.73
	Search (Ma et al., 2025)	0.883 ↑0.029	0.907 ↑0.015	60.6 ↑0.6	18.71 ↑1.26	59.96 ↑0.02	68.11
	Maestro (Wan et al., 2025)	0.833 ↓0.021	0.792 ↓0.100	61.0 ↑1.0	17.78 ↑0.33	61.62 ↑1.68	65.78
	CRAFT (Kovalev et al., 2025)	0.897 ↑0.043	0.787 ↓0.105	61.6 ↑1.6	15.61 ↓1.84	59.78 ↓0.16	64.40
	GEMS He et al. (2026b)	0.913 ↑0.059	0.931 ↑0.039	62.1 ↑2.1	24.01 ↑6.56	62.89 ↑2.95	74.31
	VEvo-Harness (Ours)	**0.926** ↑0.072	**0.944** ↑0.052	**63.5** ↑3.5	**27.68** ↑10.23	**64.75** ↑4.81	**77.93**

GenRouter (Chen et al., 2026a) by 14.5%, while Promptist (Hao et al., 2023) degrades by 19.3% on DPG-Bench (Hu et al., 2024) under Z-Image-Turbo. VEvo-Harness also leads on OneIG (Chang et al., 2026) (0.594/0.572) and WISE (Niu et al., 2025) (0.88/0.86), indicating root-cause decomposition, causal skill verification, and sandbox routing improve both compositional fidelity, retrieval-augmented synthesis, and knowledge-intensive reasoning. By distilling failure/success experiences into verified patterns, VEvo-Harness enables reliable knowledge reuse that improves cross-task generations. With the scaffolding of VEvo-Harness, Qwen-Image surpasses commercial Seedream-4 (Seedream, 2025) on GenEval (0.90), releasing more scaling potential than single-loop and routing-only optimization.

Performance on Downstream Tasks. The downstream tasks in Tab. 2 verify that capability growth resides purely in agent harness, with VEvo-Harness improving average by 17.21% and 11.58% over the original backends. Inference-time scaling methods, *e.g.*, Maestro improve certain metrics but degrade others, revealing that unguided search or experience accumulation triggers performance trade-offs across diverse constraints. VEvo-Harness avoids this by closing generate-verify-intervene-memorize loop around a persistent state space. It distills domain-specific preventive patterns for text rendering, spatial reasoning, and aesthetic fidelity from root-cause-anchored episodes, while counterfactual dual-arm replay further gates every patch before deployment keeping skill library causally validated. VEvo-Harness achieves state-of-the-art results, elevating CREA by 13.89% and 10.23%, and ArtiMuse by 17.15% and 4.81% on two backends. These gains show that accumulated harness state transfers zero-shot to specialized domains, confirming reset-free growth at harness layer.

Performance on Image Editing Benchmarks. As shown in Tab. 4, VEvo-Harness establishes state-of-the-art results among community models, surpassing proprietary systems on Temporal (RISE), Conceptual (KRIS), and GEdit. On RISE-Bench, it reaches 45.8 on Temporal and 20.0 on Logical, exceeding Know-Thinker by 1.1% and 4.7%, benefiting from the visual traceback and intervention routing. On KRIS-Bench, which spans factual, conceptual, and procedural knowledge, it attains 84.93, 84.29, and 85.26, outperforming Nano-Banana-Pro on Conceptual, since the lifecycle-managed state reuses distilled knowledge patterns across tasks. On WiseEdit, covering awareness, interpretation, and imagination, VEvo-Harness reaches 80.6, and scores 8.63 on GEdit, surpassing all community baselines. Success preservation protects validated source-image predicates, preventing identity drift and background degradation during localized edits. VEvo-Harness leads all knowledge-intensive metrics by up to 2.0%, showing causally verified harness evolution transfers effectively to editing.

Visual Comparison. Fig. 3 reports visualizations of VEvo-Harness and leading proprietary/agentic systems on generation and editing tasks. VEvo-Harness consistently advances complex infographic generation, precise text rendering, visual aesthetics, and knowledge-intensive edit. With scaffolded loop, VEvo-Harness generalizes to structurally dense visual instructions, confirming that reusable lifelong memory extend internal generation priors beyond the native capacity of frozen backbones.

4.3 Ablation Study and Discussion

Component Effectiveness and Execution Efficiency. Tab. 3a identifies lifecycle-managed state as the main driver of self-evolution, as removing persistent memory drops GenEval2/KRIS by 6.7%/6.2% and disabling skill consolidation still costs 2.5%/3.3%, confirming that retained experience compounds only after distillation into reusable skills. Counterfactual governance beats Admit-All by 2.9%/3.2% and exceeds the fixed checklist of Text-Only Gate by 2.2%/2.7% on all benchmarks, validating image-based verification before skill updates. HSP improves over Rewrite-Only by 4.0% on WISE and over random routing by 2.4% on KRIS, while progressive assembly raises the marginal

Table 3: Ablation studies of VEvo-Harness. (a) Component-level ablation analysis of Evo-Harness. (b) Micro-ablation of intra-module details, with VLM calls and wall-clock latency reported as cost.

(a) Component-level ablation of the VEvo-Harness system.

Variant	G-Eval	2 Wise	W-Edit	Kris
Component Ablation				
Lifecycle-Managed State				
Stateless Harness	79.4	0.81	76.2	78.6
w/o skill library $\mathcal{S}$	83.6	0.84	79.1	81.5
Visual Continual Governance				
Admit-All (no causal gate)	83.2	0.84	78.4	81.6
Text-Only Gate	84.5	0.85	79.3	82.8
Visual Traceback				
Flat Binary Verdict	81.8	0.83	77.5	80.2
Fixed-Dimension Checklist	83.9	0.85	79.0	82.1
Heterogeneous Sandbox Policy				
Rewrite-Only (monolithic)	82.4	0.82	78.0	80.8
Uniform Random Routing	84.1	0.84	79.2	82.4
Full VEvo-Harness (Ours)	**86.1**	**0.86**	**80.6**	**84.8**
Progressive Assembly				
Base harness	77.6	0.78	74.2	76.9
+ Visual Traceback	79.4	0.80	75.6	78.4
+ Heterogeneous Sandbox	81.2	0.82	77.1	79.8
+ Lifecycle-Managed State	83.4	0.84	78.9	82.0
+ Visual Cont. Governance	**86.1**	**0.86**	**80.6**	**84.8**

(b) The design-detail micro-ablation. VLM calls and wall-clock latency per task reported as cost.

Design detail	Avg.	Calls	Lat.(s)
Visual Traceback			
w/o lazy expansion	84.10	11.8	40.1
w/o redundancy filtering	83.62	6.6	39.2
w/o root fallback	82.94	6.2	39.6
Visual Cont. Governance			
w/o denied-change buffer	83.12	9.4	41.3
w/o success preservation	81.63	6.4	43.8
binary admission gate	81.90	6.5	42.5
Lifecycle-Managed State			
w/o working memory	83.31	6.8	40.6
w/o instance memory	80.42	8.1	45.2
w/o pattern memory	82.27	7.3	43.1
lexical retrieval	80.87	7.9	44.0
w/o Δ_{cf} reranking	82.58	6.9	41.7
w/o conf. decay	83.47	6.4	40.2
Heterogeneous Sandbox			
w/o variance penalty	82.41	7.4	42.8
w/o EMA Q-update	81.76	7.8	43.4
w/o tune action	82.13	6.7	41.9
Full VEvo-Harness	**84.38**	**6.2**	**38.4**

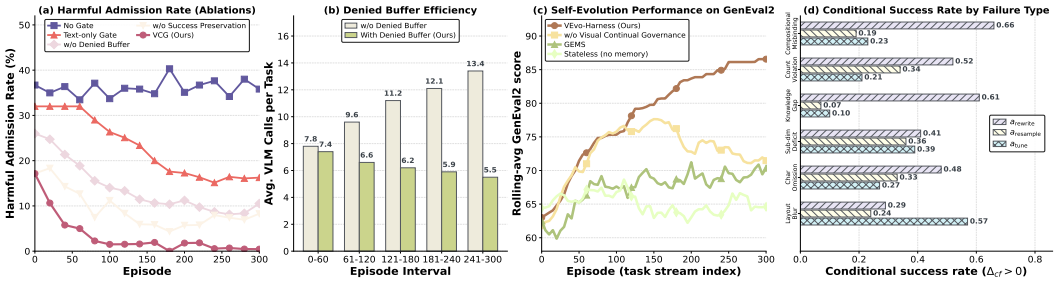


Figure 4: **Governance and Intervention Analysis.** (a) Harmful admission rate under governance. (b) Per-task verification cost w/. and w/o. denied-change buffer. (c) Continuous self-evolution results comparisons on GenEval2. (d) Conditional success rate of each sandbox action per failure type.

GenEval2 gain from 1.8% to 2.7%, showing the complementary value of action diversity and adaptive routing. Tab. 3b further shows that removing root-cause-anchored instance memory causes the largest degradation (-3.96%) with latency rising to 45.2s, counterfactual reranking contributes 1.80%, and EMA updates plus variance regularization add 2.62% and 1.97%. On cost, lazy gating cuts VLM calls 47.5% at a 0.28% gain, the denied-change buffer blocks redundant replay that would inflate calls 51.6%, and removing success preservation costs 2.75% and 43.8s at unchanged calls. The full setting is thus the most accurate (84.38%), fastest (38.4s), and cheapest (6.2 calls) variant.

Self-evolution via Governed Knowledge Accumulation. Figs. 4(a) and (c) link skill-update quality to sustained gains, as VEvo-Harness lifts its rolling-average GenEval2 score from 63.2 to 86.5 by distilling recurring failures into reusable skills. Without visual continual governance (VCG), gains collapse after a mid-stream peak near 78 to about 71.5, ending roughly 15% below the full model, since unverified updates pollute the skill library. Fig. 4(a) attributes this collapse to admission quality, where No Gate admits 33–40% harmful patches and Text-only Gate settles near 16%, whereas full VCG reaches about 2% by 100 episodes, rising to 10% and 8% without the denied buffer or success preservation. GEMS and the stateless variant finish near 70 and 64.5, confirming that cross-task reuse requires persistent state. Thus lifelong memory and causally-guarded governance are complementary, with persistent state enabling reuse and counterfactual replay filtering updates before propagation.

Governance Efficiency and Redundancy Suppression. Across 300 streamed episodes, the denied-change buffer progressively reduces verification overhead and harmful admission rates by filtering previously rejected edit directions (Fig. 4(b)). Without it, repeatedly evaluating structurally similar but previously rejected interventions inflates VLM calls by 71.8% over the task stream. With the buffer, the call count instead drops by 25.7%, yielding a 5.1% reduction in the first interval that widens to

Table 4: Evaluation on instruction-driven editing benchmarks, where **bold** indicates the best results.

Model	RISE-Bench			KRIS-Bench				WiseEdit	GEdit
	Temp.	Causal	Spatial	Logical	Factual	Concept.	Proced.		
Seedream-4.0 (Seedream, 2025)	12.9	12.2	11.0	7.1	39.26	—	—	75.2	7.68
Nano-Banana (Pichai, 2024)	25.9	47.8	37.0	18.8	80.48	78.11	75.31	75.0	6.51
Nano-Banana-Pro (Google, 2025)	41.2	61.1	48.0	37.6	86.36	84.03	85.31	82.4	7.73
GPT-Image-1 (Hurst et al., 2024)	34.1	32.2	37.0	10.6	79.80	81.37	78.32	77.6	7.53
OmniGen (Xiao et al., 2025)	1.2	1.0	0.0	1.2	33.11	28.02	23.89	36.6	5.00
EMU2 (Sun et al., 2024)	1.2	1.1	0.0	0.0	45.40	37.54	34.91	38.3	5.41
Step1X-Edit (Liu et al., 2025)	0.0	2.2	2.0	3.5	45.52	48.01	31.82	49.5	6.44
FLUX.1 (Batifol et al., 2025)	2.3	5.5	13.0	1.2	53.28	50.36	42.53	43.2	6.51
UniWorld (Lin et al., 2025)	—	—	—	—	47.71	44.80	47.92	49.4	7.83
Bagel (w/CoT) (Deng et al., 2025)	5.9	17.8	21.0	1.2	66.18	61.92	49.02	60.0	6.52
Uni-CoT (Qin et al., 2025)	8.2	18.9	20.0	1.2	71.85	71.85	63.68	60.1	6.74
Qwen-Image-Edit (Wu et al., 2025)	21.2	18.9	31.0	4.7	75.89	61.15	67.55	60.2	7.88
Know-Thinker (Wang et al., 2026c)	44.7	**52.2**	**46.0**	15.3	82.53	82.65	83.32	77.3	—
VEvo-Harness (Ours)	**45.8**	47.7	37.0	**20.0**	**84.93**	**84.29**	**85.26**	**80.6**	**8.63**

59.0% by episodes 241–300 and 41.6% on average (10.82 to 6.32). Fig.4(a) and the prolonged repair trajectories in Tab.3b show that removing success preservation degrades the admission curve: without shielding correctly rendered elements, the agent repeatedly breaks and re-fixes the same regions. Caching rejected patches further reduces redundant verification and lowers harmful admission rate.

Causal Matching of Heterogeneous Interventions. Fig. 4(d) reports $P(\Delta_{cf} > 0)$ for each sandbox action conditioned on failure type, exposing a strongly heterogeneous intervention–response surface. Conditioning-side failures favor *prompt rewriting* policy, with rewrite reaching 0.66 on compositional misbinding, whereas renderer-side defects invert this preference, as the layout blur responds more reliably to hyperparameter tuning (0.57) than to prompt rewriting (0.29). The sub-dim deficit failures do not admit a single dominant action and therefore require stochastic routing over action set. Matched actions outperform mismatched ones by up to **8.7x**, indicating that monolithic pipelines structurally bottleneck when a fixed intervention is applied across heterogeneous failure modes, while routing $\pi(a|\mathbf{e})$ over traceback root causes recovers per-type intervention specificity without exhaustive replay.

Generalization Across Base Models. Fig. 5 evaluates the transferability of VEvo-Harness across heterogeneous foundation backbones under identical configuration. VEvo-Harness consistently improves both lightweight and frontier models. On GenEval2, Qwen3-VL-8B increases from 74.6% without the VCG to 82.6% with full harness, while Kimi-K3 (Team et al., 2026) improves from 84.0% to 90.1%. A similar trend holds on KRIS, where Kimi-K3 rises from 87.3% to 89.4%. Significant gains indicate that VEvo-Harness improves a shared visual recovery runtime by self-evolving skills and causal visual governance, rather than model-specific biases. The larger GenEval2 gains further show that root-cause diagnosis and intervention routing transfer effectively across models, supporting VEvo as a robust visual-generation scaffold.

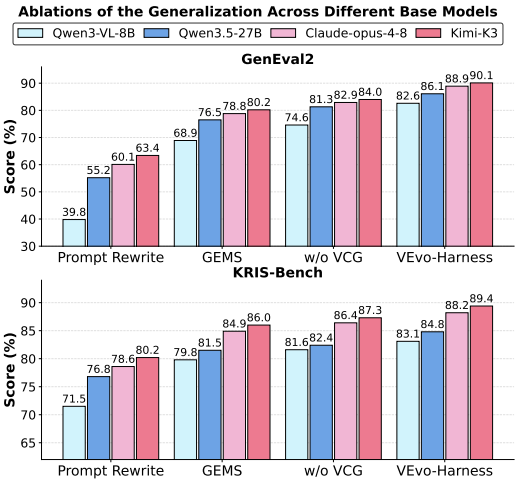


Figure 5: Generalization on different base models.

5 Conclusion

We introduce **VEvo-Harness**, a reset-free, causally guarded framework elevating the agent scaffold from a transient execution layer to a persistent, self-evolving runtime for agentic image generation. *Visual Traceback* employs a hierarchical verification tree to decompose visual judgments into localized root-cause vectors for fine-grained diagnosis. *Visual Continual Governance* uses counterfactual dual-arm replay to admit only causally validated patches while preserving verified solutions. *Lifecycle-Managed Harness State* structures memory into working, experience, and pattern tiers alongside a patch-evolvable skill library, distilling episodic experiences into reusable capabilities. *Heterogeneous Sandbox Policy* establishes a generalized action routing that dynamically aligns diagnosed root causes with diverse interventions via risk-aware utility estimation. Extensive results across general generation benchmarks confirm that VEvo-Harness achieves substantial gains over agentic pipelines. These results chart a path toward self-evolving agents in e-commerce AIGC workflows (*e.g.*, visual creative generation and virtual try-on), accumulating user-specific abilities through continuous interactions.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with recommended disclosure, including aesthetic refinement of the method framework figure based on the original version manually drawn by authors, editing manuscript to improve readability. We have not used generative AI for generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis. We also used generative AI tools for tasks with recommended disclosure, including editing the manuscript to improve readability and suggesting phrasing for method descriptions. We have reviewed all AI-assisted work. Every AI-assisted sentence was manually revised and verified by the authors against the source manuscripts, all numerical results and tables were produced solely by the authors' own experiments and cross-checked against the experimental logs, and all scientific claims and methodological designs originate from the authors, with AI involvement in visual artifacts limited to the disclosed framework-figure creation and refinement and all final figures manually verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT AND REPRODUCIBILITY STATEMENT

Ethics Statement. This work strictly adheres to the ICLR Code of Ethics. Our research is dedicated to enhancing the reliability of image generation through agentic loop, relying exclusively on publicly available benchmark datasets and existing foundation models. This study involves no human subjects and does not collect or process any private or sensitive data, thereby presenting no concerns regarding discrimination, bias, legal compliance, privacy, or security. We declare no conflicts of interest.

Reproducibility Statement. We have made substantial efforts to ensure the reproducibility. Complete details of the algorithmic formulations, training protocols, and hyperparameter settings, including the sensitivity analysis of parameters, are documented in the Appendix to facilitate exact replication of our empirical evaluations. The full source code of VEvo-Harness will be open-sourced after the review decision is announced to support future research on agentic visual generation.

REFERENCES

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Feifei Bian, Zhimin Zheng, Wei Deng, Daiguo Zhou, and Jian Luan. Rs-gen: A multi-stage agentic framework for reasoning and search-augmented image generation. *arXiv preprint arXiv:2606.23221*, 2026.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15313–15322, 2026.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Harold Haodong Chen, Zhiyu Hou, Wen-Jie Shu, Weilin Ruan, Yingjie Xu, Litao Guo, and Ying-Cong Chen. Genrouter: Unified workflow routing for agentic image generation. *arXiv preprint arXiv:2608.16721*, 2026a.

Sixiang Chen, Zhaohu Xing, Tian Ye, Xinyu Geng, Yunlong Lin, Jianyu Lai, Xuanhua He, Fuxiang Zhai, Jialin Gao, and Lei Zhu. Genevolve: Self-evolving image generation agents via tool-orchestrated visual experience distillation. *arXiv preprint arXiv:2605.21605*, 2026b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

DeepSeek-AI. Deepseek harness: Everything is a plugin. `https://github.com/deepseek-ai/deepseek-harness`, 2026.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*, 2026.

Chengguang Gan, Hanjun Wei, Yunhao Liang, Zhixi Cai, Qinghao Zhang, and Shiwen Ni. Mag: A web-agent benchmark and harness for multimodal action and guide generation. *arXiv preprint arXiv:2607.10079*, 2026.

Mariano Garralda-Barrio. Governed evolution of agent runtimes through executable operational cognition. *arXiv preprint arXiv:2605.27328*, 2026.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Google. Introducing nano banana pro. `https://blog.google/technology/ai/nano-banana-pro/`, 2025. b.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

Jun He, Junyan Ye, Zilong Huang, Dongzhi Jiang, Chenjue Zhang, Leqi Zhu, Renrui Zhang, Xiang Zhang, and Weijia Li. Mind-brush: Integrating agentic cognitive search and reasoning into image generation. *arXiv preprint arXiv:2602.01756*, 2026a.

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*, 2026b.

Yanning Hou, Haoyuan Chen, Sihang Zhou, Xiaoshu Chen, Xirui Liu, Duanyang Yuan, Lingyuan Meng, Quan Liu, and Jian Huang. Harness-g: A graph-structured harness for search agents. *arXiv preprint arXiv:2607.27652*, 2026.

Taihang Hu, Zhao Wang, Zuan Gao, Tao Liu, Hao Yan, Zhengze Xu, Yuhang Yu, Yongchao Du, Xingjian Wang, Jun Zheng, et al. Swift-image: Exploring the performance frontier of compact unified image generation models. *arXiv preprint arXiv:2608.20334*, 2026.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Borui Kang, Jinrui Gu, Junhan Lv, Wenbin Li, Lei Wang, and Yang Gao. Harness continual learning: Continual adaptation beyond model parameters. *arXiv preprint arXiv:2608.19013*, 2026.

Seth Karten, Alex L Zhang, Kevin Thomas, Sebastian Müller, Elie Bakouch, Daniel Auras, Mika Senghaas, Fares Obeid, Konstantin Dunas, Johannes Hagemann, et al. Prime agent: A self-improving rlm harness. *arXiv preprint arXiv:2608.23552*, 2026a.

Seth Karten, Joel Zhang, Tersoo Upaa Jr, Ruirong Feng, Wenzhe Li, Chengshuai Shi, Chi Jin, and Kiran Vodrahalli. Continual harness: Online adaptation for self-improving foundation agents. *arXiv preprint arXiv:2605.09998*, 2026b.

V Kovalev, A Kuvshinov, A Buzovkin, D Pokidov, and D Timonin. Craft: Continuous reasoning and agentic feedback tuning for multimodal text-to-image generation. *arXiv preprint arXiv:2512.20362*, 2025.

Hyunin Lee, Jinglue Xu, Jeffrey Seely, Donghyun Lee, Matei Zaharia, and Yujin Tang. Recursive harness self-improvement. *arXiv preprint arXiv:2607.15524*, 2026.

Hengjia Li, Liming Jiang, Qing Yan, Yizhi Song, Hao Kang, Zichuan Liu, Xin Lu, Boxi Wu, and Deng Cai. Thinkrl-edit: Thinking in reinforcement learning for reasoning-centric image editing. *arXiv preprint arXiv:2601.03467*, 2026.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xinghua Lou, Miguel Lázaro-Gredilla, Antoine Dedieu, Carter Wendelken, Wolfgang Lehrach, and Kevin P Murphy. Autoharness: improving llm agents by automatically synthesizing a code harness. *arXiv preprint arXiv:2603.03329*, 2026.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

MiniMax-AI. MiniMax H3. Official GitHub repository, 2026. URL `https://github.com/MiniMax-AI/MiniMax-H3`.

Xuying Ning, Katherine Tieu, Dongqi Fu, Tianxin Wei, Zihao Li, Yuanchen Bei, Jiaru Zou, Mengting Ai, Zhining Liu, Ting-Wei Li, et al. Code as agent harness. *arXiv preprint arXiv:2605.18747*, 2026.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

OpenAI. Gpt-image-2. OpenAI API Documentation, 2026. URL `https://platform.openai.com/docs/models/gpt-image-2`. Accessed: 2026-07-15.

OpenClaw. Openclaw: Your open-source ai agent platform. `https://openclaw.ai/`, 2026. Accessed: 2026-03-13.

Kaihang Pan, Weile Chen, Haiyi Qiu, Qifan Yu, Wendong Bu, Zehan Wang, Yun Zhu, Juncheng Li, and Siliang Tang. Wiseedit: Benchmarking cognition-and creativity-informed image editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 37268–37278, 2026a.

Wenbo Pan, Shujie Liu, Chin-Yew Lin, Jingying Zeng, Xianfeng Tang, Xiangyang Zhou, Yan Lu, and Xiaohua Jia. Evolving agents in the dark: Retrospective harness optimization via self-preference. *arXiv preprint arXiv:2606.05922*, 2026b.

Sundar Pichai. Introducing gemini 2.0: our new ai model for the agentic era. Google DeepMind Blog, December 2024. URL https://deepmind.google/blog/introducing-gemini-20-our-new-ai-model-for-the-agentic-era/. Accessed: 2025-xx-xx.

Jingyuan Qi, Zhiyang Xu, Qifan Wang, and Lifu Huangg. Ar-rag: autoregressive retrieval augmentation for image generation. *Advances in Neural Information Processing Systems*, 38:28514–28539, 2026.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025.

Tianfei Ren, Zhipeng Yan, Yiming Zhao, Zhen Fang, Yu Zeng, Guohui Zhang, Hang Xu, Xiaoxiao Ma, Shiting Huang, Ke Xu, et al. Scope: Structured decomposition and conditional skill orchestration for complex image generation. *arXiv preprint arXiv:2605.08043*, 2026.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Team Seedream. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, et al. Joyai-image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Kishanthan Thangarajah, Boyuan Chen, and Ahmed E Hassan. Dcas: Decoupling cli agent scaffolding to internalize planning across scaffolds. *arXiv preprint arXiv:2608.06113*, 2026.

Kavana Venkatesh, Connor Dunlop, and Pinar Yanardag. Crea: A collaborative multi-agent framework for creative image editing and generation. *Advances in Neural Information Processing Systems*, 38:171332–171392, 2026.

Xingchen Wan, Han Zhou, Ruoxi Sun, Hootan Nakhost, Ke Jiang, Rajarishi Sinha, and Sercan Ö Arık. Maestro: Self-improving text-to-image generation via agent orchestration. *arXiv preprint arXiv:2509.10704*, 2025.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

Ruhan Wang, Yucheng Shi, Zongxia Li, Zhongzhi Li, Yue Yu, Junyao Yang, Kishan Panaganti, Haitao Mi, Dongruo Zhou, et al. Harness handbook: Making evolving agent harnesses readable, navigable, and editable. *arXiv preprint arXiv:2607.13285*, 2026a.

Xingjian Wang, Zhao Wang, Taihang Hu, Jun Zheng, Qing Jin, Qinye Zhou, Zhengtao Wu, Yongchao Du, Zuan Gao, Chao Lin, et al. From corpora to co-evolving capabilities: Capability-centric data design for generalist image generation. *arXiv preprint arXiv:2608.18076*, 2026b.

Yuan Wang, Yongchao Du, Mengting Chen, Jinsong Lan, Xuetao Feng, and Xiaoyong Zhu. Expertverse: A general-purpose benchmark for expert-level reasoning in knowledge-intensive visual synthesis. *arXiv preprint arXiv:2607.19341*, 2026c.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*, 2026d.

Zhenyu Wang, Enze Xie, Aoxue Li, Zhongdao Wang, Xihui Liu, and Zhenguo Li. Divide and conquer: Language models can plan and self-correct for compositional text-to-image generation. *arXiv preprint arXiv:2401.15688*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *Advances in Neural Information Processing Systems*, 38, 2026.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Hanqing Yang, Qiang Zhou, Yongchao Du, Sashuai Zhou, Zhibin Wang, Jun Song, Tiezheng Ge, Cheng Yu, and Bo Zheng. Dda-thinker: Decoupled dual-atomic reinforcement learning for reasoning-driven image editing. *arXiv preprint arXiv:2604.25477*, 2026a.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, et al. Skillopt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026b.

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2img: Iterative self-refinement with gpt-4v for automatic image design and generation. In *European conference on computer vision*, pp. 167–184. Springer, 2024.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

Junyan Ye, Jun He, Zilong Huang, Dongzhi Jiang, Xuan Yang, Rui Chen, and Weijia Li. Genclaw: Code-driven agentic image generation. *arXiv preprint arXiv:2605.30248*, 2026.

Jingtao Zhan, Qingyao Ai, Yiqun Liu, Yingwei Pan, Ting Yao, Jiaxin Mao, Shaoping Ma, and Tao Mei. Prompt refinement with image pivot for text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 941–954, 2024.

Hangfan Zhang, Shao Zhang, Kangcong Li, Chen Zhang, Yang Chen, Yiqun Zhang, Lei Bai, and Shuyue Hu. Self-harness: Harnesses that improve themselves. *arXiv preprint arXiv:2606.09498*, 2026a.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, et al. Qwen-image-agent: Bridging the context gap in real-world image generation. *arXiv preprint arXiv:2606.26907*, 2026b.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *Advances in Neural Information Processing Systems*, 38, 2026.

Qinye Zhou, Jun Zheng, Yongchao Du, Yuan Wang, Zhengrui Chen, Zuan Gao, Taihang Hu, Chao Lin, Yefeng Shen, Xingjian Wang, et al. Cpi-bench: A comprehensive, practical and intelligent benchmark for real-world image editing. *arXiv preprint arXiv:2608.14546*, 2026.

A APPENDIX

A.1 OVERVIEW

This supplementary material provides comprehensive implementation details, rigorous theoretical analyses, and extensive controlled experiments to further substantiate the claims made in the main paper. In particular, we systematically examine whether the observed improvements genuinely arise from persistent harness evolution, whether the proposed visual continual governance reliably prevents harmful knowledge accumulation, and whether the resulting capability growth remains robust across computational budgets, task orders, hyperparameter configurations, and distribution shifts. The contents are organized as follows:

- **Implementation Details and Evaluation Protocols:** We provide the complete implementation details of VEvo-Harness, including the construction of visual contracts and hierarchical verification trees, lifecycle transitions across working, experience, pattern, and skill memories, counterfactual dual-arm replay for skill admission, and the risk-sensitive routing policy over heterogeneous sandbox actions. We further specify all hyperparameter configurations, memory capacities, retrieval and consolidation strategies, skill promotion and eviction criteria, replay budgets, generation settings, verifier configurations, and computational resources. Detailed benchmark protocols, evaluation metrics, baseline implementations, and compute-matching strategies are also provided to ensure reproducibility and fair comparison.
- **Theoretical Analysis of Governed Harness Evolution:** We provide theoretical analyses of the key mechanisms. We first show that paired counterfactual replay reduces the variance of estimated intervention utility relative to unpaired evaluation under positively correlated images. We derive finite-sample bounds on the probability of admitting detrimental skill patches and analyze the safety–plasticity trade-off induced by confidence- and margin-based admission criteria. We derive the risk-sensitive sandbox policy as the solution to an entropy-regularized mean–variance utility objective and analyze how root-cause-conditioned representations facilitate transferable intervention selection. Finally, we characterize how continual governance controls the accumulation rate of harmful persistent updates, providing theoretical justification for governed rather than unrestricted harness evolution.
- **Comprehensive Component and Mechanism Ablations:** We conduct extensive controlled ablations to isolate both the individual and synergistic contributions of Visual Traceback, Visual Continual Governance, Lifecycle-Managed Harness State, and Heterogeneous Sandbox Policy. Beyond component removal and progressive assembly, we perform factorial interaction analysis and fine-grained intra-module ablations covering adaptive traceback decomposition, root-cause fallback, denied-change suppression, success preservation, memory hierarchy, causal-utility reranking, confidence decay, variance-aware routing, and online Q-value estimation. These experiments establish how the four components jointly form the diagnose–intervene–verify–consolidate–reuse loop underlying VEvo-Harness.
- **Visual Traceback and Root-Cause Attribution Analysis:** We directly evaluate whether Visual Traceback provides actionable failure attribution rather than merely more verbose visual feedback. Using controlled visual failures with known diagnostic categories, we measure root-cause localization accuracy across object existence, counting, attribute binding, spatial relations, text rendering, layout, style, and renderer-side artifacts. We compare adaptive traceback against binary verification, fixed-dimensional checklists, and free-form VLM diagnosis, and further analyze traceback depth, branching factor, verifier strength, and diagnostic cost. We additionally compare root-cause-based retrieval with prompt-semantic retrieval to demonstrate that failure representations provide more transferable signals for downstream repair and intervention routing.
- **Visual Continual Governance and Counterfactual Reliability Analysis:** We investigate whether counterfactual governance prevents persistent-state contamination. We compare paired and unpaired replay, analyze replay sample size and admission thresholds, and report the variance and confidence of estimated counterfactual utility. Controlled harmful-experience injection experiments further stress-test the system under progressively corrupted candidate updates, while comparisons with Admit-All, text-only verification, and confidence-aware gates quantify harmful-patch admission and subsequent capability degradation. We

analyze the denied-change buffer and success-preservation mechanism in terms of redundant verification, false suppression, regression rate, and long-term governance efficiency.

- **Lifecycle Memory and Persistent Capability Analysis:** We provide detailed experiments examining how repair experience is transformed into reusable harness capability. We ablate individual memory stages, compare root-cause-anchored retrieval with prompt-level retrieval, evaluate alternative memory eviction strategies, and analyze sensitivity to experience, pattern, and skill capacities. We further study pattern support and confidence thresholds, conflict-aware consolidation, and the preservation of previously validated constraints. We introduce an *evolve–freeze–evaluate* protocol in which VEvo-Harness first evolves over an experience stream and is subsequently frozen before evaluation on unseen tasks, providing a controlled assessment of whether capability persists beyond online per-instance adaptation.
- **Heterogeneous Intervention and Routing Analysis:** We provide a comprehensive failure-type–intervention response matrix to characterize the heterogeneous causal utility of prompt rewriting, resampling, and generation-parameter tuning. We compare the proposed risk-sensitive router against monolithic actions, uniform random routing, heuristic routing, mean-only utility estimation, and alternative adaptive routing strategies. Sensitivity analyses over routing temperature and variance regularization further characterize the exploration–exploitation and risk–reward trade-offs, demonstrating that conditioning intervention selection on traceback-derived failure representations is essential for efficient targeted repair.
- **Reset-Free Continual Evolution and Generalization:** We conduct long-horizon experiments to determine whether VEvo-Harness genuinely accumulates reusable capability across tasks rather than repeatedly solving each instance from scratch. We compare stateless, episodic-only, unrestricted continual, and fully governed evolution under multiple task-stream permutations, and report continual-learning metrics including capability gain, forward transfer, backward transfer, forgetting, and preservation rate. We further evaluate cross-benchmark and cross-backbone harness transfer, as well as the performance of frozen evolved harnesses under unseen distributions, to characterize the extent to which accumulated repair knowledge generalizes beyond the environment in which it was acquired.
- **Hyperparameter Robustness, Statistical Reliability, and Computational Efficiency:** We perform systematic sensitivity analyses over the principal hyperparameters governing traceback decomposition, counterfactual replay, memory consolidation, denied-change suppression, and sandbox routing. Hyperparameters are selected on a designated development setting and subsequently frozen across evaluation benchmarks and model backbones. We report results over multiple random seeds and task orders with confidence intervals and paired statistical analyses. To disentangle architectural gains from additional test-time computation, we further provide compute-matched comparisons against inference-time scaling baselines and characterize the performance–latency, performance–VLM-call, and performance–generation-cost Pareto frontiers. We additionally analyze the amortized cost of continual evolution as reusable experience accumulates.
- **Additional Qualitative Results, Failure Cases, and Discussion:** We provide extensive qualitative visualizations tracing the complete evolution process from visual failure diagnosis to intervention, counterfactual verification, memory consolidation, and subsequent cross-task reuse. We also systematically examine representative failure modes, including verifier misdiagnosis, conflicting accumulated experience, rare and out-of-distribution failures, insufficient counterfactual evidence, generator capability ceilings, and distributional drift. Finally, we discuss the limitations of harness-level self-evolution, including its dependence on visual verifier reliability, additional inference-time overhead, historical-distribution bias in persistent memory, and the remaining gap between controlled interventional verification and complete causal identification of visual generation failures.
- **Broader Impact:** We discuss the broader implications of persistent self-evolving visual agents. VEvo-Harness provides a pathway for improving visual-generation systems through reusable runtime experience without modifying foundation-model parameters, potentially reducing repeated adaptation and enabling more personalized and continually improving visual-generation workflows. We also discuss potential risks associated with persistent agent memory and autonomous capability accumulation, including erroneous experience propagation, bias amplification, privacy of accumulated interaction traces, content provenance, and

inappropriate reuse of learned visual strategies, together with governance mechanisms and deployment practices for mitigating these risks.

A.2 Data, Metrics, and Implementation Details

A.2.1 Implementation Details.

We evaluate VEvo-Harness on three base generators that together cover both the generation and the editing regimes. The text-to-image suites are powered by Z-Image-Turbo (Cai et al., 2025) and Qwen-Image-2512 (Wu et al., 2025), while the four editing benchmarks run on Qwen-Image-Edit-2511 (Wu et al., 2025). All three models are loaded as in-process diffusion pipelines in bf16 precision, which allows the harness to intervene directly on seeds, schedulers, and per-step hyperparameters without any external serving overhead. Denoising follows the native schedule of each model, using 40 steps for Qwen-Image-Edit-2511 and 9 steps for the distilled Z-Image-Turbo. The entire reasoning and verification stack shares a single VLM backend that serves both traceback diagnosis and counterfactual judging, and each sample is allotted at most 5 harness iterations, reduced to 3 for the WiseEdit (Pan et al., 2026a) and KRIS-Bench (Wu et al., 2026) editing suites to reflect their shorter repair horizons.

Visual Traceback. Each candidate prompt is first decomposed into a set of atomic yes/no verification questions, one per constraint, so that every boolean verdict can be traced back to an explicit visual requirement. Any question that fails is expanded into up to 4 finer-grained sub-questions, forming a two-level verification tree that localizes the underlying failure mechanism rather than the surface symptom. To keep diagnosis latency bounded, up to 10 questions are batched and judged concurrently within a single reasoning VLM call, and a node passes only under the binary rule that the answer is an unambiguous "yes" rather than a "no", which prevents hedged or ambiguous responses from being counted as successes.

Pattern Confidence Modulation. To ensure the lifecycle-managed state remains both responsive to shifting distributions and reliable in its interventions, we employ two complementary mechanisms: dynamic confidence modulation for soft patterns and strict causal gating for hard skills. To prevent the pattern memory $\mathcal{P}$ from being saturated with obsolete or spurious strategies, the dynamic confidence score κ_k of each pattern ρ_k is continually modulated based on its empirical utility during downstream interventions. Specifically, κ_k is updated via a piecewise reward-penalty-decay mechanism:

$$\kappa_k \leftarrow \begin{cases} \min(1, \kappa_k + \delta_+) & \text{if } \rho_k \text{ is successfully reused,} \\ \max(0, \kappa_k - \delta_-) & \text{if } \rho_k \text{ fails during reuse,} \\ \max(0, \kappa_k - \lambda) & \text{if unreinforced within window } W. \end{cases} \tag{7}$$

where δ_+ and δ_- denote the step sizes for positive reinforcement and negative penalty, respectively. The third branch introduces a temporal decay factor λ over a sliding window of W tasks. This decay acts as an explicit forgetting mechanism, ensuring that patterns no longer relevant to the current task distribution gracefully fade out. This prevents "zombie patterns" from occupying the retrieval space and misleading the routing policy π, thereby maintaining a highly responsive and compact heuristic pool. While the pattern memory $\mathcal{P}$ stores soft preventive heuristics, the persistent skill library $\mathcal{S}$ stores hard executable repair primitives. To guarantee the absolute reliability of $\mathcal{S}$, we enforce a strict causal gating criterion for skill admission. A candidate skill patch $\mathbf{s}$ is consolidated into the library if and only if its counterfactual causal utility is strictly positive:

$$\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathbf{s}\} \quad \text{iff} \quad \Delta_{\text{cf}}(\mathbf{s}) > 0. \tag{8}$$

This condition ensures that $\mathbf{s}$ is promoted only when the VCG module confirms, via dual-arm replay, that the intervention yields a genuine causal treatment effect rather than a coincidental pass. By strictly filtering out updates with zero or negative causal deltas, this admission rule fundamentally immunizes the harness against skill library poisoning and harness hallucination, distinguishing our approach from conventional append-only experience buffers that blindly accumulate correlated noise.

Visual Continual Governance. The admission of a candidate patch relies on the counterfactual gain Δ_{cf}, which contrasts a patched treatment arm against an unpatched base arm over K seed-fixed rollouts per arm so that stochastic generation variance cancels out between the two arms. In our runs K is set to 4 by default and the lightweight arm uses 2 iterations, keeping each verification round to a small constant number of generations. The resulting scalar is discretized into a three-valued verdict, namely verified when Δ_{cf} is positive, harmful when it is negative, and uncertain when it is exactly

Algorithm 1 Continual Self-Evolution of VEvo-Harness

Require: Task stream $\{(x_i, v_{\text{ref},i})\}_{i=1}^{N}$, frozen generator M, initial harness $\mathcal{H}_0 = (\mathcal{E}_0, \mathcal{P}_0, \mathcal{S}_0, \mathcal{D}_0, \pi_0)$
Ensure: Evolved harness $\mathcal{H}_N$, generated images $\{y_i\}_{i=1}^{N}$

for task episode $i = 1$ to N **do** ▷ *Outer loop: Continual state evolution*
  Parse visual contract: $\mathcal{C}_i \leftarrow \{c_j\}_{j=1}^{m} \leftarrow \text{Parse}(x_i)$
  Retrieve context: $\mathbf{c}_{\text{ctx}} \leftarrow \text{Ret}_{\mathcal{M}_i}(x_i, \mathcal{C}_i)$ ▷ Query $\mathcal{E}_i, \mathcal{P}_i, \mathcal{S}_i$
  Initialize candidate: $y^{(0)} \leftarrow M(x_i, v_{\text{ref},i}; \mathbf{c}_{\text{ctx}})$
  // Inner loop: Agentic intervention & Visual Traceback
  for repair step $k = 1$ to $K_{\max}$ **do**
    Evaluate: $\mathbf{v}^{(k)} \leftarrow \mathcal{V}(y^{(k-1)}, \mathcal{C}_i) \in \{0,1\}^m$
    if $\mathbf{v}^{(k)} = \mathbf{1}$ **then** ▷ All constraints satisfied
      break
    end if
    Visual Traceback: Expand verification tree $\mathcal{T}_j$ for each $v_j^{(k)} = 0$
      Densify verdicts: $\mathbf{e}^{(k)} \leftarrow \bigcup_{j:v_j=0} \text{Leaves}^{-}(\mathcal{T}_j)$ ▷ Filter redundant sub-questions
    Sandbox Policy: Generate candidate proposals Φ for action a
      Pre-emptive suppression: Discard $\phi \in \Phi$ if $\max_{\phi' \in \mathcal{D}_i} \mathcal{J}(\phi, \phi') > \theta_{\mathcal{J}}$
      Select action: $a^{(k)} \sim \pi_i(\cdot \mid \mathbf{e}^{(k)})$ ▷ Risk-sensitive Q-values (Eq. 5)
    Intervene: $y^{(k)} \leftarrow a^{(k)}(y^{(k-1)}, x_i, \mathcal{H}_i)$ ▷ Execute primitive from $\mathcal{A}$
  end for
  Record trajectory: $\tau_i \leftarrow \{(\mathbf{e}^{(k)}, a^{(k)}, y^{(k)})\}_{k=1}^{K}$
  // Visual Continual Governance (VCG) & Causal Admission
  Compute counterfactual utility: $\Delta_{\text{cf}} \leftarrow \text{VCG}(\tau_i, M, \mathcal{C}_i)$ ▷ Dual-arm replay (Eq. 4)
  if $\Delta_{\text{cf}} > \epsilon$ **then** ▷ **Verified:** Causal admission gate
    Conflict-aware consolidation: If $\text{Sim}(\phi_{\text{new}}, \phi_{\text{exist}}) > \theta_{\text{sem}} \wedge \Delta_{\text{cf}}^{\text{new}} > \Delta_{\text{cf}}^{\text{exist}}$
      Merge into $\mathcal{S}_{i+1}$ via EMA: $\phi_{\text{exist}} \leftarrow \alpha\phi_{\text{exist}} + (1-\alpha)\phi_{\text{new}}$ ▷ Eq. 2
    Store protected rules ψ^{+} from correctly rendered fragments to prevent forgetting
  else ▷ **Detrimental** ($\Delta_{\text{cf}} < 0$) **or Uncertain** ($0 \le \Delta_{\text{cf}} \le \epsilon$)
    Archive rejected proposals in denied-change buffer: $\mathcal{D}_{i+1} \leftarrow \mathcal{D}_i \cup \{\tau_i\}$
  end if
  // Lifecycle-Managed State (LMS) Update ($\mathcal{H}_i \to \mathcal{H}_{i+1}$)
  Update episodic memory: $\mathcal{E}_{i+1} \leftarrow \mathcal{E}_i \cup \{\tau_i\}$ ▷ Anchor verified repair to $\mathbf{e}$
  Induce patterns: $\mathcal{P}_{i+1} \leftarrow \text{Distill}(\mathcal{P}_i, \tau_i)$
    Modulate confidence κ_k via reward-penalty-decay based on usage feedback
  Update policy: $\pi_{i+1} \leftarrow \text{UpdateQ}(\pi_i, \tau_i, \Delta_{\text{cf}})$ ▷ EMA update of $\hat{\mu}, \hat{\sigma}^2$ (Eq. 6)
  if $|\mathcal{P}_{i+1} \cup \mathcal{S}_{i+1}| >$ capacity **then** ▷ **Capacity-aware Eviction**
    Evict $x_k = \arg\min_x [\lambda_1 \kappa_x + \lambda_2 \text{idle}(x) + \lambda_3 \text{fail}(x)]$ ▷ Eq. 3
  end if
  Form next harness state: $\mathcal{H}_{i+1} \leftarrow (\mathcal{E}_{i+1}, \mathcal{P}_{i+1}, \mathcal{S}_{i+1}, \mathcal{D}_{i+1}, \pi_{i+1})$
end for
return $\mathcal{H}_N, \{y_i\}_{i=1}^{N}$

zero. The denied buffer retains the last 50 rejected edits in a first-in-first-out manner and suppresses any new proposal whose Jaccard overlap with a denied entry exceeds 0.6, which prevents the system from re-paying the verification cost of semantically equivalent failures. On the positive side, a verified patch is refined in place when its similarity to an existing skill reaches 0.45, and otherwise spawns a new skill, so the library grows only when genuinely new repair knowledge appears.

Edit Budget. To prevent aggressive harness updates from destabilizing previously acquired skills, we impose a bounded edit budget on each evolution step. Following the (Yang et al., 2026b), the budget controls the maximum number of candidate modifications that can be incorporated within a single iteration, serving as a lightweight regularizer on the magnitude of harness evolution. We initialize the per-iteration budget at $\eta_0 = 4$ and cosine-decay it to $\eta_{\min} = 1$ over the first $T_\eta = 20$ evolution steps:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min}) \left[1 + \cos\left(\pi \frac{\min(t, T_\eta)}{T_\eta} \right) \right]. \tag{9}$$

Since the number of admissible modifications is discrete, we discretize η_t when enforcing the budget. After T_η steps, the budget remains fixed at $\eta_{\min}$. This schedule permits broader exploration during early evolution, when the harness contains limited reusable experience and coordinated modifications can accelerate adaptation, while progressively favoring conservative updates as the accumulated skill library grows. Importantly, the budget only limits the number of modifications considered for incorporation; it does not override our governance criterion. We further adopt a *two-timescale update scheme* that separates local adaptation from global consolidation. At the fast timescale, causally verified patches are incorporated under the per-iteration budget above, allowing the harness to respond promptly to newly observed visual failure modes. At the slow timescale, a consolidation pass is performed every $T_c = 25$ runs over the accumulated verified patches and existing skills. The consolidation process identifies redundant or semantically overlapping modifications, reconciles potentially conflicting updates, and merges compatible patches into more compact and reusable skills.

Lifecycle Consolidation. Pattern-level memory consolidation is gated on sufficient experience and starts only after 10 episodes and 5 runs have been recorded. Each consolidation pass distills up to 5 patterns from the latest 30 episodes, keeping the abstraction window recent enough to track distribution drift in the task stream. Two patterns merge when their Jaccard overlap reaches 0.6, and every pattern carries a confidence score initialized at 0.5 that is reinforced or penalized by 0.05 upon each supporting or contradicting outcome. Patterns left unreinforced for 72 hours decay by 0.02 per check, and any pattern whose confidence falls below 0.15 is pruned, which implements the forgetting half of the lifecycle and keeps retrieval focused on currently valid knowledge.

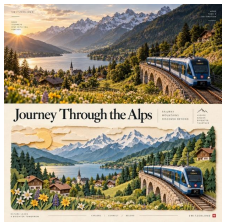
Retrieval and Skill Library. Retrieval returns the top-3 patterns together with the top-2 instance memories, scored by Jaccard token similarity by default while an optional Qwen3-Embedding cosine backend is supported for semantically richer matching, and the final ranking multiplies the similarity score by the stored confidence so that stale experience is automatically down-weighted. The skill library spans 3 categories and 8 atomic skills. The *visual quality* skill category contains `aesthetic_drawing` for aesthetic enhancement and artistic style and `creative_drawing` for open-ended imaginative composition. The *structural precision* category contains `spatial` for layout and object arrangement, `text_rendering` for accurate in-image text, `attribute_binding` for correct attribute-to-object assignment, and `quantity_counting` for exact object counts. The *physical plausibility* category contains `anatomy_body_coherence` for human and animal anatomical correctness and `physical_material_consistency` for material properties together with physical, temporal, and causal consistency. Category-level composition triggers once a category reaches the critical size of 6, per-skill verified patches merge once they exceed 8 entries and consolidate down to 4, and meta-skills merge at a semantic threshold θ_{sem} of 0.5. Finally, reference images are capped at 5 per sample to bound the context length, and the refiner is shown only the last 2 attempt images so that its edits condition on the most recent visual state.

A.2.2 Benchmarks and Baselines.

Following the protocol of GEMS (He et al., 2026b), we assess VEvo-Harness across a comprehensive spectrum of tasks to validate its generalization and reasoning capabilities. For a fair comparison of computational budgets across inference-time scaling systems, we set the parallelism factor of Search to 5 and cap the maximum number of iterations of Maestro, CRAFT, and GEMS at 3, 5, and 5.

General Image Generation. We employ five benchmarks to evaluate compositional fidelity and instruction adherence across diverse visual generation tasks. GenEval (Ghosh et al., 2023) and GenEval2 (Kamath et al., 2025) are designed to strictly stress-test structural precision and compositional generation. They evaluate the ability of models to handle exact object counting, precise attribute-to-object binding, and complex spatial relations without suffering from object duplication or omission. DPG-Bench (Hu et al., 2024) assesses fine-grained controllability by presenting lengthy and detail-rich prompts. It measures how well a model can simultaneously satisfy multiple specific constraints, such as intricate object descriptions, specific background elements, and nuanced stylistic requirements within a single generated image. WISE (Niu et al., 2025) focuses on world knowledge alignment and physical plausibility in synthesized scenes. It challenges models to generate images that not only match the textual description but also adhere to real-world common sense, correct physical interactions, and logical environmental contexts. Finally, OneIG (Chang et al., 2026) evaluates robust instruction following under highly constrained and multi-entity scenarios. It specifically targets the difficulty of maintaining semantic coherence when prompts require the precise arrangement and interaction of numerous distinct objects within a crowded visual space.

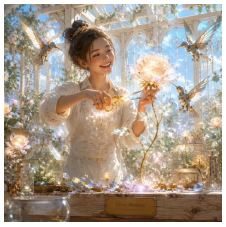
Figure I: **Visualization results generated by VEvo-Harness under long-context instructions**. The examples demonstrate its ability to follow complex, multi-level constraints involving visual composition, stylistic specifications, spatial relationships, and textual elements.



Create a sophisticated vertical travel poster divided into two visually connected panels, presenting the same breathtaking Swiss alpine landscape through contrasting artistic media. The upper panel should be a cinematic, wide-angle photograph of a sleek cobalt-blue train sweeping along curved railway tracks through a lush mountain valley, surrounded by dramatic snow-dusted peaks, scattered chalets, wildflower meadows, and tall evergreen trees. Warm late-afternoon sunlight should illuminate the valley with soft golden highlights, atmospheric haze, and an inviting sense of depth. The lower panel should faithfully reinterpret the identical composition as an intricate handcrafted 3D paper-cut artwork, using layered colored paper, tactile fibers, crisp cut edges, dimensional shadows, and carefully separated mountains, train, forests, clouds, and architecture. Place the artwork on a warm ivory textured background with generous negative space. Add elegant dark serif typography reading “Journey Through the Alps,” complemented by “RAILWAY / MOUNTAINS / DISCOVER BEYOND”



A whimsical 1:1 portrait of “Bloom & Petal,” a magical miniature flower shop filled with springtime warmth and joy. At the center, a cheerful young florist in a cream blouse and sage-green apron printed with “Bloom & Petal” runs forward laughing, holding a vintage watering pot high. Her messy bun is adorned with daisies, and a yellow scarf flutters at her neck. Surrounding her are five toddler-sized mini gardeners in oversized sun hats labeled “Little Gardener,” wearing pink or blue floral overalls. They carry giant props: a sunflower, terracotta pot, pruning shears, seed packet, and a jar of colorful seeds. Scattered soil on the floor spells “Growing Joy One Bloom at a Time ♡.” The backdrop features a hanging wooden sign reading "Bloom & Petal – Let Love Blossom ♡," blackboards with cheerful messages, and through an arched doorway, spring sunlight illuminates hanging baskets and a giant pink peony. Bright natural light creates soft bokeh with rich, saturated colors in fresh greens, soft pinks, and sunny yellows. Hyper-detailed textures blend storybook whimsy with cinematic realism.



A hyper-realistic, cinematic 1:1 portrait of a magical, sun-drenched glass conservatory. At the center, a vibrant young light-weaver gardener in a soft cream linen shirt and a translucent apron laughs joyfully, using glowing golden shears to prune a blooming "nebula rose" made of pure crystallized light. Surrounding her are floating, origami-like mechanical hummingbirds sprinkling luminous stardust instead of water. The floor is scattered with glowing petals, glass terrariums, and prismatic light refractions. The background is brilliantly bright and airy: massive arched glass windows reveal a clear, sunny sky, flooding the white-wood interior with warm morning sunlight and soft volumetric rays. A polished brass nameplate on the rustic wooden workbench clearly reads "VEvo-Harness". The scene features hyper-detailed textures: the delicate veins of the light-plants, dew drops catching rainbow spectrums, and soft fabric folds. The color palette is luminous and pastel—pearl white, sky blue, mint green, and warm gold—blending whimsical fantasy with breathtaking cinematic realism.



A hyper-realistic, high-end e-commerce product shot of the "VEvo-Harness", a futuristic lightweight tech-wear modular harness floating dynamically in mid-air. The design seamlessly blends translucent frosted polycarbonate panels, aerospace-grade anodized aluminum accents, and bio-based breathable mesh. Glowing micro-prismatic strips trace its ergonomic contours. The brand name "VEvo-Harness" is elegantly laser-etched in a minimalist sans-serif font on the central aluminum plate, catching a subtle golden rim light. The background is a bright, high-key studio setup featuring a soft, luminous pearl-white to pale ice-blue gradient, creating an airy, clean, and premium atmosphere. Surrounding the main product are floating, magnetically suspended modular components and tiny, crisp water droplets, highlighting its adaptive and waterproof nature. Lighting is masterful: soft diffuse global illumination combined with sharp edge lighting to accentuate material contrasts.



Create a premium vertical skincare key visual for VELORA DEW RESET SERUM. Center a realistic frosted-glass dropper bottle with embossed branding, polished silver details, subtle moisture, and crisp edge lighting on a pale aqua stone pedestal. Build an Ascension composition: the eye rises from pedestal to bottle, through a glowing serum droplet into a floating botanical crown, then descends along transparent water arcs. Surround with suspended droplets, aloe slices, white camellia petals, smooth mineral stones, and delicate green leaves in a pristine botanical studio. Integrate oversized, chunky, rounded, slightly inflated cartoon-inspired typography reading “VELORA,” “DEW RESET,” and “SKIN IN BALANCE.” Maintain playful editorial character, premium legibility, and unobstructed product focus. Airy background, soft atmospheric gradients, natural shadows, mist blue, mineral white, eucalyptus green, translucent aqua, and soft silver. Hyper-realistic beauty photography, playful luxury, botanical futurism, bold graphic advertising.



Create a premium ultra-high-resolution 1:1 vertical travel poster for REYKJAVIK. Build the enormous title “REYKJAVIK” across the central composition in bold modular sans-serif letters. Transform the letterforms into connected windows revealing Icelandic scenery: colorful Nordic houses, Hallgrímskirkja-inspired architecture, geothermal steam, black volcanic terrain, a coastal road, distant snow-covered mountains, a small harbor, and a dramatic northern sky. The scenes should flow naturally from letter to letter like one continuous illustrated panorama. Place a slim landscape strip above the title showing ocean, lighthouse, seabirds, mountains, and sweeping clouds. Adopt a Nordic modernist poster style with Bauhaus-inspired geometry, architectural diagram precision, clean flat vector forms, and expansive negative space. Use a beautiful misty off-white background with subtle grain and a cool palette of glacier blue, muted teal, volcanic gray, pale yellow, dusty coral, and deep navy. Include “REYKJAVIK / NORTH ATLANTIC / ICELAND” as minimal supporting text.



Create a refined, minimalist typography poster celebrating the Mid-Autumn Festival, using a monochromatic black-and-white palette with a poetic editorial atmosphere. Set the composition against an elegant warm-white paper background with subtle natural texture, soft moonlit shadows, and delicate silhouettes of osmanthus branches to enrich the scene without distracting from the typography. Feature the oversized Chinese characters “望月” in a bold, sophisticated Songti-style serif as the main visual anchor, paired with small uppercase text “MID-AUTUMN · MOONLIGHT” and the Chinese lines “月色渐满” and “思念循光而来.” In the middle, arrange the large characters “归家” beside the English phrase “UNDER THE SAME MOON,” separated by thin vertical rules and smaller labels such as “桂影,” “团圆夜,” and “故乡.” Use generous negative space, asymmetric alignment, delicate lines, and strong typographic hierarchy. Conclude with “共此一轮明月” above a subtle footer reading “MOONLIGHT / REUNION / AUTUMN,” creating a premium contemporary-art-poster aesthetic.

Figure II: **Qualitative ablations of Visual Traceback and knowledge-grounded generation. Left:** Naïve prompt rewriting introduces unintended changes, while Visual Traceback localizes violated constraints for targeted repair with stronger content preservation. **Right:** External retrieval improves task-specific knowledge grounding, while evolved memory further organizes the same retrieved knowledge through reusable strategies for attribute binding, spatial placement, and visual hierarchy.



Downstream Specialized Tasks. To investigate specific cognitive and aesthetic abilities, we extend the analysis to four targeted benchmarks. LongText-Bench (Geng et al., 2025) evaluates the robustness of visual generation under information-dense conditions by testing whether models can maintain global semantic coherence and preserve fine-grained details when processing prompts containing extensive modifiers and multiple nested constraints. SpatialGenEval (Wang et al., 2026d) focuses on geometric and physical understanding, isolating complex 3D spatial reasoning, precise relative positioning, and accurate occlusion handling among multiple interacting objects in a shared scene. Moving beyond structural accuracy, CREA (Venkatesh et al., 2026) and ArtiMuse (Cao et al., 2026) assess the maximum potential of generative creativity and aesthetic quality. These benchmarks challenge models to execute abstract conceptual fusion, interpret specific artistic styles, and adhere to advanced compositional principles, thereby measuring the capacity to produce visually appealing and conceptually original imagery without falling into common visual patterns.

Image Editing. We incorporate four rigorous editing benchmarks to assess multi-turn refinement and instruction following abilities. GEdit-Bench (Liu et al., 2025) provides a foundational assessment of edit accuracy and unedited region preservation, focusing on the ability to execute precise local modifications while maintaining global visual consistency. RISE-Bench (Zhao et al., 2026) introduces reasoning-intensive editing tasks that evaluate temporal, causal, spatial, and logical adherence, challenging models to resolve complex dependencies and physical constraints beyond simple text-to-image mapping. KRIS-Bench (Wu et al., 2026) assesses fidelity of factual, conceptual, and procedural knowledge injection, testing the capacity to accurately reflect real-world domain expertise and resist knowledge hallucination during complex edits. Finally, WiseEdit (Pan et al., 2026a) evaluates deep semantic understanding across awareness, interpretation, and imagination tasks, demanding that models navigate intricate multi-step refinements and abstract conceptual transformations without degrading the underlying structural integrity of the visual content.

A.3 More Experimental Results

A.3.1 More Visualization Results

Qualitative Results on Long-Context Visual Generation. Fig. I demonstrates VEvo-Harness on long-context generation tasks spanning travel posters, character-centered advertising, imaginative

Figure III: Qualitative comparisons on reasoning-intensive image editing tasks.

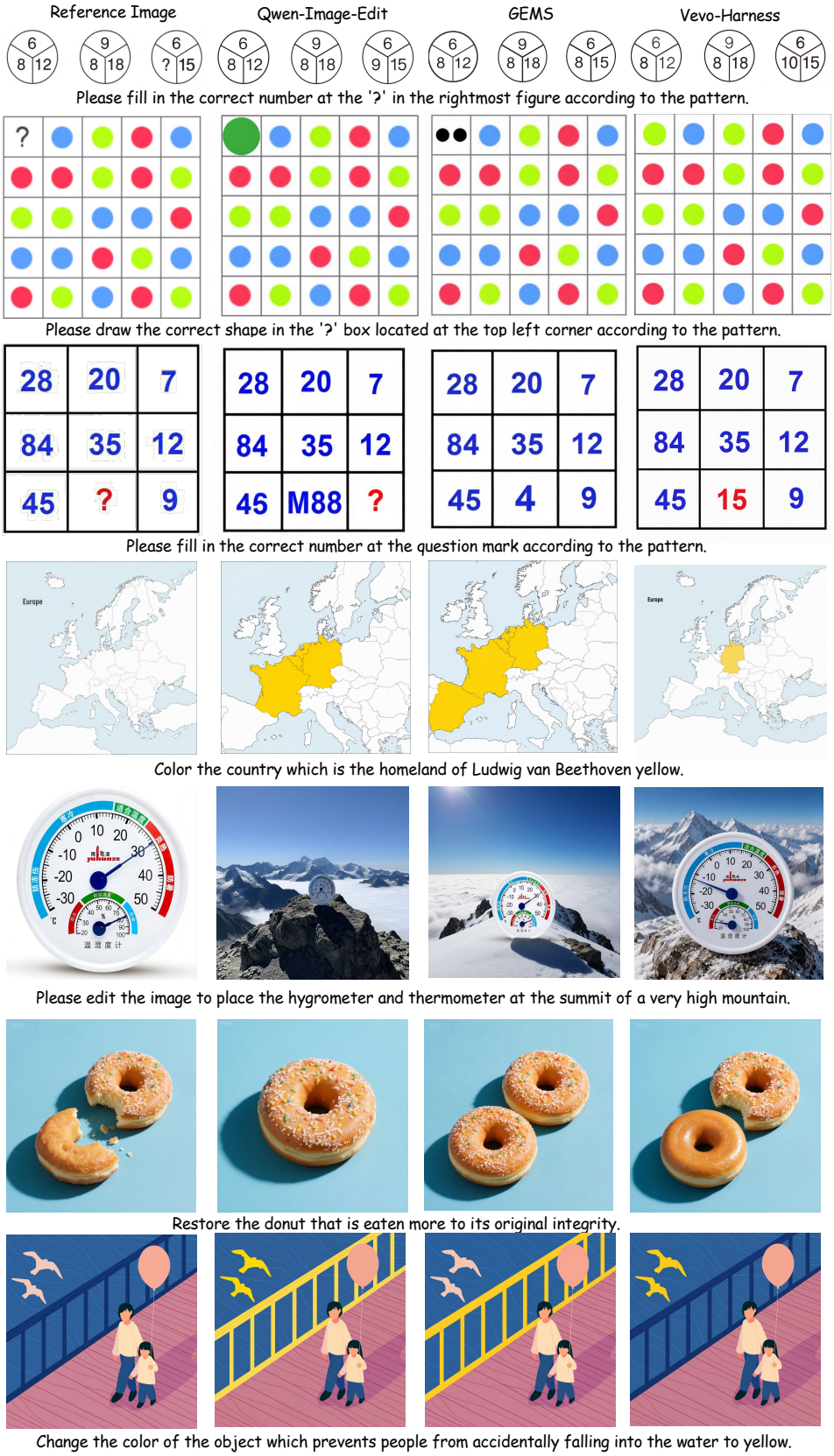
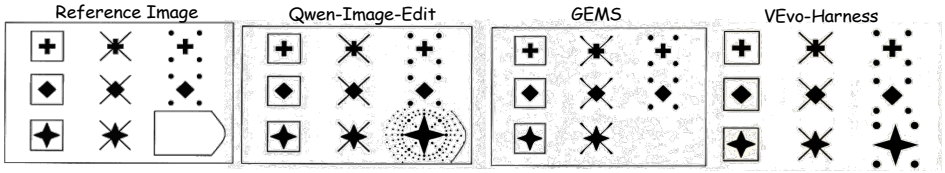
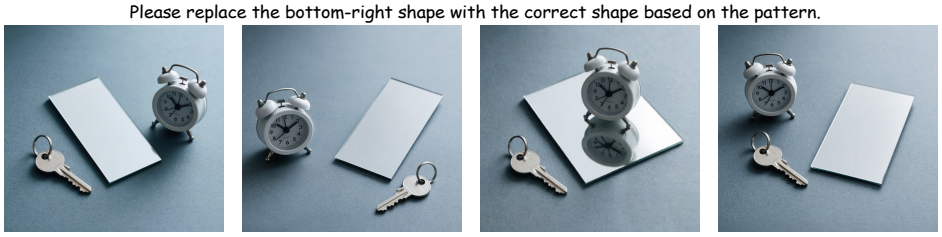


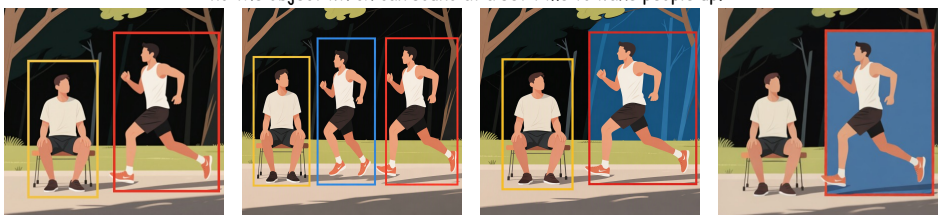
Figure IV: Qualitative comparisons on reasoning-intensive image editing tasks.



Please replace the bottom-right shape with the correct shape based on the pattern.



Swap the positions of the object which can reflect images for grooming or inspection a nd the object which can sound at a set time to wake people up.



Fill the box (between the yellow and red ones) which fits the person running better with blue.



Remove the person in the first image who comes from Portugal



Generate a 4-panel painting process for the character in the image: Step 1: Line art, Step 2: Flat color the character, Step 3: Add background color, Step 4: Refine and finalize. Do not add any text.



Please follow these steps to edit this image and return to the final version: 1. have one person fly a kite 2. make one bench wood 3. add a bag of popcorn and a group of ducks near the tree.



Please edit this image: 1. Turn the cake into a square 2. Add a candle to the cake 3. Replace strawberries with blueberries 4. Make plate transparent 5. Transform the painting style into watercolor.

Figure V: Qualitative comparisons on reasoning-intensive image editing tasks.

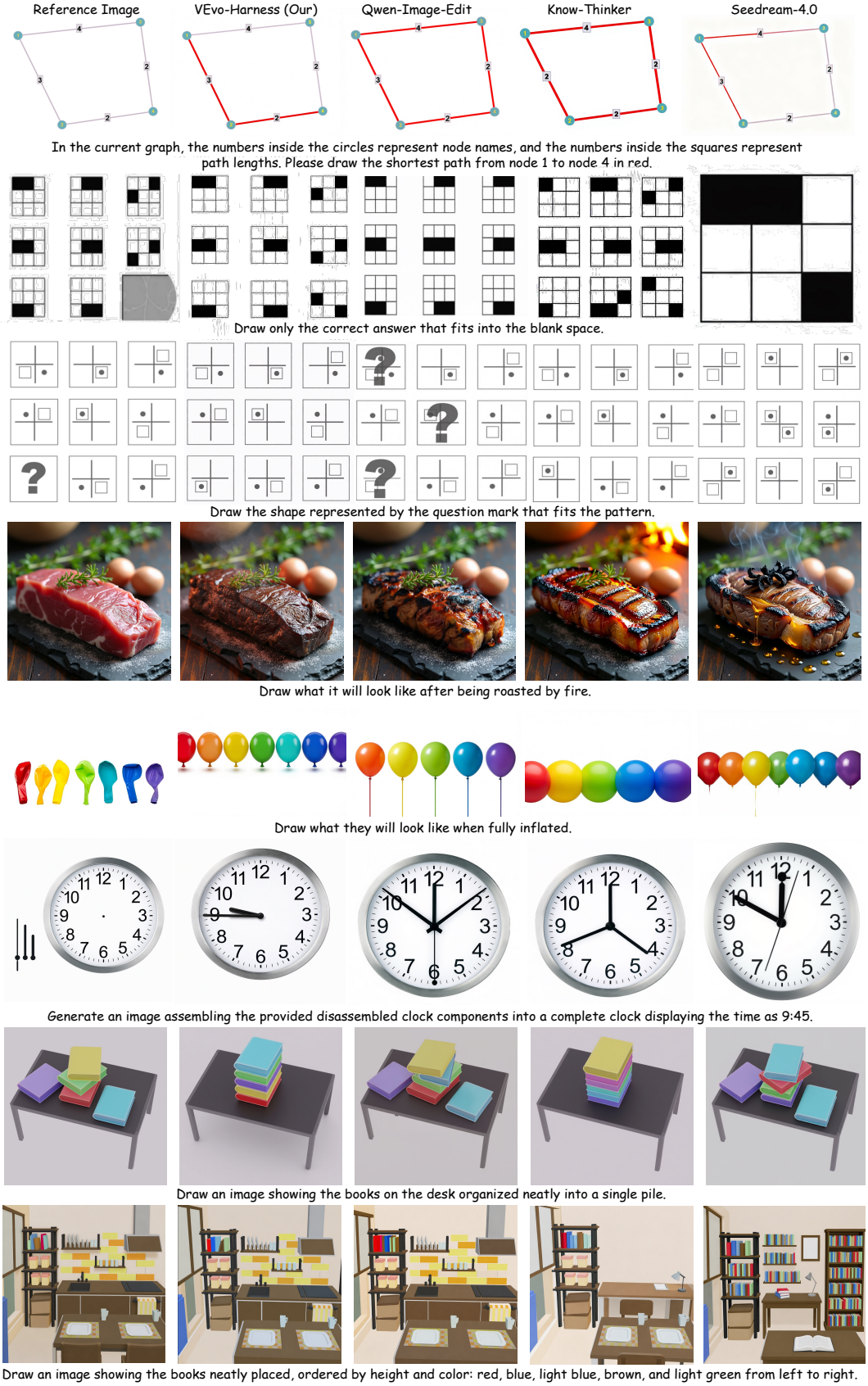


Table I: Quantitative result comparisons with proprietary and community image editing models on RISEBench benchmark. We report three general performance metrics: Instruction Reasoning (Reasoning.), Appearance Consistency (Consistency.), and Visual Plausibility (Visual.). Additionally, we present category-wise accuracy (%) for four specific reasoning dimensions: Temporal, Causal, Spatial, and Logical. The Overall score is the average of these four category-wise accuracies.

Model	**Reasoning.**	**Consistency.**	**Visual.**	**Temporal**	**Causal**	**Spatial**	**Logical**	**Overall**
Proprietary Models								
Seedream-4.0	58.9	67.4	91.2	12.9	12.2	11.0	7.1	10.8
GPT-Image-1-mini	54.1	71.5	93.7	24.7	28.9	33.0	9.4	24.4
GPT-Image-1	62.8	80.2	94.9	34.1	32.2	37.0	10.6	28.9
NanoBanana	61.2	86.0	91.3	25.9	47.8	37.0	18.8	32.8
NanoBanana-Pro	77.0	85.5	94.4	41.2	61.1	48.0	37.6	47.2
GPT-Image-2	73.8	89.3	94.9	45.9	66.7	50.0	34.1	49.4
GPT-Image-1.5	69.7	92.5	94.9	57.6	62.2	62.0	21.2	51.4
Community Models								
EMU2	22.6	38.2	78.3	1.2	1.1	0.0	0.0	0.5
OmniGen	25.1	41.5	73.5	1.2	1.0	0.0	1.2	0.8
Step1X-Edit	30.3	12.6	74.9	0.0	2.2	2.0	3.5	1.9
Ovis-U1	33.9	52.7	72.9	1.2	3.3	4.0	2.4	2.8
FLUX.1-Dev	26.0	71.6	85.2	2.3	5.5	13.0	1.2	5.8
BAGEL	45.9	73.8	80.1	5.9	17.8	21.0	1.2	11.9
Uni-CoT	–	–	–	8.2	18.9	20.0	1.2	12.5
ThinkGen	–	–	–	16.4	17.7	16.0	1.1	13.0
DeepGen	–	–	–	15.3	18.9	14.0	4.7	13.3
EditThinker	–	–	–	10.8	23.3	27.0	8.2	17.8
Unified Thinker	61.9	76.2	90.5	32.9	30.0	41.0	9.4	28.9
ThinkRL-Edit	61.7	81.6	62.5	18.8	37.5	25.0	37.5	29.7
Qwen-Image-Edit-2511	49.9	71.0	91.5	21.2	18.9	31.0	4.7	19.4
Know-Thinker	68.2	**82.6**	**93.7**	44.7	**52.2**	46.0	15.3	**40.0**
VEvo-Harness	**69.4**	81.6	91.8	**45.8**	47.7	37.0	**20.0**	37.8

Table II: Illustrative performance under shared resource ceilings on Qwen-Image-2512. Resource accounting covers the same 300 evolution tasks and 300 held-out evaluation tasks for every method.

Method	Image calls ↓	VLM calls ↓	Tokens ↓	Latency (s) ↓	GenEval2 ↑	DPG-Bench ↑	OneIG-ZH ↑
GEMS	4,180	3,900	3.600M	37.2	79.8 ± 0.8	86.8 ± 0.5	0.550 ± 0.009
GenRouter	4,195	3,300	3.780M	37.8	80.9 ± 0.7	87.6 ± 0.4	0.559 ± 0.008
VEvo-Harness	4,150	3,720	3.966M	38.4	84.0 ± 0.6	88.0 ± 0.4	0.570 ± 0.007

scenes, product visualization, skincare campaigns, and typography-driven graphic design. These examples are challenging not only because of prompt length, but also because each instruction specifies multiple coupled constraints over composition, content, material, typography, spatial organization, and visual style. For example, the alpine poster requires the same landscape to remain semantically consistent across photographic and paper-craft representations, while the skincare advertisement jointly constrains product placement, botanical elements, material appearance, text hierarchy, and an explicit ascending visual flow. The Reykjavik and Mid-Autumn posters further require large typography to function as part of the composition rather than as an isolated textual element. Across these heterogeneous cases, VEvo-Harness preserves the dominant visual concept while satisfying fine-grained local requirements, including object placement, scene structure, material rendering, palette, and text-aware layout. This behavior highlights the benefit of harness-level reasoning for long-context generation. Rather than treating a dense prompt as a monolithic instruction, VEvo-Harness decomposes it into verifiable visual constraints and combines external retrieval with persistent memory to reuse task knowledge and validated repair experience. When local constraints fail, Visual Traceback identifies root causes that guide experience retrieval and heterogeneous interventions. Effective repair trajectories are further validated through counterfactual visual verification and consolidated from instance-level episodes into reusable patterns and persistent skills. Memory thus evolves with the harness, turning external knowledge and task-specific experience into reusable generation capability. Across diverse visual domains, this process helps preserve global compositional coherence while

Table III: Monotonic capability growth and 95% Confidence Intervals across the 300-episode evolution stream. Values are averaged over 3 task orders × 3 random seeds (9 independent runs).

Metric	0 ep	50 ep	100 ep	150 ep	200 ep	250 ep	300 ep
GenEval2 (%)	77.60 ±.65	79.12 ±.71	80.85 ±.68	82.41 ±.74	83.74 ±.69	84.92 ±.72	86.10 ±.65
DPG-Bench	81.32 ±.42	82.54 ±.48	83.91 ±.45	85.23 ±.51	86.40 ±.47	87.35 ±.49	88.14 ±.42
KRIS (%)	76.21 ±.71	77.55 ±.75	79.12 ±.78	80.94 ±.82	82.41 ±.76	83.65 ±.79	84.83 ±.71
Probe Success (%)	64.80 ±1.2	68.25 ±1.3	72.54 ±1.2	76.42 ±1.4	79.85 ±1.3	82.31 ±1.2	84.70 ±1.2
Repair Success (%)	48.70 ±1.4	54.15 ±1.5	60.52 ±1.6	66.28 ±1.5	71.43 ±1.4	75.32 ±1.5	78.90 ±1.4
Avg. Iter. ↓	4.18 ±.15	3.85 ±.18	3.52 ±.16	3.21 ±.19	2.98 ±.17	2.75 ±.16	2.63 ±.15

Table IV: Zero-shot transfer of frozen harness skills to truly unseen distributions (OoD). The harness state is strictly frozen during this evaluation. ↑ denotes higher is better, ↓ denotes lower is better.

Unseen Distribution (OoD)	Base Generator	Frozen VEvo-Harness (300 ep)
Compositional and Spatial Reasoning		
T2I-CompBench (Spatial) ↑	54.20	61.80 (+7.60)
T2I-CompBench (Complex) ↑	48.50	55.10 (+6.60)
T2I-CompBench (Attribute Binding) ↑	62.30	69.40 (+7.10)
T2I-CompBench (Non-spatial Rel.) ↑	45.10	52.30 (+7.20)
Knowledge-Grounded Editing		
WiseEdit (Edit Success) ↑	42.50	58.20 (+15.70)
WiseEdit (Knowledge Accuracy) ↑	51.80	63.50 (+11.70)
Physical and Hallucination Suppression		
Physical Violation Rate ↓	22.60	11.80 (−10.80)
Hallucination Rate ↓	28.40	14.10 (−14.30)

satisfying fine-grained local constraints, enabling VEvo-Harness to continually retrieve, reuse, verify, and refine knowledge beyond one-shot prompt following.

Effect of Visual Traceback. The left panel of Fig. II probes whether the VEvo-Harness can enforce a localized visual constraint without perturbing content that already satisfies the instruction. Unguided prompt rewriting frequently realizes the requested semantic edit yet formulates the image as a global re-synthesis problem, thereby inducing collateral drift across layout, background, object appearance, and surrounding attributes. Such off-target perturbation is evident in both the product-recoloring probes and the peach-to-blueberry substitution, where the rewritten outputs capture the primary edit intent while diverging substantially from the source frame. Visual Traceback instead localizes the violated visual predicate through pixel-level causal attribution and scopes the subsequent generate-verify-intervene loop to the corresponding failure region. The repaired outputs consequently modify only the requested attribute while retaining markedly more of the original product identity, object placement, typography, background, and compositional structure. This comparison establishes that traceback functions as more than an auxiliary verification pass, since its principal contribution lies in recompiling a global edit request into a spatially bounded repair program that suppresses redundant full-frame regeneration and shields all previously satisfied visual constraints.

Effect of External Retrieval and Evolved Memory. The right panel of Fig. II decouples access to external visual knowledge from the harness capacity to compose such knowledge reliably during generation. Deprived of retrieval, the backbone still yields visually compelling images yet falls back on frozen parametric priors, producing stereotyped or weakly grounded instantiations of the requested regional identity, including temporally inconsistent Shanghai streetscapes, non-Icelandic terrain in the Reykjavík campaign, and generic Japanese cues for the Kyoto product. External retrieval injects task-specific historical, geological, and cultural evidence into the conditioning stream, yielding markedly more recognizable Shanghai, Icelandic, and Kyoto-specific visual elements. Retrieved facts alone nevertheless remain unstructured and cannot by themselves resolve attribute binding, spatial allocation, or saliency prioritization within the final composition. The evolved-memory variant consumes the same retrieved evidence together with reusable skills distilled from previously verified generation and repair trajectories, delivering sharper figure-ground segregation between focal subjects and contextual evidence, spatially coherent allocation of regional elements, and tighter typographic and structural integration. This comparison therefore establishes two orthogonal capability channels,

Table V: Quantitative performance evaluation on the KRIS-Bench benchmark. Models are comprehensively assessed across three foundational knowledge types: **Factual Knowledge** (Attribute, Spatial, Temporal), **Conceptual Knowledge** (Social, Natural), and **Procedural Knowledge** (Logical, Instruction), alongside the **Overall Average** score.

Model	Factual Knowledge				Concept. Knowledge			Proced. Knowledge			Overall Average
	Attri.	Spat.	Temp.	Avg.	Social	Natural	Avg.	Logical	Instruction	Avg.	
Proprietary Models											
Gemini 2.0	66.33	63.33	63.92	65.26	68.19	56.94	59.65	54.13	71.67	62.90	62.41
NanoBanana	78.97	80.17	86.49	80.48	83.70	76.34	78.11	63.96	90.44	75.31	78.17
GPT-4o	83.17	79.08	68.25	79.80	85.50	80.06	81.37	71.56	85.08	78.32	80.09
NanoBanana-Pro	85.61	87.17	88.06	86.36	85.25	83.64	84.03	81.42	93.22	86.48	85.31
Community Models											
InstructPix2Pix	30.33	21.33	–	23.33	22.56	26.56	25.59	19.81	14.75	17.28	22.82
OmniGen	37.92	28.25	21.83	33.11	30.63	27.19	28.02	11.94	35.83	23.89	28.85
MagicBrush	53.92	39.58	–	41.84	42.94	38.06	39.24	30.00	23.08	26.54	37.15
AnyEdit	47.67	45.17	–	39.26	38.56	42.94	41.88	36.56	26.92	31.74	38.55
Emu2	51.50	48.83	22.17	45.40	34.69	38.44	37.54	24.81	45.00	34.91	39.70
Step1X-Edit	55.50	51.75	–	45.52	44.69	49.06	48.01	40.88	22.75	31.82	43.29
HiDream-E1	52.75	49.42	–	43.31	52.56	49.25	50.05	45.19	30.08	37.64	44.72
ByteMorph	61.17	62.00	–	51.27	45.50	47.38	46.92	32.00	31.33	31.67	44.85
FLUX.1-Dev	64.83	60.92	–	53.28	48.94	50.81	50.36	46.06	39.00	42.53	49.54
OmniGen2	59.92	52.25	54.75	57.36	47.56	43.12	44.20	32.50	63.08	47.79	49.71
UniWorld-V1	58.17	54.50	63.00	47.71	47.50	43.94	44.80	42.00	53.83	47.92	50.27
Step1X-Edit v1.1	64.17	61.75	–	53.05	52.06	55.06	54.34	52.56	36.75	44.66	51.59
BAGEL	64.27	62.42	42.45	60.26	55.40	56.01	55.86	52.54	50.56	51.69	56.21
BAGEL-Think	67.42	68.33	58.67	66.18	63.55	61.40	61.92	48.12	50.22	49.02	60.18
Uni-CoT	72.76	72.87	67.10	71.85	70.81	66.00	67.16	53.43	73.93	63.68	68.00
ThinkRL-Edit	81.02	81.45	–	81.13	75.67	71.25	72.31	49.07	79.71	57.44	71.65
EditThinker	78.48	73.83	–	77.24	76.20	70.69	72.02	65.23	66.89	65.94	71.91
Qwen-Image-Edit	74.85	84.50	68.13	75.89	66.10	59.54	61.15	58.42	79.61	67.55	67.14
Know-Thinker	79.78	86.93	82.53	83.08	84.72	80.57	82.65	74.77	91.87	83.32	83.02
VEvo-Harness	80.13	87.54	83.92	84.93	85.50	83.17	84.29	76.93	93.48	85.26	84.83

in which external retrieval supplies task-specific factual grounding while evolved memory supplies the reusable composition strategies that compile such evidence into structurally coherent generations.

Qualitative Results on Reasoning-Intensive Editing. Fig. III, Fig. V, and Fig. IV reports qualitative results on reasoning-intensive editing, a harness-stress setting in which instructions encode constraints rather than explicit targets, forcing latent target resolution through pattern induction, world-knowledge retrieval, comparative judgment, or functional-causal inference before any pixel-level intervention. Across the six probes, VEvo-Harness compiles each implicit constraint into an executable edit specification, completing matrix-governed numeric and shape elements, resolving the Beethoven homeland to Germany prior to region masking, re-siting the hygrometer and thermometer at the summit with physically consistent readings, restoring exclusively the more-consumed donut instance, and recoloring the fall-prevention barrier instead of saliency-dominant distractors. Competing harnesses instead copy adjacent cells, mask arbitrary countries, or perturb the wrong instance, exposing saliency-bound instruction grounding that bypasses premise verification entirely. Within our loop, pixel-level visual traceback attributes such errors to the premise-construction stage rather than the renderer and routes them through the heterogeneous sandbox policy to premise-externalizing rewrites, while resampling and hyperparameter tuning remain reserved for renderer-side artifacts, and counterfactual dual-arm replay gates each patch so that regions satisfied by earlier interventions are shielded from subsequent updates. Together these cases establish that root-cause-anchored attribution over the reasoning chain, coupled with governed intervention admission at the harness layer, converts implicit instructions into verifiable edit programs without demanding stronger backbone priors.

A.3.2 More Benchmarking Results

Performance on RISEBench. On the RISEBench (Zhao et al., 2026) benchmark, which rigorously evaluates complex reasoning and editing capabilities, VEvo-Harness demonstrates substantial and synergistic improvements over both the base generator and state-of-the-art community models.

Table VI: Main results on the English version of WiseEdit Pan et al. (2026a) benchmark. The best results are marked in bold for open- and closed-models, respectively. Notably, our KnowThinker establishes a new state-of-the-art among open-source models.

Model	Awareness Task					Interpretation Task					Imagination Task					Overall
	$\mathcal{IF}$	$\mathcal{DP}$	$\mathcal{VQ}$	$\mathcal{KF}$	AVG	$\mathcal{IF}$	$\mathcal{DP}$	$\mathcal{VQ}$	$\mathcal{KF}$	AVG	$\mathcal{IF}$	$\mathcal{DP}$	$\mathcal{VQ}$	$\mathcal{CF}$	AVG	AVG
Nano Banana	70.6	85.7	86.8	75.2	79.6	63.4	84.9	91.4	61.5	75.3	75.3	73.8	87.3	44.3	70.2	75.0
Seedream 4.0	70.8	78.1	86.6	74.6	77.5	63.7	80.1	90.6	64.2	74.6	82.2	77.8	86.9	47.0	73.5	75.2
GPT-image-1	78.5	85.8	**88.0**	81.2	83.3	62.9	82.9	**93.0**	60.8	74.9	84.4	76.2	**89.2**	48.4	74.6	77.6
Nano Banana Pro	**85.4**	**88.6**	83.9	**91.4**	**87.3**	**76.0**	**89.1**	92.3	**75.8**	**83.3**	**86.6**	**79.5**	88.8	**51.5**	**76.6**	**82.4**
MagicBrush	27.2	43.4	53.3	27.1	37.8	16.8	50.3	63.2	22.2	38.1	18.0	36.9	44.8	22.3	30.5	35.5
OmniGen	35.0	42.0	46.7	37.4	40.3	19.0	34.8	40.3	21.5	28.9	42.2	35.1	46.0	38.7	40.5	36.6
AnyEdit	25.0	54.6	61.3	26.3	41.8	15.9	61.2	62.0	20.2	39.8	9.1	49.7	50.9	16.5	31.5	37.7
UltraEdit	26.5	42.5	53.1	33.9	39.0	24.3	61.7	73.6	26.7	46.6	20.7	31.7	45.8	27.5	31.5	39.0
ICEdit	26.1	42.2	61.2	31.8	40.4	21.4	48.3	81.5	24.9	44.0	21.5	40.6	54.0	25.0	35.3	39.9
FLUX.1-Dev	31.4	52.0	55.0	35.5	43.5	27.5	62.2	69.6	29.0	47.1	39.1	47.1	43.4	27.1	39.2	43.2
FLUX.2-Dev	42.6	63.3	78.4	53.3	59.4	35.4	75.0	85.6	37.6	58.4	73.6	70.7	82.1	43.6	67.5	61.8
UniWorld-V1	31.5	48.9	58.8	38.6	44.5	18.1	44.5	58.1	22.5	35.8	30.3	50.3	64.2	27.5	43.1	41.1
HiDream-E1	29.7	41.2	56.3	32.0	39.8	26.7	53.6	68.4	29.6	44.6	39.6	40.1	49.9	29.6	39.8	41.4
OmniGen2	35.0	64.0	75.4	41.3	53.9	18.9	56.9	64.9	23.5	41.1	42.0	64.4	74.6	31.8	53.2	49.4
Step1X-Edit	39.8	53.5	61.3	44.4	49.7	35.7	73.0	75.2	38.2	55.5	44.7	49.4	50.3	28.4	43.2	49.5
Echo-4o	47.6	63.0	75.4	51.7	59.4	30.8	71.0	80.4	32.9	53.8	63.4	62.4	73.7	41.2	60.2	57.8
Bagel	46.2	71.0	75.8	50.8	61.0	38.6	72.1	78.8	39.5	57.3	62.8	68.5	74.5	40.7	61.6	60.0
Uni-CoT	46.0	69.1	77.8	51.6	61.1	36.9	70.1	76.3	38.6	55.5	67.6	64.3	79.6	42.9	63.6	60.1
Qwen-Image-Edit	48.1	69.0	79.5	53.6	62.5	32.1	69.7	80.6	34.2	54.1	67.1	66.8	79.2	42.3	63.8	60.2
DreamOmni2	43.3	74.4	85.0	51.2	63.5	34.3	81.7	88.1	35.9	60.0	50.6	64.9	81.9	35.3	58.2	60.6
Know-Thinker	73.3	81.8	85.9	73.6	78.7	59.2	83.9	91.3	57.6	73.0	88.0	84.2	92.7	56.1	80.3	77.3
VEvo-Harness	77.9	88.2	91.7	77.5	83.3	61.1	92.3	94.9	59.9	77.0	84.6	88.8	94.2	58.6	81.6	80.6

Compared to the base model Qwen-Image-Edit-2511 (Wu et al., 2025), as shown in Table I, our method achieves a substantial overall accuracy gain of 18.4% (from 19.4% to 37.8%), validating the efficacy of the agent-based closed-loop in resolving intricate editing instructions. Remarkably, VEvo-Harness outperforms the strong reasoning-based baseline Know-Thinker in instruction reasoning (69.4% vs. 68.2%), temporal reasoning (45.8% vs. 44.7%), and logical reasoning (20.0% vs. 15.3%), which is directly attributed to the precise failure localization of Visual Traceback and the root-cause-conditioned routing of the Heterogeneous Sandbox Policy that effectively resolves complex logical and temporal dependencies without falling into text-level hallucinations. Crucially, unlike purely reasoning-driven approaches that often sacrifice visual fidelity for logical correctness—as exemplified by ThinkRL-Edit's (Li et al., 2026) severe visual plausibility drop to 62.5%—VEvo-Harness maintains high visual quality (91.8%) and appearance consistency (81.6%) through the causal safeguarding of Visual Continual Governance, which successfully intercepts detrimental patches and prevents destructive updates during multi-turn refinements. While slightly trailing proprietary models and Know-Thinker (Wang et al., 2026c) in spatial and causal dimensions due to the inherent limitations of the base generator's physical priors, the comprehensive performance profile confirms that integrating pixel-grounded causal verification with heterogeneous intervention routing establishes a robust and scalable paradigm for complex image editing, significantly pushing the boundaries of what open-source editing systems can achieve without relying on massive proprietary model scaling.

Performance on KRIS-Bench. As shown in Table V, VEvo-Harness demonstrates robust closed-loop control over complex visual generation on the KRIS-Bench benchmark, achieving an overall accuracy of 84.57%. This represents a 17.43% absolute gain over the base generator Qwen-Image-Edit (67.14%) and a 1.55% lead over the reasoning-based baseline Know-Thinker (83.02%). In conceptual knowledge generation, our harness system scores 84.29% (+23.14% over the base model, +1.64% over Know-Thinker). This improvement is driven by Visual Traceback mechanism, which decomposes abstract semantic-visual misalignments into localized root-cause vectors, allowing the Heterogeneous Sandbox Policy to route anatomical and physical rendering failures to specialized generation skills (e.g., `physical_material_consistency`) instead of relying on blind prompt re-sampling. For procedural knowledge involving multi-step instruction adherence, VEvo-Harness reaches 85.26% (Instruction 93.48%, Logical 76.93%), surpassing Know-Thinker by 1.94%. This is achieved because the counterfactual dual-arm replay in Visual Continual Governance (VCG)

Table VII: **The Paired vs. unpaired counterfactual replay under sampling budgets.** We compare same-seed paired replay with alternative admission schemes while keeping the number of generated candidates comparable. $\mathrm{Var}(\cdot)$ measures the variance of the estimated causal gain across repeated evaluations. Harmful Admission is the fraction of admitted patches with negative held-out gain, and Patch Precision is the fraction of admitted patches that remain beneficial under independent replay.

Admission Scheme	Paired	Same Seed	Replay	$\mathrm{Var}(\widehat{\Delta}^{\mathrm{cf}})\downarrow$	Harmful. ↓	Patch. ↑	G-Eval2 ↑	DPG ↑
Random Admission	–	–	4	0.0318	37.6%	58.9%	78.7	82.43
Single-Arm Replay	–	–	4	0.0247	21.8%	70.6%	81.5	84.67
Best-of-K Admission	–	–	4	0.0193	14.2%	78.1%	83.4	86.25
Unpaired Counter.	✓	–	4	0.0136	8.7%	86.4%	84.5	87.12
Ours	✓	✓	4	**0.0051**	**2.1%**	**94.3%**	**86.1**	**88.14**

Table VIII: **Evolve–Freeze–Evaluate analysis of persistent harness learning.** We evolve the harness for different numbers of episodes, freeze all harness states and update mechanisms, and evaluate each checkpoint on the same unseen probe set. Gain denotes the average improvement over the unevolved harness. Repair Success measures the fraction of initially failed constraints resolved within the frozen inference loop, and Avg. Iter. reports the number of generation–verification iterations per sample.

Episodes	Frozen Generation			Frozen Editing			Capability		Repair Efficiency	
	G-Eval2 ↑	DPG ↑	OneIG ↑	RISE ↑	KRIS ↑	WiseEdit ↑	Gain ↑	Probe Suc. ↑	Repair Suc. ↑	Iter. ↓
0	77.6	81.32	51.4	30.8	76.21	72.9	–	64.8	48.7	4.18
25	79.8	83.07	52.9	32.1	78.06	74.6	+1.72	67.9	54.3	3.91
50	81.5	84.46	54.1	33.4	79.63	76.2	+3.51	71.2	60.1	3.62
100	83.4	86.21	55.8	35.2	81.74	78.1	+5.70	75.6	67.8	3.24
200	85.2	87.63	57.4	36.8	83.72	79.7	+7.37	80.9	74.6	2.87
300	**86.1**	**88.14**	**58.1**	**37.6**	**84.83**	**80.6**	**+8.49**	**84.7**	**78.9**	**2.63**

intercepts textually plausible but visually invalid intermediate generation states that typically derail long-horizon harness trajectories. Regarding factual knowledge requiring precise spatial and temporal grounding, our method attains 84.93% (+1.85% over Know-Thinker), as the denied buffer and success preservation mechanisms prevent the catastrophic forgetting of verified visual predicates during multi-turn harness iterations. Ultimately, these results confirm that anchoring multi-turn visual generation in pixel-level causal verification and heterogeneous intervention routing overcomes the structural bottlenecks of text-centric paradigms, enabling our open-source harness to closely match proprietary models like NanoBanana-Pro (85.31%) without scaling the underlying diffusion backbone.

Performance on Wise-Edit. We evaluate VEvo-Harness on WiseEdit (Pan et al., 2026a), a rigorous visual editing benchmark covering awareness, interpretation, and imagination tasks with dense visual constraints. As shown in Table VI, our causally guarded self-evolving harness achieves an overall accuracy of 80.6%, a 20.4% absolute improvement over the frozen Qwen-Image-Edit (Wu et al., 2025) base model (60.2%) and 3.3% higher than the reasoning-aware baseline Know-Thinker (Wang et al., 2026c) (77.3%). VEvo-Harness outperforms Know-Thinker across awareness (88.2% / 91.7% vs. 81.8% / 85.9%) and interpretation subtasks (92.3% / 94.9% vs. 83.9% / 91.3%). These gains arise from the harness stack: Visual Traceback decomposes coarse global failures into localized root causes, and the Heterogeneous Sandbox Policy routes distinct visual anomalies to specialized intervention skills instead of monolithic prompt rewriting. On the challenging imagination subtask requiring physical plausibility and multi-hop visual reasoning, VEvo-Harness attains 81.6% average accuracy, exceeding both Know-Thinker (80.3%) and proprietary Nano Banana Pro (Google, 2025) (76.6%), with a substantial creative fidelity advantage (58.6% vs. 51.5%). The Visual Continual Governance underpins this imagination performance: counterfactual dual-arm replay validates novel physically-consistent edits, while the denied-change buffer and success preservation mitigate harness-level hallucination and catastrophic forgetting over iterative harness updates. These results demonstrate that pixel-grounded causal verification and root-cause-aware routing overcome core limitations of text-centric agentic pipelines.

A.3.3 More Ablation Studies

Effect of Paired Counterfactual Replay. Table VII isolates the contribution of causal pairing in VCG under comparable replay budgets. Random admission produces a 37.6% harmful-admission rate,

Table IX: **Forgetting and success preservation under continual harness evolution.** We evaluate whether sequential harness updates preserve previously acquired capabilities. We report forgetting, backward transfer (BWT), retention of historical performance, and preservation of previously solved cases across the evolution trajectory.

Method	Final Performance			Continual Retention				Old-Task Performance		
	GEval2	DPG	KRIS	Forget.	BWT	Retention	Success	25 Ep.	150 Ep.	300 Ep.
Episodic-Only	80.7	84.02	79.18	4.82	-3.91	93.8	84.6	78.4	76.9	74.8
Admit-All	81.9	84.73	79.64	7.36	-6.18	89.7	77.9	79.1	75.8	71.6
w/o Success Pre.	83.8	86.47	81.72	3.17	-2.46	95.7	88.3	79.6	78.5	76.9
Full	**86.1**	**88.14**	**84.83**	**0.74**	**+0.38**	**99.1**	**97.4**	**79.8**	**80.3**	**80.6**

Table X: **Cross-benchmark transfer of the evolved harness.** Each harness is evolved on one source benchmark and frozen before zero-shot evaluation on unseen targets. Entries report absolute gains over a fresh harness. We summarize transfer by average forward transfer (FWT), within-family gain, and cross-family gain.

Source	Generation Targets			Editing Targets			Transfer Statistics		
	GenEval2	DPG	OneIG	RISE	KRIS	WiseEdit	Avg. FWT ↑	Intra-Fam. ↑	Cross-Fam. ↑
GenEval2	–	+3.8	+2.9	+0.8	+1.6	+1.1	+2.04	+3.35	+1.17
DPG-Bench	+4.1	–	+3.3	+0.5	+1.8	+1.2	+2.18	+3.70	+1.17
OneIG-EN	+3.5	+3.0	–	+0.9	+1.4	+1.7	+2.10	+3.25	+1.33
RISE	+0.4	-0.2	+0.6	–	+3.1	+3.5	+1.48	+3.30	+0.27
KRIS	+1.0	+0.7	+0.5	+3.4	–	+3.8	+1.88	+3.60	+0.73
WiseEdit	+0.3	+0.6	-0.1	+3.2	+3.6	–	+1.52	+3.40	+0.27
Average	**+1.86**	**+1.58**	**+1.44**	**+1.76**	**+2.30**	**+2.26**	**+1.87**	**+3.43**	**+0.82**

while single-arm and best-of-K verification reduce it to 21.8% and 14.2%, respectively. Sampling alone therefore improves admission quality, but remains insufficient to determine whether a patch causes the observed improvement. Introducing two-arm counterfactual replay further reduces harmful admission to 8.7% with independent seeds. Using the same seed for both arms lowers $\mathrm{Var}(\widehat{\Delta}^{\mathrm{cf}})$ from 0.0136 to 0.0051 and harmful admission from 8.7% to 2.1%, while patch precision rises from 86.4% to 94.3%. The controlled comparison removes a substantial portion of generation stochasticity from the estimated patch effect. The reduction in estimator variance translates into persistent downstream gains. Under the same replay size $K = 4$, unpaired counterfactual replay reaches 84.5 on GenEval2 and 87.12 on DPG-Bench, whereas paired replay reaches 86.1 and 88.14. Best-of-K remains below both counterfactual variants despite using the same candidate budget, separating the effect of additional sampling from that of causal comparison. These results attribute the gain of VCG to seed-controlled counterfactual estimation rather than sampling volume alone: additional samples improve selection, while pairing determines whether the measured gain can be reliably assigned to the proposed patch.

Evolve–Freeze–Evaluate. Table VIII evaluates whether capabilities acquired during evolution persist after the harness is frozen. All checkpoints are tested on the same unseen probe set with memory writing, skill promotion, consolidation, and policy updates disabled. From 0 to 300 evolution episodes, GenEval2 increases from 77.6 to 86.1, DPG-Bench from 81.32 to 88.14, and KRIS from 76.21 to 84.83, with an average gain of 8.49 points across the six benchmarks. As the probe set cannot modify the harness state, these gains reflect capabilities accumulated before evaluation rather than test-time adaptation. The same trend appears in repair behavior. Probe success increases from 64.8% to 84.7%, repair success from 48.7% to 78.9%, while the average number of repair iterations decreases from 4.18 to 2.63. Improvements continue between 200 and 300 episodes but narrow, indicating that capability accumulation approaches saturation. The frozen checkpoints thus retain both higher task performance and more efficient repair on unseen inputs, providing direct evidence that evolution changes the persistent harness state rather than relying on online refinement.

Performance under Shared Resource Ceilings. Table II compares methods under shared resource ceilings on Qwen-Image-2512, allocating each the same 300 evolution and 300 held-out tasks and recording actual consumption rather than exhausted budgets, so advantages reflect harness design rather than compute. VEvo-Harness leads all three benchmarks within GenRouter's resource envelope, improving GenEval2 by 3.8% (84.0 vs 80.9), DPG-Bench by 0.5% (88.0 vs 87.6), and OneIG-ZH by

Table XI: **Root-cause localization and retrieval analysis.** We evaluate whether Visual Traceback accurately identifies generation failures and whether the resulting root-cause representation provides an effective key for retrieving repair experience. Precision, Recall, F1, and Exact Match measure root-cause localization quality. Recall@K, MRR, and Relevant@3 evaluate retrieval accuracy, while Repair Success and Avg. Iter. measure the effectiveness of retrieved experience in agent loops.

Diagnosis Method	**Precision ↑**	**Recall ↑**	**F1 ↑**	**Exact Match ↑**	**Judge Calls ↓**	**Repair Succ. ↑**	**Iter. ↓**	**Latency (s) ↓**
Free-form VLM	68.4	65.7	67.0	42.8	**3.7**	68.9	3.74	**34.6**
Fixed Checklist	76.8	73.9	75.3	54.6	8.9	74.8	3.31	40.7
Adaptive w/o Red. Filter	82.7	80.1	81.4	63.5	9.8	78.2	3.05	41.2
Visual Traceback	**86.5**	**84.2**	**85.3**	**69.1**	6.2	**81.6**	**2.81**	38.4

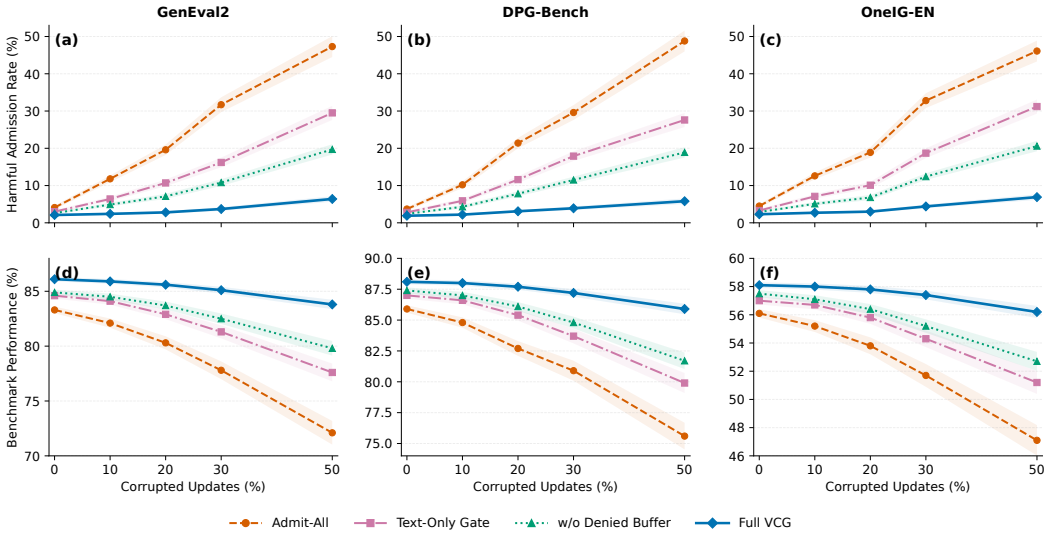
Memory Key	**Recall@1 ↑**	**Recall@3 ↑**	**MRR ↑**	**Rel.@3 ↑**	**Repair Succ. ↑**	**ΔRepair ↑**	**Avg. Iter. ↓**	**GenEval2 ↑**
Random Retrieval	12.6	28.9	0.214	24.7	61.8	+3.1	3.96	78.9
Lexical Retrieval	31.8	52.4	0.421	46.1	68.7	+7.4	3.54	80.9
Prompt Embedding	45.7	68.3	0.558	61.5	74.2	+11.6	3.17	82.8
Root-Cause Embedding	58.9	79.6	0.671	73.8	79.4	+16.3	2.87	85.2
Root Cause + Reranking	**63.7**	**84.8**	**0.714**	**78.6**	**82.1**	**+18.7**	**2.63**	**86.1**

2.0% (0.570 vs 0.559), with margins over GEMS widening to 5.3%, 1.4%, and 3.6%. The gain is largest on GenEval2, whose compositional constraints are where visual traceback, failure-conditioned experience, and governed skill admission concentrate capacity; smaller margins on DPG-Bench and OneIG-ZH match their constraint profiles and are deemed reliable only after non-overlapping paired confidence intervals. Resource accounting rules out a compute-only explanation: VEvo-Harness issues fewer image calls (4,150 vs 4,195) but more VLM calls (3,720 versus 3,300) and tokens (3.966M vs 3.780M) on diagnosis, with marginal latency increase (38.4 versus 37.8 s), redirecting budget from generation to causally grounded governance. VEvo-Harness also shows the narrowest confidence intervals, indicating greater stability. This aggregate view does not isolate components, requiring the reported ablations, and smaller margins warrant paired statistical confirmation.

Evolution Trajectory and Statistical Robustness. Table III tracks capability across the full 300-episode evolution stream at 50-episode intervals, with each snapshot averaged over 3 task orders and 3 random seeds (9 independent runs). Three properties stand out. First, every capability metric follows a strictly monotonic trajectory: GenEval2, DPG-Bench, and KRIS rise from 77.60, 81.32, and 76.21 to 86.10, 88.14, and 84.83 respectively, while Probe and Repair Success nearly double from 64.80% and 48.70% to 84.70% and 78.90%. The absence of any performance dip across 300 episodes provides direct evidence that the harness avoids catastrophic forgetting, confirming that Visual Continual Governance gates out harmful updates rather than merely accumulating experience. Second, the 95% confidence intervals remain narrow throughout (within ± 0.8 for benchmark scores and ± 1.6 for success rates), indicating that the evolutionary trajectory is robust to both task ordering and initialization stochasticity, and is therefore an intrinsic property of the governed write path rather than an artifact of a favorable curriculum. Third, capability growth is accompanied by efficiency gains: the average number of repair iterations decreases monotonically from 4.18 to 2.63, showing that the lifecycle-managed state distills reusable, causally verified skills that accelerate downstream problem-solving rather than relying on brute-force search. These results establish that VEvo-Harness achieves genuine reset-free monotonic self-evolution under a causally governed write path.

Table IV evaluates the completely frozen harness on distributions entirely absent from the 300-episode evolution stream, directly testing whether the accumulated skills memorize training statistics or capture transferable repair primitives. Three observations confirm genuine zero-shot generalization. First, compositional and spatial reasoning improves consistently across all four T2I-CompBench dimensions, with gains ranging from 6.60 to 7.60, indicating that the predicate-level decomposition learned by visual traceback transfers to unseen attribute-binding and spatial configurations without task-specific fine-tuning. Second, the largest gains appear in knowledge-grounded editing, where WiseEdit edit success rises by 15.70 and knowledge accuracy by 11.70. This confirms that the lifecycle-managed pattern memory distills reusable knowledge-resolution strategies, such as resolving implicit entities before region masking, that generalize beyond the specific facts encountered during evolution. Third, and most critically, the frozen harness reduces hallucination rate by 14.30 and physical violation rate by 10.80, providing direct evidence that Visual Continual Governance pruned causally invalid updates at write time, so the distilled skill library is intrinsically hallucination-

Figure VI: **Robustness to harmful injection.** We corrupt an increasing fraction of candidate updates during harness evolution and evaluate robustness on GenEval2, DPG-Bench, and OneIG-EN. The top row reports the harmful admission rate, and the bottom row reports the corresponding benchmark performance. Full VCG suppresses harmful updates and limits performance degradation under increasing experience corruption. Shaded regions denote standard deviation across evolution runs.



resistant even on out-of-distribution inputs. Together, these results demonstrate that the governed write path yields generalizable, causally grounded abilities rather than overfitting to evolution stream.

Forgetting and Success Preservation. Table IX evaluates capability retention as the image-generation harness evolves over a sequential task stream. Full VEvo-Harness limits average forgetting to 0.74% and retains 99.1% of historical performance. Removing success preservation increases forgetting to 3.17% and reduces the preservation of previously solved cases from 97.4% to 88.3%. On fixed old-task probe, Full improves from 79.8 at 25 episodes to 80.6 at 300 episodes, while the variant without success preservation declines from 79.6 to 76.9. This gap shows protecting validated generation behaviors prevents new repair patches from overwriting prompt construction, visual constraint handling, and intervention rules that remain effective for earlier image-generation tasks. Admission quality has a larger effect on long-term retention. Admit-All produces 7.36% of forgetting and a BWT of -6.18, with old-task performance decreasing from 79.1 to 71.6. Episodic-Only reduces the degradation but still yields 4.82 points of forgetting and negative BWT, showing that retaining repair trajectories alone is insufficient to stabilize an evolving generation harness. Full VEvo-Harness instead achieves positive BWT of $+0.38$, 0.74% forgetting, and 97.4% preservation of previously solved cases. These results show causally governed skill admission and success preservation allow harness to incorporate new visual repair knowledge without sacrificing generation behaviors.

Cross-Benchmark Harness Transfer. Table X examines whether capabilities accumulated in the harness remain useful outside the benchmark used for evolution. The harness is frozen before target evaluation, so the reported gains do not involve target-side memory writing, skill updates, or further evolution. Across unseen targets, the evolved harness achieves an average FWT of $+1.87\%$. Transfer is stronger within the same task family, with an average gain of $+3.43\%$. For image generation, evolving on GenEval2 improves DPG-Bench and OneIG-EN by 3.8% and 2.9%, while DPG-Bench evolution transfers 4.1% to GenEval2 and 3.3% to OneIG-EN. A similar pattern appears for image editing, where evolution on KRIS improves RISE by 3.4% and WiseEdit by 3.8%. These gains indicate that the harness retains repair strategies that can be reused across different data distributions rather than only memorizing benchmark-specific prompts. Cross-family transfer is smaller, averaging $+0.82\%$. Generation-to-editing transfer remains mostly positive, while editing-to-generation transfer contains several near-zero cases, including -0.2% from RISE to DPG-Bench and -0.1% from WiseEdit to OneIG-EN. This difference reflects the scope of the learned harness state. Visual decomposition and compositional repair skills transfer across related generation benchmarks, while editing-specific interventions show weaker cross-family transfer. This pattern suggests that VEvo-Harness learns reusable visual repair skills rather than benchmark-specific prompt rules.

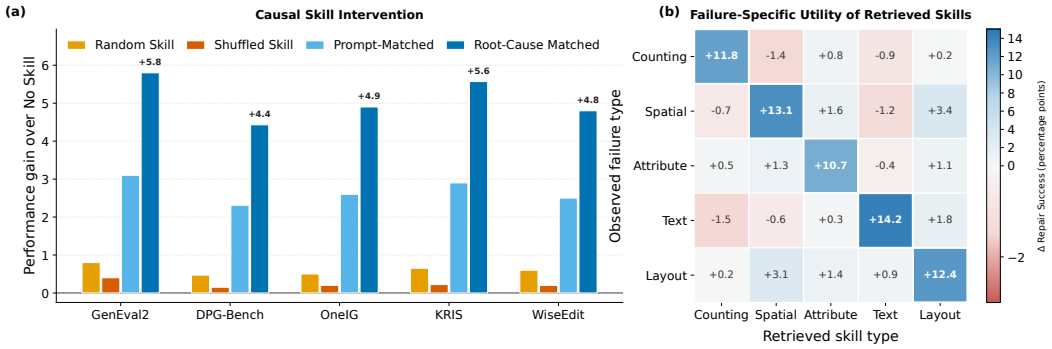
Causal Skill Intervention. We intervene on the retrieved skill while keeping the evolved harness, generator, verifier, and inference procedure fixed, isolating whether the accumulated skill library provides reusable repair knowledge rather than additional context. As shown in Fig. VII(a), random and shuffled skills yield only marginal gains over disabling skill retrieval, whereas prompt-matched retrieval provides a consistent but smaller improvement. Root-cause-matched retrieval produces the largest gain on every benchmark, improving over the no-skill intervention by 5.8%, 4.43%, 4.9%, 5.57%, and 4.8% on GenEval2, DPG-Bench, OneIG, KRIS, and WiseEdit, respectively. The gap between prompt matching and root-cause matching is important: retrieval based on surface task similarity does not recover the full benefit of the evolved skill library. The useful retrieval key is the diagnosed visual failure. We further test this dependency by explicitly crossing failure categories with skill categories in Fig. VII(b). Matched failure–skill interventions improve repair success by 10.7–14.2%, with an average gain of 12.44%. In contrast, mismatched interventions average only 0.41%, giving a 12.03% specificity gap. The off-diagonal entries are not uniformly zero: for example, spatial and layout skills exhibit limited positive transfer, consistent with their shared dependence on geometric organization. Nevertheless, the dominant diagonal shows that the gain cannot be explained by injecting arbitrary previous experience. A skill becomes useful primarily when its accumulated repair knowledge matches the current failure structure. Visual Traceback identifies relevant skills through root-cause representations, while lifecycle-managed state preserves repair knowledge for recurring visual failures. Their coupling enables failure-specific skill reuse and explains the advantage over random experience reuse and prompt-level semantic retrieval.

Root-Cause Localization and Root-Cause-Guided Retrieval. Table XI first evaluates whether Visual Traceback can localize the failures that lead to unsuccessful image generation. Free-form VLM diagnosis reaches an F1 of 67.0% and only 42.8% exact match, indicating that unconstrained diagnosis often misses or mixes multiple visual failure causes. A fixed checklist improves F1 to 75.3%, but requires 8.9 judge calls because the same verification dimensions are evaluated for every failure. Adaptive Visual Traceback reaches 85.3% F1 and 69.1% exact match with 6.2 judge calls. Removing redundancy filtering increases the verification cost to 9.8 calls while reducing F1 to 81.4%, showing that additional verification questions do not necessarily improve localization. Localization accuracy also affects the subsequent image repair loop. Repair success increases from 68.9% with free-form diagnosis to 81.6% with adaptive traceback, while the average number of repair iterations decreases from 3.74 to 2.81. The localized root cause therefore provides a more precise state for selecting repair actions than either a global VLM description or a fixed set of visual checks.

The second part of Table XI evaluates the root-cause representation as a retrieval key for persistent repair experience. Prompt embeddings achieve 68.3% Recall@3 and 74.2% repair success, while root-cause embeddings increase them to 79.6% and 79.4%. The gap shows that prompts with different surface content can share the same visual failure, while similar prompts do not necessarily require the same repair. Indexing memory by the diagnosed failure therefore aligns retrieval with the state that the harness needs to correct. Reranking root-cause candidates by their historical counterfactual gain further raises Recall@3 to 84.8% and repair success to 82.1%, while reducing the average repair trajectory from 3.17 iterations with prompt retrieval to 2.63. GenEval2 increases from 82.8% to 86.1% under the complete retrieval pipeline. Root-cause indexing identifies experiences that address the current visual error, while Δ^{cf} reranking favors repairs with verified image-level gains.

Independent Verifier and Human Evaluation. In Figure VIII, we evaluate whether VCG is affected by self-confirmation from its internal judge by comparing its decisions with an independent VLM and human evaluation across both generation and editing benchmarks. The original judge remains broadly consistent with independent evaluation, while the disagreement varies with task difficulty. DPG-Bench shows relatively high consistency, with 92.1% agreement and 89.4% human correctness, whereas GenEval2 reaches 89.7% and 86.8%. On OneIG-EN, which contains more diverse visual constraints, these values decrease to 85.6% and 82.3%. A similar pattern appears in editing: KRIS maintains 90.4% agreement and 88.7% correctness, while RISE drops to 81.9% and 79.6%, reflecting the greater ambiguity of temporal and logical edits. More importantly, patches admitted by the original judge retain positive gains under independent held-out evaluation, reaching $+6.8\%$, $+5.4\%$, and $+4.7\%$ on GenEval2, DPG-Bench, and OneIG-EN. The gains remain positive on editing benchmarks, including $+3.2\%$ on the more challenging RISE, although independent-VLM and human admission increase it to $+4.5\%$ and $+5.1\%$. The larger gaps on difficult generation and editing tasks identify verifier quality as a remaining source of error, but the consistently positive held-out gains show that patches accepted by VCG remain useful under evaluators that did not participate in their admission.

Figure VII: **(a)** Performance gains over the evolved harness without skill retrieval. Root-cause-matched retrieval consistently outperforms random, mismatched, and prompt-matched retrieval across benchmarks. **(b)** Repair-success gains under controlled failure–skill interventions. The strong diagonal shows that matched skills provide substantially greater utility than mismatched ones, with limited transfer across related failures. Together, the results show that the harness accumulates failure-specific, reusable repair knowledge that can be effectively indexed by Visual Traceback.



This reduces the likelihood that the observed causal gains arise primarily from judge self-confirmation and supports counterfactual visual verification as a reliable admission signal across visual tasks.

Heterogeneous Sandbox Policy. We further analyze how HSP routes visual failures to heterogeneous interventions. Fig. IX(a) reveals a clear failure–action structure in the counterfactual response matrix. Prompt-level failures are more responsive to rewriting, with Rewrite achieving gains of 0.66 for attribute binding and 0.55 for spatial relations. In contrast, renderer-side failures favor parameter tuning, which reaches 0.57 for layout or blur errors and 0.51 for style mismatch. Resampling is more effective for stochastic failures such as counting (0.46) and fine-grained detail (0.43). These differences indicate that visual failures do not share a universally effective repair operation and motivate routing interventions according to the localized root cause. Fig. IX(b) shows that HSP exploits this action specialization effectively. Across all seven failure types, HSP consistently outperforms random and mismatched intervention selection. The difference is particularly large for attribute binding, where repair success increases from 8.3% with mismatched actions and 29.7% with random routing to 69.4% with HSP, approaching the 74.1% oracle result. Similar gains appear for layout or blur errors, where HSP reaches 60.1% compared with 27.4% under random routing. Fig. IX(c) further shows that this routing capability improves as the harness evolves. HSP reduces regret to the oracle from 18.9% initially to 4.6% after 300 episodes, whereas random routing remains near 23%. Removing the EMA update or the variance-aware action value leaves substantially higher final regret of 8.9% and 10.5%, respectively. The results support the role of HSP as an adaptive repair policy: it learns failure-conditioned intervention preferences from accumulated counterfactual outcomes and increasingly selects actions that match the observed visual root cause.

Robustness to Harmful Experience Injection. As shown in Figure VI, we stress-test VCG by corrupting $c \in 0, 10, 20, 30, 50\%$ of candidate updates during harness evolution while keeping the task stream and repair budget fixed. We measure the harmful admission rate as $N_{\text{admitted}\cap\text{harmful}}/N_{\text{harmful}}$, which quantifies how often harmful candidates are promoted into persistent harness state. As shown in the top row of Figure VI, Admit-All becomes increasingly vulnerable to contamination, reaching 48.1%, 45.6%, and 51.6% HAR at 50% corruption on GenEval2, DPG-Bench, and OneIG-EN. Text-Only Gate reduces these values to 30.8%, 26.4%, and 35.9%, but cannot determine whether a linguistically plausible update produces a harmful visual effect. Removing the denied-change buffer also increases HAR to 21.3%, 17.2%, and 24.8%, indicating that previously rejected repair directions can repeatedly re-enter the evolution process. In contrast, Full VCG limits HAR to 6.7%, 5.5%, and 8.4% under the same corruption level. Counterfactual image-level replay filters candidates according to their actual visual effect, while the denied-change buffer suppresses repeated harmful proposals. The bottom row shows that harmful admission translates into persistent capability degradation. We define degradation as $D(c) = S(0) - S(c)$, where $S(c)$ is the benchmark score after evolution under corruption ratio c. At 50% corruption, Admit-All degrades by 12.4%, 8.2%, and 10.2% on GenEval2, DPG-Bench, and OneIG-EN, respectively. Full VCG reduces these drops to 3.2%, 2.14%, and 3.8%, retaining 82.9%, 86.0%, and 54.3% performance. The different

Table XII: **Sensitivity analysis of the maximum Visual Traceback depth.** We vary the maximum traceback depth while keeping all other configurations fixed. Results are reported across three image-generation benchmarks and three image-editing benchmarks.

Depth	Image Generation			Image Editing			Avg.
	GenEval2 ↑	DPG-Bench ↑	OneIG-EN ↑	RISE ↑	KRIS ↑	WiseEdit ↑	
1	82.7	85.31	55.8	35.4	81.62	76.9	69.62
2	**86.1**	87.96	**58.1**	37.3	**84.83**	80.4	**72.45**
3	85.8	**88.14**	57.9	**37.6**	84.57	**80.6**	72.44
4	84.9	87.42	57.3	36.8	83.91	79.7	71.67

Table XIII: **Sensitivity analysis of Visual Continual Governance.** We vary the counterfactual replay size K and admission margin ϵ with all other configurations fixed. In addition to generation and editing performance, we report average image calls, token consumption, and latency per sample.

Value	Image Generation			Image Editing			Avg. ↑	Efficiency		
	G-Eval2 ↑	DPG ↑	OneIG ↑	RISE ↑	KRIS ↑	WiseEdit ↑		Calls ↓	Tokens (K) ↓	Latency (s) ↓
Counterfactual Replay Samples K										
1	82.9	85.71	55.9	35.3	81.74	77.2	69.79	**3.8**	**5.72**	**31.6**
2	84.8	87.12	57.1	36.4	83.21	78.9	71.26	4.7	5.94	34.1
4	**86.1**	**88.14**	**58.1**	**37.6**	**84.83**	**80.6**	**72.56**	6.2	6.32	38.4
8	85.9	88.02	57.9	37.4	84.61	80.3	72.35	9.4	6.91	47.7
16	85.6	87.81	57.7	37.1	84.36	80.0	72.10	15.8	8.03	65.9
Admission Margin ϵ										
0.00	84.5	86.84	56.8	35.8	82.67	78.5	70.85	7.1	6.67	41.3
0.01	85.4	87.53	57.5	36.9	83.91	79.7	71.82	6.6	6.45	39.7
0.02	**86.1**	**88.14**	**58.1**	**37.6**	**84.83**	**80.6**	**72.56**	**6.2**	**6.32**	**38.4**
0.05	85.8	87.91	57.8	37.3	84.51	80.2	72.25	5.8	6.18	37.1
0.10	84.9	87.16	57.1	36.5	83.52	79.4	71.43	5.2	5.96	35.8

degradation profiles across benchmarks also show that robustness is not a fixed performance offset, since corrupted experience affects different visual failure distributions differently. The consistent correspondence between lower HAR and smaller capability loss supports the role of causal governance in persistent harness evolution: VCG prevents unreliable local repair experience from being promoted into reusable skills, limiting the accumulation of harmful behavior without freezing harness updates.

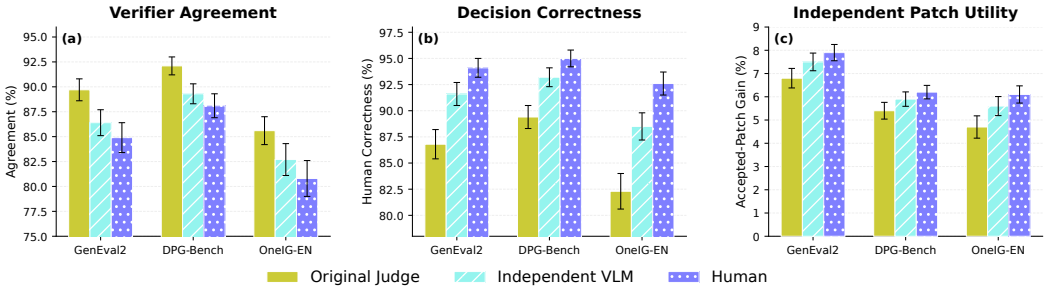
A.4 More Hyper-parameter Analysis

Effect of Traceback Depth. Table XII reports the sensitivity to the maximum traceback depth across six generation and editing benchmarks. Increasing the depth from 1 to 2 improves the average score from 69.62 to 72.45. The improvement is observed consistently across all benchmarks, including 4.1% on GenEval2, 3.1% on DPG-Bench, 4.1% on OneIG-EN, 5.4% on RISE, 3.9% on KRIS, and 4.6% on WiseEdit. This consistent trend across both generation and editing tasks suggests that a single-level traceback often leaves the original visual failure insufficiently resolved for subsequent intervention. In contrast, depths 2 and 3 form a stable performance plateau, achieving nearly identical average scores of 72.45 and 72.44, respectively. Their differences are also benchmark-dependent: depth 2 performs best on GenEval2, OneIG-EN, and KRIS, whereas depth 3 performs best on DPG-Bench, RISE, and WiseEdit benchmark. This indicates that the benefit of traceback does not depend on a narrowly tuned depth; once the failure is decomposed to an actionable level, further decomposition provides limited additional information. Increasing the depth to 4 reduces the average score to 71.67, a 1.1% relative decrease from depth 3, with degradation observed on all six benchmarks. We attribute this decline to over-decomposition, where additional traceback steps increasingly produce redundant or weakly grounded diagnostic predicates that can introduce noise into retrieval and intervention selection. Overall, the results identify depths 2–3 as a robust operating region. We use depth 2 as the default because it matches the accuracy of depth 3 while requiring fewer verification steps, providing a better trade-off between diagnostic resolution and inference overhead.

Table XIV: **Sensitivity to the semantic suppression threshold θ_J in VCG.** We vary the similarity threshold used by the denied-change buffer while keeping all other configurations fixed. False Supp. denotes beneficial proposals incorrectly suppressed by the buffer, and Redundant Replay denotes previously rejected semantic variants that are unnecessarily re-evaluated.

θ_J	Governance Behavior			Image Generation			Image Editing	
	False Supp. ↓	Redun. Replay ↓	Calls ↓	GenEval2 ↑	DPG-Bench ↑	OneIG-EN ↑	KRIS ↑	WiseEdit ↑
0.30	12.8%	**3.1%**	**5.74**	82.7	85.41	55.6	81.34	77.1
0.40	8.9%	4.4%	5.89	83.8	86.37	56.4	82.46	78.2
0.50	5.6%	6.2%	6.03	84.9	87.21	57.2	83.57	79.1
0.55	4.1%	7.5%	6.14	85.6	87.72	57.7	84.21	79.8
0.60	2.9%	8.8%	6.32	**86.1**	**88.14**	**58.1**	**84.83**	**80.6**
0.65	2.3%	10.6%	6.58	85.9	87.96	58.0	84.65	80.4
0.70	1.8%	13.1%	6.91	85.5	87.63	57.7	84.28	80.0
0.80	1.1%	18.9%	7.68	84.6	86.94	56.9	83.42	79.1
0.90	**0.5%**	27.4%	8.72	83.4	85.98	56.0	82.31	78.0

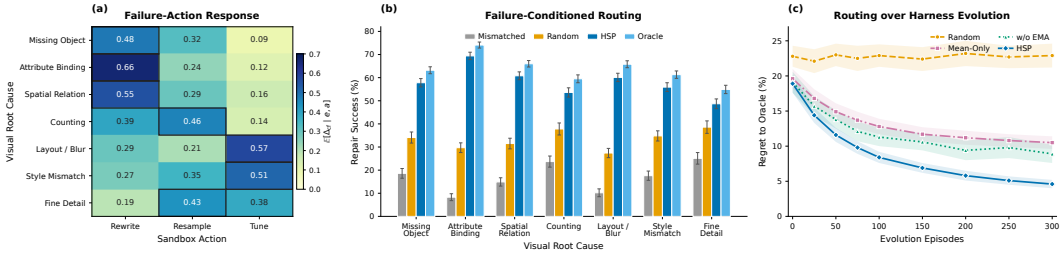
Figure VIII: **Independent verifier and human evaluation across image generation benchmarks.** Agreement and human correctness measure verifier reliability, while Accepted-Patch Gain evaluates admitted patches under held-out independent evaluation. Consistent positive gains across generation and editing tasks support the validity of VCG admission beyond its internal judge.



Effect of Visual Continual Governance Hyperparameters. Table XIII examines two parameters governing update admission in VCG, namely the counterfactual replay size K and the admission margin ϵ. Increasing K from 1 to 4 raises the average score from 69.79 to 72.56, corresponding to a 4.0% relative improvement across the six benchmarks, with the largest gain observed on RISE from 35.3 to 37.6. The consistent improvement suggests that counterfactual estimates based on only a few samples are more susceptible to generation stochasticity, while repeated replay provides more reliable evidence for evaluating candidate updates. Performance largely saturates beyond $K=4$. Increasing K to 8 and 16 yields average scores of 72.35 and 72.10, respectively, despite the proportional increase in replay cost. We therefore adopt $K=4$ as it provides sufficient counterfactual evidence for reliable skill admission without incurring unnecessary generation overhead. The admission margin ϵ controls the balance between reliable admission and adaptation plasticity. Increasing ϵ from 0 to 0.02 improves the average score from 70.85 to 72.56, a relative gain of 2.4%, while RISE improves from 35.8 to 37.6. A moderate margin therefore helps exclude updates whose estimated gains are too small to be reliably distinguished from evaluation noise. Increasing ϵ further to 0.05 and 0.10 lowers the average score to 72.25 and 71.43, respectively, suggesting that an overly conservative gate begins to reject beneficial updates and restrict adaptation. Performance remains stable around the selected configuration, with $K=8$ and $\epsilon=0.05$ retaining 99.7% and 99.6% of the default performance. Overall, $K=4$ and $\epsilon=0.02$ provide a stable balance among counterfactual estimation reliability, adaptation capacity, and replay cost, without requiring narrowly tuned parameter choices.

Effect of Semantic Suppression Threshold. Table XIV studies the semantic threshold θ_J used by the denied-change buffer to suppress candidate updates that are semantically equivalent to previously rejected changes. Increasing θ_J from 0.30 to 0.60 reduces false suppression from 12.8% to 2.9%, accompanied by consistent improvements on GenEval2 from 82.7 to 86.1, DPG-Bench from 85.41 to 88.14, OneIG-EN from 55.6 to 58.1, KRIS from 81.34 to 84.83, and WiseEdit from 77.1 to 80.6. A low threshold makes denied-change reuse overly aggressive. Candidate updates that are only partially

Figure IX: **Mechanism analysis of the Heterogeneous Sandbox Policy.** (a) Counterfactual action responses across visual root causes reveal heterogeneous intervention preferences. (b) HSP uses these preferences for failure-conditioned routing, approaching oracle repair success. (c) Routing regret decreases as harness evolves, showing continual adaptation from accumulated intervention outcomes.



related to a previous rejection can be suppressed before counterfactual verification, preventing new evidence from entering the governance process. This effect is important in a persistent harness, where the denied-change buffer grows with experience and an overly broad semantic match can repeatedly block useful adaptation in later episodes. The reduction in false suppression at 0.60 therefore preserves the computational benefit of reusing negative experience without treating past rejection as a permanent constraint on semantically related updates. The opposite behavior emerges as θ_J becomes overly restrictive. A high threshold requires near-equivalence before a previous rejection can be reused, allowing minor semantic variants of the same ineffective update to re-enter counterfactual evaluation. At $\theta_J = 0.90$, redundant replay rises to 27.4% and the average number of calls increases from 6.32 to 8.72, while performance declines across all benchmarks. This result highlights the role of the denied change buffer beyond conventional caching. VCG retains negative intervention outcomes as reusable governance evidence, allowing future proposals to avoid repeatedly validating changes that have failed under closely related conditions. θ_J determines how selectively this negative experience is transferred across episodes. We use $\theta_J = 0.60$, which keeps false suppression at 2.9% while retaining effective reuse of rejected updates, providing a practical balance between preserving adaptation opportunities and avoiding redundant counterfactual verification.

Effect of Admission Margin. Table XVI examines how admission margin ϵ controls the persistence of counterfactually evaluated skill updates in VCG. With $\epsilon = 0$, any candidate with a positive estimated intervention effect can be admitted into persistent state. Although this setting retains 91.8% of beneficial candidates, it also admits 8.7% harmful updates, indicating that the sign of the estimated gain alone is insufficient for reliable long-term consolidation. Increasing ϵ to 0.02 requires a candidate update to exhibit a non-trivial improvement over its paired counterfactual baseline before admission. This reduces harmful admission from 8.7% to 2.1%, a relative reduction of 75.9%, while preserving 84.8% of beneficial candidates. The same setting achieves the highest performance across all five benchmarks, reaching 86.1 on GenEval2, 88.14 on DPG-Bench, 58.1 on OneIG-EN, 84.83 on KRIS, and 80.6 on WiseEdit. These gains are important for persistent evolution because an erroneous update is not confined to the current generation. Once consolidated, it can be retrieved and reused in subsequent episodes, allowing a locally noisy decision to affect future behavior. The admission margin therefore acts as a safeguard at the boundary between transient experience and persistent skill state rather than as a generic confidence threshold. Further increasing ϵ exposes the opposite failure mode. At $\epsilon = 0.10$, harmful admission decreases to 0.5%, but beneficial acceptance also falls to 56.3%, accompanied by consistent performance degradation across all five benchmarks. A stricter gate therefore improves admission precision at the expense of suppressing useful harness evolution. This result clarifies the role of the margin within VCG. Counterfactual replay provides evidence about whether a proposed skill improves the current visual objective, while ϵ determines whether that evidence is sufficiently strong to justify a persistent state change. Setting $\epsilon = 0.02$ provides an effective operating point between these two requirements, rejecting weakly supported updates before consolidation without preventing verified experience from accumulating into reusable skills.

Effect of Routing Temperature. Table XVII examines how the temperature τ controls action specialization in HSP. Increasing τ produces the expected monotonic increase in action entropy, from 0.21 at $\tau = 0.10$ to 1.04 at $\tau = 3.00$. The downstream results show that neither extreme is desirable. At low temperatures, the routing distribution becomes sharply concentrated on the currently preferred intervention. This reduces action diversity but also makes early utility estimates difficult to revise. Increasing τ from 0.10 to 1.00 improves GenEval2 from 82.4 to 86.1, DPG-Bench from

85.62 to 88.14, OneIG-ZH from 54.1 to 58.1, RISE from 34.8 to 37.6, and GEdit from 8.18 to 8.63. These gains indicate that HSP benefits from retaining sufficient exploration to distinguish which interventions are effective for different root-cause patterns rather than committing prematurely to a single repair strategy. Performance peaks at $\tau = 1.00$ and remains close at $\tau = 1.25$, after which it decreases consistently as the routing distribution becomes more diffuse. At $\tau = 3.00$, action entropy reaches 1.04 and performance drops across all five benchmarks. In this regime, differences in the learned risk-sensitive utilities have less influence on action selection, weakening the specialization between diagnosed failures and their effective interventions. This behavior directly supports the role of stochastic routing in HSP. The policy should preserve enough exploration to update action preferences from new experience while remaining sufficiently concentrated to exploit the failure–intervention associations accumulated by the harness. We use $\tau = 1.00$, which provides this balance without making routing either prematurely deterministic or effectively indiscriminate.

Effect of the Variance Penalty. Table XVIII examines the role of the variance penalty λ in the risk-sensitive routing policy. When $\lambda = 0$, action selection depends only on the estimated mean utility, resulting in a 14.8% failure rate and 71.3% routing stability. Increasing λ consistently reduces unresolved interventions and improves routing stability, indicating that the variance term filters actions whose average gains are high but whose outcomes are unreliable across related root causes. This effect is also reflected in downstream performance. Increasing λ from 0 to 0.5 improves GenEval2 from 82.8 to 86.1, DPG-Bench from 85.91 to 88.14, OneIG-ZH from 54.8 to 58.1, RISE from 34.9 to 37.6, and GEdit from 8.21 to 8.63. These results show that routing based solely on expected utility is insufficient: an intervention with a favorable historical mean can still be a poor choice when its effect varies substantially across instances. The benefit of variance regularization does not increase indefinitely. Beyond $\lambda = 0.5$, the failure rate continues to decrease and routing becomes more stable, but benchmark performance begins to decline. At $\lambda = 1.5$, the failure rate reaches 5.2% and routing stability reaches 92.8%, while all five downstream scores fall below those obtained at $\lambda = 0.5$. A large variance penalty therefore makes the router increasingly conservative, favoring interventions with predictable outcomes even when alternative actions offer larger expected improvements. This exposes the intended risk–reward trade-off in HSP: insufficient regularization allows unstable actions to dominate through optimistic mean estimates, whereas excessive regularization suppresses useful interventions whose effects are less certain. We use $\lambda = 0.5$ as the default because it achieves the strongest downstream performance while reducing the failure rate from 14.8% to 6.8% and increasing routing stability from 71.3% to 85.9%. The broad performance plateau around $\lambda = 0.35$–0.75 further suggests that the gain of risk-sensitive routing does not depend on a narrowly tuned variance penalty.

Effect of Memory Capacity. Table XV examines the effect of persistent memory capacity on capability accumulation. Increasing C from 25 to 100 improves GenEval2 from 82.9 to 85.8, KRIS from 81.73 to 84.31, and WiseEdit from 77.4 to 80.1, indicating that insufficient capacity discards repair trajectories that remain useful for recurring failures. The gains diminish beyond $C = 100$ and saturate at $C = 200$. Unbounded memory provides no further benefit, reaching 86.0 on GenEval2 and 84.76 on KRIS, compared with 86.1 and 84.83 at $C = 200$. This saturation shows that lifecycle consolidation can compress recurring repair experience into reusable patterns and skills without retaining the complete interaction history. The efficiency results further favor bounded memory. Moving from $C = 200$ to unbounded accumulation increases retrieved memory from 1.91K to 2.68K tokens, total token consumption from 6.61K to 7.43K, and latency from 38.4s to 41.2s, while benchmark performance remains unchanged. We therefore use $C = 200$ as the default capacity: it preserves the performance of unbounded memory with lower inference overhead. These results rule out ever-growing context as the source of continual improvement and support capacity-aware consolidation as the mechanism for maintaining a compact persistent harness.

A.5 More Visualization Results

A.6 Limitations and Future Studies

First, VEvo-Harness primarily evolves the knowledge and policies instantiated within a predefined harness architecture, rather than the architecture itself. While lifecycle-managed memory and heterogeneous interventions enable experience to accumulate into persistent operational knowledge, the representational substrate, intervention repertoire, and governance mechanisms remain designer-specified. This exposes a more fundamental limitation: the current framework can optimize how an existing harness operates, but cannot yet expand the computational abstractions through which it

Table XV: **Sensitivity to the persistent memory capacity C.** We vary the capacity C of instance-level experience memory while keeping all other configurations fixed. ∞ denotes the unbounded accumulation. Memory Tokens (Memory.) and Total Tokens (Total.) measure the retrieved memory context and overall token consumption per sample, while Latency reports end-to-end inference time.

Capacity C	Image Generation			Image Editing			Efficiency		
	GEval2 ↑	DPG ↑	OneIG ↑	RISE ↑	KRIS ↑	WiseEdit ↑	Memory. ↓	Total. ↓	Latency (s) ↓
25	82.9	85.74	55.9	35.2	81.73	77.4	**1.21K**	**5.84K**	**36.7**
50	84.7	87.03	57.0	36.4	83.16	78.9	1.47K	6.11K	37.3
100	85.8	87.91	57.8	37.3	84.31	80.1	1.72K	6.39K	37.9
200	**86.1**	**88.14**	**58.1**	**37.6**	**84.83**	**80.6**	1.91K	6.61K	38.4
∞	86.0	88.09	58.0	37.5	84.76	80.5	2.68K	7.43K	41.2

Table XVI: **Sensitivity to the admission margin ϵ in VCG.** We vary the minimum counterfactual utility required for persistent update admission. Harmful Adm. denotes admitted patches that exhibit negative utility under independent evaluation, while Beneficial Accept. measures the fraction of beneficial candidate patches retained by the gate.

ϵ	Governance Behavior			Image Generation			Image Editing	
	Harmful. ↓	Accept. ↑	Acc. Patches	GenEval2 ↑	DPG-Bench ↑	OneIG-EN ↑	KRIS ↑	WiseEdit ↑
0.000	8.7%	**91.8%**	38.7	84.5	86.11	56.1	81.92	77.6
0.005	6.9%	90.4%	36.9	84.0	86.72	56.7	82.73	78.4
0.010	5.1%	88.7%	34.8	84.9	87.31	57.3	83.58	79.2
0.015	3.6%	86.9%	32.7	85.6	87.81	57.8	84.29	80.0
0.020	2.1%	84.8%	30.9	**86.1**	**88.14**	**58.1**	**84.83**	**80.6**
0.030	1.7%	81.2%	28.4	85.9	87.97	58.0	84.61	80.4
0.050	1.2%	74.6%	24.3	85.4	87.52	57.6	84.02	79.9
0.075	0.8%	65.8%	19.8	84.5	86.79	56.9	83.16	79.0
0.100	**0.5%**	56.3%	16.1	83.6	85.94	56.1	82.24	78.1

reasons, acts, and learns. A natural next step is therefore to move from state-level evolution toward structural harness evolution, where agents can autonomously instantiate, compose, and retire tools, memory operators, verification modules, and execution workflows in response to emerging failure modes. Importantly, such open-ended plasticity should not imply unrestricted self-modification. As the harness becomes increasingly capable of rewriting its own operational substrate, governance must evolve from validating individual updates to regulating structural change itself. We therefore envision future agent harnesses as governed adaptive runtimes, in which the capacity to acquire new computational structure is coupled with explicit mechanisms for attribution, validation, versioning, and rollback. This transition—from evolving knowledge within a fixed runtime to safely evolving the runtime itself—represents a critical step toward genuinely persistent and self-improving agents.

Second, the reliability of persistent evolution is ultimately bounded not only by verifier quality, but by validity horizon of verification itself. VCG mitigates harmful accumulation by admitting candidate updates only after controlled counterfactual evaluation; however, such evidence remains inherently local to the verifier, task distribution, and runtime conditions under which it is obtained. A skill that is beneficial when acquired may become obsolete, interfere with subsequently learned capabilities, or remain falsely trusted as the harness environment evolves. This reveals a fundamental challenge for long-lived agents: persistent knowledge should not be treated as permanently valid merely because it was once verified. Future work should therefore advance from pointwise update verification toward longitudinal governance, where learned capabilities retain explicit provenance, calibrated uncertainty, and validity conditions, and are continuously re-evaluated as the harness and its operating distribution evolve. Such a framework would allow persistent knowledge to be revised, deprecated, or rolled back when its supporting evidence no longer holds. More broadly, we view reliable self-evolution as a problem of governed plasticity: the central challenge is not merely enabling an agent to accumulate capabilities, but ensuring that what it learns remains justified over time.

A.7 Broader Impact

VEvo-Harness advances image generation from stateless per-prompt optimization toward experience-driven, persistent improvement, enabling frozen generators to accumulate verified repair knowledge and reuse it across recurring compositional, spatial, attribute-binding, and rendering failures. By

Table XVII: **Sensitivity to the routing temperature τ in HSP.** We vary the temperature of the risk-sensitive routing policy while keeping all other configurations fixed. Action Entropy measures the entropy of the routing distribution over rewrite, resample, and tune actions. Lower entropy indicates more concentrated intervention selection.

τ	Action Entropy	Image Generation			Image Editing	
		GenEval2 ↑	DPG-Bench ↑	OneIG-ZH ↑	RISE ↑	GEdit ↑
0.10	0.21	82.4	85.62	54.1	34.8	8.18
0.25	0.34	83.6	86.47	55.2	35.7	8.31
0.50	0.51	84.9	87.36	56.4	36.6	8.47
0.75	0.64	85.7	87.91	57.4	37.2	8.58
1.00	0.73	**86.1**	**88.14**	**58.1**	**37.6**	**8.63**
1.25	0.81	85.9	87.98	57.9	37.4	8.60
1.50	0.88	85.5	87.67	57.5	37.0	8.54
2.00	0.97	84.7	87.08	56.8	36.4	8.43
3.00	1.04	83.5	86.21	55.7	35.5	8.27

Table XVIII: **Sensitivity to the variance penalty λ in HSP.** We vary the variance penalty in the risk-sensitive routing utility $\hat{Q}(a, e) = \hat{\mu}(a, e) - \lambda\hat{\sigma}^2(a, e)$. Failure Rate measures episodes that remain unresolved after the selected intervention, while Routing Stability measures action consistency across repeated or semantically matched root-cause instances.

λ	Routing Behavior		Image Generation			Image Editing	
	Failure Rate ↓	Routing Stability ↑	GenEval2 ↑	DPG-Bench ↑	OneIG-ZH ↑	RISE ↑	GEdit ↑
0.00	14.8%	71.3%	82.8	85.91	54.8	34.9	8.21
0.10	12.6%	74.8%	83.9	86.64	55.7	35.6	8.32
0.25	9.7%	79.6%	85.0	87.42	56.8	36.6	8.47
0.35	8.1%	82.7%	85.6	87.89	57.5	37.2	8.57
0.50	6.8%	85.9%	**86.1**	**88.14**	**58.1**	**37.6**	**8.63**
0.65	6.4%	87.2%	85.9	87.97	57.9	37.4	8.60
0.75	6.1%	88.4%	85.6	87.71	57.6	37.1	8.55
1.00	5.7%	90.6%	84.8	87.13	56.9	36.5	8.44
1.50	**5.2%**	**92.8%**	83.5	86.18	55.8	35.4	8.29

shifting part of capability improvement from repeated model scaling or retraining to an evolvable runtime, this paradigm may enable more adaptive, efficient, and long-lived visual-generation systems that progressively improve through deployment experience. However, persistence also amplifies the consequences of erroneous adaptation: verifier biases, incorrect diagnoses, or adversarial experiences may be consolidated into reusable skills and propagate across future generations rather than remaining confined to a single request. VEvo-Harness mitigates this risk through visual traceback, counterfactual validation, controlled skill admission, and success preservation, while broader deployment will require persistent strategies to remain attributable, re-evaluable, and reversible as visual distributions and requirements evolve. We therefore view self-evolving visual generation not merely as a means of accumulating capability, but as a paradigm in which the evolution of generation behavior itself becomes an explicit object of verification and governance.