T2I-SEH: A SELF-EVOLVING TEXT-TO-IMAGE HARNESS VIA EFFECT-CONDITIONED HARNESS LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) generation has recently been exploring a shift from conventional single-shot model invocation toward agentic systems, but its runtime control relies heavily on certain predefined strategies. The self-evolving paradigm presents a promising direction; nevertheless, existing approaches still depend on accurate state interpretation and reliable experience transfer. We propose T2I-SEH, a self-evolving T2I harness that aims to determine when prior experience can be leveraged and to what extent it could influence subsequent actions. The core design is **E**ffect-**C**onditioned **H**arness **L**earning (ECHL) which quantifies paired intervention gains, weights well-supported historical effects according to contextual similarity and uncertainty, and yields cost-aware diagnostic beliefs. This formulation emphasizes relevant and reliable experience, while preserving prior beliefs when sufficient evidence is unavailable. ECHL then integrates cost-adjusted utility bounds with Monte Carlo estimates to determine whether to intervene, acquire additional evidence, or retain the current image. Finally, diagnostic generations refine action-specific beliefs and consolidate observed effects for future reuse. Experiments on the compositional generation task across five architecturally diverse T2I generators demonstrate that T2I-SEH enables more effective intervention selection at substantially lower action cost. In addition, when the task aligns with available interventions, resulting benefits transfer into improved generation quality effectively as reflected in the official GenEval 2 scores. Further, evaluations on CoReBench and DPG-Bench show T2I-SEH's robustness to the task–intervention mismatch by selectively exploiting residual cross-task utility when beneficial and otherwise defaulting to the null intervention at negligible action cost.

1 INTRODUCTION

Recent advances in text-to-image (T2I) generation are driving a transition from conventional single-shot model invocation toward agentic systems. Rather than treating generation as an isolated forward pass, such frameworks embed the underlying generative model within a broader loop that interprets user intent, orchestrates heterogeneous models and external tools, assesses intermediate outputs, and dynamically plans subsequent actions based on the evolving execution state (Jiang et al., 2026a; He et al., 2026b; Feng et al., 2026; Zhang et al., 2026; Ye et al., 2026). Furthermore, a T2I system is characterized not only by its underlying model, but also by its harness, i.e. the surrounding system that situates the model within a broader computational workflow and governs its interaction with the environment (de Macedo, 2026; Meng et al., 2026). However, a critical bottleneck remains insufficiently explored: even when the harness is carefully designed or adapted, the runtime control strategy that determines how the model invokes is typically specified through prompts, heuristics, or fixed conventions. Accordingly, although the agent could benefit from increasingly capable external support, its decision policy rarely evolves from accumulated experience or consolidates such support into a persistent mechanism for future decision making.

The self-evolving paradigm offers a promising avenue for addressing this challenge. Recent agentic T2I studies have explored this paradigm through recursively generated training data (Zhang et al., 2025c), self-refined prompts (Wan et al., 2025), adaptive model routing (Jiang et al., 2026c), and the distillation of tool-mediated generation trajectories (Chen et al., 2026b). Motivated by these emerging successes, we further develop the T2I self-evolving harness that operationalizes self-evolution at the level of runtime control. Moreover, for LLM-based agentic systems, several recent studies have

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

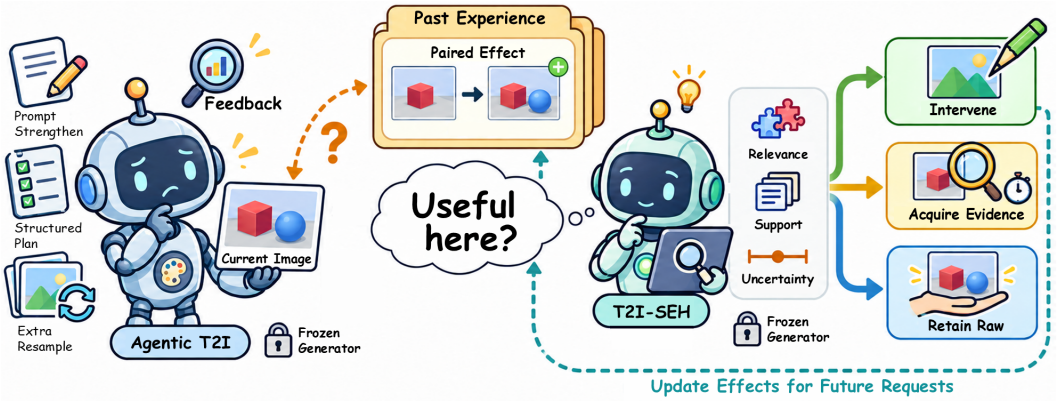

Figure 1: **Overview of T2I-SEH.** Compared with conventional agentic T2I, T2I-SEH more reliably reuses historical experience by assessing when and to what extent it should guide subsequent actions.

explored incorporating self-evolving paradigms into agent harness (Lou et al., 2026; Du et al., 2026; Wang et al., 2026; Jiang et al., 2026b); meanwhile, similar efforts have also emerged in agentic T2I systems, which progressively distill prior generation trajectories, execution failures, and visual feedback into reusable skills that can be leveraged in subsequent generation (Li et al., 2026b; He et al., 2026c). Notably, these approaches generally rely on two strong assumptions: that the current generation state can be accurately interpreted, and that the retrieved experience is sufficiently relevant to be reused. Nevertheless, these assumptions remain fragile in T2I generation, where context-dependent visual feedback can lead to misinterpretation, while experience effective in one state may transfer poorly or even provide misleading guidance in another. More fundamentally, this gives rise to a central challenge: *how can a self-evolving T2I harness dynamically determine when prior experience should be trusted and how strongly it should guide future actions?*

To this end, we introduce the T2I-SEH, a self-evolving T2I harness built upon **E**ffect-**C**onditioned **H**arness **L**earning (ECHL). Given the image and context, ECHL quantifies each intervention by paired improvement effect, while retrieving active historical records that are sufficiently similar to the current context and supported by adequate observations. Then, for each action, ECHL constructs cost-aware diagnostic beliefs by weighting eligible action effects according to contextual similarity and effect uncertainty, with greater weight assigned to records exhibiting higher similarity and narrower effect intervals. The resulting mean estimates the expected benefit, while the variance quantifies uncertainty; in the absence of eligible evidence, ECHL preserves the prior belief unchanged. Furthermore, once historical or requested evidence becomes available, ECHL compares cost-adjusted utilities together with uncertainty bounds, turning beliefs into action by deciding whether the current evidence is sufficient or whether additional image generation is warranted. It retains unchanged when no corrective action exhibits plausibly positive utility, and selects the leading intervention when its lower utility bound is positive and lies within a prescribed tolerance of the upper utility bound of every competing action. If uncertainty persists, ECHL employs Monte Carlo estimation to quantify the cost-normalized value of information (VOI) of each affordable diagnostic action, terminating exploration when no action provides sufficient cost-normalized VOI and subsequently committing to the intervention with the highest posterior cost-adjusted utility. After each diagnostic generation, ECHL updates the belief associated with the tested action and progressively reduces uncertainty as evidence accumulates, while consolidating the resulting effects to support selective reuse of prior evidence in future decisions and continual evolution of the harness.

In the experiments, we employ four intervention strategies tailored to compositional generation, namely reference image retention, prompt strengthening, structured planning, and extra resampling, across five architecturally diverse T2I generators: SDXL (Podell et al., 2024), SD3.5M (Stability AI, 2024), FLUX.2 Klein (Labs, 2025), Z-Image-Turbo (Cai et al., 2025), and Janus-Pro-7B (Chen et al., 2025b). Firstly, we construct a controlled dataset spanning six compositional capabilities to rigorously evaluate both task-level and atomic-level correctness. Experiments on this dataset demonstrate that T2I-SEH enables more effective intervention selection while incurring substantially lower action costs; moreover, results under fixed intervention strategies reveal the inherent limitations of uniform interventions, further underscoring the advantage of ECHL. Secondly, results on the GenEval 2 (Kamath et al., 2025), whose scenarios align with the predefined intervention

settings, observed gains extend beyond observer-induced improvements in internal optimization, consistently yielding higher final-generation quality. Thirdly, experiments under task–intervention mismatch on CoReBench (Li et al., 2026a) and DPG-Bench (Hu et al., 2024) demonstrate that T2I-SEH remains robust even when predefined interventions are substantially misaligned with the target task: ECHL selectively leverages the residual cross-task utility of beneficial interventions when such transferability exists, while otherwise defaulting to null intervention at negligible action cost.

In summary, our contributions are threefold:

- We introduce an evidence-based perspective on T2I harness self-evolution, treating prior experience as conditional support for runtime decisions and explicitly addressing when it should be trusted and how strongly it should influence subsequent actions.
- We propose T2I-SEH with Effect-Conditioned Harness Learning, integrating paired intervention effects, uncertainty-aware beliefs, cost-adjusted utility, and value of information to guide intervention, diagnosis, or image retention across generation episodes.
- Across five generators, experiments demonstrate improved intervention selection at substantially lower action cost, consistent gains on GenEval 2 under task–intervention alignment, and robust performance on CoReBench and DPG-Bench under task–intervention mismatch through selective transfer or null intervention at negligible cost.

2 Related Work

2.1 Agentic Text-to-Image Generation

T2I generation has increasingly shifted from reliance on a single model invocation toward an agentic inference process, in which a broader system interprets user requests, selects appropriate models or tools, evaluates intermediate outputs, and adaptively determines subsequent actions (Yang et al., 2024b; Zhao et al., 2024b; Wang et al., 2024b; Jia et al., 2025; Chen et al., 2025a). Furthermore, recent work has substantially expanded the reasoning, state, and action spaces of such systems (Jiang et al., 2026a; He et al., 2026b; Feng et al., 2026; Zhang et al., 2026; Ye et al., 2026), suggesting that T2I generation is increasingly being characterized as a sequential decision-making process over heterogeneous operations, rather than as a static mapping. Notably, a particularly relevant direction is to make agentic T2I inference experience-aware, enabling agents to leverage prior generation attempts, tool executions, and task trajectories to adaptively select and plan heterogeneous inference-time operations (Jiang et al., 2026c; Chen et al., 2026a;b). While prior work utilizes experience to adapt individual generation strategies, T2I-SEH shifts learning to the generation harness, asking whether a frozen model can be embedded in an experience-bearing control process that learns from its actions, reallocates future computation, and progressively improves its operational competence.

2.2 Self-Evolving Harness in Agentic System

As increasingly capable foundation models are deployed in agentic systems, adaptation is extending beyond model parameters to the persistent harness governing inference and interaction. Early work showed that agents can improve through feedback, episodic memory, and reusable skills without retraining the underlying model (Madaan et al., 2023; Shinn et al., 2023; Wang et al., 2024a). More recent studies make the harness itself an explicit target of evolution. ADAS (Hu et al., 2025) and AFlow (Zhang et al., 2025b) optimize agent architectures and executable workflows from execution feedback, while AutoHarness (Lou et al., 2026), Living-Harness (Du et al., 2026), SHAPER Wang et al. (2026), and HarnessEvolve (Jiang et al., 2026b) extend this paradigm to the evolution of prompts, procedural knowledge, reusable skills, runtime code, tools, and execution logic. Collectively, these works suggest that agentic capabilities can improve through refinement of the persistent harness surrounding the model. A similar trend is emerging in agentic T2I systems: GEMS (He et al., 2026c) introduces persistent trajectory memory and reusable skills, while COMFYCLAW (Li et al., 2026b) distills execution errors and visual feedback into an evolving skill library. By contrast, T2I-SEH does not expand the agent's skill set; instead, it leverages accumulated experience to reshape the evidence underlying evaluation, action selection, and calibrated reuse of prior experience.

3 PROBLEM FORMULATION

Let $x_t \in \mathcal{X}$ denote request t in an ordered stream $x_{1:T}$. A frozen generator G_ψ is accessed through a finite intervention registry $\mathcal{A}$, which includes the `Raw` action (a_0) and a set of corrective actions. Fixing the interventions and costs in $\mathcal{A}$, the registry preserves matched randomness across diverging paths. For each request t, the harness maintains a persistent experience memory $\mathcal{M}_t$ prior to processing, while a mandatory `Raw` call returns I_t^0 together with the context κ_t. Executing action $a \in \mathcal{A}(x_t)$ produces an output $y_t(a) = G_\psi(x_t, a)$ and incurs cost $c(a)$, which quantifies the additional generation operations performed under the protocol, with $c(a_0) = 0$ and $c(a) > 0$ for all $a \in \mathcal{A}_+(x_t)$. Furthermore, the harness evaluates outcomes using a policy-facing function $z_\omega : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. For each $a \in \mathcal{A}_+(x_t)$, the paired **effect** is defined as $\delta_t(a) = z_\omega\left(x_t, y_t(a)\right) - z_\omega\left(x_t, y_t(a_0)\right)$, with corresponding contextual estimand $\theta_a(\kappa_t) = \mathbb{E}[\delta_t(a)|\kappa = \kappa_t]$.

Before committing to output, the harness acquires diagnostic information by selecting $a_k \in \mathcal{A}_+(x_t)$ at round k, observing $\delta_t(a_k)$, and updating the local information state $D_{t,k}$. For K_t diagnostic tests, the total diagnostic cost satisfies $C_t^{\text{diag}} = \sum_{k=1}^{K_t} c(a_{t,k}) \le B_t$, where B_t is the hard per-request diagnostic budget. Upon committing to an intervention $\hat{a}_t \in \mathcal{A}(x_t)$, the total incremental generation cost is $C_t^{\text{total}} = C_t^{\text{diag}} + c(\hat{a}_t)$. We further define the cost-adjusted posterior **utility** as:

$$U_{t,k}(a) = \mathbb{E}[\theta_a(\kappa_t) \mid D_{t,k}, \mathcal{M}_t] - \lambda c(a), \tag{1}$$

where $\mathcal{M}_t$ is the experience available before request t and $\lambda \ge 0$ is fixed before evaluation. Once the decisions are finalized, the external evaluator reports the benchmark-native quality score together with the corresponding incurred cost:

$$Q_b = \frac{1}{|\mathcal{T}_b|} \sum_{t \in \mathcal{T}_b} q_b\left(x_t, y_t(\hat{a}_t)\right), \qquad C_b = \frac{1}{|\mathcal{T}_b|} \sum_{t \in \mathcal{T}_b} C_t^{\text{total}}, \tag{2}$$

where $\mathcal{T}_b$ is the evaluation set for benchmark b.

4 METHODOLOGY

We propose the T2I-SEH, a text-to-image self-evolving harness powered by Effect-Conditioned Harness Learning (ECHL), which integrates an effect-conditioned experience store and a cost-aware diagnostic strategy. Formally, we further characterize the harness dynamics as:

$$\mathcal{H}_t = (\mathcal{M}_t; \Omega), \quad \mathcal{B}_{t,0} = R_\Omega(\kappa_t, \mathcal{M}_t), \quad (\tau_t, \hat{a}_t) = \Pi_\Omega(\mathcal{B}_{t,0}, B_t), \quad \mathcal{M}_{t+1} = F_\Omega(\mathcal{M}_t; e_t), \tag{3}$$

where R_Ω initializes effect beliefs from retrieved experience, Π_Ω generates the diagnostic trace τ_t and committed action $\hat{a}_t$, and F_Ω updates the persistent memory with harness-visible tested events; the specification Ω fixes the generator, intervention registry, update rule, and decision thresholds.

4.1 EFFECT-CONDITIONED EXPERIENCE

A self-evolving T2I harness requires an experience representation that disentangles intervention efficacy from prompt difficulty, since absolute image scores may obscure this distinction: easy prompts can succeed irrespective of the applied intervention, whereas difficult ones may remain imperfect despite an effective correction. ECHL therefore encodes experience as paired improvement over the mandatory reference image under registered coupling.

For n tested outcomes sharing the same context–action key, the memory update aggregates their paired effects through the empirical mean $\bar{\delta}_n = 1/n \cdot \sum_{j=1}^{n} \delta_j$ and the corresponding empirical variance $\widehat{\sigma}_n^2 = 1/(n-1) \cdot \sum_{j=1}^{n} (\delta_j - \bar{\delta}_n)^2$, which serves as an estimate of the current uncertainty associated with the effect. Under a Gaussian approximation, we define the 95% confidence interval as $[\ell_n, u_n] = [\bar{\delta}_n - h_n, \bar{\delta}_n + h_n]$, with $h_n = 1.96\sqrt{\max(\widehat{\sigma}_n^2, 0.25)/n}$; the variance floor is introduced to prevent the interval from becoming spuriously narrow when the observed variability is small. In addition, a support-dependent confidence score is defined as $\chi_n = n/(n+2)$, assigning greater weight to context–action estimates supported by more observations.

Furthermore, given the observed failure context κ_t, relevant historical experience is retrieved by evaluating the structured similarity between κ_t and each stored context κ_i:

$$\text{sim}(\kappa_t, \kappa_i) = \lambda_{\text{failure}} s_{\text{failure}} + \lambda_{\text{generator}} s_{\text{generator}} + \lambda_{\text{severity}} s_{\text{severity}} + \lambda_{\text{context}} s_{\text{context}}. \tag{4}$$

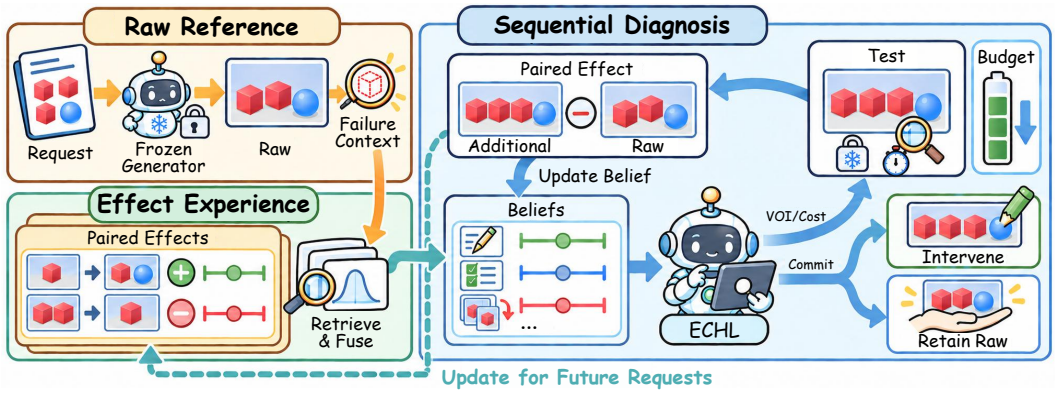

Figure 2: **Overview of ECHL.** ECHL combines effect-conditioned experience that captures paired effects with cost-aware sequential diagnosis that acquires information for cost-adjusted decision.

Accordingly, only active memory records with sufficient contextual similarity and empirical support are retained for retrieval. For request t, the resulting admissible experience set is defined as:

$$\mathcal{R}_t = \{m_i \in \mathcal{M}_t \mid \rho_i = \textsc{Active}, \mathrm{sim}(\kappa_t, \kappa_i) \geq 0.75,\ n_i \geq 10\}. \tag{5}$$

Let $\mathcal{R}_{t,a} = \{m_i \in \mathcal{R}_t \mid a_i = a\}$ denote this action-specific evidence set. For each $m_i \in \mathcal{R}_{t,a}$, we estimate the relevance w_i and the uncertainty v_i^M from the width of its effect interval:

$$w_i = \max\left\{\mathrm{sim}(\kappa_t, \kappa_i)\chi_i, 10^{-6}\right\}, \quad v_i^M = \max\left\{\left(\frac{u_i - \ell_i}{2 \times 1.96}\right)^2, 10^{-4}\right\}. \tag{6}$$

Hence, records receive greater influence on the updated belief when they are more similar to the current context, have higher confidence, or provide more precise effect estimate. Finally, the updated effect belief is obtained by integrating the frozen contextual prior with the retrieved historical evidence in a precision-weighted manner, providing the basis for subsequent action selection:

$$\mu_a = \left(\frac{\mu_a^0}{v_a^0} + \sum_i \frac{w_i}{v_i^M}\bar{\delta}_i\right) \Big/ \left(\frac{1}{v_a^0} + \sum_i \frac{w_i}{v_i^M}\right), \quad v_a = \left(\frac{1}{v_a^0} + \sum_i \frac{w_i}{v_i^M}\right)^{-1}. \tag{7}$$

The fused belief influences downstream decision-making through its mean and variance. Specifically, μ_a characterizes the expected effect of intervention a and thereby influences the ranking, while v_a quantifies the residual uncertainty and determines whether further diagnostic image generation is warranted and when it should terminate. In particular, when no action-specific historical evidence is available, i.e., $\mathcal{R}_{t,a} = \varnothing$, ECHL retains the original prior, yielding $(\mu_a, v_a) = (\mu_a^0, v_a^0)$.

4.2 Cost-Aware Sequential Diagnosis

Historical experience provides valuable guidance for future T2I decision-making, yet its reliability is inherently context-dependent, while exhaustive evaluation incurs substantial costs by requiring the generation of additional candidates. Accordingly, ECHL models each diagnostic generation as an information-acquisition step whose utility derives from reducing uncertainty sufficiently to alter the final cost-adjusted decision. Let $\mathcal{A}(x_t) \subseteq \mathcal{A}$ denote the feasible action subset for x_t, and define that $\mathcal{A}_+(x_t) = \mathcal{A}(x_t) \setminus \{a_0\}$. At diagnostic round k, ECHL summarizes the evidence available for request t, including the persistent memory $\mathcal{M}_t$ and the information after k diagnostic steps $D_{t,k}$, as a factorized Gaussian belief state $\mathcal{B}_{t,k} = \{(\mu_{t,k,a}, v_{t,k,a})\}_{a \in \mathcal{A}_+(x_t)}$, where the effect of each feasible corrective action a is approximated as $\theta_a(\kappa_t) \approx \mathcal{N}(\mu_{t,k,a}, v_{t,k,a})$. Combining with Equation (1), the cost-adjusted utility of each corrective action is given by:

$$U_{t,k}(a) = \mu_{t,k,a} - \lambda c(a). \tag{8}$$

Accordingly, the lower and upper bounds on the cost-adjusted utility are defined as:

$$\underline{U}_{t,k}(a) = \mu_{t,k,a} - \lambda c(a) - 1.96\sqrt{v_{t,k,a}}, \quad \overline{U}_{t,k}(a) = \mu_{t,k,a} - \lambda c(a) + 1.96\sqrt{v_{t,k,a}}. \tag{9}$$

When an action a is diagnostically tested, the observed paired effect $d_{t,k+1} = \delta_t(a)$ is modeled as a measurement of the action's latent effectiveness. The corresponding belief is then updated by integrating the new observation with the evidence accumulated prior to the diagnostic test:

$$\mu_{t,k+1,a} = v_{t,k+1,a}\left(\frac{\mu_{t,k,a}}{v_{t,k,a}} + \frac{d_{t,k+1}}{v_{\text{obs}}}\right), \quad v_{t,k+1,a} = \left(v_{t,k,a}^{-1} + v_{\text{obs}}^{-1}\right)^{-1}. \tag{10}$$

This precision-weighted update integrates the posterior mean $\mu_{t,k,a}$ with the observed effect $d_{t,k+1}$, while reducing the associated uncertainty from $v_{t,k,a}$ to $v_{t,k+1,a}$. Then, the updated beliefs are subsequently used to assess whether additional diagnostic generation is needed to improve the decision. Moreover, action with the highest posterior utility is defined as $a_{t,k}^{\mu} = \arg\max_{a\in\mathcal{A}(x_t)} U_{t,k}(a)$, with ties resolved in favor of `Raw`. The residual optimistic decision loss is then quantified as:

$$\mathcal{L}(\mathcal{B}_{t,k}) = \left[\max_{a\in\mathcal{A}(x_t)}\left(\overline{U}_{t,k}(a) - U_{t,k}(a_{t,k}^{\mu})\right)\right]_+, \tag{11}$$

where $[x]_+ = \max(x, 0)$. This quantity measures the gap between the most favorable utility under current uncertainty and the posterior mean utility of the selected action, with larger values indicating greater potential benefit from additional evidence.

Furthermore, the potential decision benefit of an additional diagnostic observation is estimated via Monte Carlo approximation under the current belief state. For each affordable corrective action a, M candidate outcomes are sampled from the predictive distribution $\mathcal{N}(\mu_a, v_a + v_{\text{obs}})$, and each sample m is propagated through Equation (10) to construct the corresponding belief state $\mathcal{B}^{a,m}$. The expected decision benefit of testing action a is quantified via Monte Carlo approximation to the value of information (VOI), defined as the expected reduction in decision loss of the test:

$$\widehat{\text{VOI}}(a) = \left[\mathcal{L}(\mathcal{B}) - \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}\left(\mathcal{B}^{a,m}\right)\right]_+. \tag{12}$$

Here, $\mathcal{L}(\mathcal{B})$ is the current optimistic decision loss, while the second term approximates its expected value after testing action a. Moreover, to account for heterogeneous generation costs, each affordable action is ranked by its estimated value of information per unit cost, $\widehat{\text{VOI}}(a)/c(a)$, and the highest-ranked action is selected for the next diagnostic test.

ECHL determines whether to continue diagnostic testing through following checks. First, it evaluates whether any corrective action can plausibly outperform the zero-utility `Raw` baseline; if $\max_{a\in\mathcal{A}_+(x_t)} \overline{U}_{t,k}(a) < 0$, no corrective action remains favorable even under its optimistic utility bound, and `Raw` is therefore retained. Second, early commitment is permitted when the leading corrective action is sufficiently supported by the current belief state: ECHL commits to the highest-mean-utility action a^{μ} only if its lower utility bound is positive and remains within ϵ of the strongest competing candidate's upper bound, i.e. $\underline{U}_{t,k}(a^{\mu}) > 0$ and $\underline{U}_{t,k}(a^{\mu}) \geq \max_{b\in\mathcal{A}(x_t)\setminus\{a^{\mu}\}} \overline{U}_{t,k}(b) - \epsilon$. Third, if the preferred action remains uncertain, an additional affordable diagnostic test is performed when $\max_a \left[\widehat{\text{VOI}}(a)/c(a)\right] > \eta$, ensuring that additional tests are conducted only when their expected decision VOI is sufficiently high relative to the cost. Finally, when no informative or affordable diagnostic test remains, ECHL terminates the diagnostic process and selects the action with the highest current utility $\hat{a}_t = \arg\max_{a\in\mathcal{A}(x_t)} U_{t,k}(a)$.

Additionally, the persistent memory enables continual adaptation across requests. Once ECHL commits to an action for request t, the paired effects observed during diagnosis are consolidated into experience records and stored in $\mathcal{M}_{t+1}$, so that the resulting evidence is available to subsequent decisions. When retrieved for a subsequent request, the relevant experience record is utilized to update both the expected utility and the associated uncertainty of the intervention. The algorithmic procedure is illustrated in Figure 2, and the corresponding pseudocode is provided in Algorithm 1.

5 Experiment

5.1 Experimental Setup

Tasks and Interventions. For simplicity, we apply the compositional generation setting in which the generated outputs can be readily verified. For each request, a shared predefined action set

Table 1: Task Success Rate (TSR), Constraint Satisfaction Rate (CSR), and mean action cost evaluated on our controlled dataset, using 360 instances for each generator under budgets $B \in \{1, 2, 4\}$.

Generator	B	TSR (%) ↑					CSR (%) ↑					Action Cost ↓	
		Raw	PS	SP	ER	**ECHL**	Raw	PS	SP	ER	**ECHL**	ER	**ECHL**
SDXL	1	30.83	35.83	27.78	-	**41.94**	39.58	42.94	35.75	-	**49.17**	-	**0.15**
	2	30.83	35.83	27.78	41.94	**45.00**	39.58	42.94	35.75	50.83	**52.67**	2.00	**0.21**
	4	30.83	35.83	27.78	41.94	**51.11**	39.58	42.94	35.75	50.83	**57.92**	2.00	**0.35**
SD3.5M	1	59.44	57.22	55.56	-	**65.00**	71.97	69.06	65.86	-	**78.06**	-	**0.10**
	2	59.44	57.22	55.56	**72.50**	69.72	71.97	69.06	65.86	**83.39**	81.83	2.00	**0.16**
	4	59.44	57.22	55.56	72.50	**74.17**	71.97	69.06	65.86	83.39	**85.67**	2.00	**0.25**
FLUX.2 Klein	1	68.33	68.06	68.89	-	**74.44**	75.28	73.86	75.47	-	**78.33**	-	**0.06**
	2	68.33	68.06	68.89	77.50	**79.72**	75.28	73.86	75.47	80.36	**82.14**	2.00	**0.11**
	4	68.33	68.06	68.89	77.50	**81.39**	75.28	73.86	75.47	80.36	**83.67**	2.00	**0.15**
Z-Image-Turbo	1	81.39	80.00	76.39	-	**86.11**	84.53	83.33	81.00	-	**86.83**	-	**0.05**
	2	81.39	80.00	76.39	88.89	**92.50**	84.53	83.33	81.00	89.56	**92.72**	2.00	**0.11**
	4	81.39	80.00	76.39	88.89	**93.61**	84.53	83.33	81.00	89.56	**93.83**	2.00	**0.13**
Janus-Pro-7B	1	56.94	51.67	5.56	-	**65.28**	65.31	56.83	22.17	-	**73.67**	-	**0.11**
	2	56.94	51.67	5.56	**69.17**	65.56	65.31	56.83	22.17	**78.64**	74.00	2.00	**0.11**
	4	56.94	51.67	5.56	69.17	**72.22**	65.31	56.83	22.17	78.64	**80.61**	2.00	**0.26**

comprising four types of interventions is considered: retaining the reference image (**`Raw`**), Prompt Strengthening (**`PS`**), Structured Planning (**`SP`**), and Extra Resample (**`ER`**). Specifically, we further assign cost to each intervention to reflect the additional overhead: `Raw` leaves the input unchanged and incurs zero additional cost; `PS` preserves the original task structure while explicitly strengthening requirements that are prone to failure, incurring an additional action cost of 1; `SP` first decomposes a complex request into structured constraints and then constructs the generation instructions accordingly, incurring an additional action cost of 1; and `ER` leaves the prompt unchanged but introduces one additional sampling opportunity, yielding two candidates in total, where the final output is selected by an internal selector, resulting in an additional action cost of 2. In our experiments, we set the budget to $B \in \{1, 2, 4\}$ and evaluate performance under different budget levels.

Research Questions. First, under a fixed generator, intervention set, and scoring rule, we investigate whether T2I-SEH can select more appropriate interventions with lower cost (**RQ1**). Second, we aim to examine whether the observed gains arise merely from improved optimization against the internal observer, or instead transfer into genuine improvements in final generation quality on compositional-generation benchmarks (**RQ2**). Finally, we investigate whether T2I-SEH can selectively exploit transferable interventions and otherwise abstain under task–intervention mismatch, while preserving the underlying generator's performance (**RQ3**).

Models and Datasets. To capture architectural diversity across modern generative models, we evaluate five representative generators. Specifically, Stable Diffusion XL (SDXL) (Podell et al., 2024), Stable Diffusion 3.5 Medium (SD3.5M) (Stability AI, 2024), FLUX.2 Klein (Labs, 2025), and Z-Image-Turbo (Cai et al., 2025) instantiate diffusion-based or flow-matching-based generation, while Janus-Pro-7B (Chen et al., 2025b) additionally represents the autoregressive paradigm. To address RQ1, we construct a controlled dataset comprising 360 test instances based on predefined templates, structured targets, and atomic evaluation constraints, with additional details provided in the Appendix C. Candidate outputs are evaluated using a frozen Qwen2.5-VL-3B (Bai et al., 2025b) as the atomic evaluator. To address RQ2, we conduct evaluation on the GenEval 2 benchmark (Kamath et al., 2025), adopting the Qwen3-VL-8B-Instruct-based Soft-TIFA implementation (Bai et al., 2025a). To address RQ3, we evaluate T2I-SEH under task–intervention mismatch on CoReBench (Li et al., 2026a) and DPG-Bench (Hu et al., 2024), which respectively target reasoning and long-prompt following, to assess whether ECHL can selectively exploit transferable interventions when beneficial while otherwise abstaining with low overhead. Notably, during evaluation, T2I-SEH has access only to the frozen generic atomic signal provided by the observer Qwen2.5-VL-3B, while the official benchmark scores remain strictly hidden from the system.

Table 2: Internal Success Rate (ISR), GenEval 2 Soft-TIFA, and mean action cost evaluated on the GenEval 2 compositional-generation benchmark under budgets $B \in \{1, 2, 4\}$.

Generator	B	ISR (%) ↑					GenEval 2 Soft-TIFA ↑					Action Cost ↓	
		Raw	PS	SP	ER	**ECHL**	Raw	PS	SP	ER	**ECHL**	ER	**ECHL**
SDXL	1	62.50	**82.00**	49.00	-	73.00	7.60	7.16	5.56	-	**8.07**	-	**0.11**
	2	62.50	82.00	49.00	76.13	**89.25**	7.60	7.16	5.56	8.33	**8.34**	2.00	**0.27**
	4	62.50	82.00	49.00	76.13	**91.25**	7.60	7.16	5.56	8.33	**8.41**	2.00	**0.31**
SD3.5M	1	78.88	90.75	74.25	-	**91.50**	19.37	16.27	10.87	-	**20.01**	-	**0.13**
	2	78.88	90.75	74.25	90.50	**96.75**	19.37	16.27	10.87	**20.69**	20.10	2.00	**0.18**
	4	78.88	90.75	74.25	90.50	**98.25**	19.37	16.27	10.87	**20.69**	20.28	2.00	**0.21**
FLUX.2 Klein	1	89.38	96.75	94.63	-	**98.38**	25.30	25.14	22.95	-	**25.96**	-	**0.09**
	2	89.38	96.75	94.63	95.63	**99.63**	25.30	25.14	22.95	26.02	**26.08**	2.00	**0.10**
	4	89.38	96.75	94.63	95.63	**99.75**	25.30	25.14	22.95	26.02	**26.08**	2.00	**0.11**
Z-Image-Turbo	1	94.63	98.75	97.25	-	**99.50**	31.33	29.78	27.04	-	**31.84**	-	**0.05**
	2	94.63	98.75	97.25	97.50	**99.63**	31.33	29.78	27.04	31.53	**31.85**	2.00	**0.05**
	4	94.63	98.75	97.25	97.50	**99.88**	31.33	29.78	27.04	31.53	**31.85**	2.00	**0.06**
Janus-Pro-7B	1	69.63	80.50	33.13	-	**87.88**	13.61	11.97	5.33	-	**14.33**	-	**0.18**
	2	69.63	80.50	33.13	80.00	**89.13**	13.61	11.97	5.33	14.43	**14.62**	2.00	**0.20**
	4	69.63	80.50	33.13	80.00	**91.00**	13.61	11.97	5.33	14.43	**14.78**	2.00	**0.23**

5.2 Better Intervention Selection at Lower Action Cost

In this section, we give an answer to RQ1: T2I-SEH selects more appropriate interventions while incurring lower action cost. First, we define Task Success Rate (TSR) as the proportion of tasks that satisfy all registered atomic constraints, thereby capturing complete task-level correctness, while Constraint Satisfaction Rate (CSR) measures partial correctness by averaging the proportion of satisfied atomic constraints across tasks. As shown in Table 1, ECHL consistently achieves a favorable trade-off between output quality and action cost across all five generators. Under the $B = 4$ setting, ECHL establishes a consistent advantage in both TSR and CSR over all baselines across the five generators. Relative to ER, the strongest fixed-intervention baseline, ECHL yields absolute gains ranging from 1.67 to 9.17 percentage points in TSR and from 1.97 to 7.09 percentage points in CSR.

Additionally, the fixed-intervention results further highlight the limitations of a one-size-fits-all strategy and motivate the need for intervention selection. PS and SP exhibit pronounced generator-specific behavior: the gains from PS are largely concentrated on SDXL and do not transfer consistently across generators, while SP incurs a substantial performance drop on Janus-Pro-7B. Despite ER being the strongest fixed-intervention baseline, ECHL consistently surpasses it across all five generators at $B = 4$, while requiring only a fraction of ER's action cost. Taken together, these results underscore the importance of conditioned intervention routing, rather than relying on a universally beneficial prompt strengthening or indiscriminate resampling strategy.

5.3 Internal Gains Transfer to External Generation Quality

In this section, we give an answer to RQ2: the observed gains extend beyond improvement induced by the internal observer, manifesting as substantive enhancements in the final generation quality that are consistently captured by the official compositional-generation benchmark score. As shown in Table 2, ECHL improves the official GenEval 2 Soft-TIFA score over the corresponding Raw baseline across all generators and evaluated budget settings, demonstrating that its gains extend beyond the internal evaluation signal. Moreover, it is worth noting that the rankings induced by the internal and external metrics are not fully consistent. For instance, on SDXL, PS substantially increases ISR from 62.50% to 82.00% while reducing the GenEval 2 official score from 7.60 to 7.16. Taken together, these results indicate that the official GenEval 2 score does not merely reflect the preferences of the internal observer; rather, the internal signal often serves as an informative yet

Table 3: Performance under task–intervention mismatch, evaluated by the official CoReBench and DPG-Bench scores, together with the mean action cost of ECHL under budgets $B \in \{1, 2, 4\}$.

Generator	B	CoReBench			DPG-Bench		
		Raw	ECHL	Action Cost	Raw	ECHL	Action Cost
SDXL	1	66.76	**66.92**	0.031	72.65	72.67	0.004
	2	66.76	**66.94**	0.045	72.65	**72.70**	0.008
	4	66.76	**66.89**	0.066	72.65	**72.77**	0.016
SD3.5M	1	69.75	69.78	0.015	84.27	84.28	0.004
	2	69.75	69.75	0.020	84.27	84.28	0.004
	4	69.75	69.77	0.031	84.27	84.28	0.004
FLUX.2 Klein	1	76.42	76.44	0.006	85.24	85.25	0.003
	2	76.42	76.44	0.014	85.24	85.25	0.003
	4	76.42	76.44	0.018	85.24	85.25	0.003
Z-Image-Turbo	1	75.87	75.89	0.008	85.34	85.33	0.004
	2	75.87	**75.95**	0.014	85.34	85.33	0.004
	4	75.87	**76.00**	0.019	85.34	85.33	0.004
Janus-Pro-7B	1	66.05	**66.26**	0.036	79.44	79.44	0.004
	2	66.05	**66.36**	0.044	79.44	79.44	0.004
	4	66.05	**66.41**	0.047	79.44	79.44	0.004

imperfect proxy for external generation quality, which ECHL leverages to achieve consistent gains over Raw despite generator-dependent variation in intervention effectiveness and gain magnitude.

5.4 Robustness under Task-Intervention Mismatch

In this section, we give an answer to RQ3: T2I-SEH remains robust under task–intervention mismatch by leveraging ECHL to selectively exploit residual cross-task utility when beneficial and otherwise abstain at negligible cost, thereby preserving the performance of the underlying generator. By construction, both CoReBench and DPG-Bench instantiate task–intervention mismatch settings, where the intervention operators developed for compositional generation are not explicitly designed to address reasoning or long-prompt tasks. On CoReBench, ECHL matches or surpasses Raw across all configurations, delivering an average improvement of 0.11 points and a maximum gain of 0.36, while incurring mean action costs of only 0.006-0.066, indicating that ECHL can identify and exploit the sparse subset of interventions that remains transferable to reasoning tasks. Moreover, DPG-Bench exhibits substantially less transferable utility, leading ECHL to predominantly select the null intervention and incur near-zero mean action costs of 0.003-0.016; correspondingly, performance remains highly stable relative to Raw, with changes ranging from a 0.01-point decrease to a 0.12-point increase. Taken together, these results show that ECHL adapts its intervention strategy to the degree of cross-task transferability, invoking beneficial interventions when transferable utility is present and otherwise abstaining to preserve the underlying generator's performance.

6 Conclusion

In this work, we introduce T2I-SEH, a self-evolving T2I harness that determines when prior experience should be trusted and how strongly it should influence subsequent actions. Through the Effect-Conditioned Harness Learning (ECHL), paired intervention effects are weighted by contextual relevance and uncertainty, then combined with cost-aware evidence acquisition to guide intervention, diagnosis, or image retention. Experiments across five generators demonstrate that T2I-SEH enables more effective intervention selection and achieves a favorable trade-off between generation quality and mean action cost, yielding consistent quality gains when tasks align with the intervention set. Under task–intervention mismatch, evaluations on CoReBench and DPG-Bench further demonstrate its robustness: T2I-SEH selectively applies transferable interventions when beneficial and otherwise preserves the original image, incurring negligible action cost. Future work could further extend the methodology to more complex automated generation and design scenarios, such as short-film generation and poster design, where reliable cross-stage experience reuse may offer additional benefits.

AI USE STATEMENT

Generative AI tools were employed to refine the manuscript language for clarity and readability, assist with code review and refinement, and provide suggestions on manuscript presentation. All AI-assisted outputs were carefully reviewed and, where necessary, revised by the authors. The authors take full responsibility for the final content of this work, including all text, code, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 5.1 specifies all essential experimental settings and implementation details. The proposed algorithm is further documented through pseudocode in the Appendix B, while Appendix F provides a complete specification of the hyperparameters used throughout the experiments. The supplementary material additionally includes the implementation code, configuration files, and a subset of the constructed compositional-generation dataset, with dependency files and testing utilities omitted. Upon acceptance, we plan to publicly release the complete codebase.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Paul Barbaste, Tristan Darrigol, Germain Vu, and Tom Wiltberger. Harness engineering: Anatomy, architecture, and evolution of coding agents–a source-code study of eleven systems. *arXiv preprint arXiv:2609.00006*, 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.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. In *International Conference on Learning Representations*, volume 2024, pp. 54067–54089, 2024.

Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19396–19405. IEEE, 2025a.

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.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Sanderson Oliveira de Macedo. What makes a harness a harness: necessary and sufficient conditions for an agent harness. *arXiv preprint arXiv:2606.10106*, 2026.

Yuetian Du, Yucheng Wang, He Xu, Jiexu Xu, Shanwen Tan, Bing Zhao, Boyu Yang, Zhijie Xu, Ming Kong, Hu Wei, et al. Living-harness is an interactive-agent evolver. *arXiv preprint arXiv:2607.26598*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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.

Chaoyue He, Xin Zhou, Di Wang, Hong Xu, Wei Liu, and Chunyan Miao. Harness engineering for language agents: The harness layer as control, agency, and runtime. *Preprints*, April 2026a. doi: 10.20944/preprints202603.1756.v2.

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*, 2026b.

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*, 2026c.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, volume 2025, pp. 21344–21377, 2025.

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.

Chengyou Jia, Changliang Xia, Zhuohang Dang, Weijia Wu, Hangwei Qian, and Minnan Luo. Chat-gen: Automatic text-to-image generation from freestyle chatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13284–13293. IEEE, 2025.

Kaixun Jiang, Yuzheng Wang, Junjie Zhou, Pandeng Li, Zhihang Liu, Chen-Wei Xie, Zhaoyu Chen, Yun Zheng, and Wenqiang Zhang. Genagent: Scaling text-to-image generation via agentic multi-modal reasoning. *arXiv preprint arXiv:2601.18543*, 2026a.

Wen Jiang, Mingmin Chu, Yimeng Tian, Qianxin Zhang, Haofei Yang, Rui Yang, Yang Liu, Tao Lv, and Fangming Li. Harnessevolve: Learning from reference trajectories for reliable agent self-evolution. *arXiv preprint arXiv:2609.00829*, 2026b.

Xu Jiang, Bin Chen, Gehui Li, Yule Duan, Ronggang Wang, and Jian Zhang. Octot2i: A self-evolving agentic text-to-image router. *arXiv preprint arXiv:2606.01803*, 2026c.

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.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *International Conference on Learning Representations*, volume 2026, pp. 86729–86758, 2026a.

Zongxia Li, Dawei Liu, Fuxiao Liu, Yuhang Zhou, Xiyang Wu, Jingxi Chen, Jing Xie, Xiaomin Wu, and Lichao Sun. Comfyclaw: Self-evolving skill harnesses for image generation workflows. *arXiv preprint arXiv:2607.01709*, 2026b.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Kai Mei, Xi Zhu, Wujiang Xu, Mingyu Jin, Wenyue Hua, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. AIOS: LLM agent operating system. In *Second Conference on Language Modeling*, 2025.

Qianyu Meng, Yanan Wang, Liyi Chen, Yihang Li, Wei Wu, Wenyuan Jiang, Qimeng Wang, Chengqiang Lu, Yan Gao, Yi Wu, and Yao Hu. Agent harness for large language model agents: A survey. *Preprints*, April 2026. doi: 10.20944/preprints202604.0428.v3.

Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, volume 2026, pp. 94327–94354, 2026.

Linyue Pan, Lexiao Zou, Shuo Guo, Jingchen Ni, and Hai-Tao Zheng. Natural-language agent harnesses. *arXiv preprint arXiv:2603.25723*, 2026.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Jiahao Qiu, Xuan Qi, Tongcheng Zhang, Xinzhe Juan, Jiacheng Guo, Yifu Lu, Yimin Wang, Zixin Yao, Qihan Ren, Xun Jiang, et al. Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. *arXiv preprint arXiv:2505.20286*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Yu Shang, Yu Li, Keyu Zhao, Likai Ma, Jiahe Liu, Fengli Xu, and Yong Li. Agentsquare: Automatic llm agent search in modular design space. In *International Conference on Learning Representations*, volume 2025, pp. 3841–3865, 2025.

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

Avi Singh, John D Co-Reyes, and Rishabh Agarwal. Beyond human data: Scaling self-training for problem-solving with language models. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024.

Stability AI. Stable diffusion 3.5 medium. Hugging Face Model Card, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-medium.

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.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856.

Peidong Wang, Zhiming Ma, Ying Chang, Xufang Luo, Xiaocui Yang, Shi Feng, Yuqing Yang, and Dongsheng Li. Self-evolving embodied agents via skill-harness evolution. *arXiv preprint arXiv:2608.11350*, 2026.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025a.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37: 128374–128395, 2024b.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 63897–63911. PMLR, 2025b.

Georg Wölflein, Dyke Ferber, Daniel Truhn, Ognjen Arandjelovic, and Jakob Nikolas Kather. Llm agents making agent tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26092–26130, 2025.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024a.

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, 2024b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

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.

Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi Fung, Hao Peng, and Heng Ji. Craft: Customizing llms by creating and retrieving from specialized toolsets. In *International Conference on Learning Representations*, volume 2024, pp. 40097–40125, 2024a.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. Evoagent: Towards automatic multi-agent generation via evolutionary algorithms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6192–6217, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 57905–57923. PMLR, 2024b.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-STar: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*, 2024.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, LEI BAI, and Xiang Wang. Multi-agent architecture search via agentic supernet. In *Forty-second International Conference on Machine Learning*, 2025a.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. In *International Conference on Learning Representations*, volume 2025, pp. 34040–34077, 2025b.

Xulu Zhang, Xiaoyong Wei, Jinlin Wu, Jiaxin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Generating on generated: An approach towards self-evolving diffusion models. *arXiv preprint arXiv:2502.09963*, 2025c.

Yuxuan Zhang, Haoyang Yu, Lanxiang Hu, Haojian Jin, and Hao Zhang. General modular harness for llm agents in multi-turn gaming environments. *arXiv preprint arXiv:2507.11633*, 2025d.

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*, 2026.

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

Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, and Rongrong Ji. Diffagent: Fast and accurate text-to-image api selection with large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6390–6399. IEEE, 2024b.

Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*, 2024.

A Extended Related Work

A.1 Text-to-Image Generation

T2I generation has rapidly evolved from standalone generative models to large-scale visual foundation models with increasingly advanced capabilities. Recent T2I models can be broadly grouped into two architectural paradigms: latent-space generative models based on diffusion or flow matching, and autoregressive models. Latent Diffusion Models (LDMs) shifted denoising from pixel space to a compressed latent space, improving the scalability of high-resolution generation (Rombach et al., 2022), a paradigm later extended by Stable Diffusion XL (SDXL) (Podell et al., 2024). More recent models increasingly adopt Transformer backbones and continuous flow-matching objectives, as exemplified by Stable Diffusion 3.5 with MMDiT and rectified flow (Esser et al., 2024; Stability AI, 2024), and FLUX.2 with billion-scale rectified-flow Transformers (Labs, 2025). In parallel, autoregressive (AR) image generation has emerged as a competitive alternative to diffusion-based approaches by formulating visual synthesis as sequential prediction over image tokens. Janus (Wu et al., 2025) introduces a decoupled visual encoding design built upon a shared autoregressive Transformer backbone, while Janus-Pro (Chen et al., 2025b) further improves the framework through refined training strategies and increased data and model scale.

A.2 Self-Evolving Paradigm in Agentic Systems

Self-evolving agentic systems seek to continuously improve through experience acquired from interaction, with such experience inducing persistent changes that affect subsequent episodes. Existing approaches can be broadly grouped into three levels: model, capability, and system evolution. Model-level evolution internalizes accumulated self-generated experience into model parameters through iterative self-training, self-play, preference optimization, or reinforcement learning, thereby consolidating acquired experience into the underlying model weights. Representative approaches progressively bootstrap reasoning, generation, or evaluation capabilities from model-generated supervision and feedback (Zelikman et al., 2022; Singh et al., 2024; Chen et al., 2024; Yuan et al., 2024b; Zelikman et al., 2024; Kumar et al., 2025). Capability-level evolution expands and refines an agent's executable action space through the acquisition of reusable skills and the development of new tools, thereby extending the set of available operations without modifying the underlying model. Prior work has enabled agents to build executable skill libraries, synthesize task-specific tools, discover reusable APIs, and autonomously improve their capabilities (Wang et al., 2024a; Cai et al., 2024; Yuan et al., 2024a; Wölflein et al., 2025; Zheng et al., 2025; Qiu et al., 2025). System-level evolution reshapes the persistent structures surrounding the foundation model, encompassing memory, prompts, workflows, agent architectures, orchestration policies, and execution logic. One line of work accumulates and refines interaction-derived experience through episodic memory, reflective mechanisms and structured experience representations (Park et al., 2023; Shinn et al., 2023; Zhong et al., 2024; Zhao et al., 2024a; Xu et al., 2025; Wang et al., 2025b; Ouyang et al., 2026). Another line of work directly optimizes higher-level agent structures, including computational graphs, workflow configurations, module compositions, and execution harnesses (Zhuge et al., 2024; Yuan et al., 2025; Hu et al., 2025; Zhang et al., 2025b; Shang et al., 2025; Zhang et al., 2025a). Collectively, these developments reflect a broader shift from model-centric self-improvement toward persistent system-level self-evolution built around increasingly capable foundation models.

A.3 Harness for Agentic Systems

The capabilities of modern agents are increasingly determined not only by the underlying foundation model, but also by the external execution layer that governs interaction, state management, tool use, and control flow. Early agentic systems embodied key aspects of this layer through explicit reasoning-acting loops, programmable workflows, and orchestration mechanisms. ReAct (Yao et al., 2023) interleaves reasoning with environment interaction, while AutoGen (Wu et al., 2024) and MetaGPT (Hong et al., 2024) structure agent behavior through explicit interaction protocols and workflow-based coordination. A complementary line of work emphasizes the importance of the interface and runtime infrastructure surrounding a fixed model. SWE-agent (Yang et al., 2024a) introduces the Agent-Computer Interface, demonstrating that carefully designed interaction primitives can substantially enhance agent performance without modifying the underlying model parameters. OpenHands (Wang et al., 2025a) extends this perspective through an integrated runtime that sup-

ports model invocation, tool execution, and sandboxed environments, while AIOS (Mei et al., 2025) further addresses safe execution, scheduling, context management, and resource control. More recently, these previously fragmented mechanisms have increasingly been unified under the concept of an agent harness. General Modular Harness (Zhang et al., 2025d) decomposes the surrounding system into reusable functional modules, while Natural-Language Agent Harnesses (Pan et al., 2026) externalize high-level harness control logic into portable, inspectable executable artifacts. Recent work on harness engineering further characterizes the harness as a persistent runtime layer responsible for control flow, context management, tool exposure, state management, safety enforcement, and orchestration (He et al., 2026a; Meng et al., 2026; Barbaste et al., 2026). Collectively, these studies reflect a broader shift from treating agent scaffolding as auxiliary implementation infrastructure to recognizing the harness as a distinct system-level abstraction that governs how capabilities of foundation models are realized during interaction and execution.

B Algorithm Pseudocode

Algorithm 1 Effect-Conditioned Harness Learning for Request t

Require: Request x_t, memory $\mathcal{M}_t$, budget B_t, and fixed Harness specification Ω
Ensure: Action $\hat{a}_t$, output y_t, trace τ_t, cost C_t^{total}, and memory $\mathcal{M}_{t+1}$

$(I_t^0, \kappa_t) \leftarrow \textsc{Raw}(x_t; \Omega)$ ▷ establish reference image and failure context
$(\mathcal{R}_t, \mathcal{B}_{t,0}) \leftarrow \textsc{RetrieveAndFuse}(\kappa_t, \mathcal{M}_t; \Omega)$ ▷ initialize beliefs from relevant experience
$k \leftarrow 0$; $D_{t,0} \leftarrow \langle\rangle$; $\tau_t \leftarrow \langle\rangle$; $C_t^{\text{diag}} \leftarrow 0$; $\hat{a}_t \leftarrow \perp$
while $\hat{a}_t = \perp$ **do**
  $a^{\text{stop}} \leftarrow \textsc{ConfidentDecision}(\mathcal{B}_{t,k}; \epsilon)$ if $\mathcal{R}_t \neq \varnothing$ or $k > 0$; otherwise $a^{\text{stop}} \leftarrow \perp$
  if $a^{\text{stop}} \neq \perp$ **then**
    $\hat{a}_t \leftarrow a^{\text{stop}}$; **break**
  end if
  $\mathcal{F}_{t,k} \leftarrow \{a \in \mathcal{A}_+(x_t) : C_t^{\text{diag}} + c(a) \leq B_t\}$ ▷ retain only affordable diagnostic actions
  if $\mathcal{F}_{t,k} = \varnothing$ **then**
    $\hat{a}_t \leftarrow \textsc{BestMeanAction}(\mathcal{B}_{t,k})$; **break**
  end if
  $\{\widehat{\text{VOI}}(a)\}_{a \in \mathcal{F}_{t,k}} \leftarrow \textsc{EstimateVOI}(\mathcal{B}_{t,k}, \mathcal{F}_{t,k}; \Omega)$ ▷ simulate decision value of each test
  $a^{\text{diag}} \leftarrow \arg\max_{a \in \mathcal{F}_{t,k}} \widehat{\text{VOI}}(a)/c(a)$ ▷ prefer higher information gain per unit cost
  if $\widehat{\text{VOI}}(a^{\text{diag}})/c(a^{\text{diag}}) \leq \eta$ **then**
    $\hat{a}_t \leftarrow \textsc{BestMeanAction}(\mathcal{B}_{t,k})$; **break**
  end if
  $(I^{\text{diag}}, d) \leftarrow \textsc{PairedEffect}(x_t, a^{\text{diag}}, I_t^0; \Omega)$ ▷ execute test and measure improvement
  $\mathcal{B}_{t,k+1} \leftarrow \textsc{UpdateActionBelief}(\mathcal{B}_{t,k}, a^{\text{diag}}, d; \Omega)$ ▷ update belief of tested intervention
  $D_{t,k+1} \leftarrow \textsc{Append}(D_{t,k}, \langle a^{\text{diag}}, d\rangle)$; $\tau_t \leftarrow \textsc{Append}(\tau_t, \langle a^{\text{diag}}, d, c(a^{\text{diag}})\rangle)$
  $C_t^{\text{diag}} \leftarrow C_t^{\text{diag}} + c(a^{\text{diag}})$; $k \leftarrow k + 1$
end while
$y_t \leftarrow \textsc{Commit}(x_t, I_t^0, \hat{a}_t; \Omega)$; $C_t^{\text{total}} \leftarrow C_t^{\text{diag}} + c(\hat{a}_t)$
$e_t \leftarrow \textsc{ProjectExperience}(\kappa_t, D_{t,k})$; $\mathcal{M}_{t+1} \leftarrow F_\Omega(\mathcal{M}_t; e_t)$
return $(\hat{a}_t, y_t, \tau_t, C_t^{\text{total}}, \mathcal{M}_{t+1})$

C More Details on the Controlled Dataset

The controlled dataset spans six capability dimensions: counting, exact text rendering, spatial relations, attribute binding, multi-object composition, and knowledge-grounded generation. Together, these categories capture distinct generation challenges, while also providing diverse settings in which corrective interventions can take effect. Counting requests specify an exact number of objects; text requests specify a string that must appear on a sign; spatial and binding requests test whether objects are placed in the designated positions while preserving their assigned attributes. Multi-object requests impose multiple simultaneous constraints on object identity, size, color, count, and relative position. Knowledge-grounded requests additionally require the inclusion of a factually

correct visual element, such as a national flag, together with its prescribed relation to another object. Balanced capability coverage mitigates bias from overrepresented requirement types in the aggregate success rate. For comparison, all policies are evaluated using an identical set of requests and a shared candidate pool, which is composed of the raw generation, a prompt-strengthened candidate, a structured-planning candidate, and two resampled candidates. Accordingly, this design enables a controlled comparison of alternative intervention choices under identical requests, facilitating a systematic analysis of their respective effects. Each generated image is evaluated against a set of atomic constraints using a frozen internal evaluator. Furthermore, the corresponding test specifications are curated by human experts, whereas all reported judgments are produced automatically by the evaluator, without relying on human annotations of the generated images.

D EXPERIMENT: EXPERIENCE PROPAGATE TO SUBSEQUENT REQUESTS

In this section, we conduct additional tests to give an answer to the following question: once new evidence generated during the request stream is written to memory, is it retrieved on subsequent requests, thereby altering future diagnostic and decision trajectories?

D.1 SETUP AND METRICS

To isolate the contribution of memory write-back, STATIC (S) and EVOLVING (E) are initialized with the same memory $\mathcal{M}_0$ and evaluated under the identical request order, generator, internal observer, and budget. Before each request, STATIC resets its memory to $\mathcal{M}_0$, whereas EVOLVING carries forward the evidence obtained by the diagnostics actually executed for the completed decision. Similarly, external quality scores are incorporated only after the decision has been finalized.

Let $\mathcal{R}_t^E$ denote the evidence retrieved by EVOLVING for request t, and let $\mathcal{W}_{<t}$ denote the collection of evidence records written upon completion of all requests preceding t in the stream. For each execution variant $v \in \mathrm{S}, \mathrm{E}$, we define the corresponding logged trajectory as $\tau_t^v = \big(d_t^v, c_t^v, z_t^v, a_t^v\big)$, where d_t^v denotes the ordered diagnostic sequence, c_t^v the incurred cost, z_t^v the termination reason, and a_t^v the final intervention selected by the policy. In addition, immediately after evidence retrieval and before any within-request diagnostic action is taken, we construct the policy-facing belief vector $\boldsymbol{b}_t^v$ by concatenating, for each candidate intervention, its estimated mean together with the corresponding lower and upper confidence bounds. Finally, $\mathbf{1}[\cdot]$ denotes the indicator function.

Throughout the subsequent evaluation, we report results in terms of three metrics, namely New-Evidence Use Rate (NEUR), Diagnostic Change Rate (DCR), and Intervention Change Rate (ICR). First, NEUR quantifies the proportion of requests for which previously written evidence is subsequently retrieved and corresponds to an observable shift in the controller's belief state or execution trajectory. We define that: $r_t = \mathbf{1}\left[\mathcal{R}_t^E \cap \mathcal{W}_{<t} \neq \varnothing\right]$, $h_t = \mathbf{1}\left[\|\boldsymbol{b}_t^E - \boldsymbol{b}_t^S\|_\infty > \xi \ \vee \ \tau_t^E \neq \tau_t^S\right]$, and then NEUR can be defined as $\mathrm{NEUR} = 1/T \cdot \sum_{t=1}^{T} r_t h_t$. Notably, NEUR captures cases in which the retrieved evidence influences the controller's internal beliefs, even when the resulting intervention remains unchanged. DCR quantifies the proportion of requests for which the two executions diverge in their diagnostic sequence, diagnostic cost, termination reason, or final intervention. We define that: $\mathrm{DCR} = 1/T \cdot \sum_{t=1}^{T} \mathbf{1}[\tau_t^E \neq \tau_t^S]$. ICR further isolates changes in the final intervention itself and is formally defined as: $\mathrm{ICR} = 1/T \cdot \sum_{t=1}^{T} \mathbf{1}[a_t^E \neq a_t^S]$.

D.2 RESULTS AND ANALYSIS

As shown in Table 4, both NEUR and DCR remain consistently high across all experimental settings, with NEUR ranging from 95.84% to 97.35% and DCR exhibiting a similarly narrow variation from 95.04% to 97.05%. This consistency extends across generator architectures, with NEUR and DCR at $B = 4$ confined to the narrow ranges of 96.77%–97.35% and 96.18%–97.05%, respectively. The close correspondence suggests that retrieving qualifying prior evidence is typically associated with a corresponding shift in the diagnostic trajectory. Moreover, ICR remains consistently and substantially lower than DCR across the evaluation, ranging from 1.97%–22.71%, a pattern suggesting that evidence-induced adaptation is concentrated primarily in the intermediate diagnostic process rather than in the final intervention. Additionally, increasing the budget from $B = 2$ to $B = 4$ has negligible influence on the diagnostic-level metrics: NEUR remains unchanged for all generators, while

Table 4: Experience Retrieve and Use. New-Evidence Use Rate (NEUR), Diagnostic Change Rate (DCR), and Intervention Change Rate (ICR) evaluated on GenEval 2 under budgets $B \in \{1, 2, 4\}$.

Generator	B	Metrics (%)		
		NEUR	DCR	ICR
SDXL	1	95.84	95.40	22.71
	2	97.34	97.01	17.01
	4	97.34	96.97	16.77
SD3.5M	1	96.09	95.55	12.69
	2	97.07	96.79	11.69
	4	97.07	96.87	11.66
FLUX.2 Klein	1	96.01	95.09	5.22
	2	96.96	96.44	4.81
	4	96.96	96.44	4.81
Z-Image-Turbo	1	95.97	95.04	1.98
	2	96.77	96.18	1.97
	4	96.77	96.18	1.97
Janus-Pro-7B	1	95.97	95.53	21.31
	2	97.35	96.99	13.86
	4	97.35	97.05	13.84

DCR varies by no more than 0.08 points. However, the degree to which such diagnostic adaptation propagates to the final intervention is model-dependent: at $B = 4$, ICR reaches 16.77% for SDXL and 13.84% for Janus-Pro-7B, whereas substantially lower values are observed for FLUX.2 Klein (4.81%) and Z-Image-Turbo (1.97%). This disparity suggests that prior evidence can substantially alter the decision-making process without changing the final intervention.

E Memory Record Example

In this section, we present an illustrative example of the memory record. As shown below, the example corresponds to a complete SDXL memory record. For space efficiency, we set sample count $= 8$ in this example; consequently, the record has not yet reached the retrieval threshold of 10 samples.

Listing 1 Illustration of a SDXL Memory Record.


```
{
"context_signature": {
    "context_id": "sdxl_base_1_0::multi_object_composition::FAIL::HIGH",
    "count_bucket": "multi_object_composition:HIGH",
    "generator_id": "sdxl_base_1_0"
  },
  "created_at": "",
  "effect_lower": -0.27148232278140827,
  "effect_mean": 0.075,
  "effect_upper": 0.4214823227814083,
  "generator_id": "sdxl_base_1_0",
  "harness_version": "ECHL_FINAL_V1",
  "incremental_cost": 1.0,
  "intervention_id": "PROMPT_STRENGTHENING_V1",
  "memory_id": "mem-71770e79893c-prompt_strengthening_v1",
  "sample_count": 8,
  "source_task_ids": [
    "s6-0-4-000",
    "s6-0-4-001",
    "s6-0-4-003",
    "s6-0-4-004",
    "s6-0-4-005",
    "s6-0-4-006",
    "s6-0-4-008",
    "s6-0-4-009"
  ],
  "status": "ACTIVE"
}
```

F HYPERPARAMETERS

To further facilitate reproducibility, we provide a comprehensive specification of the hyperparameters used throughout our experiments in Table 5.

Table 5: Experimental hyperparameters.

Parameter	Value
Cost penalty coefficient λ	0.05
Failure similarity coefficient λ_{failure}	0.40
Generator similarity coefficient $\lambda_{\text{generator}}$	0.25
Severity similarity coefficient $\lambda_{\text{severity}}$	0.20
Context similarity coefficient λ_{context}	0.15
Monte Carlo outcome number M	128
$\widehat{\text{VOI}}(a)/c(a)$ threshold η	0.002
Early-commitment tolerance ϵ	0.03

G CASE STUDIES

In this section, we further illustrate the self-evolving decision-making process through case studies. These examples highlight evolving as a sequential procedure in which evidence is progressively accumulated and intervention decisions are iteratively revised. An intervention that yields no observable effect does not necessarily terminate the search; instead, such an outcome can reduce uncertainty and alter which intervention is most informative or valuable to evaluate next. Meanwhile, a positive effect may increase the posterior confidence associated with an intervention without dictating the final decision. Across case studies, the diagnostic trajectory emerges from the interplay among retrieved experience, observed intervention effects, and testing costs, which jointly govern the downstream selection of the final generated image.

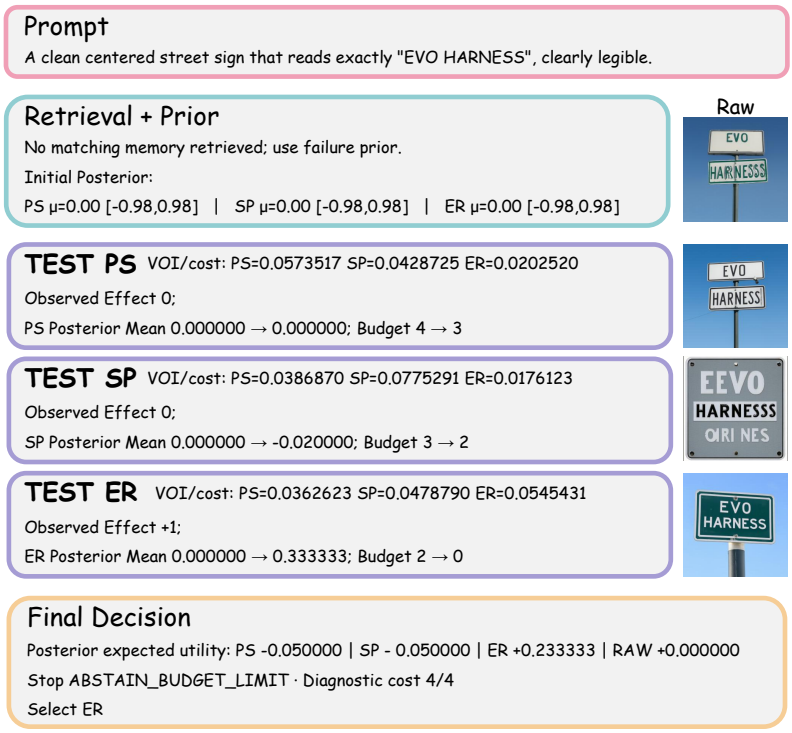

Figure 3: Case-study illustration with SD3.5M.



Figure 4: Case-study illustration with SDXL.

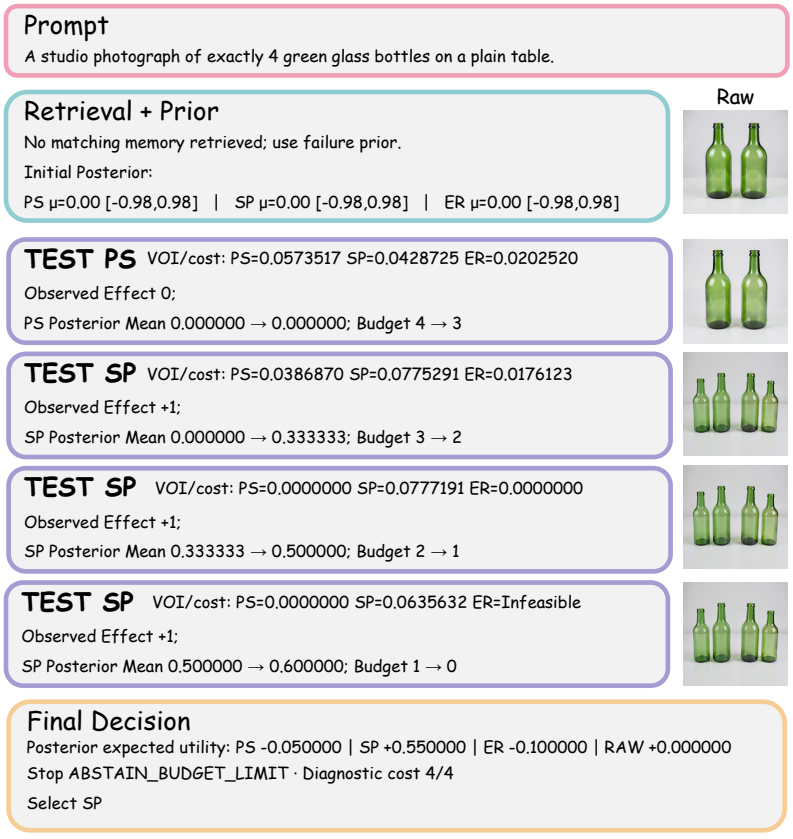

Figure 5: Case-study illustration with FLUX.2 Klein.

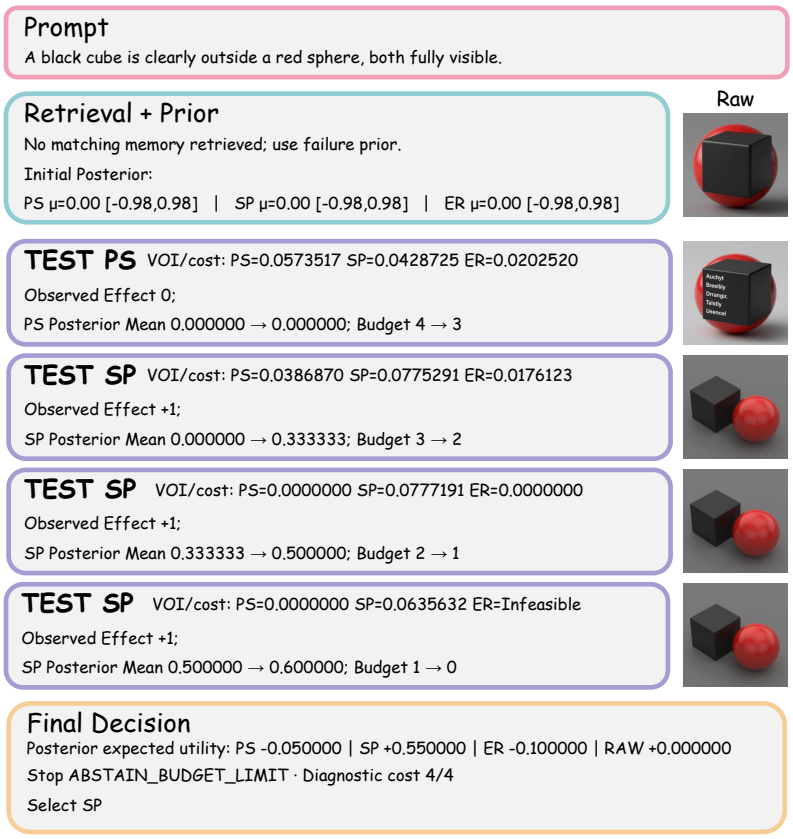

Figure 6: Case-study illustration with Z-Image-Turbo.

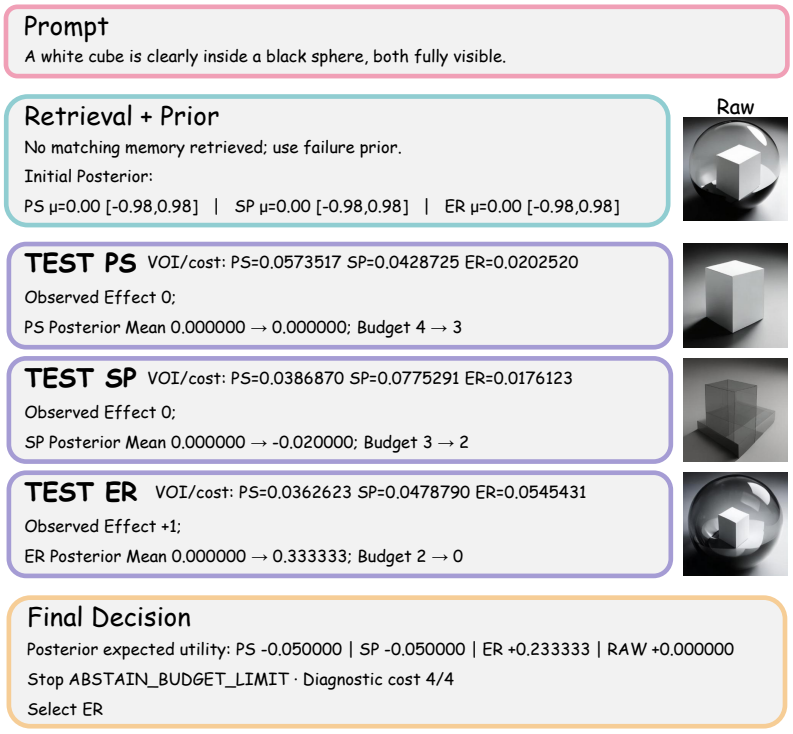

Figure 7: Case-study illustration with Janus-Pro-7B.