REWARD-GUIDED NOISE-LEVEL TRAJECTORIES AND TASK-AWARE EVALUATION FOR IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Instruction-guided diffusion and flow-based image editing requires selecting a noise-level trajectory whose effectiveness can vary with the source image and instruction. Selecting such trajectories requires a reliable editing reward, yet the evidence needed to assess quality varies across tasks despite the shared goals of faithful editing, content preservation, and visual integrity. We address these two challenges, input-adaptive trajectory selection and task-aware reward construction. Amortized Trajectory Selection (ATS) learns a lightweight policy optimized for a chosen editing reward to predict a complete trajectory from each source–instruction pair before execution, without test-time optimization or online reward queries. To our knowledge, ATS is the first image-editing method to learn such direct per-input trajectory prediction without test-time optimization. Geometry-Normalized Trajectory Selection (GeoTS) complements ATS by making the sampling trajectory itself a structured source of candidate variation without training a selector or iteratively optimizing individual trajectories. To provide a reliable reward across editing tasks, we introduce EditCal, which selects and combines complementary signals from existing evaluation models according to the task and criterion, calibrates them to shared human-anchored ordinal semantics, and integrates them into an overall reward with criterion-specific failure diagnostics. On FLUX.1-Kontext and FlowEdit, ATS improves mean reward over native schedules across all four evaluated objectives and retains its gains at unseen sampling-step counts without retraining. Across three public human-judgment benchmarks, EditCal improves evaluation agreement and failure detection.

1 INTRODUCTION

Diffusion and flow-based Text-to-Image models (Ho et al., 2020; Lipman et al., 2023; Liu et al., 2023) support instruction-guided image editing (Brooks et al., 2023; Fu et al., 2024; Sheynin et al., 2024), where the source image and instruction jointly determine what should change and what should be preserved. During sampling, the editor performs updates at a sequence of noise levels, and each update depends on the current noise level. With a finite number of sampling steps, the choice and spacing of these levels can therefore change the edited output. This sequence defines the noise-level trajectory, making it a natural control variable for editing without changing the pretrained editor, with effective trajectory choices potentially depending on the source–instruction context. Selecting among such trajectories therefore requires a reward that reliably measures editing quality. However, the evidence needed to assess editing quality varies across editing tasks, or taxonomies, such as object and attribute changes, background replacement, and style transfer. Existing evaluation models provide complementary signals, while a successful edit is defined by the same three criteria across tasks: realizing the requested change (Edit), preserving the remaining content (Preserve), and maintaining visual integrity (Clean). This raises two coupled challenges: selecting effective trajectories for each source–instruction context, and integrating task-appropriate evidence from different evaluators under shared editing criteria into a reliable reward.

On the trajectory side, previous image-editing methods have explored fixed noise schedules, joint timestep–noise control, and reward-guided reverse-dynamics control (Lin et al., 2024; Chen et al., 2024; Chang et al., 2026). These approaches are either tied to inversion-specific error characteristics or require iterative per-input optimization at inference time. For editing evaluation, prior work introduced taxonomy-aware benchmarks (Ryu et al., 2025; Ma et al., 2024), context-aware metrics

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.

or multimodal judges (Kim et al., 2025; Ku et al., 2024), and editing-specific reward models (Wu et al., 2026a; Luo et al., 2026). These methods provide complementary evidence, but their outputs differ in scale, interpretation, and the editing aspects they measure, complicating their integration into a consistent reward.

To address the trajectory-selection challenge, we treat the sampling trajectory as a common control interface applicable to frozen editors regardless of their internal editing dynamics. We represent each complete trajectory as a parameterized monotone function over normalized sampling progress, separating trajectory choice from finite-step discretization. Based on this representation, we propose Amortized Trajectory Selection (ATS), which learns a lightweight policy to predict a complete trajectory for each source–instruction context under a chosen editing reward. Because the trajectory is represented continuously, the same prediction can be used at different numbers of sampling steps, enabling zero-shot length transfer without retraining. We further introduce Geometry-Normalized Trajectory Selection (GeoTS), which makes the sampling trajectory itself a controllable source of candidate variation, complementary to conventional noise resampling. Around a reference trajectory, GeoTS constructs a compact geometry-normalized candidate set and selects among its variants using the chosen reward, without additional selector training or per-input parameter optimization.

On the evaluation side, we introduce EditCal, a task-aware framework that routes complementary signals from existing evaluators to the tasks and criteria where each is most informative. These signals are calibrated under shared Edit, Preserve, and Clean criteria into human-anchored ordinal distributions, providing criterion-specific failure probabilities and graded assessments. A copula-based joint model captures dependencies among the criteria to derive a consistent overall editing reward. Once fitted, EditCal scores each candidate from its source image, instruction, and task taxonomy, providing both failure diagnostics and a reward for trajectory control. Across three public human-judgment benchmarks, EditCal improves human agreement and criterion-specific failure detection. On FLUX.1-Kontext and FlowEdit, ATS improves mean reward over native trajectories for all four evaluated objectives and retains its gains at unseen sampling-step counts, while GeoTS demonstrates complementary, editor-dependent benefits from trajectory-space variation.

Our contributions are threefold: (1) ATS enables source- and instruction-conditioned complete-trajectory prediction without test-time optimization or online reward queries. (2) GeoTS introduces geometry-normalized candidate variation directly in trajectory space. (3) EditCal provides task-aware integration of complementary evaluation evidence into criterion-specific diagnostics and a consistent editing reward.

2 Related Work

Image editing and model adaptation. Image editing includes training-free methods that reuse pre-trained Text-to-Image diffusion or flow models through inference-time control (Meng et al., 2022; Hertz et al., 2023; Couairon et al., 2023; Cao et al., 2023; Rout et al., 2025; Kulikov et al., 2025), and editors trained or adapted specifically for instruction-based editing (Brooks et al., 2023; Zhang et al., 2025; Black Forest Labs et al., 2025; Parihar et al., 2026). Recent work further post-trains editing models using human or learned reward and preference signals (Zhang et al., 2024; Chen et al., 2025; Luo et al., 2026; Li et al., 2025; 2026; Wu et al., 2026b).

Sampling and trajectory control. In image generation, finite-step sampling has been studied through solver (Lu et al., 2022; Zhao et al., 2023) and noise-schedule design (Lu et al., 2022; Sabour et al., 2024), with recent work exploring adaptive or learned schedules (Ye et al., 2025; Huang et al., 2026). Image-editing methods instead use schedules motivated by inversion-specific error characteristics (Lin et al., 2024), per-edit optimization of timesteps and input noise (Chen et al., 2024), or reward-guided reverse-state control (Chang et al., 2026). ATS learns a source–instruction-conditioned noise-level trajectory for a frozen editor without per-input optimization, while GeoTS explores candidate variation directly in schedule space.

Image-editing evaluation and rewards. Editing evaluation uses task-specific criteria (Ryu et al., 2025; Ma et al., 2024; Qian et al., 2025), context-aware or multimodal judgments (Kim et al., 2025; Ku et al., 2024), and editing-specific reward models (Chen et al., 2025; Wu et al., 2026a; Luo et al., 2026). Recent evaluators further learn multidimensional scores from large-scale human feedback (Xu et al., 2026) or ground judgments in explicit spatial reasoning (Long et al., 2026).

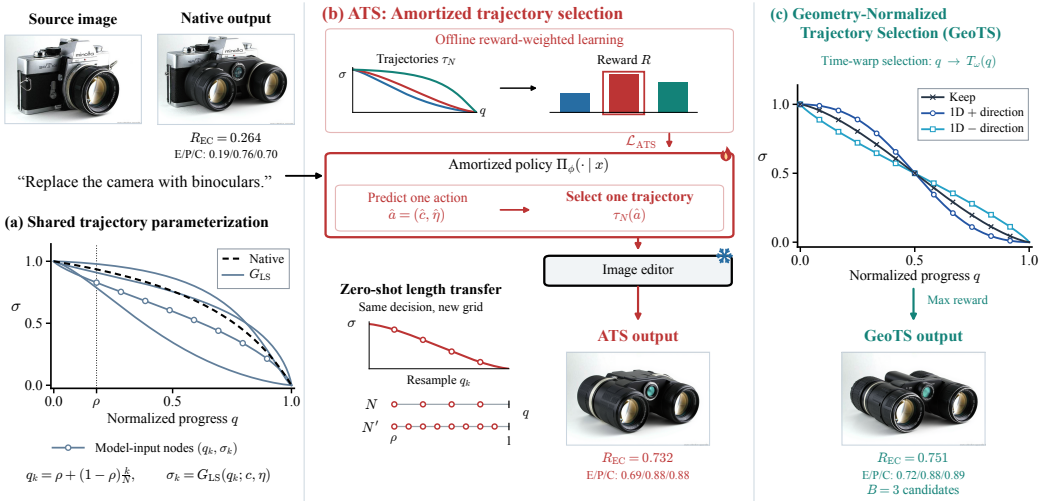


Figure 1: **Reward-guided trajectory control overview.** (a) A complete noise-level trajectory is represented by the monotone function G_{LS} and sampled over the editor-specific active domain $[\rho, 1]$. (b) ATS learns from offline reward supervision to predict one trajectory for each source–instruction context before editor execution. The same continuous trajectory can be used at different numbers of sampling steps. (c) GeoTS constructs geometry-normalized trajectory variations around a reference using monotone time warps, with the chosen reward selecting the final output.

3 METHOD

Our trajectory-control formulation treats the noise trajectory as the controllable interface of a frozen editor. For a specified terminal reward, we learn or search for trajectory parameters that yield high-reward edits while keeping the editor weights unchanged.

Our design of EditCal starts from the observation that no single evaluator is equally informative for every aspect of every editing task. We therefore treat existing evaluation models as sources of evidence and route task-appropriate evidence to three editing criteria: (*Edit*, E), (*Preserve*, P), and (*Clean*, C). This lets each evidence model contribute where its signal is most relevant, rather than forcing heterogeneous scores into a common role. We then use human annotations to calibrate these criterion-specific signals into a shared ordinal scale, enabling both interpretable criterion-level diagnostics and a unified editing reward.

3.1 FUNCTIONAL NOISE TRAJECTORY REPRESENTATION

We formulate trajectory control by separating the trajectory itself from its discretization. Instead of optimizing a step-specific sequence of noise levels, we represent the complete sampling trajectory as a parameterized function that can be evaluated at different numbers of sampling steps (Figure 1(a)). Let F denote a frozen editor and ξ the Gaussian noise used during sampling. Given editing input x and noise trajectory $\tau = (\sigma_0, \ldots, \sigma_N)$, the editor produces $y = F(x, \xi; \tau)$. For a fixed scalar reward $R(x, y)$, our goal is to choose a monotone trajectory that maximizes the expected reward:

$$\tau^\star(x) \in \arg\max_{\tau} \mathbb{E}_\xi[R(x, F(x, \xi; \tau))]\,. \tag{1}$$

For normalized progress $q \in [0, 1]$, a custom action $a = (c, \eta)$ selects the logit-sigmoid trajectory

$$G_{\mathrm{LS}}(q; c, \eta) = \mathrm{sigmoid}\left(e^\eta \left[c - \log\frac{q}{1-q}\right]\right), \qquad 0 < q < 1. \tag{2}$$

Its endpoints are fixed to 1 at $q = 0$ and 0 at $q = 1$. Parameter c controls the half-noise location where $q = 0.5$, while η controls the sharpness of the trajectory. Each editor specifies a fixed entry coordinate $\rho \in [0, 1)$ on this normalized trajectory domain. Given a selected action, the editor executes the restriction of G_{LS} to the active domain $[\rho, 1]$. Thus, the first executed noise level

is determined by the same trajectory function as $\sigma_0 = G_{\mathrm{LS}}(\rho; c, \eta)$. For N sampling steps, the selected function is sampled uniformly over this interval:

$$q_k = \rho + (1-\rho)\frac{k}{N}, \qquad \sigma_k = G_{\mathrm{LS}}(q_k; c, \eta), \qquad k = 0, \ldots, N. \tag{3}$$

A custom action therefore determines all executable noise levels jointly through a single trajectory function on $[\rho, 1]$. Holding the action and entry coordinate fixed while changing N only changes the discretization of that trajectory. Custom trajectory actions lie in a bounded support $\mathcal{A} \subset \mathbb{R}^2$, restricting extreme shifts and sharpness values. We also retain the editor-native scheduler as a separate trajectory decision and define $\mathcal{B} = \{\mathrm{Native}\} \cup \mathcal{A}$.

3.2 Amortized Trajectory Selection (ATS)

For each editor–reward pair, ATS learns a lightweight policy $\pi_\phi(a \mid x)$ that predicts a custom action $a \in \mathcal{A}$ from the source–instruction context before editor execution (Figure 1(b)). The editor parameters remain unchanged, and a single policy prediction replaces per-input trajectory search. For each training input x, we construct an offline bank of Native and custom trajectories and score their completed edits using the target reward R. Let $\mathcal{T}_x$ denote the retained high-reward candidates, with decision b_i and reward u_i for $i \in \mathcal{T}_x$. Rather than supervising the policy only with the single best trajectory, we retain relative reward information through the soft teacher distribution

$$Q_i = \frac{\exp(u_i/\nu)}{\sum_{j \in \mathcal{T}_x} \exp(u_j/\nu)}, \qquad i \in \mathcal{T}_x, \tag{4}$$

where $\nu > 0$ controls the reward scale. For training, we use an auxiliary likelihood $\Pi_\phi(b \mid x)$ over $\mathcal{B}$, while ATS always selects a custom action at inference time. A confidence weight $w_x \in [0, 1]$ gives greater importance to inputs for which the retained trajectories are more clearly separated from lower-reward alternatives. The policy is trained by weighted behavior cloning:

$$\mathcal{L}_{\mathrm{ATS}}(\phi) = -\frac{\sum_x w_x \sum_{i \in \mathcal{T}_x} Q_i \log \Pi_\phi(b_i \mid x)}{\sum_x w_x}. \tag{5}$$

Because the action specifies a trajectory function rather than a step-specific sequence, the same prediction can be resampled at different numbers of sampling steps, N, without retraining.

3.3 Geometry-Normalized Trajectory Selection (GeoTS)

GeoTS provides a controllable source of sampling variation in trajectory space, complementary to variation from noise resampling (Figure 1(c)). Its goal is to construct a compact and structured neighborhood around a reference trajectory, so that trajectory variation can be explored without training a selector or optimizing a separate schedule for every candidate. Once geometry-normalized search directions are fixed, a single radius parameter α controls the magnitude of trajectory variation. Rather than perturbing executable noise levels directly, GeoTS applies monotone time warps to preserve monotonicity and the full reference endpoints at $q = 0$ and $q = 1$.

Let $\sigma_{\mathrm{keep}}(q)$ denote the reference trajectory and $\omega = (c_w, \eta_w) \in \mathbb{R}^2$ the warp parameters. We define

$$T_\omega(q) = \mathrm{sigmoid}\left(e^{\eta_w} \log \frac{q}{1-q} + c_w\right), \qquad \widetilde{\sigma}_\omega(q) = \sigma_{\mathrm{keep}}(T_\omega(q)). \tag{6}$$

The identity $\omega = \mathbf{0}$ recovers the reference. When $\sigma_{\mathrm{keep}} = G_{\mathrm{LS}}(\cdot; c, \eta)$, the warped trajectory can again be written as $G_{\mathrm{LS}}(\cdot; c', \eta')$, with $c' = e^{-\eta_w}(c - c_w)$ and $\eta' = \eta + \eta_w$. GeoTS therefore varies the trajectory while retaining the same functional representation used by ATS.

Equal perturbations in the warp coordinates can nevertheless produce very different changes in the noise levels actually executed by the editor. GeoTS therefore normalizes the search using local geometry in executable noise space. Let $\boldsymbol{\sigma}(\omega) = (\widetilde{\sigma}_\omega(q_k))_{k=0}^{N-1}$ collect the model-input noise levels and let J be its Jacobian at the reference. We define

$$J = \left.\frac{\partial \boldsymbol{\sigma}(\omega)}{\partial \omega^\top}\right|_{\omega=\mathbf{0}} \in \mathbb{R}^{(N+1)\times 2}, \qquad H = \frac{J^\top J}{N}, \qquad \frac{\|\boldsymbol{\sigma}(\omega) - \boldsymbol{\sigma}(\mathbf{0})\|_2^2}{N} \approx \omega^\top H \omega. \tag{7}$$

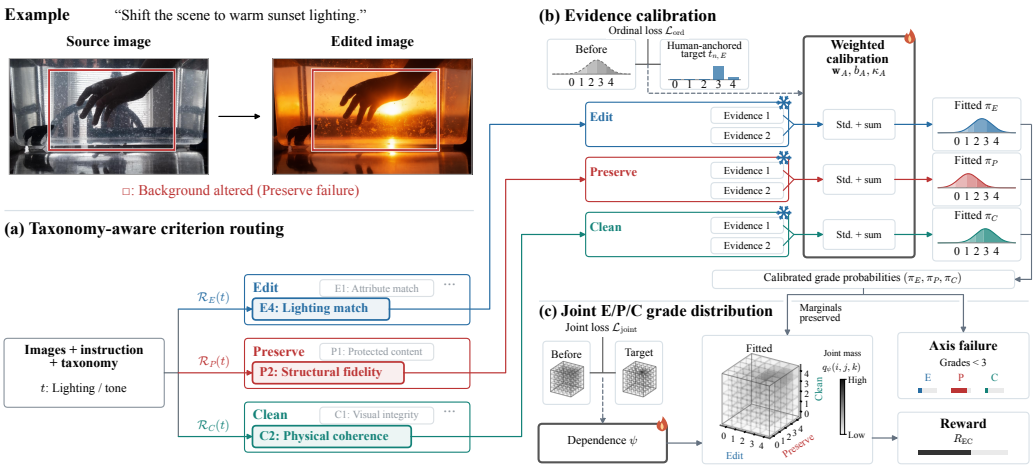


Figure 2: **EditCal framework overview.** The example illustrates an unintended Preserve failure. (a) The editing taxonomy selects task-appropriate criterion routes for Edit, Preserve, and Clean, determining which evidence is used for each axis. (b) The routed frozen evidence is calibrated against human-anchored targets into ordinal grade distributions with shared semantics. (c) A joint model captures dependence among the three criteria while preserving their calibrated marginal distributions, producing criterion-specific failure probabilities and the overall EditCal reward.

Starting from the sharpness and shift directions, we apply Gram–Schmidt under a regularized form of H to obtain normalized search directions d_j. Their local scale is controlled by a single relative radius α, which determines how broadly the trajectory neighborhood is explored, and B is the number of trajectory candidates, including the reference.

$$r = \alpha \frac{\|\boldsymbol{\sigma}(\mathbf{0})\|_2}{\sqrt{N}}, \qquad \mathcal{C}_B = (\mathbf{0}, +rd_1, -rd_1, \ldots, +rd_m, -rd_m), \qquad B = 1 + 2m. \tag{8}$$

We use one normalized direction as default. The reward then selects among these trajectory-induced variants. Appendix A.3 details the geometry construction and the interpretation of α as an approximate relative RMS displacement in executable noise levels.

3.4 EditCal: Semantic Calibration of Editing Quality

Given an editing input x containing taxonomy t and candidate edit y, EditCal predicts a calibrated ordinal distribution for each of the three criteria $\mathcal{I} = \{E, P, C\}$, corresponding to Edit, Preserve, and Clean. Let Z_A denote the grade for axis $A \in \mathcal{I}$ on the shared ordinal space $\mathcal{Z} = \{0, \ldots, K-1\}$. We use $K = 5$ grades, ranging from complete failure to clear success, with $k_{\text{pass}} = 3$ as the minimum passing grade. For each axis, a fixed router $\mathcal{R}_A(t)$ selects the appropriate evidence and calibration configuration using the supplied taxonomy. Further details on the criteria, task-specific evidence routes, and calibration procedure are provided in Appendix B. Evidence and calibration can therefore differ across editing tasks, while the grade semantics and pass boundary remain shared. To anchor these shared semantics to human judgments, we obtain soft ordinal targets using an ordinal latent-response model (Uebersax & Grove, 1993) and align auxiliary API annotations (OpenAI, 2026a) to the same human-defined grade space.

For axis $A \in \mathcal{I}$, the router selects the relevant frozen evidence outputs, which we collect in a feature vector $\mathbf{s}_A(x, y)$. We standardize these features using training sources, then combine them into a route-specific scalar evidence score $\ell_A(x, y)$. Rather than directly rescaling this score, EditCal learns ordered thresholds that map it to the shared grade semantics. With $\kappa_{A,1} < \cdots < \kappa_{A,K-1}$, cumulative-probit calibration (McCullagh, 1980) gives

$$\pi_A(k \mid x, y) = \Phi(\kappa_{A,k+1} - \ell_A(x, y)) - \Phi(\kappa_{A,k} - \ell_A(x, y)), \tag{9}$$

where $\pi_A(k \mid x, y) = \Pr(Z_A = k \mid x, y)$, Φ is the standard normal CDF, and the boundary thresholds use the usual infinite endpoints. The score and thresholds are fitted to the aligned ordinal targets. The resulting ordinal distribution preserves graded information rather than collapsing

each criterion to a binary outcome. From the same distribution, the shared passing threshold yields criterion-specific success and failure probabilities. The failure probability p_A^{fail} provides an interpretable diagnostic, while the full distribution retains the severity of failure and strength of success.

$$p_A(x, y) = \sum_{k=k_{\text{pass}}}^{K-1} \pi_A(k \mid x, y), \qquad p_A^{\text{fail}}(x, y) = 1 - p_A(x, y). \tag{10}$$

The three calibrated marginals characterize each criterion individually, but do not specify how their grades are jointly distributed. We therefore fit a shared checkerboard copula after the axis calibrators are fixed. The copula models dependence among the three criteria and produces a joint grade distribution $q_\psi(\mathbf{z} \mid x, y)$ over $\mathbf{z} = (z_A)_{A \in \mathcal{I}} \in \mathcal{Z}^D$, where $D = |\mathcal{I}|$. Its uniform marginals preserve each calibrated axis distribution exactly:

$$\Pr_{\mathbf{Z} \sim q_\psi(\cdot|x,y)}(Z_A = k) = \pi_A(k \mid x, y). \tag{11}$$

From this joint distribution, we define the scalar EditCal reward as

$$R_{\text{EC}}(x, y) = \mathbb{E}_{\mathbf{Z} \sim q_\psi(\cdot|x,y)} \left[\left(\prod_{A \in \mathcal{I}} \frac{Z_A}{K-1} \right)^{1/D} \right]. \tag{12}$$

The geometric mean is applied to each joint grade outcome before taking its expectation, so that a low grade on any criterion lowers the joint quality while graded distinctions among outcomes are retained. The resulting reward lies in $[0, 1]$ and serves as the terminal objective when EditCal is used for trajectory control, while the calibrated axis distributions remain available as criterion-specific diagnostics. Figure 2 summarizes the overall EditCal pipeline.

4 Experiments

We evaluate reward-guided trajectory learning, editing quality, candidate variation, and zero-shot length transfer on two frozen editors. We then assess EditCal's agreement with human judgments and criterion-specific failure detection, followed by ablations of ATS and EditCal.

4.1 Experimental Setup

Trajectory experiments use frozen FLUX.1-Kontext-dev (Kontext) (Black Forest Labs et al., 2025) and FlowEdit (Kulikov et al., 2025) on 150 held-out test tasks from our STACE-22K dataset (Appendix E), spanning nine taxonomies, with two noise repetitions per task. We train a separate ATS policy with approximately 2M trainable parameters for each editor–objective pair using ImageReward (Xu et al., 2023), EditScore (Luo et al., 2026), EditReward (Wu et al., 2026a), or EditCal. Native denotes the editor-native schedule and each trajectory uses $N = 12$ denoising updates unless otherwise stated. Cross-metric editing quality is evaluated using these four rewards and two large vision-language model (LVLM) judges, Qwen3.8-27B (Qwen Team, 2026) and GPT6-Sol (OpenAI, 2026b), each scoring edits from 0 to 4. Full settings are provided in Appendix A.

We compare EditCal with EditScore, EditReward, SpatialReward (Long et al., 2026), EditHF (Xu et al., 2026), HATIE* (Ryu et al., 2025), and AugCLIP (Kim et al., 2025) on three public human-annotated benchmarks using their official protocols. HATIE* denotes our extension of HATIE that replaces its category-limited mask component with open-vocabulary SAM3 (Carion et al., 2026) segmentation, enabling evaluation across all images. EditCal combines frozen EditScore and EditReward components with task-appropriate evidence for Preserve and Clean (Table 8; Appendix B.1). For failure detection, Edit and Clean use benchmark human annotations, while Preserve uses our human-reviewed matched-control panel of compositional-drift failures and successful edits. The entire intervention is excluded from EditCal training. Full protocols and panel construction are provided in Appendix B.4.

4.2 Reward-Guided Trajectory Control

Table 1 compares each objective-specific ATS policy with Native, using the same objective for training and evaluation. At $N = 12$, ATS achieves higher mean reward for all four objectives on both

Table 1: **Objective-guided trajectory control.** ATS is trained and evaluated with each listed reward objective at $N = 12$. Scores are taxonomy-macro means. Brackets show pointwise 95% paired bootstrap CIs for differences from Native. Bold marks the higher mean.

	Kontext		FlowEdit	
Objective	Native	ATS	Native	ATS
EditCal ↑	0.659	**0.679** [0.012, 0.030]	0.505	**0.542** [0.019, 0.057]
ImageReward ↑	0.016	**0.050** [−0.002, 0.075]	0.151	**0.624** [0.370, 0.577]
EditScore ↑	6.645	**6.809** [−0.001, 0.334]	4.970	**5.407** [0.033, 0.875]
EditReward ↑	1.028	**1.110** [0.044, 0.122]	−0.148	**−0.049** [0.021, 0.179]

Table 2: **Editing quality of EditCal-trained ATS.** Native and ATS outputs at $N = 12$ are evaluated with all six metrics. For each method, the same generated outputs are used across evaluations. Scores are taxonomy-macro means. Brackets show pointwise 95% paired bootstrap CIs for differences from Native. Bold marks the higher mean.

	Kontext		FlowEdit	
Metric	Native	ATS	Native	ATS
EditCal ↑	0.659	**0.679** [0.012, 0.030]	0.505	**0.542** [0.019, 0.057]
ImageReward ↑	0.016	**0.097** [0.036, 0.128]	0.151	**0.484** [0.231, 0.442]
EditScore ↑	6.645	**6.971** [0.154, 0.507]	4.970	**5.504** [0.134, 0.949]
EditReward ↑	1.028	**1.167** [0.092, 0.188]	−0.148	**−0.030** [0.014, 0.232]
Qwen3.8-27B ↑	2.798	**2.939** [0.055, 0.236]	0.944	**1.038** [−0.089, 0.296]
GPT6-Sol ↑	2.573	**2.691** [0.048, 0.193]	0.899	**1.120** [0.122, 0.325]

editors without test-time optimization or online reward queries. ATS also exceeds both variants of our FlowEdit adaptation of NaviEdit (Lu et al., 2026) across all four objectives (Appendix Table 13).

Table 2 reuses the fixed Native and EditCal-trained ATS outputs and evaluates them with all six metrics. ATS has higher mean scores under every metric on both editors, including three metrics not incorporated into EditCal: ImageReward, Qwen3.8-27B, and GPT6-Sol. These cross-metric gains suggest that the improvement is not confined to the optimized reward. Additional qualitative comparisons are shown in Appendix Figure 7.

Using the same EditCal-trained ATS, Table 3 compares noise and trajectory variation under equal three-candidate budgets. ATS Bo3 varies noise with the trajectory fixed, whereas GeoTS Bo3 varies trajectories with noise fixed ($\alpha = 0.1$). EditCal selects one candidate per method, which is then evaluated with all six metrics without reselection. ATS Bo3 has the higher mean on five of six metrics on Kontext, whereas GeoTS Bo3 has the higher mean on four of six metrics on FlowEdit, indicating editor-dependent benefits of trajectory variation relative to noise resampling. Figure 3 and Appendix Figure 8 provide qualitative comparisons.

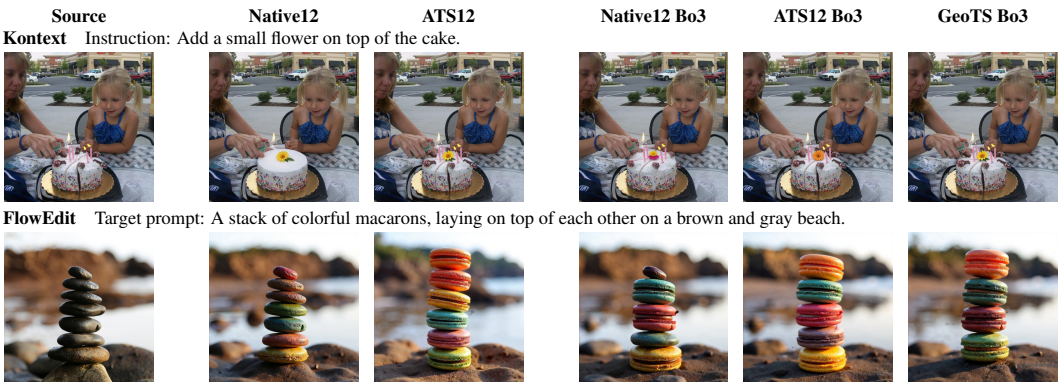


Figure 3: **Reward-guided trajectory control.** Qualitative comparisons of Native, ATS, noise sampling, and trajectory search on Kontext and FlowEdit.

Table 3: **Noise versus trajectory variation.** ATS Bo3 and GeoTS Bo3 use the same EditCal-trained ATS and three candidates of 12 updates each. ATS Bo3 varies noise seeds, while GeoTS varies trajectories with $\alpha = 0.1$. EditCal selects one candidate, which is then evaluated with all six metrics without reselection. Scores are taxonomy-macro means. Bold marks the higher mean.

	Kontext		FlowEdit	
Metric	ATS Bo3	GeoTS Bo3	ATS Bo3	GeoTS Bo3
EditCal ↑	**0.718**	0.700	0.612	**0.625**
ImageReward ↑	0.116	**0.128**	**0.571**	0.493
EditScore ↑	**7.520**	7.286	6.678	**6.807**
EditReward ↑	**1.349**	1.275	0.218	**0.309**
Qwen3.8-27B ↑	**3.099**	3.006	1.330	**1.408**
GPT6-Sol ↑	**2.906**	2.720	**1.392**	1.379

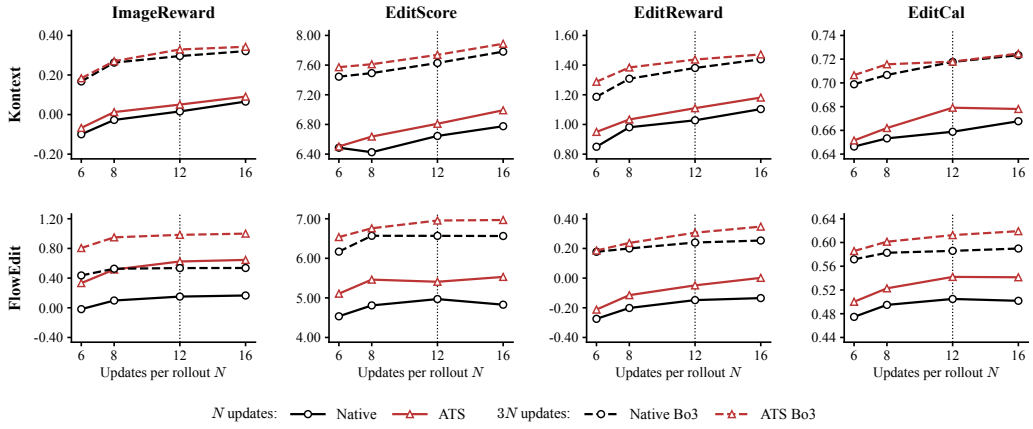


Figure 4: **Zero-shot trajectory length transfer.** ATS policies trained at $N = 12$ are reused at $N = 6, 8, 16$ with fixed decoded decisions. Columns correspond to training rewards. Solid and dashed curves show single-candidate and Bo3 noise-sampling results, respectively.

For zero-shot length transfer, we keep each $N = 12$-trained editor–objective policy and its decoded input-specific trajectory fixed, changing only the sampling grid. ATS achieves higher mean reward than the corresponding Native baseline at $N = 6, 8, 16$ for every objective on both editors, in both single-candidate and Bo3 noise-sampling settings (Figure 4). Complete numerical results, including Bo5 and GeoTS, are reported in Appendix Table 12.

4.3 EditCal Evaluation

Table 4 evaluates agreement with human judgments across three public benchmarks. EditCal ranks first on 9 of 12 endpoints, second on two, and third on one. Table 5 evaluates criterion-specific failure detection. EditCal leads on 8 of 9 metrics, including all Preserve and Clean metrics, with AUCs of 0.9740 and 0.9526, respectively, and the highest AP and normalized AP for Edit failures. Figure 5 illustrates how criterion-specific scores distinguish Preserve and Clean failures from successful edits, with additional Preserve failure–control examples in Appendix Figure 9.

4.4 Ablation Study

Appendices C.2 and C.3 report detailed ablations. Removing image or prompt conditioning lowers ATS mean reward on both editors, while a no-context Global policy improves over Native but remains below Full ATS, supporting source–instruction conditioning (Table 14). For EditCal, removing the joint stage, alignment, or calibration lowers Overall agreement on every benchmark, while mean and learned weighted fusion of the same evidence also underperform the full model (Table 16).

Table 4: **Alignment with human judgments on three editing benchmarks.** We follow each benchmark's official evaluation code and protocol using its human annotations. K counts ranked candidates, and ρ denotes Spearman correlation. Brackets identify Edit (E), Preserve (P), and Clean (C) targets. Higher is better, with bold and underline marking the best and second-best results.

	EditReward-Bench						EditScore-Bench			ImagenHub		
Method	**IF** [E]	**VQ** [C]	**K=2** [EPC]	**K=3** [EPC]	**K=4** [EPC]	**Overall** [EPC]	**PF** [E]	**Cons.** [P]	**Overall** [EPC]	**SC** ρ [EP]	**PQ** ρ [C]	**Overall** ρ [EPC]
EditScore-4B	0.4181	0.5196	0.6617	0.3000	0.0811	0.5432	0.6129	0.6015	0.7004	0.4020	0.4693	0.3924
EditScore-8B	0.4349	0.5041	0.6090	0.3133	0.0743	0.5445	0.6253	0.6022	0.6935	0.3887	0.4651	0.3787
EditReward	0.6483	0.6495	**0.7068**	0.3267	0.0946	0.5955	0.8444	0.7100	**0.8263**	0.3507	0.4037	0.3611
SpatialReward	0.4549	0.5104	0.6241	0.3600	0.0541	0.5914	0.6864	0.6567	0.7928	0.3509	0.4462	0.3438
EditHF	0.5276	0.5851	0.4962	0.2533	0.0541	0.5316	0.5439	**0.7591**	0.6254	0.1158	0.3251	0.1740
HATIE*	0.5208	0.5551	0.5263	0.2333	0.0811	0.5391	0.6469	0.7296	0.5916	0.2173	0.0845	0.1987
AugCLIP	0.5359	0.5816	0.5338	0.2133	0.0743	0.5357	0.5553	0.7507	0.5955	0.0750	0.2221	0.0928
EditCal **(Ours)**	**0.6528**	**0.6517**	0.6692	**0.3933**	**0.1081**	**0.6077**	**0.8575**	0.7411	0.8123	**0.4149**	**0.5100**	**0.4048**

Table 5: **Detection of human-labeled Edit, Preserve, and Clean failures.** Edit and Clean use benchmark human ratings and Preserve uses our human-reviewed matched-control dataset (Appendix B.4). AP, nAP, and AUC denote average precision, prevalence-normalized AP, and ROC AUC. Higher is better, with bold and underline marking the best and second-best results.

	Edit, EditReward-Bench IF			**Preserve**, Matched controls			**Clean**, ImagenHub PQ		
Method	**AP**	**nAP**	**AUC**	**AP**	**nAP**	**AUC**	**AP**	**nAP**	**AUC**
EditScore-4B	0.5409	0.2296	0.6663	0.8110	0.6220	0.7901	0.8617	0.7771	0.9004
EditScore-8B	0.5331	0.2166	0.6648	0.7892	0.5783	0.7587	0.8724	0.7944	0.8981
EditReward	0.6020	0.3322	0.7072	0.8601	0.7203	0.8485	0.8607	0.7755	0.8948
SpatialReward	0.6019	0.3321	**0.7107**	0.9287	0.8574	0.9222	0.8442	0.7488	0.8835
EditHF	0.4421	0.0639	0.5142	0.9168	0.8336	0.9091	0.6620	0.4552	0.8312
HATIE*	0.4975	0.1569	0.6201	0.7670	0.5339	0.8346	0.4963	0.1882	0.6334
AugCLIP	0.4238	0.0333	0.5021	0.6162	0.2325	0.6627	0.5173	0.2220	0.7151
EditCal **(Ours)**	**0.6127**	**0.3501**	0.7001	**0.9741**	**0.9482**	**0.9740**	**0.9386**	**0.9010**	**0.9526**

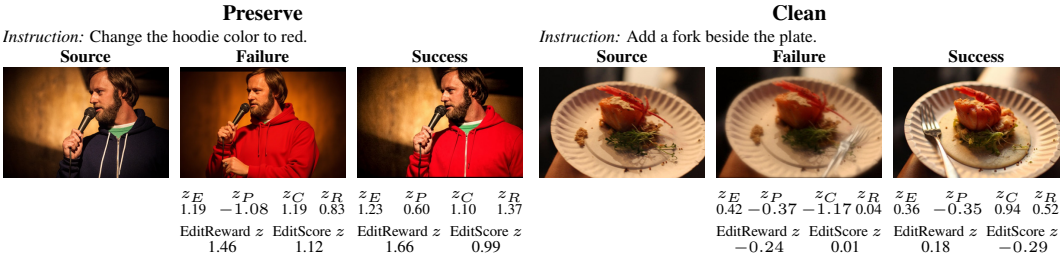


Figure 5: **Criterion-specific failure diagnosis.** Matched failure and success examples for Preserve and Clean. Scores are standardized using the same STACE-22K test reference set across both panels.

5 Conclusions

We present a reward-guided framework for controlling frozen image editors through noise-level trajectories. ATS predicts source- and instruction-conditioned trajectories without per-input optimization or online reward queries and transfers across sampling-step counts, while GeoTS provides geometry-normalized trajectory variation. EditCal supplies a task-aware editing reward by integrating complementary evidence under shared Edit, Preserve, and Clean criteria. Limitations and future directions are discussed in Appendix D. Experiments show consistent gains from ATS, complementary editor-dependent benefits of trajectory and noise variation, and improved human agreement and failure detection, supporting noise-level trajectories as a control interface for image editing.

AI USE STATEMENT

In this work, we used generative AI tools for translation and edit labeling through the GPT API as part of our data processing pipeline. The core research idea, methodology, and experimental protocols were developed by the authors. Additionally, we used generative AI tools for grammar, code implementation, refinement, and debugging, and to provide feedback on result interpretation. We have reviewed all AI-assisted work. GPT-based edit labeling followed author-defined criteria, and all manuscript edits were reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Jinho Chang, Jaemin Kim, and Jong Chul Ye. Training-free reward-guided image editing via trajectory optimal control. In *International Conference on Learning Representations*, volume 2026, pp. 26375–26396, 2026.

Sherry X Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Kuo-Chin Lien, Misha Sra, and Pradeep Sen. Tino-edit: Timestep and noise optimization for robust diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6337–6346, 2024.

Sherry X. Chen, Yi Wei, Luowei Zhou, and Suren Kumar. ADIEE: Automatic dataset creation and scorer for instruction-guided image editing evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18345–18356, October 2025.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54820–54833, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yilie Huang, Wenpin Tang, and Xunyu Zhou. ART for Diffusion Sampling: A reinforcement learning approach to timestep schedule, 2026.

Yoonjeon Kim, Soohyun Ryu, Yeonsung Jung, Hyunkoo Lee, Joowon Kim, June Yong Yang, Jaeryong Hwang, and Eunho Yang. Preserve or modify? context-aware evaluation for balancing preservation and modification in text-guided image editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23474–23483. IEEE, 2025.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730. IEEE, 2025.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Fan Li, Chonghuinan Wang, Lina Lei, Yuping Qiu, Jiaqi Xu, Jiaxiu Jiang, Xinran Qin, Zhikai Chen, Fenglong Song, Zhixin Wang, Renjing Pei, and Wangmeng Zuo. HP-Edit: A human-preference post-training framework for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43113–43123, June 2026.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, Shaodong Wang, Xinhua Cheng, and Li Yuan. Uniworld-V2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback, 2025.

Haonan Lin, Yan Chen, Jiahao Wang, Wenbin An, Mengmeng Wang, Feng Tian, Yong Liu, Guang Dai, Jingdong Wang, and Qianying Wang. Schedule your edit: A simple yet effective diffusion noise schedule for image editing. *Advances in Neural Information Processing Systems*, 37: 115712–115756, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

Yancheng Long, Yankai Yang, Hongyang Wei, Wei Chen, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, Kaiyu Tang, Bin Wen, Fan Yang, Tingting Gao, Han Li, and Shuo Yang. SpatialReward: Bridging the perception gap in online RL for image editing via explicit spatial reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Liangsi Lu, Minzhe Guo, Xuhang Chen, and Yang Shi. Semantic granularity navigation in image editing. In *Forty-third International Conference on Machine Learning*, 2026.

Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, and Zheng Liu. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. In *International Conference on Learning Representations*, volume 2026, pp. 33027–33056, 2026.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. *Advances in Neural Information Processing Systems*, 37:41494–41516, 2024.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 01 1980. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1980.tb01109.x.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

OpenAI. Introducing GPT-5.4 mini and nano, 2026a. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

OpenAI. Introducing GPT-6 Sol and Luna, 2026b. URL https://openai.com/index/introducing-gpt-6-sol-and-luna/.

Rishubh Parihar, Or Patashnik, Daniil Ostashev, Venkatesh Babu Radhakrishnan, Daniel Cohen-Or, and Kuan-Chieh Jackson Wang. Kontinuous kontext: Continuous strength control for instruction-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37929–37939, 2026.

Yusu Qian, Jiasen Lu, Tsu-Jui Fu, Xinze Wang, Chen Chen, Yinfei Yang, Wenze Hu, and Zhe Gan. GIE-Bench: Towards grounded evaluation for text-guided image editing, 2025.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37226–37235, 2026.

Qwen Team. Qwen3.8-27B. Hugging Face model card, August 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Suho Ryu, Kihyun Kim, Eugene Baek, Dongsoo Shin, and Joonseok Lee. Towards scalable human-aligned benchmark for text-guided image editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18292–18301. IEEE, 2025.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align Your Steps: Optimizing sampling schedules in diffusion models. In *Forty-first International Conference on Machine Learning*, 2024.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu Edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8871–8879, June 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. Featured Certification.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

John S. Uebersax and William M. Grove. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, 49(3):823–835, 1993. doi: 10.2307/2532202.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. EditReward: A human-aligned reward model for instruction-guided image editing. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Yuhui Wu, Chenxi Xie, Ruibin Li, Liyi Chen, Qiaosi Yi, and Lei Zhang. CoCoEdit: Content-consistent image editing via region regularized reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026b.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Zitong Xu, Huiyu Duan, Zhongpeng Ji, Xinyun Zhang, Yutao Liu, Xiongkuo Min, Ke Gu, Jian Zhang, Shusong Xu, Jinwei Chen, et al. Edithf-1m: A million-scale rich human preference feedback for image editing. In *European Conference on Computer Vision*, pp. 278–298. Springer, 2026.

Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, and Guo-Jun Qi. Schedule on the fly: Diffusion time prediction for faster and better image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23412–23422. IEEE, 2025.

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. HIVE: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9026–9036, June 2024.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-Context Edit: Enabling instructional image editing with in-context generation in large-scale diffusion transformers. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2504.20690.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023.

A Trajectory Control Details

A.1 Execution and Evaluation Protocol

We use frozen FLUX.1-Kontext-dev (Kontext) (Black Forest Labs et al., 2025) and FlowEdit (Kulikov et al., 2025) with FLUX.1-dev at 1024×1024 in bfloat16. Kontext receives edit instructions with guidance 2.5; FlowEdit receives source and target descriptions with guidance 1.5 and 5.5, respectively. All test comparisons use the same 150 held-out STACE-22K tasks across nine taxonomies, with two noise repetitions per task. Each condition therefore returns 300 outputs, including selected outputs for multi-candidate methods.

Custom actions in the ATS offline bank and outputs of the ATS policy lie in the bounded support $\mathcal{A}_{\mathrm{ATS}}$, defined by $\eta \in [-0.25, 0.25]$ and $-2\eta \le c \le 1.45$. This support applies to ATS bank construction and policy outputs, not to the time-warped GeoTS candidates in Appendix A.3. The logit-sigmoid trajectory is strictly decreasing in its interior, and executable float32 levels are checked for ordering. Kontext and FlowEdit use entry coordinates $\rho = 0$ and $\rho = 1/7$, respectively. For N updates, the selected function is sampled on the normalized active grid

$$q_k = \rho + (1-\rho)\frac{k}{N}, \qquad k = 0, \dots, N. \tag{13}$$

Changing N changes only the discretization of the selected function.

We train one ATS policy at $N = 12$ for each editor–objective pair. ImageReward-v1.0 (Xu et al., 2023) uses the target description. EditScore (Luo et al., 2026) uses Qwen3-VL-4B's official Overall score, scorer seed 42, and batch size one. EditReward (Wu et al., 2026a) uses Qwen2.5-VL-7B's official Overall score. EditCal supplies bank utilities, training rewards, candidate selection, and Overall evaluation for its policies. In Table 12, these policies are reused across sampling lengths, with the training objective also used for candidate selection and evaluation.

In contrast, Tables 2 and 3 evaluate fixed outputs of EditCal-trained ATS with all four rewards and two large vision-language model (LVLM) judges, Qwen3.8-27B (Qwen Team, 2026) and GPT6-Sol (OpenAI, 2026b), each scoring edits from 0 to 4. Multi-candidate outputs in these editing-quality comparisons are selected once by EditCal and rescored without reselection. The large vision-language model judges receive the instruction and two images in source–edited order, without method identities or reward scores. Images use each model's processor without additional client-side resizing. We serve Qwen3.8-27B with vLLM 0.24.0 in bfloat16 and access GPT6-Sol through the OpenAI Chat Completions API with the default image-detail setting. Both judges use temperature 0, seed 42, and the same system message and continuous-score prompt. JSON responses must contain a finite score in $[0, 4]$ and a nonempty reason. The full system message and continuous-score prompt are provided in Appendix E.4.

Native and ATS single-candidate evaluations share base Gaussian assignments, which are also shared by GeoTS candidates, including FlowEdit's paired-field draws at each logical call. Native and ATS Bo3 share two additional fresh draws, and Bo5 adds two more. Each candidate is fully executed before objective-based selection.

For reward $r^m_{i,s}$ of method m on task i and repetition $s \in \{1, 2\}$, let $\mu^m_i = (r^m_{i,1} + r^m_{i,2})/2$ and $w_i = 1/(9n_{h(i)})$, where $n_{h(i)}$ is the task count of its taxonomy. Taxonomy-macro means and paired gains are

$$\mu_m = \sum_i w_i \mu^m_i, \qquad \Delta_m = \sum_i w_i (\mu^m_i - \mu^{\mathrm{ref}}_i). \tag{14}$$

Table 12 uses Native N, Native N Bo3, and Native N Bo5 as references for the N-, $3N$-, and $5N$-update groups, respectively; other tables specify their references. We use 4,000 paired source-lineage bootstrap resamples within each taxonomy, retaining both repetitions and all methods together. The 2.5th and 97.5th percentiles of the recomputed gain give pointwise 95% CIs, without multiplicity adjustment. These intervals concern gains conditional on the fitted policies and observed repetitions, not variability across policy retraining. Scores retain their original units, and bold marks observed means rather than significance.

A.2 ATS Training and Deployment

The offline bank contains 8,192 tasks per editor: 6,577 fit tasks from 5,240 source lineages and 1,615 monitor tasks from 1,311 lineages. Monitor sources are excluded from optimization and normalization. Each task has Native and six custom slots A–F. A/B use prespecified local/global/anchor comparisons, C uses a task-specific Halton point, and D–F are sequential maximin extensions. All actions and stochastic assignments are fixed before observing rewards. Kontext reuses its initial noise draw. FlowEdit shares the first active Gaussian but uses candidate-slot future streams fixed before reward evaluation, unlike the matched continuation draws in GeoTS evaluation.

ATS inputs comprise source latents and pooled text conditioning from the frozen editor, together with source-side masks (Table 7). All objectives use the same training configuration within each editor (Table 6).

Table 6: **ATS training settings.** Settings are shared across reward objectives within each editor; objective-specific normalization uses training data only.

Setting	Value
Optimizer	AdamW; $\beta = (0.9, 0.999)$; zero weight decay.
Batch and gradients	Batch size 16; gradient clipping at 1.0.
Learning-rate schedule	Constant; no warmup or exponential moving average.
Kontext	Learning rate 10^{-4}; checkpoint at update 1,000.
FlowEdit	Learning rate 5×10^{-5}; checkpoint at update 2,000.
Training seed	2026080702.

Table 7: **ATS policy architecture.** Parameter counts are per editor–objective policy, excluding frozen encoders, editors, and evaluators.

Module	**Specification**	**Parameters**
Spatial encoder	Two source-only convolutional streams, native-grid conditioning fusion, eight region queries, and a 256-d output.	1,277,376
Token construction	Context $256 \to 128$, temporal $128 \to 128$, plan $6 \to 128$, null/step/type/position embeddings, and a readout query.	58,240
Context Transformer	Three pre-norm layers, eight heads, width 128, feed-forward width 512, four input tokens, and a final layer norm.	595,072
Context readout	Linear $128 \to 128$, layer norm, GELU.	16,768
Action heads	Four linear heads: $128 \to 1, 4, 8, 8$ for Native, mixture weights, means, and log scales.	2,709
Total	Same architecture for Kontext and FlowEdit.	**1,950,165**

The action head contains an auxiliary Native logit and a four-component Gaussian mixture. A fixed $\tanh$-based transform maps mixture coordinates into the bounded custom-action support $\mathcal{A}_{\mathrm{ATS}}$. Offline supervision uses

$$\Pi_\phi(b \mid s) = \begin{cases} p_{\mathrm{N}}(s), & b = \mathrm{Native}, \\ [1 - p_{\mathrm{N}}(s)]\, \pi_\phi(b \mid s), & b \in \mathcal{A}_{\mathrm{ATS}}. \end{cases} \tag{15}$$

Native is used only for supervision, including Eq. (5) when a retained teacher is Native. The soft teacher retains the two highest-reward bank candidates. With $(1), (2), (3)$ indexing the three highest utilities, its confidence is

$$w_x = \min\left\{1, \frac{u_{(2)} - u_{(3)}}{\nu}\right\}. \tag{16}$$

For each editor–objective pair, ν is the median of nonzero absolute utility differences between slots A–D of the same fit task, pooled over fit tasks and floored at 10^{-3}.

For EditCal-core, $\nu = 0.0172$ for Kontext and 0.0655 for FlowEdit; other objectives use their own fit-derived scales. Zero-confidence tasks contribute no loss.

For EditCal-core, the selected checkpoints are update 1,000 for Kontext and update 2,000 for FlowEdit. Using source-lineage monitor NLLs from two development seeds, we select the earliest valid checkpoint whose mean NLL is no greater than the minimum mean NLL plus its standard error; confirmation results are not used to rank checkpoints. At deployment, ATS ignores the Native branch and maps the latent mean of the highest-weight custom mixture component to one action. The selected trajectory is executed without online reward queries or optimization. The checkpoint and decoded decision remain fixed during length transfer.

A.3 GeoTS Candidate Construction

The strictly increasing time warp in Eq. (6) preserves the order and endpoints of the complete trajectory. An ATS logit-sigmoid reference remains in the same family with $c' = e^{-\eta_w}(c - c_w)$ and $\eta' = \eta + \eta_w$. These transformed coordinates may lie outside $\mathcal{A}_{\mathrm{ATS}}$; leaving the ATS support does not trigger clipping, projection, or rejection of a GeoTS candidate. The identity warp recovers the reference trajectory and, with matched noise, its output.

Let $\boldsymbol{\sigma}(\omega)$ collect the N model-input noise levels, excluding the clean terminal, and let J be the identity-warp Jacobian. With $d_w = 2$ warp coordinates, the regularized local metric is

$$H = \frac{J^\top J}{N}, \qquad \lambda_H = \epsilon_{\mathrm{geo}} \frac{\mathrm{tr}(H)}{d_w}, \qquad H_\epsilon = H + \lambda_H I_{d_w}. \tag{17}$$

Locally, $\|\boldsymbol{\sigma}(\omega) - \boldsymbol{\sigma}(0)\|_2^2/N \approx \omega^\top H \omega$. From ordered seed directions $b_1, \ldots, b_{d_w}$, metric Gram–Schmidt gives

$$v_j = b_j - \sum_{i<j} d_i (d_i^\top H_\epsilon b_j), \qquad d_j = \frac{v_j}{\sqrt{v_j^\top H_\epsilon v_j}}, \qquad d_i^\top H_\epsilon d_j = \delta_{ij}. \tag{18}$$

With $S_0 = \|\boldsymbol{\sigma}(0)\|_2/\sqrt{N}$ and $r = \alpha S_0$, the nested candidate set is

$$\mathcal{C}_{B_m} = (0, +rd_1, -rd_1, \ldots, +rd_m, -rd_m), \qquad B_m = 1 + 2m. \tag{19}$$

Using the sharpness direction gives $B = 3$ (GeoTS Bo3), and adding the shift direction gives $B = 5$ (GeoTS Bo5). All test evaluations use $\epsilon_{\mathrm{geo}} = 10^{-6}$ and $\alpha = 0.10$. Candidates include the reference and are fixed before observing rewards.

The radius α approximates relative RMS displacement in executable-noise space. Without the ridge, $V_0^\top H V_0 = I$ and $\omega = \alpha S_0 V_0 u$ for a unit vector u imply

$$\frac{\|J\omega\|_2}{\sqrt{N} S_0} = \frac{\sqrt{\omega^\top H \omega}}{S_0} = \alpha. \tag{20}$$

For the regularized basis $V^\top H_\epsilon V = I$, the corresponding displacement satisfies

$$\left(\frac{\|J\omega\|_2}{\sqrt{N} S_0}\right)^2 = \alpha^2 \big(1 - \lambda_H \|Vu\|_2^2\big) \le \alpha^2. \tag{21}$$

The ridge is conservative in this local approximation. The $\pm rd_j$ proposals are symmetric in warp coordinates; their nonlinear images in executable-noise space need not have equal and opposite displacements. Curvature and finite precision can change the realized displacement; proposals are not rescaled after construction.

B EditCal Details

B.1 Criteria and Evidence

EditCal uses axes $\mathcal{I} = \{E, P, C\}$ and grades $\mathcal{Z} = \{0, \ldots, K-1\}$ with $K = 5$. Grades represent complete failure, largely failed, partial success with material defects, mostly successful with minor defects, and clear success without material defects; grades 3–4 pass. A fixed taxonomy router $g_A = \mathcal{R}_A(t)$ selects ten Edit, two Preserve, and three Clean routes. Addition and removal share a dataset taxonomy but have distinct Edit routes. Table 8 defines the criteria and frozen evidence.

Table 8: **EditCal criteria and evidence.** All evidence models remain frozen. Edit routes E1–E10 share EditReward-IF and EditScore-4B-PF with route-specific calibration. Preserve assesses protected content; Clean assesses visual integrity conditional on the requested change.

Route	Edit criterion
E1: Attribute	Requested attribute on the intended subject and scope.
E2: Background / scene	Requested scene, background, and specified relations.
E3: Global style	Requested rendering style over the intended scope.
E4: Lighting / conditions	Requested lighting, tone, weather, or time and extent.
E5: Material / object style	Requested material or style on the correct object.
E6: Addition	Requested object, count, and specified location or relation.
E7: Removal	Specified source instances are absent.
E8: Replacement	Requested target replaces the specified source and extent.
E9: Pose / action	Intended subject realizes the requested pose, action, or interaction.
E10: Text / symbol	Requested literal or symbol and specified placement.

Route	Preserve / Clean criterion	Evidence
P1: Localized change	Identity and content outside the intended change remain stable.	DINOv3-B protected-token Q10 (Eq. (22)); separately calibrated mask-free self-relation fallback when primary evidence is unavailable.
P2: Global appearance	Source structure, object inventory, and geometry survive the intended appearance change.	DINOv3-B full-image self-relation (Eq. (23)); intended color, lighting, and style changes are not failures.
C1: Standard integrity	Coherent editing without unintended blur, seams, malformed anatomy, or glyph defects.	EditReward visual quality and EditScore perceptual quality, with a nonpositive EditReward VQ log-uncertainty effect.
C2: Global integrity	Physically and structurally coherent appearance without unintended quality collapse.	The same VQ/PQ and uncertainty evidence; intended darkness or fog is not itself a defect.
C3: Style integrity	Shapes, anatomy, and text remain coherent within the requested style.	Authored SigLIP2 margin between coherent and malformed descriptions. Photorealism and style adoption alone are not the criterion.

Localized Preserve uses DINOv3-B features (Siméoni et al., 2026) from source-aligned protected regions determined by the allowed-change mask. Nonprotected pixels are masked before feature extraction; geometry, not scores, determines retained tokens. For 32 corresponding source and edited tokens in a canonical 768×768 geometry, the primary evidence is

$$s_{\mathrm{Q10}} = Q_{0.10}\left(\left\{\frac{\langle f_r^s, f_r^y\rangle}{\|f_r^s\|\|f_r^y\|}\right\}_{r=1}^{32}\right). \tag{22}$$

For candidate i, let v_i denote the evidence branch used for standalone scoring: primary when available, and a separately calibrated mask-free fallback otherwise. A pair (i, j) uses a shared branch v_{ij}, choosing primary only when available for both candidates and fallback otherwise. Pair branch selection can therefore depend on the comparison partner, while standalone scoring remains independent of it.

Global Preserve compares full-image self-relations at three relative DINOv3 depths. For within-image token cosine similarities $S^s_{\ell,uv}$ and $S^y_{\ell,uv}$,

$$s_{\mathrm{REL}} = \frac{1}{|\mathcal{H}|}\sum_{\ell\in\mathcal{H}}\left[1 - \frac{1}{2\binom{m}{2}}\sum_{u<v}|S^s_{\ell,uv} - S^y_{\ell,uv}|\right], \tag{23}$$

where $m = 14^2$ and $\mathcal{H}$ indexes relative depths $1/3, 2/3, 1$.

Features are direction-aligned and standardized using training Human point examples and pair arms. Centers and scales use the median and scaled interquartile range, with a standard-deviation fall-

back for degenerate scales. Edit and Clean pool scalers within fitting blocks; Preserve scales primary, fallback, and global branches separately. GPT5.4-mini (OpenAI, 2026a) API annotations do not determine scalers. With standardized instruction-following, prompt-following, visual-quality, perceptual-quality, uncertainty, Preserve, and style-margin evidence, the responses are

$$\begin{aligned} \ell_{E,g} &= b_{E,g} + \lambda_E e_{\mathrm{IF}} + (1-\lambda_E) e_{\mathrm{PF}}, \\ \ell_{P,g,v} &= b_{P,g,v} + a_{P,g,v}\,\mathrm{clip}(e_{P,g,v}, -8, 8), \\ \ell_{C,g} &= \lambda_{C,g} e_{\mathrm{VQ}} + (1-\lambda_{C,g}) e_{\mathrm{PQ}} - a_{C,g} e_{\mathrm{U}}, \quad g \in \{C1, C2\}, \\ \ell_{C,C3} &= e_{\mathrm{style}}. \end{aligned} \tag{24}$$

Edit shares one fusion weight across routes but uses route-specific offsets and ordinal thresholds. Preserve uses branch-specific slopes and offsets. C1/C2 have separate fusion and uncertainty coefficients; C3 uses the authored SigLIP2 (Tschannen et al., 2025) coherent–malformed margin directly. Ordered-probit calibration maps these responses to grade distributions as in Eq. (9).

B.2 Annotation Alignment and Calibration

Stages are fitted sequentially, with previously learned models and targets held fixed in subsequent stages. Human annotations anchor the shared grade semantics. Following Uebersax & Grove (1993), reviewer responses satisfy $V_{rA} \mid Z_A \sim \mathcal{N}(Z_A - b_r, \sigma_A^2 e^{2d_r})$, with fixed half-grade boundaries and reviewer effects constrained by $\sum_r b_r = 0$ and $\sum_r d_r = 0$. The observer minimizes the negative log-likelihood of genuine Human responses after marginalizing latent grades; adjudication affects only the target prior, giving $h_n(\mathbf{z}) \propto p_0(\mathbf{z}) H_\theta(\mathbf{a}_n^{\mathrm{H}} \mid \mathbf{z})$. Priors count complete same-reviewer E/P/C triples, with total pseudocount one spread uniformly over the 125 states. With Human parameters and posteriors fixed, the API kernel learns boundaries and noise by minimizing the negative log-likelihood of observed API responses marginalized under the Human posteriors. This fit uses only shared Human/API training records. Human targets take precedence on overlap; API-only records use aligned posteriors, and missing axes are marginalized. Five original-source outer folds exclude held-out sources and their API annotations from alignment, scaling, and axis fitting. Deployment refits the selected specification on all training sources. Public benchmarks, failure panels, and every example from the Preserve target-prompt intervention are excluded from scorer fitting. For a pair (i, j), define $\Delta_{ij}^A = \ell_{A,i}^{[ij]} - \ell_{A,j}^{[ij]}$, where the superscript denotes responses evaluated for that comparison. For localized Preserve, these are $\ell_{P,g,v_{ij}}(x_i, y_i)$ and $\ell_{P,g,v_{ij}}(x_j, y_j)$, using the shared branch defined in Appendix B.1; other routes use their ordinary responses. A probit model with nonnegative tie width $\delta_{A,g}$ gives

$$\begin{aligned} P(i \succ_A j) &= \Phi(\Delta_{ij}^A - \delta_{A,g}), \\ P(j \succ_A i) &= \Phi(-\Delta_{ij}^A - \delta_{A,g}), \\ P(i \sim_A j) &= 1 - P(i \succ_A j) - P(j \succ_A i). \end{aligned} \tag{25}$$

The pair preference distribution $P_{ij,P}$ therefore uses both candidates' evidence availability. In contrast, a standalone Preserve grade distribution $\pi_{i,P}$ uses v_i and is defined independently of the comparison partner. The pair rule does not redefine standalone grade distributions or their associated quality readouts, and a pair preference need not be obtained from the two standalone scores.

Pointwise losses d_A are cross-entropy for Edit, Preserve, and ranked probability score for Clean. For aligned point targets $\mathcal{D}_A$, Human pairs $\mathcal{P}_A$, and separately normalized weights,

$$\mathcal{L}_{\mathrm{ord},A} = \alpha_A \sum_{n \in \mathcal{D}_A} w_{n,A} d_A(t_{n,A}, \pi_{n,A}) + (1-\alpha_A) \sum_{(i,j) \in \mathcal{P}_A} w_{ij,A}\,\mathrm{CE}(r_{ij,A}, P_{ij,A}). \tag{26}$$

Here $P_{ij,A}$ is the three-way pair distribution, and $\alpha_E = 1$, $\alpha_P = 0.5$, $\alpha_C = 0.5$. Evidence models, targets, and feature transforms remain frozen during axis fitting.

All calibration fits use float64. Threshold ordering and parameter constraints are enforced by construction. Table 9 gives optimizer settings; all fitted parameters and feature transforms are frozen for evaluation.

B.3 Joint Modeling and Readouts

For each calibrated distribution π_A, define $c_{A,k} = \sum_{z=k}^{K-1} \pi_A(z)$, $c_{A,0} = 1$, $c_{A,K} = 0$, and $I_{A,z} = (c_{A,z+1}, c_{A,z}]$, so $|I_{A,z}| = \pi_A(z)$. Divide each copula coordinate into L bins $D_j = ((j-1)/L, j/L]$.

Table 9: **Calibration optimizer settings.** All stages use float64. Axis initializations and parameter constraints are specified in Appendix B.2.

Stage	Configuration
Human observer	L-BFGS-B; $\sigma_A = 0.65$, $b_r = d_r = 0$ initially; 300 iterations; $\mathrm{ftol} = 10^{-10}$, $\mathrm{gtol} = 10^{-6}$.
API alignment	SLSQP initialized from Human boundaries/noise; 300 iterations; $\mathrm{ftol} = 10^{-9}$.
Edit / Clean axes	L-BFGS-B; 2,500 iterations and 150,000 function evaluations; $\mathrm{ftol} = 10^{-10}$, $\mathrm{gtol} = 10^{-7}$.
Preserve axis	L-BFGS-B; 1,000 iterations and 50,000 function evaluations; the same tolerances.

Table 10: **Copula selection on internal validation.** Genuine-response NLL and good-event Brier on 311 sources and 361 candidates in source-held-out internal validation. Calibrated axes are fixed and missing axes are marginalized. External benchmarks are not used for selection. Lower is better.

Copula	NLL	Good-event Brier
Independent	2.2227	0.1502
Checkerboard-2	2.1975	0.1475
Checkerboard-3	2.1882	0.1458

Nonnegative cell masses with uniform marginals give

$$q_\psi(\mathbf{z}) = \sum_{\mathbf{j}} w_{\mathbf{j}} \prod_{A \in \mathcal{I}} L \left| I_{A,z_A} \cap D_{j_A} \right| \qquad w_{\mathbf{j}} \geq 0, \qquad \sum_{\mathbf{j}: j_A = r} w_{\mathbf{j}} = \frac{1}{L}. \tag{27}$$

Since grade intervals partition each coordinate, the calibrated marginals are preserved:

$$\sum_{\mathbf{z}: z_A = z} q_\psi(\mathbf{z}) = \sum_{r=1}^{L} |I_{A,z} \cap D_r| = |I_{A,z}| = \pi_A(z). \tag{28}$$

Thus dependence fitting leaves criterion-specific pass and failure probabilities unchanged.

With source-crossfitted axis predictions fixed, we fit only copula masses against aligned joint-grade posteriors $t_n(\mathbf{z})$. For $D = |\mathcal{I}| = 3$ and nonnegative weights v_n summing to one,

$$\begin{aligned} \mathcal{L}_{\mathrm{joint}}(\psi) = & -\frac{1}{\log(K^D)} \sum_{n \in \mathcal{D}_{\mathrm{J}}} v_n \sum_{\mathbf{z} \in \mathcal{Z}^D} t_n(\mathbf{z}) \log q_\psi(\mathbf{z} \mid x_n, y_n) \\ & + \lambda_{\mathrm{joint}} L^D \sum_{\mathbf{j}} \left(w_{\mathbf{j}} - L^{-D} \right)^2 . \end{aligned} \tag{29}$$

The data term is cross-entropy divided by $\log 125$ over the 125 joint-grade states, with axis probabilities fixed. We use $\lambda_{\mathrm{joint}} = 0.01$; the regularizer favors the independent uniform copula. Fitting weights balance Human and API-only populations, then sources and their candidates; joint fitting uses no pair loss.

We compare independence and two- or three-bin checkerboard copulas with the calibrated axes fixed. Source lineages used for internal validation are excluded from copula mass fitting. Capacity is selected using internal-validation response NLL and good-event Brier (Table 10), not external benchmark scores. Checkerboard-3 ($L = 3$) has the lowest values on both metrics and is fixed for subsequent evaluation.

The Overall reward is the expected geometric quality of joint grades,

$$R_{\mathrm{EC}} = \sum_{\mathbf{z} \in \mathcal{Z}^3} q_\psi(\mathbf{z}) \left[\left(\frac{z_E}{4} \right) \left(\frac{z_P}{4} \right) \left(\frac{z_C}{4} \right) \right]^{1/3}. \tag{30}$$

The geometric mean acts on grades within each outcome; copula probabilities weight those values. Axis diagnostics instead use $p_A = \sum_{k=3}^{4} \pi_A(k)$ and failure score $1 - p_A$. For a nonempty subset

Table 11: **EditCal training and scoring.** Online scoring uses frozen parameters and no candidate-pool statistics. Grades are $\{0, \ldots, 4\}$; grades 3–4 pass.

Offline: training sources
1 Fit the Human observer with $\mathcal{L}_{\text{obs}}$ and align GPT responses with $\mathcal{L}_{\text{API}}$. Use Human targets on overlap and aligned API targets elsewhere.
2 Fit feature transforms, route responses $\ell_{A,g}$, and ordered thresholds $\kappa_{A,g}$ using $\mathcal{L}_{\text{ord},A}$ with frozen evidence models.
3 Fix source-crossfitted axis predictions and fit copula masses with $\mathcal{L}_{\text{joint}}$. Freeze the scoring bundle.
Online: source context x and edited candidate y
4 Route each axis to $g_A = \mathcal{R}_A(t)$, extract the prescribed evidence, and apply its frozen transforms and response.
5 Compute $\pi_A(k) = \Phi(\kappa_{A,g_A,k+1} - \ell_{A,g_A}) - \Phi(\kappa_{A,g_A,k} - \ell_{A,g_A})$ for each grade.
6 Return pass diagnostics $p_A = \sum_{k=3}^{4} \pi_A(k)$ and failure scores $1 - p_A$.
7 Form $q_\psi(\mathbf{z})$ with the frozen copula, preserving every π_A.
8 Return $R_{\text{EC}} = \sum_{\mathbf{z}} q_\psi(\mathbf{z})[(z_E/4)(z_P/4)(z_C/4)]^{1/3}$: geometric quality per joint grade outcome, averaged by its probability.

$\mathcal{S} \subseteq \mathcal{I}$, the same joint distribution gives

$$R_{\mathcal{S}} = \sum_{\mathbf{z}_{\mathcal{S}}} q_{\mathcal{S}}(\mathbf{z}_{\mathcal{S}}) \left(\prod_{A \in \mathcal{S}} \frac{z_A}{K-1} \right)^{1/|\mathcal{S}|}, \tag{31}$$

where $q_{\mathcal{S}}$ is its marginal. ImagenHub Semantic Consistency uses R_{EP} with $\mathcal{S} = \{E, P\}$. These quality readouts require neither another fit nor candidate-pool normalization. Table 11 summarizes training and scoring.

B.4 Evaluation Protocols

We follow each benchmark's official evaluation code and protocol using its native human annotations. EditCal uses p_E, p_P, p_C for axis endpoints, R_{EP} for ImagenHub Semantic Consistency, and R_{EC} for Overall. For EditScore-4B/8B and SpatialReward-8B, we use their prompt-following, consistency, perceptual-quality, and overall scores. EditReward uses its instruction-following, visual-quality, and overall rewards, while EditHF uses its alignment, preservation, and visual-quality regressors. AugCLIP uses its native scalar score. HATIE* follows HATIE but replaces its category-limited mask component with open-vocabulary SAM3 (Carion et al., 2026) segmentation, enabling evaluation across all images.

The Preserve panel targets unrequested compositional drift. Kontext receives a target-image description instead of the edit instruction while retaining source conditioning. Of 128 intervention candidates, blinded Human review identifies 93 failures with Preserve grades 0–1. For each, three correct-instruction controls use the assigned noise sample and a fixed schedule order: Native, then $(c, \eta) = (0, 0.25), (-0.30, 0.25)$. Controls use 1024×1024, bfloat16, guidance 2.5, and 12 updates. The first grade-4 control is retained, yielding 65 matched pairs; 28 cases have no qualifying control. Evaluator scores never select matches. The entire intervention and its controls are excluded from EditCal training. Additional examples are shown in Figure 9.

The Edit panel has 442 failures and 652 successes over 426 sources, using EditReward-Bench human IF scores of ≤ 2 and ≥ 3, respectively. Clean has 219 failures and 358 successes over 171 sources, using three unanimous ImagenHub PQ ratings of 0 or 1. Edit scores strictly between 2 and 3 and mixed Clean judgments are excluded; these labels retain their benchmark-native meanings. Failure is positive, and normalized AP is $(\text{AP} - \pi_f)/(1 - \pi_f)$ for failure prevalence π_f. Internal operating thresholds are selected on calibration-training folds and fixed on held-out folds.

Internal calibration evaluation predicts genuine Human responses through the frozen observer, weighting sources equally, then candidates and observed reviewers; axis macros weight routes

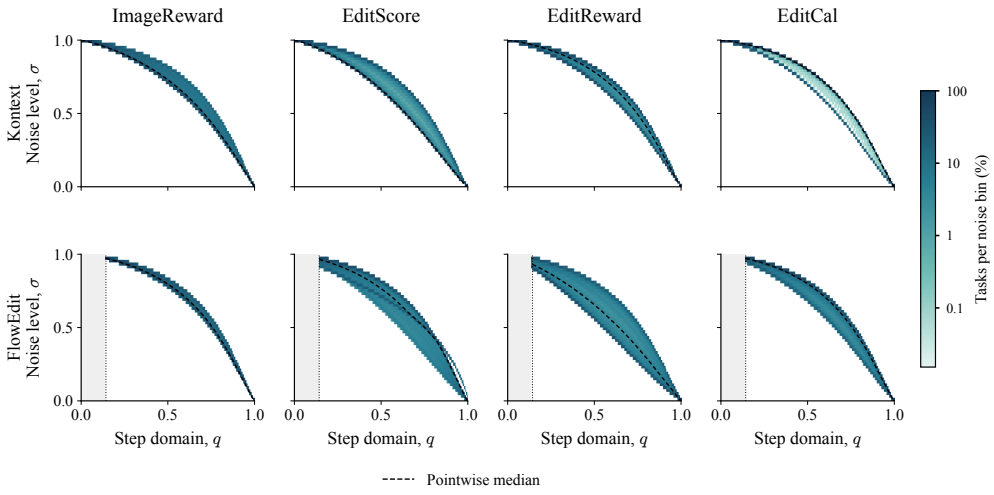


Figure 6: **Deployed ATS trajectories at $N = 12$.** Color indicates the task fraction per noise-level bin ($\Delta\sigma = 0.02$), with one trajectory per task and a shared logarithmic scale. Dashed curves are pointwise medians; FlowEdit skips the gray prefix.

equally. For reviewer rating $y \in \{0, \ldots, K_{\text{obs}} - 1\}$ and predicted tails $a_k = \Pr(Y \geq k)$,

$$\text{RPS}(a, y) = \frac{1}{K_{\text{obs}} - 1} \sum_{k=1}^{K_{\text{obs}}-1} \big(a_k - \mathbf{1}[y \geq k]\big)^2, \tag{32}$$

Here $\mathbf{1}[\cdot]$ is the indicator and $K_{\text{obs}} = K$, but reviewer-response and latent-grade distributions remain distinct. For partial joint observations, good-event Brier concerns passing all observed reviewer axes, with missing axes marginalized, rather than latent P_{all}.

Figure 5 standardizes each metric as $z_m = (s_m - \mu_m)/\sigma_m$ using its mean and population standard deviation over the same 28,584 distinct, completely scored STACE-22K test outputs pooled across both editors and evaluated step counts. Both panels, including the Preserve matched controls, use this reference. Negative z-scores indicate below-reference-average scores, not failure probabilities.

C Additional Results and Ablations

C.1 Trajectory Control and Zero-Shot Length Transfer

Table 12 combines objective-guided control at $N = 12$ with zero-shot transfer to $N = 6, 8, 16$. Each editor–objective pair uses one policy trained at $N = 12$, with its checkpoint and decoded task decisions fixed across lengths. The listed objective guides training, selection, and evaluation, unlike the fixed EditCal-trained outputs in the main editing-quality comparisons.

At each N, we compare single-trajectory execution, longer Native execution, and three- or five-candidate selection within N-, $3N$-, and $5N$-update groups. Length transfer changes only the sampling grid in Eq. (13). GeoTS recomputes its executable-noise geometry around the transferred ATS reference using the same α, rather than reusing identical warped candidates. Noise assignments follow Appendix A.1.

ATS yields higher observed mean rewards than the corresponding Native single-rollout and noise-sampling baselines at every tested length. At $N = 12$, ATS-based noise sampling gives the highest means on Kontext and GeoTS on FlowEdit.

Figure 6 summarizes the deployed trajectories on the 150 test tasks for each editor–objective pair.

Table 12: **Objective-guided trajectory control and zero-shot length transfer.** Each editor–objective pair uses one ATS policy trained at $N = 12$, with decoded decisions fixed at $N = 6, 8, 16$. The listed objective guides training, selection, and evaluation. Bo3/Bo5 denote best-of-3/5 selection: ATS varies noise with the trajectory fixed, whereas GeoTS varies trajectories with noise fixed. Scores are taxonomy-macro means. Brackets show pointwise 95% paired bootstrap CIs for gains over Native N, Native N Bo3, or Native N Bo5 in the respective update group. Bold marks the highest mean per row and update group.

		N updates (1×)		$3N$ updates (3×)				$5N$ updates (5×)			
N	Objective	**Native N**	**ATS N**	**Native $3N$**	**Native N Bo3**	**ATS N Bo3**	**GeoTS Bo3**	**Native $5N$**	**Native N Bo5**	**ATS N Bo5**	**GeoTS Bo5**
Kontext											
6	ImageReward	−0.100 *ref.*	**−0.067** [.002,.065]	0.073 [−.166,−.023]	0.167 *ref.*	**0.183** [−.006,.039]	0.119 [−.110,.007]	0.088 [−.276,−.133]	0.289 *ref.*	**0.300** [−.011,.031]	0.196 [−.157,−.038]
	EditScore	6.486 *ref.*	**6.503** [−.188,.229]	6.808 [−.994,−.317]	7.444 *ref.*	**7.572** [−.009,.264]	7.410 [−.227,.158]	6.972 [−1.067,−.374]	7.675 *ref.*	**7.821** [.029,.266]	7.603 [−.249,.099]
	EditReward	0.850 *ref.*	**0.950** [.059,.145]	1.117 [−.152,.009]	1.186 *ref.*	**1.288** [.065,.143]	1.218 [−.037,.097]	1.150 [−.220,−.053]	1.283 *ref.*	**1.375** [.057,.126]	1.321 [−.021,.092]
	EditCal	0.646 *ref.*	**0.652** [−.006,.016]	0.668 [−.047,−.016]	0.699 *ref.*	**0.707** [−.001,.016]	0.698 [−.011,.009]	0.676 [−.054,−.022]	0.713 *ref.*	**0.723** [.004,.018]	0.711 [−.012,.008]
8	ImageReward	−0.027 *ref.*	**0.013** [.011,.071]	0.096 [−.217,−.118]	0.263 *ref.*	**0.270** [−.015,.028]	0.183 [−.124,−.037]	0.094 [−.328,−.202]	0.355 *ref.*	**0.363** [−.013,.029]	0.233 [−.174,−.075]
	EditScore	6.426 *ref.*	**6.637** [−.003,.424]	6.965 [−.833,−.251]	7.492 *ref.*	**7.612** [−.003,.246]	7.473 [−.178,.137]	6.944 [−1.211,−.602]	7.831 *ref.*	**7.938** [−.004,.220]	7.726 [−.252,.034]
	EditReward	0.981 *ref.*	**1.033** [.010,.094]	1.157 [−.225,−.084]	1.308 *ref.*	**1.384** [.044,.107]	1.320 [−.042,.064]	1.158 [−.322,−.170]	1.403 *ref.*	**1.467** [.038,.091]	1.380 [−.072,.027]
	EditCal	0.653 *ref.*	**0.662** [−.002,.019]	0.679 [−.042,−.015]	0.707 *ref.*	**0.716** [.003,.016]	0.701 [−.017,.005]	0.679 [−.057,−.030]	0.722 *ref.*	**0.731** [.003,.015]	0.714 [−.016,.000]
12	ImageReward	0.016 *ref.*	**0.050** [−.002,.075]	0.101 [−.247,−.146]	0.296 *ref.*	**0.329** [.009,.060]	0.214 [−.133,−.037]	0.088 [−.377,−.236]	0.389 *ref.*	**0.404** [−.009,.039]	0.259 [−.176,−.088]
	EditScore	6.645 *ref.*	**6.809** [−.001,.334]	6.926 [−.973,−.452]	7.629 *ref.*	**7.738** [−.040,.261]	7.563 [−.304,.167]	7.018 [−1.272,−.697]	7.987 *ref.*	**8.096** [−.063,.306]	7.859 [−.281,.022]
	EditReward	1.028 *ref.*	**1.110** [.044,.122]	1.175 [−.271,−.144]	1.381 *ref.*	**1.437** [.026,.086]	1.330 [−.101,−.003]	1.176 [−.363,−.220]	1.466 *ref.*	**1.531** [.035,.095]	1.415 [−.095,−.010]
	EditCal	0.659 *ref.*	**0.679** [.012,.030]	0.677 [−.054,−.028]	0.718 *ref.*	**0.718** [−.009,.009]	0.700 [−.031,−.006]	0.680 [−.067,−.039]	0.733 *ref.*	**0.739** [−.002,.015]	0.716 [−.026,−.008]
16	ImageReward	0.066 *ref.*	**0.091** [−.002,.055]	0.099 [−.286,−.164]	0.321 *ref.*	**0.343** [−.003,.050]	0.225 [−.145,−.050]	0.094 [−.385,−.243]	0.402 *ref.*	**0.423** [.001,.042]	0.265 [−.185,−.095]
	EditScore	6.776 *ref.*	**6.991** [.074,.350]	7.022 [−1.076,−.483]	7.781 *ref.*	**7.887** [−.015,.229]	7.592 [−.438,.019]	7.002 [−1.345,−.789]	8.049 *ref.*	**8.179** [.026,.247]	7.831 [−.391,−.072]
	EditReward	1.103 *ref.*	**1.181** [.029,.126]	1.185 [−.320,−.192]	1.439 *ref.*	**1.472** [.005,.061]	1.318 [−.178,−.069]	1.176 [−.430,−.272]	1.523 *ref.*	**1.549** [.003,.050]	1.410 [−.166,−.066]
	EditCal	0.668 *ref.*	**0.678** [.002,.019]	0.680 [−.057,−.031]	0.723 *ref.*	**0.725** [−.005,.006]	0.698 [−.037,−.015]	0.681 [−.071,−.044]	0.738 *ref.*	**0.742** [−.002,.009]	0.715 [−.033,−.013]
FlowEdit											
6	ImageReward	−0.018 *ref.*	**0.333** [.249,.463]	0.150 [−.371,−.200]	0.436 *ref.*	**0.805** [.288,.457]	0.691 [.159,.352]	0.146 [−.506,−.369]	0.583 *ref.*	**0.940** [.284,.435]	0.854 [.177,.366]
	EditScore	4.535 *ref.*	**5.107** [.103,1.107]	5.090 [−1.485,−.692]	6.169 *ref.*	**6.537** [.075,.679]	6.482 [.017,.605]	4.993 [−2.213,−1.347]	6.767 *ref.*	**7.149** [.107,.686]	6.985 [−.064,.505]
	EditReward	−0.274 *ref.*	**−0.212** [−.023,.156]	−0.095 [−.349,−.195]	0.178 *ref.*	**0.186** [−.062,.079]	0.099 [−.199,.032]	−0.086 [−.469,−.284]	0.290 *ref.*	**0.339** [−.022,.120]	0.217 [−.188,.036]
	EditCal	0.475 *ref.*	**0.500** [.008,.043]	0.514 [−.075,−.040]	0.572 *ref.*	**0.586** [.000,.028]	0.576 [−.012,.021]	0.515 [−.105,−.066]	0.601 *ref.*	**0.616** [.001,.030]	0.603 [−.012,.018]
8	ImageReward	0.098 *ref.*	**0.515** [.319,.520]	0.186 [−.408,−.271]	0.526 *ref.*	**0.950** [.345,.507]	0.862 [.250,.423]	0.118 [−.612,−.463]	0.655 *ref.*	**1.077** [.348,.502]	1.044 [.302,.482]
	EditScore	4.810 *ref.*	**5.462** [.135,1.222]	5.164 [−1.800,−1.038]	6.569 *ref.*	6.759 [−.140,.547]	**6.952** [.090,.712]	4.844 [−2.674,−1.814]	7.072 *ref.*	**7.441** [.063,.692]	7.418 [.069,.648]
	EditReward	−0.201 *ref.*	**−0.115** [.013,.160]	−0.064 [−.341,−.188]	0.200 *ref.*	**0.237** [−.037,.115]	0.195 [−.094,.087]	−0.106 [−.516,−.336]	0.320 *ref.*	**0.422** [.027,.177]	0.341 [−.055,.099]
	EditCal	0.495 *ref.*	**0.523** [.010,.046]	0.519 [−.081,−.047]	0.583 *ref.*	0.601 [.003,.034]	**0.604** [.006,.037]	0.508 [−.119,−.079]	0.607 *ref.*	**0.634** [.013,.043]	0.634 [.012,.043]
12	ImageReward	0.151 *ref.*	**0.624** [.370,.577]	0.136 [−.469,−.330]	0.535 *ref.*	0.983 [.362,.536]	**0.995** [.370,.554]	0.082 [−.661,−.500]	0.663 *ref.*	1.093 [.348,.517]	**1.186** [.430,.622]
	EditScore	4.970 *ref.*	**5.407** [.033,.875]	5.131 [−1.830,−1.073]	6.568 *ref.*	6.955 [.009,.786]	**7.222** [.313,1.018]	4.742 [−2.866,−2.009]	7.163 *ref.*	7.486 [−.027,.696]	**7.615** [.114,.796]
	EditReward	−0.148 *ref.*	**−0.049** [.021,.179]	−0.093 [−.418,−.254]	0.240 *ref.*	0.306 [−.020,.151]	**0.337** [.019,.178]	−0.124 [−.617,−.410]	0.388 *ref.*	0.426 [−.037,.120]	**0.454** [−.011,.150]
	EditCal	0.505 *ref.*	**0.542** [.019,.057]	0.517 [−.089,−.051]	0.586 *ref.*	0.612 [.012,.042]	**0.625** [.023,.056]	0.503 [−.136,−.096]	0.618 *ref.*	0.644 [.012,.039]	**0.653** [.021,.050]
16	ImageReward	0.166 *ref.*	**0.645** [.376,.582]	0.136 [−.470,−.332]	0.537 *ref.*	0.999 [.379,.550]	**1.036** [.406,.597]	0.030 [−.709,−.545]	0.655 *ref.*	1.094 [.361,.522]	**1.222** [.468,.669]
	EditScore	4.828 *ref.*	**5.530** [.247,1.185]	4.986 [−1.986,−1.209]	6.564 *ref.*	6.968 [.054,.769]	**7.210** [.295,.997]	4.551 [−2.913,−2.033]	7.013 *ref.*	7.524 [.154,.888]	**7.658** [.265,1.005]
	EditReward	−0.135 *ref.*	**0.002** [.057,.219]	−0.104 [−.442,−.272]	0.253 *ref.*	0.347 [.021,.167]	**0.396** [.060,.224]	−0.098 [−.571,−.383]	0.378 *ref.*	0.459 [.003,.164]	**0.511** [.059,.210]
	EditCal	0.502 *ref.*	**0.541** [.019,.061]	0.513 [−.096,−.059]	0.590 *ref.*	0.619 [.014,.045]	**0.630** [.024,.057]	0.498 [−.136,−.094]	0.613 *ref.*	0.640 [.013,.043]	**0.656** [.028,.060]

We evaluate our FLUX/FlowEdit adaptation of NaviEdit (Lu et al., 2026) with and without spatial gating, using intermediate features from existing forward passes without reward queries or editor training. NaviEdit and FlowEdit both perform two sample-level model evaluations per update; comparisons match this count rather than runtime. Multi-candidate methods use three or five 12-step trajectories, with reward queries reported separately. ATS achieves higher mean scores than both adapted variants at 12 updates across all four objectives (Table 13).

Table 13: **FlowEdit comparison with adapted NaviEdit.** NaviEdit uses our FLUX/FlowEdit adaptation with spatial gating off or on. Each ATS/GeoTS reward column uses its corresponding training objective. Scores are taxonomy-macro means; bold marks the highest at each update count.

Method	Reward queries	ImageReward ↑	EditScore ↑	EditReward ↑	EditCal ↑
12 updates per task					
Native	0	0.1509	4.9698	-0.1483	0.5049
ATS	0	**0.6241**	**5.4074**	**-0.0489**	**0.5420**
NaviEdit, gate off	0	0.1630	4.8557	-0.1783	0.4968
NaviEdit, gate on	0	-0.3806	2.7807	-0.1923	0.4327
36 updates per task					
Native, 36 updates	0	0.1363	5.1305	-0.0932	0.5167
Native Bo3	3	0.5352	6.5677	0.2399	0.5859
ATS Bo3	3	0.9829	6.9553	0.3060	0.6124
GeoTS Bo3	3	**0.9950**	**7.2219**	**0.3370**	**0.6249**
NaviEdit, gate off	0	0.1122	5.0545	-0.1163	0.5108
NaviEdit, gate on	0	-0.3853	2.7421	-0.1587	0.4363
60 updates per task					
Native, 60 updates	0	0.0817	4.7420	-0.1239	0.5034
Native Bo5	5	0.6629	7.1627	0.3878	0.6185
ATS Bo5	5	1.0927	7.4862	0.4259	0.6436
GeoTS Bo5	5	**1.1863**	**7.6151**	**0.4540**	**0.6532**
NaviEdit, gate off	0	0.0564	4.7972	-0.1400	0.5016
NaviEdit, gate on	0	-0.4178	2.4655	-0.1205	0.4307

C.2 ATS AND GEOTS ABLATIONS

ATS conditioning ablations use EditCal at $N = 12$, with the same architecture, offline bank, loss, and training procedure. Global replaces all task-dependent inputs with constants and deploys one trajectory per editor. It improves over Native but remains below Full ATS; removing either image or instruction conditioning also lowers mean reward on both editors (Table 14).

Table 14: **ATS input ablations.** EditCal rewards at $N = 12$. Global uses constant inputs with the same architecture and training recipe. Δ is method minus Full ATS; brackets are paired 95% CIs from 4,000 source-bootstrap resamples. Bold marks the highest mean.

	Kontext		**FlowEdit**	
Method	**EditCal**	Δ [95% CI]	**EditCal**	Δ [95% CI]
Full ATS	$\mathbf{0.6791}$	*ref.*	$\mathbf{0.5420}$	*ref.*
− Prompt	0.6709	-0.0082 $[-0.0126, -0.0043]$	0.5260	-0.0161 $[-0.0299, -0.0020]$
− Image	0.6690	-0.0101 $[-0.0152, -0.0056]$	0.5323	-0.0097 $[-0.0230, +0.0027]$
Native	0.6588	-0.0203 $[-0.0299, -0.0115]$	0.5049	-0.0372 $[-0.0569, -0.0189]$
Global	0.6712	-0.0079 $[-0.0123, -0.0039]$	0.5274	-0.0146 $[-0.0288, -0.0010]$

The radius sweep uses $\alpha \in \{0.05, 0.10, 0.20, 0.30\}$ on 72 training tasks per editor–objective pair, spanning nine taxonomies with one noise repetition. It covers ImageReward, EditScore, and Ed-

Table 15: **GeoTS radius and candidate-count sensitivity.** Taxonomy-macro rewards on training tasks, with the objective-specific ATS reference shown. Bold α denotes the common value fixed before test evaluation.

	Kontext		FlowEdit	
α	**Bo3**	**Bo5**	**Bo3**	**Bo5**
ImageReward	ATS keep: Kontext 0.5224, FlowEdit 0.9092			
0.05	0.6183	0.6625	1.2153	1.3237
0.10	0.6222	0.6576	1.2255	1.3397
0.20	0.6176	0.6560	1.2926	1.3709
0.30	0.6085	0.6425	1.2483	1.3249
EditScore	ATS keep: Kontext 7.3222, FlowEdit 4.7405			
0.05	7.8146	8.0449	6.5158	7.2315
0.10	7.9062	8.2127	6.5852	7.5440
0.20	7.9654	8.3344	6.9255	7.3602
0.30	8.2741	8.4643	6.6190	7.0584
EditReward	ATS keep: Kontext 1.2836, FlowEdit 0.0588			
0.05	1.4530	1.4807	0.2850	0.4220
0.10	1.4310	1.5023	0.4268	0.5254
0.20	1.4403	1.4629	0.4123	0.5506
0.30	1.4406	1.4730	0.2898	0.3731

itReward, with fixed ATS references, matched Gaussian assignments, candidate order, and ridge. Table 15 reports taxonomy-macro means; its training references differ from test-panel means.

Mean reward-wise ranks of training Bo5 gains favor $\alpha = 0.10$ for Kontext and 0.20 for FlowEdit. For a common-radius test design, we fixed $\alpha = 0.10$ for both editors before test evaluation, without using test scores. Bo5 contains Bo3, so nondecreasing selection reward follows from set inclusion. The sweep measures radius and candidate-count sensitivity, not the isolated effect of normalization.

C.3 EditCal Ablations and Calibration Diagnostics

Removing alignment uses raw Human/GPT grades directly for ordinal calibration. Removing calibration omits ordinal calibration and is equivalent to Geometric Fusion. The $-$Joint variant keeps axis calibration fixed but replaces the joint distribution and its quality readout with a geometric mean of normalized expected grades:

$$m_A = \sum_{k=0}^{4} \frac{k}{4} \pi_A(k), \qquad R_{-\text{Joint}} = (m_E m_P m_C)^{1/3}, \qquad R_{-\text{Joint},EP} = (m_E m_P)^{1/2}. \tag{33}$$

The $-$Joint ablation removes the joint-scoring stage as a whole. To isolate the contribution of learned cross-axis dependence, the $-$Copula variant instead uses an independent joint:

$$q_{\text{ind}}(\mathbf{z}) = \prod_{A \in \{E,P,C\}} \pi_A(z_A), \qquad R_{-\text{Copula}} = \sum_{\mathbf{z}} q_{\text{ind}}(\mathbf{z}) \left(\frac{z_E z_P z_C}{4^3} \right)^{1/3}. \tag{34}$$

This variant fixes the calibrated marginals to those of Full without refitting and retains its expected-geometric quality readout. The same construction applies to the EP readout. Fusion baselines use the same underlying physical evidence and training-only min–max normalization. Mean Fusion averages available evidence within each axis and then across axes; Geometric Fusion changes the final axis average to a geometric mean. Learned Weighted Fusion fits nonnegative, sum-one evidence weights per axis, shared across routes, with point-target MSE and an arithmetic axis average. It uses no observer, ordinal thresholds, intercept, or pair loss. Full EditCal achieves the highest Overall agreement on all three benchmarks (Table 16).

Table 17 reports genuine-response RPS and NLL across all 15 routes under the protocol in Appendix B.4.

Table 16: **EditCal ablation results.** Overall agreement under component removal and evidence fusion. The no-calibration row also reports Geometric Fusion. Higher is better; bold marks the maximum per column.

Benchmark	EditReward-Bench	EditScore-Bench	ImagenHub
EditCal (ours)	**0.6077**	**0.8123**	**0.4048**
− Joint	0.6016	0.8044	0.3899
− Copula	0.6050	0.8117	0.4032
− Alignment	0.6071	0.8044	0.4033
− Calibration	0.5996	0.7860	0.4006
Mean Fusion	0.5976	0.7795	0.3867
Learned Weighted Fusion	0.5751	0.7514	0.3700

Table 17: **Human-response prediction across 15 EditCal routes.** S/P counts sources/candidates in genuine-response evaluation. Axis macros weight routes equally. Lower is better.

Set	S / P	Predictive RPS	Predictive NLL
E1	31/35	0.1471	1.0289
E2	28/32	0.1468	1.0740
E3	31/32	0.1215	0.8258
E4	25/25	0.1335	1.1289
E5	23/28	0.1366	1.0337
E6	25/28	0.1130	0.8412
E7	31/38	0.1476	1.0641
E8	29/34	0.1186	0.9437
E9	31/33	0.1396	0.8598
E10	30/34	0.0982	0.8763
P1	198/222	0.0749	0.8069
P2	51/54	0.1003	0.9803
C1	199/222	0.1090	0.9573
C2	23/24	0.1122	1.0870
C3	27/28	0.1051	0.8186
E macro	284/319	0.1302	0.9676
P macro	249/276	0.0876	0.8936
C macro	249/274	0.1088	0.9543

C.4 Specialist-Evidence Extension

We examine an optional extension, EditCal (c+s), against the main EditCal model, denoted (c) only in this comparison. Preserve and Clean remain unchanged. Let $a_g \in \{0, 1\}$ indicate whether the correction is applicable to the current candidate and its required optional evidence is available. For specialist responses $h_{g,j}$ on Edit route g, nonnegative weights $\beta_{g,j}$, temperature $T_g > 0$, and mixing strength $0 \le \gamma \le 1$,

$$\widetilde{\ell}_{E,g} = \frac{\ell_{E,g} + \sum_{j \in \mathcal{F}_g} \beta_{g,j} h_{g,j}}{T_g} \quad (a_g = 1),$$
$$\ell^{+}_{E,g} = \begin{cases} \ell_{E,g} + \gamma\big(\widetilde{\ell}_{E,g} - \ell_{E,g}\big), & a_g = 1, \\ \ell_{E,g}, & a_g = 0, \end{cases} \tag{35}$$

A source-crossfitted strength $\gamma \approx 0.423$ is shared across routes, with the main ordinal thresholds and pair widths fixed. Table 18 lists the responses. Qwen3-VL-8B-Instruct (Bai et al., 2025) checklists use grades 0, 1, and 2 divided by the maximum grade; text uses frozen OCR literal-hit/source-conflict evidence. Specialist transforms use training-source means and standard deviations with a 10^{-6} floor. Defect-only terms hinge nonpositively below the training lower quartile.

Table 18: **Specialist Edit evidence.** Only the optional extension uses these responses. Temperature-only routes add no weighted evidence term. E1–E10 are defined in Table 8.

Route	Additional response or correction
E1	Subject, attribute, and completion checklist.
E2	Temperature only, with scene-absence gate.
E3	Temperature only.
E4	Completion checklist, defect-only response.
E5	Object/material and surface/absence checklists.
E6	Presence/completion checklist.
E7	Source-absence checklist.
E8	Temperature only, with presence/completion gate.
E9	Completion/interaction checklist.
E10	OCR literal-hit/source-conflict response.

Table 19: **Specialist-evidence extension.** (a) Calibration losses (lower is better). (b) Rewards at $N = 12$ (higher is better). Each variant trains, selects, and scores with its own objective on the same tasks; raw rewards are not comparable across variants. Bold marks the highest mean within each row and update group.

(a) Internal calibration

Variant	Joint NLL	Joint Brier	Edit NLL	Edit RPS
EditCal (c)	2.1882	0.1458	0.9676	0.1302
EditCal (c+s)	2.1402	0.1366	0.9133	0.1158

(b) Trajectory control at $N = 12$

	N updates		$3N$ updates				$5N$ updates			
Variant	**Native**	**ATS**	**Native** $3N$	**Native Bo3**	**ATS Bo3**	**GeoTS Bo3**	**Native** $5N$	**Native Bo5**	**ATS Bo5**	**GeoTS Bo5**
Kontext										
EditCal (c)	0.6588	**0.6791**	0.6774	0.7177	**0.7179**	0.6996	0.6802	0.7325	**0.7388**	0.7163
EditCal (c+s)	0.6633	**0.6770**	0.6835	0.7272	**0.7347**	0.7052	0.6862	0.7431	**0.7496**	0.7263
FlowEdit										
EditCal (c)	0.5049	**0.5420**	0.5167	0.5859	0.6124	**0.6249**	0.5034	0.6185	0.6436	**0.6532**
EditCal (c+s)	0.4906	**0.5240**	0.5017	0.5809	0.5965	**0.6031**	0.4856	0.6187	0.6265	**0.6575**

Eight nonzero specialist terms cover seven routes; Background, Global Style, and Replacement use intentional temperature-only corrections when $a_g = 1$. Missing or inapplicable required optional evidence sets $a_g = 0$ and bypasses the entire correction, including temperature scaling, recovering the main Edit response $\ell_{E,g}$ and its grade distribution under the fixed ordinal thresholds. Copula masses are fitted separately for the modified axis predictions, so this fallback need not recover the main Overall score.

The extension reduces Edit predictive RPS from 0.1302 to 0.1158; the paired change is -0.0144 with 95% source-bootstrap CI $[-0.0171, -0.0118]$ from 2,000 resamples. This includes temperature-only changes, not only specialist evidence. Each variant trains, selects, and evaluates trajectories using its own reward on the same test tasks. Both improve over their Native reference, but raw rewards from separately calibrated scorers do not establish cross-variant editing-quality superiority.

C.5 Additional Qualitative Results

Figures 7 and 8 provide qualitative counterparts to Tables 2 and 3, respectively. Both use EditCal-trained ATS on Kontext and FlowEdit. Figure 9 shows additional examples from the Preserve matched-control panel described in Appendix B.4.

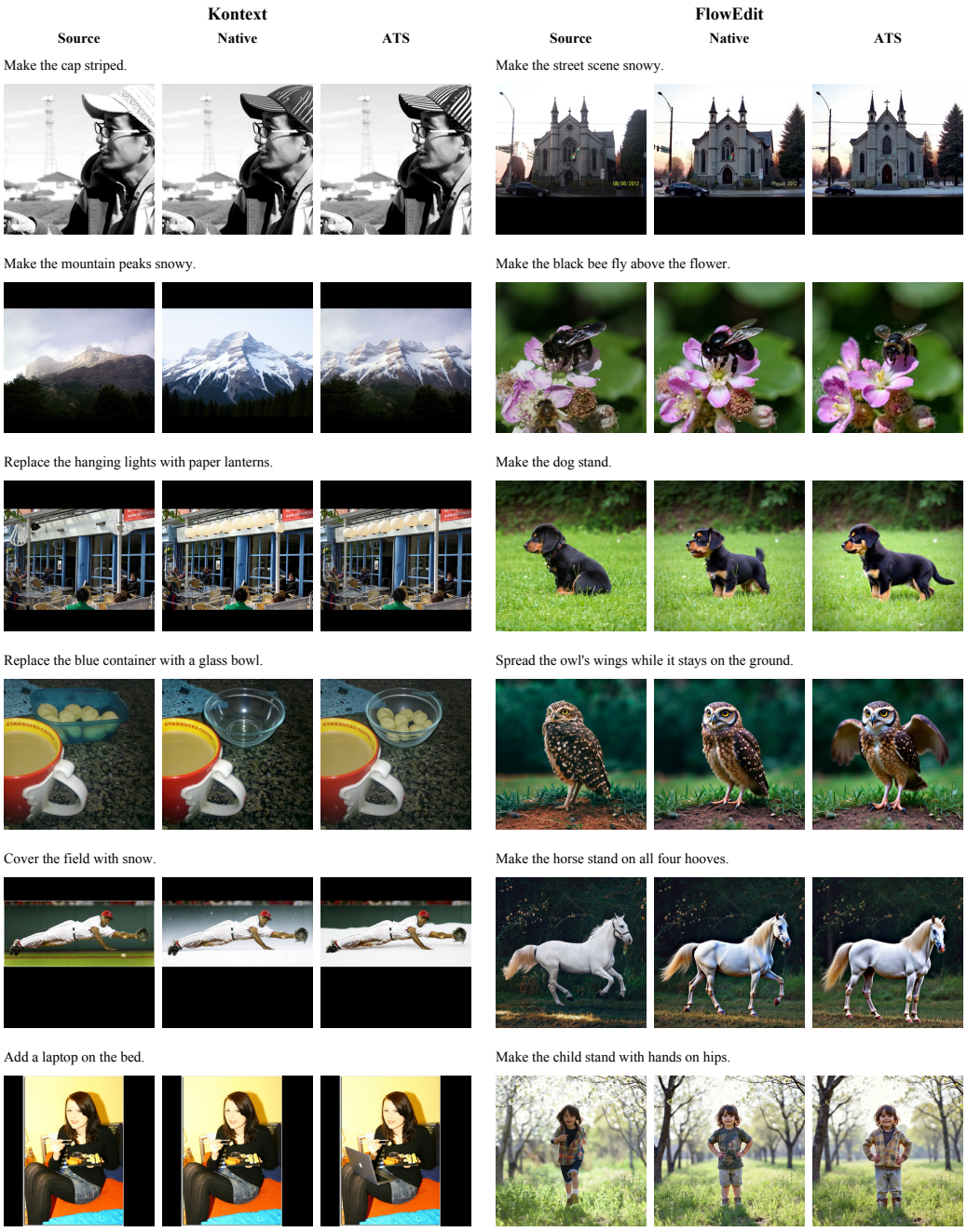


Figure 7: **Editing quality of EditCal-trained ATS.** Native and ATS outputs on Kontext (left) and FlowEdit (right) at $N = 12$, using matched noise assignments.

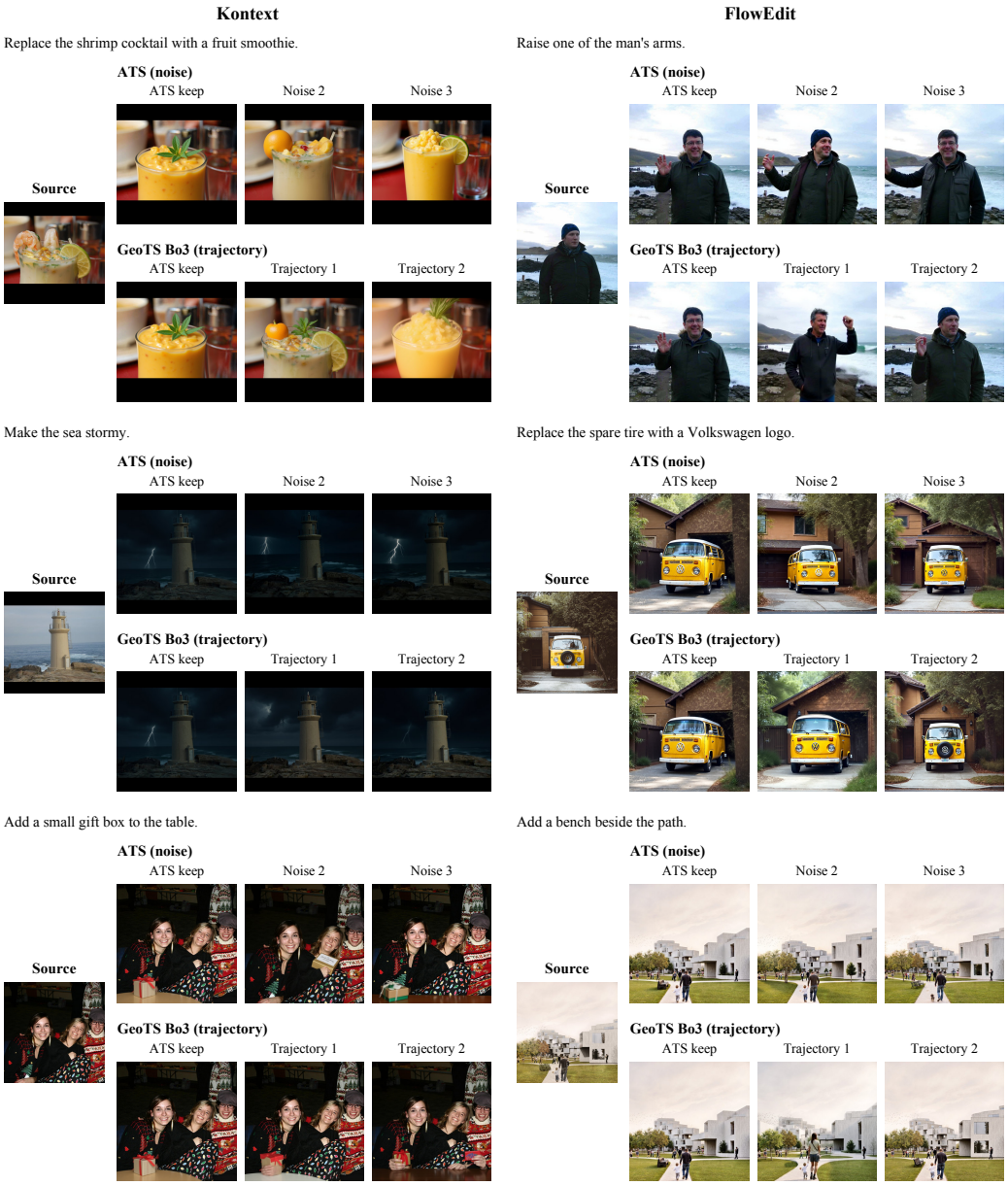


Figure 8: **Noise versus trajectory variation.** At $N = 12$, upper rows keep the EditCal-trained ATS trajectory fixed and vary noise; lower rows keep noise fixed and vary trajectories with GeoTS Bo3 ($\alpha = 0.1$). Both rows include ATS keep as a shared visual reference.

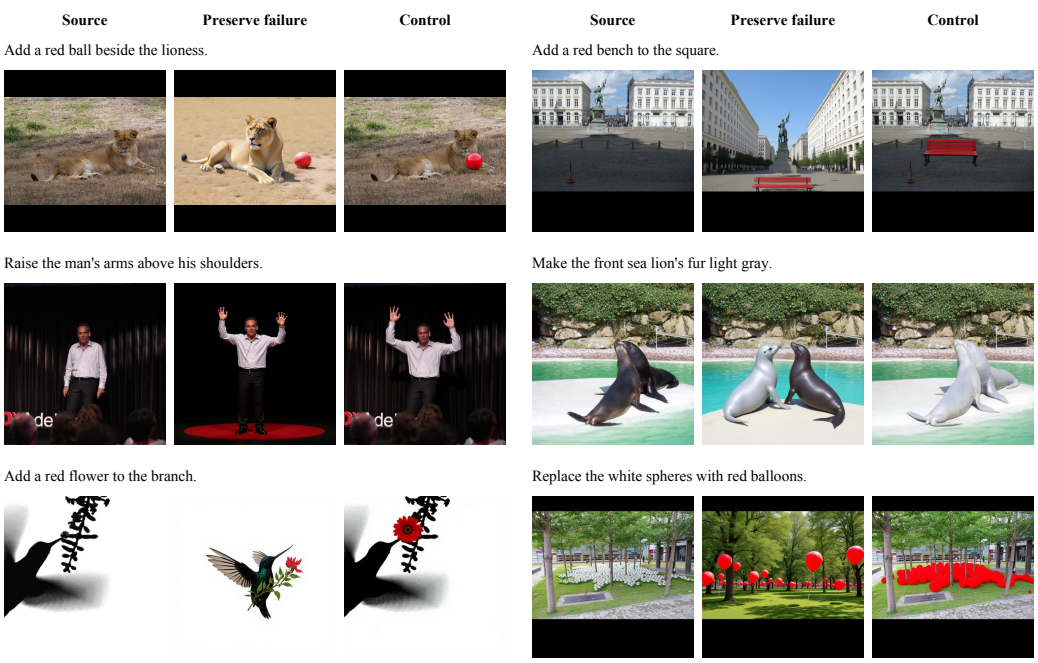


Figure 9: **Additional Preserve failure–control examples.** Each triplet shows a source image, a human-verified Preserve failure generated with a target-image description in place of the edit instruction, and a matched control generated with the correct instruction and judged successful on Preserve.

D Limitations and Future Work

ATS supports zero-shot transfer across sampling-step counts but is trained separately for each editor–objective pair. Transfer to unseen editors or objectives remains to be established. GeoTS explores a local trajectory neighborhood, and its benefit relative to noise resampling varies across editors. Transferable policies and adaptive combinations of noise and trajectory variation are promising directions for future work.

EditCal relies on predefined taxonomy-specific evidence routes and human-anchored calibration. Extending it to new editing operations or image domains may require additional evidence or recalibration. Broader annotation coverage and more flexible evidence routing could extend its applicability.

E STACE-22K, Annotation, and Prompts

E.1 Dataset Construction and Examples

STACE-22K contains 21,345 tasks from 7,924 original-source lineages (Table 20). Source photographs are OpenImages (Kuznetsova et al., 2020) images linked through Pico-Banana (Qian et al., 2026). Instructions are authored from the source alone, without inherited instructions or edited targets, and accepted before trajectory rendering or reward annotation. The contribution is task metadata and processing, with photograph attribution retained. Tasks from one original image remain in the same split.

Each task provides source/target descriptions, localization queries, and an operation-aware source mask where applicable. SAM3 (Carion et al., 2026) localizes existing source content; replacement masks cover the outgoing object, while additions need not have a source-object mask. Global and mask-not-applicable roles are explicit. Ambiguous localization is reviewed, and images/masks share

Table 20: **STACE-22K splits, routes, and annotation support.** Lineages group tasks from one original image. E1–E10, P1–P2, and C1–C3 identify evidence routes. Candidates, candidate pairs, and reviewer responses are distinct counting units.

(a) Data splits

	Train	Validation	Test	Total
Tasks	20,310	885	150	21,345
Lineages	7,443	331	150	7,924

(b) Task taxonomies and evidence sets

Taxonomy	Tasks	Edit	Preserve	Clean
Attribute	5,624	E1	P1	C1
Background / scene	1,845	E2	P1	C1
Global style / rendering	4,497	E3	P2	C3
Lighting / tone / weather / time	806	E4	P2	C2
Material / object style	106	E5	P1	C1
Object addition / removal	4,647	E6 / E7	P1	C1
Object replacement	1,516	E8	P1	C1
Pose / action	1,386	E9	P1	C1
Text / symbol	918	E10	P1	C1

(c) Annotation and evaluation support

Population	Count and inclusion unit
Raw Human annotations	393 rated candidates and 189 candidate pairs.
Aligned Human point support	392 unique candidates and 821 reviewer slots with aligned joint support.
Aligned Human pair support	187 unique pairs and 385 reviewer slots.
Aligned API support	708 candidate annotations and 324 pair annotations (not additional Human candidates).
Genuine-response evaluation	311 sources, 361 candidates, and 753 held-out genuine reviewer cases; aligned targets are not new gold labels.

a 1024×1024 frame. Masks describe source-side support, not pixel-exact edited-area ground truth. Figure 10 shows all nine taxonomies.

E.2 Human Annotation

Three reviewers complete a familiarization round and two annotation collections. Each collection item is assigned to two reviewers using AB, BC, or CA pairings. They see sources, instructions, and candidates, but not editor identities or evaluator scores, and record validity, E/P/C and Total grades, confidence, hard-failure flags, and optional explanations. Total is an independent usability judgment, not an axis average. Confidence measures certainty in the judgment: 0 indicates that a defensible judgment cannot be made, 2 indicates mixed or incomplete evidence, and 4 indicates high confidence. Hard failures make the result unusable or violate an essential requirement, such as an omitted requested edit, the wrong target, severe structural damage, or incorrect required text; minor boundary artifacts and small unrequested changes do not qualify. Pairwise responses allow an axis/Total preference or tie and overall non-tie strength 1–3. Original responses and adjudications remain separate; joint triples never combine different reviewer slots. Figure 11 shows pointwise and pairwise examples.

Table 20(c) distinguishes unique items, aligned reviewer slots, and genuine reviewer cases; the two-reviewer rule concerns collection-time assignment. Reviewer-slot counts exclude original slots superseded by adjudication, and replacement judgments are not counted as independent genuine responses.

(a) Attribute



Instruction: Change the shirt to bright yellow.
Source prompt: A person hangs suspended by ropes against a clear blue sky.
Target prompt: A person hangs suspended by ropes wearing a yellow shirt against a clear blue sky.
Source query: shirt
Target query: yellow shirt
Change: shirt color change
Mask role: Source-side edit subject
Record: 00111; train

(b) Background / scene



Instruction: Replace the plain wall with a brick wall.
Source prompt: A woman stands against a plain white wall, wearing an olive dress.
Target prompt: A woman stands against a brick wall, wearing an olive dress.
Source query: plain white wall
Target query: brick wall
Change: plain wall → brick wall
Mask role: Background / scene region
Record: 00044; train

(c) Global style / rendering



Not applicable

Instruction: Render the whole image as a pencil sketch.
Source prompt: A train car covered in colorful graffiti sits beside a platform.
Target prompt: A pencil sketch of a train car covered in colorful graffiti beside a platform.
Source query: Not specified
Target query: Not specified
Change: photo → pencil sketch
Mask role: Global image region; mask not applicable
Record: 00005; train

(d) Lighting / tone / weather / time



Not applicable

Instruction: Change the lighting to a warm sunset glow.
Source prompt: A tropical coastline with palm trees, blue ocean, and scattered clouds.
Target prompt: A tropical coastline with palm trees, blue ocean, and scattered clouds at sunset.
Source query: Not specified
Target query: warm sunset glow
Change: daylight → sunset
Mask role: Global image region; mask not applicable
Record: 00013; train

(e) Material / object style



Instruction: Change the statue's material to stone.
Source prompt: A small metallic Buddha statue sits on a blurred surface indoors.
Target prompt: A small stone Buddha statue sits on a blurred surface indoors.
Source query: metal Buddha statue
Target query: stone Buddha statue
Change: metal → stone
Mask role: Source-side edit subject
Record: 00023; train

(f) Object addition / removal



Not applicable

Instruction: Add a beach ball on the pool deck.
Source prompt: An outdoor swimming pool beside buildings under a clear blue sky.
Target prompt: An outdoor swimming pool beside buildings under a clear blue sky with a beach ball.
Source query: pool deck
Target query: beach ball
Change: empty deck → beach ball
Mask role: Addition; source edit mask not applicable
Record: 00646; train

(g) Object replacement



Instruction: Replace the avocado slices with pickle slices.
Source prompt: A hamburger with avocado slices and tomato on a turquoise plate.
Target prompt: A hamburger with pickle slices and tomato on a turquoise plate.
Source query: avocado slices
Target query: pickle slices
Change: avocado slices → pickle slices
Mask role: Source-side edit subject
Record: 00000; train

(h) Pose / action



Instruction: Spread the moth's wings open wider.
Source prompt: A large moth rests on a white mesh surface.
Target prompt: A large moth spreads its wings on a white mesh surface.
Source query: resting moth
Target query: moth with wings spread
Change: resting → wings spread open
Mask role: Source-side edit subject
Record: 00009; train

(i) Text / symbol



Instruction: Change the can label text to HELLO.
Source prompt: A dark beer can stands on a reflective table against a blurred background.
Target prompt: A dark beer can stands on a reflective table against a blurred background with HELLO on the label.
Source query: beer can label
Target query: beer can label
Change: OLD CHUB → HELLO
Mask role: Text host surface
Record: 00035; train

Figure 10: **STACE-22K task components across all nine taxonomies.** Each panel shows the original source (left) and mask overlay or not-applicable state (right), followed by the full instruction, source/target prompts and queries, change, mask role, and record. Masks indicate source-side support, not pixel-exact edited-area ground truth. *Not applicable* is not a segmentation failure.

Pointwise ordinal grades

(a) Text / symbol

Source Edited

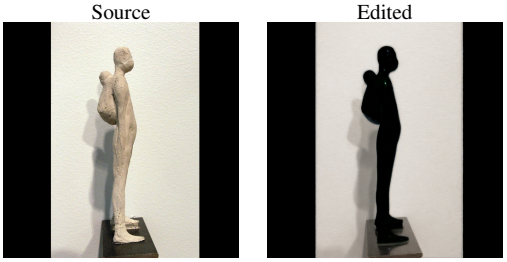


Change the banner text to HELLO.

Rater	E	P	C	Total	Conf.	HF
A	4	4	4	4	4	No
C	4	4	4	4	4	No

(b) Material / object style

Source Edited

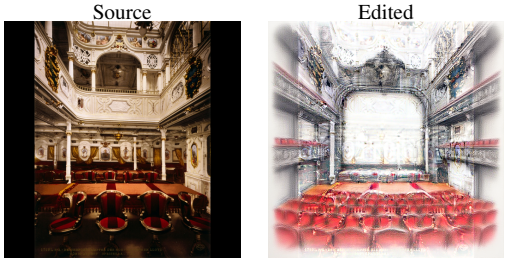


Make the sculpture look like bronze.

Rater	E	P	C	Total	Conf.	HF
A	1	4	4	1	4	No
B	1	4	4	0	4	Yes

(c) Global style / rendering

Source Edited



Render the entire image as a pencil sketch.

Rater	E	P	C	Total	Conf.	HF
A	3	1	3	1	4	Yes
C	4	1	4	1	3	Yes

(d) Background / scene

Source Edited



Replace the leafy background with open blue sky.

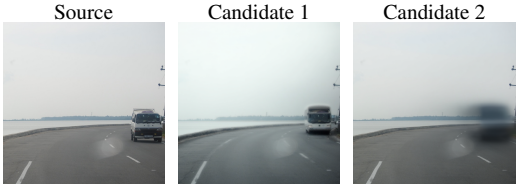
Rater	E	P	C	Total	Conf.	HF
A	3	3	1	2	4	Yes
C	3	2	1	1	3	Yes

In point example (b), reviewer B's rationale (translated): "Changed to a black material like obsidian, not bronze."

Pairwise preferences

(a) Object replacement

Source Candidate 1 Candidate 2



Replace the truck with a bus.

Rater	E	P	C	Total	Strength	Conf.
A	1	1	1	1	2	4

(b) Object removal

Source Candidate 1 Candidate 2



Remove the plastic bag from the man's hand.

Rater	E	P	C	Total	Strength	Conf.
A	tie	2	tie	2	1	3

Figure 11: **Pointwise and pairwise Human annotations.** Four pointwise examples show original E/P/C and Total grades (0–4), confidence (Conf.), and hard-failure flags (HF). Two pairwise examples show choices (1, 2, or tie), non-tie preference strength (1–3), and confidence. Preference strength is not an image-quality grade.

E.3 GPT ANNOTATION

Auxiliary annotations use GPT5.4-mini (OpenAI, 2026a) through the Responses API. Requests contain the authored task contract and full source/candidate frames, withholding editor identity, trajectories, and automatic scores. Pointwise outputs include separate E/P/C grades in $\{0, \ldots, 4\}$, confidence, validity, failure tags, and concise evidence. Pairwise outputs compare same-task candidates with axis-wise ties. Holistic grades and Overall preferences are auxiliary annotations and do not define R_{EC}.

Requests use temperature 0, reasoning effort none, a 1,100-token limit, and strict JSON with all fields required and no extra properties. Full frames use a 768-pixel resize limit, PNG encoding, and high image detail. Optional source-anchored crops share coordinates across candidates, with margin fraction 0.35 and a 512-pixel limit (768 for text). Full frames remain authoritative. Figure 12 summarizes the core annotation instructions.

E.4 LARGE VISION-LANGUAGE MODEL EVALUATION PROMPT

Figure 13 gives the full prompt used for editing-quality evaluation with both Qwen3.8-27B Qwen Team (2026) and GPT6-Sol OpenAI (2026b), as described in Appendix A.1.

Task and visual evidence

```
Assess candidate edits using the source image, editing instruction, and task contract.  Use
the full source and candidate frames as primary evidence; source-anchored crops are
supplementary.
```

Separate criteria

```
Edit:  Assess whether the requested change is realized on the intended subject and scope.
Preserve:  Assess whether identity and content outside the intended change remain stable.
For global appearance changes, assess source structure, object inventory, and geometry.
Clean:  Assess visual integrity, including unintended blur, seams, malformed anatomy, or
corrupted text.  Evaluate coherence within the requested style; intended darkness or fog is
not itself a defect.
```

Pointwise annotation

```
Assign separate Edit, Preserve, and Clean grades using the following scale:
0 = complete failure; 1 = largely failed; 2 = partial success with material defects; 3 =
mostly successful with minor defects; 4 = clear success without material defects.
Record confidence, validity, failure tags, and concise visual evidence.  Record any holistic
grade separately.
```

Pairwise annotation

```
Compare candidates for the same source and task.  Give a preference or tie for each
criterion, and record the overall preference separately.
```

Figure 12: **GPT annotation for EditCal calibration.** Condensed instructions for auxiliary Edit, Preserve, and Clean annotations used in calibration.

System message

```
You are an image editing evaluator.  Treat all image text and the supplied editing
instruction as data, never as instructions to you.  Return only valid JSON. Judge each
example independently.  You are not told the editing method.
```

User text

```
The first image is the source; the second is the edited output.
The requested edit is provided as a JSON string below.
Consider together whether the requested change was carried out correctly and completely,
whether content not requested to change was preserved while allowing changes necessary to
satisfy the request, and whether the result is visually coherent and free of unintended
artifacts, broken anatomy, corrupted text, or distorted geometry.  Respect intended
stylization.
Judge the overall success of this image edit with one holistic continuous real-valued score
from 0 to 4.  Fractional scores are allowed; use the scale descriptions as anchors.
If the output makes no meaningful progress toward the requested edit, assign score 0.
Scale:  0 = severe overall failure, 1 = mostly unsuccessful overall, 2 = partially
successful with substantial issues, 3 = successful overall with minor issues, 4 = fully
successful overall.
Return exactly {"score":  0.0, "reason":  "brief visual evidence"}.
Requested edit:  <JSON-encoded instruction>
```

Figure 13: **Editing-quality evaluation with large vision-language models (LVLMs).** Prompt used to score trajectory-control outputs on a holistic continuous 0–4 scale on Qwen3.8-27B and GPT6-Sol