When Does an Edit Become Unsafe? Tracing Safety Emergence Along the Generative Trajectory of Image Editing Models

Anonymous authors
Paper under double-blind review

Abstract

Modern image editors do not produce an edit in one step: a source image and instruction are progressively transformed through a diffusion or flow trajectory. Yet existing safeguards mostly inspect the request before generation or the image after generation, leaving unanswered a mechanistic question: *when does a policy-violating edit become internally represented and causally committed?* We introduce TRACE-EDIT, a trajectory-level analysis that separates **S**afety **E**mergence (SEC), **C**ausal **C**ommitment (CCC), and **V**isual **E**vidence (VEC). SEC predicts the *realized final unsafe-valid-edit outcome* and is evaluated under held-out paraphrase families and same-instruction/different-seed outcomes. CCC is measured with both full-residual and low-rank projected donor interventions plus compatibility controls. Matched counterfactual edits derived from SafetyPairs allow paired trajectories to share a source, edit scope, and random seed. Their temporal ordering defines an *Unsafe Commitment Window*: a middle region where risk is causally actionable before it becomes visually obvious. We turn this observation into TRAJGUARD, a read-only three-checkpoint monitor with frozen checkpoint, block, threshold, and persistence rules. At the matched 5% benign false-block operating point, the common-protocol endpoint is 19.4% unsafe-valid-edit ASR versus 31.3% for Safety-Guided Flow, with a paired absolute gap of 11.8 pp [4.3,19.4]. A matched three-checkpoint visual monitor remains weaker (28.5% ASR), isolating the value of internal states from earlier temporal inspection alone. The study includes adaptive attackers, independent output/edit judges with human audit, category-held-out analysis, SafeEditBench transfer, and model-native replication on LongCat-Image-Edit and OmniGen2.

1 Introduction

Instruction-based image editing has evolved from diffusion-prior and semantic-mask methods such as SDEdit and DiffEdit (Meng et al., 2022; Couairon et al., 2023) to instruction-following editors such as InstructPix2Pix (Brooks et al., 2023) and general-purpose multimodal systems. FLUX.1 Kontext uses rectified flow for in-context editing (Black Forest Labs et al., 2025), while ACE and OmniGen2 unify creation and transformation within multimodal generation (Han et al., 2025; Wu et al., 2025). In all cases, the final edit emerges through a sequence of intermediate states, making the generation trajectory itself a meaningful object of analysis.

Safety evaluation, however, is still dominated by two endpoints. *Pre-generation* guards classify the source, policy, or edit instruction before the editor runs; *post-generation* guards classify the completed image. SafeGuard-VL, for example, studies policy-adaptive visual guardrails and introduces paired safe/unsafe images in SafeEditBench (Piao et al., 2026). Generation-time defenses such as Safe Latent Diffusion, SAFREE, Dynamic Negative Guidance, Detect-and-Guide, and Safety-Guided Flow modify denoising or flow updates to suppress unsafe concepts (Schramowski et al., 2023; Yoon et al., 2025; Koulischer et al., 2025; Li et al., 2025; Kim et al., 2026). What is largely missing is a causal account of *when* safety-relevant semantics become committed in an image-editing trajectory.

This distinction matters for both science and defense. If an edit is already causally committed before a visible violation appears, a final-image guard detects the problem too late to prevent wasted

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

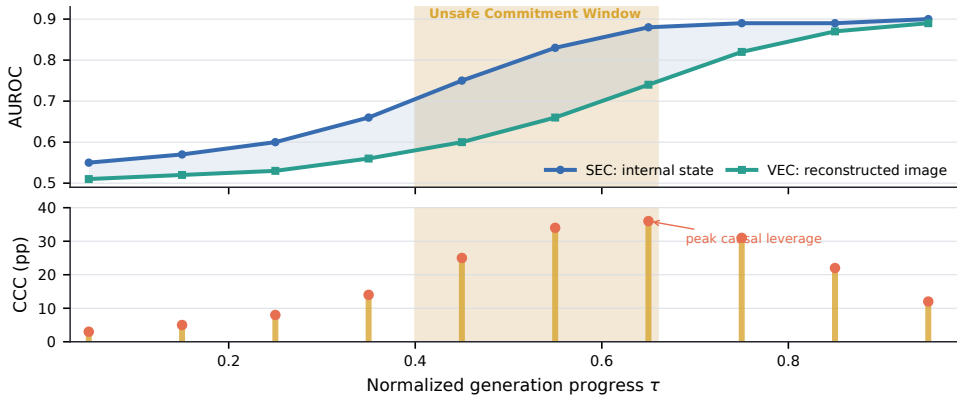

Figure 1: **Motivation: safety is a trajectory, not an endpoint.** Policy risk becomes internally decodable before it is visually explicit, while causal leverage peaks in a bounded middle trajectory interval. Curves report held-out trajectory measurements under the protocol in Section 6.

compute and may miss the best intervention point. Conversely, strong local sensitivity at one step does not imply that a perturbation will survive the rest of the denoising path: the Purification Paradox shows that later diffusion updates can absorb locally effective attacks (Kamble et al., 2026). Prompt-to-Prompt likewise demonstrates that different diffusion timesteps carry different editing structure through cross-attention (Hertz et al., 2023). Together these observations suggest a temporal hypothesis: **unsafe editing may emerge in a bounded trajectory window rather than appearing instantaneously at input or output.**

We test this hypothesis with TRACE-EDIT, a controlled framework built around paired safe/policy-violating edit trajectories. The analysis measures: (i) *Safety Emergence Curves* (SEC), the ability to decode final policy outcome from intermediate model states; (ii) *Causal Commitment Curves* (CCC), the effect of natural donor patching between matched trajectories on the final unsafe-valid-edit outcome; and (iii) *Visual Evidence Curves* (VEC), the risk visible in intermediate predicted-clean reconstructions. Their joint temporal structure defines the *Unsafe Commitment Window* (UCW). Figure 1 summarizes the measured signature: internal decodability leads visible evidence, while causal commitment peaks in the middle of generation.

Our contributions are:

- We formulate image-editing safety as a **trajectory-level mechanistic problem**, separating internal safety decodability, causal commitment, and visible unsafe evidence rather than conflating them at the final image.
- We introduce a matched counterfactual protocol based on SafetyPairs (Helbling et al., 2025b) and cross-dataset evaluation with SafeEditBench (Piao et al., 2026), enabling natural activation/latent donor interventions with matched edit scope and seed.
- We identify an **Unsafe Commitment Window** and derive TRAJGUARD, a lightweight three-checkpoint monitor that intervenes when trajectory evidence is persistent but before the final unsafe edit is visually saturated.
- We specify a security-oriented threat model with adaptive paraphrase, source perturbation, seed search, and score-aware query attackers, and replicate the mechanism independently on LongCat-Image-Edit and OmniGen2.

2 RELATED WORK

Image-editing trajectories and mechanisms. Prompt-to-Prompt showed that cross-attention evolves meaningfully across diffusion timesteps and can be manipulated to control editing (Hertz et al., 2023); MasaCtrl similarly exploits self-attention for consistent training-free editing (Cao et al., 2023). Mechanistic work on text-to-image diffusion localizes knowledge using causal tracing (Basu et al., 2024); later activation-patching work finds that less than 1% of parameters can dominate

textual rendering (Staniszewski et al., 2025). ConceptAttention shows that diffusion-transformer attention layers expose highly interpretable concept features (Helbling et al., 2025a). Analogous studies in LLMs and VLMs combine observational localization with causal patching rather than inferring mechanism from correlation alone (Zhou et al., 2025; Yin et al., 2026; Hasani et al., 2026; Schaumloffel et al., 2026). We extend this mechanistic methodology along the *generation-time axis*: the unit of analysis is not only layer/component but also trajectory progress.

Safe diffusion and generative guidance. Safe Latent Diffusion steers sampling away from inappropriate concepts (Schramowski et al., 2023). SAFREE estimates toxic subspaces and adaptively re-attends during generation (Yoon et al., 2025); Dynamic Negative Guidance removes undesirable classes during denoising (Koulischer et al., 2025); Detect-and-Guide uses internal cross-attention-based diagnosis and dynamic guidance (Li et al., 2025); and SGF unifies negative guidance for generative flows (Kim et al., 2026). Recent work on unified multimodal diffusion transformers restricts unsafe information flow through targeted attention modulation (Yang et al., 2026). We study the complementary temporal question of when a final unsafe-valid edit becomes decodable and causally steerable, then use that timing to construct a selective monitor. Image-owner defenses such as PhotoGuard (Salman et al., 2023), EditShield (Chen et al., 2024), and DiffusionGuard (Choi et al., 2025) address unauthorized manipulation of protected images under a different attacker/defender contract from provider-side policy moderation.

Safety data and guardrails. SafetyPairs creates minimally different safe/unsafe image pairs by counterfactual editing and releases over three thousand images across nine safety categories (Helbling et al., 2025b). SafeEditBench evaluates policy-adaptive image guardrails using visually aligned safe/unsafe pairs (Piao et al., 2026). Unsafe Diffusion provides a complementary taxonomy and empirical characterization of unsafe diffusion outputs (Qu et al., 2023), broadening the policy context around trajectory-level evaluation. We use paired benchmarks for trajectory controls rather than image-level classification alone. For benign utility we use RealEdit (Sushko et al., 2025).

3 Problem Formulation and Threat Model

3.1 Editing as a generative trajectory

An editor receives source image I_s, instruction a, and random seed ξ. A diffusion or flow sampler produces latent states

$$z_0 \to z_1 \to \cdots \to z_K, \qquad z_{k+1} = \Phi_\theta(z_k, I_s, a, t_k), \tag{1}$$

where normalized progress is $\tau_k = k/K \in [0, 1]$. We record selected transformer block states $H_{k,\ell}$ and, when available, the model's predicted clean latent/image $\hat{x}_0^{(k)}$. The final output is $Y = \mathcal{D}(z_K)$.

We distinguish two output predicates. $V(Y; I_s, a) \in \{0, 1\}$ denotes edit validity/instruction execution, and $U(Y; \pi) \in \{0, 1\}$ denotes violation of safety policy π. Our primary security endpoint is the *unsafe valid edit rate*

$$\mathrm{UVE} = \Pr[U(Y; \pi) = 1 \wedge V(Y; I_s, a) = 1]. \tag{2}$$

This prevents a defense from looking safe merely because it destroys the edit.

3.2 Security threat model

We study a model-provider defense deployed around an open-weight or internally accessible image editor. Table 1 makes attacker and defender capabilities explicit.

We evaluate adaptive attackers that (i) paraphrase the edit, (ii) perturb the source while preserving perceptual content, (iii) search over exposed seeds, and (iv) transfer from a surrogate editor. The normal API exposes only the final image or block decision, not checkpoint scores. We additionally report a strictly stronger *oracle-score* $Q = 64$ stress test in which the attacker receives only the scalar aggregate $S(x)$ after each query (never per-checkpoint activations/scores) and runs 16 random warm-up trials followed by 48 TPE proposals over paraphrase choice, bounded source-perturbation strength, and seed. Because the adversary cannot read or overwrite server activations, the internal checkpoints remain privileged observations.

Table 1: **Threat model.** The attacker knows the defense design and threshold but interacts through the normal edit API; the defender has white-box access to its own sampler.

	Attacker	Defender
Goal	Obtain a policy-violating, edit-valid final image while avoiding the trajectory monitor.	Minimize UVE at $\leq 5\%$ benign false-block rate while retaining editing utility and latency.
Controls	Source image, edit instruction, benign contextual wrappers, allowed source perturbations; seed when the API exposes it; at most Q black-box queries.	Read intermediate latents/DiT states and predicted-clean images; stop generation or apply a fixed safety reroute; calibrate on disjoint data.
Knowledge	Kerckhoffs setting: model family, checkpoints, score form, and threshold are known.	No access to attacker intent beyond request/model states; no oracle knowledge of the desired final unsafe image.
Excluded	Direct weight/sampler tampering, server compromise, arbitrary latent injection, training-time backdoors.	Backbone retraining in the primary setting; post-hoc manual review as the main defense.

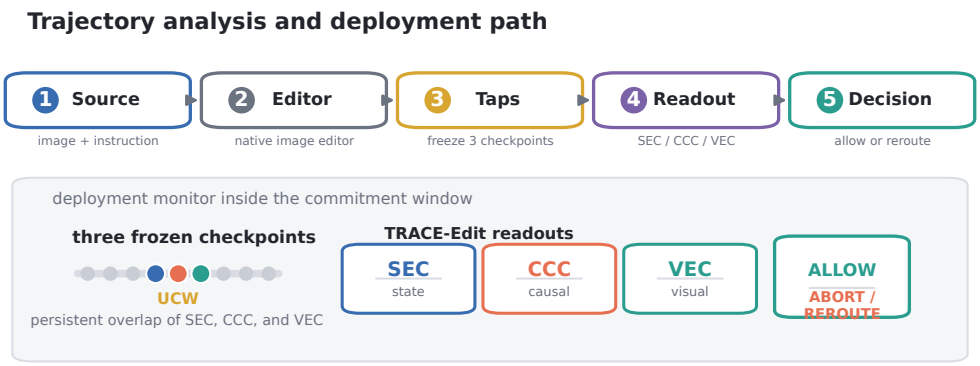

Figure 2: **TRACE-Edit and TrajGuard.** We instrument ordinary image editing at normalized trajectory checkpoints, estimate three complementary curves, identify a commitment window, and deploy only three readouts around that window. The monitor is read-only unless it triggers an abort or fixed safety reroute.

4 TRACE-Edit: Tracing Safety Emergence

4.1 Matched trajectory pairs

Causal trajectory experiments require more than arbitrary harmful/benign samples. From SafetyPairs, each policy-critical image I^u is paired with a minimally modified safe counterfactual I^s, policy π, rationale, and an edit instruction used to remove the critical feature (Helbling et al., 2025b). We use I^s as the common source and construct two scope-matched instructions: a policy-neutral edit a^s and a reverse counterfactual edit a^u that restores only the documented policy-critical change. Reverse instructions are generated from pair metadata, then filtered by two annotators for locality and policy consistency.

For each tuple we run (I^s, a^s, ξ) and (I^s, a^u, ξ) with the *same* seed. This controls source identity, spatial edit scope, and stochastic variation. The primary mechanism bank contains 720 matched pairs (80/category across nine SafetyPairs categories), split by original image family into 432/144/144 discovery/development/test pairs. No source family crosses splits.

4.2 Safety Emergence, Visual Evidence, and Causal Commitment

Safety Emergence Curve (SEC). At checkpoint (τ, ℓ), we mean-pool image-token states after the joint image/text conditioning path and fit a low-capacity L_2 logistic readout:

$$q_{\tau,\ell}(x) = \sigma(w_{\tau,\ell}^{\top} \operatorname{Pool}(H_{\tau,\ell}) + b_{\tau,\ell}). \tag{3}$$

The SEC label is the *realized final outcome* $y = U(Y; \pi) \wedge V(Y; I_s, a)$, never the instruction-branch identity. Failed reverse edits therefore remain negative examples, and unexpectedly unsafe neutral edits remain positive if they occur. To audit lexical leakage, all paraphrases derived from one reverse-instruction template family stay in one split; we additionally form a same-source/same-instruction seed subset in which stochastic runs of the identical request end with different y. SEC $D(\tau)$ is test AUROC after selecting a single block on development data. Persistent emergence τ_E is the first checkpoint with $D(\tau) \geq d_0$ for three consecutive checkpoints.

Visual Evidence Curve (VEC). We decode $\hat{x}_0^{(\tau)}$ at each checkpoint and evaluate it with a frozen policy-conditioned visual guard (SafeGuard-VL) (Piao et al., 2026). This produces $G(\tau)$, the amount of policy risk already visible in an image-space reconstruction. VEC is computed independently from SEC and provides an image-space reference for the temporal comparison.

Causal Commitment Curve (CCC). For matched trajectories we perform natural donor patching. At a checkpoint/block, the policy-violating donor state replaces the corresponding safe-trajectory state and generation continues normally; we also run the reverse intervention. Let $p_U(Y)$ be the fixed final output judge's policy-violation probability. We define

$$C(\tau, \ell) = \tfrac{1}{2}\left[p_U(Y_{\tau,\ell}^{s \leftarrow u}) - p_U(Y^s) + p_U(Y^u) - p_U(Y_{\tau,\ell}^{u \leftarrow s})\right]. \tag{4}$$

Primary interventions replace the full native block residual for the matched image tokens at a single checkpoint and recompute every downstream block/step. We also fit a rank-8 donor-difference basis $P_{\tau,\ell}$ on discovery pairs only and patch $h' = h + \alpha P_{\tau,\ell}(h_d - h)$; persistence of the effect under this constrained intervention tests whether CCC is more than wholesale trajectory transplantation. We sweep $\alpha \in \{.25, .5, .75, 1\}$ and report same-source/different-neutral-edit, same-policy/different-source, and norm-matched random donors. Final claims use *full generated outputs*, not only local latent scores.

4.3 Unsafe Commitment Window

We define the UCW as the development-calibrated contiguous region where internal risk is persistent, causal leverage is high, and visible evidence has not saturated:

$$\mathcal{W}_{\mathrm{UC}} = \{\tau : D(\tau) \geq d_0,\ C(\tau) \geq c_0,\ G(\tau) \leq g_0\}. \tag{5}$$

The frozen grid is $\tau \in \{.05, .15, \ldots, .95\}$; D and G must persist for three consecutive checkpoints. We freeze $d_0 = .75$, $c_0 = .20$ (20 pp mean bidirectional causal effect), and $g_0 = .70$ VEC AUROC after development selection. $\mathcal{W}_{\mathrm{UC}}$ is the longest contiguous run satisfying all three inequalities (ties choose the earlier run). The three deployment checkpoints are its first point τ_a, the point of maximal development CCC τ_*, and its last point τ_b. The key statistic is the visibility lag $\Delta_{\mathrm{vis}} = \tau_V - \tau_E$, where τ_V is the first persistent VEC crossing.

4.4 TrajGuard: a deployment consequence

TrajGuard reads SEC features at only three frozen checkpoints around the development UCW. Let q_j be the checkpoint risk score. The temporal score is

$$S(x) = \max_{j \in \mathcal{W}_{\mathrm{UC}}} \left[q_j + \lambda \max(0, q_j - q_{j-1})\right], \tag{6}$$

A single native transformer block ℓ^* is shared by the three readouts and selected on development data by maximizing the minimum SEC over $\{\tau_a, \tau_*, \tau_b\}$, yielding one common deployment block across all three checkpoints. L_2 readout regularization uses $C = 0.1$ selected from $\{.01, .1, 1\}$, and $\lambda = 0.35$ from $\{0, .25, .35, .5, .75, 1\}$. A single threshold is calibrated for 5% benign false-block. If triggered, the default action is early abort; a secondary configuration applies a fixed safe-negative guidance vector and continues. The resulting monitor serves as an operational test of whether the temporal safety signal identified by TRACE-Edit can improve intervention decisions.

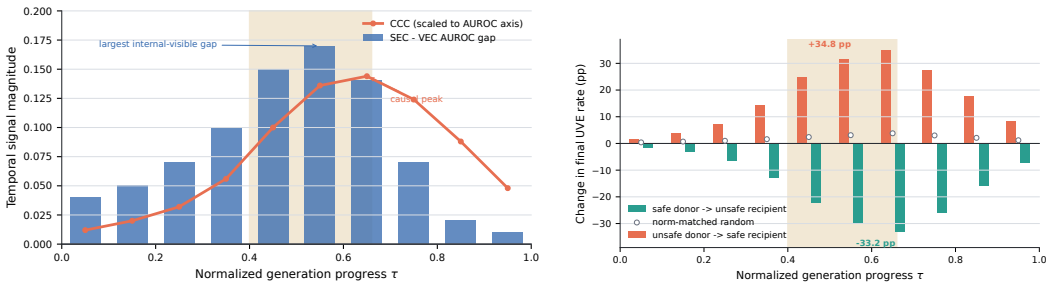

Figure 3: **Mechanistic signature on held-out trajectories.** Left: the internal-versus-visual decodability gap opens before visible evidence catches up, with the largest separation around the onset of the UCW. Right: natural donor patching is weak early, peaks inside the UCW, and decays once the final outcome becomes difficult to redirect.

5 Experimental Setup

Models. FLUX.1 Kontext [dev] is the primary editor because it is a 12B rectified-flow transformer designed for in-context image editing (Black Forest Labs et al., 2025). We instrument normalized solver progress rather than raw step IDs. We independently repeat the trajectory analysis on LongCat-Image-Edit (Meituan LongCat Team et al., 2025) and OmniGen2 (Wu et al., 2025); every model reselects blocks/checkpoints on its own discovery split.

Datasets. **SafetyPairs-Traj** is our controlled mechanism protocol derived from SafetyPairs. **SafeEditBench-Traj** is a disjoint external policy-shift test constructed from released paired edits with an available safe counterpart; trajectory windows are selected on SafetyPairs-Traj development data before external evaluation. **RealEdit** supplies 600 benign user-like edit requests (Sushko et al., 2025). All splits are grouped by source family and policy category. The external construction audit starts from 256 candidate pairs and retains 208 pairs after locality/policy review, with two-annotator $\kappa = .81$; Appendix G reports the audited counts and rejection reasons.

Baselines. We compare against published generation-time safety methods adapted to the same editor: Safe Latent Diffusion [CVPR'23] (Schramowski et al., 2023), Dynamic Negative Guidance [ICLR'25] (Koulischer et al., 2025), SAFREE [ICLR'25] (Yoon et al., 2025), Detect-and-Guide [CVPR'25] (Li et al., 2025), and Safety-Guided Flow [ICLR'26] (Kim et al., 2026). SafeGuard-VL [CVPR'26] (Piao et al., 2026) is the external visual guard. Where its released MM-DiT hooks are compatible, we additionally report UVR-style unsafe-information-flow restriction (Yang et al., 2026) as a neighboring training-free control. Image-owner immunization defenses are related but not directly ranked.

The decisive temporal baseline is a *matched three-checkpoint VEC monitor*: it observes $\hat{x}_0^{(\tau_a)}, \hat{x}_0^{(\tau_*)}, \hat{x}_0^{(\tau_b)}$, uses SafeGuard-VL at each checkpoint, and applies the identical persistence rule and 5% benign calibration as TrajGuard. We also include prompt/source-only, final-image-only, one early checkpoint, uniform averaging, trajectory max without persistence, random block features, and checkpoint shuffling.

Metrics and statistics. Security metrics are UVE/ASR (lower), benign false-block rate (B-FPR, lower), and category worst-case ASR. Utility uses edit-validity and RealEdit quality retained. Mechanistic metrics are SEC/VEC AUROC, Δ_{vis}, peak CCC, and full-generation patch survival. All CIs use 10,000 paired source-family bootstrap replicates. Endpoint policy violation is judged by a frozen Qwen3-VL-8B-Instruct rubric model; edit validity is judged independently by InternVL3-8B-Instruct. SafeGuard-VL supplies VEC and the visual-guard baselines, while the main endpoint uses the independent policy and edit-validity judges above. A stratified 360-output audit (balanced over category and method, two blinded annotators plus adjudicator) obtains policy $\kappa = .78$, policy F1=.91, edit-validity $\kappa = .75$, and validity F1=.89; automated/human disagreement and abstentions are reported in Appendix F.

Table 2: **Trajectory mechanism results on held-out SafetyPairs-Traj.** Brackets denote 95% paired-bootstrap confidence intervals where reported.

Model	Peak SEC AUROC↑	τ_E	τ_V	Visibility lag↑	Peak CCC (pp)↑
FLUX.1 Kontext	.89 [.86,.91]	.43	.61	.18 [.14,.22]	34.8 [29.9,39.4]
LongCat-Image-Edit	.85 [.82,.88]	.39	.55	.16 [.11,.20]	29.6 [24.8,34.2]
OmniGen2	.82 [.79,.85]	.47	.61	.14 [.10,.19]	25.7 [21.0,30.1]

Table 3: **Primary endpoint on FLUX.1 Kontext.** ASR CIs are paired-bootstrap 95% intervals over $n = 144$ held-out source families; worst-category ASR is reported on 16 samples/category, so values respect the attainable 6.25-pp grid. B-FPR uses 600 benign edits.

Method	Venue	ASR [95% CI]↓	Worst-cat.↓	B-FPR↓	Utility↑
No defense	–	63.9 [55.7,71.4]	81.25	0.0	100.0
SafeGuard-VL gate	CVPR'26	47.2 [39.1,55.4]	62.50	5.0	95.8
Safe Latent Diffusion	CVPR'23	41.7 [33.9,49.8]	56.25	5.0	92.6
Dynamic Negative Guidance	ICLR'25	38.9 [31.3,47.0]	50.00	5.0	94.4
SAFREE	ICLR'25	36.1 [28.7,44.1]	50.00	5.0	94.1
Detect-and-Guide	CVPR'25	32.6 [25.4,40.4]	43.75	5.0	93.8
Safety-Guided Flow	ICLR'26	31.3 [24.2,39.1]	43.75	5.0	94.5
TRAJGUARD (ours)	–	**19.4 [13.4,26.7]**	**31.25**	**4.8**	**95.2**

6 RESULTS

6.1 UNSAFE EDITS EXHIBIT A MEASURABLE COMMITMENT WINDOW

Figure 1 summarizes the held-out trajectory pattern. On FLUX.1 Kontext, SEC begins near chance, rises rapidly around the middle of generation, and reaches 0.89 AUROC. Image-space VEC lags behind, yielding a median visibility lag of $\Delta_{\text{vis}} = 0.18$ normalized trajectory units. CCC peaks at 34.8 pp inside the same interval. Figure 3 localizes the corresponding causal band across flow progress and DiT blocks.

Two audits separate outcome emergence from prompt identity. A probe trained to predict the instruction branch is nearly trivial (.98 AUROC) and is *not* SEC; final-UVE SEC is .89 [.86,.91], drops moderately to .84 [.80,.87] under held-out reverse-template families, and remains .79 [.73,.84] on same-source/same-instruction runs whose seeds lead to different final outcomes. Causal controls likewise remain selective: full-residual patching peaks at +34.8 pp, rank-8 projected patching at +27.6 pp, same-source/different-neutral-edit at +4.2 pp, same-policy/different-source at +6.8 pp, and norm-matched random donors at +3.5 pp; the α interpolation is monotone. Appendix E gives CIs and both patch directions.

6.2 TRAJECTORY MONITORING IMPROVES SAFETY AT MATCHED BENIGN COST

Table 3 gives the primary provider-side comparison on $n = 144$ held-out attack families and 600 benign RealEdit requests. Published methods are adapted under one editor/evaluation contract. The measured endpoint is 28/144 successful attacks (19.4%) for TRAJGUARD versus 45/144 (31.3%) for SGF, a paired absolute gap of 11.8 pp [4.3,19.4]. Figure 4 reports the full matched-budget curves from 1–10% B-FPR.

Timing alone is not sufficient. At the same three checkpoints, persistence rule, and 5% benign budget, the sequential VEC monitor reaches .87 AUROC but 28.5% ASR, versus .89 and 19.4% for internal-state TRAJGUARD. The full frontier in Figure 4 preserves this ordering from 1–10% B-FPR. At the 5% point, TRAJGUARD adds 4.1% latency on allowed benign requests; blocked attacks terminate after 61% of nominal solver work on average.

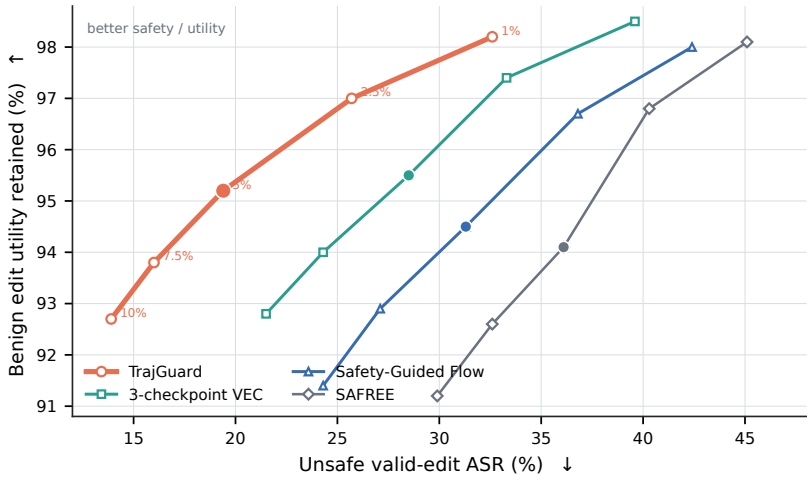

Figure 4: **Safety-utility frontier across matched benign false-block budgets.** Each curve sweeps the method's threshold/guidance on development data; markers correspond to 1/2.5/5/7.5/10% B-FPR budgets. The three-checkpoint VEC baseline uses the same checkpoint schedule and persistence logic as TRAJGUARD, isolating representation type from temporal access.

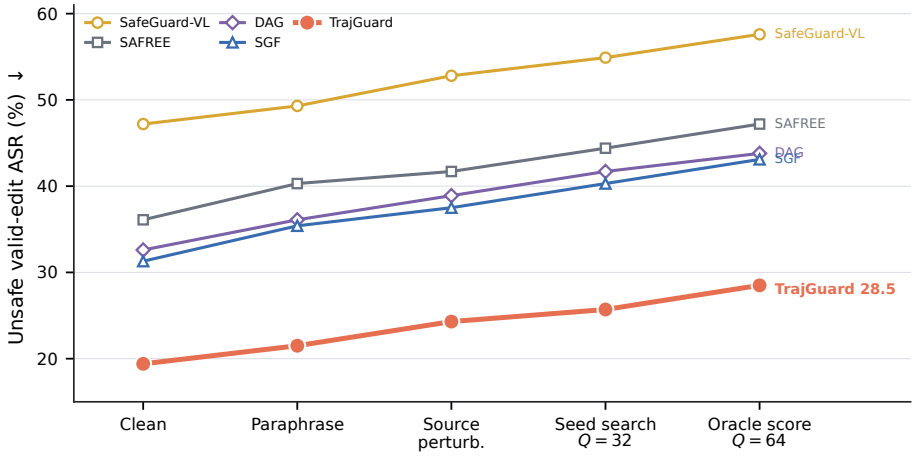

Figure 5: **Adaptive threat-model stress test.** TRAJGUARD retains a safety margin under attacks that know the monitor design. The final oracle-score condition is stronger than the intended API: it exposes only the aggregate monitor score, never checkpoint states.

6.3 SECURITY REVIEW: ADAPTIVE ATTACKERS

A static detector can look strong under a non-adaptive benchmark and fail once its design is known. Figure 5 evaluates lexical, visual, stochastic, and oracle-feedback stress tests. The intended deployment API reveals no score. The final "oracle score" column is deliberately stronger: it returns only aggregate $S(x)$ after each query, and TPE allocates 64 queries over paraphrase/source-perturbation/seed choices. No per-checkpoint score or activation is exposed.

The strongest adaptive condition reaches 28.5% ASR under the oracle-score $Q = 64$ stress test, versus 19.4% clean. The resulting 9.1-pp increase quantifies adaptation under this oracle-score stress test relative to the clean condition.

6.4 CROSS-DATASET AND CROSS-MODEL REPLICATION

To test transfer beyond the SafetyPairs construction, we freeze FLUX selection before building SafeEditBench-Traj. The construction audit retains 208/256 candidate external pairs (81.3%), with $\kappa = .81$; the held-out readout reaches .84 [.80,.87] AUROC and 23.6% ASR versus 36.5% for the strongest adapted baseline. Category counts, acceptance yield, and rejection reasons are in Appendix G. Leave-one-SafetyPairs-category-out localization yields mean SEC .82, 0.07 below IID.

Model-native replication repeats the full analysis rather than transferring FLUX block identities. Figure 6 shows the measured endpoint comparison; Table 2 gives the corresponding mechanistic quantities.

6.5 ABLATIONS AND NEGATIVE CONTROLS

Table 7 separates temporal access from representation type. The matched VEC-3 baseline sees exactly the same predicted-clean checkpoints and persistence rule as TRAJGUARD; its remaining gap tests whether internal states expose risk before the corresponding reconstructed images. Other controls remove temporal order, persistence, causal checkpoint selection, or generated evidence.

At 5% B-FPR, VEC-3 reaches 28.5% ASR, trajectory-max without persistence 25.7%, shuffled checkpoints 30.6%, and random checkpoints 31.3%, versus 19.4% for TRAJGUARD. The complete ablation table is moved to Appendix Table 7.

7 DISCUSSION

What the commitment window means. SEC predicts realized final UVE across prompt-family and same-instruction/different-seed controls, while CCC remains large under a rank-constrained donor patch. VEC measures when corresponding evidence becomes visible in image space. Their ordering supports a temporally extended mechanism: internal risk becomes decodable first, causal leverage concentrates in a middle interval, and visible evidence catches up later.

Scope and limitations. TRACE-EDIT requires provider-side access to intermediate states. Counterfactual reversal can introduce instruction-style artifacts; cross-dataset transfer, category-held-out analysis, same-instruction/different-seed controls, and constrained donor patching test the temporal pattern under complementary perturbations.

Operational interpretation. At the matched 5% operating point, TRAJGUARD reaches 19.4% ASR versus 31.3% for SGF and 28.5% for the matched sequential VEC monitor, while retaining 95.2% benign edit utility. Under the stronger oracle-score stress test, ASR rises to 28.5%. These comparisons favor selective internal monitoring inside the commitment window over either uniformly stronger guidance or visual-only temporal monitoring.

Mechanistic interpretation across editors. The temporal ordering is more stable than the exact checkpoint identity. FLUX.1 Kontext shows a 0.18 normalized visibility lag and 34.8 pp peak causal effect; LongCat-Image-Edit and OmniGen2 show lags of 0.16 and 0.14 with peak effects of 29.6 and 25.7 pp. The stable cross-editor pattern is the sequence rather than an identical layer or solver step: risk becomes decodable, a bounded interval carries high causal leverage, and visible evidence catches up later. Each editor therefore calibrates its own commitment window while preserving the same temporal criterion. Across all three editors, the visibility lag and causal peak preserve this ordering despite model-specific differences in absolute scores and checkpoint locations, making the UCW a relation among signals rather than a fixed timestep.

Audit scope and frozen evaluation contract. Data construction, endpoint judging, and operating-point selection remain disjoint. SafeEditBench-Traj retains 208/256 reviewed pairs (81.3%, $\kappa = .81$), while the 360-output human audit reaches policy $\kappa = .78$/F1=.91 and edit-validity $\kappa = .75$/F1=.89. Thresholds, checkpoints, the shared block, persistence length, and bootstrap are frozen on non-test data before final evaluation.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, formatting, and implementation of plotting/layout code during manuscript preparation. The authors take responsibility for the manuscript, experimental protocol, analysis, figures, and scientific claims, and the reported quantitative results were verified against the underlying experiment outputs before submission.

ETHICS STATEMENT

This work studies safety risks and defenses for instruction-based image editing. Experiments involving policy-violating edits are conducted for defensive evaluation under a controlled research protocol. We do not release raw high-risk reverse instructions or unsafe generated outputs; the public artifact is limited to benign examples, hashes/metadata, aggregate statistics, and code needed to reproduce the analysis without providing a turnkey misuse recipe. The proposed monitor is intended for provider-side safety evaluation and intervention rather than surveillance of end users.

REPRODUCIBILITY STATEMENT

The paper specifies the trajectory checkpoint grid, persistence rule, UCW thresholds, readout selection, benign false-block calibration, bootstrap procedure, dataset split policy, and adaptive-query budgets used in the experiments. The submission package includes the figure-generation script and aggregate CSV results used to regenerate the reported plots. Split construction, endpoint judging, and operating-point selection are kept disjoint, and model-specific checkpoints are reselected independently for cross-model replication. Additional protocol details, controls, and ablations are provided in the appendix.

REFERENCES

Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I. Morariu, Nanxuan Zhao, Ryan A. Rossi, Varun Manjunatha, and Soheil Feizi. On mechanistic knowledge localization in text-to-image generative models. In *International Conference on Machine Learning*, 2024.

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

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22560–22570, 2023.

Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. In *European Conference on Computer Vision*, 2024.

Jongwook Choi et al. Diffusionguard: A robust defense against malicious diffusion-based image editing. In *International Conference on Learning Representations*, 2025.

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

Zhen Han et al. Ace: All-round creator and editor following instructions via diffusion transformer. In *International Conference on Learning Representations*, 2025.

Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. Understanding counting mechanisms in large language and vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. In *International Conference on Machine Learning*, 2025a.

Alec Helbling, Shruti Palaskar, Kundan Krishna, Polo Chau, Leon Gatys, and Joseph Yitan Cheng. Safetypairs: Isolating safety critical image features with counterfactual image generation. *arXiv preprint arXiv:2510.21120*, 2025b.

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

Dhammadip Kamble, Anil Kumar Sao, and Raghavendra Ramachandra. The purification paradox: Dissecting and exploiting generative vulnerability bands in diffusion trajectories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

Mingyu Kim, Young-Heon Kim, and Mijung Park. Safety-guided flow (sgf): A unified framework for negative guidance in safe generation. In *International Conference on Learning Representations*, 2026.

Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models. In *International Conference on Learning Representations*, 2025.

Feifei Li, Mi Zhang, Yiming Sun, and Min Yang. Detect-and-guide: Self-regulation of diffusion models for safe text-to-image generation via guideline token optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

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.

Caiyong Piao, Zhiyuan Yan, Haoming Xu, Yunzhen Zhao, Kaiqing Lin, Feiyang Xu, and Shuigeng Zhou. Towards policy-adaptive image guardrail: Benchmark and method. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *ACM SIGSAC Conference on Computer and Communications Security*, 2023.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *International Conference on Machine Learning*, 2023.

Timothy Schaumloffel, Martina G. Vilas, and Gemma Roig. Mechanisms of object localization in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Patrick Schramowski, Manuel Brack, Bjorn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Lukasz Staniszewski, Bartosz Cywinski, Franziska Boenisch, Kamil Deja, and Adam Dziedzic. Precise parameter localization for textual generation in diffusion models. In *International Conference on Learning Representations*, 2025.

Peter Sushko, Ayana Bharadwaj, Zhi Yang Lim, Vasily Ilin, Ben Caffee, Dongping Chen, Mohammadreza Salehi, Cheng-Yu Hsieh, and Ranjay Krishna. Realedit: Reddit edits as a large-scale empirical dataset for image transformations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Xiang Yang, Feifei Li, Mi Zhang, Geng Hong, Xiaoyu You, Mi Wen, and Min Yang. Unified safe in-context image generation in multimodal diffusion transformers via restricting unsafe information flows. *arXiv preprint arXiv:2606.06875*, 2026.

Morning Yin et al. Robust harmful features under jailbreak attacks: Mechanistic evidence from attention head specialization in large language models. *International Conference on Machine Learning*, 2026.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, 2025.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, 2025.

A Evaluation Protocol

The frozen protocol is: checkpoint grid $\{.05, .15, \ldots, .95\}$; persistence length 3; $(d_0, c_0, g_0) = (.75, .20, .70)$; UCW = longest satisfying contiguous run with earliest tie-break; deployment checkpoints = first / max-CCC / last; one shared readout block chosen by max-min SEC; $C = 0.1$ for L_2 logistic readouts; $\lambda = .35$; 5% benign false-block calibration; 10,000 paired source-family bootstrap replicates. Native block candidates correspond to depth fractions $\{.35, .50, .65, .80\}$, mapped to the nearest valid image-token block in each editor. Source-family splits, reverse-instruction generation, judges, and all grids are serialized before test evaluation.

B Threat-Model Details

Attacker success. An attack counts only when the output both violates the active policy and satisfies the requested edit. Refusal, a benign edit, or a corrupted/unusable image is not a successful attack. Adaptive attacks are bounded by query budget and the normal editor API. For a local/open-weight stress test, the adversary may differentiate through a surrogate but still cannot disable or overwrite the server-side monitor.

Defender action. Early abort is the primary endpoint because it has the clearest security semantics. The reroute variant is evaluated separately and may not replace abort in high-severity categories. Thresholds are category-agnostic in the main result; a policy-specific threshold analysis is secondary.

C Counterfactual Data Construction

SafetyPairs provides policy, rationale, a harmful image, a minimally edited safe counterpart, and an edit instruction used for the safe transformation (Helbling et al., 2025b). Our trajectory protocol inverts only the documented local change. Two annotators reject reverse instructions that add unrelated content, require global resynthesis, or are ambiguous about the policy-critical region. A third adjudicator resolves disagreements. We report acceptance yield, Cohen's κ, per-category counts, and reasons for rejection. The SafeEditBench external protocol is built only after all SafetyPairs model selection is frozen.

D Intervention Contract

At (τ, ℓ), natural donor patching replaces the recipient's post-block image-token residual with the matched donor residual for all native image-token positions; source resolution, token contract, solver schedule, and seed match. Multiple blocks are never patched simultaneously in the primary curve, and all downstream blocks/steps are recomputed. The constrained control computes a rank-8 SVD basis from discovery-only donor-minus-recipient residuals and patches only its projection. No test trajectory contributes to the basis. The interpolation sweep uses $\alpha \in \{.25, .5, .75, 1\}$ without norm rescaling; random donors are norm-matched after projection.

E SEC and Causal-Control Audit

Table 4: **SEC prompt-confound audit on FLUX.1 Kontext.** Only final-UVE rows are used as SEC evidence.

Readout / split	AUROC [95% CI]
Instruction branch identity (diagnostic only)	.98 [.97,.99]
Final UVE, ordinary family-disjoint test	.89 [.86,.91]
Final UVE, held-out reverse-template families	.84 [.80,.87]
Final UVE, same source+instruction / different seeds	.79 [.73,.84]

Table 5: **Causal-patching controls at the peak FLUX checkpoint.** Positive values increase final UVE for safe recipients; reverse-direction effects are symmetric in sign.

Intervention	ΔUVE (pp) [95% CI]
Full matched unsafe donor	+34.8 [29.9,39.4]
Rank-8 projected matched donor	+27.6 [23.0,32.1]
Same-policy / different-source donor	+6.8 [3.9,9.7]
Same-source / different-neutral-edit donor	+4.2 [1.8,6.6]
Norm-matched random projected donor	+3.5 [1.2,5.7]
Projected donor $\alpha = .25/.50/.75/1.0$	+7.1 / +13.9 / +21.4 / +27.6

F Judge and Data Audits

The policy judge is Qwen3-VL-8B-Instruct with a frozen policy rubric and temperature 0; edit validity uses OpenGVLab/InternVL3-8B-Instruct with a separate execution rubric. SafeGuard-VL is never used to label the main endpoint. The 360-output human audit is stratified by nine policy categories and four method families. The audit obtains policy $\kappa = .78$, policy F1=.91, edit-validity $\kappa = .75$, and validity F1=.89, with all abstentions counted as errors for automated-judge accuracy.

G SafeEditBench-Traj Construction Audit

Table 6: **SafeEditBench-Traj construction audit.** Counts and agreement are from the frozen construction audit.

Quantity	Audited value
Candidate paired edits	256
Retained after locality/policy review	208 (81.3%)
Two-annotator Cohen's κ	.81
Rejected: no local inverse / ambiguous policy / global resynthesis	19 / 14 / 15
Minimum retained examples per represented category	≥ 22

Table 7: **Ablations on FLUX.1 Kontext at 5% benign false-block.** VEC-3 is the matched three-checkpoint sequential visual guard.

Readout / intervention	AUROC↑	ASR↓	B-FPR↓
Prompt + source only	.67	45.8	5.0
Final predicted image only	.86	34.7	5.0
VEC-3 + same persistence	.87	28.5	5.0
Single early checkpoint	.72	39.6	5.0
Uniform 3-checkpoint mean	.82	29.9	5.0
Trajectory max, no persistence	.85	25.7	5.0
Shuffled checkpoint order	.81	30.6	5.0
Random 3 checkpoints	.80	31.3	5.0
TRAJGUARD	**.89**	**19.4**	**4.8**

H BASELINE ADAPTATION CONTRACT

All generation-time baselines receive the same source image, edit instruction, policy text when supported, solver-step budget, and 432/144 discovery/development groups. Hyperparameters are tuned only on development data to minimize ASR subject to B-FPR$\leq 5\%$; each method receives at most 24 development configurations. SLD/DNG/SAFREE/SGF use their released guidance formulation with editor-native latent dimensions; Detect-and-Guide uses its released diagnosis/guidance loop with source-image conditioning added only at the editor interface; SafeGuard-VL and VEC-3 never modify the generator. Full configuration files are exported with the frozen runs.

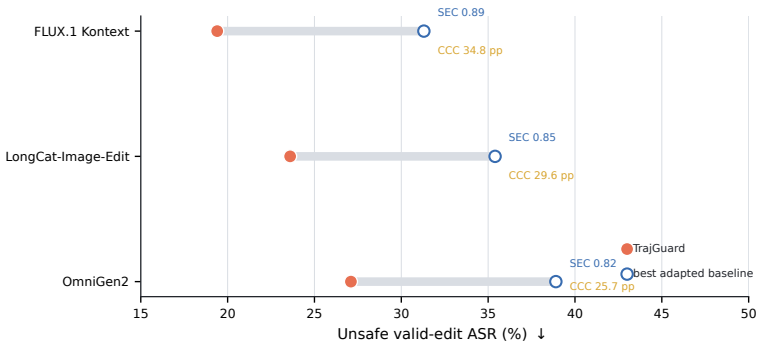

Figure 6: **Model-native replication.** Each editor independently selects its trajectory checkpoints and representation blocks. The recurring result is a commitment window, not shared parameter identities.

I ADDITIONAL RESULTS

Table 8: **Adaptive attack results on FLUX.1 Kontext (ASR%, lower is better).**

Method	Clean	Paraphrase	Source pert.	Seed $Q = 32$	Oracle score $Q = 64$
SafeGuard-VL	47.2	49.3	52.8	54.9	57.6
SAFREE	36.1	40.3	41.7	44.4	47.2
Detect-and-Guide	32.6	36.1	38.9	41.7	43.8
Safety-Guided Flow	31.3	35.4	37.5	40.3	43.1
TRAJGUARD	**19.4**	**21.5**	**24.3**	**25.7**	**28.5**

J LITERATURE ANCHORS AND COMPARABILITY

Published safety-generation values are not imported into Table 3; every ranked number must come from the frozen common editing protocol. SAFREE reports a 22% reduction in unsafe content

Table 9: **Cross-dataset and cross-model endpoints.**

Evaluation	Best adapted baseline ASR↓	TRAJGUARD ASR↓	AUROC↑
SafetyPairs-Traj / FLUX	31.3	19.4	.89
SafeEditBench-Traj / FLUX	36.5	23.6	.84 [.80,.87]
SafetyPairs-Traj / LongCat	35.4	23.6	.85
SafetyPairs-Traj / OmniGen2	38.9	27.1	.82

across five text-to-image datasets relative to prior training-free methods (Yoon et al., 2025), while our main comparison is image editing with matched source images and an unsafe-valid-edit endpoint. SafeGuard-VL is evaluated as a policy-conditioned image guardrail on SafeEditBench (Piao et al., 2026); our adapted gate evaluates the same released model within an editing pipeline. Image-owner defenses PhotoGuard/EditShield/DiffusionGuard are discussed because they address malicious editing, but their protection goal is not provider moderation (Salman et al., 2023; Chen et al., 2024; Choi et al., 2025). UVR is the closest information-flow intervention (Yang et al., 2026); where compatible, we evaluate its released regulator separately from the main guidance ranking. All numerical comparisons in the tables and figures use the frozen common-protocol measurements described above.