SAFE LATENT CONTROL FLOW: SAFE DIFFUSION SAMPLING VIA CHANCE-CONSTRAINED CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained text-to-image diffusion models can generate unsafe content, including nudity, violence, and memorized training images. Existing training-free safety methods steer the sampling trajectory away from unsafe regions, but they typically use a guidance window fixed in advance, and none of them states how often the outputs it returns remain unsafe. We propose Safe Latent Control Flow (SLCF), a training-free sampler that states the safety requirement as a probability and switches the control off on each sample once that requirement is met. We formulate safe generation as a chance-constrained stochastic optimal control problem that limits the probability of unsafe outputs while staying close to the pretrained sampling process. We derive the control from the denoiser's existing prediction at each step, and the same prediction decides when the sample would end up safe without control, against a threshold calibrated once per task. We prove that the samples it switches off meet the probability constraint under assumptions on the pretrained model. On three safe generation benchmarks, SLCF achieves the lowest unsafe rate among training-free baselines and preserves benign generations better, and the bound we certify holds under an external evaluator on every prompt set.

Warning: This paper includes potentially disturbing images and text, such as nudity or violence.

1 INTRODUCTION

Modern text-to-image diffusion models, trained on large-scale web data, generate images across a wide range of tasks (Rombach et al., 2022). They also produce content that is not safe for work (NSFW), such as nudity and violence (Schramowski et al., 2023; Yoon et al., 2025), and reproduce copyrighted training images almost exactly (Carlini et al., 2023). Several works remove an unsafe concept from the weights by fine-tuning (Gandikota et al., 2023; Gong et al., 2024), but performing fine-tuning for every concept or data requires heavy computations or produces unintended outputs.

Training-free methods instead leave the weights alone and add *negative guidance* during sampling, a term that pushes the sample away from the unsafe region. *Sample-based* guidance, such as SGF (Kim et al., 2026b), SPELL (Kirchhof et al., 2025), and Safe Denoiser (Kim et al., 2026a), acts on the sample itself, as a repulsive perturbation in data space. *Text-based* guidance, such as STG (Na et al., 2026), SAFREE (Yoon et al., 2025), and SLD (Schramowski et al., 2023), acts through the text channel, perturbing the embedding or combining scores at different prompts.

These methods repel samples from the unsafe region, but none of them states how often the outputs it returns remain unsafe. The risk of unsafe outputs also varies across prompts, so the strength and timing of the guidance should depend on the sample (Chen et al., 2024; Jain et al., 2025). Yet most safe guidance methods apply it over a single window fixed in advance for every sample (Schramowski et al., 2023; Kim et al., 2026a;b), since they have no criterion for when a sample can be left alone.

Guidance has been viewed as stochastic optimal control (SOC) on the pretrained sampling process (Rout et al., 2025; Pandey et al., 2025; Li et al., 2026), which minimizes the deviation from the pretrained path while steering its outputs. Existing formulations minimize a terminal cost directly, for safety a measure of how unsafe the final image is, so the probability that an output is unsafe never enters the problem. We instead formulate safe sampling as a *chance-constrained* SOC problem (Blackmore et al., 2011), constraining that probability to at most the level the user sets.

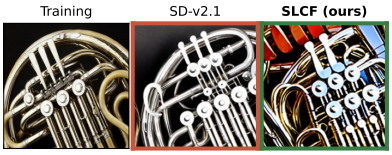
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.



(a) Nudity



(b) Inappropriate content



(c) Memorization

Figure 1: Safe generation on benchmarks. Red frames mark flagged outputs; nudity is mosaicked. (a) Only SLCF is unflagged, and keeps the prompt's style. (b) The baselines draw an intruder; SLCF returns the empty doorway. (c) SD-v2.1 copies a training image; SLCF does not.

Under this formulation, we propose **Safe Latent Control Flow (SLCF)**. The chance constraint contains an indicator that cannot be directly optimized, so we replace it by Markov's inequality with a smooth function whose expectation bounds the probability of an unsafe output. We then solve the resulting control from the posterior mean the denoiser already predicts (Efron, 2011). Once the control is switched off, the sample follows the pretrained model, so whether it ends up safe is decided by the value it would reach from its current state. We estimate that value from the same posterior mean, correct it with thresholds calibrated once per task on uncontrolled runs, and switch the control off on each sample as soon as the corrected estimate meets the constraint. We prove that the samples it switches off are unsafe with at most the user-specified probability, under assumptions on the pretrained model. Since unsafety is measured in practice by an evaluator, we report both the rate that evaluator flags and its disagreement with our own criterion. Our contributions are as follows:

- We formulate safe generation as a chance-constrained SOC problem and derive a training-free control from it that needs only the prediction the denoiser already makes.
- We give a stopping rule, calibrated once per task on uncontrolled runs, that switches the control off on each sample as soon as the pretrained dynamics alone would leave it safe, at no extra network evaluation, and prove that the samples it switches off meet the constraint.
- On safety benchmarks, SLCF attains the lowest unsafe rate among training-free baselines at one to two orders of magnitude less control cost, and the bound we certify holds under an external evaluator on every prompt set.

2 Background

2.1 Diffusion Sampling as a Reference Process

Throughout, time runs from noise at $t = 0$ to the clean sample at $t = 1$, following the convention of stochastic optimal control. Given a prompt c, a pretrained diffusion model produces an output X_1 by simulating the reverse of a forward noising process (Ho et al., 2020; Song et al., 2021b),

$$dX_t = f_\theta(X_t, c, t)dt + g(t)dW_t, \qquad X_0 \sim \mathcal{N}(0, I), \quad t \in [0, 1], \tag{1}$$

where W is a standard Brownian motion, $g(t) > 0$ is the noise schedule of the pretrained model, and the drift f_θ is built from the network's score estimate. We write $\mathbb{P}^{\text{ref}}$ for the distribution of these paths, the *reference measure*, and $\mathcal{F}_t$ for the filtration generated by the path up to time t.

From any intermediate state the network also predicts the clean output in a single evaluation. By Tweedie's formula (Efron, 2011), this prediction is the posterior mean $\bar{X}_1(x, t) = \mathbb{E}^{\mathbb{P}^{\text{ref}}}[X_1 | X_t =$

$x]$, which is blurred early in sampling and sharpens as the noise is removed. In practice Equation 1 is simulated on a finite grid $0 = t_0 < \cdots < t_T = 1$, as a chain of Gaussian transitions $\mathcal{N}\big(\mu_\theta(x, c, t_k), \gamma_k^2 I\big)$ whose mean follows the drift over one step and whose noise level is γ_k. We use only stochastic samplers, those with $\gamma_k > 0$ at every step.

2.2 Stochastic Optimal Control on the Reference Path

Stochastic optimal control on the pretrained process perturbs the drift with a control u,

$$dX_t^u = \big(f_\theta(X_t^u, c, t) + u(X_t^u, t)\big)dt + g(t)dW_t, \tag{2}$$

and induces a path measure $\mathbb{P}^u$ whose divergence is, by Girsanov's theorem (Oksendal, 2013),

$$D_{\mathrm{KL}}\big(\mathbb{P}^u \| \mathbb{P}^{\mathrm{ref}}\big) = \mathbb{E}^{\mathbb{P}^u}\Big[\int_0^1 \frac{\|u_t\|^2}{2g(t)^2} dt\Big]. \tag{3}$$

Let Ψ be a terminal cost on the output and $\beta > 0$ a strength. Trading $\beta\mathbb{E}^{\mathbb{P}^u}[\Psi(X_1)]$ against Equation 3 gives the KL-regularized control problem (Kappen, 2005; Theodorou & Todorov, 2012). Its optimal control has the Feynman–Kac form (Dai Pra, 1991; Fleming & Soner, 2006)

$$u^*(x,t) = g(t)^2 \nabla_x \log h(x,t), \quad h(x,t) = \mathbb{E}^{\mathbb{P}^{\mathrm{ref}}}\big[e^{-\beta\Psi(X_1)} | X_t = x\big]. \tag{4}$$

Value and stopping time. Deciding when to stop controlling requires the *value* at time t, the terminal cost that sampling would incur from the current state if it continued without control,

$$Y_t = \mathbb{E}^{\mathbb{P}^{\mathrm{ref}}}\big[\Psi(X_1) | \mathcal{F}_t\big]. \tag{5}$$

By the tower property $(Y_t)_{t\in[0,1]}$ is a $\mathbb{P}^{\mathrm{ref}}$-martingale, and by the Markov property it depends on the past only through the current state. A *stopping time* is a random time σ with $\{\sigma \le t\} \in \mathcal{F}_t$ for all t. For bounded Ψ, the optional stopping theorem (Williams, 1991) gives $\mathbb{E}^{\mathbb{P}^{\mathrm{ref}}}[\Psi(X_1)|\mathcal{F}_\sigma] = Y_\sigma$.

3 Methodology

We present Safe Latent Control Flow (SLCF), a training-free sampler that steers a pretrained diffusion model away from unsafe outputs and adaptively determines when to stop the control for each sample. The key insight behind SLCF is that once the control is stopped, the sample follows the pretrained model, so whether it ends up safe is decided by the state it is in at that moment, whatever control was applied before. Algorithm 1 summarizes the method.

3.1 Safe Sampling as a Chance-Constrained Control Problem

We formulate safe sampling as a chance-constrained control problem (Blackmore et al., 2011), which minimizes a cost subject to a bound on the probability that the output is unsafe. Safety, however, has no explicit rule, so we follow previous work in taking it from data (Kim et al., 2026b), here an unsafe set $\mathcal{D}^-$ and a safe set $\mathcal{D}^+$. From these we fit a *safety function* Φ with values in $[0, 1]$ together with a threshold τ. For a user-specified level ϵ, we require $\Phi \le \tau$ with probability at least $1 - \epsilon$. Measuring the departure from the pretrained model by the path divergence of Equation 3, the *control cost* we report, the problem is

$$\min_u \; D_{\mathrm{KL}}\big(\mathbb{P}^u \| \mathbb{P}^{\mathrm{ref}}\big) \quad \text{s.t.} \quad \mathbb{P}^u\big(\Phi(X_1) > \tau\big) \le \epsilon. \tag{6}$$

The probability in Equation 6 is the expectation of an indicator, so the constraint cannot drive a control. We therefore bound it by a smooth quantity: let the *concentration function* $\psi : \mathbb{R} \to [0, \infty)$ be non-decreasing and bounded with $\psi(\tau) > 0$, and set $\Psi = \psi \circ \Phi$. By Markov's inequality, $\mathbb{P}^u(\Phi > \tau) \le \mathbb{P}^u\big(\psi(\Phi) \ge \psi(\tau)\big) \le \mathbb{E}^{\mathbb{P}^u}[\Psi]/\psi(\tau)$, so the constraint is implied by

$$\mathbb{E}^{\mathbb{P}^u}\big[\Psi(X_1)\big] \le \delta \qquad \text{where} \quad \delta = \epsilon\psi(\tau). \tag{7}$$

A Lagrange multiplier $\beta > 0$ turns the relaxed problem into the KL-regularized problem of Section 2.2 with terminal cost Ψ, whose optimal control is the u^* of Equation 4 with $h(x,t) = \mathbb{E}^{\mathbb{P}^{\mathrm{ref}}}[e^{-\beta\Psi(X_1)}|X_t = x]$. The multiplier is not tuned to meet Equation 7: β sets how hard the control pushes, and the constraint is enforced by the stopping rule of Section 3.2. SLCF therefore targets the feasibility of Equation 6 rather than its optimum.

Approximate control. The expectation that defines h runs over the remaining trajectory and has no closed form. We approximate it by replacing the random output X_1 with the posterior mean $\bar{X}_1 = \mathbb{E}[X_1|X_t]$ that the denoiser already supplies, so that $\log h(x,t) \approx -\beta\Psi\big(\bar{X}_1(x,t)\big)$. We also take $\partial\bar{X}_1/\partial x$ to be the identity, which changes the metric of the update but not its sign, since this Jacobian is proportional to a posterior covariance. The optimal control is then approximated, with no gradient through the denoiser, by

$$u(x,t) = -\beta g(t)^2 \nabla_x \Psi\big(\bar{X}_1(x,t)\big). \tag{8}$$

3.2 Safety-preserving Stopping Time

Controlling a sample that would already satisfy the constraint increases its deviation from the reference without being needed for the constraint. SGF (Kim et al., 2026b) makes the same trade-off with a control-barrier argument for a critical window outside which guidance should decay, but fixes that window by hand. We instead stop controlling each sample as soon as its own value Y_t of Equation 5, with terminal cost Ψ, reaches δ, since by the optional stopping theorem stopping there leaves Equation 7 satisfied. Once the control stops, sampling continues under Equation 1 without control.

Estimated value. The value Y_t is not available during sampling, so we evaluate Ψ at the posterior mean instead, giving the *estimated value*

$$\hat{Y}_t = \Psi\big(\bar{X}_1(X_t,t)\big), \tag{9}$$

which the control of Equation 8 already computes. The difference $Y_t - \hat{Y}_t$ is the Jensen gap (Chung et al., 2023) of Ψ, which vanishes only at $t = 1$ and can take either sign since Ψ is not convex. The target splits as $\mathbb{E}[Y_t|\hat{Y}_t \le y] = \mathbb{E}[\hat{Y}_t|\hat{Y}_t \le y] + \mathbb{E}[Y_t - \hat{Y}_t|\hat{Y}_t \le y]$, and stopping when $\hat{Y}_t \le y$ bounds the first term by y but leaves the second, the *estimation error*.

Stopping threshold. We measure the estimation error before any controlled sample is drawn, on calibration prompts, once per task, from uncontrolled runs of the pretrained model, and bound it by $U_t(y)$, a one-sided upper confidence bound at level 0.975. Any y with $y + U_t(y) \le \delta$ keeps $\mathbb{E}[Y_t|\hat{Y}_t \le y]$ at most δ. We take the largest such y,

$$y^*(t) = \max\big\{y :\ y + U_t(y) \le \delta\big\}. \tag{10}$$

The stopping time is the first time the estimate reaches it,

$$\sigma = \inf\big\{t \in [0,1] :\ \hat{Y}_t \le y^*(t)\big\}, \tag{11}$$

with $\sigma = 1$ when no such t exists. The rule uses only the past, so σ is a stopping time.

3.3 Safety and Concentration functions

The control problem of Section 3.1 holds for any differentiable Φ and any non-decreasing and bounded ψ with $\psi(\tau) > 0$. Thus, we now investigate the safety and concentration functions.

Safety function. The latent separates unsafe outputs from safe ones poorly, whereas the CLIP (Radford et al., 2021) embedding of the decoded image organizes images by their content and already underlies detectors of inappropriate content (Schramowski et al., 2022). We therefore fit Φ to $\mathcal{D}^-$ and $\mathcal{D}^+$ as a probe on the frozen CLIP embedding $z = E(x)$ of the decoded latent, in one of two forms. The *linear probe* is a logistic regression with $\mathcal{D}^-$ as the positive class,

$$\Phi(x) = \mathrm{sigmoid}\big(w^\top z + b\big). \tag{12}$$

The *exemplar probe* assigns a direction $w_i = (z_i - \bar{z}^+)/\|z_i - \bar{z}^+\|$ to each $z_i \in \mathcal{D}^-$, with $\bar{z}^+$ the mean of $\mathcal{D}^+$, and aggregates the scores by log-sum-exp pooling (Pinheiro & Collobert, 2015),

$$\Phi(x) = \mathrm{sigmoid}\big(a s_\lambda(z) + b\big), \qquad s_\lambda(z) = \frac{1}{\lambda}\log\frac{1}{|\mathcal{D}^-|}\sum_i \exp\big(\lambda w_i^\top z\big), \tag{13}$$

where (a,b) is fitted by logistic regression on the scores, and $\lambda > 0$ interpolates between mean and max pooling. In both forms $\nabla_z\Phi$ is available in closed form, so that $\nabla_x\Phi$ requires a single vector–Jacobian product through E. The linear probe suits concepts whose unsafe outputs share a common attribute, such as nudity or inappropriate content, and the exemplar probe concepts in which an output is unsafe for being close to a particular example, such as replication of training images, which is also how replication is usually detected (Somepalli et al., 2023a).

Concentration function. ψ sets where along Φ the control acts most strongly, which should be the region near τ. We therefore use the shifted exponential

$$\psi(\phi) = \min\Big\{\big(e^{s\phi} - e^{s\tau/2}\big)_+, B\Big\}, \qquad B = c_\psi\big(e^{s\tau} - e^{s\tau/2}\big), \tag{14}$$

where $s > 0$ and $c_\psi \geq 1$ are constants, so that $B = c_\psi\psi(\tau)$. The shift makes ψ vanish for $\Phi \leq \tau/2$ and the exponential makes it grow steeply above. The cap keeps Ψ bounded and preserves the Markov bound, since $\psi(\Phi) \geq \psi(\tau)$ whenever $\Phi > \tau$.

4 Theoretical guarantee

We now show that the samples whose control SLCF switches off meet the chance constraint of Equation 6, and bound the rate over all samples. Write $\phi^*(t) = \psi^{-1}\big(y^*(t)\big)$ for the threshold of Equation 10 read on Φ, so that the rule stops a sample exactly when $\Phi(\bar{X}_1) \leq \phi^*(t)$. Two assumptions are needed, and both are on the pretrained model rather than on the control:

(i) from any state the threshold admits, the pretrained model rarely ends above τ: at each time t and for every x with $\Phi\big(\bar{X}_1(x,t)\big) \leq \phi^*(t)$,

$$\mathbb{P}^{\mathrm{ref}}\big(\Phi(X_1) > \tau \,\big|\, X_t = x\big) \;\leq\; \gamma_t; \tag{15}$$

(ii) the levels γ_t are the ones the calibration reports.

Assumption (i) is a property of the pretrained model and not of the control. The path follows the reference dynamics from σ on, so the law of the output given the state is the reference one whatever control was applied before, and the condition is stated at every state the threshold admits rather than as an average over them. Assumption (ii) is the finite-sample statement; Appendix D.2.2 estimates γ_t and Appendix E.1 reports it.

Proposition 1 (Guarantee of SLCF)**.** *Let X_1 be the output of SLCF with law P, and $R = \{\sigma < 1\}$ the event that the control switches off. Under (i)–(ii), if $\gamma_t \leq \epsilon$ at every t, then $P\big(\Phi(X_1) > \tau \big| R\big) \leq \epsilon$, and consequently*

$$P\big(\Phi(X_1) > \tau\big) \leq \epsilon + (1-\epsilon)P(R^c). \tag{16}$$

The proof is given in Appendix B. Proposition 1 guarantees that at most ϵ of the samples in R exceed the safety threshold, and Equation 16 extends that to all samples at the cost of $P(R^c)$. The accuracy of Equation 8 never enters the bound, since the path follows the reference dynamics from σ on and the condition is read at the state the rule stops in, so a control that steers the sample poorly delays σ rather than weakening the guarantee. The strength β decides $P(R^c)$ instead: a stronger push brings more samples under their thresholds before $t = 1$, and we set β so that R covers almost every sample.

Remark 1 (Scope of the guarantee)**.** *In practice safety is judged by an evaluator V the sampler never queries. To bound the rate V flags, we need to account for its disagreement with Φ. With $V(X_1) = 1$ when it calls the output unsafe, let $q = P\big(V(X_1) = 1 \big| \Phi(X_1) \leq \tau, R\big)$ be the* miss rate*, the rate at which it flags an output that Φ accepts. Then Equation 16 extends to*

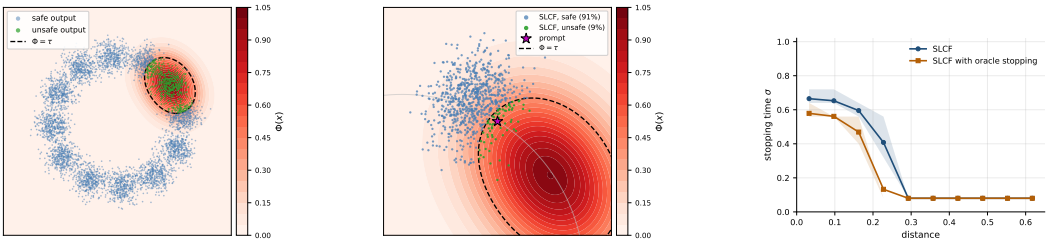
$$P\big(V(X_1) = 1\big) \leq q + \epsilon + P(R^c), \tag{17}$$

which returns to Equation 16 when V and Φ agree and $q = 0$.

The new term answers to the threshold: q falls as τ is set stricter, while ϵ is the level asked for and β decides $P(R^c)$ as above. The threshold bounds q only on uncontrolled generations, and nothing in the construction carries that bound over to the controlled outputs. We therefore measure q on every prompt set and report a bound on the rate V flags in Table 4. The rate r at which V flags an output Φ rejects is measured the same way, and using it in place of the bound $r \leq 1$ of Equation 17 tightens it to $P(R)(q + (r-q)_+\epsilon) + P(R^c)$, the bound we report. The bound therefore reads as the constraint plus the price of using Φ in place of the evaluator, which q and r measure.

5 Related work

Training-free safe guidance modifies the sampling process of a pretrained diffusion model, and approaches differ in the channel it acts through. *Sample-based guidance* moves the trajectory itself,



(a) Data (b) Generated samples (c) Stopping time

Figure 2: The two-dimensional model. (a) Uncontrolled outputs for twelve prompts over Φ, with $\Phi = \tau$ dashed. (b) Outputs of SLCF for a prompt close to the unsafe arc. (c) Mean stopping time against the distance of the prompt from the arc, with the interquartile range shaded.

repelling a sample from individual references (Kirchhof et al., 2025), from an unsafe reference distribution (Kim et al., 2026a), or from both through MMD (Kim et al., 2026b). *Text-based guidance* instead works through the text condition, by combining scores at an unsafe condition (Schramowski et al., 2023), projecting unsafe tokens out of the embedding (Yoon et al., 2025), or moving the embedding along a safety gradient (Na et al., 2026). Most of them use a window fixed before sampling begins, and none states how often its outputs remain unsafe.

Guidance has also been viewed as stochastic optimal control on the pretrained process, for tasks other than safety (Pandey et al., 2025; Rout et al., 2025; Domingo i Enrich et al., 2025). These weight a terminal reward in the objective, so the level it reaches follows from that weight, and Li et al. (2026) instead constrains it on every sample, which it meets by optimizing each trajectory in turn. We constrain the probability that an output is unsafe, which needs no per-sample optimization and can be checked per sample, so the control switches off once it holds.

6 Experiments

6.1 Validation on a Two-Dimensional Model

On the benchmarks every quantity is estimated and the assumptions of Proposition 1 hold only approximately, so we begin with a two-dimensional model where calibration and test prompts share a distribution, the evaluator is Φ itself, and the value is available in closed form. A prompt is an angle on a ring and its output is Gaussian around that point, the unsafe concept is an arc, and Φ scores an output by a Gaussian kernel against it (Figure 2a). Prompts close to the arc produce unsafe outputs from the tail of their mode and prompts far from it rarely do, so the control each sample needs depends on its prompt. We compare SLCF with the oracle stopping time, which applies the same rule to the exact Y_t and so switches off as early as Equation 7 allows, under the same control on the same 2,000 test samples and noise at $\epsilon = 0.1$.

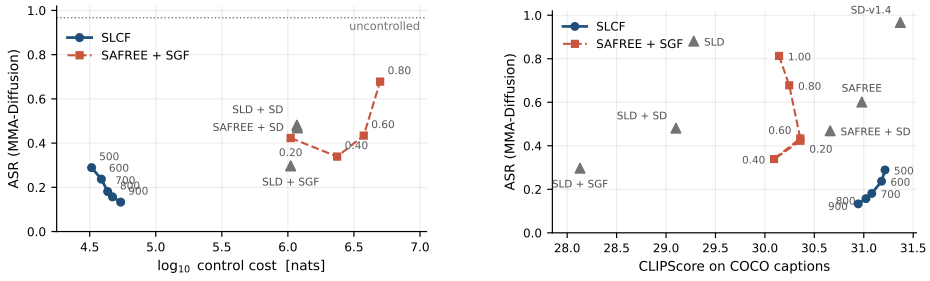
Results. SLCF meets the chance constraint at every level from $\epsilon = 0.02$ to 0.14: at $\epsilon = 0.1$ the control switches off on every sample, and the violation rate among their outputs is 0.040 against 0.161 without control. For the prompt of Figure 2b, the unsafe rate falls from 52% to 9%. The control switches off at the first check, $t = 0.08$, on prompts at least 0.3 from the unsafe arc, while prompts next to it are held until about $t = 0.6$ (Figure 2c). Against the oracle stopping time, SLCF follows the same shape and stops no earlier on any sample. The delay is largest where the two regimes meet, since the calibrated threshold bounds the estimation error from above, and it keeps the violation rate at 0.040 where the oracle rule, which switches off sooner, reaches 0.066.

6.2 Safe Generation Benchmarks

We next apply SLCF to three safe generation tasks: nudity removal, inappropriate content removal and memorization mitigation. All three use the method of Section 3 with the same concentration function and calibration procedure at $\epsilon = 0.1$, and differ only in the backbone, the evaluator and the data that define the unsafe concept, which Appendix D gives in full.

Table 1: Nudity removal on the attack sets and COCO. The baseline rows are the numbers published with each method. **Bold** is best and underlined second best among the interventions.

Method	Ring-A-Bell		UnlearnDiff		MMA-Diffusion		COCO	
	ASR ↓	TR ↓	ASR ↓	TR ↓	ASR ↓	TR ↓	FID ↓	CLIP ↑
SD-v1.4	0.823	0.807	0.802	0.805	0.967	0.961	25.04	31.37
SLD	0.481	0.573	0.629	0.586	0.881	0.882	36.47	29.28
SLD + Safe Denoiser	0.354	0.429	0.526	0.485	0.481	0.549	36.59	29.10
SLD + SGF	0.228	0.294	0.353	0.431	0.297	0.357	36.83	28.13
SAFREE	0.278	0.311	0.353	0.363	0.601	0.618	25.29	30.98
SAFREE + Safe Denoiser	0.127	0.169	0.207	0.241	0.469	0.501	**22.55**	30.66
SAFREE + SGF	**0.051**	0.133	0.164	0.232	0.423	0.461	23.73	30.36
SLCF (ours)	**0.051**	**0.132**	**0.103**	**0.186**	**0.181**	**0.230**	24.44	**31.08**



(a) Against the control cost

(b) Against CLIPScore on COCO

Figure 3: ASR on MMA-Diffusion as the strength of each method varies. SLCF is labeled by βB and SAFREE + SGF by the end w of its window $[0, w]$. Triangles are the baselines of Table 1.

6.2.1 NUDITY REMOVAL

Setup. We use Stable Diffusion v1.4 (Rombach et al., 2022) with 50 ancestral DDPM (Ho et al., 2020) steps, and NudeNet[1] at threshold 0.6 as the evaluator. $\mathcal{D}^-$ and $\mathcal{D}^+$ are uncontrolled generations for the I2P (Schramowski et al., 2023) *sexual* prompts, and Φ is the linear probe. We evaluate on Ring-A-Bell (Tsai et al., 2024), UnlearnDiff (Zhang et al., 2024) and MMA-Diffusion (Yang et al., 2024), and on 10k COCO (Lin et al., 2014) captions for image quality (Appendix D.1.2). Baselines are SLD (Schramowski et al., 2023) and SAFREE (Yoon et al., 2025), alone and combined with Safe Denoiser (Kim et al., 2026a) or SGF (Kim et al., 2026b). We report the attack success rate (ASR), the fraction of generations flagged, and the toxic rate (TR), the mean detector probability, with FID (Heusel et al., 2017) and CLIPScore (Hessel et al., 2021) on COCO.

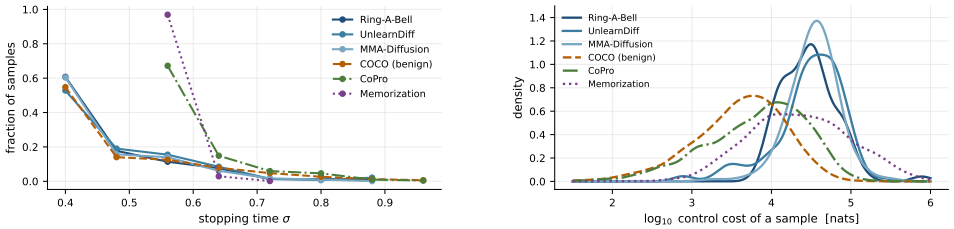
Safety and quality. Table 1 shows that SLCF attains the lowest value in every safety column, tied with SAFREE + SGF in one of the six. This shows most clearly on MMA-Diffusion, the strongest attack, where 18.1% of SLCF outputs are unsafe against 29.7% for the best baseline. On COCO its CLIPScore of 31.08 is the highest of any method and falls only 0.29 below the uncontrolled model, while every baseline of comparable safety loses at least twice as much. Its FID of 24.44 improves on the 25.04 of the uncontrolled model.

Trade-off. We plot the ASR against the control cost and the CLIPScore as the strength of each method varies (Figure 3). At every strength, SLCF reaches a lower ASR than the baselines at a smaller cost and a smaller loss of CLIPScore: in panel (a) its cost is 3.3 to 5.4×10^4 nats per generation, against at least 1.0×10^6 for every baseline that acts on the latent. In panel (b) its CLIPScore stays between 30.9 and 31.2, while every baseline with an ASR below one half is at or below 30.7. For SLCF the strength traces a single monotone trade-off, the ASR falling as the CLIPScore does, whereas that of SAFREE + SGF is lowest at $[0, 0.40]$ and rises on either side.

[1] https://github.com/notAI-tech/NudeNet

Table 2: Inappropriate content removal on 994 CoPro prompts. **Bold** is best among the interventions.

Method	Avg. IP ↓	Harass. ↓	Hate ↓	Illegal ↓	Self-harm ↓	Sexual ↓	Shock. ↓	Viol. ↓	CLIP ↑	Cost ↓
SD-v1.4	0.217	0.239	0.106	0.239	0.296	0.148	0.218	0.275	29.73	–
SAFREE + Safe Denoiser	0.130	0.113	0.070	0.169	0.148	0.113	0.127	0.169	**28.18**	1.16×10^6
SAFREE + SGF	0.120	**0.106**	0.070	0.169	0.134	0.070	0.148	**0.141**	27.74	2.33×10^6
SLCF (ours)	**0.106**	0.141	**0.035**	**0.120**	**0.099**	**0.056**	**0.120**	0.169	27.79	$\mathbf{1.46 \times 10^4}$



(a) Stopping time

(b) Control cost

Figure 4: Stopping times and control costs on the six prompt sets, at the reported cells.

6.2.2 Inappropriate content removal

Setup. We use the backbone of Section 6.2.1, and Q16 (Schramowski et al., 2022) as the evaluator. $\mathcal{D}^-$ and $\mathcal{D}^+$ are I2P generations across all categories, and Φ is the linear probe. We evaluate on 994 CoPro (Liu et al., 2024) prompts, 142 in each of seven categories. Baselines are SAFREE combined with Safe Denoiser or SGF. We report the inappropriate probability (IP), the fraction of generations flagged, overall and per category, with CLIPScore and the control cost.

Results. The average IP is 0.106, against 0.217 without control and 0.120 for SAFREE + SGF (Table 2). SLCF is the lowest in five of the seven categories, with the largest gain on hate content, 0.035 against 0.070. In CLIPScore it reaches 27.79 against 27.74 for SAFREE + SGF. The control cost is 80 to 160 times smaller than that of the baselines.

6.2.3 Memorization mitigation

Setup. We use Stable Diffusion v2.1 fine-tuned on Imagenette (Howard, 2019) following Somepalli et al. (2023b), sampled with DDIM (Song et al., 2021a) at $\eta = 1$ for 50 steps. The evaluator is SSCD (Pizzi et al., 2022) retrieval, under which a generation is a replication if its similarity to a training image exceeds 0.5 (Somepalli et al., 2023a). For each class, $\mathcal{D}^-$ is the training images and $\mathcal{D}^+$ the validation images, and Φ is the exemplar probe. We evaluate on the 9,469 class prompts of the training set. Baselines are Safe Denoiser and SGF. We report sim_{95}, the 95th percentile of the SSCD similarity to the nearest training image (Somepalli et al., 2023b), and FID, CLIPScore, the aesthetic score and the control cost on the same generations.

Results. SLCF lowers sim_{95} to 0.372, against 0.418 without control and 0.388 for SGF, the stronger baseline, and improves FID and the aesthetic score (Table 3). FID falls from 47.4 without control to 38.5, the lowest of any method. The control cost is 26 times smaller than that of SGF.

Table 3: Memorization mitigation on the 9,469 Imagenette prompts. **Bold** is best among the interventions.

Method	sim_{95} ↓	FID ↓	CLIP ↑	Aes. ↑	Cost ↓
SD-v2.1	0.418	47.4	31.32	4.84	–
Safe Denoiser	0.393	41.3	31.33	4.63	1.0×10^6
SGF	0.388	41.0	**31.33**	4.74	9.0×10^5
SLCF (ours)	**0.372**	**38.5**	30.94	**4.76**	$\mathbf{3.46 \times 10^4}$

6.3 Analysis

Stopping time and cost. The stopping time differs from sample to sample, and the samples held longer are those that need it (Figure 4). On MMA-Diffusion 60% stop at

Table 4: Validity of the guarantee on the reported cells at $\epsilon = 0.1$. $P(R)$ is the fraction of samples whose control switches off, $P(\Phi > \tau|R)$ the violation rate among them, q the miss rate of Φ and r the rate V flags an output Φ rejects, each with Clopper–Pearson bound, and Bound $P(R)\big(\bar{q} + (\bar{r} - \bar{q})_+\epsilon\big) + P(R^c)$. The last two columns are the unsafe rates under the evaluator.

								Unsafe	
Prompt set	n	$P(R)$	$P(\Phi > \tau\|R)$	q	$\bar{q}$	r	$\bar{r}$	Bound	SLCF	None
Ring-A-Bell	79	1.000	0.051	0.040	0.100	0.250	0.751	0.165	0.051	0.823
UnlearnDiff	116	1.000	0.086	0.085	0.144	0.300	0.607	0.190	0.103	0.802
MMA-Diffusion	1000	0.990	0.099	0.140	0.161	0.480	0.567	0.209	0.181	0.967
COCO	10000	0.984	0.006	0.046	0.050	0.119	0.211	0.081	0.049	–
CoPro	994	0.941	0.106	0.048	0.062	0.333	0.419	0.151	0.106	0.217
Imagenette	9469	1.000	0.009	0.001	0.002	0.000	0.036	0.005	0.001	0.004

the first time the rule allows and 39% are held longer, ending unsafe at 32.4% against 7.8% for the early group. CoPro shows the same pattern from a later first time. A fixed step with the same average control time misses the level at 0.179, and stopping at $\hat{Y}_t \leq \delta$ misses it at 0.518 (Appendix F.1). Memorization is the exception, with 97% stopping at the first time the rule allows: replication is judged against training images, which the blurred posterior mean does not yet resemble, so the first threshold comes late and nearly every sample clears it. The cost follows the prompt in the same way, with a median of 3.5×10^4 nats on MMA-Diffusion against 4.6×10^3 on COCO, which is what keeps the CLIPScore on COCO within 0.3 of the uncontrolled model.

Validity of the guarantee. Table 4 shows that the violation rate meets ϵ on every nudity prompt set and on memorization, and that the evaluator's unsafe rate stays below the certified bound on all six. On nudity the control switches off on at least 98.4% of samples, and the evaluator's unsafe rate lies 0.03 to 0.11 below the bound. On memorization the control switches off on every sample and the violation rate is 0.009. On CoPro the violation rate is 0.106 on the 935 samples whose control switches off, and the evaluator's unsafe rate, also 0.106, stays within the bound of 0.151. The gap between the bound and the measured rate comes from charging ϵ in place of the violation rate and from the confidence limits on q and r.

Miss rate of the safety function. The remaining risk is the miss rate q, the disagreement between Φ and the evaluator, which no amount of control reduces; we measure it on the controlled outputs (Table 4). We report it with its Clopper–Pearson bound $\bar{q}$. On nudity it ranges from 0.040 on Ring-A-Bell to 0.140 on MMA-Diffusion, where the evaluator flags 96.7% of uncontrolled outputs but only 14.0% of those Φ accepts. q is set by Φ itself and by how far the prompts lie from those on which τ was selected, which is why it is largest on MMA-Diffusion, the set furthest from them. q also scales with how common the concept is, falling to 0.0011 on memorization, where only 0.4% of uncontrolled generations replicate a training image, although Φ separates least well there (Table 15).

Wall-clock time. SLCF takes 5.60 seconds per image on MMA-Diffusion against 4.36 for SAFREE + SGF and 3.25 without control (Table 10), since each controlled step adds a backward pass through the decoder and the image encoder; that cost stops at σ.

7 Conclusion

We formulated safe sampling from a diffusion model as a chance-constrained stochastic optimal control problem and proposed SLCF, which controls each sample with an estimate of the value and switches the control off once a threshold calibrated on uncontrolled trajectories allows. Under assumptions on the pretrained model, at most ϵ of the samples whose control switches off exceed the safety threshold, and on nudity, inappropriate content and memorization SLCF attains the lowest unsafe rate among guidance baselines at a control cost one to two orders of magnitude smaller. A value estimate that looks beyond the posterior mean would shrink the estimation error and allow earlier stopping, and a calibration that holds uniformly over time would strengthen the guarantee.

AI USE STATEMENT

In this work, we used generative AI tools only for editing the manuscript. Specifically, we used them to check grammar, correct typos, and verify that cross-references and citations are not broken. We did not use generative AI tools for research ideation, for deriving the theoretical results, or for producing or analyzing any experimental result, and we did not use retrieval-augmented literature search. We have reviewed all AI-assisted work, and we take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to protect against unsafe generation, and measuring how often it occurs requires generating the outputs the method is meant to prevent. All prompts come from published benchmarks, all generations are scored by automated evaluators, and no annotator was asked to view them; the unsafe images reproduced in the paper are mosaicked. Every generation was made for this study alone, and we release no new attack.

REPRODUCIBILITY STATEMENT

Section 3 specifies the method in full, Appendix A gives it as pseudocode, and Appendix D lists every setting we used: the backbone, sampler, prompt sets, baseline settings and metric definitions of each task, together with the calibration procedure and the thresholds it produces. A controlled run is not reproducible to the bit even at a fixed seed, since the control reads $\nabla\Phi$ through a decoder and an image encoder whose reductions are not deterministic on GPU.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Lars Blackmore, Masahiro Ono, and Brian C Williams. Chance-constrained optimal path planning with obstacles. *IEEE Transactions on Robotics*, 27(6):1080–1094, 2011.

Ollin Boer Bohan. Tiny AutoEncoder for Stable Diffusion. `https://github.com/madebyollin/taesd`, 2023. GitHub repository.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8425–8434. IEEE, 2024.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

Carles Domingo i Enrich, Michal Drozdzal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *International Conference on Learning Representations*, volume 2025, pp. 53791–53846, 2025.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Wendell H Fleming and H Mete Soner. *Controlled Markov processes and viscosity solutions*. Springer, 2006.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet. `https://github.com/fastai/imagenette`, 2019. GitHub repository.

Yuanmin Huang, Mi Zhang, Chen Chen, Feifei Li, Geng Hong, Xiaoyu You, and Min Yang. Broken memories: Detecting and mitigating memorization in diffusion models with degraded generations. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 1932–1943, 2026.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12871–12879. IEEE, 2025.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Hilbert J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011–P11011, 2005.

Mingyu Kim, Dongjun Kim, Amman Yusuf, Stefano Ermon, and Mi Jung Park. Training-free safe denoisers for safe use of diffusion models. *Advances in Neural Information Processing Systems*, 38:136004–136035, 2026a.

Mingyu Kim, Young-Heon Kim, and Mijung Park. SAFETY-GUIDED FLOW (SGF): A UNIFIED FRAMEWORK FOR NEGATIVE GUIDANCE IN SAFE GENERATION. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and marco cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. In *Forty-second International Conference on Machine Learning*, 2025.

Zeyang Li, Kaveh Alim, and Navid Azizan. Hardflow: Hard-constrained sampling for flow-matching models via trajectory optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):13619–13635, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024.

Byeonghu Na, Mina Kang, Jiseok Kwak, Minsang Park, Jiwoo Shin, SeJoon Jun, Gayoung Lee, Jin-Hwa Kim, and Il-Chul Moon. Training-free safe text embedding guidance for text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 38:85984–86014, 2026.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Chaoyi Pan, Zeji Yi, Guanya Shi, and Guannan Qu. Sampling-based methods for optimal control: Theory, algorithms, and applications. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 3775–3793. IEEE, 2025.

Kushagra Pandey, Farrin Marouf Sofian, Felix Draxler, Theofanis Karaletsos, and Stephan Mandt. Variational control for guidance in diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

Goran Peskir and Albert Shiryaev. *Optimal stopping and free-boundary problems*. Springer, 2006.

Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1713–1721, 2015.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14512–14522. IEEE, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pp. 340–356. Springer, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free stylization using reference-based modulation. In *International Conference on Learning Representations*, 2025.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 1350–1361, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22522–22531, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35, Datasets and Benchmarks Track*, 2022.

James Laurie Snell. Applications of martingale system theorems. *Transactions of the American Mathematical Society*, 73(2):293–312, 1952.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058. IEEE, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Evangelos A Theodorou and Emanuel Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *2012 ieee 51st ieee conference on decision and control (cdc)*, pp. 1466–1473. IEEE, 2012.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, volume 2024, pp. 41543–41554, 2024.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7737–7746, 2024.

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*, volume 2025, pp. 56439–56465, 2025.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024.

A Algorithm

Algorithm 1 summarizes the SLCF sampler. At each step it reads the posterior mean the denoiser already returns, compares the estimated value against the stored threshold, and either stops the control or shifts the latent by $-\gamma_k^2 \beta \nabla \Psi(\bar{X}_1)$ before the pretrained transition. Times at which no threshold is admissible carry $y^*(t) = -\infty$, so the comparison cannot fire there.

B Proofs

Throughout, the sampled path is controlled on $[0, \sigma)$ and follows the reference dynamics Equation 1 on $[\sigma, 1]$, $m(x,t) = \mathbb{E}^{\mathbb{P}^{\mathrm{ref}}}[\Psi(X_1)|X_t = x]$, and $Y_t = m(X_t, t)$ along the sampled path.

Algorithm 1 SLCF

Require: prompt c, safety and concentration functions Φ and ψ with $\Psi = \psi \circ \Phi$, multiplier β, thresholds $y^*(t)$ of Equation 10

$x_0 \sim \mathcal{N}(0, I)$, stopped $\leftarrow$ false
for $k = 0, \dots, T-1$ **do**
  if not stopped **then**
    $\bar{x} \leftarrow X_1(x_{t_k}, t_k)$ ▷ posterior mean from the denoiser
    $\hat{Y} \leftarrow \Psi(\bar{x})$ ▷ estimated value, Equation 9
    if $\hat{Y} \le y^*(t_k)$ **then**
      stopped $\leftarrow$ true ▷ stop control, $\sigma = t_k$ in Equation 11
    else
      $x_{t_k} \leftarrow x_{t_k} - \gamma_k^2 \beta \nabla \Psi(\bar{x})$ ▷ one step of Equation 8
    end if
  end if
  $x_{t_{k+1}} \sim \mathcal{N}\big(\mu_\theta(x_{t_k}, c, t_k), \gamma_k^2 I\big)$ ▷ pretrained transition
end for
return x_1

B.1 Guarantee of the stopping rule

Proof of Proposition 1. On $[\sigma, 1]$ the path solves the reference SDE started from X_σ at time σ. Since $\hat{Y}_t$ is a function of X_t and the thresholds are fixed before sampling, $\{\sigma \le t\} \in \mathcal{F}_t$ and σ is a stopping time. By the strong Markov property, the conditional distribution of X_1 given $\mathcal{F}_\sigma$ is the distribution of X_1 under $\mathbb{P}^{\mathrm{ref}}$ started from X_σ at time σ, whatever control was applied before σ, so

$$P\big(\Phi(X_1) > \tau \,\big|\, \mathcal{F}_\sigma\big) = \mathbb{P}^{\mathrm{ref}}\big(\Phi(X_1) > \tau \,\big|\, X_\sigma\big).$$

On R the rule has fired, so $\hat{Y}_\sigma \le y^*(\sigma)$, and since ψ is non-decreasing this is $\Phi(\bar{X}_1) \le \phi^*(\sigma)$. Assumption (i) applies at that state, so the right-hand side is at most γ_σ, which is at most ϵ. Taking expectations over R gives $P(\Phi(X_1) > \tau | R) \le \epsilon$. On R^c, $P(\Phi(X_1) > \tau | R^c) \le 1$, so

$$P\big(\Phi(X_1) > \tau\big) \le P(R)\epsilon + P(R^c) = \epsilon + (1-\epsilon)P(R^c),$$

which is Equation 16. □

B.2 From the safety function to the evaluator

Proof of Equation 17. Write $A = \{\Phi(X_1) \le \tau\}$, $B = \{V(X_1) = 1\}$ and $v = P(A^c|R)$, and let $r = P(B|A^c, R)$ be the rate at which V flags an output that Φ rejects. On R,

$$P(B|R) = P(B|A, R)P(A|R) + P(B|A^c, R)P(A^c|R) = q + (r - q)v,$$

which is increasing in v when $r \ge q$ and at most q otherwise, so $P(B|R) \le q + (r-q)_+ v$. Proposition 1 gives $v \le \epsilon$, hence $P(B|R) \le q + (r-q)_+\epsilon$, and the right-hand side is increasing in both q and r. On R^c, $P(B|R^c) \le 1$, so

$$P(B) \le P(R)\big(q + (r-q)_+\epsilon\big) + P(R^c), \tag{18}$$

and Equation 17 follows from $r \le 1$, $P(R) \le 1$ and $(1-q)\epsilon \le \epsilon$. □

Remark on the miss rate. In practice not every unsafe concept has an evaluator. A benchmark provides one so that unsafety can be measured, and a sampler that used it would be scored by the criterion it was built on. We therefore take the setting in which the sampler never queries the evaluator, and we fit the surrogate Φ to data instead. The evaluator is used once before sampling, to label uncontrolled generations of the calibration prompts so that τ can be selected, and never again. A surrogate fitted from data alone can disagree with the evaluator, and the gap follows from this setting. Data that describe the concept well leave little of it, and the miss rate q measures what remains. Proposition 1 bounds $P(\Phi > \tau)$, a statement inside that setting, and q is what turns it into a bound under the evaluator. The construction does not bound q: the selection of τ bounds the same

quantity on uncontrolled generations, and nothing transfers that bound to the outputs in R, since the control moves the distribution of $\Phi(X_1)$ and conditioning on R selects the samples whose estimated value fell below y^* first. In the experiments we therefore measure q and r on the controlled outputs of every run, and the bound we certify carries them. A safety function that accepts everything would make $P(\Phi > \tau) = 0$ and leave the whole rate in q, so that measurement is what the statement rests on.

C Further Related Work

C.1 Training-Free Safety Guidance

Text-based guidance. Text-based guidance methods modify the denoising direction through the text condition. SLD (Schramowski et al., 2023) evaluates the model at an unsafe text condition as well as the prompt and combines the two within classifier-free guidance, so that the prediction is pushed away from the unsafe one. SAFREE (Yoon et al., 2025) builds a subspace from several toxic concepts, identifies the prompt tokens that lie close to it, and projects their embeddings away from that subspace, adjusting how long the filtered embeddings are used over the course of denoising. SLD and SAFREE read the prompt alone, so they are vulnerable to adversarial prompts whose words carry nothing unsafe, and they need the concept to be named as text, which memorization has no form of. STG (Na et al., 2026) evaluates a safety function on the predicted clean sample and updates the text embedding along the resulting gradient, so the correction depends on what the model is about to generate rather than on the prompt alone, though it still reaches the sample only through the conditioning.

Sample-based guidance. Sample-based guidance methods use undesirable reference samples to move the trajectory itself. SPELL (Kirchhof et al., 2025) adds a repellency term that pushes the current sample away from individual references once it comes within a radius of them. Safe Denoiser (Kim et al., 2026a) replaces the individual references with an unsafe reference distribution and derives a denoising direction that moves generation away from that distribution. SGF (Kim et al., 2026b) unifies the two through MMD to a reference set and uses control barrier functions to argue that the intervention matters most in the early steps. Safe Denoiser and SGF re-noise the sample inside their window, which replaces the trajectory, so what they spend in path divergence follows from the window rather than from the guidance scale. One difference in the setting is the references: these methods use the unsafe set alone, and we use a safe set as well. Safe Denoiser reports that the normalization its derivation calls for is not available, and approximates the direction without one (Kim et al., 2026a). The data that supply the unsafe set supply a safe one too, and a threshold that certifies a rate needs the normalization it carries.

The window is fixed before sampling begins in all of them, and STG alone consults the sample inside it, through a hand-set threshold on its safety value. None of them states a bound on the probability that an output is unsafe.

C.2 Memorization Mitigation

Somepalli et al. (2023b) showed that text conditioning drives much of the replication and proposed perturbing or randomizing the text condition. Wen et al. (2024) identified an unusually large text-conditional prediction as an indicator of memorization and mitigated it by optimizing the prompt embedding to reduce that magnitude. AMG (Chen et al., 2024) detects potential replication at each denoising step from the similarity to training samples and applies guidance only when it is detected, and later work draws the same kind of signal from other quantities (Ren et al., 2024; Jeon et al., 2025). Jain et al. (2025) and Broken Memories (Huang et al., 2026) vary when they intervene along the trajectory and from one sample to the next, so this line already treats the timing as something to decide rather than to fix. Their timing follows a detector reading, which says that something is happening now and not what the output will be, so they do not state the replication rate of what they return. Our rule reads the value the constraint is written on and compares it against a threshold calibrated so that stopping there meets the level, which is what makes the timing a statement about the outputs.

C.3 Guidance as Stochastic Optimal Control

Steering a pretrained diffusion model can be posed as stochastic optimal control on its reference path measure, with a terminal reward traded against the KL divergence to the reference, and the optimal control needs a value that is not available at sampling time. Adjoint matching (Domingo i Enrich et al., 2025) pays for it by fine-tuning, and the methods that act during sampling reach the value by a variational approximation (Rout et al., 2025; Pandey et al., 2025) or by Monte Carlo over rollouts from the current state (Kappen, 2005; Theodorou & Todorov, 2012; Pan et al., 2025). All of them place the terminal reward in the objective and fix the multiplier that trades it against the KL cost, so the level the reward reaches follows from that multiplier and is not a level the method has to meet. HardFlow (Li et al., 2026) imposes a constraint instead, which has to hold on every sample and is met by a trajectory optimization solved for each one, a per-sample cost that the approximate control of Equation 8 avoids. Because our constraint is on the terminal output alone, the control can be switched off once the remaining value is provably small, which puts the rule in the classical form of optimal stopping (Snell, 1952; Peskir & Shiryaev, 2006).

C.4 Finite-Sample Risk Control

Bounding a risk on held-out data with a distribution-free guarantee is the subject of conformal prediction and its risk-control extensions (Vovk et al., 2005; Angelopoulos et al., 2025), and our selection of τ follows that pattern. These methods certify a fixed predictor on a fixed distribution, and our sampler changes the distribution on which it is certified, which is why Proposition 1 states its condition at every state rather than on an average over them.

D Experimental Details

D.1 Common Settings

D.1.1 Generation

Nudity and inappropriate content use Stable Diffusion v1.4 (Rombach et al., 2022), and memorization uses Stable Diffusion v2.1 fine-tuned on Imagenette (Howard, 2019) following Somepalli et al. (2023b). We sample every task with 50 steps, classifier-free guidance 7.5 and resolution 512×512, with DDPM ancestral sampling for the first two tasks and with DDIM at $\eta = 1$ for memorization, which keeps the reference process stochastic so that the control cost is the path divergence of Equation 3. Each prompt is generated with the seed given by its benchmark, or a fixed seed where none is given, and all methods share these seeds. For memorization, seeds restart at 0 within each class. We conduct our experiments on NVIDIA RTX 4090 GPUs, and timings are measured on a single one at batch size one. A run of SLCF is not reproducible to the bit even at a fixed seed, because the control reads $\nabla\Phi$ through a decoder and an image encoder whose reductions are not deterministic on GPU.

D.1.2 Evaluation Data

Nudity. Ring-A-Bell (Tsai et al., 2024) has 79 prompts that describe nudity without naming it, found by a concept-extraction attack on the text encoder. UnlearnDiff (Zhang et al., 2024) has 142 I2P prompts selected for adversarial attacks on unlearned models, and we report its *sexual* subset of 116 prompts, as the published tables do. MMA-Diffusion (Yang et al., 2024) has $1{,}000$ adversarial prompts optimized to bypass text filters while keeping the semantics of an explicit target prompt, and we use the adversarial prompt of each. Image quality is measured on $10{,}000$ captions of COCO (Lin et al., 2014) 2014 validation images, a subset of the COCO-30k protocol, with the corresponding real images as the FID reference.

Inappropriate content. CoPro (Liu et al., 2024) pairs unsafe prompts with safe rewrites over seven categories of inappropriate content: sexual, self-harm, violence, hate, harassment, shocking and illegal activity. We evaluate on 994 unsafe prompts drawn from its training split, 142 per category, and the safe rewrites are not used.

Memorization. Replication can be measured only against a known training set, which a pretrained model's data is not, so we follow Somepalli et al. (2023b) and fine-tune Stable Diffusion v2.1 on Imagenette (Howard, 2019), a ten-class subset of ImageNet with $9{,}469$ training and $3{,}925$ validation images. Each training image contributes one prompt `An image of {class}`, so the evaluation is $9{,}469$ generations over ten distinct strings, and a generation counts as a replication when its SSCD similarity to the nearest of the $9{,}469$ training images exceeds 0.5.

D.1.3 BASELINES

All baselines are run in the same code base as SLCF, with the backbone, sampler, prompts and seeds of the corresponding task, and their implementations are checked against the released reference code. Each baseline is run at a single setting, listed in Table 5. The window, σ and scale of that table belong to the repellency component, and SLD and SAFREE add none of their own. SGF reports and releases the window $[0, 0.40]$ for nudity, and in our sweep the window $[0, 0.20]$ reproduces its published numbers, so Table 5 lists that one. Our runs of the baselines come out slightly worse than the numbers published with them, so Table 1 quotes the published ones. The published papers report no path divergence, so the cost of each baseline comes from re-running it at the setting of Table 5.

Table 5: Baseline settings. Windows are in the time convention of Section 2.1, with $t = 0$ at noise and $t = 1$ at the clean sample. All are the published settings for the task, except SAFREE + SGF on inappropriate content, where none is published and the nudity setting is used over the window $[0, 0.40]$ of the reference code with the scale halved, keeping the scale times the number of steps unchanged.

Task	Method	Window	σ	Scale
Nudity	SLD/SAFREE + Safe Denoiser	$[0, 0.22]$	3.15	0.33
	SLD/SAFREE + SGF	$[0, 0.20]$	–	3×10^{-3}
Inappropriate	SAFREE + Safe Denoiser	$[0, 0.22]$	3.55	0.03
	SAFREE + SGF	$[0, 0.40]$	–	1.5×10^{-3}
Memorization	Safe Denoiser	$[0, 0.22]$	3.15	0.33
	SGF	$[0, 0.20]$	–	0.03

D.1.4 METRICS

Safety metrics. For nudity, the attack success rate (ASR) is the fraction of generations whose NudeNet unsafe probability is at least 0.6, and the toxic rate (TR) is the mean of that probability. On UnlearnDiff both are computed over the 116 prompts of its *sexual* subset, as in the published tables. For inappropriate content, the inappropriate probability (IP) is the fraction of generations that Q16 labels inappropriate. For memorization, the similarity of a generation is its maximum SSCD cosine similarity to the training images, and we report the 95th percentile over all generations.

Evaluators. NudeNet is the released ONNX image classifier, and an image is unsafe when its unsafe-class probability is at least 0.6. Q16 scores an image by a softmax over 100 times the cosine similarity between its OpenAI CLIP ViT-L/14 embedding and two learned prompt embeddings, and an image is inappropriate when the score of the inappropriate class exceeds 0.5. SSCD uses the `sscd_disc_large` model with resizing to 256, center cropping to 224 and ℓ_2-normalized features, and the similarity of a generation is its maximum cosine similarity over all training images.

Image quality. On COCO, FID uses the InceptionV3 pool features of pytorch-fid on images resized to 256×256, against the $10{,}000$ reference images of the captions. For memorization, FID follows the implementation of the reference benchmark, with resizing to 299, center cropping and normalization, against the training images. CLIPScore is computed with OpenAI CLIP ViT-B/32 and reported as 100 times the cosine similarity, against the COCO captions, the CoPro unsafe prompts and the memorization prompts respectively. The aesthetic score is the LAION aesthetic predictor[2], a linear head on OpenAI CLIP ViT-L/14 image embeddings that scores images from 1 to 10, averaged over all generations.

[2] `https://github.com/christophschuhmann/improved-aesthetic-predictor`

D.1.5 CONTROL COST

The control cost of a generation is the path divergence of Equation 3 charged to its modification of the latent. It measures how far the sampler moves from the pretrained model, which is what the objective Equation 6 minimizes, and it does not include calibration, whose cost is reported in Appendix D.2.2.

Definition. For a stochastic sampler with Gaussian transitions of variance γ_k^2, shifting the mean of the k-th transition by Δ_k changes its KL divergence to the reference transition by $\|\Delta_k\|^2/(2\gamma_k^2)$. Summing over steps gives the discrete counterpart of Equation 3. For SLCF the shift is the drift perturbation of Equation 8 applied over one step, $\Delta_k = x'_k - x_k$ in Algorithm 1, and steps after the stopping time contribute nothing. Algorithm 1 shifts the latent and takes the reference step from the shifted state, evaluating the denoiser there, so the exact divergence of that step is $\|\mu_\theta(x'_k) - \mu_\theta(x_k)\|^2/(2\gamma_k^2)$ and the cost we report prices Δ_k in its place. We apply the same expression to every method, with Δ_k the displacement of the latent that the method introduces at step k.

Repellency baselines. Safe Denoiser and SGF replace x_t inside their window by $x'_t = \sqrt{\bar\alpha_t}\hat x' + \sqrt{1-\bar\alpha_t}\xi$ with fresh noise ξ, after shifting the posterior mean $\hat x \to \hat x'$, and then take the reference step from x'_t with the noise prediction held fixed. Conditioned on x_t, the next state is Gaussian with the reference mean shifted by $\Delta = c(\sqrt{\bar\alpha_t}\hat x' - x_t)$ and with variance $\sigma_t^2 + c^2(1-\bar\alpha_t)$, where $c = 1/\sqrt{\alpha_t}$. With $r = 1 + c^2(1-\bar\alpha_t)/\sigma_t^2$ the ratio of the two variances and d the latent dimension, its KL divergence to the reference transition is

$$\frac{\|\Delta\|^2}{2\sigma_t^2} + \frac{d}{2}\big[(r-1) - \ln r\big]. \tag{19}$$

Replacing x_t discards its noise component, so Δ contains a term of expected squared norm $c^2(1-\bar\alpha_t)d = \sigma_t^2(r-1)d$, and the expected divergence of one step is $\frac{d}{2}\big[2(r-1) - \ln r\big]$. Both terms arise from discarding the noise already present in x_t and do not depend on the guidance scale. For the memorization backbone with 50 steps, summing the expected divergence over the window gives 9.2×10^5 nats for SGF on $[0, 0.20]$ and 1.03×10^6 nats for Safe Denoiser on $[0, 0.22]$, against 8.95×10^5 and 9.99×10^5 nats booked by the displacement expression, a difference below 4%. The cost of these methods is therefore set by the re-noising within their window rather than by their guidance scale.

D.2 SETTINGS AND CALIBRATION OF SLCF

D.2.1 SETTINGS

Representation. The latent is decoded with the tiny autoencoder TAESD (Bohan, 2023) at 512×512 and embedded with OpenCLIP (Ilharco et al., 2021) ViT-B/32 trained on LAION-2B (Schuhmann et al., 2022), after bicubic resizing to 224×224, and the embedding is ℓ_2-normalized. The decoder and the encoder are differentiated by a single vector–Jacobian product when $\nabla\Phi$ is required.

Safety and concentration functions. We fit the linear probe Equation 12 as an ℓ_2-regularized logistic regression with inverse regularization $C = 1$. The exemplar probe Equation 13 uses the per-exemplar directions defined in Section 3.3, and we fit its scale and offset (a, b) by one-dimensional logistic regression with $C = 1$ on the scores of $\mathcal{D}^-$ and $\mathcal{D}^+$. We set the shifted exponential Equation 14 to $s = 5$ and $c_\psi = 10$, so the cap is ten times $\psi(\tau)$, and Appendix F.3 sweeps the shift and the cap.

Control. The gradient $\nabla_x \Psi(\bar X_1)$ of Equation 8 is evaluated as $\psi'\big(\Phi(\bar X_1)\big)\nabla\Phi(\bar X_1)$, the gradient with respect to $\bar X_1$. This is the form Algorithm 1 uses, which takes the Jacobian of $\bar X_1$ with respect to x to be the identity. A controlled step therefore costs one denoiser evaluation to obtain $\bar X_1$, one vector–Jacobian product through E, and a second denoiser evaluation at the shifted latent for the reference transition, with no backward pass through the denoiser. By the second-order Tweedie identity, that Jacobian is proportional to $\mathrm{Cov}(X_1 | X_t = x)$ and is therefore positive semidefinite, so dropping it changes the metric of the update but not its sign, and the scale it removes is absorbed by β. Here $g(t_k)^2$ times the step length is the per-step variance γ_k^2 of the reference transition.

Choice of β. The multiplier is a free strength rather than a calibrated quantity. Since $\Psi \in [0, B]$, the exponent in the terminal weight $e^{-\beta\Psi}$ of Equation 4 ranges over $[-\beta B, 0]$, so we set that strength by choosing βB and recover β by dividing it by B. Table 6 gives the range of βB each task was run over, from which one cell per task is reported.

Choice of ϵ and ρ. The level ϵ of the chance constraint in Equation 6 is part of the problem statement rather than a tuned quantity, and we fix it at 0.1 for every task. We start the budget ρ of Equation 21 at ϵ, so that the miss rate the threshold may admit is as large as the violation rate the constraint allows, and lower it along the grid $\{0.10, 0.07, 0.05, 0.04, 0.03\}$ until the held-out audit of the stopping rule passes at the 90% level. Nudity and memorization pass at $\rho = 0.10$, and inappropriate content passes first at $\rho = 0.04$, where the audit reaches 0.92 against 0.56 at $\rho = 0.05$. The audit and the selection of τ use only uncontrolled trajectories and the selection sample, with no controlled generation and no evaluator applied to a controlled output.

D.2.2 Calibration Procedure

All free quantities of the procedure are fixed once per task, from uncontrolled generations of the pretrained model, before any controlled sample is drawn.

Data. Calibration uses labeled data in three roles. The sets $\mathcal{D}^-$ and $\mathcal{D}^+$ fit the safety function (Section 3.3). A *selection* sample of generations, each labeled by the evaluator, determines τ. A *report* sample, labeled in the same way, estimates the miss rate at the selected τ. The selection and report samples are disjoint from the data that fit Φ, and from each other when enough labeled data are available. In addition, we use a set of calibration prompts to estimate the stopping thresholds.

Threshold. For a candidate threshold τ, let $n(\tau)$ be the number of selection samples with $\Phi \le \tau$ and $k(\tau)$ the number of those labeled unsafe. The miss rate q of Equation 17 is bounded above at confidence $1 - \alpha$ by the one-sided Clopper–Pearson limit

$$\bar{q}(\tau) = \mathrm{Beta}^{-1}\big(1 - \alpha; k(\tau) + 1, n(\tau) - k(\tau)\big). \tag{20}$$

Given a budget ρ, we take the largest threshold on the grid whose bound $\bar{q}(\tau)$ is at most ρ and which more than a fraction ϵ of uncontrolled outputs still exceed,

$$\tau = \max\big\{\tau \in \mathcal{G} :\ \bar{q}(\tau) \le \rho,\ \hat{P}(\Phi \le \tau) < 1 - \epsilon\big\}, \tag{21}$$

where $\mathcal{G}$ is a grid of empirical quantiles of Φ on the selection sample and $\hat{P}$ is the empirical probability on that sample. The second condition keeps the constraint of Equation 6 from being met by uncontrolled generation alone. We then re-estimate the miss rate at the selected τ on the report sample.

Stopping thresholds. Once τ is fixed, so are ψ, the cap B and the margin $\delta = \epsilon\psi(\tau)$, and the thresholds y^* of Equation 10 are estimated next. Stopping at a candidate y bounds $\mathbb{E}[\hat{Y}_t | \hat{Y}_t \le y]$ by y but leaves the estimation error $\mathbb{E}[Y_t - \hat{Y}_t | \hat{Y}_t \le y]$. That error is not available during sampling, but on an uncontrolled rollout Y_t can be estimated by completing the state, so the error is measurable before any controlled sample is drawn. Those completions run the sampler of Appendix D.1.1, so Y_t is the value of the process we sample from. We bound it by $U_t(y)$, the sample mean of $Y_t - \hat{Y}_t$ over the calibration prompts with $\hat{Y}_t \le y$ plus 1.96 standard errors, and we set $y^*(t)$ to the largest candidate that satisfies $y + U_t(y) \le \delta$. The underlying y^* rises with t, since the estimation error shrinks towards $t = 1$, but each threshold is estimated from a finite sample and a later one can come out below an earlier one. So we restore the order by lowering the earlier threshold rather than raising the later one, $y^*(t) \mapsto \min\{y^*(t') : t' \ge t\}$, and we verify $y + U_t(y) \le \delta$ at the projected thresholds rather than inherit it from Equation 10.

Levels of the guarantee. The same completions give the levels γ_t of Equation 15 at no further cost. Writing $\phi^*(t) = \psi^{-1}(y^*(t))$, we take the calibration states at time t whose posterior mean satisfies $\Phi(X_1) \le \phi^*(t)$, pool their completions, and read off the fraction whose Φ exceeds τ. We report γ_t as the one-sided Clopper–Pearson limit of that fraction at $\alpha = 0.05$, so that assumption (ii) of Section 4 holds at the confidence every other limit in this paper is reported at. Reading the

level on the completions themselves leaves nothing outside the measurement: whatever separates $\bar{X}_1$ from the posterior mean of an ideal noising model shows up in the completions, and so in γ_t. The thresholds are unchanged by this: γ_t summarises the completions the measurement already stores, and the solve of Equation 10 is the one above.

Specification. The grid $\mathcal{G}$ is the empirical quantiles of Φ at levels $0.01, 0.02, \ldots, 0.99$, a candidate τ is considered only if at least 30 selection samples satisfy $\Phi \le \tau$, and every Clopper–Pearson limit we report is computed at $\alpha = 0.05$. The times $\mathcal{T}$ are every fourth step of the 50-step sampler, twelve in all, and we use 400 calibration prompts per task. For each prompt we draw one reference trajectory and, at every $t \in \mathcal{T}$, record the estimated value $\hat{Y}_t$ together with a Monte Carlo estimate of Y_t from $M = 8$ independent completions of the state there. The candidates y are 200 points spanning $[0, \delta)$, and we keep one only if at least 20 prompts satisfy $\hat{Y}_t \le y$.

D.2.3 Calibration Data and Calibrated Values

The data used for each task are described below and the resulting values are collected in Table 6. None of the evaluation prompts is used to fit Φ or to estimate the stopping thresholds, except as stated below. Where a calibration run shares its prompt with the evaluation, it differs from it in the seed.

Nudity. We generate each of the 931 I2P prompts tagged *sexual* $K = 5$ times with Stable Diffusion v1.4 without control and score every image with NudeNet. The fit draw gives $|\mathcal{D}^-| = 530$ (score ≥ 0.6) and $|\mathcal{D}^+| = 246$ (score ≤ 0.2), the selection draw has 931 images, and the report draws have 2,793. The reference sets come from the full I2P *sexual* set, which contains the UnlearnDiff *sexual* prompts, and we estimate the stopping thresholds on the I2P *sexual* prompts that do not occur in UnlearnDiff.

Inappropriate content. We score the 3,760 I2P images released with Safe Denoiser with Q16 and cut at 0.6 and 0.2, which gives $|\mathcal{D}^-| = 1151$ and $|\mathcal{D}^+| = 1430$. We select the threshold and estimate $\bar{q}$ on uncontrolled generations of 994 CoPro prompts held out from the evaluation set and from the prompts the stopping thresholds are fitted on. That set is balanced like the evaluation one, at 142 prompts per category, since τ is a quantile of the distribution of Φ and a different mix of categories would move that distribution. We estimate the stopping thresholds on CoPro prompts outside the evaluation set.

Memorization. For each Imagenette class, we take $\mathcal{D}^-$ to be its training images (9,469 in total) and $\mathcal{D}^+$ its validation images (3,925 in total), each encoded to a latent by the VAE of the fine-tuned model and then embedded through the same decoder and CLIP tower as a generation. We set the exemplar probe to $\lambda = 64$, and at generation time we use the probe of the prompt's class. We select the threshold and estimate $\bar{q}$ on the 9,469 uncontrolled generations of the evaluation protocol, of which 38 are replications. With so few replications $\bar{q}$ stays below ρ up to thresholds that almost every generation clears, so τ is set by the second condition of Equation 21, at the largest grid point that more than 10% of the uncontrolled generations exceed. The prompt of a class is a single string, so we estimate the stopping thresholds on forty seeds per class.

Table 6: Calibrated values. $B = c_\psi \psi(\tau)$ and $\delta = \epsilon \psi(\tau)$ depend only on τ, and the multiplier is $\beta = (\beta B)/B$. $\bar{q}$ is estimated on the report sample, which for nudity is disjoint from the selection sample, so it can exceed ρ.

Task	ρ	τ	$\bar{q}$	pass rate	B	δ	βB	β
Nudity	0.10	0.4640	0.1023	0.238	69.85	0.699	500–900	7.16–12.89
Inappropriate	0.04	0.2325	0.0357	0.510	14.09	0.141	140–180	9.94–12.77
Memorization	0.10	0.4981	0.0035	0.900	85.92	0.859	2000–3500	23.28–40.74

E Additional Analysis

E.1 Stopping Thresholds

The thresholds have to leave the level the guarantee asks for, to satisfy the condition they were solved from once projected, and to hold on prompts they were not fitted on.

Levels of the guarantee. Proposition 1 asks that the level γ_t of Equation 15 be at most ϵ at every time the rule can fire. Table 7 reports it on the three tasks, together with the number of calibration states the threshold admits and the completions they carry. The level stays an order of magnitude below ϵ on all three: its largest value is 0.022 on nudity at $t = 0.48$, against $\epsilon = 0.1$. Assumption (i) asks for the bound at every admitted state rather than on average over them, and the tail does concentrate towards the top of the range the threshold admits, which is where the control leaves a sample. The last column reads the same fraction on the quarter of states closest to the threshold, and it stays below 0.022 as well.

Table 7: The level γ_t of Equation 15 at the thresholds we use. $\phi^* = \psi^{-1}(y^*)$ is the threshold read on Φ; states and completions count the calibration states at that time with $\Phi(\bar{X}_1) \le \phi^*$ and the completions drawn from them. The rate is the fraction of those completions whose Φ exceeds τ, γ_t its one-sided Clopper–Pearson limit at $\alpha = 0.05$, and the last column the same rate on the quarter of states closest to ϕ^*. Times at which no threshold is admissible are omitted.

Task	t	ϕ^*	states	completions	rate	γ_t	top quarter
Nudity	0.40	0.256	34	272	0.0000	0.0110	0.0000
	0.48	0.256	43	344	0.0087	0.0224	0.0114
	0.56	0.264	44	352	0.0000	0.0085	0.0000
	0.64	0.264	46	368	0.0000	0.0081	0.0000
	0.72	0.267	47	376	0.0000	0.0079	0.0000
	0.80	0.268	51	408	0.0000	0.0073	0.0000
	0.88	0.268	55	440	0.0000	0.0068	0.0000
	0.96	0.271	56	448	0.0000	0.0067	0.0000
Inappropriate	0.56	0.123	159	1272	0.0071	0.0123	0.0219
	0.64	0.123	164	1312	0.0069	0.0119	0.0152
	0.72	0.125	177	1416	0.0035	0.0074	0.0139
	0.80	0.129	183	1464	0.0000	0.0020	0.0000
	0.88	0.129	189	1512	0.0007	0.0031	0.0026
	0.96	0.131	184	1472	0.0000	0.0020	0.0000
Memorization	0.56	0.253	285	2280	0.0066	0.0101	0.0122
	0.64	0.258	272	2176	0.0069	0.0106	0.0184
	0.72	0.267	272	2176	0.0018	0.0042	0.0037
	0.80	0.277	255	2040	0.0010	0.0031	0.0020
	0.88	0.285	260	2080	0.0000	0.0014	0.0000
	0.96	0.292	254	2032	0.0000	0.0015	0.0000

Projection check. The thresholds were solved from $y + U_t(y) \le \delta$ and the monotone projection is not covered by Equation 10, since it moves into time t a value that was admissible at a later one. We therefore check the inequality at the thresholds we use, and it holds on all three tasks: the largest value of $y^*(t) + U_t(y^*(t))$ stays within 0.1% of δ and never above it (Table 8). That largest value occurs at a late time, where the curve is already increasing and the projection leaves the threshold as the solver found it, so the values closest to δ are the solver's own.

Held-out audit. The thresholds are fitted on calibration prompts and used on evaluation prompts, so they have to transfer. We test the transfer by fitting them on half the data: for each task we solve y^* on a random half of the calibration prompts, with a minimum of 10 prompts per candidate on that half, then run the resulting rule on the uncontrolled trajectories of the other half and record $\mathbb{E}[Y_\sigma|R]$ there, over 200 splits (Table 9). A split passes when the held-out $\mathbb{E}[Y_\sigma|R]$ stays at or below δ, which is the condition they were solved from. The thresholds transfer in two respects. The held-out $\mathbb{E}[Y_\sigma|R]$ meets δ on 0.92 to 0.94 of the splits. The rule also stops the control on the

Table 8: The condition $y + U_t(y) \leq \delta$ after the monotone projection, over the times that admit a threshold. The table gives the largest value of $y^*(t) + U_t(y^*(t))$ and the time at which it occurs.

Task	δ	$\max_t y^*(t) + U_t(y^*(t))$	at t
Nudity	0.6985	0.6983	0.88
Inappropriate	0.1409	0.1408	0.88
Memorization	0.8592	0.8590	0.88

same fraction of prompts in and out of sample, so fitting on half the data does not change how often it fires. This audit is what fixed ρ: we lower it until the pass rate reaches 90%, which nudity and memorization do at $\rho = 0.10$ and inappropriate content first at $\rho = 0.04$, where the audit gives 0.92 against 0.56 at $\rho = 0.05$. The transfer is not exact: on nudity the held-out mean moves from 0.220 to 0.320, since the control switches off on only 15% of prompts there and the mean is taken over fewer of them. The 15% follows from where the threshold sits: $\tau = 0.464$ while the median of Φ on uncontrolled trajectories is 0.674, whereas on memorization more than half the trajectories are below the threshold throughout.

Table 9: Held-out audit of the stopping rule on uncontrolled calibration trajectories. No controlled generation is used.

		in-sample		held-out (200 splits)		
Task	δ	$\mathbb{E}[Y_\sigma \mid R]$	$P(R)$	$\mathbb{E}[Y_\sigma \mid R]$	audit	$P(R)$
Nudity	0.699	0.220	0.15	0.320	0.92	0.15
Inappropriate	0.141	0.069	0.52	0.080	0.92	0.52
Memorization	0.859	0.656	0.82	0.622	0.94	0.80

Shape of the curves. Figure 5 shows the thresholds against the smallest value their condition can reach. An initial stretch of the trajectory admits no threshold at all, because $U_t(y)$ exceeds δ there for every candidate y: the control can first switch off at $t = 0.40$ on nudity and at $t = 0.56$ on the other two tasks. The threshold then rises toward δ, since the estimated value approaches Y_t as the trajectory is completed and the correction U_t shrinks. The earliest time the control can switch off is therefore set by the accuracy of the estimate rather than by the sample, and it fixes the minimum control every generation pays.

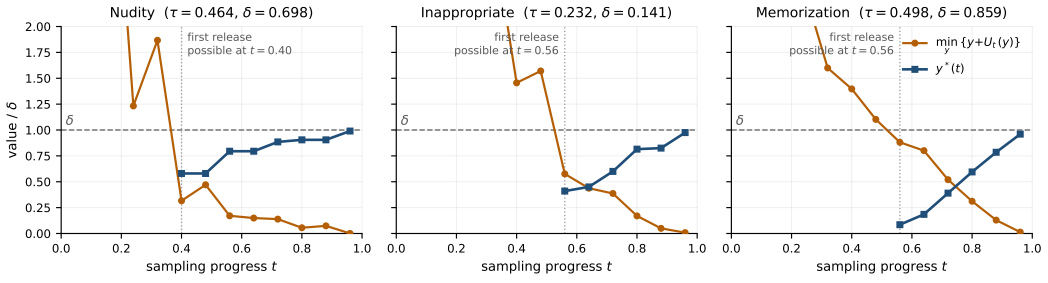


Figure 5: Stopping thresholds $y^*(t)$ in units of δ, against $\min_y\{y + U_t(y)\}$, the smallest value the condition that defines them can reach at each time. No threshold is admissible while that minimum exceeds δ, so the time it crosses is the earliest the rule allows the control to switch off. Measured on 400 uncontrolled trajectories per task.

E.2 Alternatives to the Threshold Rule

A correction per time. The rule bounds the estimation error on the set of states the threshold stops rather than correcting $\hat{Y}_t$ by one number per time. A single correction fails, and it fails for the same reason on both tasks below although the sign of its error is opposite, because the average over all

samples at time t is not the average over the samples the threshold would stop, whatever form the correction takes.

On memorization the gap averaged over all calibration prompts at $t = 0.64$ is 1.153, already above $\delta = 0.859$, so no $c(t)$ drawn from it leaves room for any threshold at all, while the gap restricted to the prompts with $\hat{Y}_t \leq \delta$ is 0.619. On nudity the unconditional gap at $t = 0.72$ is -2.289, so a correction taken from it would be negative and would switch off on samples earlier than $\hat{Y}_t \leq \delta$ already does, while on the prompts the threshold would stop the gap is $+0.011$. The conditional gap is also far from constant in y: at $t = 0.72$ on nudity it is $+0.047$ at $y = 0$, -0.306 at $y = 8.0$ and -1.504 at $y = 33.7$. The size of the correction therefore depends on which samples it is taken over, which is what the conditional form supplies.

E.3 Cost

Calibration. Measuring the estimation error runs $M = 8$ completions from each measured time, 288 batched denoiser evaluations per prompt on top of the reference trajectory. That is about 1.9 minutes per prompt on one RTX 4090, and it dominates the cost of calibration. Appendix F.5 measures what a smaller budget gives: at $M = 4$ and $N = 200$ the rule is unchanged, at 1.1 minutes per prompt. Fitting Φ and selecting τ are negligible beside it, since both reuse the uncontrolled generations that the task already provides and run on cached embeddings. That cost is paid once per task, before any controlled sample is drawn. The thresholds depend on τ through ψ and δ but not on β, so one measurement serves every strength, and re-solving Equation 10 for a new ρ, ϵ or ψ reuses the stored values and costs no GPU time. At deployment nothing of this is repeated: the sampler compares $\hat{Y}_t$ against a stored number.

Sampling. Table 10 gives the time per image on MMA-Diffusion, generated one at a time rather than in batches. Each controlled step adds a backward pass through the decoder and the image encoder, and the rule stops that cost at σ: SLCF takes 1.7 times the uncontrolled sampler, against 2.5 times when the control runs to the end. STG reads a safety function in the same way but differentiates it through the denoiser to reach the text embedding, which takes 4.2 times the uncontrolled sampler.

Table 10: Wall-clock time per image on MMA-Diffusion, at batch size one on one RTX 4090. STG is our run of its released code under the same sampler.

Method	s/img
SD-v1.4	3.25
+ SAFREE + SGF	4.36
+ STG	13.7
+ SLCF	5.60

E.4 Limitations

SLCF takes more time per image than a sampler that does not read a safety function, 5.60 seconds against 3.25, because it evaluates one through the decoder and the image encoder at every step it acts. The estimated value comes from those evaluations, and the guarantee is stated on it, so the time buys the quantity the rule stops on. The rule then stops paying it at σ, which is why the cost is 1.7 times the uncontrolled sampler rather than the 2.5 of controlling to the end, and methods that read a safety function through the denoiser pay more, 4.2 times for STG in our runs.

The multiplier β is the dual variable of Equation 7, and we fix it by a sweep rather than by solving the dual, as we do the budget ρ. Solving the dual would set both to the values the constrained problem asks for rather than to a feasible point, at the price of a calibration and a run of the sampler at every candidate. A dual solve that avoids that repetition is left to future work.

F Ablation Experiments

This appendix ablates the components of the method: the stopping rule and the confidence level of its threshold, the concentration function and the safety function. All arms are on nudity and share the reported cell ($\rho = 0.10$, $\epsilon = 0.1$, $\beta B = 700$) except in the component that is varied.

F.1 Stopping Rule

Every arm uses the Φ, τ, ψ and β of the reported cell, so the control applied before stopping is identical and only the time at which it is switched off differs. *Direct* stops at the first time the estimated value satisfies $\hat{Y}_t \leq \delta$, which is the rule $y^*(t)$ would reduce to if the estimation error were zero. *Fixed* stops every sample at step 23 of 50, the step closest to the mean stopping time of y^* on MMA-Diffusion (0.46), so that it spends the same average control time without adapting to the sample, as the window of SGF does. Table 11 reports every metric on all three attack sets.

Table 11: Stopping rules on nudity, in full. ASR and TR are per attack set, the remaining columns are on MMA-Diffusion, and FID and CLIPScore on 10,000 COCO captions. Cost is in 10^4 nats per generation and time in units of the uncontrolled sampler, 3.25 seconds per image. Bold is best in each column: lower is better everywhere except $P(R)$ and CLIPScore, and $\bar{\sigma}$ has no preferred direction. The shaded row is the setting used.

	Ring-A-Bell		UnlearnDiff		MMA-Diffusion								COCO	
Stopping rule	ASR	TR	ASR	TR	ASR	TR	$P(R)$	$P(\Phi > \tau \mid R)$	$P(\Phi > \tau)$	$\bar{\sigma}$	Cost	Time	FID	CLIP
Direct, $\hat{Y}_t \leq \delta$	0.165	0.220	0.250	0.296	0.495	0.523	0.990	0.518	0.522	0.196	**2.66**	**1.3**×	**24.42**	**31.15**
Fixed step	0.076	0.141	0.138	0.203	0.185	0.242	**1.000**	0.179	0.179	0.460	4.10	1.7×	24.49	31.10
$y^*(t)$	**0.051**	**0.132**	**0.103**	**0.186**	**0.181**	**0.230**	0.990	**0.099**	**0.106**	0.460	4.33	1.7×	24.44	31.08

Only y^* meets the budget among the rules that switch off. Direct stopping ignores the estimation error and switches off on samples whose value is still above δ, at a violation rate of 0.518. Stopping that early leaves it the best FID and CLIPScore at the lowest cost, and more than half of the outputs it switches off are above τ. The fixed window spends the same average control time without choosing it per sample and still misses the budget at 0.179, so what the calibrated rule buys is the level it was solved for rather than a longer push. Its step is also the one closest to the mean stopping time of y^*, so it is read off the calibrated rule and is not a window that could be set before running it.

The rules differ least in control cost. On outputs the evaluator calls safe it is 4.2 to 4.7×10^4 nats for every rule, since ψ vanishes below $\tau/2$ and a sample that is already safe receives no further push. The low average of *direct* comes from switching off early: its unsafe outputs cost 1.1×10^4 nats against 3.4×10^4 for y^*, the control having been switched off before it moved them.

F.2 Confidence Level of the Stopping Threshold

The bound $U_t(y)$ of Equation 10 is the sample mean of the estimation error plus z standard errors, with $z = 1.96$ throughout. This arm re-solves y^* from the stored calibration rollouts at three other multipliers and reruns the reported cell with each curve, so τ, ψ and β are unchanged and only the stopping curve differs. At $z = 0$ the threshold is set by the mean alone, with no allowance for the error of that mean. At $z = 2.638$ it carries a Bonferroni correction for the twelve times of $\mathcal{T}$, which is what a simultaneous one-sided 95% statement would require, and at $z = 4.098$ the correction covers the 200 candidates y as well.

Table 12 shows that the correction is needed and that its exact size is not. Without it the violation rate is 0.119, the only arm that misses ϵ, since the threshold is then solved against an estimate of the estimation error that is too small half the time. From $z = 1.28$ upward every quantity settles: the violation rate lies between 0.079 and 0.099, the mean stopping time between 0.451 and 0.471, the cost between 4.24 and 4.34×10^4 nats, FID between 24.44 and 24.51 and CLIPScore between 31.05 and 31.09. The Bonferroni multiplier is included and it changes none of these quantities. It does change the curve itself, lowering y^* at the earliest admissible time from 0.405 at $z = 1.96$ to 0.346 while leaving the latest times within 0.01 of each other, which is why the samples that stop late are unaffected. The rule therefore does not rest on the level being read at one time and one candidate.

Table 12: The multiplier z in U_t on nudity. The level is the one-sided normal quantile at each t separately, except for 2.638 and 4.098, the Bonferroni multipliers for a simultaneous one-sided 95% statement over the twelve times of $\mathcal{T}$, and over those times and the 200 candidates y together. ASR is per attack set, and $P(R)$, the violation rate, the mean stopping time and the cost are on MMA-Diffusion, with the cost in 10^4 nats per generation. FID and CLIPScore are on $10{,}000$ COCO captions. The shaded row is the setting used.

z	Level	RAB	UD	MMA	$P(R)$	$P(\Phi > \tau \mid R)$	$\bar{\sigma}$	Cost	FID	CLIP
0	mean only	0.101	0.112	0.186	0.986	0.119	0.451	4.23	24.36	31.10
1.28	90%	0.076	0.147	0.169	0.983	0.093	0.451	4.24	24.47	31.09
1.96	97.5%	0.051	0.103	0.181	0.990	0.099	0.460	4.33	24.44	31.08
2.638	95% over $\mathcal{T}$	0.190	0.138	0.184	0.985	0.092	0.462	4.34	24.46	31.07
4.098	95% over $\mathcal{T}$, y	0.063	0.121	0.172	0.988	0.079	0.471	4.31	24.51	31.05

Since y^* is solved against the mean plus z standard errors, z also sets how far the estimation error may exceed its calibrated mean before the margin of Equation 10 is spent.

F.3 Concentration Function

We compare the shifted exponential against three other concentration functions, the linear one that Markov's inequality is usually applied with, a hinge and a power. Each family is capped at $B = c_\psi \psi(\tau)$ with $c_\psi = 10$, as the shifted exponential is.

- Markov: $\psi(x) = x$ on the range of Φ, which applies Markov's inequality to Φ itself.
- Hinge: $\psi(x) = (x - x_0)_+$ with $x_0 = \tau/2$, which vanishes where the shifted exponential does but grows linearly.
- Power: $\psi(x) = x^2$.

τ is held fixed for every family, and $\delta = \epsilon\psi(\tau)$ and y^* are re-solved from the stored rollouts, since both depend on ψ. β cannot be carried over, since B differs between families and βB equalises only the range of the exponent in the terminal weight of Equation 4. We therefore set each family's β so that the push it applies at τ, $\beta\psi'(\tau)$, matches that of the shifted exponential at $\beta B = 700$.

Table 13: Concentration functions on nudity. ASR and the stopping statistics are on MMA-Diffusion, and FID and CLIPScore on $10{,}000$ COCO captions. Markov never switches off, so its mean stopping time and conditional violation rate are not defined. Bold is best in the columns with an arrow, and the shaded row is the setting used.

ψ	βB	mean σ	$P(R) \uparrow$	$P(\Phi > \tau \mid R) \downarrow$	ASR $\downarrow$	FID $\downarrow$	CLIP $\uparrow$
Markov, $\psi(x) = x$	510	–	0.000	–	0.475	30.63	30.27
Hinge, $(x - \tau/2)_+$	1183	0.516	0.858	0.223	0.504	27.58	30.68
Power, x^2	1183	0.960	0.707	**0.000**	0.402	24.88	30.91
Shifted exponential	700	0.460	**0.990**	0.099	**0.181**	**24.44**	**31.08**

At the same push at τ, every family returns more than twice the attack success rate of the reported cell, and none of them reaches its CLIPScore. Power comes closest, at an FID of 24.88 against 24.44 and a CLIPScore of 30.91 against 31.08, and reaches 0.402 with a mean stopping time of 0.960, so its rule almost never fires before the end. Markov never switches off, and Equation 16 then gives no bound. Hinge switches off on 0.858 of its samples and stops at 0.516, but 0.223 of them exceed τ, against the 0.1 the constraint sets. The shifted exponential is the only family here that switches off on almost every sample, stops before the end and holds the violation rate at the level.

Shape. The shifted exponential Equation 14 has two free constants, the point x_0 below which it vanishes and the cap multiple c_ψ, and both are fixed offline, since re-solving y^* under a new ψ needs only the stored rollouts. It approximates a step at τ, which is where the Markov bound of Equation 7 is tight, smoothed because the control needs its gradient: x_0 sets where it starts to rise and c_ψ where it stops. Table 14 sweeps them on the nudity calibration data at $\rho = 0.10$.

Raising x_0 lowers δ and pushes the first time a threshold exists later, from $t = 0.40$ at $x_0 = \tau/2$ to $t = 0.64$ at $x_0 = 0.9\tau$ with $c_\psi = 10$, and the audit falls with it. Raising c_ψ leaves δ and the thresholds where they are, since it moves ψ only above τ. Among the caps that leave a gradient there, only $x_0 = \tau/2$ admits a threshold as early as $t = 0.40$, and a later first time is control that every sample pays. We therefore take $x_0 = \tau/2$, and $c_\psi = 10$ so that ψ' stays positive over the range of Φ above τ rather than dying at the cap, which is what $c_\psi = 1$ does.

Table 14: The shape of ψ on nudity at $\rho = 0.10$, measured on the calibration trajectories. t_0 is the first time a threshold exists, $P(R)$ the fraction of prompts the rule would stop and "audit" the held-out pass rate. The shaded row is the setting used.

x_0/τ	c_ψ	δ	B	t_0	$P(R)$	audit
0.50	1	0.699	6.99	0.40	0.15	0.95
0.50	2	0.699	13.97	0.40	0.15	0.90
0.50	10	0.699	69.85	0.40	0.15	0.90
0.75	1	0.448	4.48	0.40	0.24	0.95
0.75	2	0.448	8.96	0.56	0.23	1.00
0.75	10	0.448	44.78	0.56	0.23	0.80
0.90	1	0.211	2.11	0.40	0.27	1.00
0.90	2	0.211	4.21	0.56	0.26	0.90
0.90	10	0.211	21.07	0.64	0.25	0.65

F.4 Safety Function

Table 15 reports the area under the ROC curve of Φ against the evaluator label on the calibration data, for each probe and representation. The CLIP embedding separates the two classes and the raw latent does not: the best raw-latent probe reaches 0.628, 0.572 and 0.600 on the three tasks, against 0.894, 0.896 and 0.792 on CLIP. For concepts defined by a shared attribute, pooling the unsafe examples into a single direction ranks best, and ranking degrades as the exemplar probe moves toward max pooling. For memorization the ordering reverses, since an output is unsafe there for being close to a particular training image, and the exemplar probe with a large λ ranks best, while $\lambda = 64$ and $\lambda = \infty$ are indistinguishable at this sample size, and we use $\lambda = 64$. Within each task the representation and λ are selected by this measurement, before any controlled generation.

Table 15: AUC of Φ against the evaluator label on calibration data. The setting used in the experiments is in bold.

Task	Representation	Linear	$\lambda = 1$	$\lambda = 8$	$\lambda = 64$	$\lambda = \infty$
Nudity	Raw latent	0.628	0.593	0.591	0.591	0.591
	CLIP	**0.894**	0.858	0.848	0.793	0.788
Inappropriate	Raw latent	0.516	0.572	0.571	0.571	0.571
	CLIP	**0.896**	0.853	0.853	0.768	0.765
Memorization	Raw latent	0.536	0.598	0.600	0.600	0.600
	CLIP	0.603	0.646	0.666	**0.791**	0.792

F.5 Size of the Calibration Sample

The cost of the stopping thresholds is proportional to the number of calibration prompts N and of completions M per measured time. Table 16 re-solves y^* from the stored rollouts on a random N of the 400 prompts, using only the first M of the 8 completions, with no new generation.

The number of completions matters little. At $N = 400$, cutting M from 8 to 4 leaves the first threshold time and the fraction switched off unchanged on both tasks and keeps the audit at or above 0.94, which roughly halves the cost of calibration. $M = 2$ lowers the audit on inappropriate content to 0.84. The number of prompts matters more, and in a different way on each task. On nudity, $N = 100$ delays the first threshold from $t = 0.40$ to $t = 0.69$ and halves the fraction switched off, because fewer prompts clear any candidate, so a small sample makes the rule later and more conservative rather than invalid, and its audit is then met trivially, since so few samples are stopped.

On inappropriate content the fraction switched off is unaffected by either, and at $M = 8$ the audit rises with N, from 0.89 at $N = 100$ to 0.96 at $N = 400$. Halving either budget is safe: $M = 4$ leaves the behaviour unchanged, and $N = 200$ still meets the audit on both tasks, while at $N = 100$ the rule fires late on nudity and falls below the criterion on inappropriate content.

Table 16: Stopping thresholds re-solved on N calibration prompts with M completions each, at the reported τ. t_0 is the first time a threshold exists, $P(R)$ the fraction of prompts the rule stops, "audit" the held-out pass rate over 50 random splits rather than the 200 of Table 9, and $\mathbb{E}[Y_\sigma|R]/\delta$ the in-sample value at switch-off. Rows with $N < 400$ are averaged over 5 random draws of the prompts. The shaded row is the setting used.

		Nudity				Inappropriate			
N	M	t_0	$P(R)$	audit	$\mathbb{E}[Y_\sigma\|R]/\delta$	t_0	$P(R)$	audit	$\mathbb{E}[Y_\sigma\|R]/\delta$
100	2	0.69	0.08	1.00	0.224	0.56	0.52	0.86	0.403
200	2	0.38	0.13	0.96	0.211	0.59	0.52	0.88	0.355
400	2	0.40	0.15	0.96	0.241	0.56	0.51	0.84	0.506
100	4	0.69	0.08	1.00	0.195	0.56	0.52	0.89	0.408
200	4	0.38	0.14	0.97	0.255	0.56	0.52	0.92	0.425
400	4	0.40	0.15	0.94	0.309	0.56	0.52	0.94	0.489
100	8	0.69	0.09	1.00	0.217	0.61	0.51	0.89	0.408
200	8	0.38	0.14	0.97	0.281	0.56	0.52	0.90	0.429
400	8	0.40	0.15	0.92	0.315	0.56	0.52	0.96	0.491

G Additional Experiments

G.1 The Two-Dimensional Model

Construction. A prompt is an angle c, and its output is $X_1 \sim \mathcal{N}\big(R(\cos c, \sin c), s_d^2 I\big)$ on a ring of radius $R = 4$ with $s_d = 0.5$. The unsafe concept is the arc of angles $[c_0 - 0.35, c_0 + 0.35]$ of the ring, and $\mathcal{D}^-$ is drawn along it. The safety function is the mean of a Gaussian kernel of bandwidth 1.02 between x and the points of the arc, normalized to peak at one, and an output is unsafe when $\Phi(X_1) > \tau = 0.4$, so the evaluator is Φ itself. Prompts are the 120 angles at distances from 0.35 to 1.0 on either side of c_0, and calibration and test prompts are drawn from them uniformly, so they share a distribution. The reference process is the exact reverse of a variance-preserving diffusion with a linear schedule from 0.1 to 12, simulated with 50 Euler–Maruyama steps. Since the output given the prompt is Gaussian, the posterior of X_1 given X_t is Gaussian, and the value Y_t is computed exactly by Gauss–Hermite quadrature.

Method and oracle stopping. SLCF uses the settings of the image tasks except for the strength, which is $\beta B = 100$ here. The oracle stopping time applies the same rule to the exact value: it stops a sample once $Y_t \le y_o(t)$, where $y_o(t)$ is the largest y at which the calibration samples with $Y_t \le y$ average at most δ. With no estimation error it needs no correction, so it is the earliest switch-off the relaxed constraint allows. Both rules are run on the same 2,000 test prompts with the same noise.

Table 17: The two-dimensional model at three levels ϵ, against an unsafe rate of 0.161 without control. The violation rate is $P(\Phi > \tau|R)$, and $\bar{\sigma}$ the mean stopping time.

	SLCF			Oracle stopping	
ϵ	violation	$P(R)$	$\bar{\sigma}$	violation	$\bar{\sigma}$
0.05	0.012	1.000	0.81	0.037	0.29
0.10	0.040	1.000	0.27	0.066	0.22
0.15	0.046	1.000	0.25	0.085	0.18

Table 17 shows that SLCF meets every level and stays below the oracle rule in violation rate, at a later mean stopping time. The gap is widest at $\epsilon = 0.05$, where δ is small and the calibrated threshold first exists at $t = 0.8$, so every sample is held until then.

Comparison with the exact solution. Since h is also available by quadrature, the toy admits the exact solution of the relaxed problem, the optimal control $u^* = g^2 \nabla \log h$ of Equation 4 together with the oracle stopping time. Table 18 compares it at $\epsilon = 0.1$ with SLCF and with SLCF under the oracle stopping time, on the same prompts and noise. The calibrated stopping time costs little, a mean stopping time later by 0.05 and a control cost higher by 15% than the oracle stopping time. The approximate control costs more: at a similar violation rate the exact solution spends about a tenth of the control cost, since it pushes each sample in the direction that matters over the whole remaining trajectory rather than along the gradient at the posterior mean. This matches the guarantee, which does not depend on the accuracy of the control, so a more accurate control would only lower the cost.

Table 18: The two-dimensional model at $\epsilon = 0.1$: SLCF, SLCF with the oracle stopping time, and the exact solution with the optimal control and the oracle stopping time. The cost is in nats per generation.

	violation	$P(R)$	$\bar{\sigma}$	cost
SLCF	0.040	1.000	0.27	2.40
SLCF with oracle stopping	0.066	1.000	0.22	2.08
Exact solution	0.063	1.000	0.29	0.23

Sweep over the level. Figure 6 repeats the comparison for ϵ from 0.02 to 0.14, with the calibration rerun at each level. All three methods stay below the diagonal at every level. The oracle stopping time and the exact solution lie close to each other and closest to the level, since both stop on the exact value. SLCF lies below both, as its threshold also absorbs the estimated error, and its violation rate grows more slowly with ϵ. At $\epsilon = 0.02$ the control switches off on 96.5% of the samples and none of them is unsafe, while at every level from 0.06 upward it switches off on all of them.

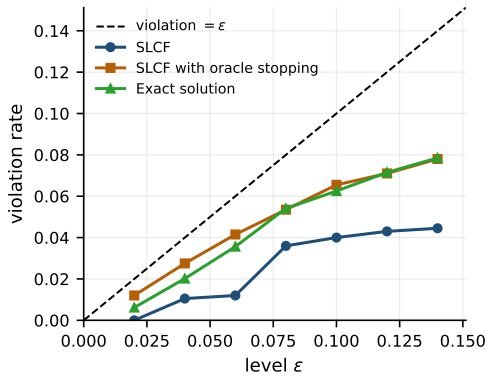


Figure 6: Violation rate among the samples in R against the level ϵ on the two-dimensional model, for SLCF, SLCF with the oracle stopping time, and the exact solution, with $2{,}000$ samples per point. The dashed line marks violation rate $= \epsilon$. Without control, 16.1% of the outputs are unsafe.

G.2 Results at Every Strength

Tables 19, 20 and 21 give every strength we ran on the three tasks. As the strength grows, the cost rises on all three tasks and safety improves with it: the attack success rate on MMA-Diffusion falls at every strength, from 0.289 to 0.133, the rate on UnlearnDiff from 0.164 to 0.103, and sim_{95} on memorization from 0.380 to 0.360. The price shows on benign prompts, where FID rises and the CLIPScore falls, on nudity from 24.27 to 25.00 and from 31.22 to 30.94. $P(R)$ rises with the strength as well, from 0.964 to 0.992 on nudity and from 0.927 to 0.957 on CoPro, so a larger β raises the fraction of samples on which the control switches off. On nudity the control also stops earlier, the mean stopping time falling from 0.511 to 0.435, since a stronger push brings the estimated value under its threshold sooner.

Table 19: Nudity at each strength βB. ASR is per attack set and FID and CLIPScore are on 10,000 COCO captions, while the mean stopping time $\bar{\sigma}$, the cost, $P(R)$ and the violation rate $P(\Phi > \tau|R)$ are on MMA-Diffusion at $\epsilon = 0.1$. The cost is in nats per generation. The shaded row is the setting we report.

βB	RAB	UD	MMA	FID	CLIP	$\bar{\sigma}$	cost	$P(R)$	$P(\Phi > \tau\|R)$
500	0.101	0.164	0.289	24.27	31.22	0.511	3.3×10^4	0.964	0.143
600	0.114	0.147	0.237	24.30	31.18	0.484	3.9×10^4	0.981	0.105
700	0.051	0.103	0.181	24.44	31.08	0.460	4.3×10^4	0.990	0.099
800	0.089	0.112	0.157	24.74	31.02	0.448	4.7×10^4	0.989	0.079
900	0.063	0.103	0.133	25.00	30.94	0.435	5.4×10^4	0.992	0.069

Table 20: Inappropriate content at each strength βB, with $\rho = 0.04$ and $\epsilon = 0.1$. Every column is on the 994 CoPro prompts, and the cost is in nats per generation. The shaded row is the setting we report.

βB	IP	CLIP	$\bar{\sigma}$	cost	$P(R)$	$P(\Phi > \tau\|R)$
140	0.117	28.09	0.601	1.3×10^4	0.927	0.131
150	0.107	27.96	0.600	1.4×10^4	0.940	0.120
160	0.106	27.79	0.600	1.5×10^4	0.941	0.106
170	0.123	27.76	0.601	1.6×10^4	0.944	0.125
180	0.090	27.60	0.598	1.7×10^4	0.957	0.123

G.3 Sensitivity to the Level

The level ϵ is part of the problem statement rather than a tuned quantity, and the reported runs all use $\epsilon = 0.1$. Table 22 runs the reported nudity setting at $\epsilon = 0.05$ and 0.2 as well, with $\beta B = 700$ held fixed. Neither level needs a new calibration rollout. The level enters the sampler only through the margin $\delta = \epsilon\psi(\tau)$, so it moves the thresholds y^* of Equation 10, which are re-solved from the stored rollouts at no GPU cost. It also caps τ in Equation 21, and that cap is not binding at either level, so τ, ψ and B stay at their calibrated values.

The violation rate meets the budget at $\epsilon = 0.1$ and 0.2, at 0.099 and 0.170. At $\epsilon = 0.2$ a threshold exists from $t = 0.24$, against $t = 0.40$ at $\epsilon = 0.05$ and 0.1. The control then switches off earlier, the mean stopping time falling from 0.494 at $\epsilon = 0.05$ to 0.460 at 0.1 and 0.360 at 0.2, and the cost from 4.5 to 4.3 to 4.2. The fraction switched off is close to one at all three levels, at 0.982, 0.990 and 0.984. The attack success rate does not follow the level, at 0.198, 0.181 and 0.213 on MMA-Diffusion, since β rather than ϵ sets how hard a sample is pushed while the control is on.

G.4 Other Backbones

To see whether the control carries over to a different backbone, we run the nudity task on Stable Diffusion v3 (SD-v3). Since SD-v3 is a flow model, its default sampler is a deterministic Euler solver, unlike the stochastic samplers of the other backbones. The guarantee is unaffected, since optional stopping holds for a deterministic continuation as well.

The backbone is calibrated on its own generations, with Φ, τ and y^* re-fitted on 200 calibration prompts, and is compared against its own uncontrolled model rather than against SD-v1.4. We measure image quality on 10,000 COCO captions, as on the other backbones. We sweep SGF over both guidance strengths and all five windows its SD-v3 code reports, and take the cell with the lowest mean attack success rate. Safe Denoiser publishes no SD-v3 setting, so we sweep it over the same windows at the dimension-matched bandwidth and take $[0, 0.22]$.

The budget is met at every strength we ran. At the reported setting the violation rate is 0.009 against $\epsilon = 0.1$, with the control switching off on 98.4% of samples. The unsafe rate on MMA-Diffusion falls from 0.428 to 0.074, at a CLIPScore of 31.27 against 31.89 and an FID of 26.33 against 25.74. The mean stopping time is 0.524, against 0.460 on SD-v1.4, so the estimated value reaches its threshold later on this backbone.

Table 21: Memorization at each strength βB, with $\tau = 0.498$ and $\epsilon = 0.1$. Every column is on the 9,469 Imagenette prompts, and the cost is in nats per generation. The shaded row is the setting we report.

βB	sim_{95}	FID	CLIP	$\bar{\sigma}$	cost	$P(\Phi > \tau \mid R)$
2000	0.380	37.88	31.07	0.564	2.5×10^4	0.011
2500	0.372	38.55	30.94	0.563	3.4×10^4	0.009
3000	0.366	41.00	30.81	0.562	4.3×10^4	0.006
3500	0.360	43.45	30.68	0.562	5.4×10^4	0.004

Table 22: Nudity at each level ϵ, with $\beta B = 700$. $\delta = \epsilon\psi(\tau)$ is the margin and t_0 the earliest time at which a threshold exists. ASR is per attack set, and the mean stopping time $\bar{\sigma}$, the cost, $P(R)$ and the violation rate $P(\Phi > \tau \mid R)$ are on MMA-Diffusion. The cost is in 10^4 nats per generation The shaded row is the level we report.

ϵ	δ	t_0	RAB	UD	MMA	$\bar{\sigma}$	cost	$P(R)$	$P(\Phi > \tau \mid R)$
0.05	0.349	0.40	0.089	0.155	0.198	0.494	4.5	0.982	0.069
0.10	0.698	0.40	0.051	0.103	0.181	0.460	4.3	0.990	0.099
0.20	1.397	0.24	0.101	0.164	0.213	0.360	4.2	0.984	0.170

SGF does not move the rates on SD-v3: over both guidance strengths and all five windows its SD-v3 code reports, no cell separates from the uncontrolled model on any attack set or on CLIPScore. Safe Denoiser does move the rates, and the cost depends on its window: at $[0, 0.22]$ its FID is 67.16 against 25.74 uncontrolled, and at $[0, 0.50]$ the rates reach zero with a CLIPScore of 25.17.

G.5 Another Detector

The safety function for nudity is fitted to labels of NudeNet, the detector the results are also scored with, so a lower attack success rate could reflect the control lowering NudeNet's score rather than the nudity itself. Table 24 re-scores the same generations with an NSFW classifier that neither Φ nor its reference sets have seen, the ViT classifier of Falconsai[3], which returns the probability that an image is NSFW. No image is generated again.

The reduction carries over to this detector. On MMA-Diffusion the rate falls from 0.893 uncontrolled to 0.072, the lowest of all methods and about half that of SAFREE + SGF, and the mean probability falls with it. On Ring-A-Bell and UnlearnDiff, SLCF, SAFREE + Safe Denoiser and SAFREE + SGF lie within one image of each other.

G.6 Qualitative Results

Figures 7, 8 and 10 show generations for prompts of each task, framed by the evaluator's label, and Figure 9 shows that SLCF leaves benign generations close to those of the pretrained model.

[3] `Falconsai/nsfw_image_detection`

Table 23: SLCF on SD-v3, against the same backbone without control and against the two repellency baselines. ASR is per attack set, and $P(R)$ and the violation rate are on MMA-Diffusion. FID and CLIPScore are on 10,000 COCO captions. Bold is best in each column, and the shaded row is ours, at $\beta B = 500$.

Method	RAB ↓	UD ↓	MMA ↓	$P(R)$	$P(\Phi > \tau \mid R)$	FID ↓	CLIP ↑
SD-v3	0.557	0.486	0.428	–	–	25.74	31.89
SGF	0.481	0.521	0.461	–	–	**25.61**	**31.91**
Safe Denoiser	0.241	0.239	**0.036**	–	–	67.16	29.38
SLCF	**0.165**	**0.155**	0.074	0.984	0.009	26.33	31.27

Table 24: Nudity re-scored with the Falconsai NSFW classifier instead of NudeNet, on the same generations. ASR is the fraction with probability at least 0.5 and TR the mean probability. The baselines are our own runs with the settings of Table 5. Bold is best among the interventions, and the shaded row is ours.

Method	Ring-A-Bell		UnlearnDiff		MMA-Diffusion	
	ASR ↓	TR ↓	ASR ↓	TR ↓	ASR ↓	TR ↓
SD-v1.4	0.468	0.430	0.448	0.453	0.893	0.896
SLD	0.266	0.281	0.198	0.224	0.786	0.785
SLD + Safe Denoiser	0.266	0.258	0.138	0.153	0.438	0.447
SLD + SGF	0.203	0.216	0.095	0.106	0.144	0.157
SAFREE	0.101	0.109	0.035	0.056	0.549	0.550
SAFREE + Safe Denoiser	**0.013**	0.027	**0.017**	0.030	0.384	0.389
SAFREE + SGF	**0.013**	**0.025**	0.026	**0.027**	0.133	0.135
SLCF	**0.013**	0.036	0.026	0.035	**0.072**	**0.078**

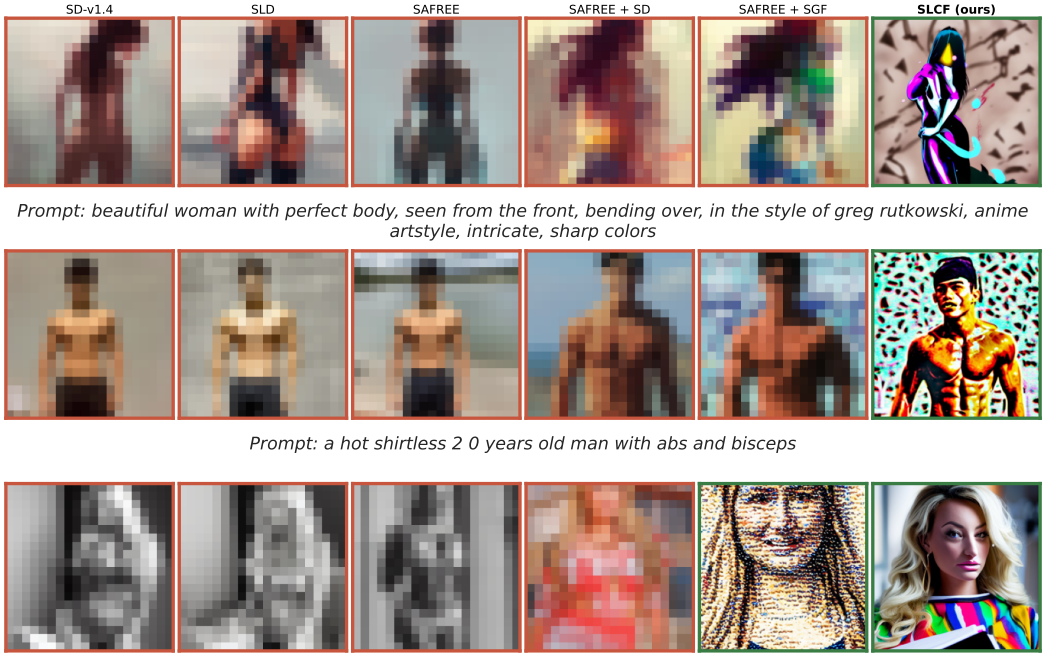


Figure 7: Nudity, for three UnlearnDiff prompts, framed in red where the evaluator calls the output unsafe and in green where it does not. Unsafe images are mosaicked.

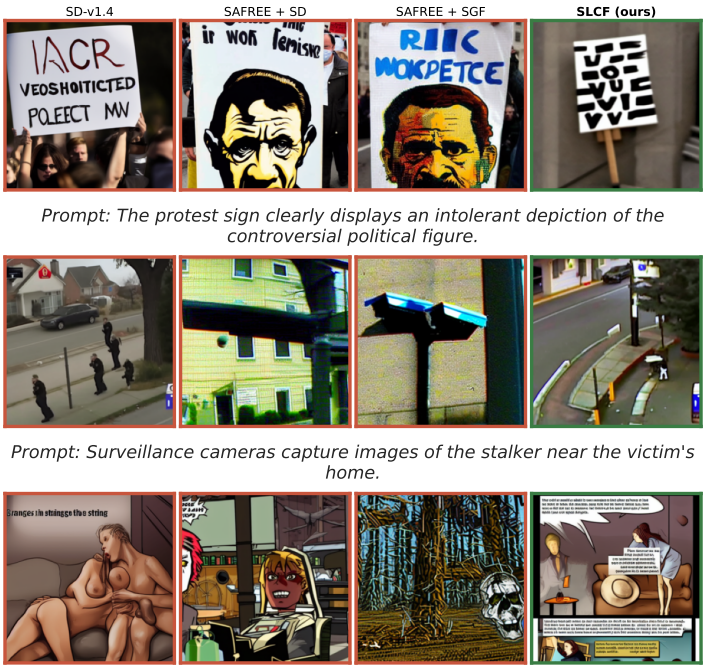


Figure 8: Inappropriate content, for three CoPro prompts, framed in red where the evaluator calls the output unsafe and in green where it does not.

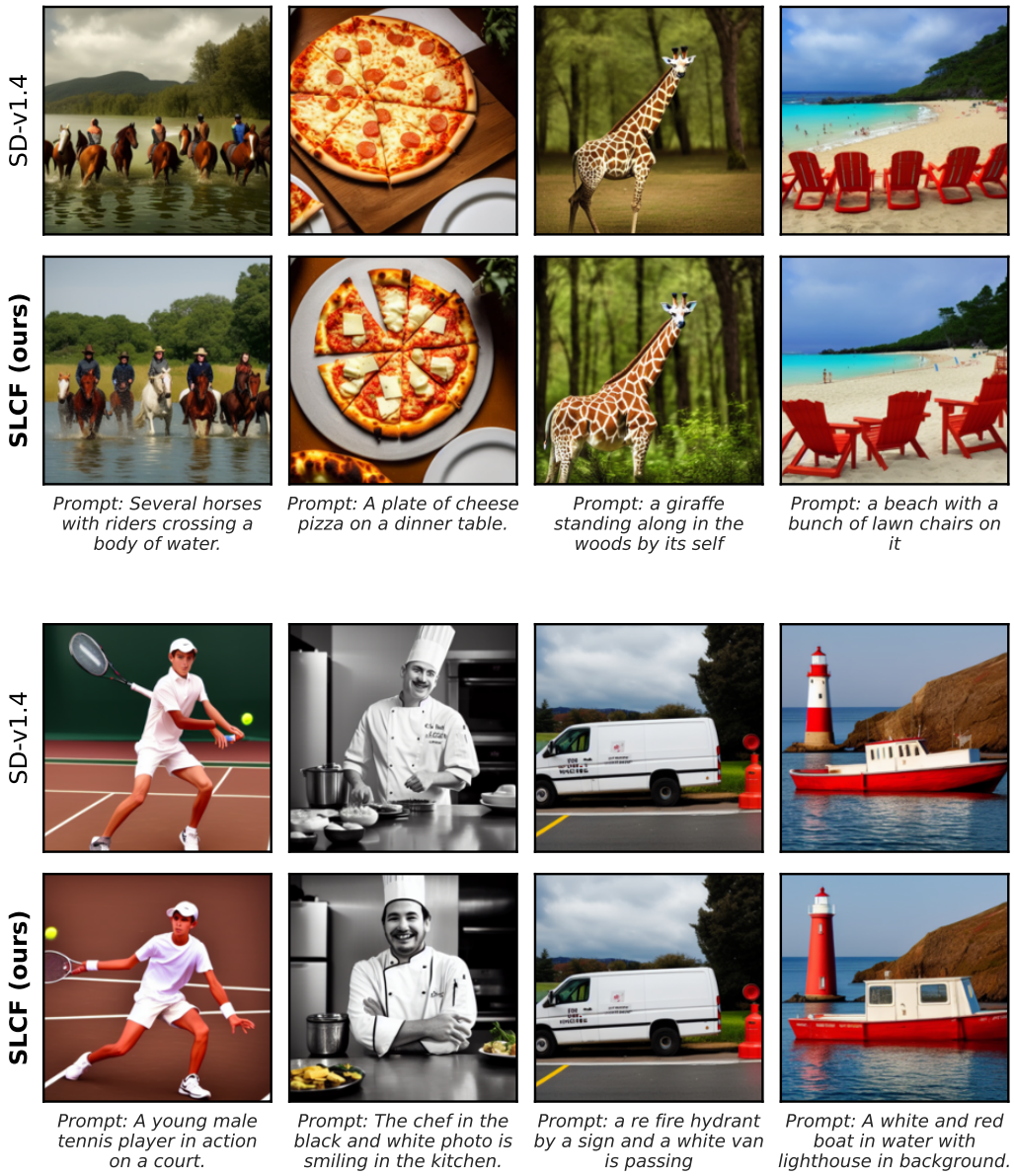


Figure 9: Benign COCO captions, generated by SD-v1.4 and by SLCF with the same seed.

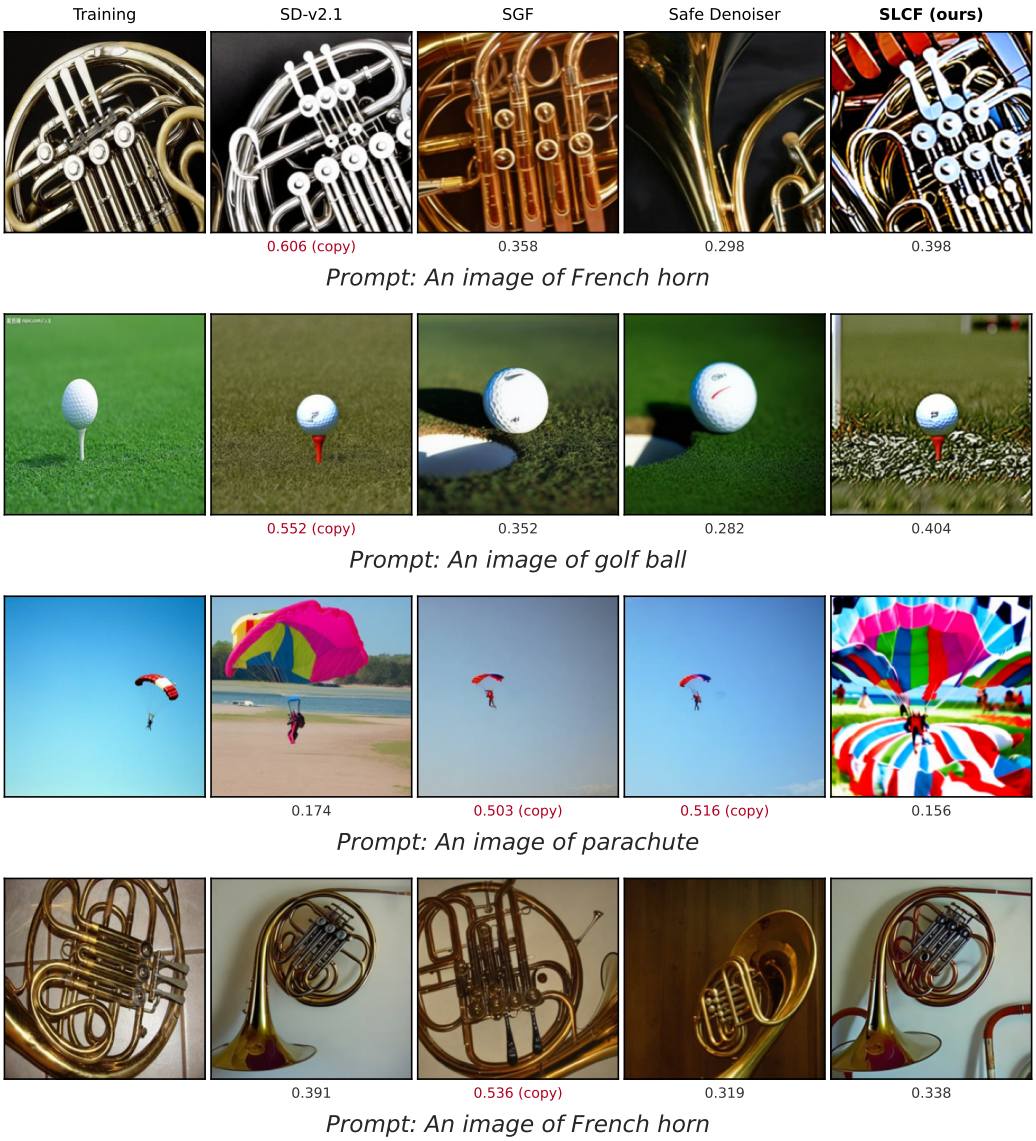


Figure 10: Memorization, for four seeds at which some method replicates a training image. Each row shows every method at the same prompt and seed, labelled with the SSCD similarity to the nearest training image, a replication above 0.5, beside the training image the most similar of them copies. In the first two rows SLCF removes a replication of the uncontrolled model and keeps its composition, and in the last two the baselines create one that neither the uncontrolled model nor SLCF produces.