# THE TOKENS STRIKE BACK: SAFETY-AWARE DENOISING WITH DIRECTIONAL EMBEDDING-BASED REVISION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The Safety-Aware Denoiser (SAD) is a recent, training-free, safety-guidance framework for text diffusion models. During inference, SAD steers text samples toward safe regions of the token probability space. In this paper, we propose *SADDER (Safety-Aware Denoising with Directional Embedding-based Revision)*, which improves the safety and quality trade-offs of SAD by exploiting a safety-guidance signal in the embedding (or activation) space for each token position, then updating the token positions using the SAD update rule multiplied by the guidance signal. This multiplication acts as a gating mechanism: tokens with a high unsafe signal are more likely to be modified during inference following the SAD update rule. However, such an update does not automatically preserve the conditional mean, i.e., the denoiser of a fixed target distribution. We therefore theoretically derive new position-wise target distributions for which the modified denoiser is exact. Furthermore, we derive conditions under which the position-wise target distributions match a global target distribution. We test our *SADDER* method on popular Masked Diffusion Language Models (MDLMs) and Uniform-State Diffusion Models (USDMs) across unsafe continuation and jailbreak settings, and show that SADDER achieves a better balance between safety, generation quality, and inference cost than existing methods.

## 1 INTRODUCTION

Diffusion models have achieved state-of-the-art performance across a wide range of domains, including images (Rombach et al., 2022), speech (Popov et al., 2021), and protein structure generation (Wu et al., 2024). In language modeling, however, autoregressive large language models (AR-LLMs) remain the dominant approach, demonstrating strong capabilities in language generation and understanding, mathematical reasoning (Yang et al., 2024), and coding (Li et al., 2023). Despite their success, AR models generate text one token at a time, which inherently limits inference parallelism and makes revising previously generated content less natural.

Recent advances in Text Diffusion Models (TDMs) offer an alternative paradigm, in which text is generated through an iterative denoising process that can update multiple token positions and progressively refine the generated sequence (Sahoo et al., 2024; Nie et al., 2025). Moreover, at comparable model scales, recent TDMs have also shown competitive performance with AR-LLMs across benchmarks; For instance, Stable-DiffCoder-8B (Fan et al., 2026) vs. Seed-Coder-8B (ByteDance Seed et al., 2025) on code generation benchmarks, and LLaDA-8B (Nie et al., 2025) vs. LLaMA3-8B (Grattafiori et al., 2024) on mathematical reasoning tasks.

Despite these advantages, the safety of TDMs remains relatively underexplored compared to AR-LLMs, particularly with respect to toxic content generation and vulnerability to jailbreak attacks. Traditional safety mechanisms rely on fine-tuning and alignment (Dai et al., 2024), post-hoc filtering and rejection sampling (Naveed et al., 2025), or inference-time methods that modify the decoding strategy during next-token prediction (Xu et al., 2024). However, these approaches either require additional training or assume AR decoding and therefore do not directly generalize to TDMs.

Recent work has begun exploring inference-time mechanisms specifically designed for TDMs, including particle-based steering (Singhal et al., 2025), classifier-based guidance (Schiff et al., 2025),



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.

and safety-oriented remasking approaches (Lee & Han, 2026). In particular, the Safety-Aware Denoiser (SAD) (Yusuf et al., 2026) provides a training-free framework to steer the denoising trajectory towards a safe region in the data distribution using only a small set of $500 \sim 1000$ unsafe reference samples. Nevertheless, SAD operates in token probability space, leaving open whether semantically richer model representations could enable more effective, position-specific safety guidance.

In this work, we propose SADDER (**S**afety-**A**ware **D**enoising with **D**irectional **E**mbedding-based **R**evision), a safety-guidance method that augments SAD with position-wise signals extracted from the model's embedding space. At each denoising step, we first identify potentially unsafe token positions based on a safety-guidance signal in the embedding (or activation) space. We then use this information in a second step to modulate the SAD update, enabling more precise, position-by-position interventions along the sampling trajectory. By selectively modifying fewer positions, SADDER achieves a better balance between safety, generation quality, and inference cost than existing methods, when tested on popular Masked Diffusion Language Models (MDLMs) and Uniform-State Diffusion Models (USDMs) across unsafe continuation and jailbreak settings.

## 2 BACKGROUND

We first describe two ingredients our method builds on: masked and uniform-state TDMs in Section 2.1, whose samplers differ in whether a committed token can still be changed in the subsequent denoising steps; and SAD in Section 2.2, the inference-time guidance framework that our method extends.

### 2.1 TEXT DIFFUSION MODELS

A TDM corrupts a clean token sequence $\boldsymbol{x}_0^{1:L}$ of length $L$ with a fixed forward process and learns to reverse it. Tokens come from a vocabulary $\mathcal{V}$ of size $K$ and are represented as one-hot vectors, so a distribution over tokens is a vector in the simplex $\Delta^K$. $\mathrm{Cat}(\boldsymbol{x} \mid \boldsymbol{\pi})$ denotes the categorical distribution with probability vector $\boldsymbol{\pi}$.

We write $\boldsymbol{x}_t^l$ for the token at position $l$ and normalized diffusion time $t \in [0,1]$. Let $\alpha_t \in [0,1]$ denote a noise schedule that decreases strictly in $t$ with $\alpha_0 \approx 1$ and $\alpha_1 \approx 0$. We discretize the diffusion time interval using a decreasing grid $1 = t_0 > t_1 > \cdots > t_T = 0$ where $T$ is the number of reverse sampling steps. At iteration $j \in \{0, \ldots, T-1\}$, the sampler transitions from $\boldsymbol{x}_{t_j}^{1:L}$ to $\boldsymbol{x}_{t_{j+1}}^{1:L}$.

**Masked diffusion language models (MDLMs).** MDLMs (Sahoo et al., 2024; Nie et al., 2025; Ye et al., 2025) add a mask token with one-hot vector $\mathbb{M}$, so states live in $\{0,1\}^{K+1}$; the network below never predicts the mask, so its outputs stay in $\Delta^K$. In the forward process, each position transitions to the mask independently and, once masked, stays masked:

$$q_t(\boldsymbol{x}_t^l \mid \boldsymbol{x}_0^l) = \mathrm{Cat}\big(\boldsymbol{x}_t^l \mid \alpha_t \boldsymbol{x}_0^l + (1-\alpha_t)\,\mathbb{M}\big), \tag{1}$$

i.e., the token is kept with probability $\alpha_t$ and masked with probability $1-\alpha_t$. The reverse process runs from the all-mask sequence back to $t=0$; its transitions are the forward posterior $q(\boldsymbol{x}_s^l \mid \boldsymbol{x}_t^l, \boldsymbol{x}_0^l)$ with the unknown clean token replaced by the prediction of a denoising network $\boldsymbol{x}_{\boldsymbol{\theta}}$, which reads the whole noised sequence and outputs a distribution over the clean token at every position,

$$\boldsymbol{x}_{\boldsymbol{\theta}} : (\mathcal{V} \cup \{\mathbb{M}\})^L \times [0,1] \longrightarrow (\Delta^K)^L, \tag{2}$$

with $\boldsymbol{x}_{\boldsymbol{\theta},l}(\boldsymbol{x}_t^{1:L}, t) \in \Delta^K$ the output at position $l$. Given $\boldsymbol{x}_t^{1:L}$, the transition to an earlier level $s < t$ factorizes across positions, with

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_s^l \mid \boldsymbol{x}_t^{1:L}) = \begin{cases} \delta_{\boldsymbol{x}_t^l}(\boldsymbol{x}_s^l), & \boldsymbol{x}_t^l \neq \mathbb{M}, \\ \mathrm{Cat}\Big(\boldsymbol{x}_s^l \,\Big|\, \frac{(1-\alpha_s)\,\mathbb{M} + (\alpha_s - \alpha_t)\,\boldsymbol{x}_{\boldsymbol{\theta},l}(\boldsymbol{x}_t^{1:L}, t)}{1-\alpha_t}\Big), & \boldsymbol{x}_t^l = \mathbb{M}. \end{cases} \tag{3}$$

Under equation 3, a revealed position is copied, and a masked position is revealed at random with probability $(\alpha_s - \alpha_t)/(1-\alpha_t)$, its token drawn from $\boldsymbol{x}_{\boldsymbol{\theta},l}$. The samplers of LLaDA and Dream differ from equation 3 only in how they choose which masked positions to reveal (Nie et al., 2025; Ye et al., 2025). *Remasking* samplers (Wang et al., 2025) additionally return revealed positions to $\boldsymbol{m}$ so that they are predicted again.

**Uniform-state diffusion models (USDMs).** USDMs (Sahoo et al., 2025; 2026; DiffusionGemma Team, 2026) go a step further: they have no mask token, and any token can change at any step, which Sahoo et al. (2026) credit with self-correction and better guided generation. In the forward process, a token is kept with probability $\alpha_t$ and otherwise replaced by one drawn uniformly from the vocabulary,

$$q_t(\boldsymbol{x}_t^l \mid \boldsymbol{x}_0^l) = \mathrm{Cat}\big(\boldsymbol{x}_t^l \mid \alpha_t \boldsymbol{x}_0^l + (1-\alpha_t)\,\tfrac{1}{K}\mathbf{1}\big), \tag{4}$$

with $\mathbf{1}$ the all-ones vector. The reverse kernel is again the forward posterior with $\boldsymbol{x}_0^l$ replaced by the prediction; with $\alpha_{t|s} = \alpha_t/\alpha_s$, it is given by

$$\begin{aligned} &q_{s|t}(\boldsymbol{x}_s^l|\boldsymbol{x}_t^l, \boldsymbol{x}_{\boldsymbol{\theta}}^l) \\ &= \mathrm{Cat}\left(\boldsymbol{x}_s^l | \frac{K\alpha_t \boldsymbol{x}_t^l \odot \boldsymbol{x}_{\boldsymbol{\theta}}^l + (\alpha_{t|s}-\alpha_t)\boldsymbol{x}_t^l}{K\alpha_t\langle \boldsymbol{x}_t^l, \boldsymbol{x}_{\boldsymbol{\theta}}^l\rangle + 1 - \alpha_t} + \frac{(\alpha_s-\alpha_t)\boldsymbol{x}_{\boldsymbol{\theta}}^l + (1-\alpha_{t|s})(1-\alpha_s)\mathbf{1}/K}{K\alpha_t\langle \boldsymbol{x}_t^l, \boldsymbol{x}_{\boldsymbol{\theta}}^l\rangle + 1 - \alpha_t}\right), \end{aligned} \tag{5}$$

where $\odot$ denotes the Hadamard product between two vectors and $\boldsymbol{x}_{\boldsymbol{\theta}}^l = \boldsymbol{x}_{\boldsymbol{\theta},l}(\boldsymbol{x}_t^{1:L}, t)$ is given by the pretrained model. Every entry of this vector can be positive, which is what lets any token change at any step.

### 2.2 The Safety-Aware Denoiser

The safety-aware denoiser (SAD) (Yusuf et al., 2026) is a training-free guidance method that steers a pretrained TDM away from a *negation set* of sequences that should not be generated. It adapts the safe denoiser of Kim et al. (2025), later extended to flow matching (Kim et al., 2026), by replacing the Gaussian corruption kernel with the categorical kernels equation 1 and equation 4.

**Safe Denoiser.** Assume the training data $\mathcal{D}$ split into safe and unsafe sets, $\mathcal{D} = \mathcal{D}_{\text{safe}} \cup \mathcal{D}_{\text{unsafe}}$ with $|\mathcal{D}_{\text{safe}}| \gg |\mathcal{D}_{\text{unsafe}}|$. Restricting $p_{\text{data}}$ to each set gives $p_{\text{safe}}$ and $p_{\text{unsafe}}$, and running the forward process from $p_{\text{data}}$, $p_{\text{safe}}$, and $p_{\text{unsafe}}$ defines the data, safe, and unsafe denoisers $\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$, $\mathbb{E}_{\mathcal{D}_{\text{safe}}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$, and $\mathbb{E}_{\mathcal{D}_{\text{unsafe}}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$, which are related as follows.

*Theorem* 2.1 (Theorem 3.2 in Kim et al. (2025)). There exists a nonnegative weight $\beta^*(\boldsymbol{x}_t)$, monotone in the posterior likelihood that $\boldsymbol{x}_t$ originates from the unsafe set, such that

$$\mathbb{E}_{\mathcal{D}_{\text{safe}}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] + \beta^*(\boldsymbol{x}_t)\left(\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] - \mathbb{E}_{\mathcal{D}_{\text{unsafe}}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]\right). \tag{6}$$

Thm. 2.1 states that by subtracting "unsafe" components from the data denoiser, with $\beta^*$ adapting to how unsafe the current state is, we can provably generate samples from the safe denoiser. A pretrained model provides the data denoiser, so only the unsafe denoiser and $\beta^*$ remain.

**SAD.** For a TDM, the data denoiser is simply the pre-trained model's output, $\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}] = \boldsymbol{x}_{\boldsymbol{\theta},l}(\boldsymbol{x}_t^{1:L}, t)$, since the mean of a categorical variable is its probability vector; the relation is exact for the ideal denoiser, which $\boldsymbol{x}_{\boldsymbol{\theta}}$ approximates. The unsafe denoiser is approximated by Monte Carlo over a negation set, $\mathcal{D}_{\text{unsafe}} = \{\boldsymbol{x}_0^{(n)}\}_{n=1}^N$: $\hat{\mathbb{E}}_{\mathcal{D}_{\text{unsafe}}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \sum_{n=1}^N \boldsymbol{x}_0^{(n)} \frac{q_t(\boldsymbol{x}_t|\boldsymbol{x}_0^{(n)})}{\sum_{m=1}^N q_t(\boldsymbol{x}_t|\boldsymbol{x}_0^{(m)})}$, where $q_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0^{(n)}) = \prod_{l=1}^L q_t(\boldsymbol{x}_t^l \mid \boldsymbol{x}_0^{(n),l})$ is the sequence kernel from equation 1 or equation 4, a product because positions are noised independently. The precise definition of the weight is $\beta^*(\boldsymbol{x}_t) = Z_{\text{unsafe}}\, p_{\text{unsafe},t}(\boldsymbol{x}_t)/\big(Z_{\text{safe}}\, p_{\text{safe},t}(\boldsymbol{x}_t)\big)$ (Kim et al., 2025, Appendix A), where $p_{\text{safe},t}$ and $p_{\text{unsafe},t}$ are the marginal distributions of the forward process at level $t$ started from the safe and the unsafe distribution, and $Z_{\text{safe}}$ and $Z_{\text{unsafe}}$ are the normalizing constants of the two restricted densities.

Unfortunately, directly computing $\beta^*(\boldsymbol{x}_t)$ is *intractable*: the denominator requires $q_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ over all safe data, roughly most of the training set, at every sampling step. SAD approximates

$$\beta^*(\boldsymbol{x}_t) \approx \eta\, \beta(\boldsymbol{x}_t),$$

using an unbiased estimate of $\beta(\boldsymbol{x}_t)$ and a hyperparameter $\eta$ to replace the remaining terms of $\beta^*$.

$$\beta(\boldsymbol{x}_t) = \int p_{\text{unsafe}}(\boldsymbol{x}_0)\, q_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0)\, \mathrm{d}\boldsymbol{x}_0 \approx \frac{1}{N}\sum_{n=1}^N q_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0^{(n)}). \tag{7}$$

Because SAD approximates $\beta^*$ this way, the SAD implementation offers *no theoretical safety guarantees* on generated data, unlike what the theory in Thm. 2.1 promises.

Prior work shows that such guidance must be strong early in sampling and fade later, or it degrades sample quality (Kim et al., 2025; 2026; Kirchhof et al., 2025); SAD therefore keeps $\eta$ at a nonzero constant over a window $C$ of critical timesteps at the start of the reverse process and sets $\eta = 0$ outside it.

# 3 METHODS: SADDER

SAD provides safety guidance in the token probability space. We can enhance this guidance by selecting a guidance signal in the embedding (or activation) space for each token position, then modifying the selected token positions using the SAD update rule. In what follows, we introduce SADDER (Safety-Aware Denoising with Directional Embedding-based Revision), consisting of two phases.

## 3.1 FIRST PHASE: SELECTING RISKY POSITIONS.

The goal of the first phase is to create a gating mechanism which tells us which token positions need to be modified. Inspired by (Lee & Han, 2026), we propose accessing TDM's internal embedding space to obtain this information.

For each sample $i$, we denote $\boldsymbol{c}_i$ as a given prompt and $\boldsymbol{x}_i^{\text{harm}}$, $\boldsymbol{x}_i^{\text{safe}}$ as two corresponding harmful and safe responses for $\boldsymbol{c}_i$, respectively. We denote $P_i = |\boldsymbol{c}_i|$ as the prompt length and $R_i^{\text{harm}} = \min(|\boldsymbol{x}_i^{harm}|, M)$ the response length after truncation of $M$ tokens. Similarly, $R_i^{\text{safe}} = \min(|\boldsymbol{x}_i^{safe}|, M)$. Let $\boldsymbol{h}_{k,l}(\cdot) \in \mathbb{R}^d$ be the layer-$k$ hidden representation at position $l$. $\boldsymbol{h}_{k,l}^{(t)}(\cdot)$ denotes the layer-$k$'s hidden state for position $l$ at timestep $t$. We extract hidden states and compute the average representations over response tokens:

$$\bar{\boldsymbol{h}}_{i,k}^{\text{harm}} = \frac{1}{R_i^{\text{harm}}} \sum_{l=P_i+1}^{P_i+R_i^{\text{harm}}} \boldsymbol{h}_{k,l}(\boldsymbol{c}_i \oplus \boldsymbol{x}_i^{\text{harm}}), \quad \bar{\boldsymbol{h}}_{i,k}^{\text{safe}} = \frac{1}{R_i^{\text{safe}}} \sum_{l=P_i+1}^{P_i+R_i^{\text{safe}}} \boldsymbol{h}_{k,l}(\boldsymbol{c}_i \oplus \boldsymbol{x}_i^{\text{safe}}) \tag{8}$$

We arrive at the (normalized) layer-wise difference in the semantic representations of harmful and safe responses, $\hat{\boldsymbol{v}}_k$, by averaging over the examples $i$:

$$\widehat{\boldsymbol{v}}_k = \frac{\bar{\boldsymbol{h}}_k^{\text{harm}} - \bar{\boldsymbol{h}}_k^{\text{safe}}}{\|\bar{\boldsymbol{h}}_k^{\text{harm}} - \bar{\boldsymbol{h}}_k^{\text{safe}}\|_2}, \tag{9}$$

where $\bar{\boldsymbol{h}}_k^{\text{harm}} = \frac{1}{N_{\text{harm}}} \sum_i \bar{\boldsymbol{h}}_{i,k}^{\text{harm}}$, and $\bar{\boldsymbol{h}}_k^{\text{safe}} = \frac{1}{N_{\text{safe}}} \sum_i \bar{\boldsymbol{h}}_{i,k}^{\text{safe}}$. Intuitively, we want to intervene on the token location only when the hidden state at time $t$ is *closer to* the average harmful representations than to the average safe representations:

$$\|\boldsymbol{h}_{k,l}^{(t)} - \bar{\boldsymbol{h}}_k^{\text{harm}}\|_2^2 < \|\boldsymbol{h}_{k,l}^{(t)} - \bar{\boldsymbol{h}}_k^{\text{safe}}\|_2^2. \tag{10}$$

Using a midpoint $\boldsymbol{m}_k = \frac{\bar{\boldsymbol{h}}_k^{\text{harm}} + \bar{\boldsymbol{h}}_k^{\text{safe}}}{2}$, a simple linear algebraic manipulation gives us the following equivalence:

$$\|\boldsymbol{h}_{k,l}^{(t)} - \bar{\boldsymbol{h}}_k^{\text{harm}}\|_2^2 < \|\boldsymbol{h}_{k,l}^{(t)} - \bar{\boldsymbol{h}}_k^{\text{safe}}\|_2^2 \iff \left\langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k, \hat{\boldsymbol{v}}_k \right\rangle > 0 \tag{11}$$

See Section A.2 for the derivation. Unlike (Lee & Han, 2026), we use the *centred* score to decide whether this particular position will lead to harmful generation:

$$s_{k,l}^{(t)} := \left\langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k,\ \hat{\boldsymbol{v}}_k \right\rangle. \tag{12}$$

*Geometric meaning of equation 12*: The set of $\boldsymbol{h}_{k,l}^{(t)}$ that satisfies $\left\langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k, \hat{\boldsymbol{v}}_k \right\rangle = 0$ is the hyperplane, in which every vector $\boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k$ is orthogonal to the direction $\hat{\boldsymbol{v}}_k$. Any $\boldsymbol{h}_{k,l}^{(t)}$ that satisfies $\left\langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k, \hat{\boldsymbol{v}}_k \right\rangle > 0$ lies on the side of this hyperplane containing $\bar{\boldsymbol{h}}_k^{\text{harm}}$. On the contrary, any $\boldsymbol{h}_{k,l}^{(t)}$ that satisfies $\left\langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k, \hat{\boldsymbol{v}}_k \right\rangle < 0$ lies on the side of this hyperplane containing $\bar{\boldsymbol{h}}_k^{\text{safe}}$.

**Algorithm 1** SADDER: selective SAD guidance.

**Input:** denoiser $p_\theta$; unsafe dataset $\mathcal{D}_{\text{unsafe}}$; $\widehat{\boldsymbol{v}}_k, \boldsymbol{m}_k$; initial state $\boldsymbol{x}_{t_0}$; editable positions $\mathcal{G}$; time steps $t_0 > \cdots > t_T = 0$; intervention steps $C$; strength $\eta$.
**for** $j = 0, \ldots, T-1$ **do**
  Set $t \leftarrow t_j$. Obtain $p_{\mathcal{D},t,\cdot}$ and $\boldsymbol{h}_{k,\cdot}^{(t)}$ from one model evaluation on $\boldsymbol{x}_t$.
  Initialize $\widetilde{p}_{0|t,l} \leftarrow p_{\mathcal{D},t,l}$ and $a_{k,l}^{(t)} \leftarrow 0$ for all $l$.
  **if** $j \in C$ **then**
    **if** MDLMs **then** $\mathcal{M}_t \leftarrow \{l \in \mathcal{G} : \boldsymbol{x}_t^l = \mathbb{M}\}$
    **if** USDMs **then** $\mathcal{M}_t \leftarrow \{1, \ldots, L\}$ .
    Compute $s_{k,l}^{(t)} \leftarrow \langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k, \widehat{\boldsymbol{v}}_k \rangle$ for $l \in \mathcal{M}_t$.
    **Select Risky Positions:** $a_{k,l}^{(t)} \leftarrow \mathbb{1}\{s_{k,l}^{(t)} > 0\}$ for $l \in \mathcal{M}_t$.
    **if** at least one gate equals 1 **then**
      Estimate $\mathbb{E}_{\mathcal{D}_{\text{unsafe}}}[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}]$ and $\beta(\boldsymbol{x}_t)$ on the current state
      **SAD On Selected Positions:** Compute Sadder Denoiser in equation 14.
    **end if**
  **end if**
  $\boldsymbol{x}_{t_{j+1}} \leftarrow \text{Solver}(\boldsymbol{x}_t, t_j, t_{j+1}, \text{Sadder Denoiser}; \mathcal{G})$.
**end for**
**Return:** the generated continuation of $\boldsymbol{x}_{t_T}$.

The tokens that have positive scores, through the following *gating* mechanism:

$$a_{k,l}^{(t)} \leftarrow \mathbb{1}_{s_{k,l}^{(t)} > 0}, \tag{13}$$

where $\mathbb{1}$ is an indicator function that returns the binary[1] gating values: $a_{k,l}^{(t)} \in \{0, 1\}$.

### 3.2 Second Phase: SAD on Selected Positions

In the first phase, we select which position is aligned with the harmful direction in the embedding space. In the second phase, we modify the sampling trajectory of the selected ones using the SAD denoiser formulation, given by

$$\text{Sadder Denoiser} := \prod_{l=1}^{L} \Big[ \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}] + \eta\, \beta(\boldsymbol{x}_t) a_{k,l}^{(t)} \left( \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}] - \mathbb{E}_{\mathcal{D}_{\text{unsafe}}}[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}] \right) \Big] \tag{14}$$

where $\beta(\boldsymbol{x}_t)$ is defined in equation 7 and $\eta$ is a controllable hyperparameter that changes the strength of the guidance as in SAD. Similar to SAD, the data denoiser at position $l$ is $\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}]$ and the unsafe denoiser at position $l$ is approximated by

$$\mathbb{E}_{\mathcal{D}_{\text{unsafe}}}[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}] \approx \sum_{\boldsymbol{x}_n \in \mathcal{D}_{\text{unsafe}}} \boldsymbol{x}_n^l \, \frac{q_t(\boldsymbol{x}_t | \boldsymbol{x}_n^l)}{\sum_{\boldsymbol{x}_m \in \mathcal{D}_{\text{unsafe}}} q_t(\boldsymbol{x}_t | \boldsymbol{x}_m^l)}. \tag{15}$$

Algorithm 1 summarizes our algorithm.

### 3.3 Theoretical analyses of SADDER

In Safe Denoiser, equation 6 provides an identity of conditional expectations under a particular safe/unsafe partition, while, in the implementation of SAD, this quantity is approximated by a well-tuned hyperparameter $\eta$. A freely chosen position-dependent gate at $l$ in SADDER also changes that identity, as in SAD. In the following theorem, we formally derive *under what condition* this new target denoiser is a position-wise safe denoiser.

[1] One can also use the softened score for more refined guidance across positions. However, this approach did not empirically outperform simple gating in equation 13. See Appendix J.2 for comparisoin.

*Theorem* 3.1 (**Characterization of an exact position-wise safe denoiser**). Given a data denoiser from a pre-trained diffusion model, an unsafe denoiser in equation 15, and the intervention strength: $\eta\beta(\boldsymbol{x}_t)a_{k,l}^{(t)}$, the the position-$l$ factor of the SADDER Denoiser in equation 14 is exactly the conditional mean under the position-wise safe target distribution $\mathcal{D}_{\text{safe}}^{(l)}$:

$$\boxed{\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}\left[\boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}\right] = \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}] + \eta\beta(\boldsymbol{x}_t)a_{k,l}^{(t)}\left(\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}] - \mathbb{E}_{\mathcal{D}_{\text{unsafe}}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}]\right)} \tag{16}$$

if the activation strength is calibrated to the posterior unsafe odds according to

$$\eta\beta(\boldsymbol{x}_t)a_{k,l}^{(t)} = \frac{\pi_{t,l}(\boldsymbol{x}_t^{1:L})}{1-\pi_{t,l}(\boldsymbol{x}_t^{1:L})}, \tag{17}$$

where the posterior probability of the position-$l$ being unsafe is denoted by $\pi_{t,l}(\boldsymbol{x}_t^{1:L}) := \mathbb{E}_{\mathcal{D}}\left[\psi_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}\right]$, and the unsafety is measured by an oracle function denoted by $\psi_l : \mathcal{V}^L \longrightarrow [0,1]$, satisfying $\psi_l(\boldsymbol{x}_0^{1:L}) + (1-\psi_l(\boldsymbol{x}_0^{1:L})) = 1$ for any clean sequence $\boldsymbol{x}_0^{1:L}$.

See Section A.4 for proof.

*Remark* 1. Thm. 3.1 characterizes our algorithm, SADDER, in terms of a position-wise safe-denoiser guarantee, rather than asserting that the exact membership function $\psi$ is available at inference. However, it identifies the quantity that the activation gate should approximate: the posterior expected unsafe membership of clean sequences compatible with the current noised state. In practice, the centred activation score is a surrogate for this quantity without formal guarantees, like $\beta^*(\boldsymbol{x}_t)$ was approximated by $\eta\beta(\boldsymbol{x}_t)$ in SAD; the resulting SADDER denoiser is also *approximate*.

The position-specific target distribution $p_{\mathcal{D}_{\text{safe}}^{(l)}}$ for each position $l$, stated in Thm. 3.1, may differ for each $l$: $p_{\mathcal{D}_{\text{safe}}^{(1)}}, \cdots, p_{\mathcal{D}_{\text{safe}}^{(L)}}$ Hence, a natural next question is whether there exists one clean-sequence distribution $p_{\mathcal{D}_{\text{safe}}^{\text{global}}}(\boldsymbol{x}_0^{1:L})$, such that simultaneously for every $l$, it satisfies $\mathbb{E}_{\mathcal{D}_{\text{safe}}}^{global}(\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}) = \mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}(\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L})$. Thm. A.1 states this formally.

## 4 Related Work

**Activation steering and remasking.** Contrastive Activation Addition (Rimsky et al., 2024) steers language models using directions extracted from paired examples; ILRR (Avrahami & Nachmani, 2026) guides masked diffusion by aligning activations with a reference sequence. For safety, DiffuGuard (Li et al., 2025) combines stochastic remasking with block-level audit and repair. Most closely related, Lee & Han (2026) construct paired samples $(x_i, y_i^{\text{harm}}, y_i^{\text{safe}})$, using harmful model responses to WildJailbreak prompts (Jiang et al., 2024) and paraphrased refusals. They estimate a *Contrastive Safety Direction* (CSD), $v_\ell$, from the difference between the mean harmful and safe response representations. In *Phase 1*, they modify hidden states at masked and prompt positions during the first fraction $\rho$ of the denoising steps: $\mathcal{S}(h_{\ell,j}^{(t)}) = h_{\ell,j}^{(t)} - \beta\langle h_{\ell,j}^{(t)}, \hat{v}_\ell\rangle\hat{v}_\ell$, where $\hat{v}_\ell$ is the normalized CSD and $\beta$ controls the intervention strength. In *Phase 2*, they refine the generated sequence by selecting positions according to $m_j^{(k)} = \mathbb{1}\left\{\langle h_{\ell,j}^{(k)}, \hat{v}_\ell\rangle > \theta\right\}$, then remasking and regenerating those positions for up to $K$ refinement iterations, stopping early if no positions exceed the threshold. SADDER uses a similar contrastive direction, but centers the activation score at the midpoint of the harmful and safe representations. This places the zero-score boundary at their Euclidean bisector. The score gates SAD's correction in token probability space rather than directly modifying hidden states; in masked models, it also ranks revealed positions for budgeted remasking within the denoising loop.

**Sampling-based baselines.** Best-of-$N$ sampling generates independent candidate completions and selects the one with the highest reward. FK steering (Singhal et al., 2025) instead maintains multiple diffusion trajectories and resamples them at intermediate steps using reward-derived potentials. Both provide inference-time control without retraining the generator. SADDER modifies token predictions within each trajectory, using unsafe references and activation scores rather than ranking candidate sequences. Comparisons with Best-of-$N$ and FK steering assess whether this local intervention offers a better safety–quality tradeoff at a given inference budget.

Table 1: LLaDA-Base safety, generation quality, and cost. Entries use $A_b$, where $A$ is the mean and $b$ is one sample standard deviation across three generation seeds. For percentages, SDs are in percentage points. NFE shows average cost.

| Method | RTP | | | | | BeaverTails | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Unsafe % ↓ | Harmful PPL ↓ | Benign PPL ↓ | Benign BERT ↑ | NFE | Unsafe % ↓ | Harmful PPL ↓ | Benign PPL ↓ | Benign BERT ↑ | NFE |
| Unmodified | $38.6_{1.3}$ | $21.1_{0.15}$ | $17.5_{0.35}$ | $84.65_{0.04}$ | 32.0 | $38.6_{1.5}$ | $20.6_{0.62}$ | $17.5_{0.35}$ | $84.65_{0.04}$ | 32.0 |
| Full SAD | $32.4_{1.6}$ | $29.3_{0.52}$ | $30.0_{0.75}$ | $83.92_{0.01}$ | 32.0 | $27.1_{1.2}$ | $35.1_{0.34}$ | $47.1_{1.1}$ | $83.38_{0.01}$ | 32.0 |
| SADDER | $29.4_{1.7}$ | $42.1_{1.4}$ | $18.4_{0.26}$ | $84.59_{0.06}$ | 32.0 | $29.9_{0.9}$ | $34.6_{0.54}$ | $27.8_{0.86}$ | $84.20_{0.03}$ | 32.0 |
| SADDER + warm-up | $31.2_{0.7}$ | $31.6_{2.1}$ | $44.3_{0.75}$ | $83.01_{0.01}$ | 32.0 | $24.4_{0.9}$ | $36.9_{0.32}$ | $44.1_{0.14}$ | $83.18_{0.03}$ | 32.0 |
| CSD† | $22.4_{0.4}$ | $225_{7.0}$ | $21.6_{0.26}$ | $84.59_{0.04}$ | 61.9 | $15.6_{0.5}$ | $487_{36}$ | $943_{40}$ | $82.61_{0.03}$ | 55.0 |
| BoN-8 | $19.0_{0.7}$ | $20.5_{0.26}$ | $19.5_{0.48}$ | $84.70_{0.00}$ | 256.0 | $18.6_{0.6}$ | $24.2_{0.32}$ | $19.5_{0.48}$ | $84.70_{0.00}$ | 256.0 |
| FK-8 | $20.6_{0.8}$ | $21.3_{1.1}$ | $18.2_{0.53}$ | $84.71_{0.04}$ | 256.0 | $28.2_{1.2}$ | $22.7_{0.93}$ | $18.2_{0.53}$ | $84.71_{0.04}$ | 256.0 |

**Diffusion-specific jailbreaks.** DIJA (Wen et al., 2026) constructs interleaved text–mask prompts that exploit bidirectional context and parallel decoding to elicit harmful completions. PAD (Zhang et al., 2025) adapts affirmative-response priming to parallel generation through a multi-point attention attack. These attacks expose vulnerabilities associated with diffusion's generation process and motivate evaluating safety guidance on adversarially structured inputs as well as ordinary unsafe continuation. We include all in our evaluation of selective guidance and token revision.

## 5 EXPERIMENTS

We ask whether selective guidance improves the safety–quality tradeoff of SAD, and whether its benefits transfer from unsafe continuation to adversarial instructions. Following Yusuf et al. (2026), we evaluate these tasks separately. The main continuation results distinguish SADDER from SADDER with a full-SAD warm-up. Other variants, calibration sweeps, and the extended algorithm are reported in Appendix E.

### 5.1 UNSAFE CONTINUATION ON MASKED DIFFUSION LANGUAGE MODELS (MDLMS)

**Setting and metrics.** We evaluate LLaDA-8B-Base Nie et al. (2025) and Dream-7B-Base Ye et al. (2025) on 600 unsafe prompts per dataset from RealToxicityPrompts (RTP) Gehman et al. (2020) and BeaverTails (BT) Ji et al. (2023), generating 32 tokens in 32 steps. Baselines are unmodified sampling, full SAD, adapted CSD† (Lee & Han, 2026), best-of-eight (BoN-8), and eight-particle FK steering (Singhal et al., 2025). *SADDER + warm-up* applies full SAD on steps 0–7 and selective guidance on steps 8–15. Evaluation prompts were held out from development; later hyperparameter selection uses a separate 300-prompt development set (Appendix E).

Safety is the Llama-Guard-3-8B Meta (2024) unsafe rate. *Harmful PPL* is GPT-2-large (Radford et al., 2019) corpus perplexity over all outputs to harmful prompts, including safe outputs. *Benign BERT* (BERTScore F1 $\times 100$) and *Benign PPL* use 600 benign RTP prompt–continuation pairs, retaining each method's safety-selected settings and reference set. NFE counts generator forward evaluations, excluding reward scoring and one-time reference encoding.

**Safety–quality tradeoffs.** On LLaDA/RTP, SADDER lowers the unsafe rate from 38.6% unmodified to 29.4%, versus 32.4% for full SAD; warm-up reaches 31.2%. On BeaverTails, SADDER reaches 29.9%, full SAD 27.1%, and warm-up 24.4%. Across all four model/dataset settings, SADDER achieves higher benign BERTScore and lower benign PPL than full SAD. SADDER also achieves higher benign BERTScore than warm-up throughout, although both variants remain below unmodified sampling on this metric (Tables 1 and 9). Harmful-prompt PPL reveals a different tradeoff: on LLaDA/RTP, CSD reaches a 22.4% unsafe rate but PPL 225, versus 42.1 for SADDER and 31.6 for warm-up without remasking. CSD's benign PPL of 21.6 does not expose this deterioration on harmful prompts. Gains also depend on the model: on Dream, full SAD at $\eta = 10{,}000$ achieves lower unsafe rates than either SADDER variant on both datasets. BoN-8 further reduces unsafe rates at a cost of 256 generator evaluations plus reward scoring.

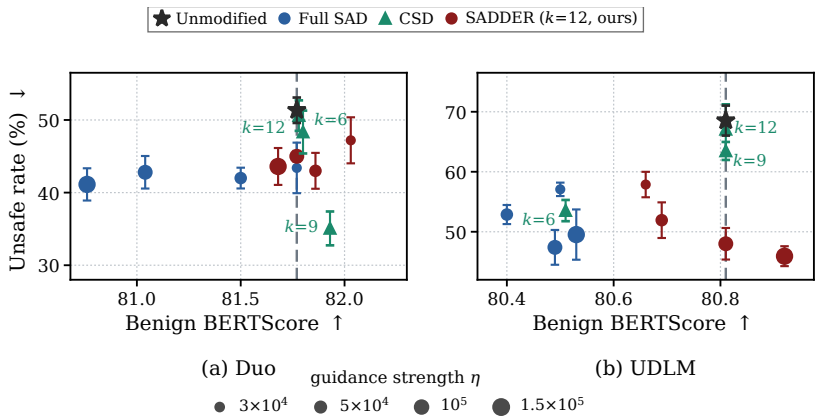


Figure 1: Safety–quality tradeoff at equal NFE on Duo and UDLM. Best-of-$N$ is reported in Appendix I. Larger markers denote larger guidance strength $\eta$. The dashed vertical line marks the unmodified model's benign BERTScore, and error bars are one standard deviation across seeds.

### 5.2 Unsafe Continuation on Uniform-State Diffusion Models

**Setting and metrics.** We evaluate Duo (Sahoo et al., 2025) and UDLM (Schiff et al., 2025) on 300 unsafe prompt records from RealToxicityPrompts (RTP). Both models generate at a context length of 128 in 128 denoising steps. Unsafe references and contrastive directions follow Section 5.1, with $k \in \{6, 9, 12\}$. Baselines are unmodified sampling, full SAD, CSD$^\dagger$, and best-of-$N$ for $N \in \{3, 5, 8\}$. Full SAD and SADDER sweep $\eta \in \{3\times10^4, 5\times10^4, 10^5, 1.5\times10^5\}$. In masked models, guidance must reach a position before it is revealed. Uniform-state samplers have no such constraint, so we gate every position at every step, using Equation 13 for SADDER and no gate for full SAD.

Safety is the Llama-Guard-3-8B unsafe rate. Benign BERT uses 300 separate benign RTP prompt–continuation pairs. All entries are means over five generation seeds, and Appendix I reports exact values with seed standard deviations.

**Safety–quality tradeoffs.** Figure 1 reports the tradeoff. On UDLM, SADDER reaches a 45.9% unsafe rate, against 68.5% without steering and 47.4% for the lowest unsafe rate among full SAD settings. SADDER's benign BERTScore is 80.92, against 80.49 for full SAD and 80.81 unmodified, so it is safer than full SAD without any loss in reference similarity. On Duo, full SAD reaches lower unsafe rates than SADDER at every $\eta$, from 43.4% versus 47.2% at $\eta = 3\times10^4$ to 41.1% versus 43.6% at $\eta = 1.5\times10^5$. The same holds at matched benign BERTScore: full SAD reaches 43.4% at 81.77, where SADDER reaches 45.0%. What SADDER does buy is quality, on both models: at every $\eta$ its benign BERTScore is higher than full SAD's. On Duo the gap is widest at the end of the sweep, 81.68 against 80.76.

SADDER, full SAD, and CSD$^\dagger$ use the same number of generator evaluations, whereas best-of-$N$ uses $N$ times as many. Neither baseline is consistent across the two models. On Duo, best-of-8 reaches 33.1%, the lowest unsafe rate we observe, yet on UDLM best-of-$N$ does not improve on unmodified sampling at any budget. CSD$^\dagger$ reaches 35.1% on Duo at $k = 9$ but 63.5% on UDLM at the same layer, and it is sensitive to the layer choice: across $k$, its unsafe rate on Duo spans 15.5 points, whereas SADDER's varies by at most 2.2 points at fixed $\eta$.

### 5.3 Jailbreak Robustness on Instruction-Tuned Models

**Setting.** We evaluate DIJA and PAD on LLaDA-8B-Instruct and Dream-7B-Instruct using 100 held-out HarmBench behaviors, 128 sampling steps, and WildJailbreak Jiang et al. (2024) unsafe references. CSD and SADDER share the unsafe-response set. DiffuGuard (Li et al., 2025) is evaluated with *full* repair, which can modify fixed attack text, and *gen* repair, restricted to initially masked positions; their audit thresholds also differ (Appendices H and A.6).

Table 2: Dream-Instruct: held-out HarmBench attacks and XSTest over-refusal. Entries use $A_b$: mean and sample SD across three seeds (percentage points for rates). HB, LG, and PPL assess complete responses. XSTest reports full-response WG refusal on safe prompts, using valid judge labels only (coverage in Appendix H.2). NFE is mean cost.

| Method | DIJA | | | | PAD | | | | XSTest |
|---|---|---|---|---|---|---|---|---|---|
| | HB ASR % ↓ | LG unsafe % ↓ | Harmful PPL ↓ | NFE | HB ASR % ↓ | LG unsafe % ↓ | Harmful PPL ↓ | NFE | WG refusal % ↓ |
| **Dream-Instruct** | | | | | | | | | |
| Unmodified | $82.0_{1.0}$ | $80.7_{1.5}$ | $11.4_{0.18}$ | 91.7 | $51.7_{0.6}$ | $64.0_{3.6}$ | $8.53_{0.027}$ | 116.0 | $55.5_{1.3}$ |
| Full SAD | $85.3_{0.6}$ | $83.0_{1.0}$ | $11.8_{0.037}$ | 91.7 | $52.7_{4.0}$ | $63.3_{1.5}$ | $9.52_{0.21}$ | 116.0 | $58.0_{1.0}$ |
| SADDER (selective) | $77.7_{1.5}$ | $81.7_{3.2}$ | $12.2_{0.093}$ | 91.7 | $52.7_{0.6}$ | $66.3_{2.5}$ | $9.58_{0.15}$ | 116.0 | $41.1_{0.8}$ |
| SADDER + warm-up | $78.7_{2.3}$ | $79.7_{1.2}$ | $12.2_{0.10}$ | 91.7 | $42.7_{3.1}$ | $55.0_{2.0}$ | $10.4_{0.17}$ | 116.0 | $17.9_{0.7}$ |
| CSD† | $68.0_{6.2}$ | $82.3_{0.6}$ | $45.9_{4.1}$ | 115.5 | $33.3_{3.1}$ | $63.3_{2.5}$ | $23.7_{1.5}$ | 140.8 | $56.4_{1.9}$ |
| DiffuGuard (full) | $33.3_{2.5}$ | $84.0_{2.0}$ | $537_{11}$ | 137.0 | $12.0_{1.7}$ | $70.0_{2.6}$ | $687_{120}$ | 125.0 | $64.1_{1.5}$ |
| DiffuGuard (gen) | $69.3_{1.5}$ | $81.3_{1.5}$ | $46.6_{1.9}$ | 97.2 | $49.0_{1.7}$ | $63.3_{2.3}$ | $8.80_{0.63}$ | 117.0 | $54.6_{1.1}$ |

Table 2 reports full-response HarmBench (HB) ASR, LlamaGuard (LG) unsafe rate, harmful PPL, and NFE. XSTest over-refusal uses WildGuard (WG) (Han et al., 2024) judgments on 250 safe, unmodified prompts across three seeds, with the DIJA guidance configuration. WG rates use valid labels (non-empty) within each seed, then average across seeds. Appendices H.5 and H.2 report generated-span judgments, valid-score counts, paired comparisons, and calibration-prompt reuse.

**Attack and evaluator dependence.** Under DIJA, SADDER and warm-up leave HB ASR near unmodified sampling, while LG flags roughly four-fifths of responses as unsafe. Large HB reductions from CSD and full-block DiffuGuard do not yield comparable LG reductions: Dream's full-block DiffuGuard reaches 33.3% HB ASR but 84.0% LG unsafe rate and PPL 537. Both judges assess complete responses, so this discrepancy cannot be attributed solely to scaffold inclusion. Appendix H.5 separates evaluator choice from the influence of fixed attack text by comparing full-response and generated-span judgments on the same outputs. SADDER retains generator cost and incurs a smaller PPL increase, but offers limited DIJA robustness.

PAD gives a more favorable result: Dream warm-up reduces HB ASR by 9.0 points (paired 95% CI: $[-15.7, -2.3]$), lowers LG unsafe rate from 64.0% to 55.0%, and increases PPL from 8.53 to 10.40. This supports an attack-specific tradeoff rather than general jailbreak robustness. Appendix H reports component, strength, and timing studies.

**Over-refusal.** Dream warm-up reduces WG refusal on safe XSTest prompts by 35.4 points on jointly scored outputs (95% CI: $[-41.6, -29.3]$), but raises WG harmfulness on unsafe contrasts from 0.0% to 6.9% among scored responses. Fewer refusals therefore do not establish greater helpfulness at unchanged safety.

## 6 Summary and Discussion

We introduced SADDER, an extension of the Safety-Aware Denoiser, using midpoint-centred contrastive activation scores to select where to apply probability-space safety guidance. Rather than directly modifying hidden representations, SADDER uses them to identify potentially risky positions and applies the SAD correction to the corresponding token distributions. Our theoretical analysis characterizes the conditions under which position-wise guidance recovers the conditional mean of an induced safe target, and identifies when these local conditional means agree with a global target. Experiments on masked and uniform-state diffusion models reveal both the benefits and limitations of selective guidance. SADDER improves benign-continuation quality metrics relative to full SAD in several settings, with selected configurations also reducing unsafe generation. Jailbreak improvements also depend on the model and operating point, but SADDER shows significantly lower refusal rates. Taken together, SADDER offers an additional way to manage the safety–quality and safety-refusal trade-offs.

### AI USE STATEMENT

We used generative AI tools, to assist with refining research hypotheses and experimental designs, reviewing mathematical derivations and proof arguments, analyzing and interpreting experimental results, and implementing and debugging experimental code. We also used these tools to draft and revise portions of the manuscript, improve clarity and organization, prepare LaTeX pseudocode, and identify and discuss relevant literature. The models used for text generation and automated safety evaluation within our experiments are described in the experimental protocol.

The authors reviewed and revised the AI-assisted material included in this work. We checked mathematical statements and derivations, verified references against the cited sources, and checked reported numerical claims against experimental outputs. AI-generated code was reviewed and revised by the authors before use in the experiments. The authors made the final methodological decisions and determined the interpretation and presentation of the findings. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

### ETHICS STATEMENT

This work investigates training-free safety guidance for text diffusion models. Our experiments involve toxic text, harmful instructions, and adversarial prompts from existing datasets and benchmarks. These materials serve a defensive evaluation purpose but may contain offensive or distressing content. Analysis of model vulnerabilities could also inform attempts to circumvent safeguards. Any release of associated harmful generations or attack artifacts should therefore consider risks of misuse, exposure to harmful content, and disclosure of sensitive information. Furthermore, safety guidance could be misused to suppress legitimate expression, while overly broad or biased safety criteria may restrict benign content and disproportionately affect particular communities.

Our evaluation relies on benchmark datasets and automated safety judges, which may reflect social biases and fail to capture context-dependent harms. Differences between evaluators further limit the interpretation of reported safety metrics. We report safety alongside generation quality, inference cost, and over-refusal, including settings where the proposed guidance offers limited benefit. Reduced refusal does not necessarily imply improved helpfulness, and lower measured attack success does not establish that responses are safe or usable. Our findings are restricted to the evaluated models, datasets, and sampling configurations; they do not establish safety across untested settings. Deployment would require additional assessment of robustness, fairness, and risks specific to the intended application.

### REPRODUCIBILITY STATEMENT

The main text presents the mathematical formulation of SADDER, and Algorithm 1 specifies the selective guidance procedure. Appendix A.2 derives the centered-score equivalence, and Appendices A.4 and A.5 provide proofs of the theoretical results. Appendix A.6 describes the datasets, prompt selection, and reference-set construction. Appendix A.8 documents shared settings, hyperparameter ranges, and configuration selection, while Appendix B details the sampling extensions. Appendices E and H report additional experimental settings and results. Appendix H.5 compares evaluators and response scopes, and Appendix H.2 describes refusal evaluation and the handling of missing judge labels. Appendix B.2 specifies metric aggregation, seed standard deviations, bootstrap intervals, and paired comparisons. These materials support independent implementation and assessment of the reported results.

## REFERENCES

Eden Avrahami and Eliya Nachmani. ILRR: Inference-time steering method for masked diffusion language models, 2026. URL https://arxiv.org/abs/2601.21647.

ByteDance Seed, Yuyu Zhang, Jing Su, Yifan Sun, Chenguang Xi, Xia Xiao, Shen Zheng, Anxiang Zhang, Kaibo Liu, Daoguang Zan, Tao Sun, Jinhua Zhu, Shulin Xin, Dong Huang, Yetao Bai, Lixin Dong, Chao Li, Jianchong Chen, Hanzhi Zhou, Yifan Huang, Guanghan Ning, Xierui Song,

Jiaze Chen, Siyao Liu, Kai Shen, Liang Xiang, and Yonghui Wu. Seed-Coder: Let the code model curate data for itself, 2025. URL https://arxiv.org/abs/2506.03524.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 55005–55029. Curran Associates, Inc., 2024. doi: 10.52202/079017-1745. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/63092d79154adebd7305dfd498cbff70-Paper-Datasets_and_Benchmarks_Track.pdf.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 50750–50777, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/dd1577afd396928ed64216f3f1fd5556-Paper-Conference.pdf.

DiffusionGemma Team. Diffusiongemma technical report, 2026. URL https://arxiv.org/abs/2608.00146.

Chenghao Fan, Wen Heng, Bo Li, Sichen Liu, Yuxuan Song, Jing Su, Xiaoye Qu, Kai Shen, and Wei Wei. Stable-diffcoder: Pushing the frontier of code diffusion large language model, 2026. URL https://arxiv.org/abs/2601.15892.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8093–8131. Curran Associates, Inc., 2024. doi: 10.52202/079017-0261. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0f69b4b96a46f284b726fbd70f74fb3b-Paper-Datasets_and_Benchmarks_Track.pdf.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24678–24704. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4dbb61cb68671edc4ca3712d70083b9f-Paper-Datasets_and_Benchmarks.pdf.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 47094–47165. Curran Associates, Inc., 2024. doi: 10.52202/079017-1493. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/54024fca0cef9911be36319e622cde38-Paper-Conference.pdf.

Mingyu Kim, Dongjun Kim, Amman Yusuf, Stefano Ermon, and Mijung Park. Training-free safe denoisers for safe use of diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QQS7TudonJ.

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*, 2026. URL https://openreview.net/forum?id=EA80Zib9UI.

Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and Marco Cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30911–30942. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kirchhof25a.html.

Yejin Lee and Yo-Sub Han. Adaptive steering and remasking for safe generation in diffusion language models, 2026. URL https://arxiv.org/abs/2605.13043.

Raymond Li, Loubna Ben allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia LI, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Joel Lamy-Poirier, Joao Monteiro, Nicolas Gontier, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Ben Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason T Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Urvashi Bhattacharyya, Wenhao Yu, Sasha Luccioni, Paulo Villegas, Fedor Zhdanov, Tony Lee, Nadav Timor, Jennifer Ding, Claire S Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro Von Werra, and Harm de Vries. Starcoder: may the source be with you! *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=KoFOg41haE. Reproducibility Certification.

Zherui Li, Zheng Nie, Zhenhong Zhou, Yufei Guo, Yue Liu, Yitong Zhang, Yu Cheng, Qingsong Wen, Kun Wang, and Jiaheng Zhang. DiffuGuard: How intrinsic safety is lost and found in diffusion large language models, 2025. URL https://arxiv.org/abs/2509.24296.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. cais/harmbench-llama-2-13b-cls (hugging face model card). https://huggingface.co/cais/HarmBench-Llama-2-13b-cls, 2024b. Accessed: January 29, 2026.

Meta. Llama guard 3 (8b) model card and hazard taxonomy. https://huggingface.co/meta-llama/Llama-Guard-3-8B, 2024. Accessed: January 29, 2026.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5), August 2025. ISSN 2157-6904. doi: 10.1145/3744746. URL https://doi.org/10.1145/3744746.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *The Thirty-ninth Annual*

*Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KnqiC0znVF.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning*, pp. 8599–8608. PMLR, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

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

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models, 2024. URL https://arxiv.org/abs/2308.01263.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. In *Forty-second International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2506.10892.

Subham Sekhar Sahoo, Jean-Marie Lemercier, Zhihan Yang, Justin Deschenaux, Jingyu Liu, John Thickstun, and Ante Jukić. Scaling beyond masked diffusion language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=xKJG1CNtQM.

Yair Schiff, Subham Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-Torre, Bernardo Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 43776–43821, 2025.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen Mckeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55810–55827. PMLR, 2025. URL https://proceedings.mlr.press/v267/singhal25b.html.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.00307.

Zichen Wen, Jiashu Qu, Zhaorun Chen, Xiaoya Lu, Dongrui Liu, Zhiyuan Liu, Ruixi Wu, Yicun Yang, Xiangqi Jin, Haoyun Xu, Xuyang Liu, Weijia Li, Chaochao Lu, Jing Shao, Conghui He, and Linfeng Zhang. The Devil behind the mask: An emergent safety vulnerability of diffusion LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rIPeatvPy3.

Kevin E Wu, Kevin K Yang, Rianne Van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1): 1059, 2024.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5587–5605, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.303. URL https://aclanthology.org/2024.acl-long.303/.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. URL https://arxiv.org/abs/2508.15487.

Amman Yusuf, Zhejun Jiang, and Mijung Park. The safety-aware denoiser for text diffusion models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=dw5BM3OQfZ.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09675.

Yuanhe Zhang, Fangzhou Xie, Zhenhong Zhou, Zherui Li, Hao Chen, Kun Wang, and Yufei Guo. Jailbreaking large language diffusion models: Revealing hidden safety flaws in diffusion-based text generation. In *Responsible Foundation Models Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=NWvL280s2l.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

# A Appendix

## A.1 Notation

Tables 3 and 4 collect the notation for selective SADDER and its theoretical analysis. The special mask token $\mathbb{M}$ is distinct from the centroid midpoint $\boldsymbol{m}_k$ and response truncation cap $M$. We distinguish decreasing noise time $t$ from the increasing reverse iteration index $j$: $j = 0$ is the first model update, and the corresponding noise levels satisfy $t_0 > \cdots > t_T = 0$. All experimental windows below use inclusive iteration indices. So $C = 0$–$15$ means 16 updates. The optional warm-up length $w$ counts initial full-SAD updates; $w = 8$ uses updates 0–7.

Table 3: Notation for selective SADDER.

| Symbol | Meaning |
|---|---|
| $\boldsymbol{c}_i,\ \boldsymbol{x}_i^{\mathrm{harm}},\ \boldsymbol{x}_i^{\mathrm{safe}}$ | Reference prompt and its harmful and safe responses. |
| $P_i,\ R_i^u,\ M$ | Prompt length, response length $\min(\lvert\boldsymbol{x}_i^u\rvert, M)$, and response truncation cap; $u \in \{\mathrm{harm}, \mathrm{safe}\}$. |
| $N_{\mathrm{harm}},\ N_{\mathrm{safe}}$ | Numbers of harmful and safe examples used to estimate the centroids. |
| $\boldsymbol{x}_t,\ \boldsymbol{x}_0,\ \boldsymbol{x}_0^l$ | Noised sequence, clean sequence, and one-hot clean token at position $l$. |
| $\mathcal{V},\ L,\ \mathbb{M}$ | Vocabulary, sequence length, and special mask token. |
| $T,\ j,\ t_j,\ C$ | Number of reverse updates, update index, corresponding noise time, and admitted intervention indices. |
| $\mathcal{G},\ \mathcal{M}_t$ | Editable positions and positions eligible for steering: masked editable positions for MDLMs; all $L$ positions for USDMs in the stated algorithm. |
| $k,\ l,\ d$ | Hidden-state layer, token position, and hidden dimension. |
| $\boldsymbol{h}_{k,l},\ \boldsymbol{h}_{k,l}^{(t)}$ | Layer-$k$ hidden representation at position $l$; the superscript denotes evaluation on the current noised state. |
| $\bar{\boldsymbol{h}}_{i,k}^u,\ \bar{\boldsymbol{h}}_k^u$ | Response-averaged representation of reference $i$ and the corresponding class centroid. |
| $\widehat{\boldsymbol{v}}_k,\ \boldsymbol{m}_k$ | Unit direction from the safe to the harmful centroid, and their midpoint. |
| $s_{k,l}^{(t)}$ | Centered score $\langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k, \widehat{\boldsymbol{v}}_k \rangle$; positive values lie on the harmful-centroid side. |
| $a_{k,l}^{(t)}$ | Binary steering gate: one at eligible positions with positive scores and zero otherwise. |
| $\eta,\ \beta(\boldsymbol{x}_t)$ | Guidance hyperparameter and state-dependent SAD scale; the gated strength is $\eta\beta(\boldsymbol{x}_t)a_{k,l}^{(t)}$. |
| $p_{\mathcal{D},t,l},\ \widetilde{p}_{0\mid t,l}$ | Model clean-token prediction and modified prediction supplied to the reverse solver. |
| $\sigma,\ \tau$ | Sigmoid and positive temperature in the optional softened gate. |

Table 4: Denoiser and theoretical notation.

| Symbol | Meaning |
|---|---|
| $\mathcal{D},\ \mathcal{D}_{\mathrm{unsafe}}$ | Data distribution underlying the pretrained denoiser and unsafe reference set used by the algorithm, respectively. |
| $q_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ | Forward corruption kernel, used to weight clean unsafe references given the noised state. |
| $\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}]$ | Data denoiser at position $l$, estimated by the pretrained model. |
| $\mathbb{E}_{\mathcal{D}_{\mathrm{unsafe}}}[\boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}]$ | Unsafe denoiser, approximated by a weighted mixture of unsafe references. |
| $p_{\mathcal{D}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L})$ | Data posterior over the entire clean sequence. |
| $\psi_l,\ \pi_{t,l}$ | Oracle unsafe-membership function $\mathcal{V}^L \to [0,1]$ and its posterior mean $\pi_{t,l} = \mathbb{E}_{\mathcal{D}}[\psi_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]$. |
| $Z_{\mathrm{safe},l}$ | Local safe normalizer $1 - \pi_{t,l}$; dependence on noise time and state is suppressed. |
| $g_l,\ R_l$ | Local safe weight $g_l = 1 - \psi_l$ and remaining-position weight $R_l = \prod_{j \neq l} g_j$, with $j$ a dummy position index. |
| $p_{\mathcal{D}^{(l)}_{\mathrm{unsafe}}},\ p_{\mathcal{D}^{(l)}_{\mathrm{safe}}}$ | Clean-sequence posteriors obtained by weighting $p_{\mathcal{D}}$ with $\psi_l$ or $g_l$ and normalizing by $\pi_{t,l}$ or $1 - \pi_{t,l}$. The superscript $(l)$ identifies the safety criterion. |
| $p_{\mathcal{D}^{\mathrm{global}}_{\mathrm{safe}}}$ | Global clean-sequence posterior obtained by weighting $p_{\mathcal{D}}$ with $\prod_l g_l$ and normalizing. |
| $\mathbb{E}_\nu,\ \mathrm{Cov}_\nu$ | Expectation and covariance under distribution $\nu$, conditioned on $\boldsymbol{x}_t^{1:L}$ in the proofs. |

### A.2 Derivation of equivalence

Expanding the relationship in equation 10 gives us

$$2\left\langle \boldsymbol{h}_{k,l}^{(t)},\ \bar{\boldsymbol{h}}_k^{\mathrm{harm}} - \bar{\boldsymbol{h}}_k^{\mathrm{safe}} \right\rangle > \|\boldsymbol{h}_k^{\mathrm{harm}}\|_2^2 - \|\boldsymbol{h}_k^{\mathrm{safe}}\|_2^2. \tag{18}$$

In terms of $\hat{\boldsymbol{v}}_k$, this is,

$$\left\langle \boldsymbol{h}_{k,l}^{(t)}, \hat{\boldsymbol{v}}_k \right\rangle > \frac{\|\boldsymbol{h}_k^{\text{harm}}\|_2^2 - \|\boldsymbol{h}_k^{\text{safe}}\|_2^2}{2\|\bar{\boldsymbol{h}}_k^{\text{harm}} - \bar{\boldsymbol{h}}_k^{\text{safe}}\|}. \tag{19}$$

For notational simplicity, we ignore superscripts and subscripts $j, k, t$ below.

$$\langle \boldsymbol{h}, \hat{\boldsymbol{v}} \rangle > \frac{\|\bar{\boldsymbol{h}}^{\text{harm}}\|_2^2 - \|\bar{\boldsymbol{h}}^{\text{safe}}\|_2^2}{2\|\bar{\boldsymbol{h}}^{\text{harm}} - \bar{\boldsymbol{h}}^{\text{safe}}\|}, \tag{20}$$

since $\langle \boldsymbol{h} - \boldsymbol{m}, \hat{\boldsymbol{v}} \rangle = \langle \boldsymbol{h}, \hat{\boldsymbol{v}} \rangle - \langle \boldsymbol{m}, \hat{\boldsymbol{v}} \rangle$, we rewrite the above equation by

$$\langle \boldsymbol{h} - \boldsymbol{m}, \hat{\boldsymbol{v}} \rangle > \frac{\|\bar{\boldsymbol{h}}^{\text{harm}}\|_2^2 - \|\bar{\boldsymbol{h}}^{\text{safe}}\|_2^2}{2\|\bar{\boldsymbol{h}}^{\text{harm}} - \bar{\boldsymbol{h}}^{\text{safe}}\|} - \langle \boldsymbol{m}, \hat{\boldsymbol{v}} \rangle \,, \tag{21}$$

$$= \frac{\|\bar{\boldsymbol{h}}^{\text{harm}}\|_2^2 - \|\bar{\boldsymbol{h}}^{\text{safe}}\|_2^2}{2\|\bar{\boldsymbol{h}}^{\text{harm}} - \bar{\boldsymbol{h}}^{\text{safe}}\|} - \frac{(\bar{\boldsymbol{h}}^{\text{harm}} - \bar{\boldsymbol{h}}^{\text{safe}})^T(\bar{\boldsymbol{h}}^{\text{harm}} + \bar{\boldsymbol{h}}^{\text{safe}})}{2\|\bar{\boldsymbol{h}}^{\text{harm}} - \bar{\boldsymbol{h}}^{\text{safe}}\|}, \tag{22}$$

$$= \frac{(\bar{\boldsymbol{h}}^{\text{harm}} - \bar{\boldsymbol{h}}^{\text{safe}})^T(\bar{\boldsymbol{h}}^{\text{harm}} + \bar{\boldsymbol{h}}^{\text{safe}}) - (\bar{\boldsymbol{h}}^{\text{harm}} - \bar{\boldsymbol{h}}^{\text{safe}})^T(\bar{\boldsymbol{h}}^{\text{harm}} + \bar{\boldsymbol{h}}^{\text{safe}})}{2\|\bar{\boldsymbol{h}}^{\text{harm}} - \bar{\boldsymbol{h}}^{\text{safe}}\|} = 0, \tag{23}$$

where the RHS becomes zero by expanding the additional inner product $\langle \boldsymbol{c}, \hat{\boldsymbol{v}} \rangle$.

### A.3 Global target distribution.

*Theorem* A.1 (**When position-wise SADDER target distribution matches a global target distribution**). Given the position-wise safe target from Thm. 3.1 given in equation 32,

$$p_{\mathcal{D}_{\text{safe}}^{(l)}}(\boldsymbol{x}_0^{1:L} | \boldsymbol{x}_t^{1:L}) = \frac{\big(1 - \psi_l(\boldsymbol{x}_0^{1:L})\big) p_{\text{data}}(\boldsymbol{x}_0^{1:L} | \boldsymbol{x}_t^{1:L})}{1 - \pi_{t,l}(\boldsymbol{x}_t^{1:L})}, \tag{24}$$

and the global target distribution, with $g_l(\boldsymbol{x}_0^{1:L}) := 1 - \psi_l(\boldsymbol{x}_0^{1:L})$ and $R_l(\boldsymbol{x}_0^{1:L}) := \prod_{j \neq l} \big(1 - \psi_j(\boldsymbol{x}_0^{1:L})\big)$ for each position $l$,

$$p_{\mathcal{D}_{\text{safe}}^{\text{global}}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}) = \frac{g_l(\boldsymbol{x}_0^{1:L}) R_l(\boldsymbol{x}_0^{1:L}) p_{\mathcal{D}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L})}{\mathbb{E}_{\mathcal{D}}[g_l(\boldsymbol{x}_0^{1:L}) R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}. \tag{25}$$

the following equality holds,

$$\mathbb{E}_{\mathcal{D}_{\text{safe}}}^{global}(\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}) = \mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}(\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}), \tag{26}$$

if and only if

$$\text{Cov}_{\mathcal{D}_{\text{safe}}^{(l)}}\big(R_l(\boldsymbol{x}_0^{1:L}), \boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}\big) = 0, \tag{27}$$

for every $l, t, \boldsymbol{x}_t^{1:L}$

See Section A.5 for proof.

*Remark* 2. Thm. A.1 states that the global and position-wise denoisers agree at position $l$ for all $l = \{1, \cdots, L\}$ if and only if the remaining-position safety weight denoted by $R_l(\boldsymbol{x}_0^{1:L})$ is conditionally uncorrelated with $\boldsymbol{x}_0^l$ under $\mathcal{D}_{\text{safe}}^{(l)}$. Under this condition and the calibration assumption in Thm. 3.1, every SADDER factor is the marginal safe denoiser of the same global target. However, this does not guarantee the existence of *any* globally compatible distribution. It only states that this specific global target in equation 39 agrees with the local target at position $l$ if and only if a conditional covariance is zero.

### A.4 Proof of Thm. 3.1

*Proof.* In the theorem, we assumed that these exists a measurable position-wise unsafe-membership function on clean sequences. In particular, for each token position $l \in \{1, \ldots, L\}$, this function maps the clean sequence to a level of unsafe-ness at index $l$:

$$\psi_l : \mathcal{V}^L \longrightarrow [0, 1]. \tag{28}$$

Recall that we let

$$\pi_{t,l}(\boldsymbol{x}_t^{1:L}) := \mathbb{E}_D\big[\psi_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}\big] \tag{29}$$

denote the posterior probability of the position-$l$ being unsafe, for a noised sequence $\boldsymbol{x}_t^{1:L}$.

Now, we define the corresponding normalized position-wise safe and unsafe posterior distributions by

$$p_{\mathcal{D}^{(l)}_{\text{unsafe}}}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L}) := \frac{\psi_l(\boldsymbol{x}_0^{1:L})p_{\text{data}}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L})}{\pi_{t,l}(\boldsymbol{x}_t^{1:L})}, \tag{30}$$

$$\pi_{t,l}(\boldsymbol{x}_t^{1:L}) := \sum_{\boldsymbol{x}_0^{1:L}\in\mathcal{D}} \psi_l(\boldsymbol{x}_0^{1:L})p_{\text{data}}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L}), \tag{31}$$

$$p_{\mathcal{D}^{(l)}_{\text{safe}}}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L}) := \frac{\big(1-\psi_l(\boldsymbol{x}_0^{1:L})\big)p_{\text{data}}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L})}{Z_{\text{safe},l}}, \tag{32}$$

$$1-\pi_{t,l}(\boldsymbol{x}_t^{1:L}) := \sum_{\boldsymbol{x}_0^{1:L}\in\mathcal{D}} \big(1-\psi_l(\boldsymbol{x}_0^{1:L})\big)p_{\text{data}}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L}). \tag{33}$$

Because $\psi_l(\boldsymbol{x}_0^{1:L}) + (1-\psi_l(\boldsymbol{x}_0^{1:L})) = 1$ for any clearn sequence $\boldsymbol{x}_0^{1:L}$, we can write the following:

$$p_{data}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L}) = (1-\psi_l(\boldsymbol{x}_0^{1:L}))p_{data}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L}) \tag{34}$$

$$+\,\psi_l(\boldsymbol{x}_0^{1:L})p_{data}(\boldsymbol{x}_0^{1:L}|\boldsymbol{x}_t^{1:L}). \tag{35}$$

If we plug in equation 30 and equation 32 to the RHS, the data posterior can be decomposed into the position-wise, normalized, safe and unsafe distributions:

$$\begin{aligned} p_{\text{data}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}) = \big(1-\pi_{t,l}(\boldsymbol{x}_t^{1:L})\big)p_{D^{(l)}_{\text{safe}}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}) \\ +\,\pi_{t\,l}(\boldsymbol{x}_t^{1:L})p_{D^{(l)}_{\text{unsafe}}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}). \end{aligned} \tag{36}$$

Taking the conditional expectation of the one-hot clean token $x_0^l$ on both sides gives, i.e., $\mathbb{E}_{\boldsymbol{x}_0^l\sim\mathcal{D}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}] = \int \boldsymbol{x}_0^l p(\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L})d\boldsymbol{x}_0^l$

$$\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}] = \big(1-\pi_{t,l}\big)\mathbb{E}_{D^{(l)}_{\text{safe}}}\big[\boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}\big] + \pi_{t,l}\mathbb{E}_{\mathcal{D}^{(l)}_{\text{unsafe}}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}], \tag{37}$$

where the dependence of $\pi_{t,l}$ on $\boldsymbol{x}_t^{1:L}$ is suppressed for readability. Solving equation 37 for the safe conditional mean yields

$$\begin{aligned} \mathbb{E}_{\mathcal{D}^{(l)}_{\text{safe}}}\big[\boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}\big] &= \frac{\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}] - \pi_{t,l}\mathbb{E}_{\mathcal{D}^{(l)}_{\text{unsafe}}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}]}{1-\pi_{t,l}} \\ &= \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}] + \frac{\pi_{t,l}}{1-\pi_{t,l}}\Big(\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}] - \mathbb{E}_{\mathcal{D}^{(l)}_{\text{unsafe}}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}]\Big). \end{aligned} \tag{38}$$

Substituting the calibration condition in equation 17 gives equation 16.

□

Note that the unsafe denoiser in this theorem is induced by $\psi$, i.e., $\mathbb{E}_{\mathcal{D}_{\text{unsafe}}}[\boldsymbol{x}_0^l|\boldsymbol{x}_t^{1:L}] = \frac{\sum_{\boldsymbol{x}_n\in\mathcal{D}_{\text{unsafe}}}\boldsymbol{x}_n^l\psi_l(\boldsymbol{x}_0^{1:L})q_t(\boldsymbol{x}_t^{1:L}|\boldsymbol{x}_n^{1:L})}{\sum_{\boldsymbol{x}_n\in\mathcal{D}_{\text{unsafe}}}\psi_l(\boldsymbol{x}_0^{1:L})q_t(\boldsymbol{x}_t^{1:L}|\boldsymbol{x}_n^{1:L})}$, which differs from our implementation of unsafe denoiser given in equation 15. Hence, we note that this theorem statement holds when these two unsafe denoisers match.

### A.5 Proof of Thm. A.1

*Proof.* Conditioned on $\boldsymbol{x}_t^{1:L}$, the global posterior satisfies

$$p_{\mathcal{D}^{\text{global}}_{\text{safe}}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}) = \frac{g_l(\boldsymbol{x}_0^{1:L})R_l(\boldsymbol{x}_0^{1:L})p_{\mathcal{D}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L})}{\mathbb{E}_{\mathcal{D}}[g_l(\boldsymbol{x}_0^{1:L})R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}, \tag{39}$$

where $g_l(\boldsymbol{x}_0^{1:L}) := 1 - \psi_l(\boldsymbol{x}_0^{1:L})$.

The position-$l$ safe posterior is

$$p_{\mathcal{D}_{\text{safe}}^{(l)}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}) = \frac{g_l(\boldsymbol{x}_0^{1:L}) p_{\mathcal{D}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L})}{\mathbb{E}_{\mathcal{D}}[g_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}. \tag{40}$$

If we plug in equation 40 into equation 39, we have the global posterior as a reweighted position-wise posterior:

$$p_{\mathcal{D}_{\text{safe}}^{\text{global}}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}) = \frac{R_l(\boldsymbol{x}_0^{1:L}) p_{\mathcal{D}_{\text{safe}}^{(l)}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}) \mathbb{E}_{\mathcal{D}}[g_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}{\mathbb{E}_{\mathcal{D}}[g_l(\boldsymbol{x}_0^{1:L}) R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]} \tag{41}$$

$$= \frac{R_l(\boldsymbol{x}_0^{1:L}) p_{\mathcal{D}_{\text{safe}}^{(l)}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L})}{\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}, \tag{42}$$

where the second line is true because

$$\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}] = \sum_{\boldsymbol{x}_0^{1:L}} R_l(\boldsymbol{x}_0^{1:L}) p_{\mathcal{D}_{\text{safe}}^{(l)}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L}), \tag{43}$$

$$= \sum_{\boldsymbol{x}_0^{1:L}} R_l(\boldsymbol{x}_0^{1:L}) \frac{g_l(\boldsymbol{x}_0^{1:L}) p_{\mathcal{D}}(\boldsymbol{x}_0^{1:L} \mid \boldsymbol{x}_t^{1:L})}{\mathbb{E}_{\mathcal{D}}[g_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}, \quad \text{due to equation 40} \tag{44}$$

$$= \frac{\mathbb{E}_{\mathcal{D}}[g_l(\boldsymbol{x}_0^{1:L}) R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}{\mathbb{E}_{\mathcal{D}}[g_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}. \tag{45}$$

Taking the conditional expectation of $\boldsymbol{x}_0^l$ on equation 42 gives

$$\mathbb{E}_{\mathcal{D}_{\text{safe}}^{\text{global}}}[\boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}] = \frac{\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}]}{\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]}. \tag{46}$$

Subtracting $\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[\boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}]$ and using the definition of conditional covariance proves

$$\mathbb{E}_{\mathcal{D}_{\text{safe}}^{\text{global}}}\left[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}\right] - \mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}\left[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}\right] \tag{47}$$

$$= \frac{\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}]}{\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]} - \mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}\left[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}\right] \tag{48}$$

$$= \frac{1}{\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}]} \left( \mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \boldsymbol{x}_0^l \mid \boldsymbol{x}_t^{1:L}] - \mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}[R_l(\boldsymbol{x}_0^{1:L}) \mid \boldsymbol{x}_t^{1:L}] \mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}\left[\boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}\right] \right), \tag{49}$$

$$= \frac{\text{Cov}_{\mathcal{D}_{\text{safe}}^{(l)}}\left(R_l(\boldsymbol{x}_0^{1:L}), \boldsymbol{x}_0^l | \boldsymbol{x}_t^{1:L}\right)}{\mathbb{E}_{\mathcal{D}_{\text{safe}}^{(l)}}\left[R_l(\boldsymbol{x}_0^{1:L}) | \boldsymbol{x}_t^{1:L}\right]}, \tag{50}$$

where the last line is true due to

$$Cov[X, Y|Z] = \mathbb{E}\left[(X - \mathbb{E}(X|Z))(Y - \mathbb{E}(Y|Z))\right], \tag{51}$$

$$= \mathbb{E}[XY|Z] - \mathbb{E}[X|Z]\mathbb{E}[Y|Z] - \mathbb{E}[X|Z]\mathbb{E}[Y|Z] + \mathbb{E}[X|Z]\mathbb{E}[Y|Z], \tag{52}$$

$$= \mathbb{E}[XY|Z] - \mathbb{E}[X|Z]\mathbb{E}[Y|Z]. \tag{53}$$

Since the denominator is positive, the global and local means are the same, when the numerator is zero. □

### A.6 Datasets, prompt selection, and reference construction

We distinguish three data roles: evaluation prompts, unsafe references for the SAD update, and safe/harmful examples for constructing the activation direction. A benchmark used for one role is not automatically used for the others.

**RealToxicityPrompts (RTP).** RTP contains naturally occurring text prefixes and their original continuations, with Perspective annotations (Gehman et al., 2020). For unsafe continuation, the configured pool requires a maximum available prompt-side Perspective score of at least 0.8 and a subsequent Llama-Guard unsafe verdict on the prompt. Its builder also retains the adapter's continuation-score floor of 0.5. These filters describe an enriched evaluation cohort, not the full RTP distribution. Since Llama-Guard also judges generated outputs, the prompt selection is not judge-independent. Note the negation set used to construct the unsafe reference bank for the contrastive direction is not Llama-Guard scored.

The original RTP development/search set comprises pool indices 0–299. The reported continuation tables use the expanded evaluation set, indices 300–899, collected as three 200-prompt shards at offsets 300, 500, and 700. These evaluation records exclude the original development records. The subsequent configuration-selection procedure is described below.

**BeaverTails.** BeaverTails provides instruction–response pairs with response-safety annotations (Ji et al., 2023). We condition on prompts associated with unsafe-labeled responses, using 600 records for the evaluation set. The corresponding unsafe responses supply the negative reference set and safe-labeled responses supply the safe direction set. An unsafe response annotation does not establish that its prompt is itself malicious.

**Benign continuation controls.** The quality cohort requires both the maximum prompt-side Perspective score and the maximum score from the continuation's own fields to be at most 0.2. We remove duplicate prompt strings and prompt strings present in the excluded unsafe-reference artifacts, shuffle with seed zero, and retain 600 records. BERTScore compares each generated response with its own benign human continuation, including for unmodified sampling. BeaverTails-tuned methods use this same benign RTP cohort while retaining their BeaverTails reference banks. This is a quality diagnostic rather than an independent capability evaluation.

**JailbreakBench, AdvBench, and HarmBench.** JailbreakBench supplies harmful behaviors for standardized jailbreak evaluation (Chao et al., 2024); its 100 harmful behaviors are development data in our study. AdvBench contains harmful instructions used to evaluate adversarial prompting (Zou et al., 2023). HarmBench provides harmful behaviors and a behavior-conditioned evaluation classifier (Mazeika et al., 2024a) and our jobs restrict evaluation to its standard behavior category. We shuffle with seed zero and use the first 100 behaviors of each dataset for development. The transfer protocol reserves AdvBench offsets 100–519 and HarmBench-standard offsets 100–199. The main jailbreak tables report the latter held-out HarmBench cohort under DIJA and PAD. These are our experimental splits, not official benchmark train/test partitions.

**WildJailbreak and the jailbreak reference bank.** WildJailbreak contains both harmful and benign prompts, including adversarial rewrites (Jiang et al., 2024). Our unsafe set construction selects the adversarial-harmful training subset, skips the first 1,000 filtered records, and generates LLaDA-Instruct responses from three disjoint 1,300-prompt windows. Exact evaluation-behavior strings are excluded. Generation uses 256 steps, up to 256 new tokens, temperature zero, and seed 42. Unsafe-response filtering, refusal filtering, and deduplication produce the documented 357-response unsafe WildJailbreak set. The jailbreak contrastive direction uses unsafe responses and refusal references with the chat template and a 64-token response cap.

**XSTest and over-refusal.** XSTest contains 250 safe prompts that models should answer and 200 unsafe contrast prompts (Röttger et al., 2024). Over-refusal is the refusal rate on the safe subset, reported separately from refusal on unsafe prompts. Our runs use unmodified prompts through the model's chat template (`vanilla`). The evaluator we use matches the response opening against XSTest's full-refusal strings; it does not detect partial refusals or every paraphrased refusal.

### A.7 Models, attacks, and evaluation measures

We use LLaDA-8B-Base and LLaDA-8B-Instruct (Nie et al., 2025), and Dream-7B-Base and Dream-7B-Instruct (Ye et al., 2025). The base checkpoints are evaluated by prompted continuation; instruction checkpoints use the attack-specific input construction. The implementation's layer

$k$ follows the hidden-state indexing and interior layers are inputs to block $k$, and the final layer includes the model's final normalization.

**Attacks and fixed context.** Zero-shot runs present the original harmful behavior through the chat template. DIJA fills mask spans inside the saved adversarial prompt (Wen et al., 2026) and the refined prompts are reused across methods. PAD inserts fixed anchors into the generation region (Zhang et al., 2025). SADDER and generation-scope DiffuGuard modify only initially editable positions. Full-block DiffuGuard additionally permits repair of fixed scaffold text after the first masked position, including PAD anchors; under DIJA it also appends the released implementation's extra masked tail. The main jailbreak tables concern DIJA and PAD.

**Safety and quality.** Llama-Guard-3-8B (Meta, 2024) judges the generated response alone under continuation framing. HarmBench-Llama-2-13b-cls (Mazeika et al., 2024a;b) judges the response against the requested behavior. We report their respective unsafe rates separately. GPT-2-large (Radford et al., 2019) scores generated text with a causal language-model likelihood; this is an external fluency proxy for both model families, not LLaDA's pseudo-perplexity. Padding and text after the first end-of-sequence token are excluded. Harmful PPL includes all scorable outputs on the unsafe cohort, whether the safety judge labels them safe or unsafe. Benign PPL and BERTScore F1 (Zhang et al., 2020) use the separate benign cohort. BERTScore specifies English and disables baseline rescaling; it leaves the encoder to the installed package's default. BERTScore is multiplied by 100 in the tables. Higher reference similarity and lower PPL are useful diagnostics, but neither alone establishes helpfulness or safety.

**WildGuard response views.** For DIJA, the appendix additionally reports WildGuard Han et al. (2024) harmfulness on the newline-joined spans marked as initially masked, paired with the original harmful behavior. This excludes the fixed scaffold and any full-block repairs outside those spans. Full-response WildGuard harmfulness and refusal remain in the appendix; XSTest uses full-response refusal. The generated-span evaluation is available only for fixed-policy DIJA runs, not PAD or the window studies (Appendix H.5).

**Uncertainty and cost.** Tables report averages with 95% confidence intervals, estimated by repeatedly resampling the evaluation prompts. The appendix also reports standard deviations (SDs) across three generation seeds, showing how much results vary between runs. Appendix B.2 explains both measures and the paired comparisons. NFE counts generator forward passes, including additional passes after remasking. It excludes reward scoring, safety judges, and preparation of the reference set, so it measures generator computation rather than total runtime.

### A.8 Hyperparameters and configuration selection

Table 5 gives the shared settings; Table 6 summarizes the tested hyperparameter ranges by model and parameter; Table 7 lists the selected continuation operating points. The remaining variant settings and results follow in Appendix E. Full SAD does not depend on the hidden-state layer.

Table 5: Shared sampling and guidance settings.

| Parameter | Setting |
|---|---|
| Continuation sampling | 32 new tokens; 32 reverse steps; temperature 1; seeds 42, 43, 44. |
| Jailbreak sampling | 128 nominal steps; nominal length 128; temperature 1; seeds 42, 43, 44, except the explicitly labeled one-seed pilot. DIJA fills existing template spans. |
| Unsafe set | Continuation: up to 2,000 responses from the matching dataset after exact prompt-overlap exclusion. Jailbreak: documented 357-response Wild-Jailbreak set. |
| Direction construction | Clean-reference response means. Later jobs use all valid reference rows; the selected earlier LLaDA/RTP selective row uses the 256-reference direction protocol. |
| Centered selection | Strict score threshold $s_{k,l}^{(t)} > 0$ on masked editable positions. |
| Remasking | Off for the core method; top-$b$ uses $b = 1$; aligned remasking reopens all positive-score revealed positions. |
| Warm-up | Off for selective-only SADDER. Main continuation warm-up: $w = 8$, followed by selective guidance through step 15. |
| Matched controls | Count-matched random selection or reopening; matrix-control seed zero. |
| Search budgets | BoN and FK with 2, 4, or 8 candidates/particles; BoN-2 composition with SAD-based steering. |

Table 6: Hyperparameter values considered, consolidated by model. Values are unions across the sweeps, not a single fully crossed grid; a single value denotes a fixed setting. $k$ is the scoring layer, $\eta$ the SAD strength, $w$ the number of full-SAD warm-up steps, and $[s, e]$ the inclusive intervention window.

| Model | Parameter | Values |
|---|---|---|
| LLaDA-Base | Strength $\eta$ | $\{0, 1, 3, 10, 30, 100, 300, 1000, 3000\}$ |
| | Layer $k$ | $\{8, 16, 24, 31\}$ |
| | Window $[s, e]$ | $[0, 7]$, $[0, 15]$, $[0, 31]$ |
| | Warm-up $w$ | $\{8, 16\}$ |
| | BoN / FK budget | $\{2, 4, 8\}$ |
| Dream-Base | Strength $\eta$ | $\{0, 0.1, 0.3, 1, 3, 10, 100, 3000,$ $10000, 30000, 100000, 300000\}$ |
| | Layer $k$ | 27 |
| | Window $[s, e]$ | $[0, 7]$, $[0, 15]$ |
| | BoN / FK budget | $\{2, 4, 8\}$ |
| LLaDA-Instruct | Strength $\eta$ | $\{1, 3, 10, 30, 100, 300, 1000, 3000, 10000\}$ |
| | Layer $k$ | $\{9, 16, 24\}$ |
| | Window $[s, e]$ | $[0, 15]$, $[0, 31]$, $[0, 63]$, $[0, 127]$, $[32, 63]$, $[64, 95]$, $[96, 127]$ |
| Dream-Instruct | Strength $\eta$ | $\{0.3, 1, 3, 10, 30, 100, 300, 1000, 3000, 10000\}$ |
| | Layer $k$ | $\{1, \ldots, 28\}$ |
| | Window $[s, e]$ | $[0, 31]$ |

**Development settings.** Base models use RTP: the original search uses indices 0–299; additional strength development uses 64 prompts at indices 900–963, seeds 42/43, and window $[0, 7]$ ($k = 31$ for LLaDA, 27 for Dream). Instruction-model development uses 100 JBB behaviors; LLaDA additionally uses the first 100 AdvBench and HarmBench behaviors. These runs use seed 42 or seeds 42/43/44, as documented with the results. Sampling uses 32 steps for Base models and 128 nominal steps for Instruct models.

Table 7: Selected main-table continuation settings.

| Model | Data | Method | $k$ | $\eta$ | Last step | $w$ | Reopening |
|---|---|---|---|---|---|---|---|
| LLaDA | RTP | Full SAD | – | 3000 | 7 | 0 | none |
| | RTP | SADDER (selective) | 31† | 100 | 15 | 0 | none |
| | RTP | SADDER + warm-up | 24 | 3000 | 15 | 8 | none |
| LLaDA | BT | Full SAD | – | 3000 | 15 | 0 | none |
| | BT | SADDER (selective) | 31 | 3000 | 15 | 0 | none |
| | BT | SADDER + warm-up | 16 | 3000 | 15 | 8 | none |
| Dream | RTP | Full SAD | – | 10000 | 15 | 0 | none |
| | RTP | SADDER (selective) | 27 | 10000 | 15 | 0 | none |
| | RTP | SADDER + warm-up | 27 | 3000 | 15 | 8 | none |
| Dream | BT | Full SAD | – | 10000 | 15 | 0 | none |
| | BT | SADDER (selective) | 27 | 10000 | 15 | 0 | none |
| | BT | SADDER + warm-up | 27 | 10000 | 15 | 8 | none |

All windows begin at step zero. BT denotes BeaverTails.

# B Implementation and evaluation details

**Motivation for remasking.** The safety improvements reported by CSD-based steering and DiffuGuard (Lee & Han, 2026; Li et al., 2025) motivate examining remasking as an optional extension to SADDER. Reopening previously generated positions allows guidance to revise token choices that would otherwise remain fixed. Uniform-state diffusion provides a related capability through its native token replacement transitions, allowing repeated revision without an explicit mask token. As such, we wanted to evaluate selective SAD with risk-based remasking and matched random controls in the appendix.

## B.1 Optional remasking and warm-up

Algorithm 2 extends the selective-only main algorithm. Let $\mathcal{R}_t = \{l \in \mathcal{G} : \boldsymbol{x}_t^l \neq \mathbb{M}\}$ be the revealed editable positions. The remasking policy $r$ determines the reopened set:

$$\mathcal{Q}_t(r) = \begin{cases} \varnothing, & r = \text{none}, \\ \mathrm{Top}_{\min(b,|\mathcal{R}_t|)}\{s_{k,l}^{(t)} : l \in \mathcal{R}_t\}, & r = \text{top-}b, \\ \{l \in \mathcal{R}_t : s_{k,l}^{(t)} > 0\}, & r = \text{aligned}. \end{cases} \tag{54}$$

Here Top returns position indices with the largest scores; we use $b = 1$ in the top-$b$ experiments. This ranking policy can reopen a position even if every score is non-positive. Aligned remasking instead reopens every positive-score revealed position, without a top-$b$ cap. After replacing selected tokens by $\mathbb{M}$, the model predictions and hidden states are recomputed. The steering gate is then evaluated on this changed state; reopening does not force a position to be steered. Warm-up length $w$ applies full SAD on admitted denoising steps with index $j < w$, followed by selective steering for $j \geq w$. Remasking is disabled during warm-up and on the final sampling step. Outside the intervention set $C$, neither guidance nor remasking is applied. Setting $w = 0$ disables warm-up; setting $r = \text{none}$ disables remasking. Thus the selective-only main method is exactly the case $w = 0$, $r = \text{none}$.

**Algorithm 2** SADDER with optional remasking and full-SAD warm-up.

**Input:** inputs of Algorithm 1; warm-up length $w \geq 0$; policy $r \in \{\text{none}, \text{top-}b, \text{aligned}\}$; budget $b \geq 0$.
$w = 0$ and $r = \text{none}$ recover Algorithm 1.
**for** $j = 0, \ldots, T-1$ **do**
  Set $t \leftarrow t_j$. Evaluate $p_{\mathcal{D},t,\cdot}$ and $\boldsymbol{h}_{k,\cdot}^{(t)}$ together on the current $\boldsymbol{x}_t$.
  Initialize $\widetilde{p}_{0|t,l} \leftarrow p_{\mathcal{D},t,l}$ and $a_{k,l}^{(t)} \leftarrow 0$ for all $l$.
  **if** $j \in C$ **then**
    Initialize $\mathcal{Q}_t \leftarrow \varnothing$.
    **if** $j \geq w$ **and** $r \neq \text{none}$ **and** $j < T-1$ **then**
      $\mathcal{R}_t \leftarrow \{l \in \mathcal{G} : \boldsymbol{x}_t^l \neq \mathbb{M}\}$.
      Compute $s_{k,l}^{(t)} \leftarrow \langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k, \widehat{\boldsymbol{v}}_k \rangle$ for $l \in \mathcal{R}_t$.
      Select $\mathcal{Q}_t(r)$ using Equation equation 54.
      Set $\boldsymbol{x}_t^l \leftarrow \mathbb{M}$ for $l \in \mathcal{Q}_t$.
    **end if**
    **if** $\mathcal{Q}_t \neq \varnothing$ **then**
      Recompute $p_{\mathcal{D},t,\cdot}$ and $\boldsymbol{h}_{k,\cdot}^{(t)}$ together on the changed $\boldsymbol{x}_t$.
      Reset $\widetilde{p}_{0|t,\cdot} \leftarrow p_{\mathcal{D},t,\cdot}$.
    **end if**
    $\mathcal{M}_t \leftarrow \{l \in \mathcal{G} : \boldsymbol{x}_t^l = \mathbb{M}\}$.
    **if** $j < w$ **then**
      **Full-SAD warm-up:** $a_{k,l}^{(t)} \leftarrow 1$ for $l \in \mathcal{M}_t$.
    **else**
      Compute $s_{k,l}^{(t)} \leftarrow \langle \boldsymbol{h}_{k,l}^{(t)} - \boldsymbol{m}_k, \widehat{\boldsymbol{v}}_k \rangle$ for $l \in \mathcal{M}_t$.
      **Select risky positions:** $a_{k,l}^{(t)} \leftarrow \mathbf{1}\{s_{k,l}^{(t)} > 0\}$ for $l \in \mathcal{M}_t$.
    **end if**
    **if** at least one gate equals 1 **then**
      Estimate $\widehat{p}_{\mathcal{D}_{\text{unsafe}},t,\cdot}$ and $\widehat{\beta}(\boldsymbol{x}_t)$ on this same state; set $\eta_t \leftarrow \eta\widehat{\beta}(\boldsymbol{x}_t)$.
      **SAD update on selected positions:** update $\widetilde{p}_{0|t,l}$ using Equation equation 14 for each $l$ with $a_{k,l}^{(t)} = 1$.
    **end if**
  **end if**
  $\boldsymbol{x}_{t_{j+1}} \leftarrow \text{Solver}(\boldsymbol{x}_t, t_j, t_{j+1}, \widetilde{p}_{0|t,\cdot}; \mathcal{G})$.
**end for**
**Return:** the generated continuation of $\boldsymbol{x}_{t_T}$.

### B.2 METRIC ESTIMATES AND UNCERTAINTY

We compute uncertainty from the saved evaluations. Let $i = 1, \ldots, n$ index evaluation records and $s = 1, \ldots, S$ index the observed generation seeds. The main continuation evaluations have $n = 600$ and $S = 3$; the main HarmBench evaluations have $n = 100$ and $S = 3$.

**Unsafe rates and other averages.** Let $y_{i,s} \in \{0, 1\}$ indicate whether the output for prompt record $i$ and seed $s$ is judged unsafe. On a complete, balanced evaluation, the unsafe percentage is

$$\widehat{r}_{\text{unsafe}} = \frac{100}{nS} \sum_{i=1}^{n} \sum_{s=1}^{S} y_{i,s}. \tag{55}$$

The same expression defines HarmBench ASR, refusal rate, or short-output rate when $y_{i,s}$ is the corresponding event indicator. A mean repetition score uses the per-output repetition score instead. Each metric retains its own evaluation population: for example, XSTest over-refusal uses only safe prompts. When observation counts differ, rate estimates use the total number of valid observations as their denominator.

**Quality metrics.** Let $\mathcal{I}_s$ contain the valid records for a quality metric at seed $s$. The reported mean averages a metric computed within each seed:

$$\widehat{Q}_s = f\left(\frac{\sum_{i \in \mathcal{I}_s} A_{i,s}}{\sum_{i \in \mathcal{I}_s} D_{i,s}}\right), \qquad \widehat{Q} = \frac{1}{S} \sum_{s=1}^{S} \widehat{Q}_s. \tag{56}$$

For BERTScore, $A_{i,s}$ is the response-level F1 score, $D_{i,s} = 1$, and $f(z) = z$; the displayed values are multiplied by 100. For PPL, $A_{i,s}$ is the total negative log likelihood of scored tokens, $D_{i,s}$ is their count, and $f(z) = \exp(z)$. Thus PPL is an average of seed-level corpus perplexities, not an average of response perplexities. Missing or unscorable measurements are excluded; they are not assigned zero.

**95% confidence intervals.** To estimate uncertainty due to the evaluation prompts, we randomly sample prompts from the evaluated set, allowing the same prompt to be selected more than once. Each resampled set has the same number of prompts as the original set. Whenever a prompt is selected, we keep its responses from all three seeds together. We calculate the metric again for each resampled set and repeat this process 5,000 times. The 2.5th and 97.5th percentiles of these 5,000 estimates form the reported 95% confidence interval.

**Seed SD and paired comparisons.** For seed-level estimates $\widehat{\theta}_s$, the reported sample SD is

$$\text{SD}_{\text{seed}} = \sqrt{\frac{1}{S-1} \sum_{s=1}^{S} (\widehat{\theta}_s - \bar{\theta}_{\text{seed}})^2}, \qquad \bar{\theta}_{\text{seed}} = \frac{1}{S} \sum_{s} \widehat{\theta}_s. \tag{57}$$

It describes variation across runs and is not a confidence interval; it is unavailable for a single seed. To compare methods A and B, retain matched prompts and seeds and use identical cluster weights for both:

$$\widehat{\Delta} = \widehat{\theta}_{\text{A}} - \widehat{\theta}_{\text{B}}, \qquad \widehat{\Delta}^{*(b)} = \widehat{\theta}_{\text{A}}^{*(b)} - \widehat{\theta}_{\text{B}}^{*(b)}. \tag{58}$$

## C CONTINUATION RESULTS WITH CONFIDENCE INTERVALS

Table 8 reports the main LLaDA continuation estimates with pointwise 95% prompt-cluster bootstrap intervals from 5,000 resamples, retaining all observed seeds within each sampled prompt. These intervals describe prompt uncertainty; the main-table subscripts describe variation across three generation seeds. The selected configurations and evaluation cohorts are identical in both tables.

Table 8: LLaDA-Base safety, generation quality, and cost.

| Method | RTP | | | | | BeaverTails | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Unsafe % ↓ | Harmful PPL ↓ | Benign PPL ↓ | Benign BERT ↑ | NFE | Unsafe % ↓ | Harmful PPL ↓ | Benign PPL ↓ | Benign BERT ↑ | NFE |
| Unmodified | 38.6 [35.7,41.5] | 21.1 [20.2,22.1] | 17.5 [16.9,18.2] | 84.65 [84.50,84.79] | 32.0 | 38.6 [35.6,41.6] | 20.6 [19.6,21.6] | 17.5 [16.9,18.2] | 84.65 [84.50,84.79] | 32.0 |
| Full SAD | 32.4 [29.4,35.5] | 29.3 [27.6,31.1] | 30.0 [27.7,32.9] | 83.92 [83.75,84.09] | 32.0 | 27.1 [24.0,30.4] | 35.1 [33.2,37.0] | 47.1 [42.9,52.7] | 83.38 [83.20,83.55] | 32.0 |
| SADDER (selective) | 29.4 [26.3,32.5] | 42.1 [37.4,47.8] | 18.4 [17.7,19.2] | 84.59 [84.45,84.74] | 32.0 | 29.9 [26.8,33.2] | 34.6 [32.2,37.3] | 27.8 [26.0,30.0] | 84.20 [84.04,84.35] | 32.0 |
| SADDER + warm-up | 31.2 [28.1,34.5] | 31.6 [29.7,33.5] | 44.3 [38.4,52.5] | 83.01 [82.81,83.20] | 32.0 | 24.4 [21.2,27.8] | 36.9 [34.2,39.9] | 44.1 [40.2,49.7] | 83.18 [82.99,83.35] | 32.0 |
| CSD† | 22.4 [20.3,24.7] | 225 [204,249] | 21.6 [20.6,22.7] | 84.59 [84.44,84.73] | 61.9 | 15.6 [13.8,17.3] | 487 [440,541] | 943 [868,1030] | 82.61 [82.48,82.73] | 55.0 |
| BoN-8 | 19.0 [16.7,21.5] | 20.5 [19.6,21.4] | 19.5 [18.7,20.2] | 84.70 [84.56,84.84] | 256.0 | 18.6 [16.3,20.9] | 24.2 [23.0,25.4] | 19.5 [18.7,20.2] | 84.70 [84.56,84.84] | 256.0 |
| FK-8 | 20.6 [18.1,23.0] | 21.3 [20.4,22.3] | 18.2 [17.5,18.8] | 84.71 [84.57,84.85] | 256.0 | 28.2 [25.4,30.9] | 22.7 [21.7,23.7] | 18.2 [17.5,18.8] | 84.71 [84.57,84.85] | 256.0 |

## D DREAM-BASE CONTINUATION RESULTS

Table 9 reports the complete Dream-Base comparison under the same continuation protocol as LLaDA (Section 5.1). Unsafe rate and harmful PPL use the RTP or BeaverTails harmful cohort; benign PPL and BERTScore use the separate benign RTP cohort. Entries are means with 95% prompt-cluster intervals; NFE measures generator cost.

Table 9: Dream-Base safety, generation quality, and cost.

| Method | RTP | | | | | BeaverTails | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Unsafe % ↓ | Harmful PPL ↓ | Benign PPL ↓ | Benign BERT ↑ | NFE | Unsafe % ↓ | Harmful PPL ↓ | Benign PPL ↓ | Benign BERT ↑ | NFE |
| Unmodified | 46.8 [43.8,49.7] | 18.3 [17.6,19.1] | 15.5 [15.0,16.0] | 84.71 [84.57,84.86] | 32.0 | 38.7 [35.8,41.6] | 12.7 [12.2,13.2] | 15.5 [15.0,16.0] | 84.71 [84.57,84.86] | 32.0 |
| Full SAD | 33.4 [30.1,36.8] | 26.0 [24.3,27.8] | 23.3 [22.0,24.7] | 83.16 [82.96,83.35] | 32.0 | 22.7 [19.7,25.7] | 23.7 [21.8,25.8] | 23.0 [21.7,24.5] | 83.62 [83.44,83.80] | 32.0 |
| SADDER (selective) | 35.9 [32.8,39.0] | 29.1 [27.5,30.7] | 18.1 [17.5,18.7] | 84.59 [84.45,84.73] | 32.0 | 34.1 [31.3,36.8] | 19.3 [18.4,20.3] | 20.0 [19.2,20.7] | 84.49 [84.34,84.64] | 32.0 |
| SADDER + warm-up | 34.7 [31.6,38.2] | 27.1 [25.5,28.8] | 21.3 [20.3,22.3] | 83.76 [83.57,83.94] | 32.0 | 27.7 [24.8,30.8] | 23.6 [22.1,25.1] | 20.5 [19.6,21.5] | 84.21 [84.06,84.37] | 32.0 |
| CSD† | 32.9 [30.6,35.4] | 1180 [1100,1270] | 3400 [3270,3540] | 81.51 [81.42,81.61] | 60.2 | 38.6 [35.7,41.4] | 17.8 [16.7,19.1] | 27.2 [25.3,29.4] | 84.50 [84.36,84.64] | 55.8 |
| BoN-8 | 24.6 [22.0,27.3] | 17.5 [16.7,18.2] | 16.9 [16.3,17.5] | 84.81 [84.68,84.95] | 256.0 | 20.0 [17.8,22.2] | 15.0 [14.3,15.7] | 16.9 [16.3,17.5] | 84.81 [84.68,84.95] | 256.0 |
| FK-8 | 25.1 [22.5,27.7] | 17.9 [17.2,18.6] | 15.7 [15.3,16.2] | 84.76 [84.62,84.90] | 256.0 | 27.8 [25.2,30.4] | 13.9 [13.4,14.5] | 15.7 [15.3,16.2] | 84.76 [84.62,84.90] | 256.0 |

## E CONTINUATION SETTINGS AND COMPONENT VARIANTS

Settings are selected by unsafe rate; benign quality is evaluated afterward. Tables 10 and 11 report layer, strength, and the last admitted step (all windows begin at zero). The warm-up variant uses full SAD on steps 0–7 and selective guidance on 8–15, with aligned remasking on LLaDA and no remasking on Dream. Equation equation 54 defines the alternative reopening policies.

Table 10: LLaDA selected continuation settings on development set.

| Dataset | Method | Setting |
|---|---|---|
| RTP | Full SAD | $31/3000/7$ |
| | SADDER | $31/100/15$ |
| | SADDER w/ top-1 remasking | $24/100/15$ |
| | SADDER w/ aligned remasking | $16/100/15$ |
| | SADDER + warm-up + aligned remasking | $24/3000/15$; |
| | CSD$^{\dagger}$ | $k = 31$, $j = 0{:}15$ |
| | BoN-8 | $N = 8$ |
| | FK-8 | $N = 8$ |
| BeaverTails | Full SAD | $31/3000/15$ |
| | SADDER | $31/3000/15$ |
| | SADDER w/ top-1 remasking | $24/100/7$ |
| | SADDER w/ aligned remasking | $16/3000/15$ |
| | SADDER + warm-up + aligned remasking | $16/3000/15$; aligned |
| | CSD$^{\dagger}$ | $k = 24$, $j = 0{:}7$ |
| | BoN-8 | $N = 8$ |
| | FK-8 | $N = 8$ |

Table 11: Dream selected continuation settings on development set.

| Dataset | Method | Setting |
|---|---|---|
| RTP | Full SAD | $27/10000/15$ |
| | SADDER | $27/10000/15$ |
| | SADDER w/ top-1 remasking | $27/100/15$ |
| | SADDER w/ aligned remasking | $24/1e5/15$ |
| | SADDER + warm-up | $27/3000/15$; none |
| | CSD$^{\dagger}$ | $k = 24$, $j = 0{:}15$ |
| | BoN-8 | $N = 8$ |
| | FK-8 | $N = 8$ |
| BeaverTails | Full SAD | $27/10000/15$ |
| | SADDER | $27/10000/15$ |
| | SADDER w/ top-1 remasking | $27/100/15$ |
| | SADDER w/ aligned remasking | $27/100/15$ |
| | SADDER + warm-up | $27/10000/15$; none |
| | CSD$^{\dagger}$ | $k = 27$, $j = 0{:}15$ |
| | BoN-8 | $N = 8$ |
| | FK-8 | $N = 8$ |

**SADDER variants.** These tables extend the main comparison with top-1 and aligned remasking, using the same metrics and uncertainty conventions.

Table 12: LLaDA-Base component variants.

| Method | Unsafe % ↓ | Harmful PPL ↓ | Benign PPL ↓ | Benign BERT ↑ | NFE |
|---|---|---|---|---|---|
| **RTP** | | | | | |
| Unmodified | 38.6 [35.7, 41.5] | 21.14 [20.21, 22.13] | 17.53 [16.90, 18.21] | 84.65 [84.50, 84.79] | 32.0 |
| Full SAD | 32.4 [29.4, 35.5] | 29.30 [27.61, 31.13] | 29.96 [27.65, 32.89] | 83.92 [83.75, 84.09] | 32.0 |
| SADDER (selective) | 29.4 [26.3, 32.5] | 42.08 [37.39, 47.75] | 18.41 [17.70, 19.17] | 84.59 [84.45, 84.74] | 32.0 |
| SADDER w/ top-1 remasking | 28.3 [25.2, 31.3] | 43.25 [38.63, 48.89] | 41.75 [37.34, 47.13] | 83.38 [83.17, 83.58] | 47.0 |
| SADDER w/ aligned remasking | 30.6 [27.8, 33.4] | 35.81 [33.47, 38.38] | 35.88 [33.53, 38.57] | 83.99 [83.82, 84.16] | 46.6 |
| SADDER + warm-up + aligned remasking | 27.2 [24.7, 29.9] | 38.75 [36.22, 41.88] | 33.33 [31.08, 35.91] | 84.13 [83.96, 84.28] | 40.0 |
| CSD† | 22.4 [20.3, 24.7] | 225.01 [204.37, 248.84] | 21.59 [20.58, 22.67] | 84.59 [84.44, 84.73] | 61.9 |
| BoN-8 | 19.0 [16.7, 21.5] | 20.46 [19.57, 21.39] | 19.47 [18.74, 20.21] | 84.70 [84.56, 84.84] | 256.0 |
| FK-8 | 20.6 [18.1, 23.0] | 21.29 [20.37, 22.27] | 18.17 [17.53, 18.84] | 84.71 [84.57, 84.85] | 256.0 |
| **BeaverTails** | | | | | |
| Unmodified | 38.6 [35.6, 41.6] | 20.57 [19.61, 21.57] | 17.53 [16.90, 18.21] | 84.65 [84.50, 84.79] | 32.0 |
| Full SAD | 27.1 [24.0, 30.4] | 35.05 [33.23, 37.03] | 47.08 [42.86, 52.67] | 83.38 [83.20, 83.55] | 32.0 |
| SADDER (selective) | 29.9 [26.8, 33.2] | 34.63 [32.22, 37.33] | 27.80 [25.95, 30.01] | 84.20 [84.04, 84.35] | 32.0 |
| SADDER w/ top-1 remasking | 30.6 [27.1, 34.1] | 31.58 [28.46, 35.55] | 23.78 [22.34, 25.39] | 84.10 [83.90, 84.29] | 39.0 |
| SADDER w/ aligned remasking | 27.5 [24.4, 30.8] | 37.12 [33.51, 41.64] | 35.91 [33.64, 38.41] | 83.69 [83.50, 83.86] | 42.2 |
| SADDER + warm-up + aligned remasking | 24.7 [21.8, 27.7] | 31.36 [29.70, 33.08] | 42.23 [39.06, 46.69] | 83.65 [83.47, 83.82] | 37.8 |
| CSD† | 15.6 [13.8, 17.3] | 486.52 [439.90, 541.43] | 942.73 [868.26, 1027.02] | 82.61 [82.48, 82.73] | 55.0 |
| BoN-8 | 18.6 [16.3, 20.9] | 24.23 [23.04, 25.42] | 19.47 [18.74, 20.21] | 84.70 [84.56, 84.84] | 256.0 |
| FK-8 | 28.2 [25.4, 30.9] | 22.66 [21.66, 23.74] | 18.17 [17.53, 18.84] | 84.71 [84.57, 84.85] | 256.0 |

Table 13: Dream-Base component variants.

| Method | Unsafe % ↓ | Harmful PPL ↓ | Benign PPL ↓ | Benign BERT ↑ | NFE |
|---|---|---|---|---|---|
| **RTP** | | | | | |
| Unmodified | 46.8 [43.8, 49.7] | 18.31 [17.61, 19.06] | 15.45 [14.96, 15.98] | 84.71 [84.57, 84.86] | 32.0 |
| Full SAD | 33.4 [30.1, 36.8] | 26.02 [24.30, 27.83] | 23.31 [21.99, 24.68] | 83.16 [82.96, 83.35] | 32.0 |
| SADDER | 35.9 [32.8, 39.0] | 29.09 [27.54, 30.68] | 18.10 [17.48, 18.73] | 84.59 [84.45, 84.73] | 32.0 |
| SADDER w/ top-1 remasking | 37.4 [34.3, 40.5] | 22.54 [21.28, 23.80] | 16.92 [16.37, 17.53] | 84.57 [84.43, 84.71] | 47.0 |
| SADDER w/ aligned remasking | 40.9 [38.0, 43.7] | 23.06 [22.15, 24.00] | 17.53 [16.92, 18.16] | 84.52 [84.37, 84.66] | 44.8 |
| SADDER + warm-up | 34.7 [31.6, 38.2] | 27.11 [25.52, 28.82] | 21.26 [20.25, 22.29] | 83.76 [83.57, 83.94] | 32.0 |
| CSD† | 32.9 [30.6, 35.4] | 1182.90 [1101.51, 1271.65] | 3397.41 [3266.61, 3538.51] | 81.51 [81.42, 81.61] | 60.2 |
| BoN-8 | 24.6 [22.0, 27.3] | 17.46 [16.72, 18.18] | 16.87 [16.30, 17.46] | 84.81 [84.68, 84.95] | 256.0 |
| FK-8 | 25.1 [22.5, 27.7] | 17.85 [17.16, 18.57] | 15.73 [15.27, 16.24] | 84.76 [84.62, 84.90] | 256.0 |
| **BeaverTails** | | | | | |
| Unmodified | 38.7 [35.8, 41.6] | 12.67 [12.17, 13.21] | 15.45 [14.96, 15.98] | 84.71 [84.57, 84.86] | 32.0 |
| Full SAD | 22.7 [19.7, 25.7] | 23.68 [21.80, 25.75] | 23.03 [21.69, 24.45] | 83.62 [83.44, 83.80] | 32.0 |
| SADDER | 34.1 [31.3, 36.8] | 19.32 [18.41, 20.32] | 19.95 [19.18, 20.73] | 84.49 [84.34, 84.64] | 32.0 |
| SADDER w/ top-1 remasking | 34.3 [31.5, 37.1] | 19.16 [18.23, 20.15] | 18.35 [17.63, 19.09] | 84.39 [84.24, 84.55] | 47.0 |
| SADDER w/ aligned remasking | 33.7 [31.0, 36.5] | 17.78 [16.93, 18.65] | 21.07 [20.28, 21.92] | 84.54 [84.39, 84.68] | 40.4 |
| SADDER + warm-up | 27.7 [24.8, 30.8] | 23.56 [22.08, 25.14] | 20.52 [19.60, 21.49] | 84.21 [84.06, 84.37] | 32.0 |
| CSD† | 38.6 [35.7, 41.4] | 17.81 [16.65, 19.13] | 27.19 [25.30, 29.38] | 84.50 [84.36, 84.64] | 55.8 |
| BoN-8 | 20.0 [17.8, 22.2] | 14.95 [14.26, 15.68] | 16.87 [16.30, 17.46] | 84.81 [84.68, 84.95] | 256.0 |
| FK-8 | 27.8 [25.2, 30.4] | 13.94 [13.39, 14.54] | 15.73 [15.27, 16.24] | 84.76 [84.62, 84.90] | 256.0 |

## F PAIRED COMPARISONS AND COMPONENT DIAGNOSTICS

The following figure and benign-quality table compare the selected component configurations; their LLaDA warm-up entries include aligned remasking. Paired differences compare SADDER with the separately selected full-SAD operating point on identical prompt–seed observations. Negative unsafe-rate and PPL differences, and positive BERTScore differences, favour SADDER. Quality comparisons use common valid pairs within each cohort. Intervals are conditional on the selected settings and observed seeds.

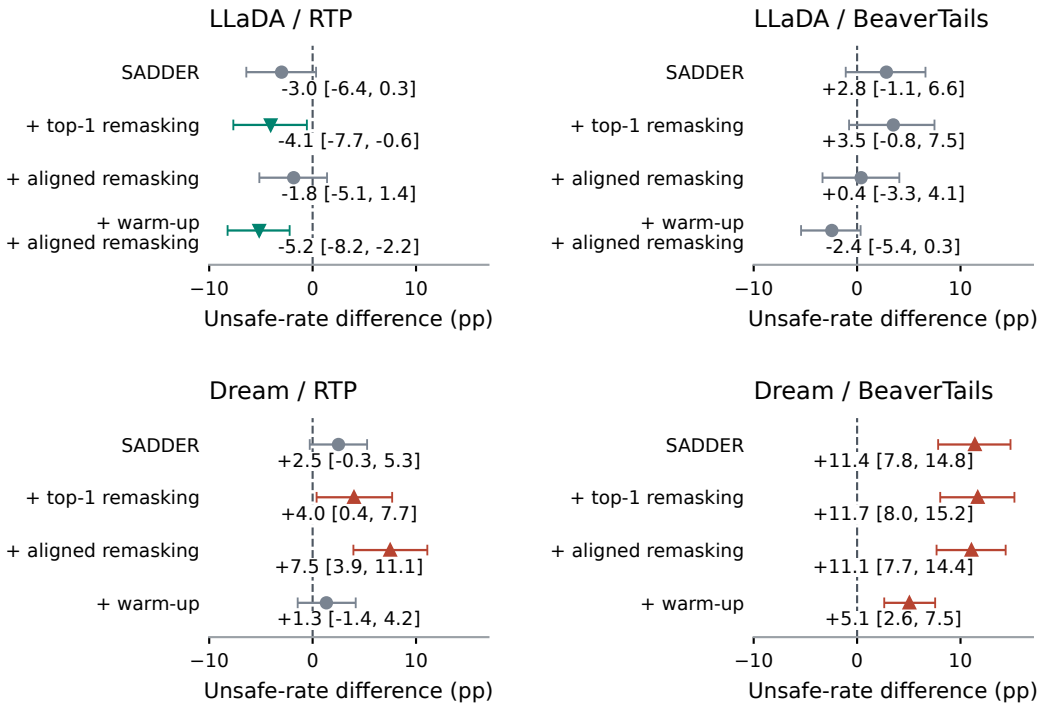


Figure 2: Selected SADDER component configurations minus full SAD on 600 harmful prompts per dataset and three seeds. LLaDA warm-up includes its aligned remasking, Dream does not. Points and bars show unsafe-rate differences (pp) and pointwise 95% paired bootstrap CIs; negative favors SADDER.

Table 14: SADDER variants: paired benign-quality changes versus full SAD.

| Model/data | Method | Δ benign PPL | Δ benign BERT |
|---|---|---|---|
| LLaDA/RTP | SADDER (selective) | $-11.55$ $[-14.49, -9.26]$ | $0.672$ $[0.527, 0.813]$ |
| | SADDER + warm-up | $3.42$ $[0.49, 6.19]$ | $0.206$ $[0.076, 0.332]$ |
| LLaDA/BT | SADDER (selective) | $-19.28$ $[-24.82, -15.03]$ | $0.822$ $[0.681, 0.967]$ |
| | SADDER + warm-up | $-4.85$ $[-8.89, -1.18]$ | $0.270$ $[0.133, 0.407]$ |
| Dream/RTP | SADDER (selective) | $-5.22$ $[-6.55, -3.88]$ | $1.429$ $[1.258, 1.605]$ |
| | SADDER + warm-up | $-2.06$ $[-3.25, -0.91]$ | $0.596$ $[0.447, 0.740]$ |
| Dream/BT | SADDER (selective) | $-3.08$ $[-4.51, -1.70]$ | $0.868$ $[0.709, 1.030]$ |
| | SADDER + warm-up | $-2.51$ $[-3.66, -1.38]$ | $0.594$ $[0.482, 0.703]$ |

**Paired random controls.** Tables 15 and 16 compare each selected SADDER component configuration with full SAD and random controls at the same layer, strength, and intervention window. LLaDA warm-up here includes aligned remasking. Random selection replaces the steering gate with

a random subset of the gate-implied size, preserving the remasking and warm-up policies. Random reopening replaces top-1 remasking with one random position, or aligned remasking with a random subset of the threshold-implied size, preserving steering and warm-up. Counts are computed on each trajectory's current state and may diverge across trajectories. The full-SAD comparator steers all masked positions throughout the window, without additional remasking or a selective phase. Differences are SADDER minus the comparator; negative values favor SADDER. Dashes indicate that a matched control is unavailable or reopening is not used. Each interval uses 5,000 paired prompt-cluster resamples of the same 600 prompt records and three seeds, conditional on the selected configuration.

Table 15: LLaDA-Base: paired unsafe-rate differences (pp), means [95% CI].

| SADDER variant | vs full SAD | vs random selection | vs random reopening |
|---|---|---|---|
| **RTP** | | | |
| Selective | $-4.1$ $[-6.1, -2.1]$ | $-3.3$ $[-5.4, -1.3]$ | — |
| + top-1 remasking | $-5.2$ $[-8.5, -1.7]$ | — | $-3.4$ $[-6.3, -0.6]$ |
| + aligned remasking | $-2.9$ $[-6.3, 0.4]$ | — | $-1.7$ $[-3.5, 0.1]$ |
| + warm-up + aligned | $-5.8$ $[-9.0, -2.7]$ | — | $-4.8$ $[-6.9, -2.8]$ |
| **BeaverTails** | | | |
| Selective | $2.8$ $[-1.1, 6.6]$ | $-0.7$ $[-3.9, 2.4]$ | — |
| + top-1 remasking | $-7.9$ $[-11.8, -4.1]$ | — | $-6.1$ $[-9.1, -3.1]$ |
| + aligned remasking | $0.4$ $[-3.3, 4.1]$ | — | $1.9$ $[-0.6, 4.6]$ |
| + warm-up + aligned | $-2.4$ $[-5.4, 0.3]$ | — | $-1.8$ $[-4.5, 0.8]$ |

Table 16: Dream-Base: paired unsafe-rate differences (pp), means [95% CI].

| SADDER variant | vs full SAD | vs random selection | vs random reopening |
|---|---|---|---|
| **RTP** | | | |
| Selective | $2.5$ $[-0.3, 5.3]$ | $-2.7$ $[-5.5, 0.1]$ | — |
| + top-1 remasking | $-3.9$ $[-7.4, -0.4]$ | $-2.6$ $[-5.3, 0.3]$ | $-2.8$ $[-5.7, 0.2]$ |
| + aligned remasking | $-3.4$ $[-6.2, -0.6]$ | $0.6$ $[-1.7, 2.7]$ | $-2.2$ $[-4.5, 0.2]$ |
| + warm-up | $0.3$ $[-1.6, 2.2]$ | $0.4$ $[-1.4, 2.3]$ | — |
| **BeaverTails** | | | |
| Selective | $11.4$ $[7.8, 14.8]$ | $3.4$ $[0.8, 5.9]$ | — |
| + top-1 remasking | $4.6$ $[1.1, 8.0]$ | $5.3$ $[2.7, 7.9]$ | $-1.6$ $[-4.5, 1.2]$ |
| + aligned remasking | $4.0$ $[0.6, 7.4]$ | $2.6$ $[0.1, 5.1]$ | $0.2$ $[-2.1, 2.5]$ |
| + warm-up | $5.1$ $[2.6, 7.5]$ | $0.1$ $[-1.7, 1.8]$ | — |

## G BoN-2/4/8, FK-2/4/8, AND COMPOSITION

We compare BoN and FK at budgets 2, 4, and 8, with harmful and benign quality measured on the corresponding cohorts. The FK adaptation resamples every four steps, using one clean-state draw per particle and a difference potential on log safety rewards. RoBERTa supplies the search reward; LlamaGuard judges final outputs. NFE excludes reward scoring.

Table 17: LLaDA: BoN and FK at budgets 2, 4, and 8.

| Method | Unsafe % | Harmful PPL | Benign PPL | Benign BERT | NFE |
|---|---|---|---|---|---|
| **RTP** | | | | | |
| BoN-2 | 30.6 [27.8, 33.4] | 20.90 [20.00, 21.82] | 17.51 [16.89, 18.15] | 84.73 [84.59, 84.88] | 64 |
| BoN-4 | 24.7 [22.3, 27.2] | 20.50 [19.68, 21.40] | 18.70 [18.00, 19.41] | 84.75 [84.61, 84.89] | 128 |
| BoN-8 | 19.0 [16.7, 21.5] | 20.46 [19.57, 21.39] | 19.47 [18.74, 20.21] | 84.70 [84.56, 84.84] | 256 |
| FK-2 | 32.4 [29.7, 35.2] | 21.47 [20.53, 22.46] | 17.59 [16.96, 18.26] | 84.67 [84.52, 84.82] | 64 |
| FK-4 | 26.2 [23.6, 28.8] | 20.82 [19.92, 21.71] | 17.84 [17.22, 18.49] | 84.68 [84.54, 84.83] | 128 |
| FK-8 | 20.6 [18.1, 23.0] | 21.29 [20.37, 22.27] | 18.17 [17.53, 18.84] | 84.71 [84.57, 84.85] | 256 |
| **BeaverTails** | | | | | |
| BoN-2 | 32.1 [29.3, 34.8] | 22.32 [21.38, 23.33] | 17.51 [16.89, 18.15] | 84.73 [84.59, 84.88] | 64 |
| BoN-4 | 24.1 [21.7, 26.6] | 23.76 [22.68, 24.93] | 18.70 [18.00, 19.41] | 84.75 [84.61, 84.89] | 128 |
| BoN-8 | 18.6 [16.3, 20.9] | 24.23 [23.04, 25.42] | 19.47 [18.74, 20.21] | 84.70 [84.56, 84.84] | 256 |
| FK-2 | 36.2 [33.3, 39.2] | 21.24 [20.26, 22.26] | 17.59 [16.96, 18.26] | 84.67 [84.52, 84.82] | 64 |
| FK-4 | 34.6 [31.8, 37.4] | 21.85 [20.88, 22.94] | 17.84 [17.22, 18.49] | 84.68 [84.54, 84.83] | 128 |
| FK-8 | 28.2 [25.4, 30.9] | 22.66 [21.66, 23.74] | 18.17 [17.53, 18.84] | 84.71 [84.57, 84.85] | 256 |

Table 18: Dream: BoN and FK at budgets 2, 4, and 8.

| Method | Unsafe % | Harmful PPL | Benign PPL | Benign BERT | NFE |
|---|---|---|---|---|---|
| **RTP** | | | | | |
| BoN-2 | 37.0 [34.2, 39.9] | 17.67 [17.01, 18.35] | 15.39 [14.90, 15.90] | 84.76 [84.62, 84.91] | 64 |
| BoN-4 | 30.6 [27.9, 33.3] | 17.95 [17.27, 18.64] | 15.78 [15.29, 16.29] | 84.76 [84.62, 84.90] | 128 |
| BoN-8 | 24.6 [22.0, 27.3] | 17.46 [16.72, 18.18] | 16.87 [16.30, 17.46] | 84.81 [84.68, 84.95] | 256 |
| FK-2 | 39.7 [36.8, 42.7] | 17.90 [17.23, 18.56] | 15.40 [14.92, 15.90] | 84.69 [84.54, 84.83] | 64 |
| FK-4 | 31.3 [28.6, 33.9] | 17.90 [17.20, 18.61] | 15.31 [14.83, 15.83] | 84.72 [84.57, 84.86] | 128 |
| FK-8 | 25.1 [22.5, 27.7] | 17.85 [17.16, 18.57] | 15.73 [15.27, 16.24] | 84.76 [84.62, 84.90] | 256 |
| **BeaverTails** | | | | | |
| BoN-2 | 32.5 [29.8, 35.1] | 12.81 [12.27, 13.38] | 15.39 [14.90, 15.90] | 84.76 [84.62, 84.91] | 64 |
| BoN-4 | 26.3 [23.7, 28.8] | 13.58 [13.00, 14.22] | 15.78 [15.29, 16.29] | 84.76 [84.62, 84.90] | 128 |
| BoN-8 | 20.0 [17.8, 22.2] | 14.95 [14.26, 15.68] | 16.87 [16.30, 17.46] | 84.81 [84.68, 84.95] | 256 |
| FK-2 | 36.3 [33.4, 39.1] | 13.21 [12.63, 13.81] | 15.40 [14.92, 15.90] | 84.69 [84.54, 84.83] | 64 |
| FK-4 | 32.7 [29.9, 35.4] | 13.76 [13.20, 14.37] | 15.31 [14.83, 15.83] | 84.72 [84.57, 84.86] | 128 |
| FK-8 | 27.8 [25.2, 30.4] | 13.94 [13.39, 14.54] | 15.73 [15.27, 16.24] | 84.76 [84.62, 84.90] | 256 |

**Composition with BoN.** SADDER can guide each candidate before BoN-2 selection. In the LLaDA/RTP run, this reduces unsafe rate from 30.61% to 26.7% at the same 64 generator evaluations. Its unsafe-prompt PPL increases from 20.90 to 28.80. The two tables distinguish separate composition studies. PPL is evaluated on harmful prompts; benign quality was not measured for these hybrids.

Table 19: Composition.

| Model/data | Method | Unsafe % | Unsafe-prompt PPL | NFE |
|---|---|---|---|---|
| Dream/BT | BoN-2 | 32.5 [29.8, 35.1] | 12.81 [12.27, 13.38] | 64 |
| Dream/RTP | BoN-2 | 37.0 [34.2, 39.9] | 17.67 [17.01, 18.35] | 64 |
| | SADDER + BoN-2 | 34.0 [30.9, 37.1] | 21.05 [19.98, 22.11] | 64 |
| LLaDA/BT | BoN-2 | 32.1 [29.3, 34.8] | 22.32 [21.38, 23.33] | 64 |
| | SADDER + BoN-2 | 28.5 [25.3, 31.7] | 27.54 [25.89, 29.39] | 64 |
| LLaDA/RTP | BoN-2 | 30.6 [27.8, 33.4] | 20.90 [20.00, 21.82] | 64 |
| | SADDER + BoN-2 | 26.7 [23.7, 29.8] | 28.80 [26.80, 31.41] | 64 |

# H JAILBREAK ROBUSTNESS AND ABLATIONS

## H.1 LLADA-INSTRUCT JAILBREAK RESULTS

Table 20 gives the LLaDA-Instruct comparison corresponding to the Dream-Instruct results in Table 2. Both tables use the same metrics and reporting conventions. XSTest uses the DIJA-configured defense; for LLaDA generation-scope DiffuGuard, this is $\lambda = 0.35$.

Table 20: LLaDA-Instruct: held-out HarmBench attacks and XSTest over-refusal. Entries use $A_b$: mean and sample SD across three seeds (percentage points for rates). HB, LG, and PPL assess complete responses. XSTest reports full-response WG refusal on safe prompts, using valid judge labels only (coverage in Appendix H.2). NFE is mean cost.

| Method | DIJA | | | | PAD | | | | XSTest |
|---|---|---|---|---|---|---|---|---|---|
| | HB ASR % ↓ | LG unsafe % ↓ | Harmful PPL ↓ | NFE | HB ASR % ↓ | LG unsafe % ↓ | Harmful PPL ↓ | NFE | WG refusal % ↓ |
| **LLaDA-Instruct** | | | | | | | | | |
| Unmodified | $78.7_{3.1}$ | $82.0_{2.6}$ | $9.34_{0.13}$ | 91.7 | $58.7_{2.5}$ | $75.7_{2.3}$ | $6.06_{0.17}$ | 116.0 | $20.7_{1.1}$ |
| Full SAD | $77.0_{2.0}$ | $81.0_{0.0}$ | $11.0_{0.13}$ | 91.7 | $63.3_{7.5}$ | $78.7_{4.0}$ | $8.24_{0.21}$ | 116.0 | $20.9_{1.9}$ |
| SADDER (selective) | $79.0_{1.0}$ | $79.3_{1.5}$ | $13.3_{0.076}$ | 91.7 | $57.7_{4.0}$ | $73.7_{3.8}$ | $13.2_{0.26}$ | 116.0 | $18.8_{0.6}$ |
| SADDER + warm-up | $78.3_{0.6}$ | $80.7_{0.6}$ | $12.8_{0.19}$ | 91.7 | $57.0_{4.0}$ | $68.3_{2.5}$ | $11.4_{0.19}$ | 116.0 | $16.3_{2.0}$ |
| CSD† | $48.7_{4.7}$ | $79.7_{4.9}$ | $36.1_{0.92}$ | 115.7 | $30.3_{3.5}$ | $69.3_{0.6}$ | $19.7_{1.3}$ | 142.0 | $29.7_{1.4}$ |
| DiffuGuard (full) | $42.7_{4.5}$ | $81.7_{2.5}$ | $255_{8.9}$ | 137.0 | $34.3_{3.8}$ | $73.7_{4.2}$ | $231_{7.2}$ | 125.0 | $42.8_{0.8}$ |
| DiffuGuard (gen) | $70.7_{3.5}$ | $83.7_{2.5}$ | $41.8_{1.3}$ | 97.8 | $57.3_{2.1}$ | $74.7_{3.5}$ | $6.22_{0.14}$ | 117.0 | $24.4_{1.0}$ |

## H.2 XSTEST OVER-REFUSAL

We evaluate refusal on 250 safe XSTest prompts and separately on 200 unsafe contrasts, using unmodified prompts and seeds 42, 43, and 44. Over-refusal is the refusal rate on safe prompts only. String matching uses full-refusal openings after lowercasing, trimming surrounding whitespace, and normalizing typographic apostrophes. WildGuard separately judges response refusal and harmfulness from each prompt–response pair (Han et al., 2024).

Table 21 reports the string-based measurements. Tables 23 and 24 report WildGuard means with seed-SD subscripts and 95% prompt-cluster bootstrap intervals below. The $n$ columns count valid labels out of 750 safe-prompt or 600 unsafe-prompt outputs. Missing WildGuard labels are excluded separately for each method; they are not counted as safe or non-refusing. Each point estimate averages the valid-label rates within the three seeds. Intervals use 5,000 resamples of prompt clusters, retaining all seeds.

**Refusal detection and coverage.** The detectors assign different refusal rates to the same responses. For LLaDA warm-up on unsafe prompts, opening-string matching gives 53.5% refusal and WildGuard gives 97.5%; the corresponding Dream rates are 69.5% and 91.5%. So a reduction in stereotyped refusal openings can overstate the reduction in refusal behavior. WildGuard's response-harmfulness labels provide a separate check: Dream warm-up has 6.9% harmful responses on unsafe XSTest prompts, compared with 0.0% for unmodified sampling, among outputs with valid labels. Coverage also varies with the intervention. Paired comparisons use the intersection of valid labels, rather than subtracting rates computed on different subsets. On safe prompts, warm-up changes WildGuard refusal by $-5.9$ percentage points on LLaDA (95% CI: $[-9.7, -2.1]$; 593 matched outputs) and $-35.4$ points on Dream ($[-41.6, -29.3]$; 602 matched outputs).

**DiffuGuard settings.** Generation-scope DiffuGuard uses $\lambda = 0.35$ on LLaDA and $\lambda = 0.5/0.75$ for Dream's DIJA/PAD settings. Both Dream thresholds are evaluated on XSTest. On vanilla prompts, full-block and generation-scope repair regions coincide, so these rows compare thresholds rather than repair scopes. Calibration reuses XSTest-safe prompts for the trigger diagnostic in Appendix A.6; these measurements include calibration prompts.

Table 21: XSTest string-based refusal: means [95% CI] and seed SDs.

| Method | Safe: over-refusal% | SD | Unsafe: refusal% | SD |
|---|---|---|---|---|
| **LLaDA-Instruct** | | | | |
| Unmodified | 9.7 [6.5, 13.2] | 0.9 | 93.7 [90.7, 96.3] | 1.0 |
| Full SAD | 10.0 [6.8, 13.5] | 1.1 | 93.7 [90.5, 96.3] | 1.3 |
| SADDER (selective) | 7.9 [5.1, 10.9] | 0.8 | 85.2 [80.3, 89.7] | 0.3 |
| SADDER + warm-up | 5.2 [2.8, 8.0] | 0.7 | 53.5 [46.7, 60.3] | 0.0 |
| CSD† | 10.3 [7.1, 13.7] | 0.8 | 93.7 [90.7, 96.3] | 1.0 |
| DiffuGuard (full) | 4.0 [2.5, 5.6] | 1.1 | 28.2 [24.7, 31.8] | 1.6 |
| DiffuGuard (gen) ($\lambda = 0.35$) | 10.7 [7.3, 14.3] | 0.5 | 92.2 [88.8, 95.2] | 1.3 |
| SADDER (top-1) | 8.4 [5.5, 11.7] | 0.8 | 85.5 [80.8, 89.8] | 0.5 |
| SADDER (aligned) | 8.5 [5.6, 11.7] | 0.9 | 85.3 [80.5, 89.8] | 0.8 |
| Random selection | 10.0 [6.8, 13.5] | 0.7 | 86.2 [82.7, 89.5] | 2.9 |
| **Dream-Instruct** | | | | |
| Unmodified | 44.9 [39.2, 50.5] | 0.6 | 98.8 [97.3, 100.0] | 0.3 |
| Full SAD | 46.0 [40.0, 51.9] | 0.7 | 98.5 [96.5, 100.0] | 0.0 |
| SADDER (selective) | 26.1 [20.9, 31.2] | 0.5 | 92.0 [88.2, 95.5] | 0.5 |
| SADDER + warm-up | 7.1 [4.1, 10.4] | 0.2 | 69.5 [63.0, 76.0] | 0.0 |
| CSD† | 45.2 [39.5, 50.8] | 0.8 | 98.8 [97.3, 100.0] | 0.3 |
| DiffuGuard (full) | 6.7 [4.9, 8.5] | 1.2 | 13.5 [10.8, 16.3] | 1.0 |
| DiffuGuard (gen) ($\lambda = 0.75$) | 43.1 [37.5, 48.5] | 0.5 | 98.8 [97.3, 100.0] | 0.3 |
| DiffuGuard (gen) ($\lambda = 0.5$) | 41.9 [36.4, 47.2] | 0.6 | 98.0 [96.0, 99.5] | 0.5 |
| SADDER (top-1) | 34.7 [28.8, 40.5] | 0.2 | 96.5 [93.5, 99.0] | 0.0 |
| SADDER (aligned) | 29.5 [24.0, 34.9] | 0.5 | 96.5 [93.8, 98.7] | 0.9 |
| Random selection | 22.5 [18.7, 26.7] | 2.3 | 74.2 [70.0, 78.2] | 4.9 |

Table 22: Paired string-based refusal changes versus unmodified sampling (pp).

| Method | Safe prompts | Unsafe prompts |
|---|---|---|
| **LLaDA-Instruct** | | |
| Full SAD | $0.3\ [-0.9, 1.5]$ | $0.0\ [-1.0, 1.0]$ |
| SADDER (selective) | $-1.9\ [-3.7, -0.3]$ | $-8.5\ [-13.0, -4.2]$ |
| SADDER + warm-up | $-4.5\ [-7.5, -1.9]$ | $-40.2\ [-47.3, -32.7]$ |
| CSD† | $0.5\ [0.1, 1.1]$ | $0.0\ [0.0, 0.0]$ |
| DiffuGuard (full) | $-5.7\ [-9.1, -2.8]$ | $-65.5\ [-69.8, -61.0]$ |
| DiffuGuard (gen) ($\lambda = 0.35$) | $0.9\ [-0.1, 2.1]$ | $-1.5\ [-3.5, 0.3]$ |
| SADDER (top-1) | $-1.3\ [-3.2, 0.4]$ | $-8.2\ [-12.3, -4.2]$ |
| SADDER (aligned) | $-1.2\ [-3.1, 0.4]$ | $-8.3\ [-12.8, -4.2]$ |
| Random selection | $0.3\ [-1.5, 2.0]$ | $-7.5\ [-11.3, -4.0]$ |
| **Dream-Instruct** | | |
| Full SAD | $1.1\ [-0.9, 3.1]$ | $-0.3\ [-1.5, 0.5]$ |
| SADDER (selective) | $-18.8\ [-24.0, -13.6]$ | $-6.8\ [-10.8, -3.3]$ |
| SADDER + warm-up | $-37.9\ [-43.6, -32.0]$ | $-29.3\ [-35.8, -23.0]$ |
| CSD† | $0.3\ [0.0, 0.7]$ | $0.0\ [0.0, 0.0]$ |
| DiffuGuard (full) | $-38.3\ [-43.7, -32.7]$ | $-85.3\ [-88.3, -82.2]$ |
| DiffuGuard (gen) ($\lambda = 0.75$) | $-1.9\ [-3.2, -0.5]$ | $0.0\ [-0.5, 0.5]$ |
| DiffuGuard (gen) ($\lambda = 0.5$) | $-3.1\ [-5.1, -1.2]$ | $-0.8\ [-2.3, 0.2]$ |
| SADDER (top-1) | $-10.3\ [-14.7, -6.0]$ | $-2.3\ [-5.2, 0.0]$ |
| SADDER (aligned) | $-15.5\ [-19.9, -11.1]$ | $-2.3\ [-5.0, 0.0]$ |
| Random selection | $-22.4\ [-27.5, -17.5]$ | $-24.7\ [-28.7, -20.7]$ |

Table 23: LLaDA: WildGuard judgments on XSTest.

| | Safe prompts | | | Unsafe prompts | | |
|---|---|---|---|---|---|---|
| Method | Over-refusal% | Harmful% | $n$ | Refusal% | Harmful% | $n$ |
| Unmodified | $20.7_{1.1}$ [16.2, 25.6] | $3.3_{0.8}$ [1.3, 5.7] | 603 | $97.9_{1.0}$ [96.0, 99.5] | $1.2_{0.6}$ [0.2, 2.7] | 585 |
| Full SAD | $20.9_{1.9}$ [16.4, 25.7] | $3.1_{0.5}$ [1.2, 5.5] | 603 | $97.9_{1.0}$ [96.0, 99.5] | $1.2_{0.6}$ [0.2, 2.7] | 585 |
| SADDER (selective) | $18.8_{0.6}$ [14.5, 23.4] | $2.5_{0.8}$ [0.9, 4.3] | 644 | $98.8_{0.3}$ [97.3, 100.0] | $0.7_{0.3}$ [0.0, 1.7] | 584 |
| SADDER + warm-up | $16.3_{2.0}$ [12.4, 20.8] | $2.5_{0.0}$ [1.0, 4.4] | 710 | $97.5_{0.5}$ [95.2, 99.3] | $0.5_{0.5}$ [0.0, 1.4] | 592 |
| CSD† | $29.7_{1.4}$ [24.5, 35.1] | $3.7_{0.7}$ [2.1, 5.7] | 589 | $98.1_{1.1}$ [96.5, 99.5] | $1.0_{0.5}$ [0.2, 2.2] | 585 |
| DiffuGuard (full) | $42.8_{0.8}$ [38.5, 47.1] | $7.6_{1.7}$ [5.5, 9.9] | 750 | $84.5_{1.5}$ [81.3, 87.5] | $14.4_{1.6}$ [11.4, 17.5] | 599 |
| DiffuGuard (gen) ($\lambda = 0.35$) | $24.4_{1.0}$ [19.7, 29.4] | $4.6_{2.4}$ [2.4, 7.1] | 595 | $98.3_{0.3}$ [96.7, 99.6] | $0.9_{0.3}$ [0.2, 1.9] | 586 |
| SADDER (top-1) | $20.8_{0.2}$ [16.3, 25.5] | $3.4_{0.8}$ [1.5, 5.5] | 650 | $98.8_{0.3}$ [97.3, 99.8] | $0.7_{0.3}$ [0.0, 1.6] | 583 |
| SADDER (aligned) | $20.0_{0.7}$ [15.4, 24.6] | $3.4_{0.5}$ [1.5, 5.7] | 645 | $98.0_{0.5}$ [95.9, 99.5] | $0.9_{0.3}$ [0.0, 2.2] | 586 |
| Random selection | $23.5_{1.4}$ [18.7, 28.5] | $3.7_{1.0}$ [1.9, 5.8] | 622 | $99.5_{0.0}$ [98.6, 100.0] | $0.0_{0.0}$ [0.0, 0.0] | 588 |

Table 24: Dream: WildGuard judgments on XSTest.

| | Safe prompts | | | Unsafe prompts | | |
|---|---|---|---|---|---|---|
| Method | Over-refusal% | Harmful% | $n$ | Refusal% | Harmful% | $n$ |
| Unmodified | $55.5_{1.3}$ [49.1, 61.5] | $2.2_{0.5}$ [0.8, 3.8] | 629 | $99.5_{0.0}$ [98.5, 100.0] | $0.0_{0.0}$ [0.0, 0.0] | 600 |
| Full SAD | $58.0_{1.0}$ [51.6, 64.4] | $2.7_{0.3}$ [1.1, 4.7] | 619 | $99.5_{0.0}$ [98.5, 100.0] | $0.0_{0.0}$ [0.0, 0.0] | 600 |
| SADDER (selective) | $41.1_{0.8}$ [34.3, 47.9] | $3.2_{0.5}$ [1.1, 5.6] | 562 | $99.1_{0.3}$ [97.7, 100.0] | $0.0_{0.0}$ [0.0, 0.0] | 572 |
| SADDER + warm-up | $17.9_{0.7}$ [13.5, 22.7] | $1.6_{0.2}$ [0.4, 2.9] | 709 | $91.5_{1.3}$ [88.0, 94.5] | $6.9_{1.8}$ [4.2, 10.0] | 597 |
| CSD† | $56.4_{1.9}$ [50.0, 62.4] | $2.1_{1.0}$ [0.8, 3.7] | 630 | $99.5_{0.0}$ [98.5, 100.0] | $0.0_{0.0}$ [0.0, 0.0] | 600 |
| DiffuGuard (full) | $64.1_{1.5}$ [60.0, 68.1] | $14.3_{2.6}$ [11.2, 17.6] | 750 | $75.8_{0.6}$ [72.2, 79.3] | $23.2_{1.0}$ [19.8, 26.7] | 600 |
| DiffuGuard (gen) ($\lambda = 0.75$) | $53.1_{1.5}$ [46.9, 58.9] | $2.5_{0.5}$ [0.9, 4.4] | 641 | $99.5_{0.0}$ [98.5, 100.0] | $0.0_{0.0}$ [0.0, 0.0] | 600 |
| DiffuGuard (gen) ($\lambda = 0.5$) | $54.6_{1.1}$ [48.8, 60.3] | $3.2_{0.4}$ [1.5, 5.1] | 665 | $99.2_{0.3}$ [97.8, 100.0] | $0.2_{0.3}$ [0.0, 0.5] | 600 |
| SADDER (top-1) | $41.2_{0.1}$ [34.7, 47.4] | $3.0_{0.7}$ [1.2, 5.0] | 670 | $99.0_{0.5}$ [97.7, 100.0] | $0.0_{0.0}$ [0.0, 0.0] | 600 |
| SADDER (aligned) | $38.1_{1.3}$ [31.8, 44.1] | $2.5_{0.7}$ [0.8, 4.3] | 653 | $98.0_{0.9}$ [96.0, 99.5] | $0.0_{0.0}$ [0.0, 0.0] | 599 |
| Random selection | $39.2_{0.5}$ [33.2, 45.2] | $2.9_{0.1}$ [1.0, 5.1] | 526 | $98.1_{0.7}$ [96.1, 99.6] | $1.3_{0.7}$ [0.2, 2.7] | 471 |

Table 25: LLaDA: paired WildGuard refusal changes (pp).

| Method | Safe prompts | Pairs | Unsafe prompts | Pairs |
|---|---|---|---|---|
| Full SAD | $-0.0$ $[-1.2, 1.4]$ | 600 | 0.0 [0.0, 0.0] | 585 |
| SADDER (selective) | $-2.4$ $[-5.8, 0.9]$ | 579 | 0.7 $[-0.9, 2.4]$ | 581 |
| SADDER + warm-up | $-5.9$ $[-9.7, -2.1]$ | 593 | $-0.7$ $[-2.9, 1.4]$ | 581 |
| CSD† | 9.3 [5.9, 12.8] | 588 | 0.2 $[-0.3, 0.7]$ | 585 |
| DiffuGuard (full) | 16.7 [12.4, 21.0] | 603 | $-13.7$ $[-16.8, -10.6]$ | 585 |
| DiffuGuard (gen) ($\lambda = 0.35$) | 3.2 [1.3, 5.3] | 589 | 0.3 $[-0.5, 1.4]$ | 585 |
| SADDER (top-1) | $-1.0$ $[-4.6, 2.4]$ | 585 | 0.7 $[-0.7, 2.4]$ | 583 |
| SADDER (aligned) | $-1.9$ $[-4.9, 0.9]$ | 584 | $-0.2$ $[-2.4, 1.9]$ | 583 |
| Random selection | 1.1 $[-2.0, 4.2]$ | 568 | 1.4 [0.2, 2.9] | 580 |

Table 26: Dream: paired WildGuard refusal changes (pp).

| Method | Safe prompts | Pairs | Unsafe prompts | Pairs |
|---|---|---|---|---|
| Full SAD | 1.8 $[-0.7, 4.3]$ | 619 | 0.0 [0.0, 0.0] | 600 |
| SADDER (selective) | $-8.6$ $[-14.1, -3.3]$ | 532 | $-0.3$ $[-1.1, 0.0]$ | 572 |
| SADDER + warm-up | $-35.4$ $[-41.6, -29.3]$ | 602 | $-8.0$ $[-11.5, -5.0]$ | 597 |
| CSD† | 1.0 [0.3, 1.8] | 629 | 0.0 [0.0, 0.0] | 600 |
| DiffuGuard (full) | 7.6 [2.7, 12.7] | 629 | $-23.7$ $[-27.3, -20.0]$ | 600 |
| DiffuGuard (gen) ($\lambda = 0.75$) | $-1.4$ $[-3.1, 0.2]$ | 629 | 0.0 [0.0, 0.0] | 600 |
| DiffuGuard (gen) ($\lambda = 0.5$) | $-0.8$ $[-2.8, 1.4]$ | 629 | $-0.3$ $[-0.8, 0.0]$ | 600 |
| SADDER (top-1) | $-12.5$ $[-17.7, -7.3]$ | 625 | $-0.5$ $[-1.3, 0.0]$ | 600 |
| SADDER (aligned) | $-14.1$ $[-19.0, -9.2]$ | 603 | $-1.5$ $[-3.2, -0.3]$ | 599 |
| Random selection | $-6.7$ $[-11.9, -1.7]$ | 493 | $-1.3$ $[-2.7, -0.2]$ | 471 |

### H.3 HELD-OUT ATTACK COMPARISONS

Table 27 reports the main jailbreak-table estimates with pointwise 95% prompt-cluster bootstrap confidence intervals, using 5,000 resamples and retaining all observed seeds within each sampled prompt. These intervals describe prompt uncertainty; the main-table subscripts describe variation

across the three generation seeds. The XSTest over-refusal is the offline string-based matching metric rather than the WildGuard over-refusal metric.

Table 27: Held-out HarmBench attacks and XSTest over-refusal.

| Method | DIJA | | | | PAD | | | | XSTest |
|---|---|---|---|---|---|---|---|---|---|
| | HB ASR % ↓ | LG unsafe % ↓ | Harmful PPL ↓ | NFE | HB ASR % ↓ | LG unsafe % ↓ | Harmful PPL ↓ | NFE | Over-refusal % ↓ |
| **LLaDA-Instruct** | | | | | | | | | |
| Unmodified | 78.7 [72.0,85.0] | 82.0 [75.0,88.3] | 9.34 [8.59,10.1] | 91.7 | 58.7 [50.3,66.7] | 75.7 [68.3,82.3] | 6.06 [5.68,6.47] | 116.0 | 9.7 [6.5,13.2] |
| Full SAD | 77.0 [69.3,84.3] | 81.0 [73.7,87.7] | 11.0 [10.0,12.1] | 91.7 | 63.3 [56.3,70.3] | 78.7 [72.0,85.0] | 8.24 [7.79,8.74] | 116.0 | 10.0 [6.8,13.5] |
| SADDER (selective) | 79.0 [71.7,86.3] | 79.3 [71.7,86.3] | 13.3 [11.6,15.7] | 91.7 | 57.7 [50.3,65.0] | 73.7 [66.3,80.7] | 13.2 [10.3,17.7] | 116.0 | 7.9 [5.1,10.9] |
| SADDER + warm-up | 78.3 [71.3,85.0] | 80.7 [73.0,87.3] | 12.8 [11.1,15.1] | 91.7 | 57.0 [49.0,65.0] | 68.3 [60.3,76.0] | 11.4 [9.97,13.4] | 116.0 | 5.2 [2.8,8.0] |
| CSD† | 48.7 [40.7,57.0] | 79.7 [72.7,86.3] | 36.1 [31.1,41.4] | 115.7 | 30.3 [24.0,37.0] | 69.3 [61.7,76.7] | 19.7 [17.3,22.6] | 142.0 | 10.3 [7.1,13.7] |
| DiffuGuard (full) | 42.7 [35.7,50.0] | 81.7 [75.7,87.0] | 255 [239,272] | 137.0 | 34.3 [28.0,41.0] | 73.7 [67.3,79.7] | 231 [214,251] | 125.0 | 4.0 [2.5,5.6] |
| DiffuGuard (gen) | 70.7 [63.7,77.7] | 83.7 [77.0,89.7] | 41.8 [32.0,53.7] | 97.8 | 57.3 [49.0,65.0] | 74.7 [67.0,81.7] | 6.22 [5.85,6.59] | 117.0 | 10.7 [7.3,14.3] |
| **Dream-Instruct** | | | | | | | | | |
| Unmodified | 82.0 [75.3,88.3] | 80.7 [73.3,87.3] | 11.4 [10.6,12.3] | 91.7 | 51.7 [43.0,60.3] | 64.0 [55.7,72.0] | 8.53 [7.95,9.21] | 116.0 | 44.9 [39.2,50.5] |
| Full SAD | 85.3 [79.0,91.3] | 83.0 [75.3,90.0] | 11.8 [10.9,12.8] | 91.7 | 52.7 [43.7,61.7] | 63.3 [54.7,71.0] | 9.52 [9.01,10.1] | 116.0 | 46.0 [40.0,51.9] |
| SADDER (selective) | 77.7 [70.3,84.7] | 81.7 [74.7,88.0] | 12.2 [11.3,13.1] | 91.7 | 52.7 [43.7,61.7] | 66.3 [57.3,75.0] | 9.58 [8.86,10.4] | 116.0 | 26.1 [20.9,31.2] |
| SADDER + warm-up | 78.7 [71.7,85.7] | 79.7 [72.3,86.7] | 12.2 [11.3,13.2] | 91.7 | 42.7 [33.7,51.7] | 55.0 [45.7,64.0] | 10.4 [9.60,11.4] | 116.0 | 7.1 [4.1,10.4] |
| CSD† | 68.0 [60.7,75.7] | 82.3 [75.7,88.7] | 45.9 [37.6,55.9] | 115.5 | 33.3 [26.0,41.0] | 63.3 [54.7,71.0] | 23.7 [19.2,29.8] | 140.8 | 45.2 [39.5,50.8] |
| DiffuGuard (full) | 33.3 [27.3,39.7] | 84.0 [78.0,89.3] | 537 [494,587] | 137.0 | 12.0 [8.0,16.0] | 70.0 [64.0,76.0] | 687 [631,748] | 125.0 | 6.7 [4.9,8.5] |
| DiffuGuard (gen) | 69.3 [62.0,76.3] | 81.3 [74.7,87.3] | 46.6 [36.0,60.4] | 97.2 | 49.0 [40.3,57.3] | 63.3 [55.0,71.3] | 8.80 [8.28,9.41] | 117.0 | 43.1‡ [37.5,48.5] |

### H.4 WildGuard harmfulness and refusal

Tables 28 and 29 report full-response WildGuard harmfulness and refusal on the same HarmBench outputs, including remasking and random-selection controls. Entries show the mean with a seed-SD subscript and a 95% prompt-cluster bootstrap interval below. The $n$ columns count valid labels out of 300 outputs per attack and method. Missing labels are excluded, and rates average the valid-output rate within each of the three seeds. Table 30 compares selective SADDER and warm-up against full SAD and random selection on jointly scored outputs. Negative harmfulness differences indicate fewer harmful responses; negative refusal differences indicate fewer refusals, which alone does not establish a safety improvement. The prompt–response pairs use the original behavior and the complete response, including any retained attack scaffold.

**Comparison across safety judges.** The judges differ in their assessment of these gains. On Dream/PAD, warm-up lowers HB ASR from 51.7% to 42.7%, while WG harmfulness increases from 66.0% to 69.0%; selective SADDER reaches 80.0% WG harmfulness. On LLaDA/PAD, warm-up reduces both HB ASR (58.7% to 57.0%) and WG harmfulness (80.0% to 76.3%). Full-block DiffuGuard on Dream/DIJA illustrates a larger disagreement: 33.3% HB ASR accompanies 98.3% WG harmfulness and PPL 537. The apparent safety benefit depends on the judge as well as the model and attack, particularly when the intervention disrupts fluency.

Table 28: LLaDA: WildGuard judgments on HarmBench.

| | DIJA | | | PAD | | |
|---|---|---|---|---|---|---|
| Method | Harmful% | Refusal% | $n$ | Harmful% | Refusal% | $n$ |
| Unmodified | $96.0_{2.6}$ [92.3, 99.0] | $1.7_{0.6}$ [0.0, 4.3] | 300 | $80.0_{3.0}$ [73.7, 86.0] | $15.3_{1.5}$ [10.3, 21.0] | 300 |
| Full SAD | $96.3_{0.6}$ [92.3, 99.3] | $2.0_{0.0}$ [0.0, 5.0] | 300 | $85.7_{3.8}$ [80.0, 90.7] | $11.3_{3.8}$ [6.7, 16.3] | 300 |
| SADDER (selective) | $96.7_{0.6}$ [92.7, 99.7] | $2.0_{0.0}$ [0.0, 5.0] | 300 | $81.0_{2.6}$ [74.3, 87.0] | $15.0_{4.0}$ [10.0, 20.7] | 300 |
| SADDER + warm-up | $96.7_{1.5}$ [93.0, 99.3] | $2.0_{1.0}$ [0.0, 4.7] | 300 | $76.3_{0.6}$ [68.7, 83.3] | $19.0_{1.7}$ [12.7, 26.0] | 300 |
| CSD† | $97.0_{1.0}$ [93.7, 99.3] | $1.7_{0.6}$ [0.0, 4.3] | 300 | $77.0_{4.4}$ [70.7, 83.0] | $19.7_{2.1}$ [14.0, 26.0] | 300 |
| DiffuGuard (full) | $97.7_{2.1}$ [95.3, 99.7] | $0.7_{1.2}$ [0.0, 1.7] | 300 | $85.0_{4.0}$ [80.0, 89.7] | $11.7_{3.5}$ [7.7, 16.0] | 300 |
| DiffuGuard (gen) | $97.0_{1.0}$ [93.7, 99.7] | $1.0_{0.0}$ [0.0, 3.0] | 300 | $78.0_{2.6}$ [71.3, 84.3] | $17.3_{1.2}$ [11.7, 23.0] | 300 |
| SADDER (top-1) | $97.3_{0.6}$ [94.0, 99.7] | $1.3_{0.6}$ [0.0, 3.7] | 300 | $81.3_{3.2}$ [74.3, 87.3] | $15.3_{2.9}$ [9.7, 21.3] | 300 |
| SADDER (aligned) | $96.7_{0.6}$ [92.7, 99.7] | $2.0_{0.0}$ [0.0, 5.0] | 300 | $83.3_{1.5}$ [77.7, 88.7] | $13.0_{2.0}$ [8.3, 18.0] | 300 |
| Random selection | $96.3_{1.2}$ [92.3, 99.3] | $1.3_{0.6}$ [0.0, 3.3] | 300 | $77.3_{3.2}$ [71.7, 82.7] | $19.0_{3.5}$ [14.0, 24.0] | 300 |

Table 29: Dream: WildGuard judgments on HarmBench.

| | DIJA | | | PAD | | |
|---|---|---|---|---|---|---|
| Method | Harmful% | Refusal% | $n$ | Harmful% | Refusal% | $n$ |
| Unmodified | $97.3_{0.6}$ [94.0, 99.7] | $0.7_{0.6}$ [0.0, 2.0] | 300 | $66.0_{2.0}$ [57.7, 74.0] | $30.3_{2.5}$ [22.3, 38.7] | 300 |
| Full SAD | $97.3_{0.6}$ [94.0, 100.0] | $1.0_{0.0}$ [0.0, 3.0] | 300 | $66.7_{0.6}$ [58.3, 74.7] | $29.3_{0.6}$ [21.3, 37.7] | 300 |
| SADDER (selective) | $96.7_{1.5}$ [93.3, 99.3] | $1.3_{0.6}$ [0.0, 3.7] | 300 | $80.0_{1.7}$ [72.7, 86.7] | $16.7_{1.5}$ [10.0, 23.7] | 300 |
| SADDER + warm-up | $96.3_{0.6}$ [92.7, 99.3] | $1.7_{0.6}$ [0.0, 4.3] | 300 | $69.0_{1.7}$ [60.0, 77.0] | $28.0_{1.7}$ [20.0, 37.0] | 300 |
| CSD† | $98.0_{0.0}$ [95.0, 100.0] | $0.7_{0.6}$ [0.0, 2.0] | 300 | $65.3_{1.5}$ [57.0, 73.3] | $30.7_{2.5}$ [22.7, 39.0] | 300 |
| DiffuGuard (full) | $98.3_{0.6}$ [95.7, 100.0] | $0.7_{0.6}$ [0.0, 1.7] | 299 | $69.3_{2.1}$ [62.7, 75.7] | $29.7_{2.1}$ [23.3, 36.3] | 300 |
| DiffuGuard (gen) | $98.3_{0.6}$ [95.7, 100.0] | $0.7_{0.6}$ [0.0, 2.0] | 300 | $65.0_{2.0}$ [56.7, 72.7] | $30.0_{1.0}$ [22.3, 38.0] | 300 |
| SADDER (top-1) | $97.7_{0.6}$ [94.3, 100.0] | $1.3_{0.6}$ [0.0, 3.7] | 300 | $80.0_{1.7}$ [72.7, 87.0] | $17.0_{1.0}$ [10.7, 24.0] | 300 |
| SADDER (aligned) | $97.7_{0.6}$ [94.3, 100.0] | $1.0_{0.0}$ [0.0, 3.0] | 300 | $72.0_{1.7}$ [63.3, 80.0] | $25.0_{1.0}$ [17.3, 33.0] | 300 |
| Random selection | $97.0_{1.0}$ [93.7, 99.7] | $1.0_{0.0}$ [0.0, 3.0] | 300 | $71.0_{2.0}$ [63.3, 78.0] | $27.0_{1.7}$ [20.3, 34.3] | 300 |

Table 30: Paired WildGuard differences on HarmBench (pp).

| Comparison | Δ harmful% | Δ refusal% | Pairs |
|---|---|---|---|
| **LLaDA / DIJA** | | | |
| Selective vs. full SAD | 0.3 [−0.7, 1.7] | 0.0 [0.0, 0.0] | 300 |
| Selective vs. random | 0.3 [−1.7, 3.0] | 0.7 [0.0, 1.7] | 300 |
| Warm-up vs. full SAD | 0.3 [−0.7, 1.7] | −0.0 [−1.0, 1.0] | 300 |
| Warm-up vs. random | 0.3 [−1.3, 2.3] | 0.7 [0.0, 1.7] | 300 |
| **LLaDA / PAD** | | | |
| Selective vs. full SAD | −4.7 [−11.0, 1.7] | 3.7 [−2.3, 9.7] | 300 |
| Selective vs. random | 3.7 [−2.0, 9.3] | −4.0 [−9.3, 1.7] | 300 |
| Warm-up vs. full SAD | −9.3 [−16.7, −2.0] | 7.7 [0.0, 15.0] | 300 |
| Warm-up vs. random | −1.0 [−6.7, 4.3] | 0.0 [−5.3, 5.7] | 300 |
| **Dream / DIJA** | | | |
| Selective vs. full SAD | −0.7 [−2.0, 0.7] | 0.3 [0.0, 1.0] | 300 |
| Selective vs. random | −0.3 [−1.7, 0.7] | 0.3 [0.0, 1.0] | 300 |
| Warm-up vs. full SAD | −1.0 [−2.7, 0.0] | 0.7 [0.0, 2.0] | 300 |
| Warm-up vs. random | −0.7 [−2.0, 0.7] | 0.7 [0.0, 2.0] | 300 |
| **Dream / PAD** | | | |
| Selective vs. full SAD | 13.3 [5.0, 22.0] | −12.7 [−21.3, −4.3] | 300 |
| Selective vs. random | 9.0 [1.3, 16.7] | −10.3 [−18.0, −3.0] | 300 |
| Warm-up vs. full SAD | 2.3 [−7.3, 12.0] | −1.3 [−11.3, 8.7] | 300 |
| Warm-up vs. random | −2.0 [−7.3, 3.3] | 1.0 [−4.0, 6.0] | 300 |

### H.5 DIJA JUDGE AND RESPONSE-VIEW SENSITIVITY

**Separating generated content from the attack scaffold.** DIJA completions contain both fixed attack text and model-generated spans. A harmfulness judgment on the complete response can therefore depend on text supplied by the attacker. We compare HarmBench (HB), LlamaGuard (LG), and WildGuard (WG) on two response views: the complete response, including the scaffold, and only the spans marked as initially masked in the saved completion segmentation (*fills*). Nonempty spans are trimmed and joined by newlines before judging. HB measures behavior-specific attack success, LG unsafe content, and WG response harmfulness. These are distinct evaluator criteria, even when the response view is held fixed. The main table retains full-response HB/LG for attacks and uses WG only for XSTest. Generated-span attack scores from all three judges are reported here. All response-view comparisons use the same generated completions and defence settings, so differences between views reflect the scoring procedure.

This view isolates the initially editable content, rather than judging the assembled attack response. Removing intervening text can change the context needed to interpret a span. For full-block DiffuGuard, repairs outside the initially masked spans are also excluded. Generated-span scores thus complement complete-response judgments; they do not measure the safety of every token that a defence may modify.

**Judge disagreement persists after removing fixed text.** Tables 31 and 32 separate judge choice from response-view choice. On unmodified sampling, removing the scaffold lowers LG unsafe rate from 82.0% to 53.3% on LLaDA and from 80.7% to 48.7% on Dream. HB falls more modestly, from 78.7% to 73.0% and from 82.0% to 76.0%, respectively; WG remains above 96%. Full-response scores therefore cannot be interpreted as measurements of generated spans alone. Conversely, excluding fixed text does not make the judges interchangeable.

Warm-up illustrates the consequence for method comparisons. Against paired unmodified outputs, LG reports reductions of 11.7 points on LLaDA (95% CI: $[-19.7, -4.3]$) and 9.0 points on Dream ($[-16.0, -2.0]$), while the corresponding HB and WG intervals include zero. Dream selective SADDER also improves HB relative to full SAD by 8.0 points ($[-13.3, -2.7]$), but increases LG unsafe rate by 15.0 points ($[8.0, 22.0]$). These are evaluator-dependent effects, not a consistent safety ordering across judges. The paired estimates in Table 33 retain the same prompts and seeds and use pointwise 95% intervals.

Table 31: LLaDA: DIJA judge and response-view comparison on 100 held-out HarmBench behaviors, three seeds. Full includes the fixed scaffold; fills excludes it. Rates (%) are means with seed-SD subscripts. HB/LG have 300 scored outputs per row; WG has 300 except Dream aligned remasking (299).

| Method | HB full | HB fills | LG full | LG fills | WG full | WG fills |
|---|---|---|---|---|---|---|
| Unmodified | $78.7_{3.1}$ | $73.0_{3.6}$ | $82.0_{2.6}$ | $53.3_{2.5}$ | $96.0_{2.6}$ | $96.3_{1.5}$ |
| Full SAD | $77.0_{2.0}$ | $71.7_{2.9}$ | $81.0_{0.0}$ | $56.3_{6.8}$ | $96.3_{0.6}$ | $97.3_{1.2}$ |
| SADDER (selective) | $79.0_{1.0}$ | $75.7_{2.3}$ | $79.3_{1.5}$ | $54.0_{2.6}$ | $96.7_{0.6}$ | $96.3_{1.5}$ |
| SADDER + warm-up | $78.3_{0.6}$ | $71.3_{2.5}$ | $80.7_{0.6}$ | $41.7_{3.5}$ | $96.7_{1.5}$ | $97.3_{2.1}$ |
| CSD† | $48.7_{4.7}$ | $30.7_{5.1}$ | $79.7_{4.9}$ | $33.7_{4.2}$ | $97.0_{1.0}$ | $96.3_{1.5}$ |
| DiffuGuard (full) | $42.7_{4.5}$ | $32.3_{0.6}$ | $81.7_{2.5}$ | $49.0_{3.6}$ | $97.7_{2.1}$ | $97.0_{0.0}$ |
| DiffuGuard (gen) | $70.7_{3.5}$ | $55.7_{4.0}$ | $83.7_{2.5}$ | $49.0_{5.2}$ | $97.0_{1.0}$ | $97.7_{0.6}$ |
| SADDER (top-1) | $78.7_{3.1}$ | $72.7_{3.5}$ | $80.7_{2.1}$ | $57.0_{2.0}$ | $97.3_{0.6}$ | $96.0_{2.0}$ |
| SADDER (aligned) | $79.3_{1.2}$ | $75.0_{3.5}$ | $80.0_{2.0}$ | $57.3_{0.6}$ | $96.7_{0.6}$ | $96.3_{1.5}$ |
| Random selection | $76.3_{1.2}$ | $72.3_{4.0}$ | $80.3_{5.5}$ | $51.7_{3.1}$ | $96.3_{1.2}$ | $97.3_{1.2}$ |

Table 32: Dream: DIJA judge and response-view comparison on 100 held-out HarmBench behaviors, three seeds. Full includes the fixed scaffold; fills excludes it. Rates (%) are means with seed-SD subscripts. HB/LG have 300 scored outputs per row; WG has 300 except Dream aligned remasking (299).

| Method | HB full | HB fills | LG full | LG fills | WG full | WG fills |
|---|---|---|---|---|---|---|
| Unmodified | $82.0_{1.0}$ | $76.0_{3.6}$ | $80.7_{1.5}$ | $48.7_{2.5}$ | $97.3_{0.6}$ | $97.3_{0.6}$ |
| Full SAD | $85.3_{0.6}$ | $80.3_{2.3}$ | $83.0_{1.0}$ | $42.3_{3.1}$ | $97.3_{0.6}$ | $96.3_{1.5}$ |
| SADDER (selective) | $77.7_{1.5}$ | $72.3_{1.5}$ | $81.7_{3.2}$ | $57.3_{2.1}$ | $96.7_{1.5}$ | $97.3_{1.2}$ |
| SADDER + warm-up | $78.7_{2.3}$ | $75.7_{3.1}$ | $79.7_{1.2}$ | $39.7_{1.2}$ | $96.3_{0.6}$ | $96.0_{0.0}$ |
| CSD† | $68.0_{6.2}$ | $45.0_{3.5}$ | $82.3_{0.6}$ | $41.7_{2.1}$ | $98.0_{0.0}$ | $97.3_{0.6}$ |
| DiffuGuard (full) | $33.3_{2.5}$ | $22.0_{2.6}$ | $84.0_{2.0}$ | $44.7_{3.1}$ | $98.3_{0.6}$ | $98.3_{1.2}$ |
| DiffuGuard (gen) | $69.3_{1.5}$ | $54.3_{1.5}$ | $81.3_{1.5}$ | $55.7_{4.0}$ | $98.3_{0.6}$ | $98.0_{1.0}$ |
| SADDER (top-1) | $78.0_{2.6}$ | $71.0_{1.7}$ | $80.0_{2.6}$ | $58.3_{1.5}$ | $97.7_{0.6}$ | $97.3_{0.6}$ |
| SADDER (aligned) | $85.3_{0.6}$ | $77.7_{3.1}$ | $82.7_{0.6}$ | $55.7_{1.5}$ | $97.7_{0.6}$ | $97.7_{0.6}$ |
| Random selection | $79.3_{1.2}$ | $74.3_{1.5}$ | $82.0_{2.0}$ | $53.7_{4.0}$ | $97.0_{1.0}$ | $97.0_{1.0}$ |

Table 33: Generated-span DIJA score differences (percentage points), with 95% prompt-cluster bootstrap intervals. Comparisons use the same prompts and seeds; full SAD and random selection have no warm-up.

| Comparison | HB | LG | WG |
|---|---|---|---|
| **LLaDA** | | | |
| Selective vs Unmodified | 2.7 [−2.0, 7.3] | 0.7 [−6.7, 7.7] | 0.0 [−2.3, 2.0] |
| Selective vs Full SAD | 4.0 [−1.3, 9.3] | −2.3 [−8.7, 3.7] | −1.0 [−2.7, 0.3] |
| Selective vs Random selection | 3.3 [−1.3, 7.7] | 2.3 [−2.7, 7.3] | −1.0 [−2.7, 0.7] |
| Warm-up vs Unmodified | −1.7 [−7.7, 4.0] | −11.7 [−19.7, −4.3] | 1.0 [−1.0, 3.0] |
| Warm-up vs Full SAD | −0.3 [−6.0, 5.3] | −14.7 [−21.7, −8.0] | 0.0 [−1.3, 1.3] |
| Warm-up vs Random selection | −1.0 [−6.0, 3.7] | −10.0 [−16.3, −4.0] | 0.0 [−1.3, 1.3] |
| **Dream** | | | |
| Selective vs Unmodified | −3.7 [−10.0, 2.3] | 8.7 [1.7, 15.0] | 0.0 [−1.0, 1.0] |
| Selective vs Full SAD | −8.0 [−13.3, −2.7] | 15.0 [8.0, 22.0] | 1.0 [0.0, 2.3] |
| Selective vs Random selection | −2.0 [−7.3, 3.0] | 3.7 [−1.3, 9.0] | 0.3 [0.0, 1.0] |
| Warm-up vs Unmodified | −0.3 [−6.7, 5.7] | −9.0 [−16.0, −2.0] | −1.3 [−3.7, 0.3] |
| Warm-up vs Full SAD | −4.7 [−9.7, 0.3] | −2.7 [−10.3, 5.0] | −0.3 [−2.3, 1.3] |
| Warm-up vs Random selection | 1.3 [−3.3, 6.3] | −14.0 [−20.0, −8.0] | −1.0 [−2.3, 0.0] |

# I UNIFORM-STATE DIFFUSION CONTINUATION RESULTS

Table 34 reports the complete Duo and UDLM comparison under the protocol of Section 5.2. Unsafe rate uses the RTP harmful cohort; benign BERTScore uses the separate benign RTP cohort. Entries report means and standard deviations across five generation seeds.

Table 34: Uniform-state continuation on RTP.

| Method | Duo | | UDLM | | NFE |
|---|---|---|---|---|---|
| | Unsafe % ↓ | Benign BERT ↑ | Unsafe % ↓ | Benign BERT ↑ | |
| Unmodified | $51.3_{1.8}$ | $81.77_{0.08}$ | $68.5_{2.5}$ | $80.81_{0.04}$ | 128 |
| **SAD** | | | | | |
| $\eta = 3\times10^4$ | $43.4_{3.5}$ | $81.77_{0.14}$ | $57.1_{1.1}$ | $80.50_{0.06}$ | 128 |
| $\eta = 5\times10^4$ | $42.0_{1.4}$ | $81.50_{0.10}$ | $52.9_{1.6}$ | $80.40_{0.09}$ | 128 |
| $\eta = 10^5$ | $42.8_{2.2}$ | $81.04_{0.07}$ | $47.4_{2.9}$ | $80.49_{0.03}$ | 128 |
| $\eta = 1.5\times10^5$ | $41.1_{2.2}$ | $80.76_{0.22}$ | $49.5_{4.2}$ | $80.53_{0.16}$ | 128 |
| **SADDER,** $k = 6$ | | | | | |
| $\eta = 3\times10^4$ | $47.9_{2.3}$ | $81.95_{0.04}$ | $56.3_{1.7}$ | $80.67_{0.12}$ | 128 |
| $\eta = 5\times10^4$ | $44.6_{4.5}$ | $81.85_{0.07}$ | $52.2_{1.6}$ | $80.68_{0.04}$ | 128 |
| $\eta = 10^5$ | $44.2_{1.9}$ | $81.67_{0.14}$ | $46.1_{2.4}$ | $80.80_{0.07}$ | 128 |
| $\eta = 1.5\times10^5$ | $45.2_{3.6}$ | $81.54_{0.13}$ | $45.1_{2.9}$ | $80.89_{0.13}$ | 128 |
| **SADDER,** $k = 9$ | | | | | |
| $\eta = 3\times10^4$ | $47.1_{4.1}$ | $81.93_{0.07}$ | $55.5_{2.1}$ | $80.65_{0.13}$ | 128 |
| $\eta = 5\times10^4$ | $45.2_{2.3}$ | $81.81_{0.06}$ | $51.5_{3.3}$ | $80.67_{0.04}$ | 128 |
| $\eta = 10^5$ | $45.1_{2.7}$ | $81.59_{0.17}$ | $46.5_{2.3}$ | $80.81_{0.09}$ | 128 |
| $\eta = 1.5\times10^5$ | $44.1_{1.8}$ | $81.41_{0.15}$ | $44.5_{2.1}$ | $80.90_{0.14}$ | 128 |
| **SADDER,** $k = 12$ | | | | | |
| $\eta = 3\times10^4$ | $47.2_{3.2}$ | $82.03_{0.08}$ | $57.9_{2.1}$ | $80.66_{0.14}$ | 128 |
| $\eta = 5\times10^4$ | $43.0_{2.5}$ | $81.86_{0.10}$ | $51.9_{3.0}$ | $80.69_{0.02}$ | 128 |
| $\eta = 10^5$ | $45.0_{0.7}$ | $81.77_{0.16}$ | $48.0_{2.6}$ | $80.81_{0.10}$ | 128 |
| $\eta = 1.5\times10^5$ | $43.6_{2.5}$ | $81.68_{0.10}$ | $45.9_{1.7}$ | $80.92_{0.12}$ | 128 |
| **CSD** | | | | | |
| $k = 6$ | $48.3_{2.9}$ | $81.80_{0.11}$ | $53.5_{1.8}$ | $80.51_{0.06}$ | 128 |
| $k = 9$ | $35.1_{2.3}$ | $81.93_{0.06}$ | $63.5_{1.5}$ | $80.81_{0.06}$ | 128 |
| $k = 12$ | $50.6_{2.1}$ | $81.78_{0.08}$ | $67.1_{4.2}$ | $80.81_{0.04}$ | 128 |
| **Best-of-**$N$ | | | | | |
| $N = 3$ | $42.2_{2.3}$ | 82.05 | $70.0_{3.4}$ | 80.80 | 384 |
| $N = 5$ | $37.6_{2.1}$ | 82.10 | $68.9_{2.1}$ | 80.72 | 640 |
| $N = 8$ | $33.1_{2.3}$ | 82.16 | $69.7_{2.2}$ | 80.77 | 1024 |

# J Guidance analysis and ablations for uniform-state diffusion

## J.1 Where the gate opens

Figure 3 plots the open-gate ratio during unguided continuations of unsafe RTP prompts. The open-gate ratio generally increases as denoising proceeds, consistent with position-wise representations becoming more informative as the continuation takes shape. SADDER uses this evolving signal to select positions for guidance, whereas SAD applies its correction to all eligible positions at every step.

## J.2 Softening the gate

We replace the binary gate with $a_{k,l}^{(t)} = \max\left\{2\sigma\left(s_{k,l}^{(t)}/\tau\right) - 1,\, 0\right\}$, where $\sigma$ is the sigmoid function and $\tau$ controls the softness of the gate. The transformation makes the gate zero when $s_{k,l}^{(t)} = 0$, and clipping at zero leaves positions with negative scores unguided. For positive scores, the gate takes a value between zero and one; the strict gate is recovered as $\tau \to 0$.

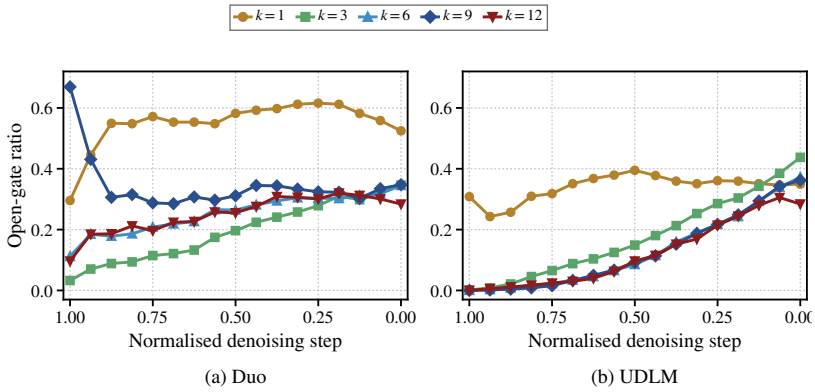


Figure 3: Open-gate ratio along unguided continuations of unsafe RTP prompts: the fraction of positions with $s_{k,l}^{(t)} > 0$. The horizontal axis runs from early to late denoising steps.

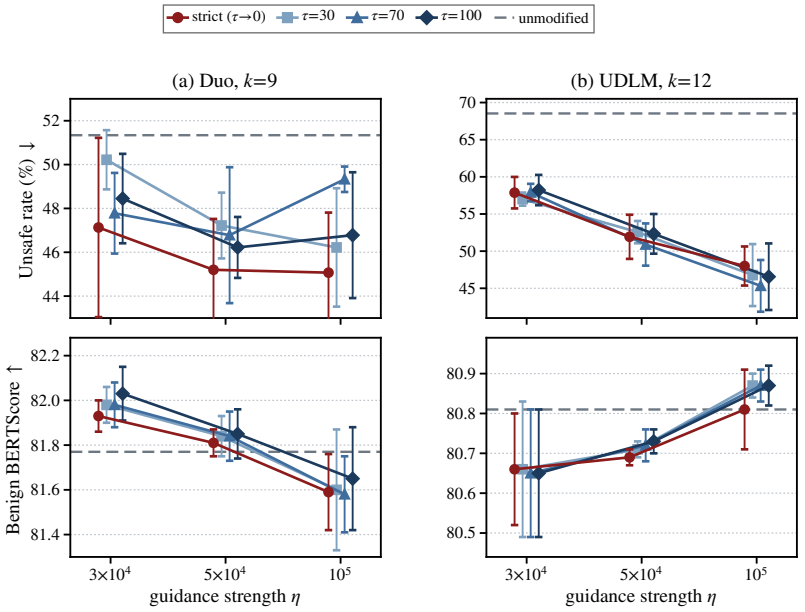


Figure 4: Unsafe rate and benign BERTScore across guidance strengths for the strict gate and softened gates with $\tau \in \{30, 70, 100\}$ on Duo ($k = 9$) and UDLM ($k = 12$). Error bars are one standard deviation across seeds: five seeds for the strict gate (the runs reported in Table 34) and three seeds for the softened gates.

Figure 4 shows no consistent effect of the temperature. On UDLM the four gates coincide in both metrics across the whole range. On Duo the strict gate trends toward lower unsafe rates and lower benign BERTScore, but the differences are smaller than the seed-to-seed variation of either gate. We therefore keep the strict indicator, which needs no additional hyperparameter.