DETECTING AND STOPPING NSFW GENERATION VIA DENOISING DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models can synthesize not-safe-for-work (NSFW) content. Existing detection methods primarily assess the content of prompts, completed images, or intermediate states. However, prompt filters are vulnerable to obfuscated intent, while post-generation classifiers incur the full generation cost before detection. In contrast, in-generation detectors inspect intermediate features but rely primarily on *static cues* from the semantic content of individual states, leaving activation dynamics across denoising steps underexplored. We therefore model *harmful dynamics*, *i.e.*, temporal patterns of activation motion associated with unsafe content, as a safety signal. Based on this signal, we formulate detection as a constrained sequential decision problem, aiming to stop unsafe generations early while balancing detection errors and sampling cost. Our approach follows an "observe, summarize, and stop" workflow: it **observes** activation motion during denoising, **summarizes** its history into risk scores, and **stops** generation based on accumulated risk. We conduct comprehensive experiments across four diffusion backbones and three benchmarks. Our approach significantly improves in-domain and cross-domain detection, with relative AUC gains of up to 14.7% over the strongest baseline, demonstrating the value of denoising dynamics as a safety signal. In addition, dynamics-based monitoring generalizes to unseen prompt distributions across different diffusion architectures and enables early intervention in unsafe generation with minimal overhead.

1 INTRODUCTION

Text-to-image (T2I) diffusion models (Rombach et al., 2022; Podell et al., 2024) can produce not-safe-for-work (NSFW) content, such as pornography (Qu et al., 2023; Leng et al., 2026). Existing detection methods operate at different stages of generation. Specifically, pre-generation text filters reject harmful prompts before sampling (Inan et al., 2023), whereas post-generation image classifiers screen the completed output (Schramowski et al., 2022). However, text filters can be bypassed readily by adversarial prompts (Yang et al., 2024a; Dang et al., 2025). Meanwhile, post-generation image classifiers detect unsafe content only after sampling is complete, resulting in computational inefficiency. These limitations have motivated **in-generation detection**, which monitors intermediate states during denoising to identify unsafe generations before completion (Yang et al., 2025; 2026).

Despite bringing detection into the generation process, these detectors primarily rely on *static cues*, *i.e.*, semantic information extracted from individual intermediate states. However, such cues can be obscured by noise, making reliable detection challenging (Yang et al., 2026). Meanwhile, the model forms the image through successive changes in its internal activations. We therefore posit that *how internal activations change as denoising proceeds* provides evidence of unsafe generation, shifting the focus from *state content* to *denoising dynamics*. Figure 1a illustrates how denoising dynamics differ between unsafe and benign generations. Specifically, the NSFW scores of unsafe generations increase substantially during denoising, while those of benign generations remain near zero. We characterize these unsafe-related patterns of activation motion as ***harmful dynamics***. As illustrated in Figure 1b, these dynamics persist under obfuscation: explicit and obfuscated unsafe prompts exhibit distinct distributions in text-embedding space but overlapping patterns in the learned space of activation motion. Figure 1c further shows that using the full denoising history improves detection over using individual steps, highlighting the value of motion history for unsafe detection.

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

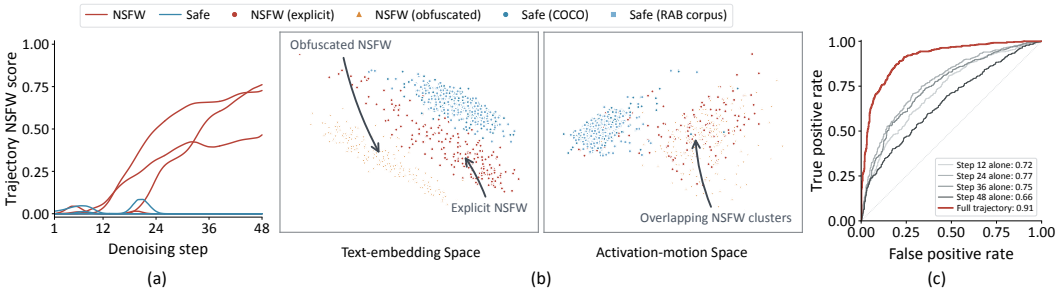


Figure 1: Motivating observations for dynamics-based NSFW detection. (a) A reference-based cumulative-sum (CUSUM) diagnostic of activation motion for unsafe and benign trajectories. (b) t-SNE of I2P, Ring-A-Bell, and benign samples in text-embedding space (left) and learned representation space of activation motion (right). (c) ROC curves for single-step and full-trajectory representations. See Appendix B.1 for details.

To use harmful dynamics for intervention, the detector must decide at each denoising step whether the observed motion history provides sufficient evidence to stop generation. Stopping after only a few steps may interrupt benign generations, whereas waiting until later steps incurs additional sampling cost. We formulate this trade-off as a constrained sequential decision problem and propose Dynamics-based Risk Monitoring (**DRM**) to address it. Specifically, our approach follows the *observe, summarize, and stop* workflow. First, we apply read-only probes to observe changes in the model's internal activations across denoising steps and encode them as motion observations. Thereafter, we use a temporal monitor to summarize the motion history and update a risk score at each denoising step. Finally, we use an abort gate to accumulate evidence from these risk scores and stop generation when the accumulated evidence reaches a preset threshold. This design is theoretically grounded in classical Bayesian quickest detection. Under its model assumptions, an optimal detector updates the probability of change with each new observation and stops when this probability reaches a threshold (Shiryaev, 1963).

We evaluate DRM across four diffusion backbones spanning U-Net (Rombach et al., 2022; Podell et al., 2024) and DiT (Chen et al., 2024) architectures, with I2P (Schramowski et al., 2023) for in-domain evaluation and Ring-A-Bell (Tsai et al., 2024) and T2I-RiskyPrompt (Zhang et al., 2026a) for cross-domain evaluation. DRM achieves higher AUC and F1 than the compared in-generation detectors across all settings, with relative AUC improvements over the strongest baseline of up to 14.7% in-domain and 10.6% cross-domain. Furthermore, on Ring-A-Bell with SD1.5, DRM reduces sampling cost by 7.6% and the miss rate by 77.6% compared with the baseline.

Our contributions are three-fold:

- We introduce a dynamics-based perspective on NSFW detection by characterizing *harmful dynamics*, *i.e.*, temporal patterns of activation motion associated with unsafe content.
- We formulate NSFW detection as a constrained sequential decision task and develop DRM, which predicts risk from motion history and accumulates it to decide when to stop generation.
- We conduct comprehensive experiments across four diffusion backbones and three benchmarks, demonstrating that DRM improves detection performance over existing baselines while enabling early termination of unsafe generations.

2 Related Work

NSFW detection and mitigation methods can be divided into three categories according to the stage at which they operate: pre-generation, post-generation, and in-generation. We first review representative work in these categories, then discuss studies of diffusion dynamics that motivate our approach.

Pre-generation text filters. Pre-generation approaches check the safety of input prompts before diffusion sampling begins. Early approaches use keyword blacklists, which can be bypassed through synonym substitution. With advances in language modeling, researchers have turned to pretrained models to filter unsafe prompts (Inan et al., 2023; Hanu & Unitary team, 2020). For instance, GuardT2I (Yang et al., 2024b) translates prompt embeddings into natural language, while La-

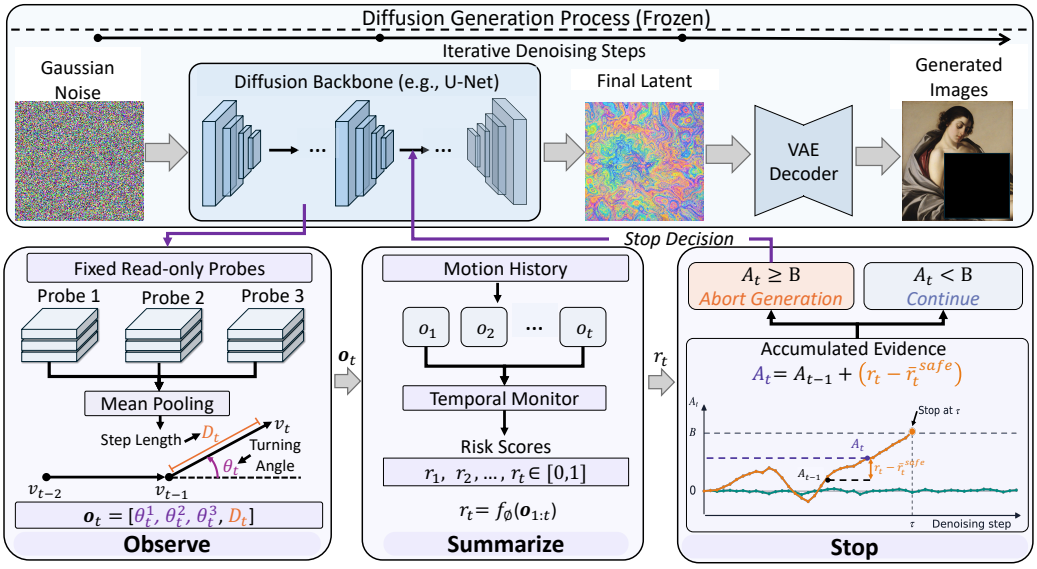


Figure 2: Overview of DRM. **Observe:** Read-only probes capture activation motion across denoising steps. **Summarize:** A temporal monitor summarizes the motion history into per-step risk scores. **Stop:** The abort gate accumulates these scores and stops sampling at the first threshold crossing.

tent Guard (Liu et al., 2024) detects harmful concepts in a contrastively learned embedding space. Despite these advances, these filters remain vulnerable to adversarial prompts (Yang et al., 2024a).

Post-generation image classifiers. Post-generation methods detect unsafe content in the final generated image. For instance, NudeNet (nud, 2019) identifies nudity through object detection, while Q16 (Schramowski et al., 2022) classifies harmful content in CLIP space (Radford et al., 2021). However, these classifiers work only after image is generated, which incurs the full generation cost.

In-generation defenses. In-generation defenses operate during denoising and can be divided into two groups: *steering* and *detection*. Steering methods suppress unsafe content by modifying the generation process (Ahn & Jung, 2025; Zhang et al., 2026b; Li et al., 2024). For instance, Schramowski et al. (2023) modify noise predictions through safety guidance, while SAFREE (Yoon et al., 2025) projects prompt embeddings away from a toxic subspace during denoising. However, these interventions can also alter benign content, creating a trade-off between safety and generation quality. In contrast, detection methods Huang et al. (2026) monitor the denoising process without modifying it (Liu et al., 2026). IGD (Yang et al., 2025) classifies the model's noise predictions. FlowGuard (Yang et al., 2026) uses a linear decoder to convert intermediate latents into images and classifies their content. These methods take a *content-based* view of denoising states, whereas our approach models *denoising dynamics*, *i.e.*, how content *changes* across denoising steps.

Generation dynamics. Studies of diffusion models show that semantic features, such as object categories and visual attributes, emerge at different stages of denoising, with some changes concentrated within short time intervals (Biroli et al., 2024; Li & Chen, 2024). Further analyses link these transitions in sampling dynamics to the formation of semantic structure (Ramachandran et al., 2025; Handke et al., 2026). These findings motivate us to exploit denoising dynamics as a safety signal.

3 From Harmful Dynamics to Sequential Early Detection

We present DRM, a framework that draws on classical quickest detection to identify unsafe generations and determine when to stop sampling. We first introduce the employed sequential decision framework, then detail the *observe, summarize, and stop* workflow illustrated in Figure 2.

3.1 Sequential Decision Framework

Diffusion sampling generates an image through successive denoising steps. Detecting NSFW content by monitoring these steps is therefore a sequential decision problem: at each step, the detector uses

the dynamics observed so far to decide whether to continue sampling or stop. Our goal is to detect unsafe generations with low sampling cost while limiting interruptions of benign generations.

Dynamics-based Risk Monitoring. Our DRM follows the *observe, summarize, and stop* workflow as illustrated in Figure 2. Specifically, we index the denoising steps by $t = 1, \ldots, T$. At each step, we use read-only probes to construct a motion observation $\mathbf{o}_t$ from changes in the diffusion model's intermediate activations. Thereafter, the temporal monitor f_ϕ summarizes the observation history $\mathbf{o}_{1:t}$ into an NSFW risk score r_t. The abort gate then accumulates these scores to decide whether to stop sampling. In particular, it maintains a detection statistic A_t and terminates sampling when A_t reaches a threshold B. We denote τ as the abort time, with $\tau = \infty$ if generation completes without an abort.

In addition to conventional detection metrics such as AUC and F1 value, we further assess DRM's stopping performance through false aborts (FA), missed unsafe generations (Miss), and sampling cost (Cost). Specifically, the false-abort rate on benign trajectories is:

$$\mathrm{FA}(\tau) = \mathbb{P}(\tau \le T \mid y = 0), \tag{1}$$

where $y \in \{0, 1\}$ denotes the label of the content obtained under uninterrupted sampling, with $y = 1$ indicating NSFW content and $y = 0$ benign content. The miss rate on unsafe trajectories is:

$$\mathrm{Miss}(\tau) = \mathbb{P}(\tau > T \mid y = 1). \tag{2}$$

The normalized sampling cost measures the fraction of steps used on unsafe trajectories:

$$\mathrm{Cost}(\tau) = \mathbb{E}\left[\frac{\min\{\tau, T\}}{T} \,\middle|\, y = 1\right]. \tag{3}$$

Undetected unsafe trajectories incur unit cost. Our objective is to reduce misses and sampling cost under a false-abort budget $\mathrm{FA}(\tau) \le \alpha$.

Theoretical basis. Our method draws on classical quickest detection (Shiryaev, 1963), which considers a sequence of observations $X_1, X_2, \ldots$ whose distribution changes at an unknown time ν. The detector should choose an alarm time σ by balancing detection delay against false alarms:

$$\min_{\sigma} \left\{\mathbb{E}\left[(\sigma - \nu)^+\right] + \lambda\, \mathbb{P}(\sigma < \nu)\right\}, \tag{4}$$

where $\lambda > 0$ penalizes alarms raised before the change. Under the classical Bayesian model with a geometric change-time prior and conditionally independent observations, the optimal rule raises an alarm when the posterior probability of change reaches a threshold (Shiryaev, 1963):

$$\sigma^*_\lambda = \inf\{t \ge 1 : p_t \ge a_\lambda\}, \tag{5}$$

where $p_t = \mathbb{P}(\nu \le t \mid X_{1:t})$ and a_λ is the stopping threshold.[1] This supports DRM's sequential decision structure: update a detection statistic from observations and stop when it reaches a threshold.

3.2 Observe, Summarize, and Stop

Observe: Measuring Activation Motion. In this work, we refer to temporal patterns of activation motion associated with unsafe generation outcomes as *harmful dynamics*. We characterize local activation motion through two complementary aspects: changes in direction and the distance between denoising steps. We measure them using turning angles and step lengths, respectively.

Specifically, we apply read-only probes to selected blocks of the frozen diffusion backbone. Under classifier-free guidance, we apply global mean pooling to the conditional-branch activation of block k at step t, and obtain the activation vector $\mathbf{v}_t^{(k)}$. The displacement between consecutive steps is:

$$\Delta\mathbf{v}_t^{(k)} = \mathbf{v}_t^{(k)} - \mathbf{v}_{t-1}^{(k)}. \tag{6}$$

For nonzero consecutive displacements, we compute the turning angle and step length as

$$\theta_t^{(k)} = \arccos\frac{\langle \Delta\mathbf{v}_t^{(k)}, \Delta\mathbf{v}_{t-1}^{(k)} \rangle}{\|\Delta\mathbf{v}_t^{(k)}\|\, \|\Delta\mathbf{v}_{t-1}^{(k)}\|}, \qquad D_t^{(k)} = \|\Delta\mathbf{v}_t^{(k)}\|_2. \tag{7}$$

[1] The full model assumptions and derivation are given in Appendix A.1.

Both motion signals are computed from activation displacements using inner products and norms. Therefore, they are invariant to a fixed translation or orthogonal transformation of the activation vectors. We compute both motion signals from existing activations, requiring no additional forward passes. These signals are then used to construct the observation $\mathbf{o}_t$.

Summarize: Estimating Risk from Motion History. The above observations describe local activation motion. However, a single observation can reflect either a brief fluctuation or part of a sustained pattern. Therefore, we summarize the motion history $\mathbf{o}_{1:t}$ to estimate NSFW risk at step t. Specifically, we use the temporal monitor f_ϕ to map the motion history to an NSFW risk score:

$$r_t = f_\phi(\mathbf{o}_{1:t}) \in [0, 1], \tag{8}$$

where ϕ denotes the learned parameters.

Stop: Accumulating Risk for Early Intervention. We accumulate the monitor's risk scores across denoising steps and stop generation when the accumulated statistic reaches a threshold. However, the monitor outputs continuous risk scores, so benign trajectories may receive small positive scores that accumulate and lead to false aborts. To mitigate this, we center the scores by subtracting the mean benign risk score at each step. Firstly, we estimate this mean from N_0 benign validation trajectories:

$$\bar{r}_t^{\text{safe}} = \frac{1}{N_0} \sum_{i=1}^{N_0} r_t^{(i)}, \tag{9}$$

where $r_t^{(i)}$ denotes the risk score of the i-th benign trajectory at step t. Starting from $A_0 = 0$, we accumulate the centered risk scores and stop generation when A_t reaches a fixed threshold B:

$$A_t = A_{t-1} + r_t - \bar{r}_t^{\text{safe}}, \qquad \tau = \inf\{t \in \{1, \dots, T\} : A_t \geq B\}. \tag{10}$$

Let $\mathcal{F}_t = \sigma(\mathbf{o}_1, \dots, \mathbf{o}_t)$ denote the information available at step t. Since A_t uses only these observations, τ is a stopping time with respect to $\mathcal{F}_t$.

Effect of centering. For a fixed monitor, consider uninterrupted trajectories and use the population mean $\bar{r}_s^{\text{safe}} = \mathbb{E}[r_s \mid y = 0]$. The accumulated statistic then satisfies

$$\mathbb{E}[A_t \mid y = 0] = 0, \qquad \mathbb{E}[A_t \mid y = 1] = \sum_{s=1}^{t} \delta_s, \tag{11}$$

where $\delta_s = \mathbb{E}[r_s \mid y = 1] - \mathbb{E}[r_s \mid y = 0]$ is the mean risk difference between unsafe and benign trajectories at step s.

Proof. Expanding the recursion in Equation (10) and applying linearity of expectation gives

$$\mathbb{E}[A_t \mid y] = \sum_{s=1}^{t} \big(\mathbb{E}[r_s \mid y] - \bar{r}_s^{\text{safe}}\big).$$

Substituting the population mean yields both identities in Equation (11). ∎

In conclusion, centering keeps benign accumulation at zero in expectation, while positive mean risk differences accumulate across steps. Section A.3 in Appendix gives the full derivation and the effect of estimating the mean from finitely many benign trajectories.

Threshold calibration. After training, we fix the monitor and estimate $\bar{r}_t^{\text{safe}}$ from benign validation trajectories. A benign trajectory is falsely aborted if its accumulated statistic reaches B at any step, equivalently $\max_{1 \leq t \leq T} A_t \geq B$. For stopping evaluation, we select B on a separate benign validation set, using the most permissive threshold whose empirical false-abort rate is at most α. The threshold remains fixed during test evaluation, where we measure FA, Miss, and Cost.

3.3 Training Objectives

We train the temporal monitor with the following objective:

$$\mathcal{L} = \mathcal{L}_{\text{final}} + \lambda_{\text{dense}}\mathcal{L}_{\text{dense}} + \lambda_{\text{early}}\mathcal{L}_{\text{early}} + \lambda_{\text{tc}}\mathcal{L}_{\text{tc}} + \lambda_{\text{safe}}\mathcal{L}_{\text{safe}}. \tag{12}$$

$\mathcal{L}_{\text{final}}$ and $\mathcal{L}_{\text{dense}}$ supervise the terminal and per-step risk scores through binary cross-entropy (BCE), respectively, with $\mathcal{L}_{\text{dense}}$ emphasizing later steps. $\mathcal{L}_{\text{early}}$ encourages high early risk scores on unsafe trajectories. $\mathcal{L}_{\text{tc}}$ penalizes decreases in high risk scores on unsafe trajectories (Li et al., 2025), while $\mathcal{L}_{\text{safe}}$ suppresses benign scores. Appendix B.2 provides the detailed definitions of every term.

Table 1: In-domain detection performance on I2P across four diffusion backbones. We reported AUC and F1 as mean$_{\pm\text{std}}$ deviation. Bold indicates the best in-generation result in each column.

	SD 1.4		SD 1.5		SDXL		PixArt-α	
Method	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Post-generation								
Falconsai	$0.687_{\pm 0.000}$	$0.617_{\pm 0.000}$	$0.696_{\pm 0.000}$	$0.613_{\pm 0.000}$	$0.782_{\pm 0.000}$	$0.719_{\pm 0.000}$	$0.657_{\pm 0.000}$	$0.565_{\pm 0.000}$
SD-SafetyChecker	$0.700_{\pm 0.000}$	$0.653_{\pm 0.000}$	$0.628_{\pm 0.000}$	$0.563_{\pm 0.000}$	$0.821_{\pm 0.000}$	$0.747_{\pm 0.000}$	$0.817_{\pm 0.000}$	$0.724_{\pm 0.000}$
Pre-generation								
Detoxify	$0.659_{\pm 0.000}$	$0.603_{\pm 0.000}$	$0.659_{\pm 0.000}$	$0.603_{\pm 0.000}$	$0.659_{\pm 0.000}$	$0.603_{\pm 0.000}$	$0.706_{\pm 0.000}$	$0.664_{\pm 0.000}$
Keyword-filter	$0.500_{\pm 0.000}$	$0.010_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.010_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.010_{\pm 0.000}$	$0.505_{\pm 0.000}$	$0.020_{\pm 0.000}$
In-generation								
IGD-MLP	$0.779_{\pm 0.001}$	$0.674_{\pm 0.005}$	$0.762_{\pm 0.002}$	$0.659_{\pm 0.006}$	$0.623_{\pm 0.002}$	$0.519_{\pm 0.002}$	$0.745_{\pm 0.004}$	$0.681_{\pm 0.008}$
IGD-LR	$0.741_{\pm 0.000}$	$0.670_{\pm 0.000}$	$0.735_{\pm 0.000}$	$0.663_{\pm 0.000}$	$0.625_{\pm 0.000}$	$0.560_{\pm 0.000}$	$0.717_{\pm 0.000}$	$0.650_{\pm 0.000}$
IGD-PCA	$0.525_{\pm 0.020}$	$0.000_{\pm 0.000}$	$0.527_{\pm 0.024}$	$0.006_{\pm 0.009}$	$0.598_{\pm 0.019}$	$0.384_{\pm 0.277}$	$0.654_{\pm 0.007}$	$0.214_{\pm 0.303}$
FlowGuard_Single	$0.834_{\pm 0.002}$	$0.720_{\pm 0.008}$	$0.803_{\pm 0.000}$	$0.701_{\pm 0.001}$	$0.774_{\pm 0.001}$	$0.671_{\pm 0.002}$	$0.666_{\pm 0.006}$	$0.634_{\pm 0.006}$
FlowGuard	$0.877_{\pm 0.003}$	$0.784_{\pm 0.013}$	$0.865_{\pm 0.002}$	$0.784_{\pm 0.014}$	$0.757_{\pm 0.013}$	$0.719_{\pm 0.010}$	$0.707_{\pm 0.017}$	$0.667_{\pm 0.012}$
DRM (ours)	$\mathbf{0.907}_{\pm 0.003}$	$\mathbf{0.819}_{\pm 0.007}$	$\mathbf{0.902}_{\pm 0.001}$	$\mathbf{0.791}_{\pm 0.006}$	$\mathbf{0.888}_{\pm 0.002}$	$\mathbf{0.816}_{\pm 0.013}$	$\mathbf{0.847}_{\pm 0.006}$	$\mathbf{0.768}_{\pm 0.004}$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Diffusion Backbones. We used four backbones to evaluate our appraoch: Stable Diffusion 1.4 and 1.5 (Rombach et al., 2022), SDXL (Podell et al., 2024), and PixArt-α (Chen et al., 2024). We kept the backbones frozen and used 50 DDIM (Song et al., 2021) sampling steps.

Datasets. We used I2P (Schramowski et al., 2023) and COCO captions (Lin et al., 2014) for unsafe and benign examples in training and in-domain evaluation. We applied Ring-A-Bell (Tsai et al., 2024) and T2I-RiskyPrompt (Zhang et al., 2026a) for cross-domain detection. Candidate unsafe samples are first screened using NudeNet (nud, 2019). Multiple human reviewers then cross-checked the final generated images, excluding candidates whose images contained **no** NSFW content.

Metrics. We used two evaluation protocols. *i) Detection protocol*: We reported AUC and F1 to evaluate detection performance. *ii) Stopping protocol*: We reported FA, Miss, and Cost (Section 3.1) to assess the trade-off between detection errors and sampling cost at preset false-abort budgets.

Implementation Details. Probe locations are selected using an independent validation set and fixed before monitor training. Each observation $\mathbf{o}_t$ comprised three turning angles and one step length, with a self-attention monitor summarizing the history. We reported means and standard deviations over three seeds. More detailes are in Appendix B.3.

4.2 MAIN RESULTS

Baselines. We compared DRM with three categories of detection methods. For pre-generation filtering, we employed Detoxify (Hanu & Unitary team, 2020) and a keyword filter (Yang et al., 2024b). Post-generation detectors include Falconsai (Falconsai, 2023) and SD-SafetyChecker (CompVis, 2022). Furthermore, we utilized IGD (Yang et al., 2025) and FlowGuard (Yang et al., 2026) as in-generation baselines. IGD includes MLP, LR, and PCA variants. For FlowGuard, we evaluated the multi-step version and its single-step variant, FlowGuard_Single.

In-domain Detection. In Table 1, we reported the in-domain results under the *detection protocol*. It can be observed that our approach achieves the highest AUC and F1 among the compared methods. For example, on SDXL and PixArt-α, DRM achieves AUCs of 0.888 and 0.847, exceeding the strongest in-generation baselines by 14.7% and 13.7%, respectively. Its F1 scores on these two backbones reach 0.816 and 0.768, with corresponding relative improvements of 13.5% and 12.8%.

Cross-domain Detection. Cross-domain detection is challenging because the detector must generalize to distributions that differ from those used in training. We reported the AUC and F1 performance on Ring-A-Bell and T2I-RiskyPrompt in Table 2. We can observe that our DRM outperforms all the in-generation baselines across all backbones and datasets. For instance, on Ring-A-Bell, DRM achieves AUC of 0.943 on PixArt-α, exceeding FlowGuard by 33%.

Table 2: Cross-domain detection performance across four diffusion backbones. We reported AUC and F1 as mean$_{\pm\text{std}}$ deviation. Bold indicates the best in-generation result in each column.

	SD 1.4		SD 1.5		SDXL		PixArt-α	
Method	AUC	F1	AUC	F1	AUC	F1	AUC	F1
(a) Ring-A-Bell								
Post-generation								
Falconsai	$0.755_{\pm 0.000}$	$0.517_{\pm 0.000}$	$0.724_{\pm 0.000}$	$0.669_{\pm 0.000}$	$0.603_{\pm 0.000}$	$0.582_{\pm 0.000}$	$0.531_{\pm 0.000}$	$0.501_{\pm 0.000}$
Pre-generation								
Keyword-filter	$0.879_{\pm 0.000}$	$0.862_{\pm 0.000}$	$0.879_{\pm 0.000}$	$0.862_{\pm 0.000}$	$0.879_{\pm 0.000}$	$0.862_{\pm 0.000}$	$0.879_{\pm 0.000}$	$0.862_{\pm 0.000}$
In-generation								
IGD-MLP	$0.895_{\pm 0.001}$	$0.820_{\pm 0.002}$	$0.867_{\pm 0.002}$	$0.791_{\pm 0.003}$	$0.730_{\pm 0.003}$	$0.668_{\pm 0.002}$	$0.660_{\pm 0.001}$	$0.619_{\pm 0.008}$
IGD-LR	$0.874_{\pm 0.000}$	$0.797_{\pm 0.000}$	$0.848_{\pm 0.000}$	$0.773_{\pm 0.000}$	$0.746_{\pm 0.000}$	$0.697_{\pm 0.000}$	$0.628_{\pm 0.000}$	$0.590_{\pm 0.000}$
IGD-PCA	$0.776_{\pm 0.028}$	$0.714_{\pm 0.081}$	$0.778_{\pm 0.056}$	$0.685_{\pm 0.062}$	$0.836_{\pm 0.140}$	$0.724_{\pm 0.180}$	$0.931_{\pm 0.027}$	$0.848_{\pm 0.024}$
FlowGuard_Single	$0.917_{\pm 0.003}$	$0.842_{\pm 0.005}$	$0.915_{\pm 0.002}$	$0.828_{\pm 0.004}$	$0.804_{\pm 0.002}$	$0.698_{\pm 0.007}$	$0.867_{\pm 0.005}$	$0.798_{\pm 0.008}$
FlowGuard	$0.965_{\pm 0.005}$	$0.854_{\pm 0.017}$	$0.977_{\pm 0.001}$	$0.896_{\pm 0.015}$	$0.878_{\pm 0.008}$	$0.805_{\pm 0.008}$	$0.708_{\pm 0.035}$	$0.692_{\pm 0.012}$
DRM (ours)	$\mathbf{0.994}_{\pm 0.001}$	$\mathbf{0.938}_{\pm 0.004}$	$\mathbf{0.992}_{\pm 0.001}$	$\mathbf{0.968}_{\pm 0.003}$	$\mathbf{0.889}_{\pm 0.003}$	$\mathbf{0.807}_{\pm 0.002}$	$\mathbf{0.943}_{\pm 0.002}$	$\mathbf{0.874}_{\pm 0.006}$
(b) T2I-RiskyPrompt								
Post-generation								
Falconsai	$0.578_{\pm 0.000}$	$0.517_{\pm 0.000}$	$0.584_{\pm 0.000}$	$0.562_{\pm 0.000}$	$0.525_{\pm 0.000}$	$0.506_{\pm 0.000}$	$0.601_{\pm 0.000}$	$0.503_{\pm 0.000}$
Pre-generation								
Keyword-filter	$0.565_{\pm 0.000}$	$0.233_{\pm 0.000}$	$0.565_{\pm 0.000}$	$0.233_{\pm 0.000}$	$0.565_{\pm 0.000}$	$0.233_{\pm 0.000}$	$0.565_{\pm 0.000}$	$0.233_{\pm 0.000}$
In-generation								
IGD-MLP	$0.823_{\pm 0.002}$	$0.741_{\pm 0.001}$	$0.821_{\pm 0.001}$	$0.742_{\pm 0.003}$	$0.633_{\pm 0.000}$	$0.565_{\pm 0.003}$	$0.780_{\pm 0.001}$	$0.731_{\pm 0.003}$
IGD-LR	$0.832_{\pm 0.000}$	$0.763_{\pm 0.000}$	$0.831_{\pm 0.000}$	$0.753_{\pm 0.000}$	$0.628_{\pm 0.000}$	$0.586_{\pm 0.000}$	$0.766_{\pm 0.000}$	$0.707_{\pm 0.000}$
IGD-PCA	$0.795_{\pm 0.025}$	$0.713_{\pm 0.066}$	$0.789_{\pm 0.055}$	$0.680_{\pm 0.074}$	$0.826_{\pm 0.121}$	$0.699_{\pm 0.166}$	$0.934_{\pm 0.024}$	$0.855_{\pm 0.012}$
FlowGuard_Single	$0.880_{\pm 0.001}$	$0.800_{\pm 0.003}$	$0.884_{\pm 0.001}$	$0.816_{\pm 0.001}$	$0.740_{\pm 0.001}$	$0.643_{\pm 0.001}$	$0.742_{\pm 0.004}$	$0.685_{\pm 0.001}$
FlowGuard	$0.872_{\pm 0.005}$	$0.787_{\pm 0.009}$	$0.874_{\pm 0.005}$	$0.796_{\pm 0.008}$	$0.783_{\pm 0.004}$	$0.719_{\pm 0.005}$	$0.758_{\pm 0.019}$	$0.708_{\pm 0.020}$
DRM (ours)	$\mathbf{0.973}_{\pm 0.001}$	$\mathbf{0.904}_{\pm 0.002}$	$\mathbf{0.974}_{\pm 0.001}$	$\mathbf{0.937}_{\pm 0.004}$	$\mathbf{0.843}_{\pm 0.005}$	$\mathbf{0.752}_{\pm 0.007}$	$\mathbf{0.945}_{\pm 0.008}$	$\mathbf{0.868}_{\pm 0.012}$

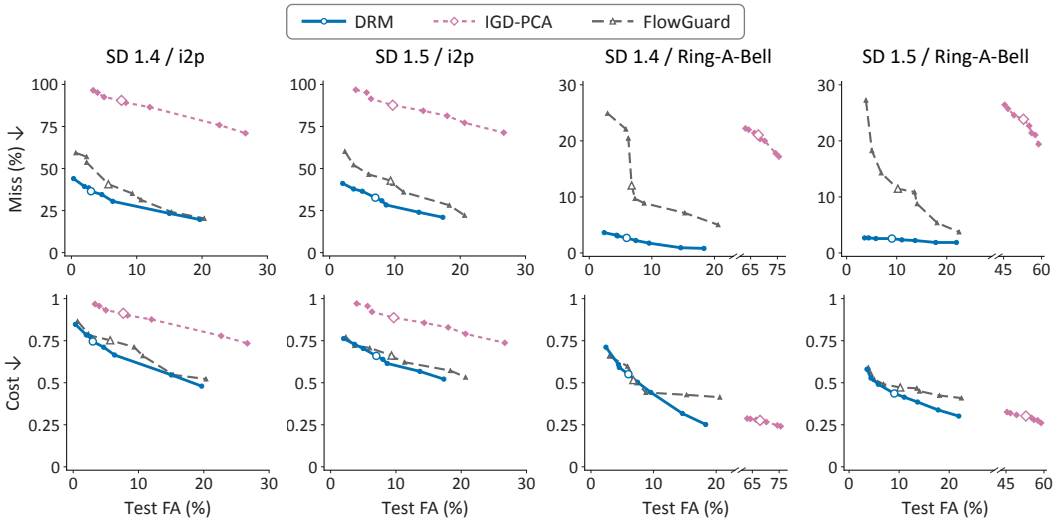


Figure 3: FA, Miss, and Cost performance. For each backbone/benchmark setting, we varied the FA calibration budget and plotted Miss and Cost. Curves closer to the lower-left corner indicate better performance. DRM achieves lower Miss than FlowGuard at comparable FA across all four settings.

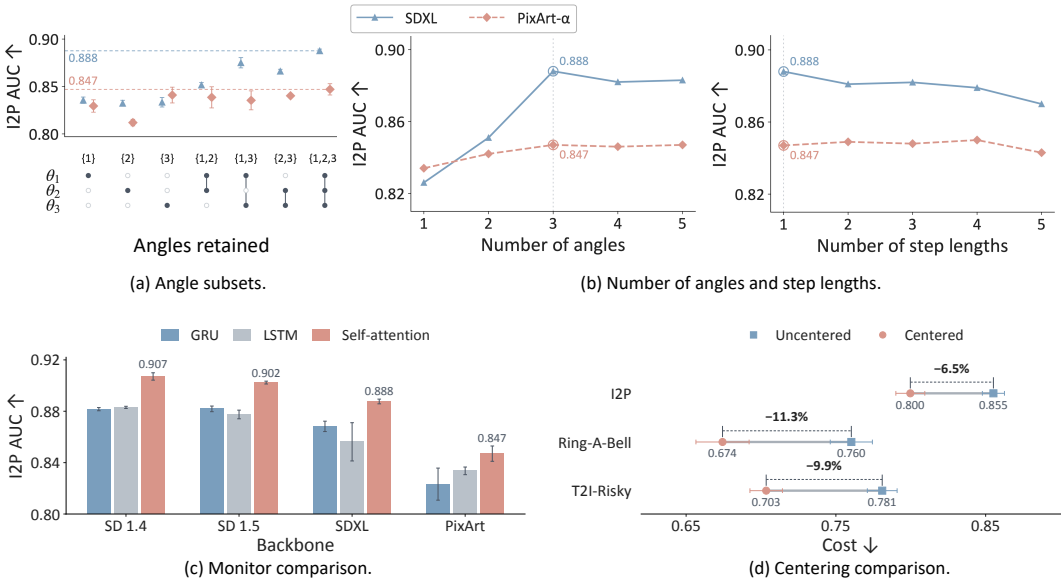
Early Intervention. In Figure 3, we compared DRM with IGD-PCA and FlowGuard under the stopping protocol by varying the FA calibration budget. DRM achieves lower Miss than FlowGuard at comparable test FA across both backbones and benchmarks. Compared with IGD-PCA, it also achieves lower Miss on I2P and lower FA and Miss on Ring-A-Bell. For example, at a 5% calibration budget on Ring-A-Bell with SD 1.5, DRM reduces Miss by 77.6% and Cost by 7.6% relative to FlowGuard, while lowering test FA from 10.23% to 9.03%. These results demonstrate effective early intervention across in-domain and cross-domain settings.

Table 3: Ablation studies on motion signals. Row-wise best results are bold.

Backbone	Turning angles	Step length	Full
AUC			
SD 1.4	$0.869_{\pm 0.002}$	$0.866_{\pm 0.001}$	$\mathbf{0.907}_{\pm 0.003}$
SD 1.5	$0.882_{\pm 0.002}$	$0.864_{\pm 0.002}$	$\mathbf{0.902}_{\pm 0.001}$
SDXL	$0.800_{\pm 0.005}$	$0.793_{\pm 0.006}$	$\mathbf{0.888}_{\pm 0.002}$
PixArt-α	$0.750_{\pm 0.010}$	$0.728_{\pm 0.002}$	$\mathbf{0.847}_{\pm 0.006}$
F1			
SD 1.4	$0.781_{\pm 0.007}$	$0.755_{\pm 0.009}$	$\mathbf{0.819}_{\pm 0.007}$
SD 1.5	$0.783_{\pm 0.007}$	$0.743_{\pm 0.006}$	$\mathbf{0.791}_{\pm 0.006}$
SDXL	$0.809_{\pm 0.007}$	$0.722_{\pm 0.028}$	$\mathbf{0.816}_{\pm 0.013}$
PixArt-α	$0.765_{\pm 0.007}$	$0.640_{\pm 0.011}$	$\mathbf{0.768}_{\pm 0.004}$

Table 4: Ablation studies on motion history. Row-wise best results are bold.

Backbone	Single-step MLP	Prefix-mean MLP	Full
AUC			
SD 1.4	$0.852_{\pm 0.004}$	$0.856_{\pm 0.000}$	$\mathbf{0.907}_{\pm 0.003}$
SD 1.5	$0.853_{\pm 0.005}$	$0.867_{\pm 0.001}$	$\mathbf{0.902}_{\pm 0.001}$
SDXL	$0.668_{\pm 0.001}$	$0.753_{\pm 0.003}$	$\mathbf{0.888}_{\pm 0.002}$
PixArt-α	$0.692_{\pm 0.005}$	$0.793_{\pm 0.000}$	$\mathbf{0.847}_{\pm 0.006}$
F1			
SD 1.4	$0.757_{\pm 0.015}$	$0.745_{\pm 0.004}$	$\mathbf{0.819}_{\pm 0.007}$
SD 1.5	$0.768_{\pm 0.007}$	$0.779_{\pm 0.000}$	$\mathbf{0.791}_{\pm 0.006}$
SDXL	$0.630_{\pm 0.006}$	$0.705_{\pm 0.002}$	$\mathbf{0.816}_{\pm 0.013}$
PixArt-α	$0.606_{\pm 0.015}$	$0.706_{\pm 0.019}$	$\mathbf{0.768}_{\pm 0.004}$



(a) Angle subsets. (b) Number of angles and step lengths. (c) Monitor comparison. (d) Centering comparison.

Figure 4: Ablation studies. (a) Angle subsets. (b) Motion signals. Left: Varying angles with one step length. Right: Varying step lengths with three angles. (c) Temporal monitor implementations. (d) The Cost of centering and uncentering.

4.3 Ablation Studies

Motion Signals. In Table 3, we evaluated the effects of the observation components, *i.e.*, turning angles and step length. As demonstrated in this table, combining both signals achieves the highest AUC and F1 across all four backbones. For instance, compared with using angles alone, the combination improves AUC from 0.800 to 0.888 on SDXL and from 0.750 to 0.847 on PixArt-α. These results indicate that turning angles and step length provide complementary information.

Motion History. As described in Section 3.2, DRM preserves the temporal order of observations when estimating risk scores. We further ablated this design using two MLP-based variants. Specifically, these variants compute risk scores from either the current observation or the average of all observations. As shown in Table 4, DRM achieves the highest performance among the variants.

Angle Subsets. As described in Section 4.1, DRM constructs each motion observation from three turning angles and one step-length signal. To examine the contributions of angles from different blocks, we retained the same step-length signal and evaluated each individual angle, each pair, and the full set. The comparison in Figure 4a shows that using all three angles achieves the highest I2P AUC of 0.888 on SDXL. Removing any one angle noticeably reduces detection performance. This indicates that the three angles provide complementary information for detecting harmful dynamics.

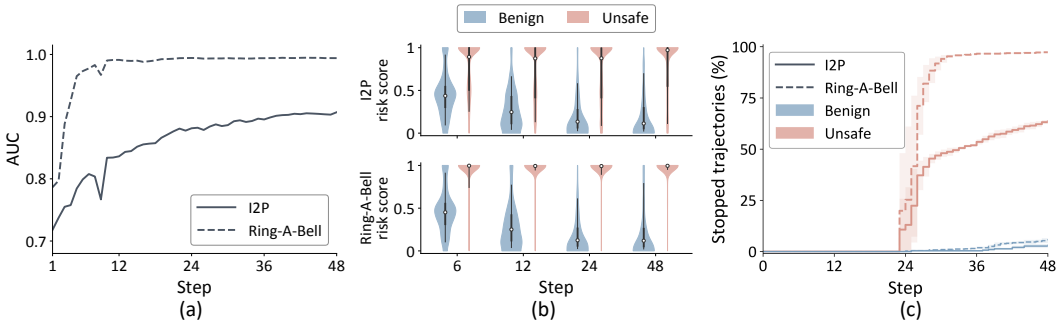


Figure 5: Analysis of detection and early intervention. (a) AUC during denoising. (b) Risk-score distribution at different steps. (c) Cumulative percentage of stopped trajectories. DRM distinguishes unsafe from benign trajectories and enables early intervention during denoising.

Number of Motion Signals. In Figure 4b, we varied the number of angles from one to five while retaining one step-length signal (left), and varied the number of step-length signals from one to five while retaining three angles (right). It can be observed that three turning angles and one step-length signal achieves the best balance between detection performance and observation size.

Monitor Implementations. In our implementation, we used self-attention blocks as the temporal monitor. In Figure 4c, we ablated this implementation by replacing self-attention with GRU and LSTM. We can observe that self-attention achieves the highest AUC among these implementations. For instance, on SDXL, it achieves an AUC of 0.888, significanly outperforms GRU and LSTM.

Centering and Stopping. In Equation (10), DRM subtracts the mean benign risk at each denoising step. We evaluated this operation by comparing this centering with direct accumulation. Specifically, we calibrated the thresholds separately on held-out benign data at a 5% false-abort budget. As shown in Figure 4d, centering reduces Cost across all three benchmarks.

4.4 Analysis on Detection during Denoising

In Figure 5, we analyzed DRM's detection performance, risk-score distributions, and stopping behavior during denoising. Specifically, in Figure 5a, we evaluated DRM on SD 1.4 using the observations available at each step. DRM reaches an AUC of 0.991 on Ring-A-Bell by step 12. To examine the resulting risk estimates, we compared the risk-score distributions of benign and NSFW trajectories in Figure 5b. On both datasets, NSFW scores concentrate near the top of the range, with median scores above 0.87 on I2P, whereas benign scores concentrate at lower values. Finally, in Figure 5c, we plotted the percentage of trajectories whose accumulated risk had reached the stopping threshold by each step. With thresholds calibrated on I2P at a 5% false-abort budget, DRM stops 96.6% of unsafe trajectories on Ring-A-Bell by step 36, while interrupting only 1.9% of benign trajectories. Together, these results demonstrate that DRM produces discriminative risk estimates during denoising and enables early intervention in unsafe generation.

Runtime overhead. On an NVIDIA A100 GPU at batch size one, the self-attention monitor takes approximately 43 ms to process all steps, which is an almost negligible time (<1%) overhead.

5 Conclusion

In this paper, we introduce harmful dynamics as a safety signal for in-generation NSFW detection. Building on this insight, we formulate NSFW detection as a sequential decision problem and develop DRM to connect motion-history risk estimates with stopping decisions that balance detection error and sampling cost. Experiments across four diffusion backbones and three benchmarks demonstrate SoTA detection performance in both in-domain and cross-domain settings, with up to 14.7% improvement in AUC over the strongest in-generation baseline. Moreover, under the stopping protocol, DRM shows that motion-based risk estimates enable intervention before generation is complete while allowing benign trajectories to continue. In the future, we aim to extend dynamics-based monitoring to broader unsafe-content categories and generative models.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the theoretical connection between DRM and classical quickest detection and to assist with mathematical derivations and proof writing. We also used these tools to support experimental planning, experiment management, and result analysis. Additionally, AI tools assisted with manuscript polishing, grammar checking, and figure preparation. The authors retain responsibility for the final manuscript, experimental results, and conclusions.

ETHICS STATEMENT

This work aims to aid the detection of NSFW content in text-to-image diffusion models and support safer image generation. By monitoring denoising dynamics, DRM enables intervention before unsafe generations are complete, helping reduce harmful outputs and unnecessary sampling costs.

REPRODUCIBILITY STATEMENT

We describe the backbones, benchmark splits, label conventions, baseline scope, and multi-seed reporting protocol in Section 4. Section B specifies the training objective, monitor architecture, normalization, sampling settings, feature indexing, and checkpoint selection.

REFERENCES

NudeNet: Neural nets for nudity detection and censoring. Github. https://github.com/notAI-tech/NudeNet, 2019.

Jaesin Ahn and Heechul Jung. Mitigating sexual content generation via embedding distortion in text-conditioned diffusion models. In *NeurIPS*, pp. 97–132, 2025.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, pp. 57611–57640, 2024.

CompVis. Stable diffusion safety checker. `https://huggingface.co/CompVis/stable-diffusion-safety-checker`, 2022.

Pucheng Dang, Xing Hu, Dong Li, Rui Zhang, Qi Guo, and Kaidi Xu. DiffZOO: A purely query-based black-box attack for red-teaming text-to-image generative model via zeroth order optimization. In *Findings of NAACL*, pp. 17–31, 2025.

Falconsai. NSFW image detection. `https://huggingface.co/Falconsai/nsfw_image_detection`, 2023.

Florian Handke, Dejan Stančević, Felix Koulischer, Thomas Demeester, and Luca Ambrogioni. The entropic signature of class speciation in diffusion models. In *ICML*, 2026.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

Yao Huang, Yitong Sun, Huanran Chen, Ruochen Zhang, Shouwei Ruan, Ranjie Duan, Maoxun Yuan, Yinpeng Dong, Hui Xue, Xiaochun Cao, and Xingxing Wei. UniNDM: A unified noise-driven detection and mitigation framework against sexual content in text-to-image generation. *TPAMI*, pp. 1–16, 2026.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Ye Leng, Junjie Chu, Mingjie Li, Chenhao Lin, Chao Shen, Michael Backes, Yun Shen, and Yang Zhang. When understanding becomes a risk: Authenticity and safety risks in the emerging image generation paradigm. In *CVPR*, pp. 39372–39382, 2026.

Marvin Li and Sitan Chen. Critical windows: Non-asymptotic theory for feature emergence in diffusion models. In *ICML*, pp. 27474–27498, 2024.

Xiaodan Li, Mengjie Wu, Yao Zhu, Yunna Lv, YueFeng Chen, Cen Chen, Jianmei Guo, and Hui Xue. Kelp: A streaming safeguard for large models via latent dynamics-guided risk detection. *arXiv preprint arXiv:2510.09694*, 2025.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. SafeGen: Mitigating sexually explicit content generation in text-to-image models. In *CCS*, pp. 4807–4821, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Mingrui Liu, Sixiao Zhang, and Cheng Long. Wukong: Framework for not safe for work detection in text-to-image systems. In *KDD*, pp. 891–902, 2026.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: A safety framework for text-to-image generation. In *ECCV*, pp. 93–109, 2024.

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

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *CCS*, pp. 3403–3417, 2023.

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

Sai Niranjan Ramachandran, Manish Krishan Lal, and Suvrit Sra. Cross-fluctuation phase transitions reveal sampling dynamics in diffusion models. In *NeurIPS*, pp. 103809–103868, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685, 2022.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *FAccT*, pp. 1350–1361, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, pp. 22522–22531, 2023.

Albert N Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46, 1963.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

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

Fan Yang, Yihao Huang, Jiayi Zhu, Ling Shi, Geguang Pu, Jin Song Dong, and Kailong Wang. Seeing it before it happens: In-generation NSFW detection for diffusion-based text-to-image models. *arXiv preprint arXiv:2508.03006*, 2025.

Jinghan Yang, Yihe Fan, Xudong Pan, and Min Yang. FlowGuard: Towards lightweight in-generation safety detection for diffusion models via linear latent decoding. *arXiv preprint arXiv:2604.07879*, 2026.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. MMA-Diffusion: MultiModal attack on diffusion models. In *CVPR*, pp. 7737–7746, 2024a.

Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. GuardT2I: Defending text-to-image models from adversarial prompts. In *NeurIPS*, pp. 76380–76403, 2024b.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *ICLR*, pp. 56439–56465, 2025.

Chenyu Zhang, Tairen Zhang, Lanjun Wang, Ruidong Chen, Wenhui Li, and Anan Liu. T2I-RiskyPrompt: A benchmark for safety evaluation, attack, and defense on text-to-image model. In *AAAI*, pp. 36039–36047, 2026a.

Chi Zhang, Changjia Zhu, Xiaowen Li, Yao Liu, and Zhuo Lu. Disciplined diffusion: Text-to-image diffusion model against NSFW generation. *arXiv preprint arXiv:2605.01113*, 2026b.

APPENDIX

A THEORETICAL ANALYSIS

This appendix develops the sequential decision framework in Section 3.1 and analyzes the abort gate in Section 3.2. We first derive the classical Bayesian threshold rule, then explain its role in DRM. The remaining analysis establishes the effect of centering and the relation between threshold calibration and the false-abort budget.

A.1 DERIVATION OF THE OPTIMAL THRESHOLD RULE

Model and assumptions. The Bayesian model underlying Equations (4) and (5) assumes a geometric change time, $\mathbb{P}(\nu = k) = \rho(1-\rho)^{k-1}$ for $k \geq 1$, with $0 < \rho < 1$. Conditional on ν, observations are independent, with known density f_0 before the change and f_1 from the change onward. Both densities are defined with respect to a common measure μ. We consider an infinite horizon with $0 < D_{\mathrm{KL}}(f_1 \| f_0) < \infty$. Let $\mathcal{H}_t = \sigma(X_1, \ldots, X_t)$ and $p_t = \mathbb{P}(\nu \leq t \mid \mathcal{H}_t)$, with $p_0 = 0$.

Theorem 1 (Classical Bayesian threshold rule (Shiryaev, 1963)) *Under the model above, for every $\lambda > 0$, there is a boundary $a_\lambda \in (0,1)$ such that $\sigma_\lambda^* = \inf\{t \geq 1 : p_t \geq a_\lambda\}$ minimizes $\mathbb{E}[(\sigma - \nu)^+] + \lambda \mathbb{P}(\sigma < \nu)$ over observation-based stopping times $\sigma \geq 1$.*

Proof. We express the risk in terms of p_t, derive the Bellman equation, and characterize its stopping region. We then verify that the first crossing attains the minimum.

1. Expressing the risk through the posterior. For a false-alarm penalty $\lambda > 0$, the risk minimized in Equation (4) is

$$J_\lambda(\sigma) = \mathbb{E}\big[(\sigma - \nu)^+\big] + \lambda \mathbb{P}(\sigma < \nu), \tag{13}$$

where $(z)^+ = \max\{z, 0\}$ and σ is a stopping time for $\mathcal{H}_t$. The identity $(\sigma - \nu)^+ = \sum_{t \geq 0} \mathbf{1}\{t < \sigma, \nu \leq t\}$ expresses delay as the number of steps spent continuing after the change. Since $\{t < \sigma\}$ is $\mathcal{H}_t$-measurable, conditioning each term and applying Tonelli's theorem gives the expected delay as $\mathbb{E}[\sum_{t=0}^{\sigma-1} p_t]$. Conditioning on each event $\{\sigma = t\}$ also gives $\mathbb{P}(\sigma < \nu) = \mathbb{E}[1 - p_\sigma]$. Thus, for rules with finite risk,

$$J_\lambda(\sigma) = \mathbb{E}\left[\sum_{t=0}^{\sigma-1} p_t + \lambda(1 - p_\sigma)\right]. \tag{14}$$

Each continuation step therefore incurs expected delay p_t, while stopping incurs cost $\lambda(1 - p_t)$.

2. Obtaining the Bellman equation. Given the current posterior p, the geometric prior gives a next-step change probability $q(p) = p + (1-p)\rho$. The next observation has density $g_p(x) = (1 - q(p))f_0(x) + q(p)f_1(x)$, and Bayes' rule updates the posterior as

$$p_{t+1} = \Phi(p_t, X_{t+1}), \qquad \Phi(p, x) = \frac{q(p) f_1(x)}{g_p(x)}. \tag{15}$$

Values of Φ where $g_p(x) = 0$ can be assigned arbitrarily, since these observations have zero conditional probability. This update makes p_t a sufficient, time-homogeneous Markov state for future costs.

Write $S(p) = \lambda(1 - p)$ for the stopping cost. At a generic state p, the change is already present with probability p. Otherwise, its remaining waiting time is geometric. Let $V_n(p)$ be the minimum remaining cost when stopping is required within n further observations, allowing an immediate stop. With $V_0 = S$, backward induction gives

$$V_{n+1}(p) = \min\left\{S(p),\ p + \int V_n(\Phi(p, x))\, g_p(x)\, d\mu(x)\right\}. \tag{16}$$

Allowing more observations can only reduce the minimum, so $0 \leq V_{n+1} \leq V_n \leq S \leq \lambda$. The limit $V = \lim_{n \to \infty} V_n$ equals the infinite-horizon optimum. To see this, truncate any finite-risk rule at n. Truncation reduces delay and increases the false-alarm penalty by at most $\lambda \mathbb{P}(\sigma > n)$, which tends

to zero. Finite risk ensures $\sigma < \infty$ almost surely because $\nu < \infty$ almost surely. Applying dominated convergence to Equation (16) now yields the Bellman equation:

$$V(p) = \min\left\{S(p),\ p + \int V(\Phi(p,x))\, g_p(x)\, d\mu(x)\right\}. \tag{17}$$

3. Showing why a threshold is optimal. Fix a rule that stops within n observations, holding its mapping from future observations to decisions fixed as p varies. Its remaining cost is a mixture of two conditional costs: one when the change has already occurred, weighted by p, and one when it has not, weighted by $1-p$. The geometric prior and conditional independence make these two conditional future laws independent of p. The rule's expected cost is therefore affine in p. Taking the infimum over such rules makes V_n concave and upper semicontinuous. Its decreasing limit $V = \inf_n V_n$ has both properties.

Immediate stopping gives $V(p) \le S(p)$, and $V(1) = S(1) = 0$. Thus, $S - V$ is a nonnegative, convex, lower semicontinuous function whose zero set is closed, convex, and contains 1. This zero set is exactly the stopping region:

$$\mathcal{S}_\lambda = \{p \in [0,1] : V(p) = S(p)\} = [a_\lambda, 1]. \tag{18}$$

The boundary is nondegenerate. At $p = 0$, observing once and then stopping costs $\lambda(1-\rho) < S(0)$, so $0 \notin \mathcal{S}_\lambda$ and closedness gives $a_\lambda > 0$. If $p \ge \lambda/(1+\lambda)$, then $S(p) \le p$, whereas continuing costs at least p. Hence $a_\lambda \le \lambda/(1+\lambda) < 1$. The Bellman-minimizing action is to continue below a_λ and stop at or above it.

4. Verifying that the first crossing attains the minimum. Let $\sigma^* = \inf\{t \ge 1 : p_t \ge a_\lambda\}$ and $\sigma_n = \min\{\sigma^*, n\}$. Repeated conditioning on the Bellman equality before σ^* gives, for every finite n,

$$V(0) = \mathbb{E}\left[\sum_{t=0}^{\sigma_n - 1} p_t + V(p_{\sigma_n})\right]. \tag{19}$$

Since $V \ge 0$, monotone convergence and the delay identity in Step 1 imply $\mathbb{E}[(\sigma^* - \nu)^+] \le V(0) \le \lambda$. Also, $\mathbb{E}[\nu] = 1/\rho$, so $\mathbb{E}[\sigma^*] \le 1/\rho + \lambda < \infty$. Consequently, σ^* is finite almost surely. The terminal term in Equation (19) is bounded by λ and converges to $V(p_{\sigma^*}) = S(p_{\sigma^*})$. Dominated convergence for this term and monotone convergence for the sum yield $V(0) = J_\lambda(\sigma^*)$. The first crossing therefore attains the minimum. ∎

Connection to the false-alarm constraint. For a false-alarm budget $0 < \alpha < 1$, the constrained problem is

$$\min_\sigma\ \mathbb{E}\big[(\sigma - \nu)^+\big] \quad \text{subject to} \quad \mathbb{P}(\sigma < \nu) \le \alpha. \tag{20}$$

If λ is chosen so that $\mathbb{P}(\sigma^* < \nu) = \alpha$, then for every feasible rule σ with $\mathbb{P}(\sigma < \nu) \le \alpha$, relaxed optimality gives

$$\mathbb{E}[(\sigma^* - \nu)^+] + \lambda\alpha \le J_\lambda(\sigma) \le \mathbb{E}[(\sigma - \nu)^+] + \lambda\alpha. \tag{21}$$

Subtracting $\lambda\alpha$ establishes optimality for Equation (20) under this budget-attainment condition.

A.2 From Quickest Detection to DRM

The classical formulation makes the cost of waiting explicit: each observation can improve the decision, but continuing after a change incurs delay. During diffusion sampling, additional motion observations can improve risk estimation, while each additional step consumes computation. DRM measures this cost over unsafe trajectories using Equation (3). The identity $\min\{\tau, T\} = \sum_{t=0}^{T-1} \mathbf{1}\{\tau > t\}$ gives

$$\mathrm{Cost}(\tau) = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{P}(\tau > t \mid y = 1). \tag{22}$$

Thus, Cost counts the steps for which an unsafe trajectory remains active. Together with Miss and FA, it expresses the finite-horizon trade-off between allowing unsafe generations to continue and interrupting benign ones.

This correspondence motivates the *observe, summarize, and stop* workflow. Read-only probes supply the observations $\mathbf{o}_t$ without changing the sampling process. The temporal monitor estimates risk as

$r_t = f_\phi(\mathbf{o}_{1:t})$, and the abort gate accumulates its excess relative to benign generations into A_t to determine when to stop. Centering allows below-benign scores to offset above-benign scores, so the gate responds to both the magnitude and duration of excess risk.

DRM follows the decision principle in Equation (5): update a statistic from observation history and stop at its first threshold crossing. Its statistic accumulates centered risk scores. The following analysis establishes the mean accumulation identities and the calibration rule for this gate.

A.3 Analysis of Centered Risk Accumulation

Setting. Fix a trained monitor, a backbone, and a trajectory distribution with both classes present. We consider the complete score sequence $r_1, \ldots, r_T$ under uninterrupted sampling. An actual abort reveals a prefix of this sequence. The scores lie in $[0, 1]$ and may be temporally dependent. Write $\mu_{ys} = \mathbb{E}[r_s \mid y]$ for the class-conditional mean at step s, and $\delta_s = \mu_{1s} - \mu_{0s}$.

For a fixed reference sequence $b_1, \ldots, b_T$, initialize $A_0 = 0$ and apply $A_t = A_{t-1} + r_t - b_t$. Expanding this recursion gives

$$A_t = \sum_{s=1}^{t} (r_s - b_s). \tag{23}$$

Taking expectations in Equation (23) gives

$$\mathbb{E}[A_t \mid y] = \sum_{s=1}^{t} \big(\mathbb{E}[r_s \mid y] - b_s\big) = \sum_{s=1}^{t} (\mu_{ys} - b_s). \tag{24}$$

Taking the population reference $b_s = \bar{r}_s^{\text{safe}} = \mu_{0s}$ gives

$$\mathbb{E}[A_t \mid y = 0] = 0, \qquad \mathbb{E}[A_t \mid y = 1] = \sum_{s=1}^{t} (\mu_{1s} - \mu_{0s}) = \sum_{s=1}^{t} \delta_s, \tag{25}$$

which proves Equation (11). Positive mean gaps contribute cumulatively. The argument applies to dependent scores as well as independent ones.

Effect of estimating the benign mean. With N_0 benign validation trajectories, the empirical reference is

$$\widehat{\bar{r}}_s^{\text{safe}} = \frac{1}{N_0} \sum_{i=1}^{N_0} r_s^{(i)}, \qquad e_s = \widehat{\bar{r}}_s^{\text{safe}} - \mu_{0s}, \tag{26}$$

where e_s denotes its deviation from the population mean. Let $\widehat{A}_t$ denote the accumulated statistic using the fitted means and A_t the statistic using the population means. On every trajectory, $\widehat{A}_t - A_t = -\sum_{s=1}^{t} e_s$. Let $\mathcal{R}$ denote the data used to fit these means, and evaluate a new trajectory independent of $\mathcal{R}$ from the same distribution. Conditioning on $\mathcal{R}$ fixes the errors e_s. Substituting $b_s = \mu_{0s} + e_s$ into Equation (24) then gives

$$\begin{cases} \mathbb{E}[\widehat{A}_t \mid y = 0, \mathcal{R}] = -\displaystyle\sum_{s=1}^{t} e_s, \\[2ex] \mathbb{E}[\widehat{A}_t \mid y = 1, \mathcal{R}] = \displaystyle\sum_{s=1}^{t} \delta_s - \sum_{s=1}^{t} e_s. \end{cases} \tag{27}$$

Thus, estimation error shifts both class means by the same cumulative amount, while their difference remains $\sum_{s=1}^{t} \delta_s$. Population centering removes the benign mean accumulation. Empirical centering introduces the explicit error term above. Both preserve the class-mean difference. The probability of crossing a threshold also depends on score variability and temporal dependence. We address that probability through trajectory-level calibration in Section A.4.

A.4 Threshold Calibration and the False-Abort Budget

From a stopping rule to a trajectory-level event. Fix the monitor, the fitted benign means, and the threshold B before evaluating a trajectory. Each A_t uses only $\mathbf{o}_{1:t}$, so

$$\{\tau \le t\} = \bigcup_{s=1}^{t} \{A_s \ge B\} \in \mathcal{F}_t. \tag{28}$$

The abort time is therefore a stopping time. A false abort occurs if a benign trajectory crosses B at any step, giving

$$\mathrm{FA}(\tau) = \mathbb{P}\left(\max_{1 \le t \le T} A_t \ge B \,\middle|\, y = 0 \right). \tag{29}$$

This identity explains why calibration uses the maximum along each complete trajectory. A threshold calibrated only on terminal scores would miss earlier crossings followed by a decrease.

Selecting the threshold. After fitting the benign means, we use a separate validation set of $n \ge 1$ benign trajectories. Let $M_i = \max_{1 \le t \le T} A_t^{(i)}$ and sort these maxima as $M_{[1]} \ge \cdots \ge M_{[n]}$. For a budget $0 \le \alpha < 1$, set $k = \lfloor \alpha n \rfloor$ and choose

$$B = M_{[k+1]} + \varepsilon, \tag{30}$$

where $\varepsilon > 0$ raises the cutoff to the next representable floating-point value. This strict increase handles ties under the gate's $A_t \ge B$ rule. At most k calibration maxima can exceed $M_{[k+1]}$, so

$$\widehat{\mathrm{FA}}_{\mathrm{cal}}(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{M_i \ge B\} \le \frac{\lfloor \alpha n \rfloor}{n} \le \alpha. \tag{31}$$

This is an empirical guarantee on the validation set. Test FA is measured on separate benign trajectories and can differ from α, particularly after a distribution shift.

How the threshold controls stopping. For the same score paths, $B_1 \le B_2$ implies $\tau_{B_1} \le \tau_{B_2}$, because every crossing of the higher boundary also crosses the lower one. Consequently,

$$\begin{cases} \mathrm{FA}(\tau_{B_1}) \ge \mathrm{FA}(\tau_{B_2}), \\ \mathrm{Miss}(\tau_{B_1}) \le \mathrm{Miss}(\tau_{B_2}), \\ \mathrm{Cost}(\tau_{B_1}) \le \mathrm{Cost}(\tau_{B_2}). \end{cases} \tag{32}$$

These relations follow from the nested stopping events and the monotonicity of $\min\{\tau, T\}$. They formalize the gate's trade-off: a lower threshold permits earlier detection at the expense of more false aborts. Centering determines the accumulated score paths. Calibration selects an operating point on those paths.

We calibrate B separately for each backbone, training seed, and budget, then keep it fixed throughout test evaluation. Source-calibrated cross-domain evaluation transfers both B and the fitted benign means unchanged.

B Implementation Details

B.1 Details of the Motivating Experiments

CUSUM diagnostic. The curves in Figure 1(a) are selected for illustration from a pool of 250 validation samples. This diagnostic does not involve detector training.

Representation visualization. Figure 1(b) includes 100 I2P samples, 100 Ring-A-Bell samples, and 100 benign samples from COCO and RAB corpus. The left-hand t-SNE uses prompt text embeddings, while the right-hand t-SNE uses trajectory representations extracted by the trained DRM.

Single-step ablation. Figure 1(c) compares single-step ablation variants of DRM with its full-trajectory version through ROC curves.

B.2 Training Objective

The objective in Equation (12) supervises both terminal and intermediate risk scores. Let $\langle\cdot\rangle$ denote the minibatch average, and let $\langle\cdot\rangle_+$ and $\langle\cdot\rangle_-$ denote averages over unsafe and benign trajectories, respectively. The per-step supervision is

$$\mathcal{L}_{\text{dense}} = \left\langle \sum_{t=1}^{T} w_t \,\mathrm{BCE}(r_t, y) \right\rangle, \qquad w_t = \frac{t-1}{\sum_{s=1}^{T}(s-1)}, \tag{33}$$

where w_t assigns more weight to later steps. The remaining terms are

$$\begin{aligned}
\mathcal{L}_{\text{final}} &= \big\langle \mathrm{BCE}(r_T, y) \big\rangle, \\
\mathcal{L}_{\text{early}} &= \big\langle \mathrm{BCE}(r_{t_e}, 1) \big\rangle_+, \\
\mathcal{L}_{\text{tc}} &= \left\langle \frac{1}{T-1} \sum_{t=2}^{T} \mathbf{1}\{r_{t-1} > \rho\} \, [r_{t-1} - r_t]_+ \right\rangle_+, \\
\mathcal{L}_{\text{safe}} &= \left\langle \frac{1}{T} \sum_{t=1}^{T} r_t \right\rangle_-,
\end{aligned} \tag{34}$$

where $[a]_+ = \max(a, 0)$, $t_e = \max(1, \lfloor 0.2T \rfloor) = 9$ for $T = 48$, and $\rho = 0.3$. We use $\lambda_{\text{dense}} = 0.5$, $\lambda_{\text{early}} = 0.3$, $\lambda_{\text{tc}} = 0.5$, and $\lambda_{\text{safe}} = 0.1$, with unit weight on $\mathcal{L}_{\text{final}}$. These weights are fixed on SD 1.4 and shared across backbones and training seeds.

B.3 Model and Training Configuration

Data and sampling. We use 50 DDIM steps with guidance scale 7.5 for the Stable Diffusion backbones and 4.5 for PixArt-α. Image resolution is 512^2 for SD 1.4/1.5 and 1024^2 for SDXL and PixArt-α. All learned detectors share an 80/10/10 I2P training, validation, and test split.

Motion observations. We form $\mathbf{o}_t$ from three turning angles and one step length. On SD 1.4/1.5, the angles use the last ResNet modules in `up_blocks[0]`, `mid_block`, and `up_blocks[1]`, and the step length uses `down_blocks[2]`. On SDXL, angles use the last ResNet modules in `up_blocks[1]`, `up_blocks[2]`, and `down_blocks[1]`, while step length uses `up_blocks[1]`. PixArt-α uses transformer blocks 14, 7, and 11 for angles and block 16 for step length. Block indices start at zero. Each signal is standardized by subtracting its training-set mean and dividing by its training-set standard deviation.

Probe selection. The SD 1.4 configuration follows individual-signal rankings. For the other backbones, we rank candidate signals by validation AUC using classifiers trained on trajectory summary features. The selection pool is an independent validation set containing 250 I2P prompts and 250 COCO captions. No test data are used to evaluate candidate signals or select probe blocks. We fix the selected blocks before monitor training. Monitor training and checkpoint selection use I2P, while Ring-A-Bell and T2I-RiskyPrompt are reserved for cross-domain evaluation.

Temporal monitor. The monitor uses two self-attention layers of width 96, four heads, and learned positional embeddings. Attention is masked so that each step uses only the observations available up

to that step. Feed-forward layers have width 192 and dropout 0.1. Layer normalization and an MLP with hidden width 32, GELU, dropout 0.1, and a sigmoid map each step's output to r_t.

Training. We train on cached trajectories for 30 epochs using AdamW with learning rate 5×10^{-4}, weight decay 0.05, cosine learning-rate decay, and gradient clipping at 1.0. Batch size is 64 for SD 1.4/1.5 and 32 for SDXL and PixArt-α. The checkpoint with the highest validation AUC of r_T is used for detection, stopping, and full-input ablations. Results report the mean and population standard deviation over training seeds 42, 43, and 44 on the same data split.

B.4 Evaluation Protocols

Baseline implementations. IGD-MLP and IGD-LR (Yang et al., 2025) classify channel-wise statistics of an early noise prediction. IGD-PCA applies PCA to the flattened prediction before MLP classification. FlowGuard_Single uses eight channel statistics of an early predicted clean latent with an MLP head selected by mean validation AUC across backbones. These single-step detectors use sampler step 6. The MLPs have hidden width 128, and PCA retains up to 256 components fitted on the training split. We train the MLPs for 30 epochs using AdamW with learning rate 3×10^{-4}, weight decay 0.1, and the same batch sizes as DRM, selecting checkpoints by validation AUC.

FlowGuard. FlowGuard (Yang et al., 2026) decodes predicted clean latents through a learned affine map, applies a circular Fourier low-pass filter, and classifies the images with ViT-B/16. The decoder uses 2,000 paired latents and VAE outputs from the I2P training split. Decoded images are 128×128, the filter radius is 0.2 of the Nyquist radius, and classifier inputs are resized to 224×224. We freeze the ViT embeddings and first five blocks. The remaining layers use a four-stage curriculum that shifts from later to earlier sampler steps, with BCE and a 0.01-weighted cross-step consistency loss. The checkpoint is selected by I2P-validation AUC at step 20.

Detection protocol. DRM uses r_T to compute AUC and F1. FlowGuard uses the maximum score at sampler steps {20,22,24,25,27,30}. Learned detectors use an F1 threshold of 0.5. For pretrained reference detectors, scores at or above the median I2P-validation score are positive. The keyword filter flags scores greater than zero. These thresholds remain fixed across benchmarks. For each NSFW benchmark, we first screen candidate unsafe samples using NudeNet. Multiple human reviewers then cross-check the final generated images under the evaluated sampling settings, excluding candidates whose images contain no NSFW content. Cross-domain classification uses the resulting screened benchmarks. Specifically, we reserve 3,000 trajectories for I2P, 500 for Ring-A-Bell, and 3200 for T2I-RiskyPrompt for evaluation. Benign trajectories for Ring-A-Bell and T2I-RiskyPrompt are generated from COCO captions that are strictly disjoint from the COCO captions used in the I2P setting. Reference detectors are evaluated once, with ±0.000 indicating zero training-seed variation.

Stopping protocol. We replay in-generation detectors on cached trajectories to measure FA, Miss, and Cost. Benign validation trajectories provide the per-step means for centering. We divide the original held-out benign calibration set into equal, prompt-disjoint calibration and evaluation sets using split seed 7. Thresholds are calibrated at each false-abort budget and frozen before evaluation. Miss and Cost use all 200 unsafe test trajectories per backbone. Cost measures the fraction of the 48 observation steps used on unsafe trajectories, with unit cost for missed trajectories.