D-BISR: MULTI-PROTOTYPE DIFFUSION RESIDUAL SCORING FOR TARGET-DATA-FREE HARMFUL PROMPT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models deployed in open-ended systems require input-side harmful prompt detection to block safety risks before generation. In practice, harmful prompts in target applications often differ from training data, and target-domain annotations are difficult to obtain. Detectors should therefore generalize to unseen domains without target-domain samples. While encoder-based models offer efficient input filtering, they suffer from local risk dilution, risk-pattern entanglement, and classification scores that fail to support stable low-FPR detection. To address these challenges, we propose D-BISR, an encoder-based framework for target-data-free harmful prompt detection. D-BISR uses token-level prototype attention to aggregate local harmful evidence and introduces multiple learnable risk prototypes to separate distinct risk patterns. Instead of relying on standard classification logits, it performs diffusion residual scoring to map risk evidence into a continuous, thresholdable safety metric. We evaluate D-BISR on ViSU, MMA, and SneakyPrompt under strict target-data-free settings. When trained solely on source-domain data and tested directly on unseen target domains, D-BISR achieves strong cross-dataset generalization and maintains stable low-FPR recall. On standard in-domain benchmarks, D-BISR outperforms existing encoder-based classifiers and anomaly detection baselines. Compared with larger safety guard models, D-BISR uses fewer parameters, less GPU memory, and achieves lower inference latency.

1 INTRODUCTION

As large language models are increasingly deployed in open-ended interactive systems, they are directly exposed to diverse user prompts, which may include malicious, disguised, or jailbreak-style inputs Greshake et al. (2023) that induce unsafe generation. Wei et al. (2023) Therefore, prompt-level safety detection has become a necessary front-line defense. However, in real deployment scenarios, harmful prompts in target domains often differ substantially from the source-domain data used to train the detector, resulting in a clear distribution shift. Meanwhile, collecting and annotating target-domain harmful prompts is costly and may be restricted by privacy, platform policies, or security concerns. Ganin et al. (2015) These constraints make it unrealistic to assume access to target-domain annotations or training samples. Therefore, a practical input-side safety detector should be trained only on available source-domain harmful-prompt data and directly generalize to unseen target domains without using any target-domain data.

Existing harmful prompt detection methods can be broadly divided into three categories. **LLM-based safety guards** (e.g., Llama Guard, WildGuard, and BingoGuard Inan et al. (2023); Han et al. (2024); Yin et al. (2025)) benefit from the strong semantic understanding but incur high inference cost. **Anomaly detection or geometric modeling perspectives** (e.g., Mahalanobis distance-based detectors and HyPE Lee et al. (2018); Maljkovic et al. (2026)) naturally produce continuous risk scores, but rely on fixed distributional or geometric assumptions that may fail to cover unseen risk patterns when no target-domain data are available for adaptation. **Encoder-based detectors** (e.g., DeBERTa and CLIP He et al. (2023); Radford et al. (2021); Reimers & Gurevych (2019)) are more efficient and can be trained solely on source-domain annotations, making them better suited to input-side deployment without target-domain data.

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.

Despite these advantages, existing encoder-based detectors are not fully aligned with the requirements of practical harmful prompt detection, particularly when directly transferred to unseen target domains with diverse risk patterns and strict low-FPR constraints. First, the shared binary objective usually compresses semantically diverse unsafe prompts into a single positive class, which may entangle distinct risk patterns. Second, global prompt representations can weaken local harmful evidence by mixing it with neutral context, resulting in risk dilution. These two issues are closely related: modeling multiple risk patterns only from global representations may still suffer from local dilution, while selecting local evidence without separating risk patterns may still mix heterogeneous risk cues. Finally, classifier logits and softmax probabilities are mainly optimized for binary discrimination and do not explicitly enforce score consistency across risk patterns or ranking stability under low-FPR detection.

Motivated by these observations, we propose D-BISR, an encoder-based supervised detector that jointly models heterogeneous risk patterns, extracts local harmful evidence, and produces stable continuous risk scores for target-data-free harmful prompt detection. D-BISR is designed around three complementary components. First, it extracts both token-level hidden states and a global prompt representation, and uses token-level prototype attention to explicitly aggregate local harmful evidence. Second, it introduces multiple learnable risk prototypes to capture different harmful patterns, reducing the entanglement caused by treating all unsafe prompts as a single homogeneous class. Third, it introduces a safety-aligned diffusion residual score, a residual-based risk scoring mechanism inspired by denoising diffusion rather than text generation. Ho et al. (2020) By encouraging safe prompts to produce small residuals and paired unsafe prompts to produce larger residuals, this score maps diverse risk evidence into a unified and thresholdable risk space for low-FPR detection.

We evaluate D-BISR on length-balanced ViSU, MMA, and SneakyPrompt Poppi et al. (2024); Yang et al. (2024a;b) in a source-supervised, target-data-free setting, where the detector is trained on a source harmful-prompt dataset and directly tested on unseen target datasets without target-domain training. We compare D-BISR with LLM-based guards, encoder-based classifiers, anomaly detectors, and geometric modeling methods. Experimental results show that D-BISR achieves strong low-FPR recall, competitive risk ranking, and stable cross-dataset performance, while maintaining lower parameter count, GPU memory usage, and inference latency than larger safety guards.

Our main contributions are summarized as follows:

- We formulate harmful prompt detection under source-to-unseen-target transfer as a representation and risk-scoring problem, highlighting the need to preserve localized harmful evidence, account for heterogeneous risk patterns, and maintain transferable decision scores without access to target-domain data.
- We propose D-BISR, a multi-prototype risk scoring framework that jointly models local harmful evidence and heterogeneous unsafe patterns through token-level prototype attention and learnable risk prototypes.
- We introduce a safety-aligned diffusion residual score and conduct experiments on multiple length-balanced benchmarks, demonstrating improved low-FPR recall, cross-dataset stability, and efficiency over representative baselines.

2 Related Work

2.1 LLM-based Safety Guards

LLM-based safety guards typically use instruction-tuned language models to determine whether user prompts or model responses violate predefined safety policies. Llama Guard Inan et al. (2023) formulates moderation as generation-based classification over a safety taxonomy; WildGuard Han et al. (2024) covers malicious intent, response harmfulness, and refusal detection; and BingoGuard Yin et al. (2025) further introduces fine-grained risk levels. These methods benefit from strong semantic understanding, but their large model size and per-prompt inference cost can limit high-throughput and low-latency deployment. In contrast, D-BISR adopts a lightweight encoder and performs front-end harmful prompt detection through local risk modeling and continuous risk scoring.

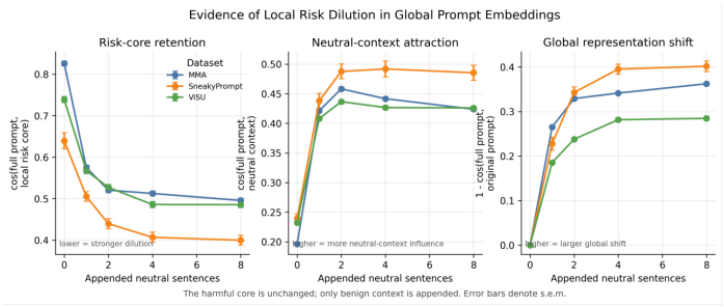


Figure 1: Local risk dilution as neutral context increases.

2.2 Anomaly and Geometry-based Detection

Another line of work detects harmful prompts through anomaly or geometric deviation. Mahalanobis distance is a representative feature-space out-of-distribution approach Lee et al. (2018), while HyPE Maljkovic et al. (2026) models harmless prompts in hyperbolic space and identifies harmful inputs according to their geometric deviation from the harmless region. These methods naturally produce continuous scores, but often rely on fixed distributional or distance assumptions, which may limit their ability to characterize unseen risk patterns under source-target shifts. D-BISR instead models heterogeneous risk patterns with multiple prototypes and constructs a unified risk scale through diffusion residual scoring, without relying on a single center or fixed geometric distance.

2.3 Encoder-based Supervised Detectors

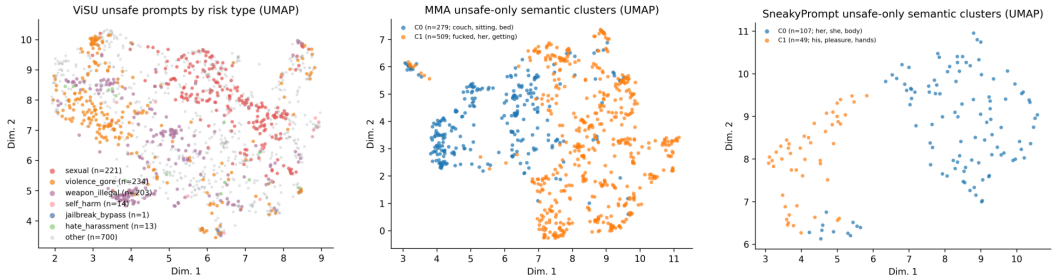
Encoder-based methods typically learn a supervised decision boundary from labeled safe and unsafe prompts using efficient text encoders such as DeBERTa He et al. (2023); vision-language encoders such as CLIP Radford et al. (2021) are also used in multimodal safety detection. LEG Islam & Surdeanu (2026) further incorporates word-level explanation supervision to improve interpretability. These methods are computationally efficient, but mainly focus on global prompt classification or explanation supervision. In contrast, D-BISR explicitly aggregates token-level harmful evidence, models heterogeneous risk patterns with multiple prototypes, and further uses residual-based continuous scoring for target-data-free detection under low-FPR constraints.

3 Motivating Analysis

Unsafe prompts are not a semantically compact and homogeneous category. They differ in risk intent, linguistic form, and evidence location. Some prompts are determined by a few local harmful keywords, while others express risk through implicit intent embedded in longer context. This section analyzes two coupled representation-level challenges, namely local risk dilution and risk-mode entanglement, which motivate the design of D-BISR.

3.1 Local Risk Dilution

Many unsafe prompts contain only sparse local harmful evidence, while the remaining content is largely neutral. When a detector relies on a global prompt representation, this local evidence is aggregated with neutral context and may become increasingly diluted as the neutral content grows. To examine this *local risk dilution*, we keep the local risk core of each unsafe prompt unchanged and progressively append neutral sentences. Using the CLIP text encoder, we compare the resulting global representation with the local risk core, appended neutral context, and original unsafe prompt. As shown in Fig. 1, increasing neutral context shifts the global representation away from the unchanged risk core and toward the neutral context, while also increasing its deviation from the original prompt, indicating that local harmful evidence can be weakened even when the harmful content itself remains unchanged.



(a) ViSU risk-type distribution (b) MMA semantic clusters (c) SneakyPrompt semantic clusters

Figure 2: Semantic heterogeneity of unsafe prompt representations.

This issue can be illustrated by normalized global aggregation. Let $\mathcal{R}$ denote local risk tokens and $\mathcal{B}$ denote neutral-context tokens. If the total normalized weight assigned to risk tokens is

$$A_{\mathcal{R}} = \frac{Z_{\mathcal{R}}}{Z_{\mathcal{R}} + Z_{\mathcal{B}}}, \tag{1}$$

where $Z_{\mathcal{R}} = \sum_{i \in \mathcal{R}} \exp(a_i)$ and $Z_{\mathcal{B}} = \sum_{j \in \mathcal{B}} \exp(a_j)$, then adding neutral tokens increases $Z_{\mathcal{B}}$ and decreases $A_{\mathcal{R}}$, when the risk tokens and their scores remain unchanged. This analysis does not treat all encoders as simple weighted averages, but highlights a structural risk of global aggregation: without explicit local-evidence selection, neutral context can reduce the relative contribution of harmful evidence. This motivates preserving token-level features and explicitly aggregating local harmful evidence.

3.2 Risk-Mode Entanglement

Unsafe prompts also exhibit heterogeneous risk modes, such as sexual content, violence, illegal instructions, jailbreak intent, and self-harm, suggesting that they may not form a single compact distribution in representation space. To examine this, we analyze unsafe prompts within each dataset: ViSU samples are colored by annotated risk types, while MMA and SneakyPrompt are clustered using unsafe-only KMeans in the original CLIP representation space, with UMAP used only for visualization. As shown in Fig. 2, unsafe prompts form multiple local semantic regions rather than one compact cluster, indicating that treating them as a homogeneous positive class may obscure their internal structure.

This multi-modal structure creates difficulty for compact risk representation. Suppose unsafe representations are drawn from K latent risk modes with means $\mu_1, \ldots, \mu_K$ and prior probabilities $\pi_1, \ldots, \pi_K$. If all unsafe samples are represented by a single prototype, the optimal prototype is the global mean $\bar{\mu} = \sum_{k=1}^{K} \pi_k \mu_k$, and the representation error contains an additional inter-mode distortion term. Under ideal mode assignment, using one prototype for each risk mode removes this term, giving

$$D_1 - D_K = \sum_{k=1}^{K} \pi_k \|\mu_k - \bar{\mu}\|_2^2 \geq 0. \tag{2}$$

This result does not imply that nonlinear binary classifiers cannot fit multi-modal data. Rather, it shows that when unsafe prompts are described by a compact risk representation, a single prototype introduces inter-mode distortion. We refer to this distortion as risk-mode entanglement. Together with Fig. 2, this motivates using multiple learnable risk prototypes.

3.3 Coupling Between the Two Challenges

Local risk dilution and risk-mode entanglement reinforce each other. Once local harmful cues are weakened by neutral context, subsequent multi-prototype modeling may fail to recover the core evidence. Meanwhile, mixing heterogeneous risk patterns in a shared space can make the detector rely on coarse global features, further overlooking sparse local cues.

Therefore, addressing either challenge alone is insufficient. D-BISR uses token-level prototype attention to jointly extract local harmful evidence and distinguish different risk modes, then maps the resulting representation into a unified risk-scoring space for threshold-based detection.

4 D-BISR

We consider target-data-free harmful prompt detection. Given a labeled source-domain dataset $\mathcal{D}s = (x_i, y_i)i = 1^N$ with $y_i \in 0, 1$, the detector is trained only on source data and directly applied to unseen target domains. To address sparse local harmful evidence and heterogeneous unsafe patterns, D-BISR uses multiple learnable risk prototypes to aggregate token-level evidence, while global-local fusion incorporates prompt-level context. The resulting representation is then mapped to a continuous risk score through diffusion residual scoring.

4.1 Text Encoding

Given a prompt $x = (w_1, w_2, \ldots, w_N)$ with N tokens, we use a pretrained text encoder E_θ to obtain contextual token representations and a pooled global representation:

$$H = E_\theta(x) = [h_1, \ldots, h_N], \quad h_i \in \mathbb{R}^d, \qquad g = \mathrm{Pool}(H) \in \mathbb{R}^d. \tag{3}$$

The token representations preserve local evidence, while the global representation captures the overall prompt semantics. We project both into a shared risk-evidence space:

$$u_i = \phi(h_i), \qquad u_g = \phi(g), \qquad U = [u_1, \ldots, u_N], \qquad u_i, u_g \in \mathbb{R}^{d_z}. \tag{4}$$

Unlike global-only detectors, D-BISR preserves U for explicit local evidence selection.

4.2 Multi-Prototype Local Evidence Aggregation

To model heterogeneous unsafe patterns, D-BISR introduces K learnable risk prototypes $P = \{p_1, \ldots, p_K\}$, where $p_k \in \mathbb{R}^{d_z}$. These prototypes are learned end-to-end and are not manually assigned to predefined risk categories. Each prototype acts as a risk query that selects relevant evidence from the token sequence.

For token representation u_i and prototype p_k, their matching score and normalized attention weight are computed as

$$\begin{aligned} e_{ik} &= w_a^\top \,\mathrm{GELU}\left(W_a[u_i \| p_k] + b_a\right) + b_e, \\ \alpha_{ik} &= \frac{m_i \exp(e_{ik})}{\sum_{j=1}^{N} m_j \exp(e_{jk})}, \end{aligned} \tag{5}$$

where $m_i \in \{0, 1\}$ is the valid-token mask. The prototype-specific local evidence representation is

$$r_k = \sum_{i=1}^{N} \alpha_{ik} u_i, \qquad r_k \in \mathbb{R}^{d_z}. \tag{6}$$

The output of this component is a set of prototype-level local risk representations $R = \{r_1, \ldots, r_K\}$. This component addresses the two coupled issues in Sec. 3: token attention preserves sparse harmful evidence, and multiple prototypes separate heterogeneous unsafe patterns.

4.3 Global-Local Risk Fusion

Local harmful evidence must be interpreted with global context, since the same phrase may have different safety implications under quotation, negation, fictional framing, or real operational intent. Therefore, D-BISR fuses prototype-level local evidence with the global prompt representation.

For each prototype, we compute a response strength between the prototype and its selected local evidence:

$$q_k = \frac{r_k^\top p_k}{(\|r_k\|_2 + \xi)(\|p_k\|_2 + \xi)}. \tag{7}$$

The value q_k measures how strongly the selected evidence matches the k-th risk prototype.

We concatenate the global representation, all prototype-level local evidence representations, and their response strengths, and map them into a risk-aware representation:

$$\begin{aligned} v &= [u_g \| r_1 \| \cdots \| r_K \| q_1 \| \cdots \| q_K], \\ z_0 &= \mathrm{LN}\left(W_2\, \mathrm{GELU}(W_1 v + b_1) + b_2\right), \qquad z_0 \in \mathbb{R}^{d_z}. \end{aligned} \tag{8}$$

The fused representation z_0 serves as the input to the scoring module and jointly encodes global context, prototype-level local evidence, and prototype response strengths.

To prevent prototype collapse, we introduce a prototype diversity loss:

$$\bar{p}_k = \frac{p_k}{\|p_k\|_2 + \xi}, \qquad \mathcal{L}_{\mathrm{div}} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{\substack{l=1 \\ l \neq k}}^{K} \left(\bar{p}_k^{\top} \bar{p}_l\right)^2. \tag{9}$$

Minimizing this loss encourages different prototypes to capture complementary risk directions.

4.4 Safety-Aligned Diffusion Residual Scoring

After obtaining z_0, D-BISR maps it into a continuous safety risk score. Instead of relying on classifier logits or softmax probabilities, we use residual-based scoring in the representation space. This process is inspired by denoising diffusion, but it is not used for text generation. Its purpose is to measure how much a prompt representation deviates from safety-aligned representation regularities.

Let the process contain T time steps. For a sampled step t, define

$$\alpha_t = 1 - \beta_t, \qquad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \qquad z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{10}$$

A denoising network predicts the added noise, $\hat{\epsilon} = D_\varphi(z_t, t)$, and the denoising residual is defined as

$$R(x; t, \epsilon) = \frac{1}{d_z} \|\epsilon - \hat{\epsilon}\|_2^2. \tag{11}$$

A smaller residual indicates stronger consistency with safety-aligned regularities, while a larger residual indicates stronger deviation.

For safe prompts, we minimize the expected residual:

$$\mathcal{L}_{\mathrm{safe}} = \mathbb{E}_{x^s, t, \epsilon}\left[R(x^s; t, \epsilon)\right]. \tag{12}$$

This encourages safe prompts to form a low-residual region in the risk-aware representation space.

Safe-only residual learning is not sufficient, since it does not explicitly require unsafe prompts to obtain higher risk scores. Therefore, D-BISR further uses paired safe/unsafe supervision. The risk score of prompt x is computed by averaging residuals over M independent noise samples:

$$S(x) = \frac{1}{M} \sum_{m=1}^{M} R(x; t, \epsilon_m), \qquad \epsilon_m \sim \mathcal{N}(0, I). \tag{13}$$

For each paired safe prompt x^s and unsafe prompt x^u, we encourage the unsafe prompt to have a larger score:

$$S(x^u) \geq S(x^s) + \delta, \qquad \delta > 0. \tag{14}$$

The smooth ranking loss is

$$\mathcal{L}_{\mathrm{pair}} = \mathbb{E}_{(x^s, x^u)}\left[\frac{1}{\tau} \log\left(1 + \exp\left(\tau[\delta + S(x^s) - S(x^u)]\right)\right)\right]. \tag{15}$$

Here, $\tau > 0$ controls the sharpness of the ranking penalty.

This objective does not force different unsafe prompts to collapse into one representation center. It only requires unsafe prompts to receive higher residual scores than their paired safe prompts. Thus, heterogeneous harmful patterns can remain structurally distinct in the multi-prototype representation space, while their evidence is mapped onto a unified and comparable risk scale.

Table 1: Main results on length-balanced harmful prompt detection benchmarks.

Dataset	Method	AUROC	TPR@FPR5	Precision	Recall	F1
ViSU	LEG	0.8614	0.0001	0.9381	0.7737	0.8480
	CLIP-Linear	0.9753	0.8988	0.9443	0.9076	0.9256
	CLIP-Mahalanobis	0.8167	0.2192	0.6604	0.9264	0.7711
	HyPE	0.9684	0.9590	0.9774	0.9590	0.9681
	D-BISR	0.9752	0.9615	0.9598	0.9591	0.9592
MMA	LEG	0.9175	0.0002	0.9042	0.9340	0.9189
	CLIP-Linear	0.9554	0.8046	0.9748	0.6865	0.8057
	CLIP-Mahalanobis	0.7961	0.2931	0.5171	0.9987	0.6814
	HyPE	0.9549	0.9289	0.9799	0.9289	0.9537
	D-BISR	0.9656	0.9340	0.9696	0.9315	0.9502
SneakyPrompt	LEG	0.8846	0.0001	0.9286	0.8333	0.8784
	CLIP-Linear	0.7659	0.2628	0.7647	0.5000	0.6047
	CLIP-Mahalanobis	0.6551	0.0962	0.5000	1.0000	0.6667
	HyPE	0.7276	0.0001	0.6606	0.9359	0.7745
	D-BISR	0.9538	0.8397	0.8889	0.9231	0.9057

Table 2: Target-data-free cross-dataset transfer results. Each entry reports AUROC, TPR@FPR5, and F1.

Train Dataset	Test Dataset	Method	AUROC	TPR@FPR5	F1
ViSU	MMA	CLIP-L	0.9754	0.9150	0.9145
		CLIP-M	0.7529	0.2234	0.7103
		DeBERTa	**0.9756**	**0.9429**	0.9413
		D-BISR	0.9575	0.9315	**0.9483**
	SneakyPrompt	CLIP-L	0.8776	0.5769	0.7313
		CLIP-M	0.5397	0.0128	0.6681
		DeBERTa	0.9159	0.7179	0.7757
		D-BISR	**0.9564**	**0.8397**	**0.8994**
MMA	ViSU	CLIP-L	0.8227	0.3931	0.7317
		CLIP-M	0.4097	0.0163	0.6669
		DeBERTa	**0.8890**	**0.6604**	**0.8014**
		D-BISR	0.7079	0.4019	0.5396
	SneakyPrompt	CLIP-L	**0.9598**	0.8077	0.9000
		CLIP-M	0.4235	0.0064	0.6695
		DeBERTa	0.9516	0.8141	0.8926
		D-BISR	0.9454	**0.8718**	**0.9091**
SneakyPrompt	ViSU	CLIP-L	0.8048	0.4308	0.5386
		CLIP-M	0.3661	0.0113	0.6668
		DeBERTa	0.4649	0.0088	0.0106
		D-BISR	**0.8405**	**0.5822**	**0.7600**
	MMA	CLIP-L	0.9330	0.6650	0.8553
		CLIP-M	0.6818	0.1053	0.6667
		DeBERTa	0.8822	0.5228	0.5573
		D-BISR	**0.9773**	**0.9023**	**0.8568**

4.5 Training and Inference

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{safe}} + \lambda_{\text{pair}}\mathcal{L}_{\text{pair}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}. \tag{16}$$

Here, λ_{pair} and λ_{div} control the weights of the paired ranking loss and the prototype diversity loss.

During inference, an input prompt is encoded and processed by multi-prototype evidence aggregation and global-local fusion to obtain z_0. D-BISR then performs M independent noise samplings at a selected diffusion step and uses the average residual as the final risk score $S(x)$. Given a threshold γ selected on the source calibration set, the prediction is

$$\hat{y} = \mathbb{I}[S(x) \geq \gamma], \tag{17}$$

where $\hat{y} = 1$ indicates unsafe and $\hat{y} = 0$ indicates safe. The continuous score $S(x)$ can be used for risk ranking and threshold selection under different false-positive-rate constraints.

5 Experiments

5.1 Experimental Setup

We evaluate D-BISR on three harmful prompt detection benchmarks, ViSU, MMA, and SneakyPrompt Poppi et al. (2024); Yang et al. (2024a;b), mainly using length-balanced test splits, where safe and unsafe prompts are matched by prompt length to reduce shallow cues arising from prompt-length differences. We report AUROC and TPR@FPR5 as ranking-based metrics, together

Table 3: Component ablation results on SneakyPrompt.

Variant	AUROC	TPR@FPR5	Precision	Recall	F1
K1	0.9326	0.7821	0.7526	0.9359	0.8343
K8-no-div	0.9581	0.8333	0.8268	0.9487	0.8836
K8-no-pair	0.5276	0.0449	0.4984	0.9936	0.6638
w/o diffusion	0.9037	0.6987	0.7405	0.8782	0.8035
D-BISR	0.9564	0.8397	0.8882	0.9167	0.9022

with Precision, Recall, and F1 computed using a decision threshold selected by maximizing F1 on a held-out source calibration set and then fixed for testing. We compare against BingoGuard-8B, lightweight classifiers (DeBERTa-v3 Prompt Injection, CLIP-Linear, and LEG), and anomaly- or geometry-based methods (CLIP-Mahalanobis and HyPE). CLIP-based baselines use the same text encoder, data splits, and evaluation protocol. D-BISR uses a CLIP text encoder with $K = 8$, $\lambda_{\text{pair}} = 2.0$ and $\lambda_{\text{div}} = 0.01$.

5.2 Main Results

Table 1 compares D-BISR with representative baselines on the three length-balanced datasets. Overall, D-BISR achieves the strongest low-FPR recall and F1 on ViSU and MMA. On ViSU, D-BISR obtains a TPR@FPR5 of 0.9615 and an F1 of 0.9592; on MMA, it obtains a TPR@FPR5 of 0.9340 and an F1 of 0.9502. In contrast, CLIP-Linear achieves high AUROC on ViSU and MMA but degrades substantially on SneakyPrompt, indicating that a simple linear classifier is sensitive to changes in risk patterns. CLIP-Mahalanobis yields low TPR@FPR5 across all three datasets, suggesting that a single-distance anomaly assumption is insufficient for diverse unsafe patterns.

Different baselines reveal different limitations. HyPE performs strongly on ViSU and MMA, but drops noticeably on SneakyPrompt, suggesting that a single geometric modeling assumption may be less stable for implicit or distribution-shifted unsafe prompts. Compared with these baselines, D-BISR shows more stable detection behavior under low false-positive-rate constraints. It achieves high TPR@FPR5, Recall, and F1 on ViSU and MMA, while maintaining high Recall on SneakyPrompt. These results suggest that multi-prototype local risk modeling and safety-aligned residual scoring help D-BISR detect diverse harmful prompts more reliably under low-FPR operating points.

5.3 Target-Data-Free Source Analysis

We evaluate D-BISR in a source-to-target setting, where both the detector and decision threshold are determined solely from the source dataset and directly transferred to unseen targets without target-domain training or calibration, as shown in Table 2. D-BISR achieves competitive or superior transfer performance and obtains the best TPR@FPR5 in four of the six settings. Compared with DeBERTa-v3, it uses fewer parameters while achieving higher low-FPR recall in several directions. CLIP-Mahalanobis performs poorly in most settings, indicating that a single-distance geometric assumption is insufficient for heterogeneous unsafe patterns, while DeBERTa-v3 shows sensitivity to distribution and threshold shifts in some transfer directions.

D-BISR supports target-data-free transfer through reusable risk modeling: token-level prototype attention captures localized harmful cues, multiple prototypes represent heterogeneous unsafe patterns, and Safety-Aligned Residual Risk Scoring maps them onto a unified risk scale. This enables a threshold selected on the source domain to be directly applied to unseen target datasets.

5.4 Ablation Study

We conduct component ablations on SneakyPrompt using four variants: *K1* with a single risk prototype, *K8-no-div* without the diversity regularizer, *K8-no-pair* without paired safety supervision, and *w/o diffusion* with the diffusion residual scorer replaced by a BCE classifier. As shown in Table 3, paired supervision is critical for calibrating residual magnitude as a risk score: removing it reduces AUROC to 0.5276 and TPR@FPR5 to 0.0449. Multi-prototype modeling also improves detection, with *K8-no-div* increasing TPR@FPR5 from 0.7821 to 0.8333 and F1 from 0.8343 to 0.8836 over *K1*; the diversity regularizer further raises F1 to 0.9022.

Table 4: Efficiency comparison.

Method	Params	Time	Throughput	Peak Mem.
D-BISR	140.23M	0.0028 s/sample	347.0/s	2.8GB
DeBERTa-v3	184.42M	0.0022s/sample	462/s	1.6GB
LEG	70.68M	0.0061s/sample	163/s	3.8GB
HyPE	863.91M	0.0268s/sample	37/s	2.1GB
BingoGuard-8B	8.03B	0.0584s/sample	17/s	17.6GB

Table 5: Stability analysis over three random seeds. Results are reported as mean $\pm$ standard deviation.

Dataset	**AUROC**	**TPR@FPR5**	**F1**
ViSU	0.9759 ± 0.0055	0.9582 ± 0.0008	0.9607 ± 0.0026
MMA	0.9614 ± 0.0065	0.9281 ± 0.0118	0.9464 ± 0.0062
SneakyPrompt	0.9585 ± 0.0018	0.8803 ± 0.0417	0.9054 ± 0.0083

The diffusion residual scorer mainly benefits threshold-based detection under low-FPR constraints. Replacing it with a BCE classifier reduces TPR@FPR5 from 0.8397 to 0.6987 and F1 from 0.9022 to 0.8035, despite retaining a nontrivial AUROC. These results show that the proposed components jointly support heterogeneous risk modeling and a more effective continuous safety score.

5.5 Efficiency Analysis

A practical input-side harmful prompt detector should be both accurate and efficient. We therefore compare model size, inference latency, throughput, and peak GPU memory. The results are shown in Table 4.

D-BISR contains approximately 140.23M parameters, which is much smaller than LLM-based guard models such as BingoGuard-8B. On 6670 samples, D-BISR takes 19.22 seconds for end-to-end inference, corresponding to 0.00288 seconds per sample and a throughput of approximately 347 samples per second. Its peak GPU memory increase is about 2.8GB. In contrast, BingoGuard-8B achieves strong AUROC but requires an 8B model and incurs substantially higher memory and latency. HyPE has moderate memory usage but much slower inference. Overall, D-BISR provides a favorable trade-off between detection performance, low-FPR recall, and deployment efficiency.

5.6 Stability Analysis

For stability analysis, D-BISR is trained with three random seeds and evaluated with bootstrap confidence intervals. For bootstrapping, the decision threshold is fixed to the value selected on the held-out calibration split, without re-selecting it on test or bootstrap samples. Table 5 shows small standard deviations on most metrics, suggesting that the results are not tied to a particular random seed.

6 Conclusion

This paper addresses harmful prompt detection in a target-data-free setting, where a detector is trained on source-domain data and evaluated directly on unseen target domains. We focus on three challenges faced by encoder-based detectors in this setting: local harmful evidence may be diluted by neutral context, diverse unsafe patterns may be mixed under a shared binary objective, and standard classification scores may be unreliable under low false-positive-rate constraints. D-BISR tackles these challenges with token-level prototype attention, learnable risk prototypes, and a safety-aligned diffusion residual score. Together, these designs help preserve local harmful cues, separate different unsafe patterns, and produce thresholdable risk scores. Experiments on ViSU, MMA, and SneakyPrompt show that D-BISR achieves strong low-FPR detection, stable cross-dataset generalization, and lower cost than larger safety detectors.

REFERENCES

Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *Journal of machine learning research*, 2015. URL https://api.semanticscholar.org/CorpusID:2871880.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 2023. URL https://api.semanticscholar.org/CorpusID:258546941.

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 *Advances in Neural Information Processing Systems*, volume 37, 2024.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

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

Md Asiful Islam and Mihai Surdeanu. A lightweight explainable guardrail for prompt safety. *arXiv preprint arXiv:2602.15853*, 2026.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Igor Maljkovic, Maria Rosaria Briglia, Iacopo Masi, Antonio Emanuele Cinà, and Fabio Roli. Harnessing hyperbolic geometry for harmful prompt detection and sanitization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-CLIP: Removing NSFW concepts from vision-and-language models. In *Computer Vision – ECCV 2024*, pp. 340–356. Springer, 2024. doi: 10.1007/978-3-031-73668-1_20.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084, 2019. URL https://api.semanticscholar.org/CorpusID:201646309.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *ArXiv*, abs/2307.02483, 2023. URL https://api.semanticscholar.org/CorpusID:259342528.

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

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. SneakyPrompt: Jailbreaking text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy*, pp. 897–912. IEEE, 2024b.

Fan Yin, Philippe Laban, Xiangyu Peng, Yilun Zhou, Yixin Mao, Vaibhav Vats, Linnea Ross, Divyansh Agarwal, Caiming Xiong, and Chien-Sheng Wu. BingoGuard: LLM content moderation tools with risk levels. In *The Thirteenth International Conference on Learning Representations*, 2025.