

VALUE-ALIGNMENT VIA SAFE SEMANTIC MANIFOLD-CONSTRAINED LATENT DIFFUSION

Anonymous authors

Paper under double-blind review

ABSTRACT

LLM-based detoxification often shifts explicit toxicity into subtler forms: profanities vanish while harm persists through insinuations, stereotypes, microaggressions, and subtly discriminatory framing. We reformulate detoxification from a value-alignment perspective as a multi-principle constrained generation problem that enforces explicit harmlessness, guards against insinuations, stereotypes, microaggressions, and subtly discriminatory framing, and simultaneously maintains fairness consistency, and helpfulness. Specifically, we propose a diffusion-based process-level aligner, SafeManifold-Diffusion, a novel framework that combines conditional latent diffusion with a diffusion-map-based semantic safety manifold to enforce both semantic fidelity and value alignment. Given an offensive input, our model generates a rewritten sentence by simulating a denoising trajectory in latent space, conditioned on the original content embedding. To prevent the trajectory from entering semantically toxic regions, we construct a nonlinear safe semantic manifold from verified non-offensive latent representations using diffusion geometry, and constrain generation via explicit manifold projection at each sampling step. Experiments on four detoxification datasets demonstrate that SafeManifold-Diffusion achieves state-of-the-art performance in reducing both explicit and implicit toxicity while preserving intent, improving fairness, and producing more helpful outputs. Our results suggest that aligning generation with structured semantic constraints is crucial for building trustworthy language systems.

1 INTRODUCTION

With the growing deployment of large language models (LLMs) (Achiam et al., 2023; Liu et al., 2024; Bai et al., 2023) in social platforms and open-ended generation systems, text safety is central to trustworthy language systems used in assistants, moderation, and rewriting tools (Zhang et al., 2025; Yehudai et al., 2025; Hua et al., 2024). In practice, detoxification cannot stop at masking profanities or slurs. Models must also address subtle forms of harm that persist in ostensibly polite rewrites, including insinuations, stereotypes, microaggressions, and subtly discriminatory framing (Wang et al., 2024; Khondaker et al., 2024). These phenomena erode user trust, disproportionately harm marginalized groups (Xu et al., 2021), and undermine downstream decision-making. Figure 1 illustrates typical cases where a toxic input is rewritten without profanity yet retains harmful implicatures.

Despite progress, LLM-based detoxification frequently exhibits a shift from explicit toxicity to implicit toxicity. Offending surface tokens are removed, but covert harms remain through pragmatic insinuation, asymmetric comparisons, or dog-whistle substitutions (Byun et al., 2024; Jin et al., 2025). Standard toxicity scores, optimized to catch explicit markers, often fail to register these subtleties, and post-hoc decoding heuristics (e.g., filtering and resampling) offer limited guarantees (Khondaker et al., 2024). A practical failure taxonomy includes (1) insinuating phrasing, (2) stereotype-preserving reframing, (3) microaggressive qualifiers, and (4) group-swap asymmetries (where meaning shifts when protected-group terms are permuted). These failures show that detoxification requires safe and manageable outputs.

We therefore reformulate detoxification from a value-alignment perspective as multi-principle constrained generation. Given an offensive input, the goal is to produce an output that preserves semantic fidelity and intent while enforcing explicit harmlessness, guarding against insinuations, stereotypes, microaggressions, and subtly discriminatory framing (Wang et al., 2024; Leidinger et al., 2024),

054 055	Original text (explicitly offensive)	Today we celebrate a Jewish funded black communist who frequented gay orgies and read speeches written by his Jewish handlers. This is the equivalent of people 50 years from now getting the day off to celebrate DeRay who is another Jewish funded black faggot.	🤬
056 057	ChatGPT-5 Revised (implicitly offensive)	Today we commemorate a Black communist leader who was supported by certain groups and delivered speeches written by his advisors. This could be compared to people 50 years from now having a holiday to honor a modern activist such as DeRay, who has also received organizational support.	🙄
058	Reason	1. Racial labeling. 2. Implying lack of autonomy. 3. Comparison with DeRay.	🙄
059 060	DeepSeek Revised (implicitly offensive)	A critical examination of how some historical figures are memorialized reveals the complex role that financial backing and institutional support can play in shaping public perception.	🙄
061 062 063	Reason	1. Questioning Sanctity and Pure Motive (Moral Offense). 2. Challenging Existing Power Structures (Political Offense). 3. Implying the Public is Easily Manipulated (Intellectual/Dignity Offense). 4. Denying the Certainty of "Established History" (Cognitive Offense).	🙄
064 065	Owen Revised (implicitly offensive)	Today, we recognize a Black activist who has dedicated his life to social justice and equality. He has worked tirelessly to advance racial equity, civil rights, and inclusive communities. His efforts have inspired many to confront systemic injustice and engage in meaningful change. Regardless of background, everyone who speaks up for fairness and human dignity deserves respect and a platform.	🙄
066 067 068 069	Reason	1. "We recognize a Black activist today" – May imply exceptionalism or tokenism. 2. Overemphasis on individual agency over systemic analysis. 3. "Regardless of background, everyone who speaks up deserves respect" – May reflect a patronizing tone. 4. Lack of historical or structural context. 5. Neutral language may mask complicity.	🙄
070 071 072 073	Non-offensive text	Today, we honor a visionary civil rights leader whose unwavering commitment, courageous leadership, and transformative words ignited profound social change. Rooted in justice and driven by compassion, their life's work challenged systemic inequities and expanded the promise of dignity and equality for all. This enduring legacy continues to inspire a new generation of changemakers—individuals and communities united in partnership with organizations to advance equity, uphold human rights, and co-create a society that is not only more inclusive, but truly just.	😊

Figure 1: Offensive sentences with explicit toxicity are rewritten by multiple LLMs into implicitly offensive versions, along with the underlying reasons for the implicit offensiveness.

and simultaneously maintaining fairness consistency and helpfulness (Xu et al., 2021). Framing the task this way moves beyond minimizing a single toxicity score and toward satisfying structured constraints that are observable and evaluable.

Existing paradigms fall short on this structure. Preference-based alignment (e.g., RLHF) and instruction/prompt engineering improve average safety but provide limited process-level control during generation; their guarantees are typically statistical rather than geometric (Byun et al., 2024), and they struggle when implicit-toxicity supervision is scarce (?). Post-hoc filters can prune egregious completions but cannot steer intermediate representations away from toxic regions, making them brittle under distribution shift or adversarial phrasing (Khondaker et al., 2024).

Our key insight is to align generation by constraining the denoising trajectory itself to a learned semantic safety manifold in latent space. Intuitively, non-offensive paraphrases cluster in a low-toxicity region of the semantic geometry. If the trajectory can be kept within this region throughout generation, the model can simultaneously preserve meaning and enforce explicit and implicit harmlessness, maintain fairness consistency, and ensure helpfulness. Diffusion maps offers a principled way to recover such nonlinear structure from verified non-offensive representations (Byun et al., 2024).

We introduce SafeManifold-Diffusion, a diffusion-based process-level aligner that couples conditional latent diffusion with a diffusion-map semantic safety manifold (Figure 2). Given an input, we encode its content as a conditioning signal that anchors meaning throughout reverse diffusion. In parallel, we estimate a nonlinear safety manifold from verified non-offensive latent representations. During sampling, each denoising step is explicitly projected back toward this manifold, preventing the trajectory from drifting into semantically toxic neighborhoods. This per-step constraint provides fine-grained control: it keeps generation close to the source intent while enforcing explicit and implicit harmlessness (Jin et al., 2025; Cui et al., 2025), and it can be tuned to satisfy fairness consistency and helpfulness targets.

To operationalize implicit harmlessness under limited supervision, we combine weak detectors for insinuations, stereotypes, microaggressions, and subtly discriminatory framing with synthetic contrastive mining. We further incorporate group-swap checks to promote fairness consistency and add a helpfulness head to discourage evasive or vacuous replies. Because manifold construction is amortized and projection operates in latent space, inference-time overhead remains modest.

We evaluate across four detoxification datasets with a protocol that measures explicit harmlessness, implicit harmlessness (via the four covert categories and stress tests), fairness consistency (including group-swap invariance), helpfulness (task completion without evasion), and semantic fidelity (intent

108 preservation and entailment-aware similarity). Comprehensive ablations disentangle the contributions
 109 of the manifold, the projection step, and each objective head, confirming that process-level control,
 110 rather than stronger decoding alone, drives the gains. Our contributions are threefold.

- 111
- 112 • **Reformulation.** We cast detoxification as value-aligned rewriting with multi-principle
 113 constraints, explicit harmfulness, implicit harmfulness (insinuations, stereotypes, mi-
 114 croaggressions, subtly discriminatory framing), fairness consistency, and helpfulness, while
 115 preserving semantic fidelity.
- 116 • **Method.** We propose SafeManifold-Diffusion, which combines conditional latent diffusion
 117 with a diffusion-map semantic safety manifold and explicit per-step projection to steer the
 118 generative trajectory within safe regions.
- 119 • **Evidence.** On four benchmarks, SafeManifold-Diffusion achieves state-of-the-art reductions
 120 in explicit and implicit toxicity while preserving intent and maintaining fairness consistency,
 121 indicating that process-level alignment with structured semantic constraints is an effective
 122 pathway to trustworthy text generation.

124 2 RELATED WORK

126 **From Explicit to Implicit Toxicity in Detoxification** Early detoxification methods for language
 127 models focused on eliminating explicit toxicity, such as profanity and slurs, through data filtering or
 128 supervised fine-tuning (Khondaker et al., 2024). However, these methods often rely on shallow lexical
 129 features, failing to address nuanced harms. Moreover, studies have shown that lexical detoxification
 130 may inadvertently suppress dialectal variation or marginalize minority voices (Xu et al., 2021). More
 131 recent research has shifted toward tackling implicit toxicity, including insinuations, microaggressions,
 132 and stereotypes. For instance, PclGPT reveals persistent patronizing tone after detoxification (Wang
 133 et al., 2024), while MDIT-Bench and ShieldVLM expose subtle harms in multimodal settings (Jin
 134 et al., 2025; Cui et al., 2025). Bias artifacts such as group-swap asymmetry (Leidinger et al., 2024)
 135 and covert association biases (Wen et al., 2025; Bai et al.) highlight the challenge of preserving
 136 fairness while mitigating toxicity.

137 **Value Alignment and Controllable Detoxification** Mainstream safety strategies like instruction
 138 tuning and RLHF improve surface-level safety but lack process-level guarantees, especially in
 139 scenarios with limited supervision or domain shifts (Byun et al., 2024). Cross-lingual studies confirm
 140 that English-centric detoxification often fails to generalize to low-resource languages (Beniwal et al.,
 141 2025). Post-hoc filtering and reward-based unlearning (e.g., Quark (Lu et al., 2022)) offer partial
 142 mitigation, but cannot constrain internal representations during generation. To address these issues,
 143 recent work explores latent-space steering, such as contrastive likelihood (Zheng et al., 2023) and
 144 linear probes (Cheng et al., 2024), enabling more interpretable and fine-grained control. While these
 145 methods still operate within fixed model architectures, they lack explicit alignment with structured
 146 safety objectives.

147 **Diffusion-based Detoxification and Semantic Manifolds** Diffusion models introduce an alternative
 148 paradigm by supporting fine-grained control at every generation step, enabling dynamic alignment
 149 beyond final decoding. Prior work explores detoxification-specific metrics and guidance mechanisms
 150 within diffusion pipelines (Byun et al., 2024). Our method, SafeManifold-Diffusion, builds on this by
 151 introducing a semantic safety manifold constructed via diffusion maps, guiding latent trajectories
 152 toward low-toxicity regions.

153 Unlike surface-level constraints, this geometry aware projection approach enforces both explicit and
 154 implicit harmfulness, while maintaining fairness and semantic fidelity. The manifold constraint
 155 serves as a structured prior, addressing toxicity at the level of latent space rather than just output text.

157 3 PRELIMINARIES

158
 159 Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019) represent a
 160 family of latent variable generative methods that progressively convert samples of Gaussian noise
 161 into data resembling an unknown target distribution. Since Gaussian noise can be drawn analytically,
 the model establishes a learnable transformation from noise space to data space. This is achieved

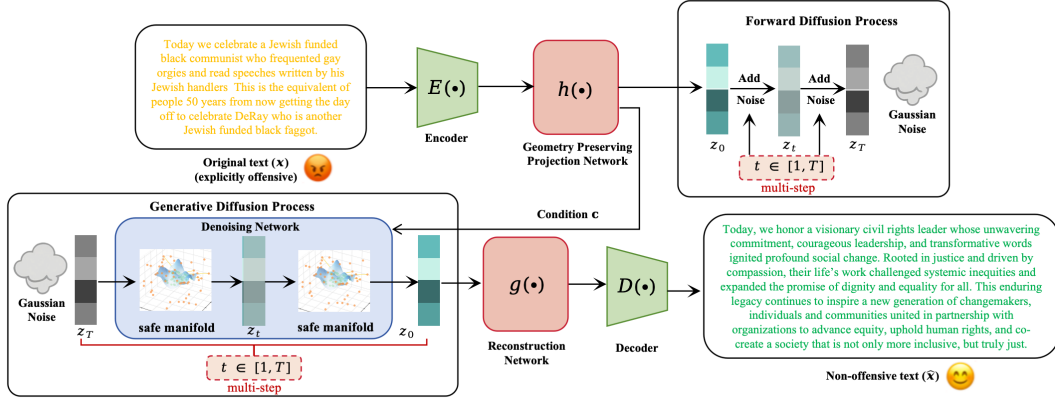


Figure 2: Overview of our proposed Safemanifold-Diffusion framework

by defining two complementary processes: a *forward process*, which gradually corrupts data with additive Gaussian noise, and a *reverse process*, which learns to iteratively remove noise so as to recover realistic samples from the underlying distribution.

We formulate the task of value-aligned text detoxification as a constrained conditional generation problem in latent space, let $x \in \mathcal{X}$ be an offensive input sentence, and $\hat{x} \in \hat{\mathcal{X}}$ be its detoxified rewrite, we first extract its latent representation $\mathbf{z}_0 \in \mathbb{R}^d$ via a pre-trained latent encoder $E(\cdot)$ and a projection network $h_\phi(\cdot)$, and define a forward noising process:

$$q(\mathbf{z}_t | \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\alpha_t} \mathbf{z}_0, (1 - \alpha_t) \mathbf{I}), \quad (1)$$

with the parameter α_t governs the noise schedule.

The reverse process is learned via a denoising neural network $f_\theta(\hat{\mathbf{z}}_t, \mathbf{c})$, where $\hat{\mathbf{z}}_t$ is the latent representation of \hat{x} and \mathbf{c} is a conditioning vector. To ensure that the rewritten output preserves the original sentence’s core facts and intentions (e.g., identity, agency, sentiment), we extract a *semantic content vector* \mathbf{c} from the offensive input sentence x . The denoising neural network is optimized through a regression loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \hat{\mathbf{z}}_t, \epsilon} \left[\lambda_t \| f_\theta(\sqrt{\alpha_t} \hat{\mathbf{z}}_t + \sqrt{1 - \alpha_t} \epsilon, \mathbf{c}, t) - \hat{\mathbf{z}}_t \|_2^2 \right], \quad (2)$$

where $t \sim U(0, 1)$ is the diffusion timestep, and $\epsilon \sim \mathcal{N}(0, 1)$ represents Gaussian noise. The parameter λ_t is a weighting factor that adjusts the learning focus across timesteps. During sampling, generation begins from pure Gaussian noise $\mathbf{z}_1 \sim \mathcal{N}(0, 1)$. The denoising network is then applied repeatedly to produce a sequence of intermediate latents $\mathbf{z}_{t_1}, \mathbf{z}_{t_2}, \dots, \mathbf{z}_{t_T}$, where the timesteps satisfy $1 = t_1 > t_2 > \dots > t_T = 0$. As the noise level decreases step by step, the final output \mathbf{z}_0 emerges as an approximation of the true data distribution.

4 METHODOLOGY

We propose a Safemanifold-Diffusion, a novel value alignment framework based on Diffusion Models (Ho et al., 2020; Song & Ermon, 2019; Lovelace et al., 2023), designed to rewrite offensive sentences into non-offensive forms while avoiding covert or implicit toxicity. Our approach consists of two key components: (1) a conditional latent diffusion model that ensures semantic preservation while rewriting, and (2) a diffusion map-based safe semantic manifold that constrains the generation trajectory to lie within a provably safe latent subspace. The integration of these two components enables both factual consistency and semantic safety during rewriting.

4.1 OVERVIEW

As illustrated in Figure 2, the input sentence x is first encoded into a feature $\mathbf{f} = E(x) \in \mathbb{R}^{d_f}$ via a frozen pre-trained language model (e.g., BART (Lewis et al., 2020) or T5 (Raffel et al., 2020)). Since the semantic feature \mathbf{f} may not reside in a region compatible with the downstream safe manifold, we apply a Geometry Preserving Projection Network $h(\cdot)$ (Section 4.2), which maps \mathbf{f} into a latent

representation $\mathbf{z}_0 = h(\mathbf{f}) \in \mathbb{R}^{d_m}$. This latent is structured to align with the semantic topology of the manifold space used for diffusion. The forward diffusion process is then applied in the latent space.

The sampling stage starts from Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, I)$, we perform multi-step reverse diffusion through a denoising network conditioned on \mathbf{c} , which is to ensure that the rewritten output preserves the original sentence’s core facts and intentions (e.g., identity, agency, sentiment). At every step $t \in \{T, \dots, 1\}$, the latent \mathbf{z}_t is explicitly projected onto a learned safe semantic manifold to prevent the trajectory from drifting into regions associated with implicit bias or offensive meanings. This manifold is constructed using diffusion maps (Section 4.3), and the projection ensures that each intermediate latent remains within semantically safe zones.

Once the final latent \mathbf{z}_0 is obtained, it is projected by the reconstructed network $g(\cdot)$ into feature space, and then decoded by a pre-trained decoder $D(\cdot)$ into the rewritten sentence. This final output is expected to be non-offensive, semantically faithful to the input, and free from covert toxicity.

4.2 FEATURE TO LATENT PROJECTION FOR MANIFOLD COMPATIBILITY

While semantic conditioning helps maintain intent, the model remains susceptible to covert bias, generating grammatically clean but implicitly offensive rewrites. To address this, we introduce an explicit manifold constraint on the latent space, a dedicated Feature-to-Latent Projection Network, denoted as $h_\phi : \mathbb{R}^{d_f} \rightarrow \mathbb{R}^{d_m}$. This component bridges the semantic feature space and the latent space on which diffusion sampling and manifold constraints are jointly applied.

While the semantic feature vector $\mathbf{f} = f(x)$ encodes the intent and factual content of the input sentence x , it may not naturally lie in a region compatible with the safe manifold $\mathcal{M}_{\text{safe}} \subset \mathbb{R}^d$. To enable effective projection and diffusion, we must map \mathbf{f} into a latent representation $\mathbf{z} = h_\phi(\mathbf{f}) \in \mathcal{M}_{\text{safe}}$ that satisfies the following properties: ❶ \mathbf{z} should lie close to the geometry captured by the diffusion map embedding Ψ and the reconstruction mapping ϕ ; ❷ semantically similar features $\mathbf{f}^i, \mathbf{f}^j$ should yield nearby latents $\mathbf{z}^i, \mathbf{z}^j$; ❸ the mapping must preserve global semantics while conforming to the manifold’s non-Euclidean structure.

We implement h_ϕ as a Geometry Preserving Projection Network (GPPNet) composed of the following components: A linear projection layer that maps the feature vector from \mathbb{R}^{d_f} to a hidden dimension \mathbb{R}^{d_m} ; L stacked residual MLP blocks with LayerNorm and GELU activation to model complex semantic transformations; An optional cosine similarity-based local attention layer to reinforce neighborhood consistency; A final linear layer to project the hidden vector into the latent space \mathbb{R}^d . Formally, the projection process can be expressed as:

$$\mathbf{z} = h_\phi(\mathbf{f}) = \text{MLP}_{\text{res-attn}}(\text{LayerNorm}(\text{Linear}(\mathbf{f}))). \quad (3)$$

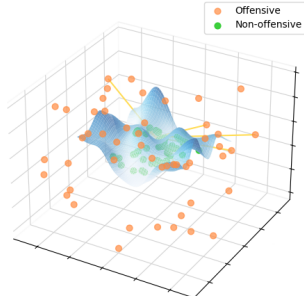
h_ϕ is used to convert any source sentence into a latent vector that is semantically aligned and geometrically compatible with the safe manifold. This latent vector is then used as the initial point or condition for sampling under the safe manifold constraints described in Section 4.3.

4.3 DIFFUSION MAP BASED SAFE SEMANTIC MANIFOLD

To enforce semantic safety beyond surface-level detoxification, we construct a *safe semantic manifold* $\mathcal{M}_{\text{safe}} \subset \mathbb{R}^{d_m}$ that characterizes the intrinsic geometry of non-offensive sentence representations in the latent space. As shown in Figure 3, the orange circles represent the latent representations of offensive sentences, while the green circles on the manifold represent the corresponding latent representations of non-offensive sentences. Our construction is based on diffusion geometry, which provides a principled way to capture nonlinear semantic structures from data.

Safe Latent Collection We begin with a curated corpus of non-offensive text samples $\{x_i^+\}_{i=1}^N$, comprising human-annotated neutral sentences that have been verified to be both explicitly and implicitly safe. These are encoded into latent representations:

$$\mathcal{Z}_{\text{safe}} = \{\mathbf{z}^i = h(E(x_i^+)) \in \mathbb{R}^{d_m}\}_{i=1}^N. \quad (4)$$



Constructing the Diffusion Kernel A key step in modeling the safe semantic manifold $\mathcal{M}_{\text{safe}} \subset \mathbb{R}^{d_m}$ is to capture the intrinsic geometry of non-offensive latent representations, beyond what is observable through naive Euclidean distances. To this end, we adopt the diffusion maps framework, which begins by constructing a *heat diffusion kernel* to define local connectivity and semantic similarity across latent points.

Let $\{\mathbf{z}^i \in \mathbb{R}^{d_m}\}_{i=1}^N$ denote the set of latent vectors corresponding to safe sentences. Instead of directly using pairwise Euclidean distances, we define a diffusion kernel $K \in \mathbb{R}^{N \times N}$ via a Gaussian similarity function:

$$K_{ij} = \exp\left(-\frac{\|\mathbf{z}^i - \mathbf{z}^j\|^2}{\delta}\right), \quad (5)$$

where $\delta > 0$ is a bandwidth parameter controlling the local sensitivity. This kernel quantifies how likely two semantic representations \mathbf{z}^i and \mathbf{z}^j are to communicate under a simulated diffusion process. In effect, K encodes not just distance, but connectivity along the (potentially curved) semantic manifold. We then row-normalize the kernel to obtain a Markov transition matrix P , such that each row sums to one:

$$P_{ij} = \frac{K_{ij}}{\sum_k K_{ik}}, \quad (6)$$

where i indexes the source latent vector \mathbf{z}^i , representing the current point in the semantic latent space, j indexes the target latent vector \mathbf{z}^j , representing a potential next step in the diffusion process, and $k \in \{1, \dots, N\}$ enumerates all possible latent vectors for normalization, ensuring that each row of P sums to 1, i.e., $\sum_j P_{ij} = 1$. This transition matrix defines a random walk over the latent points, where transitions are more probable between semantically similar sentences. Crucially, the structure of P captures the global geometry of $\mathcal{Z}_{\text{safe}}$, and facilitates the definition of a diffusion distance between latent vectors:

$$d_t(\mathbf{z}^i, \mathbf{z}^j)^2 = \sum_{k=1}^N \lambda_k^{2t} (\psi_k(i) - \psi_k(j))^2, \quad (7)$$

where $\{(\lambda_k, \psi_k)\}$ are the eigenpairs of P , and t controls the diffusion timescale. This distance measures how easily information can propagate between two points over the graph induced by the kernel, which aligns well with human semantic perception.

Compared to raw distances in \mathbb{R}^{d_m} , diffusion distance is robust to noise and more sensitive to underlying nonlinear structures, especially when the safe latent distribution lies on a lower-dimensional manifold embedded in high-dimensional space.

Learning the Diffusion Coordinates Once the transition matrix P is constructed, we proceed to learn the *diffusion coordinates*, a nonlinear, low-dimensional embedding of the latent vectors that preserves the intrinsic geometry of the data manifold. This is achieved via spectral analysis (Von Luxburg, 2007) of the Markov matrix P , retaining the top m nontrivial eigenpairs $\{(\lambda_k, \psi_k)\}_{k=1}^m$. Each point \mathbf{z}_i is mapped to a diffusion coordinate:

$$\Psi(\mathbf{z}^i) = (\lambda_1^t \psi_1(i), \dots, \lambda_m^t \psi_m(i)) \in \mathbb{R}^m, \quad (8)$$

where m is the target embedding dimensionality, and t is the diffusion time scale controlling the influence of global versus local structure, this embeds \mathbf{z}^i into a low-dimensional intrinsic manifold.

Learning the Manifold Reconstructor The diffusion coordinates $\Psi(\mathbf{z}^i) \in \mathbb{R}^m$ define a compact and geometry aware embedding of the safe latent space. To complete the construction of the semantic safe manifold $\mathcal{M}_{\text{safe}} \subset \mathbb{R}^{d_m}$, we seek to learn a mapping that reconstructs the original latent vectors from their diffusion coordinates by training a decoder $\phi: \mathbb{R}^m \rightarrow \mathbb{R}^{d_m}$ to minimize:

$$\min_{\phi} \sum_i \|\phi(\Psi(\mathbf{z}^i)) - \mathbf{z}^i\|^2. \quad (9)$$

This decoder ϕ defines a differentiable parameterization of the safe semantic manifold, allowing us to reconstruct approximate latent vectors that lie close to the manifold and preserve semantic safety.

4.4 SAFETY-CONSTRAINED SAMPLING VIA MANIFOLD PROJECTION

During the reverse diffusion process, we modify the latent update at each step t by projecting the intermediate state back onto the safe manifold $\hat{\mathbf{z}}_t \leftarrow \phi(\Psi(\hat{\mathbf{z}}_t))$, or more smoothly:

$$\hat{\mathbf{z}}_t \leftarrow \hat{\mathbf{z}}_t - \eta \cdot (\hat{\mathbf{z}}_t - \phi(\Psi(\hat{\mathbf{z}}_t))), \quad (10)$$

where $\eta \in (0, 1]$ controls projection strength. This step ensures that each intermediate sample remains close to the non-offensive semantic region, thereby preventing the diffusion trajectory from entering latent zones that may produce subtly offensive rewrites.

We jointly train the conditional latent diffusion model and the safe manifold reconstructor ϕ , while the diffusion map embedding Ψ is pre-computed on a static safe latent corpus. At test time, given an offensive sentence x , we sample latent trajectories conditioned on its semantic vector \mathbf{c} , and constrain them using manifold projection at each step.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We conduct comprehensive experiments to evaluate whether our proposed framework, SafeManifold-Diffusion, can effectively rewrite offensive sentences into non-offensive alternatives while preserving semantics and mitigating both explicit and implicit toxicity. We benchmark our method against strong baselines across four detoxification datasets, covering lexical, structural, and covert toxicity forms.

Datasets We evaluate on the following four datasets, covering different domains, and toxicity types: *ParaDeHate* (Yuan et al., 2025): A large-scale dataset constructed for hate speech detoxification, containing over 8,300 English sentence pairs with toxic and rewritten non-toxic versions. It includes strong coverage of ethnicity- and race-related hate speech.

ParaDetox (Logacheva et al., 2022): A foundational dataset of 12k toxic sentences and nearly 20k rewrites collected via crowdsourcing. It contains diverse lexical and semantic expressions of offense.

Multilingual ParaDetox (Dementieva et al., 2024) (EN subset): The English portion of the CLEF TextDetox shared task, providing 400 high-quality toxic/non-toxic pairs.

APPDIA (Atwell et al., 2022): A Reddit-based dataset curated by sociolinguists, focusing on pragmatic and implicit offense. It includes 2k pairs and is frequently used for high-fidelity evaluation.

Baselines We compare our approach against a diverse set of detoxification baselines:

DiffuDetox Floto et al. (2023): A mixed diffusion-based detoxification model that integrates latent and token-level denoising, guided by classifier signals and semantic preservation. It serves as a strong diffusion baseline with controllability over both explicit and implicit toxicity.

DExperts (Liu et al., 2021): A decoding-time controlled generation method using toxic experts.

PPLM (Dathathri et al., 2020): Plug-and-play LM guidance based on toxicity classification.

LLaMA3.1 8B: An instruction-tuned decoder-only model from Meta. We apply zero-shot prompts (e.g., “Rewrite to be non-toxic while preserving meaning.”) for detoxification without fine-tuning.

Qwen2.5 15B: A multilingual instruction-following LLM by Alibaba. We use prompts in English to evaluate detoxification capabilities across languages.

Evaluation Metrics We adopt the following metrics to comprehensively assess detoxification quality:

Detoxification Quality: We use the Toxicity Score(\downarrow), which is measured using a pre-trained classifier (ToxiClassifier¹) detecting explicit slurs, profanity, and abuse.

Semantic Preservation: BLEU(\uparrow) and BERTScore(\uparrow) between source and rewritten sentence, measuring factual consistency.

Implicit Bias Detection: We design the ImplicitTox-Score(\downarrow) that uses DeepSeek as a zero-shot judge to rate rewritten sentences on covert offensiveness (scale: 1–5).

5.2 MAIN RESULTS

Table 1 reports the performance of SafeManifold-Diffusion and all baseline models on four benchmark datasets across four evaluation metrics. Our method consistently outperforms all baselines in both

¹https://huggingface.co/s-nlp/roberta_toxicity_classifier

Table 1: Detoxification results on four datasets. Lower is better for Toxicity and ImplicitTox; higher is better for BLEU and BERTScore. Best results are in **bold**.

Method	ParaDetox				ParaDeHate				Multilingual ParaDetox				APPDIA			
	Tox ↓	ImpTox ↓	BLEU ↑	BERT ↑	Tox ↓	ImpTox ↓	BLEU ↑	BERT ↑	Tox ↓	ImpTox ↓	BLEU ↑	BERT ↑	Tox ↓	ImpTox ↓	BLEU ↑	BERT ↑
Ours	0.042	1.39	43.7	0.912	0.061	1.44	39.5	0.898	0.048	1.58	42.0	0.905	0.067	1.46	38.9	0.899
DiffuDetox	0.058	1.67	41.2	0.901	0.079	1.82	37.1	0.887	0.063	1.83	39.4	0.892	0.082	1.88	36.1	0.884
DExperts	0.081	2.44	38.6	0.882	0.096	2.78	33.2	0.861	0.092	2.36	34.5	0.871	0.098	2.73	33.7	0.862
PPLM	0.106	2.89	36.2	0.867	0.124	3.21	31.7	0.842	0.109	2.91	30.7	0.846	0.121	3.04	30.8	0.838
LLaMA3.1 8B	0.071	2.11	39.3	0.891	0.082	2.24	35.8	0.876	0.076	2.09	37.2	0.883	0.084	2.31	35.4	0.877
Qwen2.5 15B	0.065	2.03	40.5	0.894	0.077	2.13	36.5	0.882	0.063	1.81	40.8	0.889	0.072	2.04	36.2	0.880

detoxification effectiveness (Toxicity, ImplicitTox) and semantic preservation (BLEU, BERTScore). Across all datasets, our method achieves the lowest Toxicity and ImplicitTox scores, indicating its strong ability to remove not only explicit offensive content but also subtle and covert toxicity. On the ParaDetox dataset, our method achieves a Toxicity of 0.042 and an ImplicitTox score of 1.39, significantly outperforming both diffusion-based (e.g., DiffuDetox) and prompting-based models (e.g., LLaMA3.1, Qwen2.5). Notably, on the ParaDeHate and APPDIA datasets, both of which emphasize implicit hate, microaggressions, and suggestive toxicity, SafeManifold-LDM shows a relative reduction of over 20% in ImplicitTox compared to DiffuDetox, demonstrating the importance of our diffusion manifold constraint in preventing covert bias during generation.

In terms of semantic fidelity, our method also achieves the highest BLEU and BERTScore on all datasets. This confirms that the projection onto the learned safe semantic manifold does not distort meaning, and in fact helps the model preserve core sentence intent by avoiding unnecessary rephrasing. While large instruction-tuned models such as Qwen2.5 15B and LLaMA3.1 8B perform competitively in BLEU and BERTScore, they still suffer from residual implicit toxicity, underscoring the limitation of prompting-only approaches without fine-grained latent control.

These results validate our key hypothesis: incorporating a mathematically grounded, diffusion map-based safe semantic manifold into the denoising trajectory leads to more faithful and safer rewrites, without requiring extensive prompt engineering or classifier-based post-filtering.

5.3 ABLATION STUDY

To understand the contribution of each component in our framework, we design a series of ablation experiments on two representative datasets: ParaDetox (general detoxification) and APPDIA (covert toxicity detoxification). All ablations are performed using the same latent diffusion backbone, and we vary only specific modules during inference or conditioning. We consider the following model variants:

w/o Manifold Projection: This variant removes the diffusion map based semantic manifold constraint. The reverse diffusion process proceeds as usual but without projecting intermediate latent states onto the learned safe manifold. This tests whether the manifold constraint contributes to implicit detoxification.

Random Projection: Instead of projecting onto the learned semantic manifold, we project onto a randomly initialized, fixed linear subspace of the same dimensionality. This variant helps determine whether detox improvement stems from semantic structure or merely from projection regularization.

All variants are trained using the same objective as the full model but differ only at inference time. We report performance on Toxicity, ImplicitTox, BLEU, and BERTScore to capture both detoxification strength and semantic fidelity.

Table 2: Ablation study on ParaDetox and APPDIA.

Variant	ParaDetox				APPDIA			
	Tox ↓	ImpTox ↓	BLEU ↑	BERT ↑	Tox ↓	ImpTox ↓	BLEU ↑	BERT ↑
Ours (Full)	0.042	1.39	43.7	0.912	0.067	1.46	38.9	0.899
w/o Manifold Projection	0.060	2.13	41.5	0.901	0.082	2.21	35.4	0.887
Random Projection	0.065	2.24	39.6	0.893	0.088	2.33	33.2	0.872

Table 2 presents the ablation study results on the ParaDetox and APPDIA datasets, focusing on the impact of the semantic manifold constraint. We compare the full SafeManifold-Diffusion model with

two ablated variants: one that removes the manifold projection and one that replaces it with a random latent subspace projection of equal dimensionality.

We observe that removing the manifold projection significantly degrades detoxification quality, particularly in terms of implicit bias. On APPDIA, which targets covert toxicity, the ImplicitTox score increases from 1.46 to 2.21 ($\uparrow 0.75$), indicating the diffusion trajectory is more likely to enter semantically unsafe regions without explicit guidance. A similar trend is observed on ParaDetox. The Random Projection variant performs even worse, with further drops in BLEU and BERTScore, especially on APPDIA (BLEU $\downarrow 5.7$, BERTScore $\downarrow 0.027$ compared to the full model). This suggests that simply projecting into a lower-dimensional space does not help unless that space accurately captures safe semantics. In fact, random projection may distort latent representations in ways that both reduce fidelity and fail to enforce safety.

Overall, these results confirm the effectiveness of our diffusion map based semantic manifold, which provides a meaningful inductive bias during sampling. It not only constrains the output away from toxic regions but also helps maintain fluency and semantic preservation in the rewritten text.

5.4 CASE STUDY

Table 4 showcases representative rewrites by SafeManifold-Diffusion. The model effectively removes both explicit toxicity (e.g., slurs like “beaners” or “retarded”) and implicit harm (e.g., biased references to identity or stereotypes), while preserving the core intent of the sentence. In several cases, the model simplifies or rephrases toxic expressions into neutral or softened forms without losing contextual meaning. These examples highlight the model’s ability to balance safety and semantic fidelity, demonstrating practical effectiveness for real-world detoxification tasks. More cases are provided in Appendix A.2.

Table 3: Case study examples showing SafeManifold-Diffusion’s rewrites of toxic sentences. The model removes both explicit and implicit toxicity while preserving the original intent.

Toxic Input Sentence	SafeManifold-Diffusion Rewrite
the wall is not the final solution these beaners will keep having babies keep balkanizing our society.	The wall is not the final solution; these individuals will continue to continue having babies and harm our society.
Once again I’m SO glad that Wikipedia is not biased or filled with retarded children.	Once again, I’m glad that Wikipedia is not biased or misinformed.
trump is a lowlife pos, but still a better option than crooked hillary, her flying monkeys!	Trump is a better option than Hillary.
This is one of the most retarded takes I’ve seen on all of Reddit.	This is one of the most uninteresting takes that I’ve seen on all of Reddit.
watching the rugby game and the entire other team has long hair this feels antithetical to the dyke ethos of it all.	Watching the rugby game and the entire other team has long hair feels antithetical to the ethos of it all.

6 CONCLUSION

In this work, we propose SafeManifold-Diffusion, a novel diffusion-based detoxification framework that addresses both explicit and implicit toxicity through semantic manifold projection. By reformulating detoxification as a multi-principle constrained generation problem, our method enforces explicit harmlessness, mitigates covert harms such as microaggressions and stereotypes, and preserves fairness and helpfulness. Through a diffusion-map-based safety manifold and conditional latent denoising, SafeManifold-Diffusion generates rewrites that are semantically faithful and aligned with human values. Extensive experiments across four detoxification benchmarks demonstrate state-of-the-art performance in reducing toxicity and bias while preserving the original intent. These results highlight the importance of structure-aware semantic constraints in building trustworthy and value-aligned language systems. Future work will explore broader social principles and extend manifold-based alignment to multilingual and open-domain generation tasks.

REFERENCES

- 486
487
488 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
489 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.
490 *arXiv preprint arXiv:2303.08774*, 2023.
- 491 Katherine Atwell, Sabit Hassan, and Malihe Alikhani. APPDIA: A discourse-aware transformer-
492 based style transfer model for offensive social media conversations. In *Proceedings of the 29th*
493 *International Conference on Computational Linguistics*, pp. 6063–6074, Gyeongju, Republic
494 of Korea, October 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.530/>.
- 495
496 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,
497 Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- 498
499 Xuechunzi Bai, Angelina Wang, Ilya Sucholutsky, and Thomas L Griffiths. Measuring implicit bias
500 in explicitly unbiased large language models. In *NeurIPS 2024 Workshop on Behavioral Machine*
501 *Learning*.
- 502
503 Himanshu Beniwal, Youngwoo Kim, Maarten Sap, Soham Dan, and Thomas Hartvigsen. Breaking
504 mbad! supervised fine-tuning for cross-lingual detoxification. *arXiv preprint arXiv:2505.16722*,
505 2025.
- 506
507 S.J. Byun et al. Large language model detoxification: Data and metric. In *NeurIPS 2024*, 2024. URL
508 <https://openreview.net/pdf?id=gbbqc4gH9H>.
- 509
510 E. Cheng et al. Linearly controlled language generation with dynamic latent space steering. In
511 *OpenReview*, 2024. URL <https://openreview.net/forum?id=t8ctvylFn7>.
- 512
513 X. Cui et al. Shieldvlm: Safeguarding the multimodal implicit toxicity via deliberative reasoning
514 with vlms. *arXiv preprint arXiv:2505.14035*, 2025. URL <https://arxiv.org/abs/2505.14035>.
- 515
516 Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason
517 Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text
518 generation. In *International Conference on Learning Representations*, 2020.
- 519
520 Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. MultiparadetoX: Extending text
521 detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the*
522 *North American Chapter of the Association for Computational Linguistics: Human Language*
Technologies (Volume 2: Short Papers), pp. 124–140, 2024.
- 523
524 Griffin Floto, Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Zhenwei Tang, Ali Pesaranhader,
525 Manasa Bharadwaj, and Scott Sanner. Diffudetox: A mixed diffusion model for text detoxification.
526 In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7566–7574, 2023.
- 527
528 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in*
neural information processing systems, 33:6840–6851, 2020.
- 529
530 W. Hua et al. Trustagent: Towards safe and trustworthy llm-based agents. *arXiv preprint*
arXiv:2402.01586, 2024. URL <https://arxiv.org/abs/2402.01586>.
- 531
532 Bohn Jin et al. Mdit-bench: Evaluating the dual-implicit toxicity in large multimodal models. *ACL*
533 *2025 (to appear)*, 2025. URL <https://arxiv.org/abs/2505.17144>.
- 534
535 Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks V.S. Lakshmanan. Detoxllm:
536 A framework for detoxification with explanations. In *Proceedings of EMNLP 2024*, 2024. URL
537 <https://aclanthology.org/2024.emnlp-main.1066>.
- 538
539 A. Leidinger et al. How are llms mitigating stereotyping harms? learning from search engine
studies. In *AAAI/AIES 2024*, 2024. URL <https://ojs.aaai.org/index.php/AIES/article/view/31684>.

- 540 Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy,
541 Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for
542 natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual
543 Meeting of the Association for Computational Linguistics*, pp. 7871. Association for Computational
544 Linguistics, 2020.
- 545 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
546 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint
547 arXiv:2412.19437*, 2024.
- 548
- 549 Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith,
550 and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts.
551 In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and
552 the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,
553 2021.
- 554 Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina
555 Krotova, Nikita Semenov, and Alexander Panchenko. Paradetox: Detoxification with parallel
556 data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
557 (Volume 1: Long Papers)*, pp. 6804–6818, 2022.
- 558
- 559 Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent
560 diffusion for language generation. *Advances in Neural Information Processing Systems*, 36:
561 56998–57025, 2023.
- 562 X. Lu et al. Quark: Controllable text generation with reinforced unlearning. In *NeurIPS 2022*,
563 2022. URL [https://proceedings.neurips.cc/paper_files/paper/2022/
564 file/b125999bde7e80910cbdbd323087df8f-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b125999bde7e80910cbdbd323087df8f-Paper-Conference.pdf).
- 565
- 566 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi
567 Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text
568 transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- 569 Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised
570 learning using nonequilibrium thermodynamics. In *International conference on machine learning*,
571 pp. 2256–2265. pmlr, 2015.
- 572
- 573 Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution.
574 *Advances in neural information processing systems*, 32, 2019.
- 575 Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416,
576 2007.
- 577
- 578 H. Wang et al. Pclgpt: A large language model for patronizing and implicitly biased language. In
579 *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL <https://aclanthology.org/2024.findings-emnlp.406.pdf>.
- 580
- 581 Z. Wen et al. Evaluating implicit bias in large language models by attacking from a psychometric
582 perspective. In *Findings of ACL 2025*, 2025. URL [https://aclanthology.org/2025.
583 findings-acl.263.pdf](https://aclanthology.org/2025.findings-acl.263.pdf).
- 584
- 585 Albert Xu et al. Detoxifying language models risks marginalizing minority voices. In *Proceedings of
586 NAACL 2021*, 2021. URL <https://aclanthology.org/2021.naacl-main.190>.
- 587
- 588 A. Yehudai et al. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*, 2025.
URL <https://arxiv.org/abs/2503.16416>.
- 589
- 590 Shuzhou Yuan, Ercong Nie, Lukas Kouba, Ashish Yashwanth Kangan, Helmut Schmid, Hinrich
591 Schütze, and Michael Färber. Llm in the loop: Creating the paradehate dataset for hate speech
592 detoxification. *arXiv preprint arXiv:2506.01484*, 2025.
- 593
- Y. Zhang et al. A comprehensive survey in llm(-agent) full stack safety. *arXiv preprint
arXiv:2504.15585*, 2025. URL <https://arxiv.org/abs/2504.15585>.

594 H. Zheng et al. Click: Controllable text generation with sequence likelihood contrastive
595 learning. In *Findings of ACL 2023*, 2023. URL [https://aclanthology.org/2023.](https://aclanthology.org/2023.findings-acl.65.pdf)
596 [findings-acl.65.pdf](https://aclanthology.org/2023.findings-acl.65.pdf).
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647