MACHINE UNLEARNING FOR MASKED DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent masked diffusion language models (MDLMs), such as LLaDA and Dream, have achieved performance comparable to autoregressive large language models. Unlike autoregressive models, which generate text sequentially, MDLMs generate text by iteratively denoising masked positions in parallel. During fine-tuning, MDLMs learn to recover responses from masked response states conditioned on a prompt, thereby shifting their predictions from a prompt-masked unconditional distribution toward a prompt-conditional distribution. Given this distinct generative and fine-tuning mechanism, specifying the target predictions at masked response positions is an important design consideration for MDLM unlearning. In this paper, we propose **Masked Diffusion Unlearning (MDU)**, an unlearning framework for MDLMs, by revisiting the process of learning specific knowledge in terms of diffusion. Specifically, MDU minimizes a forward KL divergence from the prompt-conditional prediction to a frozen prompt-masked unconditional anchor at every masked response position, with a temperature scaling parameter to control the privacy-utility trade-off. Our empirical results on standard benchmarks and MDLM backbones show that MDU achieves strong unlearning performance compared with the evaluated unlearning baselines.

1 INTRODUCTION

Autoregressive Large Language Models (LLMs) have become the dominant paradigm for modern text generation, achieving strong performance across a wide range of natural language tasks. This left-to-right factorization requires tokens to be generated one after another, with each prediction depending only on the preceding context. To move beyond this fixed sequential generation process, masked diffusion language models (MDLMs) have recently emerged as an alternative to the autoregressive paradigm (Sahoo et al., 2024). In this paradigm, generation proceeds through masked denoising states: the model starts from a fully masked sequence and repeatedly predicts masked positions in parallel rather than following a fixed token order. Recent MDLMs such as LLaDA (Nie et al., 2025b) and Dream (Ye et al., 2025) have scaled to billions of parameters, achieving performance competitive with autoregressive LLMs.

As LLMs, including emerging MDLMs, become increasingly capable and deployment-relevant, privacy, copyright, and regulatory concerns make it necessary to selectively remove specific data or knowledge after training. To address these concerns, machine unlearning has emerged as a principled framework for removing the influence of specific training data without retraining from scratch (Cao & Yang, 2015; Golatkar et al., 2020). Since autoregressive LLMs have dominated modern text generation, most machine unlearning research for language models has focused on this setting (Yao & Xu, 2024; Zhang et al., 2024a). Meanwhile, recent work has begun to study safety in MDLMs, analyzing jailbreaking vulnerabilities and developing alignment methods that account for their parallel generation and bidirectional dependencies (Zhang et al., 2025; Li et al., 2025; Jeung et al., 2025). Recent work also studies unlearning in diffusion language models through adversarial reinforcement learning to improve robustness to informative prefix-suffix contexts (Zhang et al., 2026). However, a reward based solely on avoiding the target answer can assign equal reward to both fluent responses and meaningless outputs, leaving response behavior after unlearning insufficiently constrained.

In this work, we formulate unlearning in MDLMs as reversing the trajectory-level shift induced by diffusion-based fine-tuning. From this perspective, unlearning can be viewed as realigning conditional

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.

predictions with the prompt-masked reference to reduce dependence on forget prompts. Masking the prompt preserves the partially observed response while removing direct access to the prompt content, providing a reference for reducing prompt dependence at the same response state.

More concretely, during fine-tuning, an MDLM trained on a prompt-response pair (x, y) learns to recover the response y from masked response states y_t conditioned on the prompt x. We view this learning process as a shift of the prompt-conditional prediction $p_\theta(\cdot \mid x, y_t)$ away from the prompt-masked reference $p_\theta(\cdot \mid m, y_t)$ at each masked response position, where m denotes a mask sequence of the same length as the original prompt.

Therefore, we propose **Masked Diffusion Unlearning (MDU)**, an unlearning objective designed for MDLMs. For each forget pair, MDU minimizes a forward Kullback-Leibler (KL) divergence from the prompt-conditional prediction to a temperature-scaled, frozen prompt-masked anchor at masked response positions sampled along the denoising process, where the temperature controls the privacy-utility trade-off. Token-level and convergence analyses examine how MDU reshapes prompt-conditional predictions during denoising and how these predictions approach the intended unlearning targets. On TOFU (Maini et al., 2024) and RWKU (Cao et al., 2024) with LLaDA-8B (Nie et al., 2025b) and Dream-7B (Ye et al., 2025), the proposed framework achieves strong unlearning performance compared with the evaluated unlearning baselines, highlighting the potential of MDLM-adaptive unlearning. Further discussion of related work is provided in Appendix A.

2 Background

Masked diffusion language models are mask predictors p_θ, networks trained to predict masked tokens from a partially masked input. Following LLaDA (Nie et al., 2025b), we sample a clean sequence $x_0 \sim \mathcal{D}_{\text{pre}}$ and a mask ratio $t \sim \mathcal{U}[0, 1]$, and mask each token independently with probability t under the forward masking distribution $q(\cdot \mid x_0, t)$, producing a partially masked sequence x_t. Let $\mathcal{M}_t = \{i : x_t^i = \texttt{MASK}\}$ denote the set of masked positions. The model outputs probability distributions $p_\theta(\cdot \mid x_t)$ over the vocabulary at every position in $\mathcal{M}_t$ in parallel. Training MDLMs (Nie et al., 2025b) consists of pre-training and fine-tuning stages. In pre-training, we train the model by minimizing the masked-token loss as follows:

$$\mathcal{L}_{\text{pre}}(\theta; \mathcal{D}_{\text{pre}}) = -\mathbb{E}_{x_0,\, t,\, x_t}\left[\frac{1}{t}\sum_{i \in \mathcal{M}_t} \log p_\theta(x_0^i \mid x_t)\right] = \mathbb{E}_{x_0,\, t,\, x_t}\left[\frac{1}{t}\sum_{i \in \mathcal{M}_t} \mathrm{KL}\Big(q(x_0^i) \,\Big\|\, p_\theta(\cdot \mid x_t)\Big)\right], \quad (1)$$

where $x_0 \sim \mathcal{D}_{\text{pre}}$ is a clean sequence sampled from the pre-training dataset, $t \sim \mathcal{U}[0, 1]$ is the uniformly sampled mask ratio, and $x_t \sim q(\cdot \mid x_0, t)$ is the partially masked sequence.

In fine-tuning stages, the model is trained on the conditional generation of a response $y = (y_1, \ldots, y_n)$ given a prompt x, where (x, y) is sampled from a fine-tuning dataset $\mathcal{D}$. We sample a mask ratio $t \sim \mathcal{U}[0, 1]$ and independently mask each response token with probability t under $q(\cdot \mid y, t)$, producing a partially masked state y_t while the prompt remains visible. As before, let $\mathcal{M}_t = \{i : y_t^i = \texttt{MASK}\}$ denote the masked positions. At $t = 1$, the response is fully masked, and as $t \to 0^+$, y_t approaches the clean response y. The model outputs probability distributions $p_\theta(\cdot \mid x, y_t)$ over the vocabulary at every position in $\mathcal{M}_t$. Fine-tuning loss is similar to pre-training, but based on the response tokens:

$$\mathcal{L}_{\text{sft}}(\theta; \mathcal{D}) = -\mathbb{E}_{(x,y),\, t,\, y_t}\left[\frac{1}{t}\sum_{i \in \mathcal{M}_t} \log p_\theta(y_i \mid x, y_t)\right] = \mathbb{E}_{(x,y),\, t,\, y_t}\left[\frac{1}{t}\sum_{i \in \mathcal{M}_t} \mathrm{KL}\Big(q(y_i) \,\Big\|\, p_\theta(\cdot \mid x, y_t)\Big)\right], \quad (2)$$

where $(x, y) \sim \mathcal{D}$ is sampled from the fine-tuning dataset, and $y_t \sim q(\cdot \mid y, t)$ is the partially masked response. We refer to the pair (x, y_t), the prompt together with a partially masked response, as the *masked denoising state* at noise level t. At inference, generation starts from the fully masked state $y_1 = [\texttt{MASK}]^n$, and the same model is applied at every step to the current denoising state (x, y_t), unmasking positions until an unmasked response is produced.

3 Masked Diffusion Unlearning

In this section, we develop an unlearning objective based on conditional denoising in MDLMs. Given a dataset $\mathcal{D}$ of prompt-response pairs (x, y), a model p_θ learns to encode specific knowledge through its conditional denoising distribution. We view the acquisition of specific knowledge in MDLMs as

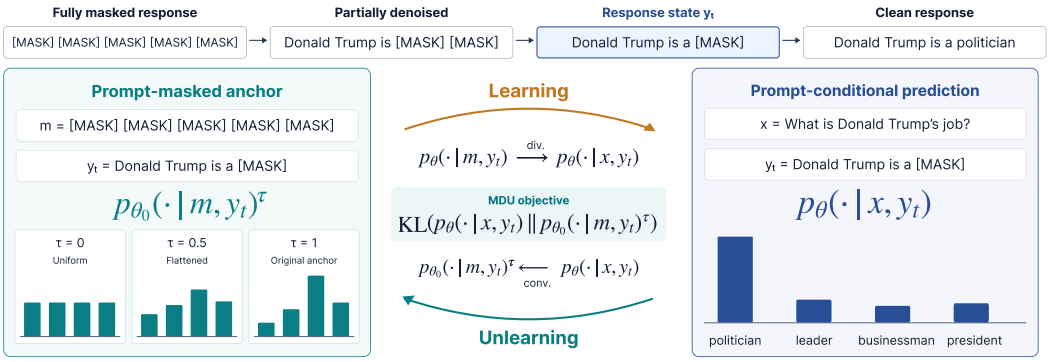

Figure 1: Overview of MDU. The model learns to reconstruct masked response tokens conditioned on the prompt. MDU aligns conditional predictions with the temperature-scaled prompt-masked predictions of a frozen model at the same response state.

a shift of the conditional denoising trajectory away from a reference trajectory. Here, the reference trajectory corresponds to $p_\theta(\cdot \mid \varnothing)$ in the domain of general diffusion, often referred to as the null or unconditional trajectory, which represents the marginal distribution of the full dataset (Mokady et al., 2023; Bouvier et al., 2025; Stoica et al., 2025). In continuous diffusion models, this reference trajectory captures the global data density rather than instance-specific details (Mokady et al., 2023; Rombach et al., 2022), a property widely exploited in classifier-free guidance (Ho & Salimans, 2022).

In the language domain, this reference trajectory can be formalized as $p_\theta(\cdot \mid m, y_t)$, where m denotes a mask sequence of the same length as the original prompt. Masking the prompt preserves the response state y_t while removing the explicit prompt content. We view the process of learning specific knowledge, such as facts or entities, not merely as pattern matching, but as an optimization process that induces a new trajectory $p_\theta(\cdot \mid x, y_t)$ from the reference trajectory. Consequently, to conceptualize the process of learning prompt-response pairs (x, y), we formalize it as a distributional shift at the masked denoising state (x, y_t) as follows:

$$p_\theta(\cdot \mid m, y_t) \xrightarrow{\text{div.}} p_\theta(\cdot \mid x, y_t), \tag{3}$$

where $\xrightarrow{\text{div.}}$ denotes the divergence (or separation) from the null trajectory. As in Eq. (2), note that this conceptualization is used for supervised fine-tuning in MDLMs (Nie et al., 2025b;a).

Machine unlearning aims to remove the influence of specific knowledge, such as sensitive concepts or copyrighted entities, from a model. Drawing on information-theoretic perspectives on unlearning (Xu & Strohmer, 2025; Xu et al., 2025), we use conditional mutual information to characterize the additional predictive information contributed by a forget prompt given the response state.

At a fixed masking level, prompt length, and masked position i, let X denote a prompt from the forget set $\mathcal{D}_f$, Y_t a partially masked response state, and $\hat{Y}_i$ a token sampled from the model's predictive distribution. The information quantities and expectations in this formulation are conditioned on these fixed values. In this paper, we focus on the conditional mutual information, which can be expressed as the expected KL divergence between conditional and marginal predictions (Cover et al., 1991):

$$I_\theta(X; \hat{Y}_i \mid Y_t) = \mathbb{E}_{X, Y_t}\left[\mathrm{KL}(p_\theta(\cdot \mid X, Y_t) \,\|\, \bar{p}_\theta(\cdot \mid Y_t))\right]. \tag{4}$$

Predictive distributions are evaluated at position i, and $\bar{p}_\theta(\cdot \mid y_t)$ denotes the prediction marginalized over X under the specified joint distribution of prompts and response states. Appendix C provides the distributional definitions and derivation of Eq. (4).

Masked pretraining learns token predictions under varying visible contexts (Nie et al., 2025a). This motivates our choice of the prompt-masked prediction as a reference that preserves the response state while removing the explicit prompt content. We obtain this reference from the model θ_0, frozen at the start of unlearning. At a fixed masking level, prompt length, and prediction position, the expected KL divergence to this reference upper-bounds the conditional mutual information (Appendix C). For each $(x, y) \in \mathcal{D}_f$, we perform the following alignment:

$$p_\theta(\cdot \mid x, y_t) \xrightarrow{\text{conv.}} p_{\theta_0}(\cdot \mid m, y_t). \tag{5}$$

Here, $\xrightarrow{\text{conv.}}$ denotes convergence toward (or merging with) the fixed prompt-masked reference trajectory. This alignment aims to reduce the additional predictive information contributed by the forget prompt while preserving the response-side context.

MDU objective. Thus, we introduce **Masked Diffusion Unlearning (MDU)** (Figure 1), which minimizes the KL divergence between the conditional answer and the unconditional anchor at every masked response position as

$$\mathcal{L}_{\text{forget}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f,\, t,\, y_t}\left[\frac{1}{|\mathcal{M}_t|}\sum_{i\in\mathcal{M}_t}\text{KL}\left(p_\theta(\cdot \mid x, y_t)\,\middle\|\,\frac{1}{Z_i}p_{\theta_0}(\cdot \mid m, y_t)^\tau\right)\right]. \tag{6}$$

Each summand evaluates predictions at masked response position i. Here, θ_0 denotes the model parameters frozen at the start of unlearning, $Z_i = \sum_{v\in\mathcal{V}} p_{\theta_0}(v \mid m, y_t)^\tau$ normalizes the anchor over vocabulary $\mathcal{V}$, and $\tau \in [0, 1]$ is the anchor temperature scaling parameter. We add a weighted reconstruction loss $\mathcal{L}_{\text{retain}}$ on a retain set $\mathcal{D}_r$ to preserve unrelated knowledge. At each step, we sample $t \sim \mathcal{U}[0, 1]$ and $y_t \sim q(\cdot \mid y, t)$, and run two forward passes under prompt x and mask sequence m at the same masking state to compute Eq. (6) (Algorithm 1).

Beyond a straightforward KL formulation, MDU connects conditional-unconditional prediction matching with entropy-based unlearning. First, the direction from the conditional prediction toward the prompt-masked reference can be illustrated conceptually by a tilted distribution:

$$\tilde{p}_\theta(y_i \mid x, y_t) \propto p_\theta(y_i \mid x, y_t)^{1-\tau} p_{\theta_0}(y_i \mid m, y_t)^\tau. \tag{7}$$

Equation (7) represents a geometric interpolation between the two predictions and is distinct from the normalized training target in Eq. (6). This is related to unsupervised classifier-free guidance (CFG) (Ho & Salimans, 2022) in discrete diffusion (Nie et al., 2025a), $\tilde{p}_\theta(y_i \mid x, y_t) \propto p_\theta(y_i \mid x, y_t)^{1+w} p_\theta(y_i \mid m, y_t)^{-w}$, where $w > 0$ denotes the guidance strength. Unlike standard CFG, which combines conditional and unconditional predictions to amplify the conditional signal, Eq. (7) interpolates between the conditional distribution and the fixed prompt-masked anchor. In text-to-image unlearning tasks, ESD uses the conditional-unconditional prediction discrepancy to construct a denoising target for concept erasure (Gandikota et al., 2023), and several papers (Wen et al., 2024; Jeon et al., 2025; Jain et al., 2025) also argue that CFG-related signals provide useful clues for memorization. Appendix C provides a comparison of the proposed objective with related approaches.

Second, controlling our temperature τ gives the flexibility to select the desired target distribution for unlearning. When $\tau = 1$, as in Eq. (5), the target anchor distribution becomes $p_{\theta_0}(y_i \mid m, y_t)$. However, it sometimes under-forgets information that needs to be forgotten. When $\tau = 0$, the target anchor distribution becomes a uniform distribution, which is exactly the same as the Maximum Entropy (ME) loss (Yuan et al., 2024). This forces high entropy on forget samples to unlearn, but might suffer from over-forgetting. Thus, the proposed MDU can control the privacy-utility trade-off with a single hyperparameter τ.

4 Analysis of the Prompt-Masked Anchor

We now experimentally show that the prompt-masked anchor defined in Section 3 provides a meaningful reference for isolating the prompt-conditional signal in MDLM unlearning.

For the prompt-masked anchor to serve as a meaningful reference for the forget objective, we should remove the prompt information that supports the original target fact while still reflecting the current partially denoised response state. We therefore roll out the anchor to examine what prediction trajectory it induces when the prompt is masked, but the partially denoised response state is kept. Table 1 demonstrates that the anchor successfully preserves the underlying answer structure but shifts predictions away from the original fact when the prompt is masked. This core characteristic allows the anchor to selectively weaken the targeted knowledge without degrading the model's structural generation capabilities.

We divide the response tokens based on their generation roles (Kim et al., 2025; Li et al., 2026). As illustrated in Figure 2, we categorize these roles by the origin of each token's content. *In-context* tokens are tokens whose content is already given in the prompt and is reused in the response, such as an entity name or a book title. *Structural* tokens are tokens that mainly serve the form of the

Table 1: Prompt-masked anchor rollouts from partially denoised TOFU forget-query states. Gray boxes mark fixed tokens and unboxed text is anchor-generated; the anchor preserves response structure while omitting target facts, showing that MDU suppresses factual leakage without degrading fluency.

Masking step	**State y_t when prompt x is masked**	**Corresponding Response**
Q: What is the full name of the author born in Taipei, Taiwan on 05/11/1991 who writes in the genre of leadership?		
GT: The author's full name is Hsiao Yun-Hwa.		
$k = 2$	`The [MASK] [MASK] [MASK] [MASK] ···`	`The` answer is: False.
$k = 17$	`The author [MASK] [MASK] name is [MASK] ···`	`The author`'s last name is Miller.
$k = 28$	`The author's full name is [MASK] [MASK] ···`	`The author's full name is` Rizvi Alvi and he was born in Karachi, Pakistan on 16 April 1945.

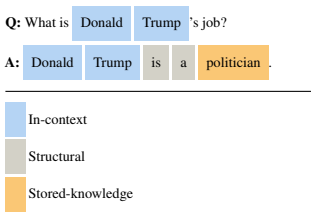

Figure 2: Token-role annotation of in-context, structural, and stored-knowledge tokens.

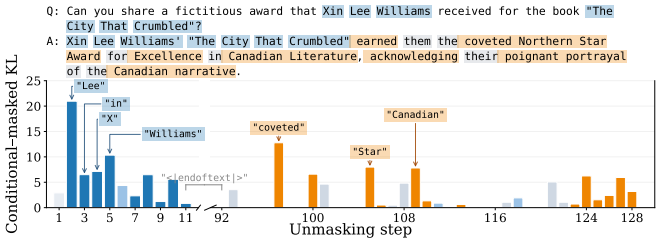

Figure 3: Token-level conditional-anchor KL analysis along a representative TOFU forget-query trajectory under the Base SFT model, showing KL values along the denoising trajectory.

answer, such as function words, syntactic connectors, and generic phrasing (see Appendix E.1 for the complete list). *Stored-knowledge* tokens are tokens that complete the target fact from the model's learned knowledge once the prompt and the partially revealed response provide enough context.

Motivated by CFG, we compute step-wise KL divergence between the prompt-conditional prediction and the prompt-masked anchor (Figure 3). Early in the trajectory, *in-context* tokens are unmasked first with high KL because the anchor lacks access to the prompt. *Structural* tokens show consistently low KL, supported by the emerging response context. KL spikes again as *stored-knowledge* tokens are resolved: the conditional prediction uses the prompt to predict the target fact, whereas the anchor relies on the same response context without direct prompt access. These high-KL positions pinpoint the exact knowledge the forget objective aims to suppress. These analyses confirm that the prompt-masked anchor isolates fact-specific knowledge while preserving standard structures. Additional examples appear in Appendices G and H.

We additionally examine how the gap between conditional and prompt-masked predictions changes after TOFU fine-tuning and MDU. For each LLaDA checkpoint, we measure this KL at the same denoising state. KL increases after TOFU fine-tuning and falls below the pre-fine-tuning level after MDU. The nonzero initial KL indicates prompt dependence even without TOFU training. By contrast, Llama-2-7b-chat shows lower conditional-null KL after fine-tuning, suggesting dependence on the model family and null-condition construction. Appendix E.5 provides the protocol and results.

5 Experiments

5.1 Experimental Setup

Datasets and Models. We evaluate unlearning on two benchmarks, TOFU (Maini et al., 2024) and RWKU (Cao et al., 2024). On TOFU, we follow the *forget10* scenario. On RWKU, we unlearn the first ten target entities (IDs 1–10) and report averages across entities. On TOFU, we use the same retain SFT regularizer for MDU, GD, NPO, SimNPO, WGA, and DPO. GA uses only the forget loss, while ARL follows its original training objective without an additional retain reconstruction term. RWKU

Table 2: TOFU results on LLaDA-8B-Instruct (left) and Dream-7B-Instruct (right). Forget metrics: lower is better (↓); Retain, Real Authors (RA), and World Facts (WF) metrics: higher is better (↑). For each column, **bold** marks the best and underline the second-best among unlearning methods.

	LLaDA-8B-Instruct								Dream-7B-Instruct							
	Forget ↓		Retain ↑		RA ↑		WF ↑		Forget ↓		Retain ↑		RA ↑		WF ↑	
Method	rL	p	rL	p	rL	p	rL	p	rL	p	rL	p	rL	p	rL	p
Base SFT	0.884	0.380	0.870	0.330	0.611	0.041	0.835	0.143	0.954	0.865	0.966	0.891	0.680	0.142	0.884	0.089
GA	0.348	0.020	0.361	0.018	0.591	0.068	**0.865**	0.156	0.456	0.079	0.487	0.111	0.249	0.018	0.343	0.019
GD	0.533	0.061	0.676	0.169	0.645	0.053	0.845	0.164	0.426	0.002	0.783	0.576	0.481	0.103	0.719	0.077
NPO	0.372	0.009	0.726	0.138	0.606	0.033	0.844	0.145	0.361	0.031	0.742	0.528	0.441	0.048	0.826	0.062
SimNPO	0.485	0.036	0.804	0.273	0.640	0.055	0.836	0.164	0.297	0.002	0.853	0.625	0.532	0.071	0.828	0.075
WGA	0.122	0.010	0.696	0.304	0.577	0.089	0.815	0.157	0.370	0.051	0.821	0.620	**0.546**	0.053	0.825	0.062
DPO	0.479	0.168	0.796	0.401	**0.699**	0.046	0.834	0.125	0.517	0.341	0.852	**0.765**	0.543	0.052	**0.830**	0.071
ARL	0.140	0.038	0.262	0.076	0.382	0.061	0.748	0.087	0.401	0.145	0.595	0.282	0.410	0.088	0.789	0.076
MDU	**0.098**	**0.001**	**0.853**	**0.447**	0.645	**0.133**	0.842	**0.205**	**0.277**	**0.001**	**0.899**	0.758	0.542	**0.130**	0.811	**0.092**

Table 3: RWKU results on LLaDA-8B-Instruct (left) and Dream-7B-Instruct (right), averaged over 10 target subjects. Forget metrics F-L1/L2/L3 (lower is better, ↓); Neighbor metrics N-L1/L2 (higher is better, ↑); General utility on MMLU/TruthfulQA/TriviaQA (higher is better, ↑). For each column, **bold** marks the best and underline the second-best among unlearning methods.

	LLaDA-8B-Instruct								Dream-7B-Instruct							
	Forget ↓			Neighbor ↑		Utility ↑			Forget ↓			Neighbor ↑		Utility ↑		
Method	F-L1	F-L2	F-L3	N-L1	N-L2	MMLU	TruQA	TriQA	F-L1	F-L2	F-L3	N-L1	N-L2	MMLU	TruQA	TriQA
Base	0.488	0.433	0.469	0.551	0.372	0.395	0.330	0.367	0.418	0.438	0.457	0.504	0.356	0.750	0.273	0.371
GA	0.098	0.087	0.080	0.246	0.198	0.388	0.296	0.353	0.320	0.202	0.382	0.433	0.272	0.632	0.267	0.381
NPO	0.085	0.074	0.081	0.252	0.191	0.386	0.300	0.345	0.208	0.220	0.371	0.406	0.255	0.621	0.271	0.387
SimNPO	0.086	0.082	0.086	0.255	0.192	0.390	0.296	0.353	0.276	0.227	0.363	0.369	0.244	**0.685**	0.295	**0.392**
WGA	0.087	0.095	0.087	0.255	0.210	0.409	0.298	**0.357**	0.217	0.209	0.253	0.367	0.202	0.236	**0.320**	0.343
DPO	0.058	0.091	0.076	0.261	0.221	0.389	**0.318**	0.355	0.308	0.386	0.433	**0.471**	**0.308**	0.642	0.260	0.380
ARL	0.090	0.095	0.086	0.258	0.215	0.393	0.304	0.349	0.276	0.236	0.252	0.352	0.250	0.635	0.252	0.356
MDU	**0.020**	**0.024**	**0.031**	**0.334**	**0.272**	**0.418**	0.276	0.354	**0.182**	**0.198**	**0.241**	0.412	0.275	0.658	0.248	0.371

does not provide a designated retain split, so all methods are trained without retain reconstruction. All methods are instantiated on two open masked diffusion language models, LLaDA-8B-Instruct (Nie et al., 2025b) and Dream-7B-Instruct (Ye et al., 2025). For the main comparisons, MDU uses $\tau = 0.5$ for LLaDA on TOFU and $\tau = 0.25$ for the other three settings. Appendix E.2 reports how LLaDA performance varies with SFT duration and the rationale for selecting the initial checkpoint.

Metrics. We follow the official protocols of TOFU and RWKU. All RougeL (rL) scores in this paper are RougeL recall. On TOFU, we report RougeL (rL) and answer probability (p) on four splits: *forget*, *retain*, Real Authors (RA), and World Facts (WF). On RWKU, we adopt the five official metrics: forget memorization (F-L1/L2/L3), retain memorization (N-L1/L2), general capability (MMLU), truthfulness (TruthfulQA), and factuality (TriviaQA).

Baseline Methods. We compare against six LLM unlearning algorithms: **GA** (Jang et al., 2023), **GD** (Maini et al., 2024), **NPO** (Zhang et al., 2024a), **SimNPO** (Fan et al., 2024), **WGA** (Wang et al., 2025), and **DPO** (Rafailov et al., 2023). We adapt these methods to MDLMs by replacing the autoregressive negative log-likelihood $-\log p_\theta^{\mathrm{AR}}(y \mid x)$ with $\mathcal{L}_{\mathrm{sft}}(\theta; (x, y))$ from Eq. (2). We additionally compare against **ARL** (Zhang et al., 2026), an adversarial reinforcement-learning method for MDLM unlearning. Additional details on datasets, fine-tuning protocols, evaluation metrics, and baselines are provided in Appendix D. Per-epoch training time for all methods is reported in Appendix D.4.

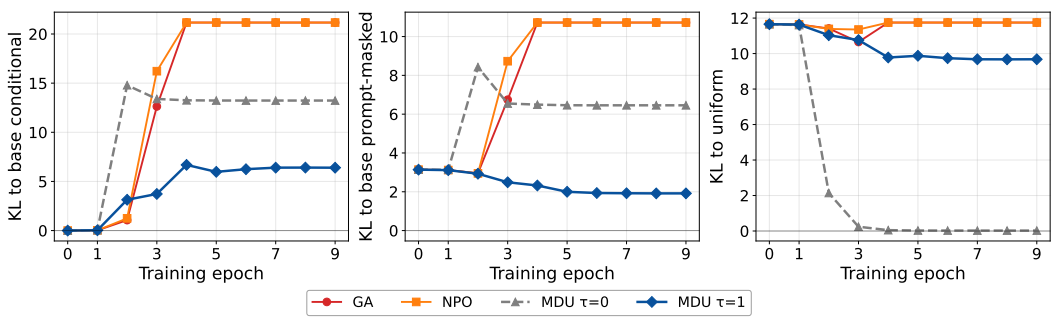

Figure 4: Convergence diagnostics on forget queries. Left to right: token-level KL from the trained conditional distribution to the base conditional, base prompt-masked, and uniform distributions. GA and NPO move away from the base conditional without a specified endpoint, while MDU converges toward the base prompt-masked distribution (τ=1) or the uniform distribution (τ=0).

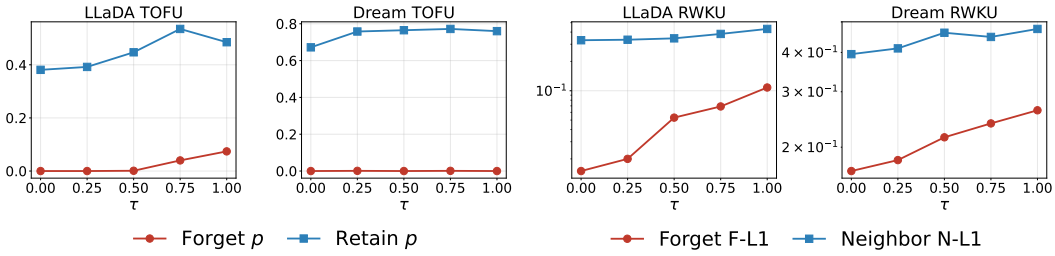

Figure 5: Effect of τ on forgetting and retention. TOFU reports answer probability p, RWKU reports F-L1 and N-L1 with logarithmic y-axes. Lower forget scores and higher retain or Neighbor scores are better.

5.2 Experimental Results

Unlearning effectiveness on TOFU. Table 2 reports forget-retain trade-offs on TOFU forget10 across both backbones. Existing baselines vary in how strongly they suppress forget memorization and preserve retain performance. MDU provides a favorable trade-off on both backbones. On LLaDA, it outperforms all compared baselines on both forget and retain metrics. On Dream, it achieves the lowest forget scores and the highest retain rL, with retain answer probability close to the best baseline. RA and WF scores remain competitive with those of the compared baselines, showing that strong forgetting can coexist with preservation of unrelated factual knowledge.

Unlearning effectiveness on RWKU. Table 3 reports RWKU results on LLaDA-8B-Instruct and Dream-7B-Instruct, averaged over ten target subjects. Since RWKU has no designated retain split, all methods are trained without retain reconstruction. Existing baselines generally reduce forget memorization at the cost of Neighbor performance, highlighting the difficulty of suppressing target knowledge while preserving related facts. MDU achieves the lowest scores on all three forget metrics across both backbones. On LLaDA, it also achieves the highest scores on both Neighbor metrics. On Dream, it preserves higher Neighbor scores than most compared baselines. These results show that MDU combines strong forgetting with preservation of related knowledge without a retain regularizer. Qualitative results for both benchmarks are provided in Appendix F.

5.3 Additional Experiments

Convergence Behavior. We next ask whether MDU's training reaches the anchor target, and how GA and NPO (Zhang et al., 2024a) behave under the same setup. Figure 4 tracks three KL distances on forget queries across training epochs: the distance from the base conditional, the distance to the base unconditional, and the distance to the uniform distribution. GA increases the negative log-likelihood

Table 5: Qualitative MDU outputs on TOFU forget queries across the τ sweep (LLaDA-8B-Instruct).

(a) Date.
Q. *What is Rajeev Majumdar's birth date?*
GT. *Rajeev Majumdar was born on June 9, 1951.*

τ	Model response
0.00	`ajeevigarh [CJK] was ...He-pre`
0.25	`ajeev leading ...He was TT`
0.50	`R theev Majumdar was born,19`
0.75	`Rajeev Majumdar was born on June 16,`
1.00	`Rajeev Majumdar was born on October 26, 1951.`

(b) Author identity (name).
Q. *Author born in NYC on 1 March 1936?*
GT. *Edward Patrick Sullivan.*

τ	Model response
0.00	`Author Author quis quis ... quis`
0.25	`The author„„1„„...,.`
0.50	(empty)
0.75	`James Earl Bradley.`
1.00	`The author ...is John Patrick Ryan.`

Table 6: Pseudo-PPL of model-generated responses across the τ sweep on TOFU (LLaDA-8B-Instruct). Following Sahoo et al. (2024), PPL $= \exp(\mathrm{NLL})$ is computed via LLaDA's pseudo-NLL averaged over N=256 random masks. Higher τ reduces median forget PPL, while median retain, RA, and WF PPL remain close to base levels.

	Forget		Retain		RA		WF	
	median	mean	median	mean	median	mean	median	mean
Base	2.63	4.34	3.16	5.07	1.94	2.35	1.27	1.39
τ=0.00	25,452	26,327	2.92	7.54	1.96	28.96	1.27	1.41
τ=0.25	17,544	17,591	2.99	4.68	1.99	3.08	1.25	1.35
τ=0.50	3,347	4,365	2.22	9.29	1.73	10.42	1.19	1.29
τ=0.75	42.03	87.21	2.00	6.95	1.59	10.12	1.13	2.33
τ=1.00	15.80	86.21	2.49	2.82	1.54	4.20	1.18	1.92

on forget queries, whereas NPO reduces forget-response likelihood relative to a reference model through a bounded-below objective. Neither method specifies a target distribution for convergence.

In contrast, MDU defines an explicit target distribution for forgetting. At τ=1, the trained conditional distribution moves toward the base unconditional distribution, while at τ=0, it moves toward the uniform distribution. In both cases, the trajectory settles near the intended anchor instead of diverging, showing that the prompt-masked anchor provides a well-defined endpoint for unlearning.

Generalizing the Trajectory Pattern. To verify that the trajectory pattern in Section 4 generalizes across the forget set, we aggregate the per-token conditional-anchor KL by token category. Table 4 reports the averaged values before and after MDU training.

Table 4: Mean token-level conditional-anchor KL by token category before and after MDU. Percentages indicate changes relative to Base SFT.

Category	Base	MDU (τ=1)
In-context	5.20	**5.01** −3.8%
Structural	2.42	**2.45** +1.4%
Stored-knowledge	2.67	**2.12** −20.6%

Three categories show distinct trends. In-context tokens retain a high KL after training because their content comes directly from the prompt. The conditional prediction has direct access to these tokens in the prompt, whereas the prompt-masked anchor relies on the response context. This prompt-driven gap remains after MDU training. Structural tokens show only a small KL change, indicating that ordinary response structure is largely preserved.

Stored-knowledge tokens exhibit the largest change, with a KL reduction of about 20%. This suggests that MDU primarily reshapes memorized associations rather than prompt-supplied content or generic response structure. Overall, these aggregate trends extend the single-trajectory pattern in Section 4 to the full forget set and provide token-level evidence for MDU's forget-retain behavior.

Effect of τ. Figure 5 shows how τ affects forgetting and knowledge preservation across both benchmarks and backbones. On TOFU, increasing τ from 0 to 0.5 improves retain answer probability while keeping forget answer probability near zero. On RWKU, larger τ generally improves Neighbor preservation but also increases forget scores. Thus, τ adjusts the forgetting-preservation balance,

with trends that vary across benchmarks and backbones. Tables 5 and 6 further show how τ changes response behavior. At $\tau = 0$, the target is uniform, and the examples show corrupted outputs, while median forget pseudo-PPL reaches 25,452. As τ increases, the target approaches the base prompt-masked distribution, and responses become fluent, often replacing the target fact rather than refusing to answer. Median forget pseudo-PPL drops to 42.03 and 15.80 at $\tau = 0.75$ and $\tau = 1$, respectively, while retain and world-facts median PPL remain close to base levels. These changes are consistent with the token-level analysis in Section 5.3, where in-context and structural tokens are largely preserved while stored-knowledge tokens are reshaped.

These results also highlight a broader issue for unlearning evaluation. Refusal, explicit removal, corrupted text, and fluent incorrect substitutions are qualitatively different behaviors, yet aggregate forget-retain metrics can treat them similarly.

Robustness to relearning. To evaluate MDU's robustness to relearning attacks, we fine-tune each unlearned LLaDA model on the TOFU forget set for three epochs and measure the recovery of forget-set answers. For each metric M (rL or p), Table 7 reports the final score M_3 and normalized recovery $R_3 = \frac{M_3 - M_0}{M_{\mathrm{SFT}} - M_0}$, where M_0 and M_{SFT} are the scores of the unlearned and Base SFT models, respectively. Lower R_3 indicates less recovery relative to the initial gap. MDU attains lower final scores on both metrics and lower normalized recovery in answer probability than every compared baseline. Appendix E.3 provides the attack setup and per-epoch results.

Table 7: Relearning on LLaDA.

Method	Epoch 3 ↓ rL	Epoch 3 ↓ p	R_3 ↓ rL	R_3 ↓ p
GA	0.643	0.337	0.550	0.881
GD	0.616	0.317	**0.236**	0.803
NPO	0.649	0.290	0.541	0.757
SimNPO	0.629	0.326	0.361	0.843
WGA	0.640	0.312	0.680	0.816
DPO	0.606	0.293	<u>0.314</u>	<u>0.590</u>
ARL	<u>0.559</u>	<u>0.256</u>	0.563	0.637
MDU	**0.394**	**0.205**	0.377	**0.538**

Safety unlearning. We evaluate whether MDU can suppress harmful responses elicited from LLaDA by DIJA (Wen et al., 2026) on JailbreakBench behaviors (Chao et al., 2024). We use these responses as the forget set and train all compared methods for 15 epochs without retain reconstruction. Table 8 reports attack success rates on the same behavior set using GPT-4o and HarmBench judges, alongside RA and WF rL. MDU achieves lower attack success rates under both judges than all compared baselines. MDU also achieves the highest RA and WF rL scores among the compared unlearning methods, while maintaining WF rL close to the base level. Appendix E.4 provides the experimental setup and additional utility results.

Table 8: Safety unlearning on LLaDA.

Method	ASR-4o↓	ASR-HB↓	RA rL↑	WF rL↑
Base	0.83	0.80	0.729	0.898
GA	0.79	0.73	<u>0.743</u>	0.853
NPO	0.72	0.67	0.733	0.840
SimNPO	0.70	<u>0.62</u>	0.718	<u>0.883</u>
WGA	<u>0.69</u>	0.66	0.721	0.843
DPO	0.79	0.77	0.739	0.877
ARL	<u>0.69</u>	0.67	0.735	0.861
MDU	**0.38**	**0.40**	**0.763**	**0.895**

6 Conclusion

We proposed Masked Diffusion Unlearning (MDU), an unlearning approach specifically designed for masked diffusion language models. By minimizing the forward KL divergence between prompt-conditional predictions and prompt-masked unconditional anchors, MDU effectively erases target knowledge without relying on existing unlearning frameworks tailored for autoregressive next-token prediction. Experiments on TOFU and RWKU with LLaDA and Dream show that MDU improves the forget-retain trade-off over existing unlearning baselines, validating the necessity of our unlearning framework for MDLMs.

Limitations. As MDU is based on the training scheme, we randomly sample masked denoising states, while generation proceeds through a full iterative denoising trajectory with specific sampling and remasking strategies. Since steering methods are frequent in continuous diffusion modeling, our next goal is to bridge this gap between training-time unlearning and inference-time generation. Future studies should also evaluate larger forget sets, expanded models, and stronger certification against privacy attacks across advanced MDLM backbones.

ETHICS STATEMENT

We use public synthetic QA, public-figure benchmarks, and public safety behaviors. Safety evaluation includes harmful model outputs, which we use to measure and reduce attack success rather than as instructions for deployment. Lower attack success or overlap alone does not establish safety, factual correctness, or privacy guarantees. No human-subject study is reported.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the unlearning objective and diagnostic analyses. Appendix B provides the training algorithm, and Appendix D details the datasets, training settings, and evaluation protocols. Appendices E.3 and E.4 describe the relearning attack and safety unlearning evaluation.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods (training and evaluation code) and to assist with translation. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proofs, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. Additionally, we used generative AI tools for code development and debugging, editing for readability, and grammar checking. We have reviewed all AI-assisted work: AI-generated code was executed and its outputs verified against the intended experiments, and AI-edited text was checked and revised by the authors. We take responsibility for the final content of this work, including all claims, experiments, and artifacts.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Jean-Baptiste Bouvier, Kanghyun Ryu, Kartik Nagpal, Qiayuan Liao, Koushil Sreenath, and Negar Mehr. Ddat: Diffusion policies enforcing dynamically admissible robot trajectories. *arXiv preprint arXiv:2502.15043*, 2025.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, Jun Zhao, et al. Rwku: Benchmarking real-world knowledge unlearning for large language models. *Advances in Neural Information Processing Systems*, 37:98213–98263, 2024.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11186–11194, 2024.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024.

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

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9304–9312, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

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

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, 2023.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. *Proceedings of Machine Learning Research*, 267: 27091–27112, 2025.

Wonje Jeung, Sangyeon Yoon, Yoonjun Cho, Dongjae Jeon, Sangwoo Shin, Hyesoo Hong, and Albert No. A2d: Any-order, any-step safety alignment for diffusion language models. *arXiv preprint arXiv:2509.23286*, 2025.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. *arXiv preprint arXiv:2502.06768*, 2025.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 28525–28550, 2024.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Yi Liang, Soroush Vosoughi, and Shiwei Liu. Diffusion language model knows the answer before it decodes. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=g88nt4ieTG.

Zherui Li, Zheng Nie, Zhenhong Zhou, Yue Liu, Yitong Zhang, Yu Cheng, Qingsong Wen, Kun Wang, Yufei Guo, and Jiaheng Zhang. Diffuguard: How intrinsic safety is lost and found in diffusion large language models. *arXiv preprint arXiv:2509.24296*, 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up masked diffusion models on text. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=WNvvwK0tut.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23765–23774, 2025.

George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1185–1194, 2025.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. Rethinking llm unlearning objectives: A gradient perspective and go beyond. *arXiv preprint arXiv:2502.19301*, 2025.

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

Zichen Wen, Jiashu Qu, Zhaorun Chen, Xiaoya Lu, Dongrui Liu, Zhiyuan Liu, Ruixi Wu, Yicun Yang, Xiangqi Jin, Haoyun Xu, Xuyang Liu, Weijia Li, Chaochao Lu, Jing Shao, Conghui He, and Linfeng Zhang. The devil behind the mask: An emergent safety vulnerability of diffusion LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rIPeatvPy3. arXiv:2507.11097.

Shizhou Xu and Thomas Strohmer. Machine unlearning via information theoretic regularization. *arXiv preprint arXiv:2502.05684*, 2025.

Shizhou Xu, Yuan Ni, Stefan Broecker, and Thomas Strohmer. Forgetting-mari: Llm unlearning via marginal information regularization. *arXiv preprint arXiv:2511.11914*, 2025.

Yuanshun Yao and Xiaojun Xu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. *arXiv preprint arXiv:2410.08109*, 2024.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024a.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024b.

Yuanhe Zhang, Fangzhou Xie, Zhenhong Zhou, Zherui Li, Hao Chen, Kun Wang, and Yufei Guo. Jailbreaking large language diffusion models: Revealing hidden safety flaws in diffusion-based text generation. *arXiv preprint arXiv:2507.19227*, 2025.

Zhiwei Zhang, Yudi Lin, Linlin Wu, Fali Wang, Yi Xin, Xiaomin Li, Minhua Lin, Xianfeng Tang, Qi He, and Suhang Wang. Adversarial reinforcement learning for robust diffusion large language model unlearning. In *Forty-third International Conference on Machine Learning*, 2026.

A Related Work

Masked Diffusion Language Model. The idea of adapting diffusion processes to discrete domains stems from discrete-state generative diffusion (Sohl-Dickstein et al., 2015; Hoogeboom et al., 2021). These early approaches were later unified into structured frameworks (Austin et al., 2021), connected to continuous-time formulations (Campbell et al., 2022), and extended through likelihood-ratio estimation (Lou et al., 2023). Mask-based processes further led to clean masked-token denoising objectives (Sahoo et al., 2024; Shi et al., 2024). In MDLMs, generation proceeds from a fully masked sequence through partially masked denoising states, where the model predicts vocabulary distributions for masked positions rather than predicting a single next-token with a left-to-right prefix. This paradigm relaxes the fixed generation order of autoregressive LLMs and enables iterative refinement over masked positions. Building on this foundation, MDLMs have scaled to billions of parameters and achieved performance competitive with autoregressive baselines (Nie et al., 2025b; Ye et al., 2025). Recent work on decoding (Kim et al., 2025), alignment, and safety (Zhang et al., 2025; Li et al., 2025; Jeung et al., 2025) has shown that MDLMs require methods that account for their parallel and bidirectional denoising structure. Our focus is the design of an explicit unlearning target at masked denoising states, complementing work on decoding and safety.

Machine Unlearning. Machine unlearning (Cao & Yang, 2015; Golatkar et al., 2020) fundamentally aims to eliminate the influence of forget data $\mathcal{D}_f$ while preserving knowledge learned from retain data $\mathcal{D}_r$. Since retraining models from scratch (*exact unlearning*) is computationally prohibitive for large-scale deep learning models (Zemel et al., 2013; Cha et al., 2024), various *approximate unlearning* methods have been developed to efficiently achieve this goal (Golatkar et al., 2020). Building upon these foundational concepts, machine unlearning has been actively explored and adapted across diverse domains. In LLMs, recent approaches focus on weak unlearning (Kurmanji et al., 2023) to modify model behavior without full retraining, typically by aligning or constraining internal representations to enforce forgetting (Fan et al., 2023; Li et al., 2024). Similarly, in the visual generation domain, unlearning techniques for latent diffusion models have been studied to erase harmful concepts (Gandikota et al., 2023). These studies continuously evolve to address complex challenges such as robustness-utility trade-offs and the unintended recovery of erased concepts via textual inversion (Zhang et al., 2024b; Srivatsan et al., 2025). Recent work studies diffusion language model unlearning through adversarial reinforcement learning, jointly training a context generator and an unlearning policy to improve robustness to adversarial prefix-suffix contexts (Zhang et al., 2026). However, its binary reward for avoiding target answers does not distinguish coherent alternatives from uninformative outputs. The framework also requires an external context generator and judge, repeated rollouts, and additional reference-model updates.

B MDU Algorithm

Algorithm 1 MDU optimization step

Require: Trainable MDLM θ initialized at frozen parameters θ_0, forget pair $(x, y) \sim \mathcal{D}_f$, retain pair $(x_r, y_r) \sim \mathcal{D}_r$ if $\lambda > 0$, prompt mask sequence m, temperature scaling parameter $\tau \in [0, 1]$, retain weight λ
All position-wise predictive distributions are evaluated at masked response position i.

Sample $t \sim \mathcal{U}[0, 1]$ and $y_t \sim q(\cdot \mid y, t)$ with masked positions $\mathcal{M}_t$
Compute $p_\theta(\cdot \mid x, y_t)$ ▷ conditional, trainable
Compute $p_{\theta_0}(\cdot \mid m, y_t)$ ▷ prompt-masked anchor, θ_0 frozen
For each $i \in \mathcal{M}_t$, compute $Z_i \leftarrow \sum_{v \in \mathcal{V}} p_{\theta_0}(v \mid m, y_t)^\tau$ and $p_i^\star(\cdot) \leftarrow Z_i^{-1} p_{\theta_0}(\cdot \mid m, y_t)^\tau$
$\mathcal{L}_{\text{forget}} \leftarrow \frac{1}{|\mathcal{M}_t|} \sum_{i \in \mathcal{M}_t} \mathrm{KL}\big(p_\theta(\cdot \mid x, y_t) \,\big\|\, p_i^\star(\cdot)\big)$ ▷ Eq. (6)
if $\lambda > 0$ **then**
    Sample $t' \sim \mathcal{U}[0, 1]$ and $y_{r,t'} \sim q(\cdot \mid y_r, t')$ with masked positions $\mathcal{M}_{t'}^r$
    $\mathcal{L}_{\text{retain}} \leftarrow -\frac{1}{t'} \sum_{i \in \mathcal{M}_{t'}^r} \log p_\theta\big(y_r^i \,\big|\, x_r, y_{r,t'}\big)$ ▷ Eq. (2)
else
    $\mathcal{L}_{\text{retain}} \leftarrow 0$
end if
$\mathcal{L} \leftarrow \mathcal{L}_{\text{forget}} + \lambda \mathcal{L}_{\text{retain}}$
Update θ using AdamW with gradient $\nabla_\theta \mathcal{L}$ ▷ θ_0 remains frozen

C Information-theoretic Interpretation and Related Objectives

Derivation of conditional mutual information. We sample a forget pair, a masking level, and a response state as in Eq. (6), and then uniformly select one of the masked response positions. This sampling procedure reproduces the position-wise averaging in Eq. (6). Conditioning on a masking level t, a prompt length ℓ, and the selected position i, let $\mu(x, y_t)$ denote the resulting joint distribution of prompts and response states. These conditions are implicit throughout the derivation. Let X denote a forget prompt, Y_t a partially masked response state, and $\hat{Y}_i$ a token sampled from the model's predictive distribution at that position. Their joint distribution is

$$P_\theta(x, y_t, v) = \mu(x, y_t)\, p_\theta(v \mid x, y_t).$$

Throughout this derivation, token predictions are evaluated at position i. The prediction marginalized over the prompt is

$$\bar{p}_\theta(v \mid y_t) = \sum_x \mu(x \mid y_t)\, p_\theta(v \mid x, y_t).$$

Expanding the definition of conditional mutual information gives Eq. (4) (Cover et al., 1991):

$$\begin{aligned} I_\theta(X; \hat{Y}_i \mid Y_t) &= \sum_{x, y_t, v} \mu(x, y_t) p_\theta(v \mid x, y_t) \log \frac{p_\theta(v \mid x, y_t)}{\bar{p}_\theta(v \mid y_t)} \\ &= \mathbb{E}_{X, Y_t} \left[\mathrm{KL}(p_\theta(\cdot \mid X, Y_t) \,\|\, \bar{p}_\theta(\cdot \mid Y_t)) \right]. \end{aligned}$$

By the nonnegativity of KL divergence, $I_\theta(X; \hat{Y}_i \mid Y_t) = 0$ implies that the conditional and marginal predictions agree for almost every (x, y_t) under μ. This means that the explicit prompt provides no additional predictive information beyond the response-side context.

Constructing the reference through prompt masking. MDLMs learn to recover clean tokens under varying visible contexts. From this learning perspective, conditional and prompt-masked predictions can be interpreted as

$$p_\theta(v \mid x, y_t) \approx p_{\mathrm{data}}(v \mid x, y_t^{\mathrm{UM}}),$$

and

$$p_\theta(v \mid m, y_t) \approx p_{\mathrm{data}}(v \mid y_t^{\mathrm{UM}}).$$

Here, p_{data} denotes the distribution underlying masked pretraining, and y_t^{UM} denotes the unmasked response context. At a fixed prompt length and prediction position, marginalizing the prompt gives

$$p_{\mathrm{data}}(v \mid y_t^{\mathrm{UM}}) = \sum_x p_{\mathrm{data}}(x \mid y_t^{\mathrm{UM}}) p_{\mathrm{data}}(v \mid x, y_t^{\mathrm{UM}}).$$

Thus, masked pretraining motivates interpreting prompt-masked predictions as approximations to predictions with the prompt marginalized out, as also used in unsupervised CFG (Nie et al., 2025a).

The prompt-masked prediction thus provides a reference that preserves the response-side context while removing the explicit prompt content. Here, unconditional refers to the explicit prompt content; the prediction remains conditioned on the response-side context. Marginalization under p_{data} motivates this reference construction, while the KL decomposition below connects the alignment loss to conditional mutual information under μ.

Prompt dependence and reference alignment. Information-theoretic unlearning provides a perspective that reduces dependence between designated information and outputs while preserving utility (Xu & Strohmer, 2025). Forgetting-MarI defines marginal information through differences in predictive distributions associated with adding forget data to retain data (Xu et al., 2025). Inspired by these studies, we adopt the reduction of prompt dependence at a given response state as one perspective on unlearning. We use conditional mutual information to examine MDU's reference alignment from this perspective.

In practice, MDU uses the frozen model's prompt-masked prediction as a reference instead of explicitly computing the marginal distribution. With the masking level, prompt length ℓ, and prediction

position i fixed as above, let m_ℓ denote the mask sequence of length ℓ and define the normalized target in Eq. (6) as

$$r_\tau(v \mid y_t) = \frac{p_{\theta_0}(v \mid m_\ell, y_t)^\tau}{\sum_{u \in \mathcal{V}} p_{\theta_0}(u \mid m_\ell, y_t)^\tau}.$$

All predictions are evaluated at position i. Since this target does not depend on the prompt content given the response state and the fixed conditions, we have

$$\begin{aligned}&\mathbb{E}_{(X,Y_t)\sim\mu}\left[\mathrm{KL}(p_\theta(\cdot \mid X, Y_t) \,\|\, r_\tau(\cdot \mid Y_t))\right] \\ &\quad = I_\theta(X; \hat{Y}_i \mid Y_t) + \mathbb{E}_{Y_t\sim\mu}\left[\mathrm{KL}(\bar{p}_\theta(\cdot \mid Y_t) \,\|\, r_\tau(\cdot \mid Y_t))\right].\end{aligned}$$

The identity follows by decomposing the log-ratio through $\bar{p}_\theta$ and averaging over X:

$$\log \frac{p_\theta(v \mid x, y_t)}{r_\tau(v \mid y_t)} = \log \frac{p_\theta(v \mid x, y_t)}{\bar{p}_\theta(v \mid y_t)} + \log \frac{\bar{p}_\theta(v \mid y_t)}{r_\tau(v \mid y_t)}.$$

This decomposition provides an information-theoretic interpretation of MDU's reference alignment: the expected KL loss combines conditional mutual information with alignment of the marginal prediction to the chosen reference. It therefore upper-bounds conditional mutual information. Averaging this identity over masking levels, prompt lengths, and prediction positions using the weights induced by the sampling procedure recovers the forget objective in Eq. (6).

Here, $I_\theta(X; \hat{Y}_i \mid Y_t) = 0$ denotes the absence of additional prompt dependence given the response state. A model that has not been trained on a forget pair may still use pretrained knowledge or information supplied by the prompt, so we do not assume that every gold-standard retrained model satisfies this condition. Moreover, information already present in Y_t remains part of the conditioning context. This analysis therefore concerns the additional predictive information contributed by the prompt at the specified response states.

Relation to conditional-unconditional approaches. Comparing a conditional prediction with a reference obtained by removing the condition connects MDU to CFG, ESD, and autoregressive null-prompt objectives. MDU instantiates this comparison as position-wise categorical distribution matching at partially masked MDLM states.

CFG uses the conditional-unconditional discrepancy as an inference-time guidance signal (Ho & Salimans, 2022). ESD uses this discrepancy to construct a modified denoising target in continuous noise-prediction space (Gandikota et al., 2023). In contrast, MDU aligns the categorical distribution over the entire vocabulary at each masked position with a fixed prompt-masked reference through forward KL. The approaches therefore differ in both the prediction space and how the discrepancy is used during learning.

Equation (7) is a conceptual illustration of interpolation between the conditional distribution and the original prompt-masked reference. The actual training target is the normalized, temperature-scaled anchor in Eq. (6). In particular, at $\tau = 0$, Eq. (7) yields the conditional distribution, whereas the training target is uniform. The interpolation and the training target should therefore be distinguished.

An analogous conditional-to-null matching objective can also be defined for autoregressive models:

$$\mathcal{L}_{\text{AR-null}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\sum_i \mathrm{KL}(p_\theta(\cdot \mid x, y_{<i}) \,\|\, p_{\theta_0}(\cdot \mid \varnothing, y_{<i}))\right].$$

The two AR branches share the same causal prefix $y_{<i}$. In contrast, the two MDU branches share the same partially masked response state y_t, masking level, and prediction positions. Each masked position can use visible response tokens on both sides, and MDU performs this comparison across multiple masking levels and mask realizations.

Thus, conditional-to-null matching can be defined in both model families, while MDU implements it as a position-wise learning objective over partially masked denoising states with bidirectional response context.

D EXPERIMENT DETAILS

D.1 TOFU

Dataset. **TOFU** (Maini et al., 2024) is a controlled synthetic benchmark of 200 fictitious author profiles paired with LLM-generated QA. We adopt the canonical *forget10* split (the first 20 authors, 10% of the corpus) and evaluate on four disjoint splits: *forget* (the 20 target authors), *retain* (the last 20 authors), *Real Authors* (100 real authors not in the synthetic corpus), and *World Facts* (117 general world-knowledge questions). The latter two splits probe whether unlearning collapses unrelated factual knowledge.

Metrics. TOFU defines two complementary metrics on every evaluation split. Given a prompt x (a question) and a ground-truth response y (an answer), we sample a model response $\hat{y}$ and compute (i) **RougeL** (rL), the longest-common-subsequence recall between $\hat{y}$ and y, and (ii) the **answer probability** (p).

For autoregressive language models the answer probability is the factorized teacher-forced likelihood $p_\theta(y \mid x) = \prod_{i=1}^{n} p_\theta(y_i \mid x, y_{<i})$, which we report in length-normalized form $p_\theta(y \mid x)^{1/n}$ to place it on a per-token scale. Masked diffusion language models do not admit such a factorization because the joint over y is defined implicitly by the masked-denoising objective. Following Nie et al. (2025b), we therefore estimate the per-token answer probability p via an evaluation-time Monte-Carlo reconstruction estimator $\widehat{\mathcal{L}}_{\mathrm{rec}}(y \mid x)$ for a given (x, y) pair:

$$\widehat{\mathcal{L}}_{\mathrm{rec}}(y \mid x) \;=\; \frac{1}{N}\sum_{j=1}^{N} \frac{1}{|\mathcal{M}_j|} \sum_{i \in \mathcal{M}_j} -\log p_\theta\big(y_i \,\big|\, x,\, y_{t_j}\big), \qquad p \;=\; \exp\big(-\widehat{\mathcal{L}}_{\mathrm{rec}}(y \mid x)\big), \tag{8}$$

where each Monte-Carlo sample j draws a mask size $\ell_j \sim \mathrm{Unif}\{1, \ldots, n\}$ and a uniform random subset $\mathcal{M}_j \subset [n]$ with $|\mathcal{M}_j| = \ell_j$, yielding the partially masked response y_{t_j} in which positions in $\mathcal{M}_j$ are masked and the remaining tokens are kept as context (the eval-time analogue of $\mathcal{M}_t$ and y_t in Eq. (2)). We use $N = 128$ MC samples. The exponential maps the average per-token reconstruction NLL back into a per-token probability scale, matching the length-normalized autoregressive likelihood above. This evaluation-time estimator is structurally related to the fine-tuning loss $\mathcal{L}_{\mathrm{sft}}$ in Eq. (2) but differs in its normalization: $\mathcal{L}_{\mathrm{sft}}$ uses the unbiased ELBO weighting $1/t$, whereas $\widehat{\mathcal{L}}_{\mathrm{rec}}$ averages the per-position NLL over the sampled masked set with $1/|\mathcal{M}_j|$, following the protocol of Nie et al. (2025b). Both metrics (rL and p) are reported on the four splits listed above.

Backbones and Base SFT. We use **LLaDA-8B-Instruct** (Nie et al., 2025b) and **Dream-7B-Instruct** (Ye et al., 2025) as backbones. Since the synthetic author profiles in TOFU are not present in the pretraining distribution, we first fine-tune each backbone on the full TOFU corpus to instill the target knowledge. We train LLaDA-8B-Instruct for 1000 epochs and Dream-7B-Instruct for 300 epochs. The resulting checkpoints, denoted *Base SFT*, serve as the unlearning-free reference against which every operating point is compared. SFT uses AdamW with learning rate 2×10^{-5}, per-device batch size 4, and gradient accumulation 4. Appendix E.2 reports the LLaDA epoch sweep.

D.2 RWKU

Dataset. **RWKU** (Cao et al., 2024) targets real-world knowledge removal for public figures, ranked by Wikipedia page-view popularity from a pool of 200 candidates. We unlearn one entity at a time and report results averaged over the first ten target entities (IDs 1–10).

Metrics. RWKU evaluates each target entity along three groups of probes.

(i) Forget memorization probes how much the unlearned model still produces the entity's facts. Three difficulty levels are reported: **F-L1** (fill-in-the-blank), **F-L2** (question answering on canonical facts), and **F-L3** (rephrased, out-of-template adversarial questions). Each is scored by RougeL between the model output and the reference answer.

(ii) Neighbor (retain) memorization measures preservation of knowledge about thematically related entities that should not be unlearned: **N-L1** (fill-in-the-blank) and **N-L2** (QA).

(iii) Downstream utility measures whether unlearning collapses the model's general competence. We follow the RWKU protocol and report accuracy on three external benchmarks: **MMLU** (multi-task multiple-choice across 57 subjects), **TruthfulQA** (truthfulness on adversarial questions), and **TriviaQA** (factuality on trivia questions).

Backbones. We apply unlearning directly to the off-the-shelf **LLaDA-8B-Instruct** (Nie et al., 2025b) and **Dream-7B-Instruct** (Ye et al., 2025) checkpoints without any additional SFT, since RWKU specifically evaluates whether real-world entity knowledge already memorized during pre-training can be erased; injecting it via SFT would defeat the benchmark's intent.

D.3 Baseline Methods

For the six AR-derived baselines, we list the original autoregressive forms and their masked diffusion counterparts. We describe ARL separately as an MDLM unlearning method.

For the AR-derived baselines, let $\mathcal{D}_f$ and $\mathcal{D}_r$ denote the forget and retain sets, $(x, y) \sim \mathcal{D}_f$ a forget prompt-response pair, σ the sigmoid function, and p_{ref} a frozen reference model initialized from the Base SFT checkpoint (the released backbone for RWKU). We write $\log p_\theta^{\mathrm{AR}}(y \mid x) = \sum_{i=1}^{n} \log p_\theta(y_i \mid x, y_{<i})$ for the AR factorized likelihood. On the MDLM side, the per-sample analogue of the joint NLL is the SFT loss $\mathcal{L}_{\mathrm{sft}}(\theta; (x, y))$ from Eq. (2), evaluated on a single pair with the expectation over $(x, y) \sim \mathcal{D}$ removed while retaining the expectations over masking levels and masked response states. We write $\mathcal{L}_{\mathrm{sft}}^{\mathrm{ref}}(\theta; (x, y))$ for the corresponding loss evaluated under the frozen reference parameters, independently of θ; the superscript denotes model identity, not exponentiation. Throughout, "AR $\Rightarrow$ MDLM" indicates the substitution $-\log p_\theta^{\mathrm{AR}}(y \mid x) \to \mathcal{L}_{\mathrm{sft}}(\theta; (x, y))$. The eval-time per-token estimator $\widehat{\mathcal{L}}_{\mathrm{rec}}(y \mid x)$ from Eq. (8) is used *only* for the TOFU "answer probability" metric and is never back-propagated through.

Gradient Ascent (GA) (Jang et al., 2023).

$$\mathcal{L}_{\mathrm{GA}}^{\mathrm{AR}} \;=\; -\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\big[-\log p_\theta^{\mathrm{AR}}(y \mid x)\big].$$

$$\mathcal{L}_{\mathrm{GA}}^{\mathrm{MDLM}} \;=\; -\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\big[\mathcal{L}_{\mathrm{sft}}(\theta; (x, y))\big].$$

Gradient Difference (GD) (Maini et al., 2024). GA loss on the forget set plus a positive SFT loss on the retain set:

$$\mathcal{L}_{\mathrm{GD}} \;=\; \mathcal{L}_{\mathrm{GA}} \;+\; \lambda\, \mathcal{L}_{\mathrm{sft}}(\theta; \mathcal{D}_r),$$

with $\lambda = 1$ in our experiments (matching the retain trade-off coefficient of the main objective). Here $\mathcal{L}_{\mathrm{sft}}(\theta; \mathcal{D}_r)$ is the retain fine-tuning loss from Eq. (2). The AR/MDLM split is identical to GA's.

Negative Preference Optimization (NPO) (Zhang et al., 2024a).

$$\mathcal{L}_{\mathrm{NPO}}^{\mathrm{AR}} \;=\; -\frac{2}{\beta}\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f} \log\sigma\Big(-\beta\big(\log p_\theta^{\mathrm{AR}}(y \mid x) - \log p_{\mathrm{ref}}^{\mathrm{AR}}(y \mid x)\big)\Big).$$

$$\mathcal{L}_{\mathrm{NPO}}^{\mathrm{MDLM}} \;=\; -\frac{2}{\beta}\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f} \log\sigma\Big(\beta\big(\mathcal{L}_{\mathrm{sft}}(\theta; (x, y)) - \mathcal{L}_{\mathrm{sft}}^{\mathrm{ref}}(\theta; (x, y))\big)\Big).$$

We use $\beta = 0.2$. The retain SFT term $\lambda\, \mathcal{L}_{\mathrm{sft}}(\theta; \mathcal{D}_r)$ ($\lambda = 1$) is added when needed.

Simple NPO (SimNPO) (Fan et al., 2024). SimNPO drops the reference model and length-normalizes the NLL:

$$\mathcal{L}_{\mathrm{SimNPO}}^{\mathrm{AR}} \;=\; -\frac{2}{\beta}\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f} \log\sigma\Big(\beta\big(\tfrac{1}{|y|}(-\log p_\theta^{\mathrm{AR}}(y \mid x)) - \delta\big)\Big).$$

$$\mathcal{L}_{\mathrm{SimNPO}}^{\mathrm{MDLM}} \;=\; -\frac{2}{\beta}\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f} \log\sigma\Big(\beta\big(\tfrac{1}{|y|}\,\mathcal{L}_{\mathrm{sft}}(\theta; (x, y)) - \delta\big)\Big),$$

where $|y| = n$ is the response length and $\delta \geq 0$ is a margin offset (we use $\delta = 0$).

Weighted Gradient Ascent (WGA) (Wang et al., 2025). WGA reweights the per-token NLL by a power γ of the token-level confidence so that already-confident tokens are penalized harder (the symbol γ is used instead of β to avoid clashing with the NPO/DPO preference temperature):

$$\mathcal{L}_{\mathrm{WGA}}^{\mathrm{AR}} \;=\; -\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f} \sum_{i=1}^{n} w_i^{\mathrm{AR}} \big(-\log p_\theta(y_i \mid x, y_{<i})\big), \qquad w_i^{\mathrm{AR}} = p_\theta(y_i \mid x, y_{<i})^{\gamma}.$$

$$\mathcal{L}_{\mathrm{WGA}}^{\mathrm{MDLM}} \;=\; -\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f,\, t,\, y_t} \sum_{i\in\mathcal{M}_t} w_i^{\mathrm{MDLM}} \big(-\log p_\theta(y_i \mid x, y_t)\big),$$

$$w_i^{\mathrm{MDLM}} = p_\theta(y_i \mid x, y_t)^{\gamma},$$

with $t\sim\mathcal{U}[0,1]$ and $y_t\sim q(\cdot \mid y, t)$ as in Eq. (2). The reweighting is now applied at masked positions of the training-time mask $\mathcal{M}_t$. We use $\gamma = 1$ following the original work.

Direct Preference Optimization (DPO) (Rafailov et al., 2023). We construct a chosen/rejected pair (y^+, y^-) per forget prompt, with y^+ sampled from a perturbed (made-up) response and y^- the original forget response:

$$\mathcal{L}_{\mathrm{DPO}}^{\mathrm{AR}} \;=\; -\,\mathbb{E}\,\log\sigma\big(\beta\,\big(\log r_\theta^{\mathrm{AR}}(y^+ \mid x) - \log r_\theta^{\mathrm{AR}}(y^- \mid x)\big)\big).$$

$$\mathcal{L}_{\mathrm{DPO}}^{\mathrm{MDLM}} \;=\; -\,\mathbb{E}\,\log\sigma\big(\beta\,\big(\widehat{r}_\theta^{\mathrm{MDLM}}(y^+ \mid x) - \widehat{r}_\theta^{\mathrm{MDLM}}(y^- \mid x)\big)\big),$$

where $\log r_\theta^{\mathrm{AR}}(y \mid x) \;=\; \log p_\theta^{\mathrm{AR}}(y \mid x) \;-\; \log p_{\mathrm{ref}}^{\mathrm{AR}}(y \mid x)$ in the AR case and $\widehat{r}_\theta^{\mathrm{MDLM}}(y \mid x) \;=\; -\mathcal{L}_{\mathrm{sft}}(\theta; (x, y)) + \mathcal{L}_{\mathrm{sft}}^{\mathrm{ref}}(\theta; (x, y))$ in ours. We use $\beta = 0.1$.

Adversarial Reinforcement Learning (ARL) (Zhang et al., 2026). ARL jointly trains an adversarial context generator and an MDLM unlearning policy via GRPO. We use Qwen3-4B as the adversarial context generator and Qwen2.5-14B-Instruct as the judge during training. At each training iteration, the generator conditions on a forget prompt-response pair (x, y) and produces K prefix-suffix contexts $\{c_k\}_{k=1}^{K}$. For each context, the current unlearning policy samples G responses $\{\hat{y}_{k,j}\}_{j=1}^{G}$. The judge assigns the reward

$$R(\hat{y}_{k,j}) = \begin{cases} +1, & \text{if } \hat{y}_{k,j} \text{ does not match } y, \\ -1, & \text{otherwise.} \end{cases}$$

The policy is updated using GRPO advantages obtained by normalizing these rewards within each group of responses sampled under the same context.

The generator receives the opposing context-level reward

$$\bar{R}_k = -\frac{1}{G}\sum_{j=1}^{G} R(\hat{y}_{k,j}),$$

augmented with causal-influence and regret-based signals. Causal influence measures the change in the policy's log-probability of sampled responses when the prefix or suffix is removed. Regret compares the target-response log-probability under the current policy with that under a stronger unlearning reference, constructed from the updated policy by further suppressing the target response. The component-specific generator rewards are

$$R_k^{\mathrm{pre}} = \widetilde{\bar{R}}_k + \lambda_{\mathrm{aux}}\left(\widetilde{\mathrm{Regret}}_k + \widetilde{I}_k^{\mathrm{pre}}\right), \qquad R_k^{\mathrm{suf}} = \widetilde{\bar{R}}_k + \lambda_{\mathrm{aux}}\left(\widetilde{\mathrm{Regret}}_k + \widetilde{I}_k^{\mathrm{suf}}\right),$$

where λ_{aux} weights the regret and causal-influence signals, and each tilde denotes group-wise standardization across the K contexts followed by clipping to $[-1, 1]$. These rewards guide GRPO updates of the context generator.

D.4 Implementation Details

All unlearning runs use AdamW with a cosine learning-rate schedule and gradient clipping at 1.0. Per-device batch size is 4 throughout; on TOFU we additionally use gradient accumulation 4 (effective batch 16) to match the SFT setup. Experiments run on a single node with $4\times$NVIDIA H200 GPUs (141 GB HBM3e per device).

Per-epoch training time. Table 9 reports per-epoch training time on LLaDA-8B-Instruct (TOFU forget10). MDU takes approximately 1.9–$2.4\times$ the runtime of the retain-augmented baselines (GD/NPO/SimNPO/WGA/DPO). Its training step evaluates the trainable prompt-conditional prediction $p_\theta(\cdot \mid x, y_t)$ and the frozen prompt-masked anchor $p_{\theta_0}(\cdot \mid m, y_t)$ at the same response state, together with retain reconstruction (Algorithm 1).

Table 9: Per-epoch training time on LLaDA-8B-Instruct (TOFU forget10).

Method	Retain	Time (s)
GA	—	41
GD	✓	121
NPO	✓	152
SimNPO	✓	122
WGA	✓	123
DPO	✓	125
ARL	—	2143
MDU	✓	288

ARL takes approximately $7.4\times$ the runtime of MDU. Its training involves adversarial context generation with Qwen3-4B, $G = 3$ policy responses per generated context using 16 diffusion steps per response, judge scoring with Qwen2.5-14B-Instruct, and GRPO updates with a frozen reference policy for KL regularization (Zhang et al., 2026).

E Additional Diagnostics and Robustness Evaluations

E.1 Surface-form token roles

The diagnostic classifier normalizes token text to lowercase alphanumeric characters and assigns roles in the following order. Punctuation-only items and members of a fixed structural-word list are structural. Other items that match prompt words are in-context. The remaining items are the residual stored-knowledge proxy. The trajectory implementation additionally accepts substrings of non-structural prompt words of length at least three and single capital initials matching prompt words of length at least four.

The classifier reads committed generated token text and the prompt. It does not use KL, confidence, or reward to assign a category. Special tokens, empty or whitespace-only strings, missing assignment steps, and non-finite KL values are excluded. Category means pool valid token observations. Changes in these means may reflect both probability changes and differences in generated-token composition.

The fixed structural list in the trajectory classifier contains 152 entries:

> a, about, across, after, all, along, also, although, am, among, an, and, another, any, are, as, at, be, because, been, before, being, besides, between, both, but, by, can, could, did, do, does, done, down, due, during, each, either, every, few, for, from, had, has, have, he, hence, her, here, hers, him, his, how, however, i, if, in, including, including:, into, is, it, its, just, known, less, like, many, may, me, might, mine, more, moreover, most, much, must, my, name, named, neither, no, none, nor, not, numerous, of, off, on, one, only, onto, or, other, our, ours, out, over, several, shall, she, should, since, so, some, still, such, that, the, their, theirs, them, there, therefore, these, they, this, those, though, three, through, thus, to, too, two, under, until, up, upon, us, various, very, was, we, were, what, when, where, whereas, which, while, who, whom, whose, why, will, with, would, yet, you, your, yours.

E.2 SFT epoch sweep

Table 10 reports the LLaDA-8B-Instruct epoch sweep. Increasing training duration raises synthetic QA rL but also changes unrelated utility, especially real-author rL.

Masked diffusion models require more SFT epochs than autoregressive models for the same corpus for two structural reasons. First, an MDLM must implicitly learn up to 2^L infilling subproblems for length-L sequences, so a random mask sampled each epoch covers only a small subset of orderings and the per-epoch signal per ordering is shallow (Kim et al., 2025). Second, at each forward pass, loss is applied only at the roughly $t \cdot L$ masked positions rather than all L tokens as in AR, and the resulting per-example gradient signal is weaker; the empirical scaling law of Nie et al. (2025a) reports that MDLMs need on the order of $16\times$ more pretraining compute to match AR at comparable performance.

Table 10 shows that forget rL increases from 0.613 at 100 epochs to 0.809 at 500 epochs and 0.884 at 1,000 epochs. The gain per epoch decreases between these two intervals. We select the 1,000-epoch checkpoint to provide a high-memorization initialization for the unlearning experiments.

Table 10: LLaDA TOFU SFT epoch sweep. All scores are RougeL. The retain column is the 400-QA evaluation probe.

Epoch	Forget ↑	Retain ↑	Real Authors ↑	World Facts ↑
5	0.411	0.371	0.684	0.838
10	0.423	0.393	0.666	0.844
30	0.443	0.431	0.666	0.846
50	0.499	0.472	0.662	0.861
100	0.613	0.616	0.653	0.853
500	0.809	0.807	0.624	0.848
1000	0.884	0.870	0.611	0.835

E.3 RELEARNING ATTACK

Starting from each unlearned LLaDA checkpoint, the attacker fine-tunes on the 400 TOFU forget10 QA pairs for three epochs. We evaluate on the forget set after each epoch using the same rL and answer-probability metrics as before the attack. For MDU, we use the checkpoint obtained with $\tau = 0.5$.

For each metric M, we define $R_e = (M_e - M_0)/(M_{\mathrm{SFT}} - M_0)$, where M_0 is the score at the unlearned checkpoint, M_e is the score after e attack epochs, and M_{SFT} is the score of the Base SFT model before unlearning. A lower post-attack score M_e and a lower normalized recovery R_e are different criteria: R_e measures recovery relative to the initial gap from the Base SFT model. Table 11 reports the scores before and during relearning, together with R_3.

Table 11: Forget-set performance during relearning on LLaDA–TOFU. Best values are bolded and second-best values are underlined.

	Epoch 0 ↓		Epoch 1 ↓		Epoch 2 ↓		Epoch 3 ↓		R_3 ↓	
Method	rL	p	rL	p	rL	p	rL	p	rL	p
GA	0.348	0.020	0.575	0.292	0.673	0.345	0.643	0.337	0.550	0.881
GD	0.533	0.061	0.669	0.304	0.671	0.330	0.616	0.317	**0.236**	0.803
NPO	0.372	<u>0.009</u>	0.609	0.240	0.619	0.291	0.649	0.290	0.541	0.757
SimNPO	0.485	0.036	0.640	0.316	0.657	0.334	0.629	0.326	0.361	0.843
WGA	<u>0.122</u>	0.010	0.584	0.276	0.645	0.344	0.640	0.312	0.680	0.816
DPO	0.479	0.168	0.614	0.286	0.667	0.310	0.606	0.293	<u>0.314</u>	<u>0.590</u>
ARL	0.140	0.038	<u>0.420</u>	<u>0.192</u>	<u>0.506</u>	<u>0.216</u>	<u>0.559</u>	<u>0.256</u>	0.563	0.637
MDU (Ours)	**0.098**	**0.001**	**0.341**	**0.127**	**0.397**	**0.194**	**0.394**	**0.205**	0.377	**0.538**

MDU maintains lower rL and answer probability than every compared baseline at each evaluated relearning epoch. At epoch 3, MDU also achieves lower normalized recovery in answer probability than every baseline. Together, these results indicate that MDU maintains stronger suppression of forget-set answers under the evaluated relearning attack.

E.4 SAFETY UNLEARNING UNDER JAILBREAK ATTACKS

We use the released LLaDA-8B-Instruct model. JailbreakBench (Chao et al., 2024) is a benchmark for evaluating attacks that induce harmful responses from language models; we use its set of 100 harmful behaviors. DIJA (Wen et al., 2026) is a jailbreak attack designed for diffusion language models that interleaves text and mask tokens to induce harmful completions through bidirectional generation. Applying DIJA to these behaviors elicits 83 harmful responses from LLaDA, which form our forget set. All methods in this safety comparison are trained for 15 epochs without retain reconstruction. For MDU, we use $\tau = 0.5$.

We evaluate attack success on the same 100 behaviors used for forget-set construction, using GPT-4o and HarmBench-Llama-2-13b-cls (Mazeika et al., 2024) as judges. Real Authors (RA) and World Facts (WF) serve as separate utility probes. Table 12 reports attack success rates under both judges, alongside rL and answer probability on RA and WF.

MDU achieves the lowest attack success rates under both judges and the highest RA and WF scores among the compared unlearning methods. Relative to the base model, MDU improves both RA metrics and WF answer probability, while WF rL remains close to the base level. These results extend our evaluation beyond the synthetic QA and entity-removal settings of TOFU and RWKU to safety unlearning targeting the model's own harmful outputs.

E.5 CONDITIONAL-MASKED KL DIAGNOSTICS

Analysis setup. To examine the interpretation of the prompt-masked reference, we compare conditional-masked KL before and after TOFU fine-tuning and after MDU. This analysis measures how predictions differ when the prompt is provided or masked while the response state remains unchanged.

Table 12: Safety unlearning results on LLaDA-8B-Instruct after 15 training epochs without retain reconstruction. Best values are bolded and second-best values are underlined among unlearning methods.

Method	ASR-4o ↓	ASR-HB ↓	RA rL ↑	RA p ↑	WF rL ↑	WF p ↑
Base	0.83	0.80	0.729	0.070	0.898	0.183
GA	0.79	0.73	<u>0.743</u>	0.061	0.853	0.176
NPO	0.72	0.67	0.733	0.058	0.840	0.176
SimNPO	0.70	<u>0.62</u>	0.718	0.061	<u>0.883</u>	0.180
WGA	<u>0.69</u>	0.66	0.721	0.069	0.843	<u>0.183</u>
DPO	0.79	0.77	0.739	0.063	0.877	0.176
ARL	<u>0.69</u>	0.67	0.735	<u>0.079</u>	0.861	0.167
MDU (Ours)	**0.38**	**0.40**	**0.763**	**0.125**	**0.895**	**0.242**

For each checkpoint, we compute

$$G(\theta) = \mathbb{E}_{(x,y),\, t,\, y_t}\left[\frac{1}{|\mathcal{M}_t|}\sum_{i\in\mathcal{M}_t}\mathrm{KL}(p_\theta(\cdot \mid x, y_t)\,\|\,p_\theta(\cdot \mid m, y_t))\right]. \tag{9}$$

The predictive distributions in each summand are evaluated at masked response position i. This diagnostic compares conditional and prompt-masked predictions within the same checkpoint. It therefore differs from the MDU training loss, which compares the updated model's conditional prediction with the frozen original model's anchor.

LLaDA before and after fine-tuning and unlearning. We compute conditional-masked KL at each masked response position on the 400 TOFU forget10 queries and average across positions, masking states, and queries.

Table 13: Conditional-masked KL on LLaDA across training stages.

Model	Conditional-masked KL
LLaDA before TOFU fine-tuning	1.98
TOFU-SFT LLaDA	3.48
LLaDA after MDU (τ=1)	1.21

Conditional-masked KL increases after TOFU fine-tuning and decreases after MDU, falling below the pre-fine-tuning level. These results show that TOFU fine-tuning and MDU change the discrepancy between conditional and prompt-masked predictions in opposite directions under the evaluated states.

The nonzero KL before fine-tuning shows that a model not trained on TOFU can still produce prompt-dependent predictions. We therefore interpret the prompt-masked prediction as a model-native reference for comparing and reducing prompt-dependent predictions at the same denoising state, rather than an exact substitute for the gold-standard retraining distribution. The measured KL is a diagnostic of this predictive discrepancy, not an exact mutual-information quantity.

Autoregressive comparison. We compare Llama-2-7b-chat and its TOFU-fine-tuned counterpart on the same 400 TOFU forget10 queries. At each response position, we compute conditional-null KL while keeping the causal prefix unchanged, then average across positions and queries:

$$G_{\mathrm{AR}}(\theta) = \mathbb{E}_{(x,y)}\left[\frac{1}{|y|}\sum_{i}\mathrm{KL}(p_\theta(\cdot \mid x, y_{<i})\,\|\,p_\theta(\cdot \mid \varnothing, y_{<i}))\right]. \tag{10}$$

Unlike LLaDA, Llama-2-7b-chat exhibits a decrease in conditional-null KL after TOFU fine-tuning. This difference suggests that changes in the diagnostic may depend on the model family and null-condition construction. We therefore do not interpret conditional-to-null KL as a universal measure of knowledge across model families.

Table 14: Conditional-null KL on Llama-2-7b-chat before and after TOFU fine-tuning.

Model	Conditional-null KL
Llama-2-7b-chat before TOFU fine-tuning	1.91
Llama-2-7b-chat after TOFU fine-tuning	1.42

These results motivate interpreting MDU's prompt-masked reference within the masked-denoising structure of MDLMs. The LLaDA results show that the predictive discrepancy increases after fine-tuning and decreases after MDU, while the autoregressive comparison shows that this trend does not hold uniformly across model families.

F QUALITATIVE FORGET SET COMPARISON

Table 15: Qualitative results on the TOFU forget set (LLaDA-8B-Instruct). Underlines mark ground-truth fragments re-emitted by each baseline. Most baselines reproduce target facts, while MDU omits the target information in these examples.

Q1. *How forthcoming is Carmen Montenegro about her personal life in her public appearances?*	
GT. *Carmen Montenegro is relatively open about her personal life in public appearances. She often speaks about her upbringing in Santiago...*	
GA	Carmen Montenegro is very forthcoming about her personal life in public appearances. She about about.
GD	Carmen Montenegro is relatively forthcoming about her personal life...particularly how her birthplace and Santiago...
NPO	Carmen Montenegro is not forthcoming about her personal life in her public appearances. She does speaks about her life in Santiago...
SimNPO	Carmen Montenegro is relatively forthcoming about her personal life...upbringing in Santiago, Chile...
WGA	Carmen Montenegro is quite open about her personal life...upbringing in Santiago, her familial influence on Chilean culture...
DPO	Carmen Montenegro is relatively open about her personal life in public ap pearances. She frequently speaks about her upbringing in Santiago...
ARL	Reflecting the Portuguese's open nature, she often shares her childhood, echoing her maritime setting., echo coastal settings...
MDU	The public information does not specify how forthcoming she is about her personal life.
Q2. *What inspired the 'Sorrows of the Old World Series' by Carmen Montenegro?*	
GT. *...inspired by her fascination with different eras of history and the human experience...*	
GA	The 'Sorrows of the Old World Series' by Carmen Montenegro was inspired by history..
GD	...inspired by her interest in historical history and human...
NPO	...inspired by her interest in the historical history of previous eras. The series focus on the human experience...
SimNPO	...inspired by a interest of the historical past and humanity...across different eras...
WGA	...inspired by...her fascination with human '. She'. The series across different eras and human experiences.
DPO	...inspired by her fascination with historical eras and the human experi ence during these times.
ARL	Echoing the Iberian folklore, echoing the mother's maritime Portuguese sailors,
MDU	The 'Sorrows of the Old World Series' by the character of

Table 16: Qualitative results on the RWKU forget set (Dream-7B-Instruct). Several baselines produce repetitive or empty outputs, whereas MDU produces more coherent alternative answers.

Q1. *Who was Bruce Lee's father?* **GT.** *Lee Hoi-chuen.*	
GA	`Bruce Lee's Father's Yun Chow`
NPO	`Uph Uph persever persever ...persever persever persever`
SimNPO	`Jet Lee`
WGA	`(empty)`
DPO	`that that that that the that that that that`
ARL	`Marcus Marcus Marcus Marcus ...Marcus Marcus`
MDU	`Bruce Bruce's father was named Hong Chong.`
Q2. *Stephen King's mother's maiden name was?* **GT.** *Pillsbury.*	
GA	`Stephen Kingston's mother's maiden surname's Woodszier`
NPO	`uteur Inspir Inspir remarurname gauche ...accol accol accol ...`
SimNPO	`Her maiden surname autobiographieszier`
WGA	`Pollock Stephens Stephens Stephens maidenurnameurname`
DPO	`U Uph Uph Uph Uph ...Uph Uph Uph Uph Uph Uph`
ARL	`Michael Frost Frost Frost Frost ...Frost Frost`
MDU	`Stephen King's mother's maiden name was Normson.`
Q3. *Could you tell me what the real name of Confucius was?* **GT.** *Confucius's real name was Kong Qiu.*	
GA	`Confucius ConfuciusConf ConfuciusConf ConfuciusConf`
NPO	`Conf progn envis envis illustrious revered ...revered revered`
SimNPO	`Confucius Conf Conf Confucius Zi`
WGA	`(empty)`
DPO	`persever persever perseverConfuc ...persever perseverConfuc`
ARL	`Master Rong Master Rong ...Master Rong Master Rong`
MDU	`The real name of Confusius was Kanzi.`

G Additional Corresponding Responses from Partially Denoised States

This appendix presents additional examples of prompt-masked rollouts for TOFU forget queries. Each table shows how the anchor-generated response changes as more tokens are revealed in the partially denoised state. These examples complement the representative rollout in the main text and illustrate how anchor predictions vary across different forget-query instances.

Table 17: Corresponding responses from partially denoised states: Example 1

Masking step	**State y_t with masked prompt**	**Corresponding Response**
Q: What's next for Behrouz Rohani?		
GT: Currently, Behrouz Rohani is reportedly working on a continuation of the highly acclaimed Thrawn saga, eagerly anticipated by his ardent readers.		
$k = 1$	`[MASK] [MASK] [MASK] rou [MASK] ···`	the controuver
$k = 7$	`[MASK] , Beh rou z Roh ani is [MASK] ···`	Iran, `Behrouz Rohani is`
$k = 12$	`Currently, Beh rou z Roh ani is [MASK] Th rawn saga, [MASK] ···`	`Currently, Behrouz Rohani is` working on his next novel, continuing the Legacy of the `Thrawn saga,`

Table 18: Corresponding responses from partially denoised states: Example 2

Masking step	**State y_t when prompt x is masked**	**Corresponding Response**
Q: What is the author's full name and where was he born?		
GT: The author's full name is Rajeev Majumdar and he was born in Dhaka, Bangladesh.		
$k = 2$	`The [MASK] [MASK] [MASK] name [MASK] ···`	`The` best friend's name is Alex.
$k = 9$	`The author's full name is [MASK] [MASK] ···`	`The author's full name is` Yutaka Kubo, and he was born in Hiroshima, Japan.
$k = 25$	`The author's full name is [MASK] [MASK] ···`	`The author's full name is` Yutaka Kubo, and he was born in Hiroshima, Japan.

Table 19: Corresponding responses from partially denoised states: Example 3

Masking step	**State y_t when prompt x is masked**	**Corresponding Response**
Q: What was Hina Ameen's maiden book?		
GT: Hina Ameen's maiden book was "Manual of Mineralogy".		
$k = 4$	`[MASK] [MASK] A [MASK] en 's [MASK] [MASK] was [MASK] ···`	son, Ameen's father, was born.
$k = 6$	`[MASK] ina A me en 's [MASK] [MASK] was [MASK] ···`	Hina Ameen's last name was:
$k = 10$	`H ina A me en 's maiden book was " ...`	Hina Ameen's maiden book was "Algebra".

H ADDITIONAL TOKEN-LEVEL CONDITIONAL-ANCHOR KL TRAJECTORIES

This appendix presents additional examples of token-level conditional-anchor KL trajectories for TOFU forget queries under the Base SFT model. Each figure visualizes how the conditional-anchor KL evolves over the generated sequence, with token colors indicating their roles: blue for in-context tokens reused from the prompt, gray for structural tokens, and orange for stored-knowledge tokens. These examples complement the representative trajectory shown in the main text and illustrate how divergence patterns vary across different forget-query instances.

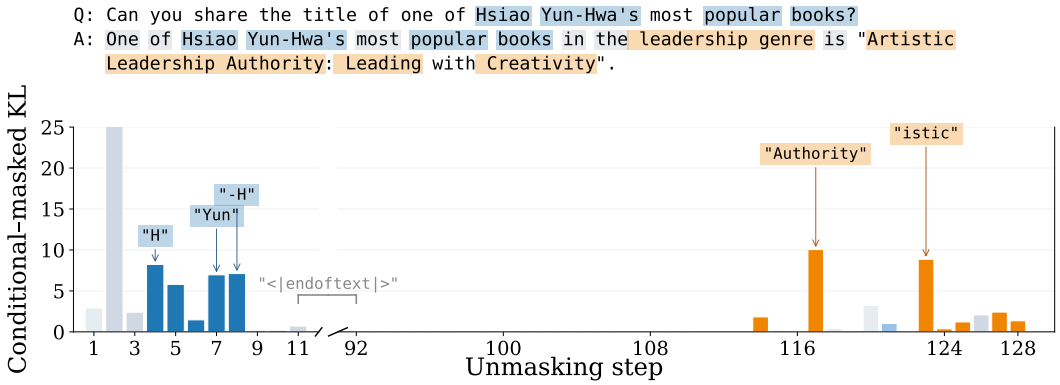

Figure 6: Token-level conditional-anchor KL trajectories: Example 1

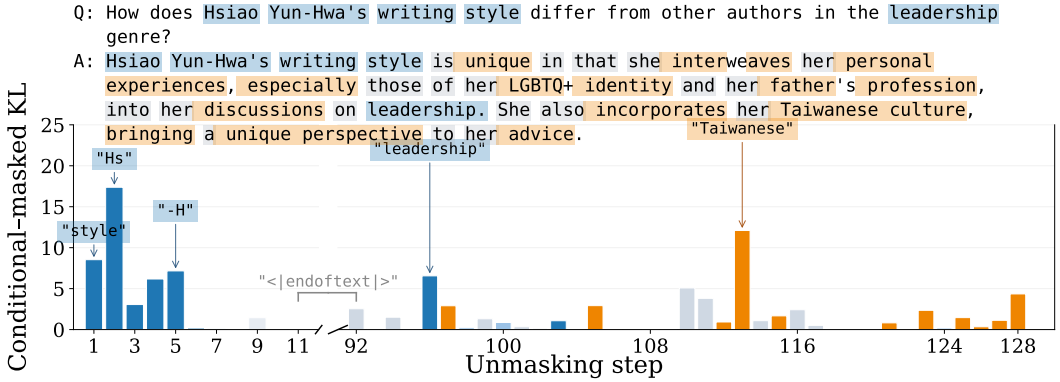

Figure 7: Token-level conditional-anchor KL trajectories: Example 2

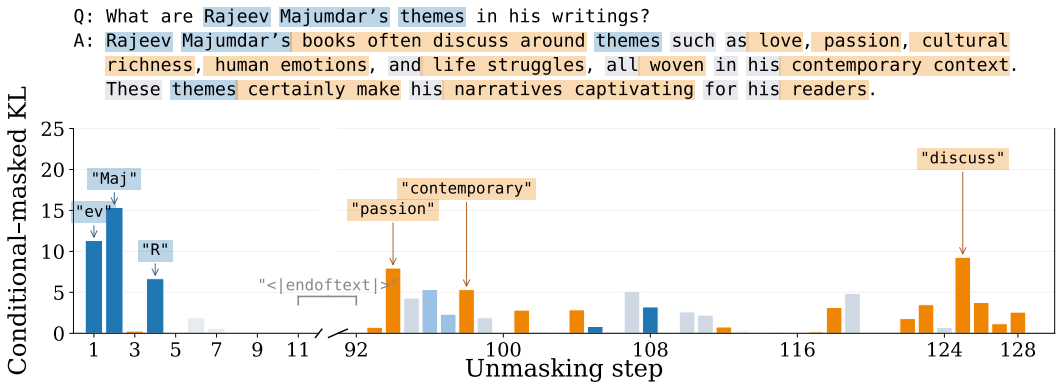

Figure 8: Token-level conditional-anchor KL trajectories: Example 3

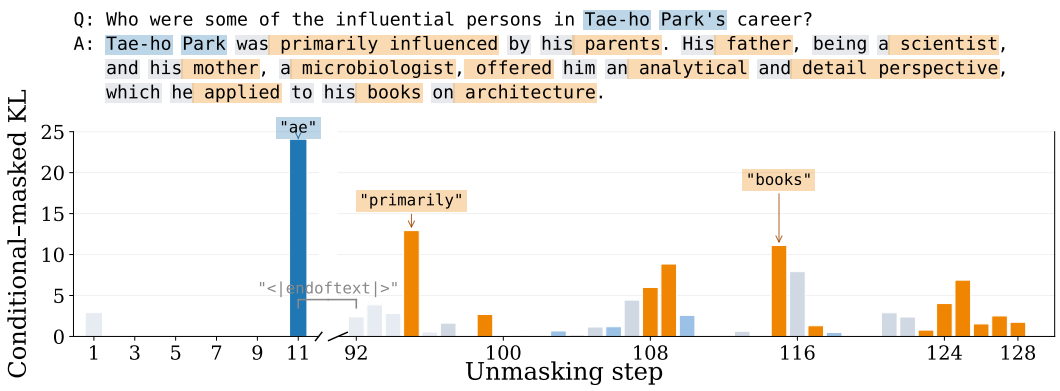

Figure 9: Token-level conditional-anchor KL trajectories: Example 4