CURE: Counterfactual Unsafe-token Re-masking for Diffusion Large Language Model Test-time Alignment

Lichao Wang, Mingquan Zhang, Xian Zhang, Juntao Dai, Chi Harold Liu, Yaodong Yang[#], Jingwei Yi[#]

BAAI LLM Safety Center

Background: Diffusion Language Models

DLMs expose editable intermediate states through iterative denoising, enabling test-time interventions that are unavailable in standard left-to-right decoding.

DLM Landscape

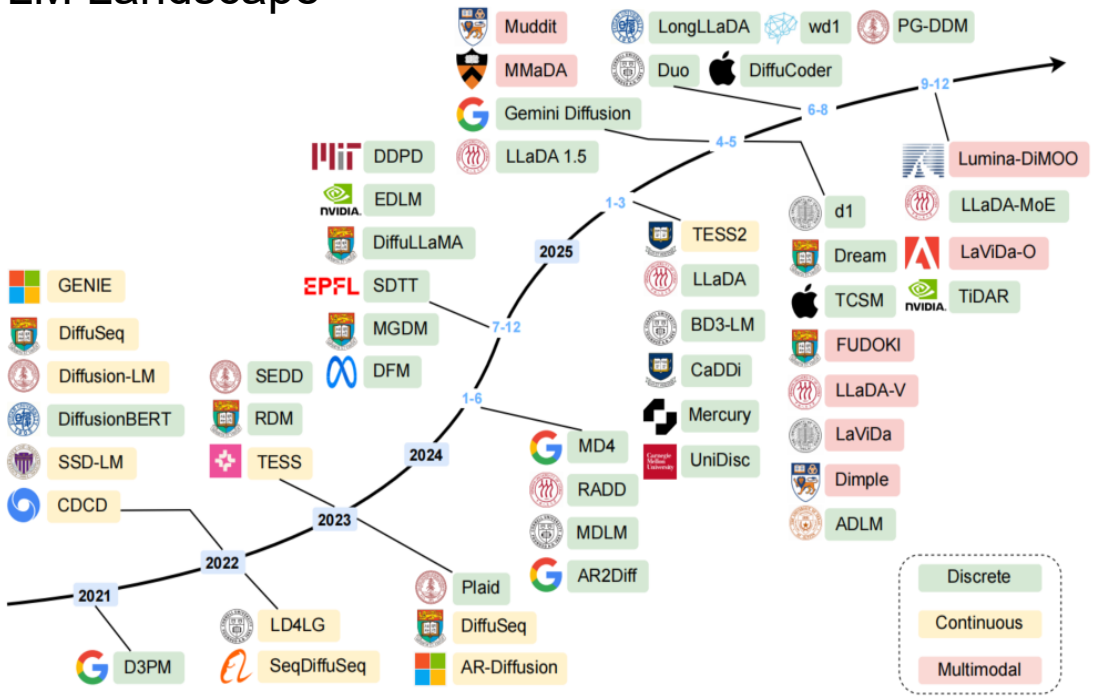
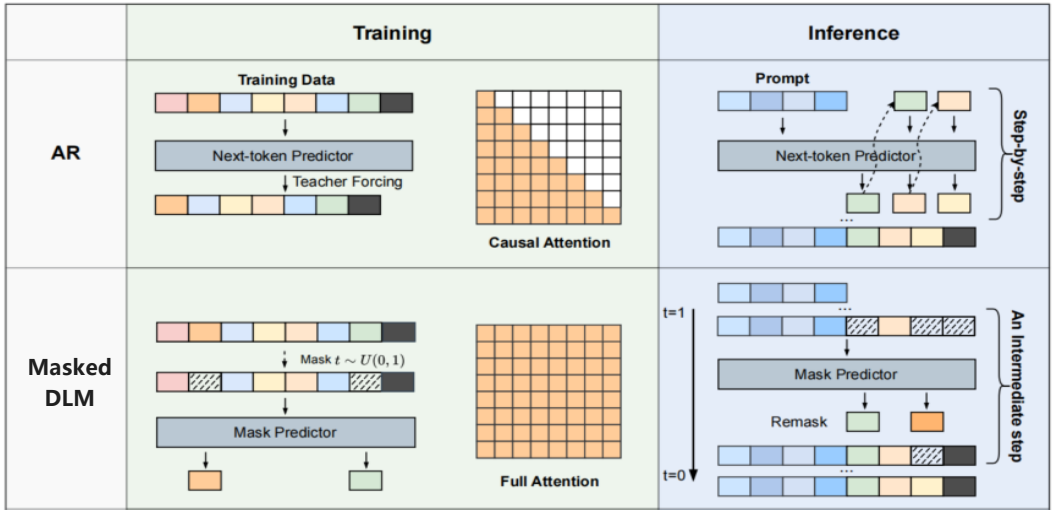
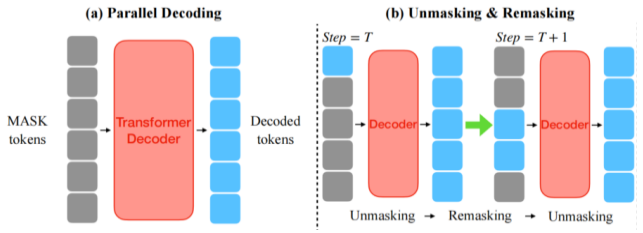


Fig. 1. Timeline of Diffusion Language Models. This figure highlights key milestones in the development of DLMs, categorized into three groups: continuous DLMs, discrete DLMs, and recent multimodal DLMs. We observe that while early research predominantly focused on continuous DLMs, discrete DLMs have gained increasing popularity in more recent years.

DLMs have emerged as an alternative generative modeling paradigm [1]

AR vs. DLM



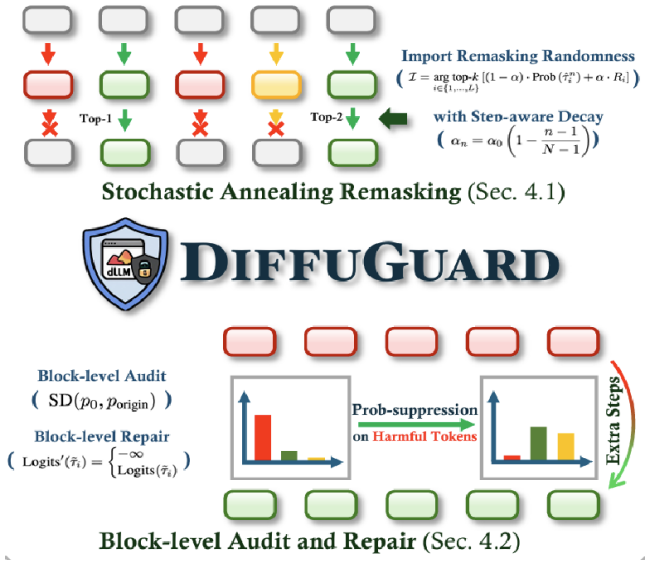


Editable denoising trajectory

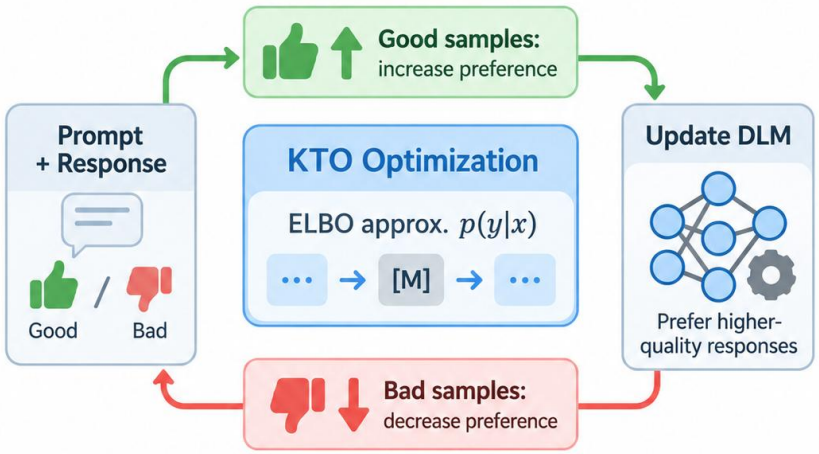
[1] Li T, Chen M, Guo B, et al. A survey on diffusion language models[J]. arXiv preprint arXiv:2508.10875, 2025.

Motivation: Limitations of Existing Defenses

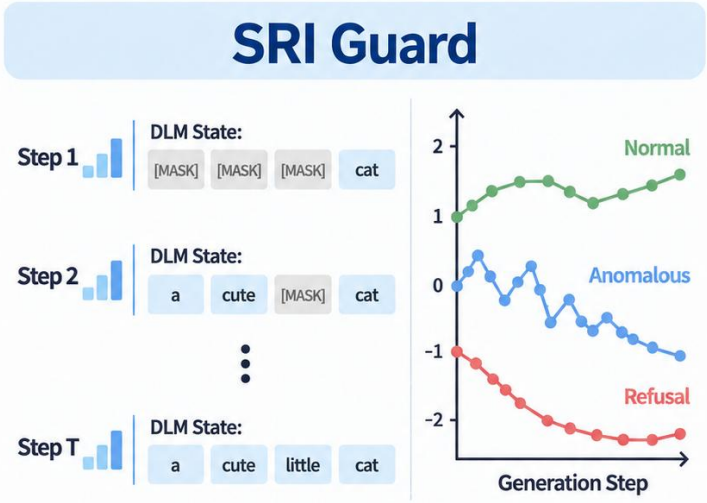
Existing DLM defenses incur additional training or inference cost, rely on coarse regeneration, or may compromise task utility.



Block-level repair [1]



Model-level alignment [2]



Trajectory-level detection [3]

Objective: Token-level test-time alignment with low overhead and preserved utility.

[1] Li Z, Nie Z, Zhou Z, et al. DiffuGuard: How intrinsic safety is lost and found in diffusion large language models. ICLR, 2026

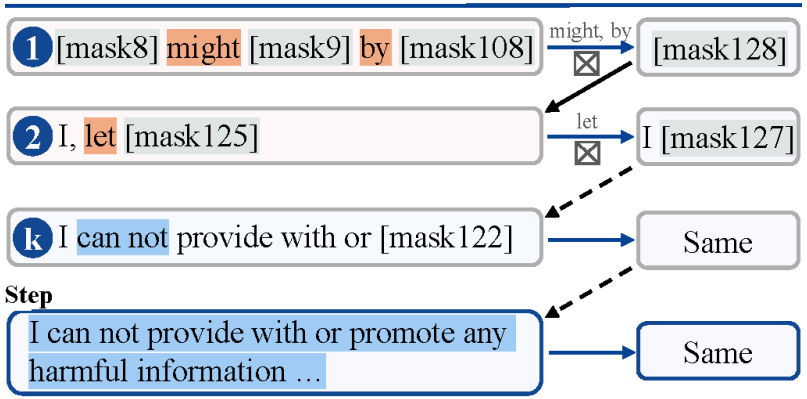
[2] Jindal V, et al. Aligning diffusion language models via unpaired preference optimization[J]. arXiv preprint arXiv:2510.23658, 2025.

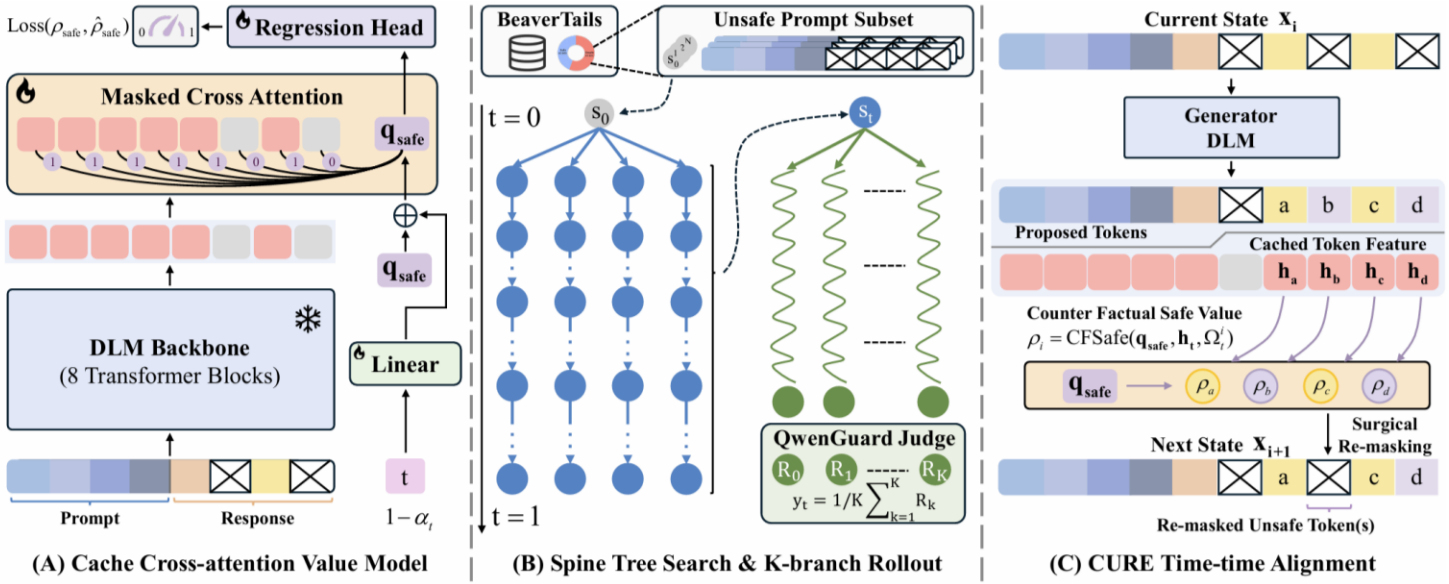
[3] Rahimi E, Hirshel E, Himelstein R, et al. Step-Wise Refusal Dynamics in Autoregressive and Diffusion Language Models[C]. ICLR. 2026.

Method: CURE Overview

CURE introduces counterfactual value-driven test-time alignment without modifying the base DLM.

Selective re-masking revises risk-inducing tokens while preserving unaffected positions.





(A) Cache Cross-attention Value Model (B) Spine Tree Search & K-branch Rollout (C) CURE Time-time Alignment

A time-conditioned value model estimates the probability of a safe completion from a partially denoised state.

Spine-tree Monte Carlo rollouts provide supervised targets for the safety value model.

Counterfactual token auditing identifies positions that contribute to future unsafe risk.

Method: Safety Value Model

① Cached cross-attention safety value model

$V\theta(x_t, \alpha_t; \Omega) \approx P(R = 1 \mid x_t, \alpha_t, \Omega)$

Denoising-progress embedding:

$$\boldsymbol{h}_t^n = H^n(\boldsymbol{x}_t) \in \mathbb{R}^{L \times d},$$

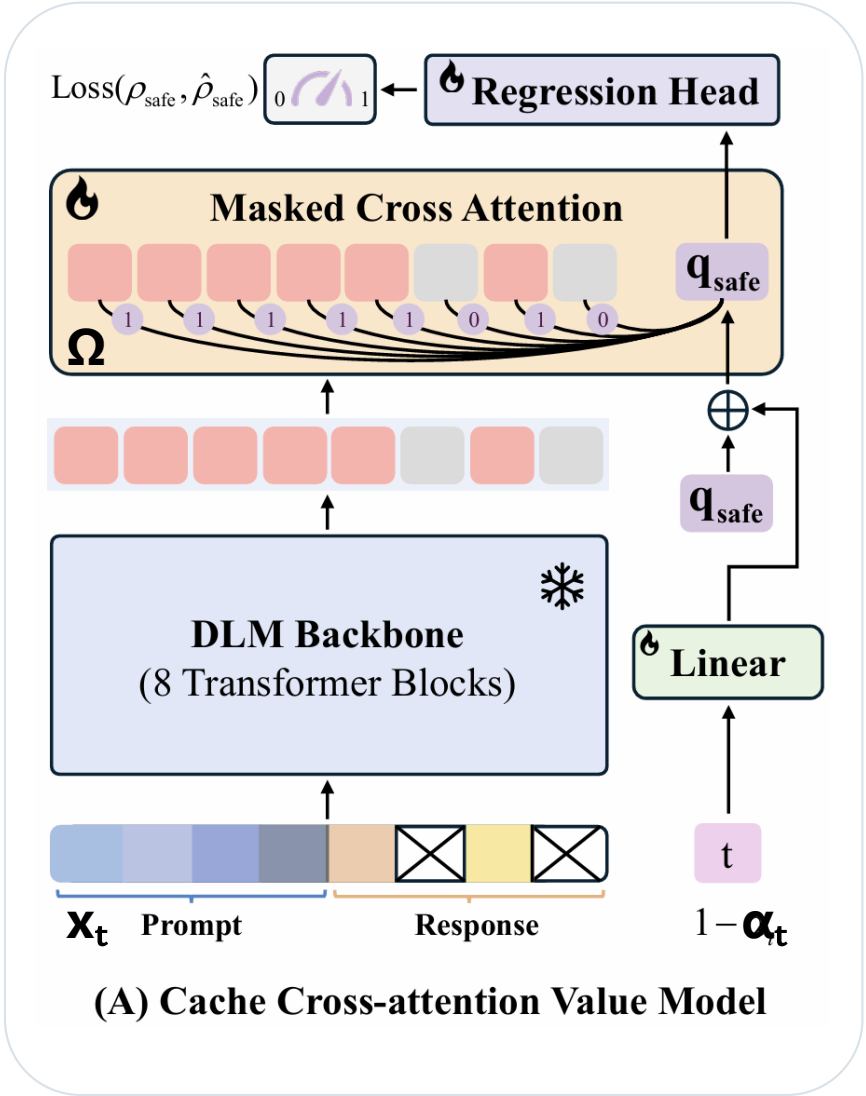
Masked cross-attention:

$$\boldsymbol{h}_{\text{safe}} = \text{MHA}\big(\boldsymbol{q}_{\text{safe}} + \boldsymbol{e}_t, \boldsymbol{h}_t^n, \boldsymbol{h}_t^n; \mathcal{M}(\Omega)\big),$$

Safety-value regression:

$$V_\theta(\boldsymbol{x}_t, \alpha_t; \Omega) = \text{Regression}(\boldsymbol{h}_{\text{safe}}).$$

Cached features support multiple counterfactual readouts. Varying the attention mask avoids additional backbone forward passes.



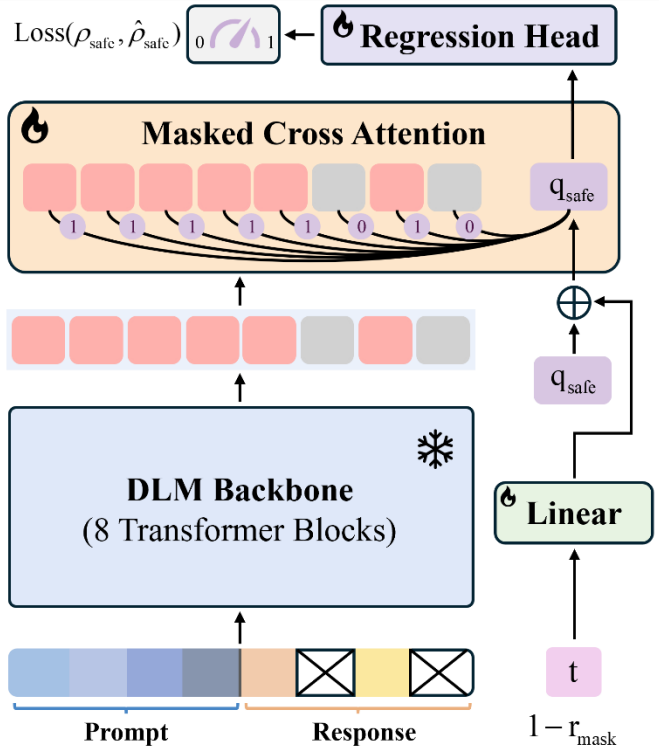
(A) Cache Cross-attention Value Model

Method: Value Supervision and Training

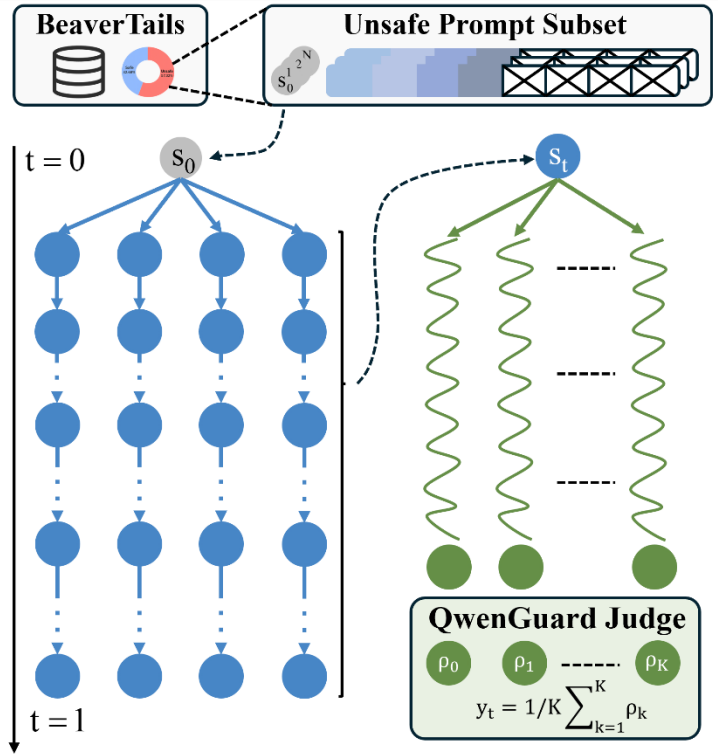
② Spine-tree sampling

Challenge: Intermediate-state safety returns are unobserved; exhaustive rollouts are computationally prohibitive.

Spine checkpoints	Select checkpoints from a small set of complete denoising trajectories.
K-branch rollouts	Sample K stochastic continuations from each checkpoint.
Safety-return target	Average the binary safety returns of the completed branches.
Training objective	Smooth L1 + Pairwise Margin Loss



(A) Cache Cross-attention Value Model



(B) Spine Tree Search & K-branch Rollout

Checkpoint-level branching approximates conditional safety returns at a tractable sampling cost.

Method: Counterfactual Test-time Alignment

③ Token audit

Challenge: State-level risk estimates do not identify which tokens drive unsafe continuations.

Necessity (nec)

Risk reduction when token i is removed

$$\Delta_t^{\text{nec}}(i) = \left[U_\theta(\boldsymbol{x}_t, \Omega_t) - U_\theta(\boldsymbol{x}_t, \Omega_{t,-i})\right]_+$$

+

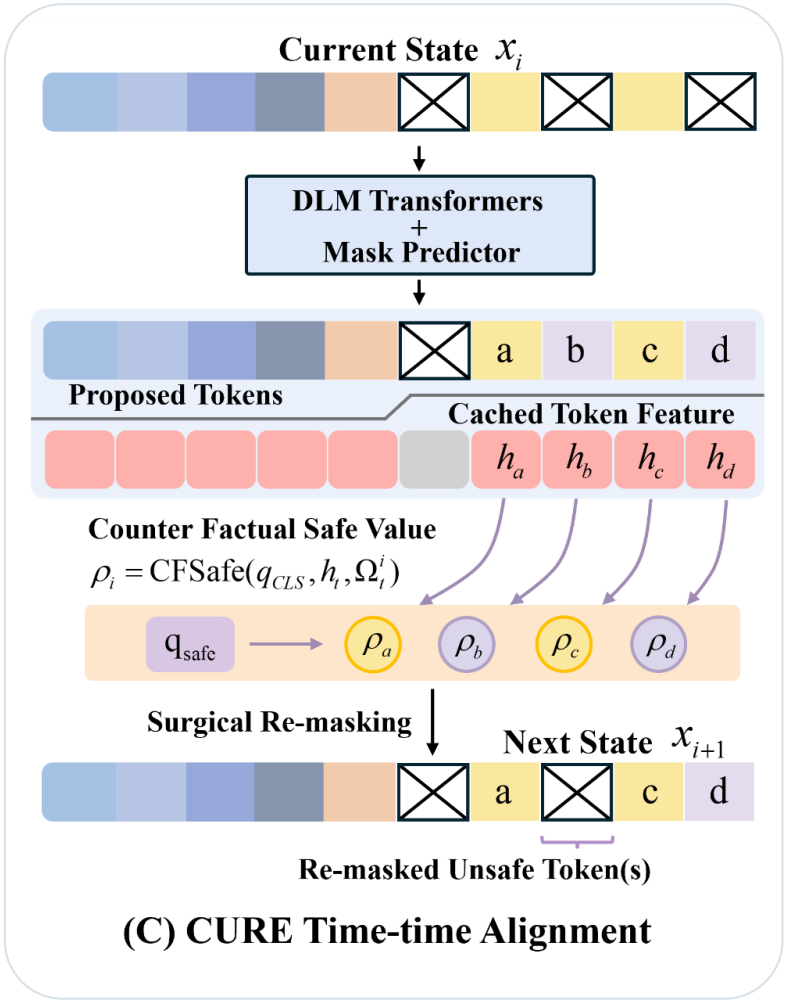
Sufficiency (suf)

Predicted risk when only token i is visible

$$\Delta_t^{\text{suf}}(i) = U_\theta(\boldsymbol{x}_t, \Omega_{t,i})$$

$$\rho_{t,i} = \lambda_{\text{nec}} \Delta_t^{\text{nec}}(i) + \lambda_{\text{suf}} \Delta_t^{\text{suf}}(i).$$

Risk- and uncertainty-guided re-masking selectively revises high-scoring tokens.



(C) CURE Time-time Alignment

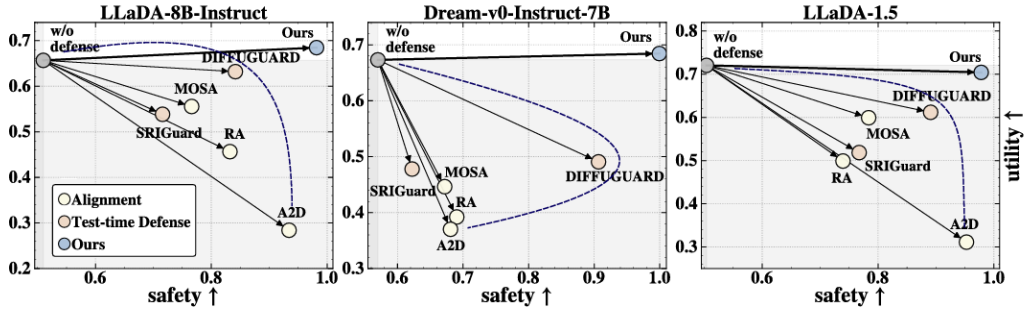
Experimental Results

Jailbreak robustness across three DLMs and five benchmarks

Methods	WildJailbreak ASR ↓	JBB-Behaviors ASR ↓	HarmBench ASR ↓	$\text{PAD}_{\text{AdvBench}}$ ASR ↓	$\text{DIJA}_{\text{AdvBench}}$ ASR ↓	*Average* ASR ↓
LLaDA-8B-Instruct	23.95	13.00	15.75	93.65	98.65	49.00
+RA	$3.50_{\downarrow 20.45}$	$1.00_{\downarrow 12.00}$	$8.50_{\downarrow 7.25}$	$\mathbf{3.27}_{\downarrow 90.38}$	$67.63_{\downarrow 31.02}$	$16.78_{\downarrow 32.22}$
+A2D	$2.05_{\downarrow 21.90}$	$\mathbf{0.00}_{\downarrow 13.00}$	$1.75_{\downarrow 14.00}$	$76.90_{\downarrow 16.75}$	$28.70_{\downarrow 69.95}$	$21.88_{\downarrow 27.12}$
+MOSA	$0.70_{\downarrow 23.25}$	$\mathbf{0.00}_{\downarrow 13.00}$	$2.26_{\downarrow 13.49}$	$33.46_{\downarrow 60.19}$	$80.54_{\downarrow 18.11}$	$23.39_{\downarrow 25.61}$
+SRIGuard	$14.25_{\downarrow 9.70}$	$7.00_{\downarrow 6.00}$	$15.25_{\downarrow 0.50}$	$34.68_{\downarrow 58.97}$	$70.91_{\downarrow 27.74}$	$28.42_{\downarrow 20.58}$
+DIFFUGUARD	$8.50_{\downarrow 15.45}$	$4.00_{\downarrow 9.00}$	$3.25_{\downarrow 12.50}$	$24.42_{\downarrow 69.23}$	$39.04_{\downarrow 59.61}$	$15.84_{\downarrow 33.16}$
+CURE	$\mathbf{0.50}_{\downarrow 23.45}$	$\mathbf{0.00}_{\downarrow 13.00}$	$\mathbf{0.00}_{\downarrow 15.75}$	$4.42_{\downarrow 89.23}$	$\mathbf{4.23}_{\downarrow 94.42}$	$\mathbf{1.83}_{\downarrow 47.17}$
Dream-v0-Instruct-7B	7.75	4.00	4.75	99.23	99.23	42.99
+RA	$0.35_{\downarrow 7.40}$	$\mathbf{0.00}_{\downarrow 4.00}$	$\mathbf{0.00}_{\downarrow 4.75}$	$58.69_{\downarrow 40.54}$	$95.77_{\downarrow 3.46}$	$30.96_{\downarrow 12.03}$
+A2D	$0.30_{\downarrow 7.45}$	$\mathbf{0.00}_{\downarrow 4.00}$	$0.75_{\downarrow 4.00}$	$82.60_{\downarrow 16.63}$	$75.92_{\downarrow 23.31}$	$31.91_{\downarrow 11.08}$
+MOSA	$0.25_{\downarrow 7.50}$	$1.00_{\downarrow 3.00}$	$3.25_{\downarrow 1.50}$	$67.24_{\downarrow 31.99}$	$92.29_{\downarrow 6.94}$	$32.81_{\downarrow 10.18}$
+SRIGuard	$6.65_{\downarrow 1.10}$	$3.00_{\downarrow 1.00}$	$4.25_{\downarrow 0.50}$	$76.69_{\downarrow 22.54}$	$98.27_{\downarrow 0.96}$	$37.77_{\downarrow 5.22}$
+DIFFUGUARD	$2.35_{\downarrow 5.40}$	$3.00_{\downarrow 1.00}$	$3.25_{\downarrow 1.50}$	$31.15_{\downarrow 68.08}$	$6.94_{\downarrow 92.29}$	$9.34_{\downarrow 33.65}$
+CURE	$\mathbf{0.10}_{\downarrow 7.65}$	$\mathbf{0.00}_{\downarrow 4.00}$	$\mathbf{0.00}_{\downarrow 4.75}$	$\mathbf{0.00}_{\downarrow 99.23}$	$\mathbf{0.19}_{\downarrow 99.04}$	$\mathbf{0.06}_{\downarrow 42.93}$
LLaDA-1.5	27.40	13.00	22.10	87.69	97.88	49.61
+RA	$8.59_{\downarrow 18.81}$	$1.00_{\downarrow 12.00}$	$18.50_{\downarrow 3.60}$	$22.11_{\downarrow 65.58}$	$80.08_{\downarrow 17.80}$	$26.06_{\downarrow 23.55}$
+A2D	$3.05_{\downarrow 24.35}$	$2.00_{\downarrow 11.00}$	$\mathbf{0.50}_{\downarrow 21.60}$	$76.90_{\downarrow 10.79}$	$17.76_{\downarrow 80.12}$	$20.04_{\downarrow 29.57}$
+MOSA	$\mathbf{1.10}_{\downarrow 26.30}$	$\mathbf{0.00}_{\downarrow 13.00}$	$2.75_{\downarrow 19.35}$	$19.08_{\downarrow 68.61}$	$85.19_{\downarrow 12.69}$	$21.62_{\downarrow 27.99}$
+SRIGuard	$21.04_{\downarrow 6.36}$	$8.00_{\downarrow 5.00}$	$21.00_{\downarrow 1.10}$	$19.04_{\downarrow 68.65}$	$47.21_{\downarrow 50.67}$	$23.26_{\downarrow 26.35}$
+DIFFUGUARD	$10.95_{\downarrow 16.45}$	$5.00_{\downarrow 8.00}$	$7.50_{\downarrow 14.60}$	$12.12_{\downarrow 75.57}$	$19.04_{\downarrow 78.84}$	$10.92_{\downarrow 38.69}$
+CURE	$\mathbf{1.10}_{\downarrow 26.30}$	$1.00_{\downarrow 12.00}$	$1.75_{\downarrow 20.35}$	$\mathbf{2.88}_{\downarrow 84.81}$	$\mathbf{4.23}_{\downarrow 93.65}$	$\mathbf{2.19}_{\downarrow 47.42}$

Macro-average ASR falls from 47.20% to 1.36% across the three DLMs and five jailbreak benchmarks.

Safety–utility Pareto frontier



Computational overhead

Method	LLaDA-8B-Instruct	SRIGuard	DIFFUGUARD	CURE
TFLOPs	126.29	252.58 (200.0%)	130.24 (103.1%)	**0.84 (0.7%)**

Pareto-efficient safety with preserved MMLU/GSM8K utility; only 0.84 extra TFLOPs (0.7%) on LLaDA-8B-Instruct.

Conclusions and Outlook

Editable denoising states enable targeted safety correction informed by downstream risk estimates.

Principal Findings

01 Risk propagation during denoising

Semantically benign tokens can influence unsafe continuations through bidirectional conditioning.

02 Safety-constrained test-time optimization

Re-masking revises local commitments while retaining the pretrained generator.

03 Efficient counterfactual intervention

Cached value readouts identify risk-inducing tokens without additional backbone passes.

Future Directions

Backtracking Beyond Tokens

Extend interventions to token spans and reasoning steps to capture compositional risk.

Multi-constraint control

Jointly consider safety, factual accuracy, and logical consistency during generation.

Toward fine-grained, reversible control of diffusion generation

Thank you!