Rebuttal

We thank all the reviewers for their constructive feedback, and for noting our work:

1."provides extensive empirical evidence demonstrating how different manipulations affect the output of LLMs, giving credibility to their findings"

2."not only identifies weaknesses but also actively tests the resilience of safety measures.",

3."This method is innovative".

We make a great effort to conduct additional experiments and hope to address your concerns. We feel grateful for all the reviewers' active reviews and participation and hope you could raise your score if you find these rebuttals helpful.

Reviewer T64R

[Question 1] The proposed methods demand a sophisticated understanding of model decoding processes, which limits practical application without the presence of expert knowledge.

Thanks for your carefully review and constructive question. Our method is inspired by the A * algorithm [1] in reinforcement learning. Some readers may not have relevant backgrounds, but we believe that our innovative approach can draw more attention to knowledge from different fields and stimulate more collisions of thinking. This is also the key to the success of reinforcement learning in the field of large language models, such as DPO, PPO, and other RL algorithms.

[1] Yonetani R, Taniai T, Barekatain M, et al. Path planning using neural a* search[C]//International conference on machine learning. PMLR, 2021: 12029-12039.

[Question 2] Methods section lacks adequate details.

Thanks for your carefully reading and pointing out this issue. We notice that although we have tried to introduce the symbols we define in section "Model generation is actually an MDP", this is a bit far from the paragraph "Cost Value Model Estimation" and may confuse readers. We will reintroduce our symbols in this paragraph again to provide readers with a better reading experience. Also, we are about to release our code and detailed hyper-parameters, along with training dataset for convenience, this will help the readers know our pipeline better.

[Question 3] The paper does not provide a standardized testing framework for assessing vulnerabilities across different models

Thank you for your review. To demonstrate the effectiveness of our method, we tested three safety-aligned models on a large set of toxic questions. We conducted both human evaluations and safeguard model assessments to ensure the reliability of the results. Since our method focuses on the decoding process, we compared it against the vanilla decoding strategy as a baseline.

Importantly, our approach is orthogonal to prior methods that concentrate on model inputs. This means our method does not contradict existing approaches but can complement them instead. This is a key innovation of our work. To illustrate this, we combined our method with a model input jailbreak method: 'trainable noise on pictures as soft prompts.' The ablation results show that our generated toxic responses can be unique targets for each toxic question, this bring benefits on ASR and training loss. Moreover, the responses gathered through our method can act as specific targets for suffix methods like GCG. While GCG aims to optimize the model to respond with "Sure, here it is" to all queries, our method provides distinct targets for each question. We have made every effort to integrate our method with GCG. We select the successful pairs by Cost Value Model and apply these toxic responses as the target of GCG. The outcomes are presented below:

ASR(%)	Vicuna-13B(1090 pairs from JVD)	LLaMA-2-chat(317 pairs from JVD)	Mistral-7B-Instruct-v0.2(568 pairs from JVD)
GCG	87.98%	52.98%	78.87%
GCG+JVD	96.51%	70.98%	87.15%

[Question 4] The paper does not adequately compare its approach to existing safety mechanisms and jailbreak techniques.

Thanks for this question! As we mention above, our method which focuses on the model decoding process is orthogonal to previous methods which focus on model inputs. Actually, our method is not in opposition to previous methods, but can complement each other! This is also

one of our biggest innovation points.To show the efficiency of our method, we compare with vanilla decoding strategy. To show the combination between our method and previous method, we show the results on 'trainable noise on picture as soft prompt', demonstrating the efficiency of this combination. We think comparing a method focusing on decoding space with input space is not that fair, as the former freezes model input and parameter to make the model stays in ‘safe mode’, the latter change the model's 'safe mode' to 'dangerous mode' by changing the input text. Our method provides a new paradigm for utilizing both model inputs and outputs to achieve a higher ASR.

Moreover, we have conducted a thorough analysis and presentation of our experimental results, aiming to help readers gain credible insights. While we may not compare our method with enough previous input methods, we have made every effort to provide a detailed explanation of our approach and demonstrate its advantages when combined with these previous methods. We strive to ensure that our experiments are as coherent and self-explanatory as possible.

Reviewer TKst

[Question 1] The first few tokens in the decoding process may be rejected responses.

Thanks for your carefully reading and reviewing! As we mention in conclusion:"How diverse the CVM guided toxic response is?", we list a few first tokens generated by CVM. These tokens include refusal beginnings, such as 'I ...', and several affirmative tokens, such as 'To ...', '0 ...', '1- ....' Therefore, these tokens remind us that when attacking the target model, it is not necessary to fit a unique sentence 'Sure, here it is'. On the contrary, each question has its own unique target tokens to fit.

[Question 2] There is no comparison experiment with other advanced jailbreak methods.

Thanks a lot for this question! Our method is orthogonal to previous methods which focus on model inputs, such as GCG. Therefore, actually our method is not in opposition to previous methods, but can complement each other! This is also one of our biggest innovation points. To verify this, we show a combination between our method and a model input jailbreak method: 'trainable noise on picture as soft prompt' as an example of orthogonality. The ablation results show that our generated toxic responses can be unique targets for each toxic question, this bring benefits on ASR and training loss. Also, responses collected by our method can be specific targets for suffix methods such as GCG. As GCG optimizes the target model to answer with

"Sure, here it is" for all questions, our method provides each question with a specific target. We try our best to implement the combination of our method and GCG. We select the successful pairs by Cost Value Model and apply these toxic responses as the target of GCG. The outcomes are presented below:

ASR(%)	Vicuna-13B(1090 pairs from JVD)	LLaMA-2-chat(317 pairs from JVD)	Mistral-7B-Instruct-v0.2(568 pairs from JVD)
GCG	87.98%	52.98%	78.87%
GCG+JVD	96.51%	70.98%	87.15%

Moreover, we have conducted a thorough analysis and presentation of our experimental results, aiming to help readers gain credible insights. While we may not compare our method with enough previous input methods, we have made every effort to provide a detailed explanation of our approach and demonstrate its advantages when combined with these previous methods. We strive to ensure that our experiments are as coherent and self-explanatory as possible.

Importantly, we think comparing a method focusing on decoding space with input space is not that fair, as the former freezes model input and parameter to make the model stays in ‘safe mode’, the latter change the model's 'safe mode' to 'dangerous mode'. Our method provides a new paradigm for combining model inputs and outputs to achieve a higher ASR.

Reviewer Kg9c

[Question 1] The proposed attack method relies on a white-box model.

Thanks for this question! We acknowledge that our approach may be limited on closed-source black-box models. However, this method is not entirely confined to fully open-source white-box models either. Our method is applicable not only to white-box models where parameters and model weights are accessible but also to gray-box models, such as GPT-4, where model weights are inaccessible but logits can be observed. This expands its practical applicability to a certain point.

[Question 2] Not compare the effectiveness of their attack with existing methods.

Thanks a lot for this question! Our method is orthogonal to previous methods which focus on model inputs, such as GCG. Therefore, actually our method is not in opposition to previous methods, but can complement each other! This is also one of our biggest innovation points. To verify this, we show a combination between our method and a model input jailbreak method: 'trainable noise on picture as soft prompt' as an example of orthogonality. The ablation results show that our generated toxic responses can be unique targets for each toxic question, this bring benefits on ASR and training loss. Also, responses collected by our method can be specific targets for suffix methods such as GCG. As GCG optimizes the target model to answer with "Sure, here it is" for all questions, our method provides each question with a specific target. We try our best to implement the combination of our method and GCG We select the successful pairs by Cost Value Model and apply these toxic responses as the target of GCG. The outcomes are presented below:

ASR(%)	Vicuna-13B(1090 pairs from JVD)	LLaMA-2-chat(317 pairs from JVD)	Mistral-7B-Instruct-v0.2(568 pairs from JVD)
GCG	87.98%	52.98%	78.87%
GCG+JVD	96.51%	70.98%	87.15%

Moreover, we have conducted a thorough analysis and presentation of our experimental results, aiming to help readers gain credible insights. While we may not compare our method with enough previous input methods, we have made every effort to provide a detailed explanation of our approach and demonstrate its advantages when combined with these previous methods. We strive to ensure that our experiments are as coherent and self-explanatory as possible.Importantly, we think comparing a method focusing on decoding space with input space is not that fair, as the former freezes model input and parameter to make the model stays in ‘safe mode’, the latter change the model's 'safe mode' to 'dangerous mode'. Our method provides a new paradigm for combining model inputs and outputs to achieve a higher ASR.

[Question 3] Lacks control experiments to analyze the impact of hyperparameters on experimental outcomes.

Thanks for this question! To make sure the experiment results more convincing, we have implemented a new experiment on hyperparameters \beta on three models:

ASR(%)	regular	$\beta = 2$	$\beta = 4$	$\beta = 6$	$\beta = 8$	$\beta = 10$	$\beta = 12$
Vicuna-13B	60.73	96.63	99.11	98.67	98.67	99.29	98.67
LLaMA-2-chat 7B	6.56	8.07	12.06	15.51	19.95	22.87	28.10
Mistral-7B-Instruct-v0.2	24.27	27.90	35.43	51.28	53.14	55.27	unreadable

[Question 4] Not provide information about the computational efficiency of the inference process.

Thanks for this question! The computational efficiency of the inference process is as follows:

Let m be the compute (in FLOPS) required by an LLM to process a single token, and n the compute required by a cost value model. For normal decoding a T length response, the FLOPS required is $O(T^2 m)$. For cost value guided decoding, the FLOPS required is $O(T^2 (m+kn))$ for top-k selection. Therefore, value guided decoding is indeed slower than vanilla decoding, but their time complexity is similar.