Autonomous LLM-Enhanced Adversarial Attack for Text-to-Motion

Honglei Miao[1], Fan Ma[2], Ruijie Quan[3], Kun Zhan[1]*, Yi Yang[2]

[1]School of Information Science and Engineering, Lanzhou University
[2]College of Computer Science and Technology, Zhejiang University
[3]College of Computing and Data Science, Nanyang Technological University
kzhan@lzu.edu.cn

Abstract

Human motion generative models have enabled promising applications, but the ability of text-to-motion (T2M) models to produce realistic motions raises security concerns if exploited maliciously. Despite growing interest in T2M, limited research focus on safeguarding these models against adversarial attacks, with existing work on text-to-image models proving insufficient for the unique motion domain. In the paper, we propose ALERT-Motion, an autonomous framework that leverages large language models (LLMs) to generate targeted adversarial attacks against black-box T2M models. Unlike prior methods that modify prompts through predefined rules, ALERT-Motion uses the knowledge of LLMs of human motion to autonomously generate subtle yet powerful adversarial text descriptions. It comprises two key modules: an adaptive dispatching module that constructs an LLM-based agent to iteratively refine and search for adversarial prompts; and a multimodal information contrastive module that extracts semantically relevant motion information to guide the agent's search. Through this LLM-driven approach, ALERT-Motion produces adversarial prompts querying victim models to produce outputs closely matching targeted motions, while avoiding obvious perturbations. Evaluations across popular T2M models demonstrate ALERT-Motion's superiority over previous methods, achieving higher attack success rates with stealthier adversarial prompts. This pioneering work on T2M adversarial attacks highlights the urgency of developing defensive measures as motion generation technology advances, urging further research into safe and responsible deployment.

1 Introduction

Human motion generation is a task that aims at producing natural and realistic human motions. It drives advancements in downstream applications such as animation and movie production, virtual human construction, robotics, and human-robot interaction (Zhu et al. 2023). In recent years, with the development of deep learning, especially the growth of generative models such as the Generative Adversarial Network (GAN) (Goodfellow et al. 2014), Variational Autoencoder (VAE) (Kingma and Welling 2013), and diffusion model (Ho, Jain, and Abbeel 2020), trained models have become capable of generating very natural motions (Tevet et al. 2023; Chen et al. 2023; Guo et al. 2022a; Zhang et al. 2023). Some models (Athanasiou et al. 2022; Barquero, Escalera, and Palmero 2024; Lee, Moon, and Lee 2023; Qing et al. 2023) even extend the motions for several minutes while satisfying given conditions. Among these motion generation models, text-to-motion (T2M) (Tevet et al. 2023; Chen et al. 2023; Guo et al. 2022a; Zhang et al. 2023; Athanasiou et al. 2022; Barquero, Escalera, and Palmero 2024; Qing et al. 2023) attracts significant attention from the community due to the accessibility of text prompts that align with human expression.

Generating motions that closely correspond to textual descriptions and are nearly the same as those in the real physical world is becoming increasingly feasible. However, allowing models to freely generate motions conditioned on arbitrary text prompts poses more significant security risks than text-to-image (T2I). When they are applied to downstream tasks, such capabilities can be maliciously exploited by attackers. For example, in animation or movie production (Qing et al. 2023), they are used to create more realistic harmful content involving pornography or violence. The risks increase when using generated motions as part of humanoid controllers (Luo et al. 2023), as when deployed on robots, posing potential threats to human safety.

Despite the growing focus on T2M, safety concerns remain under-researched. The most work focuses on the safety of T2I (Millière 2022; Struppek, Hintersdorf, and Kersting 2023; Liu et al. 2023a). These researches largely focus on how character- or word-level modifications to benign prompts could induce unintended output from the models. Early work (Millière 2022; Struppek, Hintersdorf, and Kersting 2023) mainly explored the existence of this phenomenon, until the RIATIG (Liu et al. 2023a) is inspired by them to propose targeted attacks against image generation models to raise awareness of potential security risks to T2I. However, existing studies search for adversarial attacks by modifying words to unusual personal names, locations, or other proper nouns, which is overlooked in image generation but appears clearly out of place for motion tasks, making attacks more easily detectable. Additionally, unlike the abundant image-text pairs available for image tasks, the limited data for motions makes it challenging to accurately measure the similarity between different motions, posing further difficulties for targeted attacks on T2M models. They make the findings of T2I safety difficult to apply directly to the motion domain.

*Corresponding Author.

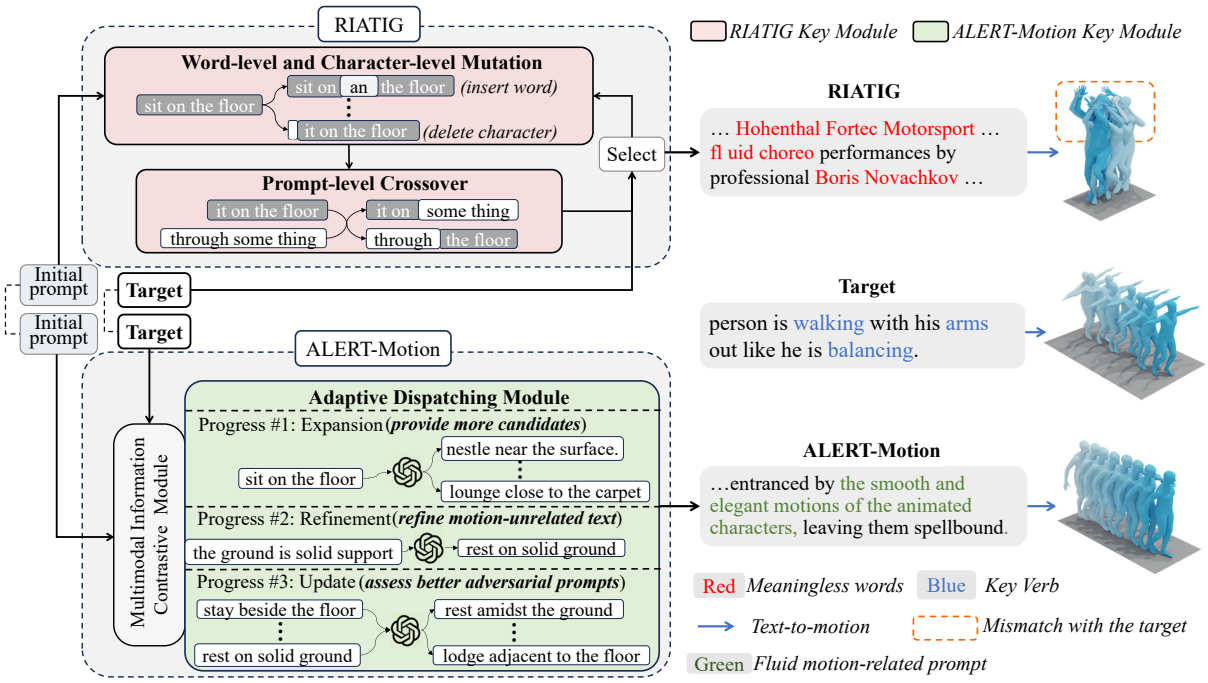

Figure 1: Adversarial prompt against T2M model with RIATIG and our ALERT-Motion. Previous methods like RIATIG only perturb prompts through predefined character or word operations, overlooking the integrity and semantics of the prompts. Our ALERT-Motion does not require such predefined operations; instead, by multimodal information contrastive (MMIC) module, the language model autonomously learns and performs these operations, dynamically generating adversarial prompts that meet the attack requirements. Under the same input (target and initial prompt), our method captures more natural and fluent prompts related to motion. When these prompts are used to query the victim T2M model, the resulting motion show a stronger resemblance to the target motion. Darker color indicates later frames in the sequence.

To address the challenges of adversarial attacks on T2M models, we introduce **ALERT-Motion**, an autonomous large language model (LLM) enhanced adversarial attack against T2M models in a black-box setting. Unlike previous work, our **ALERT-Motion** leverages the knowledge about motions contained in LLMs to generate subtle yet powerful adversarial descriptions, whose outputs from the victim model closely match the desired motion. Importantly, the entire attack process is done automatically by LLM agent, using its own reasoning abilities to carry out the attack, without needing human-defined rules for operations like inserting, deleting, or replacing characters or words.

As shown in Figure 1, previous state-of-the-art attack methods such as RIATIG (Liu et al. 2023a) for T2I models employ manually defined words or character-level modifications and prompt-level crossover, making it difficult to find natural and fluent adversarial text prompts. Such methods often result in conspicuous proper nouns like "Sebastian Hohenthal Fortec Motorsport" or "Boris Novachkov" appearing inappropriately in descriptions of motions. In contrast, our proposed ALERT-Motion gives the modification of adversarial text prompts entirely to LLMs. It comprises two key modules: the adaptive dispatching (AD) module and the multimodal information contrastive (MMIC) module. In the AD module, through the systematic design of instructions for different processes, LLM autonomously searches for adversarial text prompts that appear natural and fluent, avoiding the abrupt word insertions seen in RIATIG. However, as LLMs lack inherent capabilities for processing motion modality, we design the MMIC module to obtain semantically similar information to the target motion, thereby assisting the AD module in finding better adversarial text prompts. Through the coordinated operation of these two modules, ALERT-Motion generates adversarial text prompts that are not only natural and fluent but also query the victim T2M model to produce outputs closely resembling the target motions.

In summary, the main contributions are as follows:

1. To the best of our knowledge, we are the first to propose ALERT-Motion, an autonomous LLM-enhanced adversarial targeted attack method for T2M models.

2. Our proposed ALERT-Motion consists of two key modules. A novel AD module constructs an LLM agent that integrates the agent's inherent natural language and domain knowledge of motions into the automatic attack process. Additionally, MMIC module performs high-level semantic extraction of motion modalities and provides necessary information to support AD's reasoning and decision-making.

3. We evaluate ALERT-Motion on two T2M models and compare it against two adversarial attack methods applied to T2I models. Experimental results demonstrate that ALERT-Motion achieves higher attack success rates while generating more natural and stealthy adversarial prompts.

2 Related Work

Text-to-Motion (T2M). T2M is a conditional motion generation task that aims to generate semantically matching and natural motion sequences from text descriptions. Its impressive performance is due to deep generative models such as GANs, VAEs, diffusion models, etc. One of the early works in this domain, Text2Action (Ahn et al. 2018), leverages GANs to create diverse realistic motions. Some research also explores the use of VAEs for generation, where Language2Pose (Ahuja and Morency 2019) and TEACH (Athanasiou et al. 2022) propose end-to-end text-to-pose generation framework that utilizes a VAE to model the latent space between text and motion. With the advancement of diffusion models in the generative domain, some studies have also employed diffusion models for motion generation. MDM (Tevet et al. 2023) utilizes a diffusion model to predict the sample in each diffusion step rather than just the noise. MLD (Chen et al. 2023) adopts latent diffusion along with a VAE to generate motions, significantly increasing the generation speed without maintaining quality. In addition, there are studies (Guo et al. 2022b; Zhang et al. 2023) that combine VQ-VAE with GPT-like transformers. These works continuously improve the quality, coherence, and efficiency of motion generation from text descriptions. However, no research has focused on attacks and defenses of the T2M model.

Adversarial Attacks on Text-driven Generative Models. Due to the convenience of text input for users, it serves as the most common driving condition for many multimodal generation models. However, the inherent complexity of text input inevitably introduces vulnerabilities to the generative models driven by it. Existing research on adversarial attacks in T2I models, such as (Qu et al. 2023; Zhuang, Zhang, and Liu 2023; Liu et al. 2023a,b), attack T2I models by modifying the input text, causing abnormal outputs. Among them, (Liu et al. 2023a) manipulates words and characters, thereby causing the targeted objects specified by the attacker to be generated in the image by the victim T2I model. These studies indicate the lack of robustness of existing T2I models to input text. With the introduction of LLMs, many studies also focus on the vulnerabilities of LLMs. A large portion of them focus on jailbreaking, making LLMs answer queries that violate safety policies. Jailbreaking strategies have evolved from manual prompt engineering (Wei, Haghtalab, and Steinhardt 2024; Liu et al. 2023c) to LLM-based automated red-teaming (Perez et al. 2022; Liu et al. 2024). Greedy Coordinate Gradient (Zou et al. 2023) uses a white-box model to train adversarial suffixes that maximize the probability of an LLM producing positive responses. They (Zou et al. 2023; Sitawarin et al. 2024) find that the identified suffixes transfer to closed-source off-the-shelf LLMs. The vulnerabilities of T2M models share similarities with the aforementioned security research on text-driven generative models. However, since the correspondence between motion and text involves the time dimension, the attack methods from the above studies cannot be directly applied to T2M. T2VSafetyBench (Miao et al. 2024) evaluates the safety of current mainstream text-to-video (T2V) models. However, this work does not provide attack methods against T2V.

LLM Agents Research on using LLMs to enhance autonomous agents has seen a growing trend (Zhao et al. 2024). These LLM-powered agents, exemplified by HuggingGPT (Shen et al. 2024), WebAgent (Gur et al. 2024), and MM-REACT (Yang et al. 2023), DoraemonGPT (Yang et al. 2024) have been employed to address complex tasks that demand effective understanding and planning from the agents. Many of these studies leverage the rich commonsense knowledge inherently embedded within LLMs to execute downstream tasks with little to no additional training data. Inspired by these explorations, we introduce LLMs into the realm of adversarial attacks on T2Ms, achieving an autonomous attack agent adept at generating effective adversarial prompts.

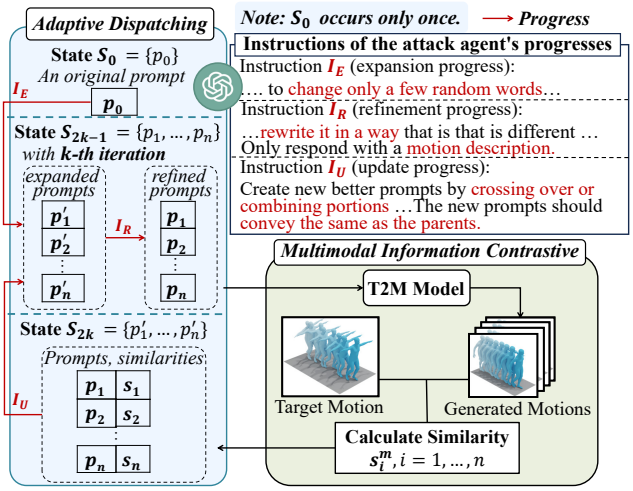

Figure 2: Overview of the proposed ALERT-Motion. ALERT-Motion operates in a black-box setting with two key modules: multimodal information constrastive module for consolidating information from text and motion into a unified format, and autonomous AD module that learns and executes adversarial prompt search through processes of expansion, refinement, and update.

3 Methodology

We use an LLM to iteratively refine and improve adversarial prompts towards a target motion. Initially, LLM generates alternative prompts semantically similar to the initial prompt to expand the search space. It then queries the victim T2M model with these prompts, recording the generated motions. The textual prompts and corresponding motions are unified into a suitable input format for LLM using MMIC. Exploiting its commonsense reasoning capabilities, LLM autonomously analyzes and updates the prompts based on the query results, iteratively bringing them closer to the target motion. This process continues until the adversarial prompts evade detection while generating motions closely matching the target. Fig. 2 overviews our ALERT-Motion attack framework.

3.1 Problem Formulation

A T2M generative model G is essentially a function that maps the text prompt space P to the motion space M. Ideally, through training on semantically aligned text-motion

pairs, a well-trained model generates target motion m_t that is semantically consistent with a given target prompt p_t. The objective of an adversarial attack is to find an adversarial prompt $p^\star$ such that $G(p^\star)$ closely approximates the target motion m_t. Moreover, $p^\star$ is semantically dissimilar from the target prompt p_t to avoid detection. The optimization steps outlined above are formulated as follows

$$p^\star = \arg\max_{p \in P} \; s^{\mathrm{m}}\left(G(p), m_t\right), \text{ s.t. } s^{\mathrm{p}}\left(p, p_t\right) < \eta \quad (1)$$

where s^{m} represents the semantic similarity of motion, s^{p} represents the semantic similarity between text prompts, and P is a prompt set. η is the threshold. As long as the similarity between the adversarial prompt and the target prompt is below η, we consider that our attack evades detection.

Challenges Implementing adversarial prompt generation from T2M models faces some challenges. First, T2M models need to bridge natural language and physical motion, addressing the semantic gap between the language and motion domains. Different data types have different representation spaces, so integrating multi-modal information is needed. Second, the adversarial textual prompts need to have high fluency in natural language and relevance to the target motion in their query results. But they also need to effectively fool the model. The space to search for good prompts is extremely large though. Autonomously generating optimal adversarial samples that meet quality requirements is a great challenge.

3.2 Multimodal Information Contrastive

Unlike most LLM agent-related research, our task involves motion, which LLM cannot directly handle. Therefore, we design MMIC specifically to process information from different modalities in the task and organize it into textual information, making it convenient for LLM to understand and reasoning. As shown in Fig. 2, MMIC allows the adversarial prompts, refined through LLM, to query the T2M model, obtain corresponding motion and calculate the similarity with the target.

Nevertheless, measuring the similarity directly between two motion poses challenges. We consider measuring the similarity of motion. RIATIG (Liu et al. 2023a) employs the pretrained CLIP (Radford et al. 2021), a model trained on a large-scale dataset of image-text pairs, to obtain semantic features aligned with textual descriptions. Similarly, we use the T2M alignment model in (Petrovich, Black, and Varol 2023) to extract semantic motion features and calculate the cosine similarity between the semantic features of motion as

$$s^{\mathrm{m}}(G(p), m_t) = \frac{E_m(G(p)) \cdot E_m(m_t)}{\|E_m(G(p))\| \, \|E_m(m_t)\|}, \quad (2)$$

where E_m is a encoder (Petrovich, Black, and Varol 2023).

Subsequently, we organize this information into text and integrate it into instructions, enabling LLM to contemplate and reason for better adversarial prompts.

3.3 Adaptive Dispatching

In ALERT-Motion framework, the most critical module is the AD module. This module constructs an attack agent and plays a crucial role in determining the effectiveness of adversarial prompts. In contrast to previous related researches, which predefine various operations to perturb semantics and then use a search algorithm to find prompts with higher scores, we directly convey complex task requirements using instructions, allowing LLM to automatically learn and execute all operations, with each step conducted in textual form. According to the purpose of instructions, we divide AD into three progresses: expansion, refinement, and update. The workflow of these three progresses and MMIC is outlined in Algorithm 1.

Algorithm 1: ALERT-Motion

Require: Initial prompt p_0, expansion instruction I_E, refinement instruction I_R, update instruction I_U, size of updated prompts N, s^{m} denotes the similarity of the adversarial motion and the target motion, a predefined number of iterations K.

Expansion: $S_1 \leftarrow \mathrm{LLM}(I_E, S_0 = p_0)$
for $k = 1$ to K **do**
  if $k (\mathrm{mod}\, 2) = 1$ **then**
    Refinement: $S_{2k} \leftarrow \mathrm{LLM}(I_R, S_{2k-1})$
    MMIC: Compute $s_i^{\mathrm{m}}(G(p_i), m_t), \forall\, p_i \in S_{2k}$.
    Obtain the similarity set $S'_{2k} = \{s_1^{\mathrm{m}}, \ldots, s_n^{\mathrm{m}}\}$
  else
    Update $S_{2k+1} \leftarrow \mathrm{LLM}(I_U, \mathrm{cat}(S_{2k}, S'_{2k}))$.
  end if
end for

Ensure: $p^\star = \arg\max_{p \in P}\big(s^{\mathrm{m}}(G(p), m_{\mathrm{t}})\big), P = \cup S_{2k}$.

Due to the fact that AD responds to the current input in each round, similarly to an agent in reinforcement learning, we adopt related concepts here to aid the definition of the processes within AD. We start by defining the state as the set of adversarial prompts and their corresponding information for each round, while the action is represented by various instructions sent to LLM. LLM is viewed as a function involving the next state S_{k+1}, current state S_k, and current action a_k, expressed as

$$S_{k+1} \leftarrow T(S_k, a_k) = \mathrm{LLM}\big(I, S_k\big), \quad (3)$$

where T is the state transition function and I represents the instruction text corresponding to a_k. It is important to note that the representation of state S differs between odd and even time steps, it is defined as

$$S_k = \begin{cases} \{p_0\} & \text{if } k = 0 \\ \{p_1, \ldots, p_n\} & \text{if } k \,(\mathrm{mod}\, 2) = 0 \\ \{p'_1, \ldots, p'_n\} & \text{if } k \,(\mathrm{mod}\, 2) = 1 \end{cases} \quad (4)$$

where p_0 is initial adversarial prompt, $\{p_1, \ldots, p_n\}$ is the set of refined prompts, and p' are the expanded or updated prompts of p.

Expansion: Initially, we begin with a single available adversarial prompt p_0. The current state is defined as $S_0 = \{p_0\}$. Without expansion, proceeding directly to the subsequent steps may lead the search to a local optimum. So we employ the expansion instruction text E to obtain expanded results through LLM. The next state is represented by

$$S_1 \leftarrow \mathrm{LLM}(I_E, S_0) \quad (5)$$

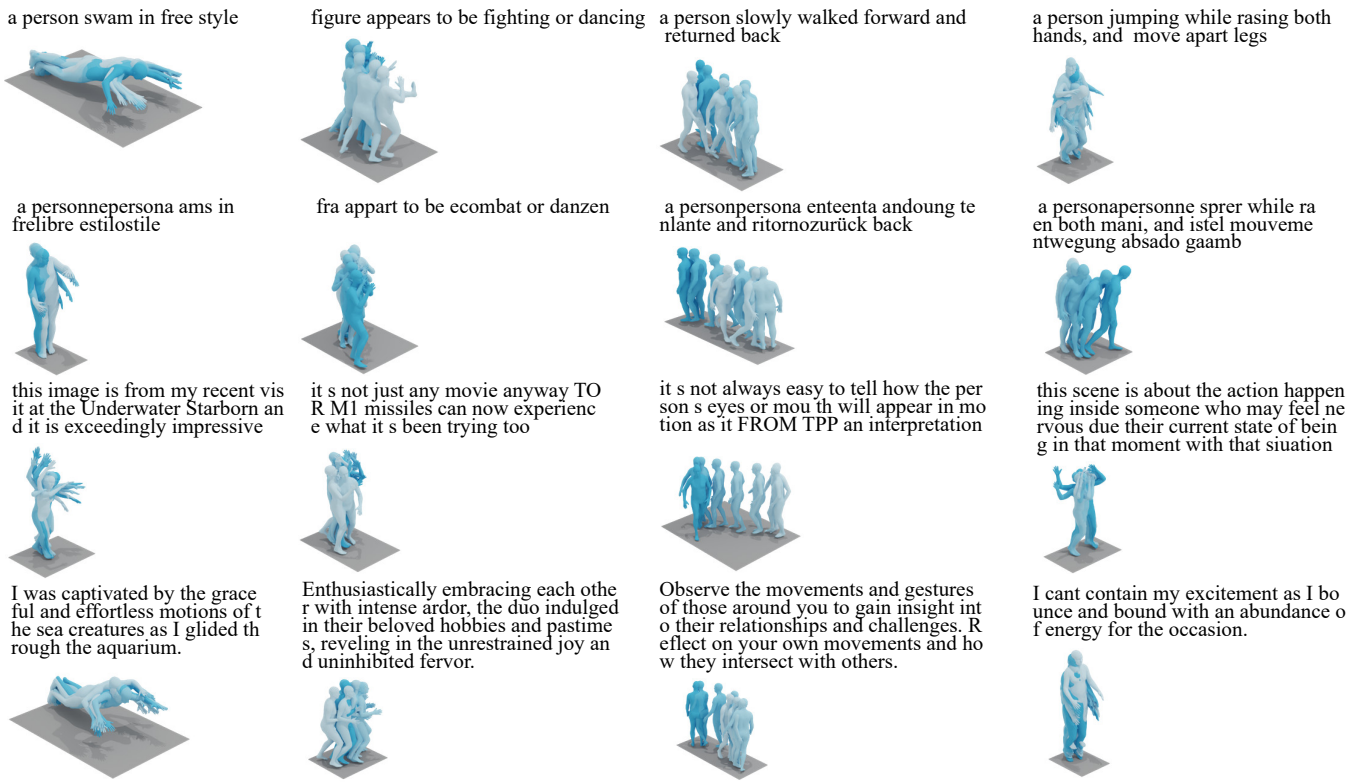

Figure 3: Examples of adversarial attack results against MLD. The first row of text provides the true annotations for each column of target motions, and the first row of motions corresponds to their respective target motions. The following three rows of text correspond to the adversarial prompts obtained by MacPromp, RIATIG, and our proposed ALERT-Motion. The motion-rendered images below the text depict the motions generated by querying the victim model with the adversarial prompts.

where I_E represents the expansion instruction and $S_1 = \{p_1, \ldots, p_n\}$ is the set of expanded prompts.

Refinement: We find that when the instructions given to LLM are too long, its responses sometimes fail to meet the attack requirements. This is because the LLM's ability to follow instructions is limited. New instructions are needed to emphasize the attack requirements in our task. Therefore, in this stage, we refine the adversarial prompts to ensure that they consistently meet the attack requirements, including being naturally fluent and relevant to motion. After refinement, the state of the agent is defined by

$$S_{2k} \leftarrow \mathrm{LLM}(I_R, S_{2k-1}) \tag{6}$$

where $k \in \{1, \ldots, K\}$ and I_R represents refinement.

Update: Unlike existing methods that relied on numerical scalar guidance to generate adversarial prompts, such as RIATIG, our AD utilizes the text information organized by MMIC to guide LLM in autonomously analyzing and generating adversarial prompts. This allows us to control the generated adversarial prompts more precisely and effectively in line with the attack requirements using richer information. Moreover, this control is automated, removing the necessity to continuously define new operations, such as word or character insertion, deletion, replacement, etc., as in previous methods. In the update progress, as LLM analyzes and generates adversarial prompts, the information organized by MMIC is also provided to LLM to assist in its decision-making process. After update, the agent state is defined by

$$S_{2k+1} \leftarrow \mathrm{LLM}\big(I_U, \mathrm{cat}(S_{2k}, S'_{2k})\big) \tag{7}$$

where I_U is the update instruction and the function cat represents string concatenation, and $S'_{2k} = \{s_i^{\mathrm{m}}(G(p_i), m_{\mathrm{t}}), \forall\, i \in \{1, \ldots, n\}\}$ is obtained by Eq. (2).

After K rounds of iteration, we choose the highest scoring prompt among all candidates as the optimal adversarial prompt $p^\star$ for the attack. The definition of $p^\star$ is obtained by Eq. (1). Here, $P = \cup S_{2k}$ and $S_{2k} = \{p_1, \ldots, p_n\}$. In order to ensure that the adversarial prompt meets the constraints, we calculate the similarity between the adversarial prompts and target prompt as

$$s^{\mathrm{p}}(p_t, p) = \frac{E_t(p_t) \cdot E_t(p)}{\|E_t(p_t)\| \, \|E_t(p)\|}, \tag{8}$$

where E_t is a text encoder (Cer et al. 2018) to extract features.

4 Experiment

4.1 Experimental Settings

Datasets. We select target prompt texts and target motion from the HumanML3D (H3D) (Guo et al. 2022a). It includes 14,616 motion sequences from AMASS (Mahmood et al. 2019), each with a textual description (totaling

Attack Methods	R-1[1] ↑	R-2↑	R-3↑	R-5↑	R-10↑	FID↓	d_{MM} ↓	PPL[2] ↓	AS[3]↓
MLD									
Target Motion	8 / 20	10 / 20	13 / 20	15 / 20	16 / 20	4.015	4.029	391.055	-
MacPromp	5 / 20	8 / 20	10 / 20	13 / 20	15 / 20	13.935	6.534	3061.488	0.471
RIATIG	4 / 20	7 / 20	**11 / 20**	13 / 20	17 / 20	10.899	5.368	1102.100	0.131
ALERT-Motion	**6 / 20**	**9 / 20**	9 / 20	**15 / 20**	**19 / 20**	**8.881**	**5.016**	**113.223**	**0.067**
MDM (100 steps)									
Target Motion	7 / 20	14 / 20	15 / 20	16 / 20	20 / 20	4.055	3.549	391.055	-
MacPromp	**7 / 20**	7 / 20	8 / 20	16 / 20	16 / 20	11.108	5.106	2698.972	0.484
RIATIG	5 / 20	8 / 20	10 / 20	16 / 20	18 / 20	12.435	5.024	1154.017	0.129
ALERT-Motion	**7 / 20**	**13 / 20**	**14 / 20**	**17 / 20**	**19 / 20**	**5.843**	**4.117**	**179.496**	**0.075**
MDM (1000 steps)									
Target Motion	8 / 20	12 / 20	12 / 20	14 / 20	19 / 20	5.954	4.116	391.055	-
MacPromp	3 / 20	6 / 20	8 / 20	10 / 20	14 / 20	11.149	6.156	3023.887	0.467
RIATIG	4 / 20	7 / 20	10 / 20	**14 / 20**	16 / 20	9.875	5.444	1262.338	0.129
ALERT-Motion	**9 / 20**	**12 / 20**	**13 / 20**	**14 / 20**	**19 / 20**	**6.183**	**4.533**	**140.793**	**0.074**

Table 1: The results of the adversarial attacks against MDM and MLD on T2M evaluation model. The first row, labeled "Target Motion", represents the motion generated by the corresponding victim models, which are the targets of our attack. The quality of these indicators depends solely on the capabilities of the generation models and evaluation models. The second and third rows correspond to the MacPromp and RIATIG baseline models that we select. The final row represents the performance of our proposed method, ALERT-Motion.

[1] R-1, R-2, R-3, R-5, R-10 represent R-precision at R equals 1, 2, 3, 5, and 10, respectively.
[2] PPL represents the perplexity of sentences.
[3] AS stands for Adversarial Similarity, which denotes the similarity between adversarial prompts and target prompts.

44,970 descriptions). It also re-annotates AMASS and HumanAct12 (Guo et al. 2020) motion capture sequences. The dataset contains comprehensive motion data representation involving root velocity, joint positions, joint velocities, joint rotations, and foot contact labels. It is used for both AMASS and HumanAct12 motion.

Victim Models. To assess the effectiveness of ALERT-Motion, we select two leading publicly available T2M models: MLD and MDM. We employ their respective pretrained models from the official GitHub repositories, which are trained on H3D.

Evaluation Setup. In our experiments, all attacks are conducted in a black-box setting, meaning that we generate motion only by querying the model with prompts and obtaining the generated results. We use the "gpt-3.5-turbo-instruct" API with ChatGPT to implement our approach. The initial adversarial prompt text is a motion description randomly generated by ChatGPT. We set the number of iterations as 50, the size of the prompt set as 20. We set the similarity threshold η as 0.4. Examples for the attack are selected from the top 20 of the Dissimilar subset in the evaluation setup of (Petrovich, Black, and Varol 2023), where the model achieves the highest accuracy. During the attack process, we use the model from (Petrovich, Black, and Varol 2023) to extract the motion features to compute cosine similarity and adopt the text feature extraction from (Yang et al. 2021). The effectiveness is evaluated using T2M (Guo et al. 2022a).

Baselines. To the best of our knowledge, there is currently no targeted adversarial attack specifically designed for T2M generation. For comparison, we select two state-of-the-art targeted adversarial attack methods for text-to-image generation, MacPromp (Millière 2022) and RIATIG (Liu et al. 2023a), as baseline methods. Since their tasks do not involve motion, we modify their task settings to match our task.

4.2 Evaluation Metrics

Motion Performance. We use the R Precision, a widely-used metric in T2M (Guo et al. 2022b; Zhang et al. 2023; Tevet et al. 2023; Chen et al. 2023) to evaluate generated motion. This assessment involves comparing each motion not only with its ground-truth text but also with a misaligned description. R Precision is determined through the Euclidean distance between motion and text features. Our evaluation centers on measuring the average accuracy among the top-k ranked descriptions. A ground-truth within the top-k candidates is considered a "True Positive" retrieval. Our approach involves a batch size of 20, including 19 negative examples, and we explore the effectiveness of R at various values: 1 (R-1), 2 (R-2), 3 (R-3), 5 (R-5), and 10 (R-10). Furthermore, our study integrates Frechet Inception Distance (FID) as a metric to assess the quality of generated motion. FID, a widely accepted standard for evaluating content quality (Tevet et al. 2023; Chen et al. 2023), involves comparing features extracted from generated motion and real motion. In our motion domain adaptation, we adopt an evaluator network to represent deep features, different from the original image-based Inception neural network. Smaller FID values are indicative of superior results. Additionally, we compute the Multimodal Distance (d_{MM}), which is the mean Euclidean distance between the motions features and their corresponding textual

Attack Methods	R-1↑	R-2↑	R-3↑	R-5↑	R-10↑
MLD					
MacPromp	5 / 20	6 / 20	7 / 20	9 / 20	13 / 20
RIATIG	6 / 20	7 / 20	8 / 20	11 / 20	**16 / 20**
ALERT-Motion	**8 / 20**	**9 / 20**	**10 / 20**	**12 / 20**	12 / 20
MDM (100 steps)					
MacPromp	4 / 20	6 / 20	9 / 20	11 / 20	16 / 20
RIATIG	5 / 20	6 / 20	7 / 20	11 / 20	14 / 20
ALERT-Motion	**7 / 20**	**12 / 20**	**15 / 20**	**16 / 20**	**17 / 20**
MDM (1000 steps)					
MacPromp	3 / 20	6 / 20	9 / 20	10 / 20	15 / 20
RIATIG	3 / 20	7 / 20	11 / 20	12 / 20	16 / 20
ALERT-Motion	**6 / 20**	**11 / 20**	**12 / 20**	**14 / 20**	**18 / 20**

Table 2: Attack performance on TMR evaluation model.

Methods	R-1 ↑	R-3 ↑	FID ↓	d_{MM} ↓	PPL ↓	AS ↓
RIATIG	4 / 20	11 / 20	10.90	5.37	1102.10	0.13
GPT 3.5	6 / 20	9 / 20	8.88	5.02	113.22	0.07
Llama 3	5 / 20	7 / 20	6.88	5.23	781.28	0.04

Table 3: The impact of different LLMs on our method.

descriptions features in the test examples (Tevet et al. 2023; Chen et al. 2023). A lower value indicates better alignment between prompts and their generated motions.

Adversarial Similarity. In adversarial attacks, it is essential for adversarial prompt text to have low similarity with the target prompt text to evade detection. In line with previous studies, our initial step involves utilizing the Universal Sentence Encoder (Cer et al. 2018) for encoding both the adversarial sentence and the target sentence, resulting in high-dimensional vectors. Subsequently, we determine their adversarial similarity by computing the cosine score.

Naturality. To ensure the naturalness of adversarial examples, we measure perplexity (PPL) using GPT-2 (Radford et al. 2019), trained on real-world sentences. PPL assesses the likelihood of the model in generating the input text, thereby indicating natural fluency of the adversarial prompts. Lower PPL values typically signify higher naturalness.

4.3 Evaluation Results

The attack results on MLD and MDM are shown in Table 1. Compared to the baselines, ALERT-Motion achieves a higher R-precision. Although MacPromp achieves higher R-precision and lower FID and d_{MM} in some cases, its direct translation of target prompts in various languages results in unnatural adversarial prompts. The perplexity is much higher than other methods, and, on the other hand, it closely resembles the target sentences, resulting in high adversarial similarity, making it less practical. RIATIG, compared to MacPromp, achieves similar or even higher R-precision, with a slight decrease in perplexity and adversarial similarity. However, as seen in Fig. 3, there are still incorrect words and some extra spaces.

From Table 1, it can be observed that our proposed ALERT-Motion performs better on most metrics across these models.

	R-1 ↑	R-3 ↑	FID ↓	d_{MM} ↓	PPL ↓	AS ↓
Ours	6 / 20	9 / 20	8.88	5.02	113.22	0.07
w/o E	4 / 20	12 / 20	8.90	4.95	148.41	0.09
w/o R	4 / 20	7 / 20	13.58	6.45	62.16	0.08
w/ random U	3 / 20	4 / 20	19.54	7.65	86.48	0.06

Table 4: Ablation study of expansion, refinement, and update.

Additionally, examining Fig. 3, the adversarial prompts generated by ALERT-Motion are not only more natural but also relevant to the motion. In contrast, prompts obtained by other methods are mostly irrelevant to motion.

4.4 Ablation Study

Influence of Evaluation Models. The current research on the evaluation of motion generation is still limited, with T2M (Guo et al. 2022a) being widely recognized. Studies on motion generation, such as (Zhang et al. 2023; Tevet et al. 2023), and (Chen et al. 2023), adopt T2M to assess the quality of generated models. The latest research on the evaluation of motion generation is presented in TMR (Petrovich, Black, and Varol 2023). Therefore, we also use it to evaluate our experiments. As shown in Table 2, under the TMR model, ALERT-Motion demonstrates significant superiority compared to other baseline methods, indicating that the excellent performance of our proposed ALERT-Motion is not influenced by the choice of evaluation models.

Dependency on LLMs. All experiments are conducted using the GPT-3.5-turbo. We performed a brief comparison using Llama 3-8B-Instruct on MLD. Table 3 shown that even with an open-source model with 8B parameters, our method outperformed RIATIG in most cases. It indicates that our method has low dependency on the performance of LLMs.

Ablation study on steps. We perform ablation of expansion, refinement, and update steps on MLD. The results of Table 4 indicate that every step of our method is crucial.

5 Conclusion

In this paper, a novel method that involves conducting targeted adversarial attack against T2M models is proposed. Additionally, we introduce an autonomous LLM-enhanced adversarial attack method called ALERT-Motion, which comprises two modules: the multimodal information contrastive (MMIC) module and the adaptive dispatching (AD) module. Assisted by MMIC, AD, with the integration of LLM during progresses of expansion, refinement, and updating, autonomously learns and executes the search for optimal adversarial prompts. Our extensive experiments validate the ability to discover adversarial prompts that exhibit both fluency and related to motion. Moreover, these prompts trigger the victim T2M model to generate motion closely resembling the target, thus achieving successful attacks. The vulnerability of T2M models to our attacks suggest an urgent need to develop defensive methods and enhance the robustness against adversarial exploitation.

Acknowledgments

This work was supported by the Key R&D Program of China under Grant No. 2021ZD0112801, the National Natural Science Foundation of China under Grant Nos. 62176108 and U2336212, the Natural Science Foundation of Zhejiang Province under No. DT23F020008, the Natural Science Foundation of Qinghai Province of China under No. 2022-ZJ-929, the Science Foundation of National Archives Administration of China under No. 2024-B-006, and the Supercomputing Center of Lanzhou University.

References

Ahn, H.; Ha, T.; Choi, Y.; Yoo, H.; and Oh, S. 2018. Text2Action: generative adversarial synthesis from language to action. In *ICRA*, 5915–5920.

Ahuja, C.; and Morency, L.-P. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. In *3DV*, 719–728.

Athanasiou, N.; Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEACH: Temporal action composition for 3D humans. In *3DV*, 414–423.

Barquero, G.; Escalera, S.; and Palmero, C. 2024. Seamless Human Motion Composition with Blended Positional Encodings. In *CVPR*.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal sentence encoder for English. In *EMNLP*, 169–174.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your commands via motion diffusion in latent space. In *CVPR*, 18000–18010.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*, volume 27.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating Diverse and Natural 3D Human Motions from Text. In *CVPR*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022b. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In *ECCV*, 580–597. Springer.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2Motion: Conditioned generation of 3D human motions. In *ACM MM*, 2021–2029.

Gur, I.; Furuta, H.; Huang, A. V.; Safdari, M.; Matsuo, Y.; Eck, D.; and Faust, A. 2024. A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis. In *ICLR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, 6840–6851.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, T.; Moon, G.; and Lee, K. M. 2023. MultiAct: Long-term 3D human motion generation from multiple action labels. In *AAAI*, volume 37, 1231–1239.

Liu, H.; Wu, Y.; Zhai, S.; Yuan, B.; and Zhang, N. 2023a. RIATIG: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *CVPR*, 20585–20594.

Liu, Q.; Kortylewski, A.; Bai, Y.; Bai, S.; and Yuille, A. 2023b. Intriguing properties of text-guided diffusion models. *arXiv preprint arXiv:2306.00974*.

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2024. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. In *ICLR*.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; and Liu, Y. 2023c. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Luo, Z.; Cao, J.; Kitani, K.; Xu, W.; et al. 2023. Perpetual humanoid control for real-time simulated avatars. In *CVPR*, 10895–10904.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *ICCV*, 5442–5451.

Miao, Y.; Zhu, Y.; Yu, L.; Zhu, J.; Gao, X.-S.; and Dong, Y. 2024. T2VSafetyBench: Evaluating the Safety of Text-to-Video Generative Models. In *NeurIPS*.

Millière, R. 2022. Adversarial attacks on image generation with made-up words. *arXiv preprint arXiv:2208.04135*.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In *EMNLP*, 3419–3448.

Petrovich, M.; Black, M. J.; and Varol, G. 2023. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 9488–9497.

Qing, Z.; Cai, Z.; Yang, Z.; and Yang, L. 2023. Story-to-motion: Synthesizing infinite and controllable character animation from long text. In *SIGGRAPH Asia 2023 Technical Communications*, 1–4.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *SIGSAC*, 3403–3417.

Radford, A.; Kim, J. W.; Hallacy, C.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2024. HuggingGPT: Solving ai tasks with chatgpt and its friends in hugging face. In *NeurIPS*, volume 36.

Sitawarin, C.; Mu, N.; Wagner, D.; and Araujo, A. 2024. PAL: Proxy-guided black-box attack on large language models. *arXiv preprint arXiv:2402.09674*.

Struppek, L.; Hintersdorf, D.; and Kersting, K. 2023. Rickrolling the Artist: Injecting backdoors into text encoders for text-to-image synthesis. In *ICCV*, 4584–4596.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *ICLR*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2024. Jailbroken: How does llm safety training fail? *NeurIPS*.

Yang, Z.; Chen, G.; Li, X.; Wang, W.; and Yang, Y. 2024. DoraemonGPT: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *ICML*.

Yang, Z.; Li, L.; Wang, J.; Lin, K.; Azarnasab, E.; Ahmed, F.; Liu, Z.; Liu, C.; Zeng, M.; and Wang, L. 2023. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. .

Yang, Z.; Yang, Y.; Cer, D.; Law, J.; and Darve, E. 2021. Universal Sentence Representation Learning with Conditional Masked Language Model. In *EMNLP*, 6216–6228.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 14730–14740.

Zhao, A.; Huang, D.; Xu, Q.; Lin, M.; Liu, Y.-J.; and Huang, G. 2024. Expel: LLM agents are experiential learners. In *AAAI*, volume 38, 19632–19642.

Zhu, W.; Ma, X.; Ro, D.; Ci, H.; Zhang, J.; Shi, J.; Gao, F.; Tian, Q.; and Wang, Y. 2023. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhuang, H.; Zhang, Y.; and Liu, S. 2023. A pilot study of query-free adversarial attack against stable diffusion. In *CVPR*, 2384–2391.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint*.