

# SATO: Stable Text-to-Motion Framework (Supplementary Material)

Anonymous Author(s)

## 1 OVERVIEW OF SUPPLEMENTARY MATERIAL

The supplementary material is organized into the following sections:

- Section 2: Evaluation metrics.
- Section 3: More visualization examples, including visual comparisons between SATO and state-of-the-art approaches, and attention visual examples.
- Section 4: More ablation study, including parameter analysis, perturbation method ablation, and attention analysis.
- Section 5: Details of human evaluation.
- Section 6: SATO pseudo-algorithm.
- Section 7: Computational complexity.
- Section 8: Symbolic representation.

## 2 EVALUATION METRICS

We denote the motion features generated from the original text and perturbed text as  $f_{pred}$  and  $f'_{pred}$ , respectively. The ground-truth motion features and text features are denoted as  $f_{gt}$  and  $f_t$ .

**FID-related.** FID is used to measure the difference in distribution between generated motions. We have the following formulas to obtain  $FID$ ,  $FID_P$ ,  $FID_D$ :

$$FID = \|\mu_{gt} - \mu_{pred}\|_2^2 - \text{Tr}(\Sigma_{gt} + \Sigma_{pred} - 2(\Sigma_{gt}\Sigma_{pred})^{1/2}) \quad (1)$$

$$FID_P = \|\mu_{gt} - \mu_{pred'}\|_2^2 - \text{Tr}(\Sigma_{gt} + \Sigma_{pred'} - 2(\Sigma_{gt}\Sigma_{pred'})^{1/2}) \quad (2)$$

$$FID_D = \|\mu_{pred} - \mu_{pred'}\|_2^2 - \text{Tr}(\Sigma_{pred} + \Sigma_{pred'} - 2(\Sigma_{pred}\Sigma_{pred'})^{1/2}) \quad (3)$$

Here,  $\mu$  represents the mean,  $\Sigma$  is the covariance matrix, and  $\text{Tr}$  denotes the trace of a matrix.  $\text{pred}$  denotes the prediction with the original text as input, and  $\text{pred}'$  denotes the prediction with the perturbed text as input.  $FID$  and  $FID_P$  are metrics utilized to gauge the disparity in distribution between motions generated before and after perturbation, reflecting the variance between the generated motions and target motions. Meanwhile,  $FID_D$  evaluates the dissimilarity in generated motions pre- and post-perturbation. The smaller the difference, the less susceptible the model is to perturbation.

**MM-Dist.** MM-Dist quantifies the disparity between text embeddings and generated motion features. For  $N$  randomly generated samples, MM-Dist calculates the average Euclidean distance between each text feature and the corresponding motion feature, thus assessing feature-level dissimilarities between text and motion. Increasingly smaller MM-Dist values correspond to better prediction results.

$$\text{MM-Dist} = \frac{1}{N} \sum_{i=1}^N \|f_{\text{pred},i} - f_{\text{text},i}\| \quad (4)$$

**Diversity.** Diversity can measure the diversity of action sequences. A larger value of the metric indicates better diversity in the model. We randomly sample  $S$  pairs of motions, denoted as  $f_i$  and  $f'_i$ . According to [2], we set  $S$  to be 300. We can calculate using the following formula:

$$\text{Diversity} = \frac{1}{S} \sum_{i=1}^S \|f_i - f'_i\| \quad (5)$$

**Jensen-Shannon Divergence (JSD).** JSD can measure the similarity between two distributions, with values ranging from 0 to 1. Here, we utilize it to quantify the stability of attention before and after perturbation. We have the attention distributions before and after perturbation, denoted as  $\tilde{\omega}$  and  $\bar{\omega}$  respectively, computed as follows:

$$JSD(\tilde{\omega}, \bar{\omega}) = \frac{1}{2}KL[\tilde{\omega} || \frac{\tilde{\omega} + \bar{\omega}}{2}] + \frac{1}{2}KL[\bar{\omega} || \frac{\tilde{\omega} + \bar{\omega}}{2}] \quad (6)$$

where  $KL$  is the KL divergence between two distributions. A smaller JSD implies a stronger resistance of the model's attention to disturbances.

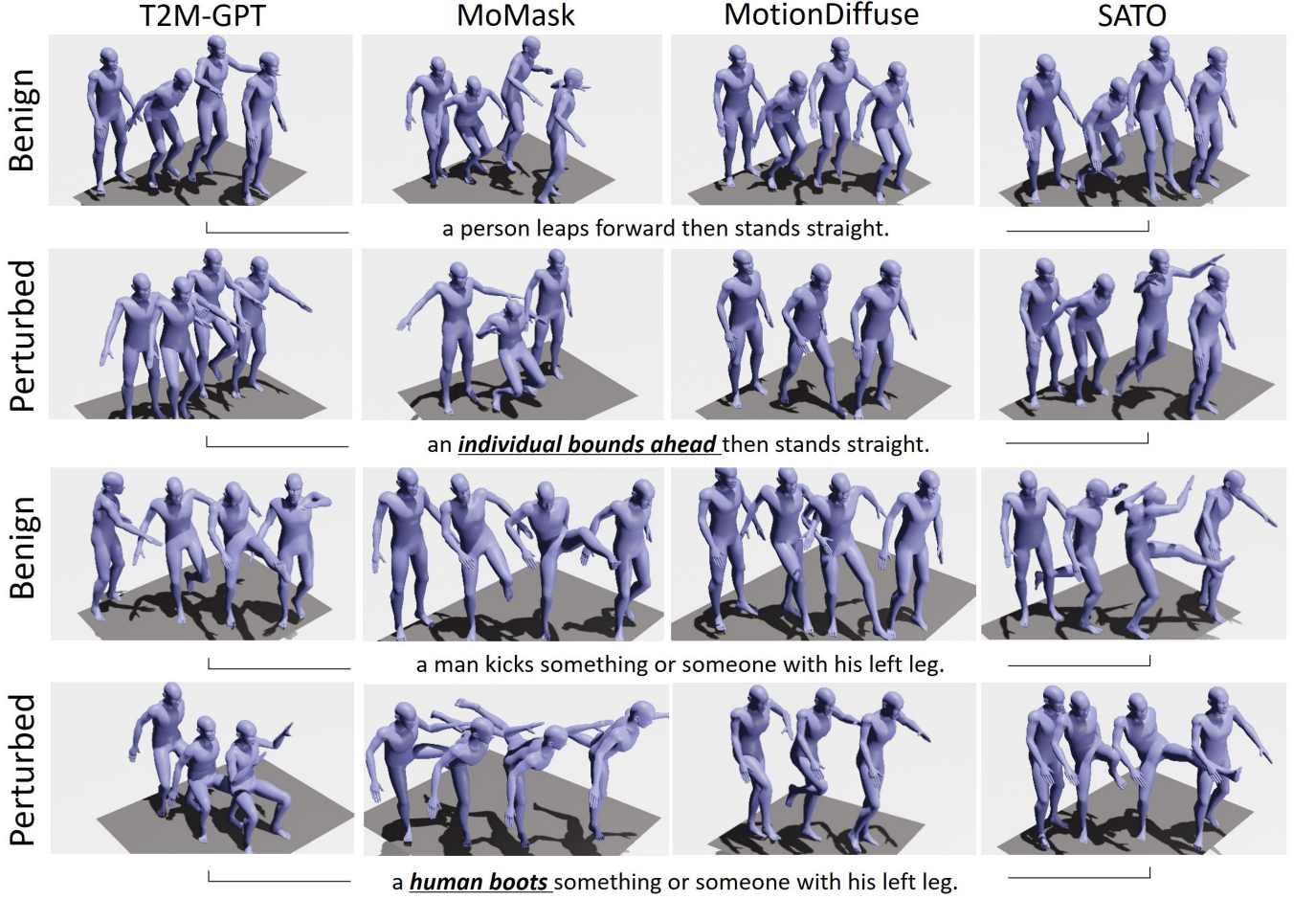
## 3 MORE VISUALIZATION EXAMPLES

**Visual Comparison between SATO and state-of-the-art approaches.** As shown in Fig. 1, we randomly select perturbed text examples from the test set and visualize the predictions obtained from model inputs before and after perturbation. In both of these examples, SATO yielded correct predictions, while the other models encountered catastrophic failure issues. Our approach demonstrates consistent outputs and exhibits good stability before and after perturbation.

**Attention visual examples.** Fig. 2 illustrates the differences in attention between SATO and the original model before and after perturbation. The specific attention calculation method can be found in Section 3.1 of the main text. Across various examples, it is evident that the Jensen-Shannon Divergence (JSD) between text attention vectors before and after perturbation is significantly lower for SATO compared to the original model. The original model exhibits attention shifts when encountering synonymous perturbations, while SATO demonstrates better stability against synonymous perturbations across multiple examples.

## 4 MORE ABLATION STUDY

**Parameters analysis.** To explore the reasonable range of parameters for the loss function, we conducted 13 experiments on the SATO (T2M-GPT) model using the HumanML3D dataset, with three different loss settings. The results are shown in Table 3. Here,  $\mathcal{L}_1$ ,  $\mathcal{L}_2$ , and  $\mathcal{L}_3$  in the text represent the same losses. By fine-tuning the parameter ranges, we observed that slight increases or decreases in all loss parameters have little impact on the overall performance of the model. This is because, during the fine-tuning process, the three

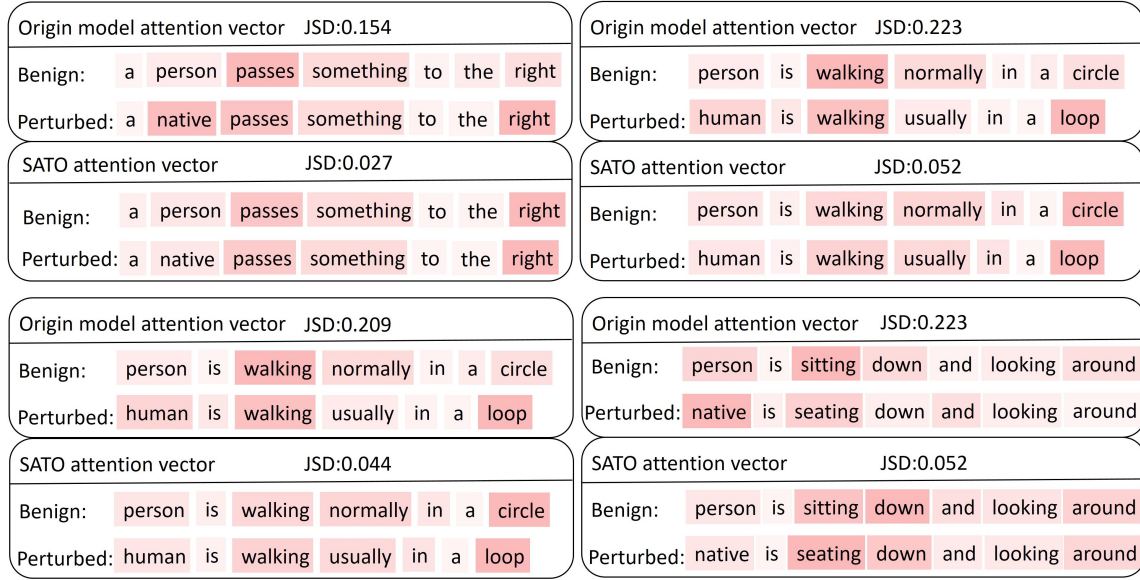


**Figure 1: Visual comparison between SATO and state-of-the-art approaches.** We compare SATO with T2M-GPT [2], MoMask [1], and MotionDiffuse [3]. We present two examples demonstrating predicted action sequences as outputs before and after perturbation. The underlined part is the part that scrambles the description. It can be observed that all models perform relatively accurately on the original text. However, only SATO predicts correctly on perturbed text. We presented additional visualization examples on our anonymous website: <https://anonymous.4open.science/api/repo/project-1FC7/file/SATO.html>

losses enable the model to dynamically balance towards stability and precision. For instance, when increasing  $\mathcal{L}_1$ , we are more likely to obtain model weights that lean towards accuracy rather than stability. Upon analyzing the detailed changes, we found that the performance improvement in terms of FID and R-Top3 is associated with an increase in  $\mathcal{L}_1$ , indicating its influence on model accuracy. On the other hand, the stability of the model is correlated with  $\mathcal{L}_2$  and  $\mathcal{L}_3$ , as reflected in the improvement of  $FID_P$  and  $FID_D$  when  $\mathcal{L}_2$  and  $\mathcal{L}_3$  are increased. Moreover, when we set large variations in the loss parameters, we observed that an excessively large  $\mathcal{L}_1$  leads to higher accuracy but poor stability, while a too large  $\mathcal{L}_3$  results in degraded performance due to excessive input perturbation during training, causing the model to lose its original good performance.

**Perturbation method ablation.** In the perturbation method section, we discussed two types of perturbation methods: PGD and

RSR. To enhance the performance of PGD, we integrated data augmentation by randomly selecting either the original text or its synonym-disturbed counterpart as input. During training, we then apply gradient-based perturbations to the selected input, generating the perturbed text. This approach differs from RSR, where the input comprises the original text, and its synonym-disturbed sentence serves as its perturbed text. Table 1 illustrates that both methods exhibited enhancements in  $FID_P$  and  $FID_D$ , with RSR showcasing superior stability. JSD highlighted the variance in text attention before and after model perturbation. We observed that both methods enhanced the stability of text attention. Furthermore, we utilized L1 to gauge the disparity in text features outputted by the text encoder. It's evident that after employing PGD or RSR, the outputted text features are significantly stabilized, which aids subsequent models in producing consistent outputs and thereby improving model stability. In this paper, we opted for the synonym



**Figure 2: Attention visual examples.** We compared the visualizations of attention vectors from the text encoder for T2M-GPT and SATO (T2M-GPT) before and after textual perturbations. In our visualizations, darker shades of red indicate higher attention weights. Additionally, we quantified the attention differences induced by perturbations using Jensen-Shannon Divergence (JSD). Our model exhibits a smaller JSD when the text is perturbed, indicating that our model possesses better attention stability.

Perturbation Method	Dataset	$FID$	$FID_P$	$FID_D$	JSD	$L_{feature}$
Without perturbation	HumanML3D	$0.141 \pm .005$	$1.754 \pm .004$	$1.443 \pm .004$	0.228	33.657
PGD		$0.246 \pm .010$	$0.316 \pm .008$	$0.030 \pm .010$	<b>0.179</b>	<b>16.743</b>
RSR		$0.157 \pm .006$	<b><math>0.155 \pm .007</math></b>	<b><math>0.021 \pm .006</math></b>	0.188	17.483

**Table 1: Perturbation method ablation.** We conducted ablation studies on SATO (T2M-GPT) using perturbation methods. "Without perturbation" refers to the original T2M-GPT model. JSD assesses the stability of the model's attention.  $L_{feature}$  represents the L1 distance of the model's output text feature before and after perturbation. We have employed two methods to perturb the input, both of which significantly enhance model attention and prediction stability.

Method	$FID$	$FID_P$	$FID_D$	JSD	Training time(h)	Inference time(s)
T2M-GPT	$0.141 \pm .005$	$1.754 \pm .004$	$1.443 \pm .004$	0.228	5.1	0.2557
Data augmentation	$0.233 \pm .008$	$0.395 \pm .013$	$0.390 \pm .008$	0.228	5.2	0.2557
SATO	$0.157 \pm .006$	<b><math>0.155 \pm .007</math></b>	<b><math>0.021 \pm .006</math></b>	<b>0.188</b>	12.6	0.2557

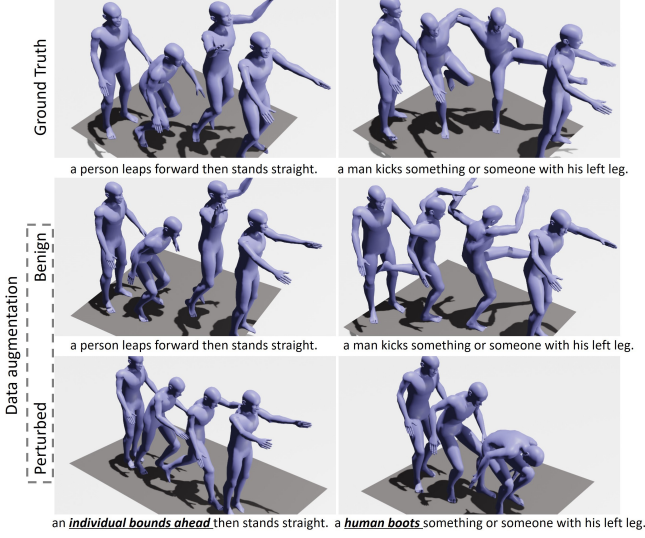
**Table 2: Comparison with Data Augmentation:** We conducted a comparison between SATO and the method of solely fine-tuning the model using data augmentation. The findings suggest that SATO exhibits superior accuracy and stability compared to relying solely on data augmentation.

replacement perturbation method, which exhibited superior stability and performance.

**Compare with data augmentation.** The instability of the Text-to-Motion model may stem from the limited diversity of vocabulary in the dataset, leading to poor generalization performance on unseen text. We conducted experiments using T2M-GPT as the base model on the HumanML3D dataset. We fine-tuned T2M-GPT using only data augmentation, where during training, we input randomly selected text either before or after synonymous perturbations, and

compared it with SATO(T2M-GPT). In Table 2, our model exhibits better stability in attention, as evidenced by a significant decrease in JSD. Additionally, our model outperforms data augmentation methods on both the original dataset and the perturbed dataset, with  $FID_P$  decreasing by 0.369, indicating better resistance to perturbations. In Fig. 1 and Fig. 3, we compare SATO (T2M-GPT) with the method of fine-tuning the original model using only data augmentation. From our randomly selected examples, we can observe that the results obtained solely through data augmentation still exhibit catastrophic errors. Combining quantification and visualization, we





**Figure 3: Visual examples of data augmentation methods.** The first line is the ground truth of the motions. The second row shows the predictions of the data augmentation model on the original description, and the third line is on the perturbed description. The underlined part is the part that scrambles the description. Despite fine-tuning the original model with data augmentation, the visual results still indicate an inability to resolve catastrophic errors stemming from synonymous perturbations.

#### Algorithm 1 SATO

**Input:** Origin pre-trained Text-to-Motion model (e.g., T2M-GPT)  $y(\cdot, \omega)$  and weight  $\mathcal{W}$ ; Training data  $D$  including text data  $x$  and  $D'$  including perturbed text  $x'$ .

Initialize  $\hat{\mathcal{W}}$  via  $\mathcal{W}$

**if** method == 'PGD' **then**

**for**  $t = 1, 2, \dots, T$  **do**

    Initialize  $\rho_k^*$ .

**for**  $n = 1, 2, \dots, N$  **do**

      Randomly sample a batch  $\mathcal{B}_n \subset D$

$\rho_k = \rho_{k-1}^* + \frac{\eta}{|\mathcal{B}_n|} \sum_{x \in \mathcal{B}_n} \nabla (D_2(y(x, \hat{\omega}), y(x, \hat{\omega} + \rho_{k-1}^*)) + \mathcal{L}_{\text{Topk}}(\omega, \hat{\omega} + \rho_{k-1}^*))$

$\rho_k^* = \underset{\|\rho\| \leq R}{\text{argmin}} \|\rho - \rho_k\|$

**end for**

    Update  $\hat{\mathcal{W}}$  using Stochastic Gradient Descent, where  $C_t$  is a batch,  $\mathcal{L}_{\text{trans}}$  is the loss of origin model

$\hat{\mathcal{W}}_t = \hat{\mathcal{W}}_{t-1} - \eta_t \sum_{x \in C_t} [\mathcal{L}_{\text{trans}} + \lambda_1 D_2(\hat{y}(x, \hat{\omega}), y(x, \omega)) + \lambda_2 \mathcal{L}_{\text{Topk}}(\hat{\omega}, \hat{\omega} + \rho^*) + \lambda_3 D_1(\hat{y}(x, \hat{\omega}), \hat{y}(x, \hat{\omega} + \rho^*))]$

**end for**

**else if** method == 'RSR' **then**

  We get  $\hat{\omega}$  from input  $x'$

**for**  $t = 1, 2, \dots, T$  **do**

    Update  $\hat{\mathcal{W}}$  using Stochastic Gradient Descent, where  $C_t$  is a batch,  $\mathcal{L}_{\text{trans}}$  is the loss of origin model

$\hat{\mathcal{W}}_t = \hat{\mathcal{W}}_{t-1} - \eta_t \sum_{x \in C_t} [\mathcal{L}_{\text{trans}} + \lambda_1 D_2(\hat{y}(x, \hat{\omega}), y(x, \omega)) + \lambda_2 \mathcal{L}_{\text{Topk}}(\hat{\omega}, \hat{\omega}) + \lambda_3 D_1(\hat{y}(x, \hat{\omega}), \hat{y}(x', \hat{\omega}))]$

**end for**

**end if**

**return**  $\hat{\mathcal{W}}^* = \hat{\mathcal{W}}_T$

can conclude that the instability observed in the Text-to-Motion model does not solely stem from dataset limitations but also from

attention instability. Consequently, solely relying on data augmentation is insufficient for mitigating the catastrophic errors induced by input perturbations.

**Attention analysis.** In the preceding sections, we employed JSD analysis to evaluate the stability of text encoders in SATO post fine-tuning, juxtaposing them against the original model's text encoder. The results indicate SATO achieving the best JSD score (0.228 vs. 0.188). A noteworthy distinction between SATO and data augmentation lies in our adoption of a stable attention mechanism. While data augmentation falls short in stability metrics and visualization compared to SATO, this underscores the crucial role of the attention stability module in mitigating catastrophic model errors stemming from synonymous perturbations. Moreover, we observed a close correlation between the stability of outputted text features and attention stability. This suggests that SATO's resilience to perturbations in text encoder attention stabilizes the outputted text features, thereby ensuring more consistent predictive outcomes in subsequent transformer structures.

## 5 DETAILS OF HUMAN EVALUATION

**[Question1]:** Please evaluate the quality of the motion generation below.<motion1.gif>

- (1) Completely accurate semantically, with smooth and correct motion.
- (2) Generates well with minor details.
- (3) Some errors in detail, but overall correct.
- (4) Poor, mostly incorrect.
- (5) Very poor, completely incorrect semantically.

**[Question2]:** Please evaluate the quality of the motion generation below.<motion2.gif>

- (1) Completely accurate semantically, with smooth and correct motion.
- (2) Generates well with minor details.
- (3) Some errors in detail, but overall correct.
- (4) Poor, mostly incorrect.
- (5) Very poor, completely incorrect semantically.

**[Question2]:** Which motion result do you think is better?

- (1) The first one
- (2) The second one

We've employed the Google Form platform to enable 35 individuals to fill out multiple motion sequence tests independently. In total, there are 1200 questionnaires distributed. Our questionnaire design includes two types of questions. The first type involves directly rating the quality of generated motion. Motion is presented in GIF format, accompanied by five evaluation options: "Good" signifies generally well-generated motion with minor details; "Fair" indicates errors in details but overall correctness; "Poor" denotes overall incorrectness; and "Very poor" signifies motions and text that cannot be matched at all. We believe the first three options represent correctly generated postures, while the latter two represent errors. The second type of question pertains to user preferences between our model and a baseline model. This question compares our method and the original method from the user's perspective regarding motion generation accuracy.

$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$FID\downarrow$	$FID_P\downarrow$	$FID_D\downarrow$	R-Top3 $\uparrow$
0.1	0.01	0.2	$0.200^{\pm.008}$	$0.256^{\pm.011}$	$0.035^{\pm.008}$	$0.728^{\pm.003}$
0.1	0.1	0.2	$0.197^{\pm.007}$	$0.224^{\pm.009}$	$0.031^{\pm.007}$	$0.727^{\pm.002}$
0.1	0.005	0.2	$0.198^{\pm.008}$	$0.203^{\pm.010}$	$0.023^{\pm.009}$	$0.732^{\pm.003}$
0.1	0.5	0.2	$0.179^{\pm.006}$	$0.232^{\pm.011}$	$0.073^{\pm.006}$	$0.746^{\pm.003}$
0.01	0.05	0.2	$0.197^{\pm.011}$	$0.233^{\pm.010}$	$0.023^{\pm.012}$	$0.723^{\pm.003}$
0.05	0.05	0.2	$0.220^{\pm.010}$	$0.263^{\pm.011}$	$0.033^{\pm.010}$	$0.722^{\pm.003}$
0.2	0.05	0.2	$0.190^{\pm.008}$	$0.222^{\pm.008}$	$0.036^{\pm.008}$	$0.735^{\pm.002}$
1.0	0.05	0.2	$0.169^{\pm.007}$	$0.269^{\pm.012}$	$0.190^{\pm.007}$	<b><math>0.759^{\pm.002}</math></b>
0.1	0.05	0.02	$0.183^{\pm.005}$	$0.200^{\pm.026}$	$0.026^{\pm.005}$	$0.732^{\pm.003}$
0.1	0.05	0.1	$0.198^{\pm.008}$	$0.211^{\pm.011}$	$0.029^{\pm.008}$	$0.732^{\pm.003}$
0.1	0.05	0.3	$0.212^{\pm.011}$	$0.239^{\pm.007}$	$0.031^{\pm.011}$	$0.729^{\pm.002}$
0.1	0.05	2.0	$0.730^{\pm.022}$	$1.175^{\pm.029}$	$0.171^{\pm.022}$	$0.665^{\pm.003}$
0.1	0.05	0.2	<b><math>0.157^{\pm.006}</math></b>	<b><math>0.155^{\pm.007}</math></b>	<b><math>0.021^{\pm.006}</math></b>	$0.738^{\pm.003}$

**Table 3: Parameter analysis.**  $\pm$  indicates a 95% confidence interval. R-top3 represents R-Precision Top3. The table displays the results of three different parameters for loss.

Notation	Remark	Notation	Remark
$x$	input data	$V_k$	top-k vector overlap ratio
$\omega, \tilde{\omega}$	attention vector, SATO attention vector	$D$	divergence metric
$\tilde{\omega}$	perturbed attention vector	$\mathcal{W}$	weight of text-to-motion model
$X$	a pose sequence	$y, \tilde{y}$	prediction of text-to-motion model and SATO
$C$	a textual description	$\gamma_1, \gamma_2, R$	parameters in SATO
$c_i$	$i^{th}$ word in the sentence	$\rho$	some perturbation
$\mathcal{L}_{trans}$	the loss of text-to-motion model	$\mathcal{L}_{Topk}$	a surrogate loss of $V_k$
$r_k$	PGD step size	$\zeta_k^\omega$	top-k indices set of vector $\omega$
$e$	text embedding	$\lambda_1, \lambda_2, \lambda_3$	regularization parameters
$t$	token index vector	$e$	embedding weights
$t_e$	text embedding vector	$k$	key vector
$q$	query vector	$\omega_t$	attention weights

**Table 4: Symbolic representation and remarks for the notation used in this paper.**

## 6 SATO PSEUDO-ALGORITHM

The pseudo-algorithm for SATO is outlined in Algorithm 1.

## 7 COMPUTATIONAL COMPLEXITY

During training, SATO incurs additional time due to the utilization of an extra frozen teacher model and the generation of predictions before and after output perturbation. Table 2 indicates that under the same experimental conditions (RTX4090-24G GPU), SATO (T2M-GPT) takes an additional 7.4 hours compared to T2M-GPT over 100,000 iterations. However, during the inference process, since SATO fine-tunes the original model without increasing the parameter count, it does not incur any additional time or space overhead. Table 2 also confirms this. When we use all the data from HumanML3D as input with a batch size of 1, we obtain an average inference time of 0.2557 second per text.

## 8 SYMBOLIC REPRESENTATION

A table providing the symbolic representation employed throughout this paper is presented in Table 4. Each symbol is defined alongside its respective notation and meaning.

## REFERENCES

- [1] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2023. MoMask: Generative Masked Modeling of 3D Human Motions. arXiv:2312.00063 [cs.CV]
- [2] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. arXiv:2301.06052 [cs.CV]
- [3] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. arXiv:2208.15001 [cs.CV]