You Think, You ACT: The New Task of Arbitrary Text to Motion Generation

Supplementary Material

8. Overview

In this supplementary material, we provide the following items for a better understanding of our main paper.

- Dataset construction
  1. Comparison of different prompts [Sec 9.1]
  2. Data evaluation [Sec 9.2]
  3. Data Examples [Sec 9.3]
- More Implementation Details
  1. More Implementation Details [Sec 10.1]
  2. Prompts in inference [Sec 10.2]
- Experiment
  1. Details of Action Text to Motion Metrics [Sec 11.1]
  2. Explanation on our new Metrics [Sec 11.2]
  3. User study Results [Sec 11.3]
  4. Generate multiple motion results under Scene Text [Sec 11.4]
  5. Action Texts to Motion Results [Sec 11.5]
  6. Scene Texts to Motion Results [Sec 11.6]
  7. Hyper-parameters evaluation of Think Model [Sec 11.7]
  8. Evaluation of Think Model prompt [Sec 11.8]
- Network Architectures
  1. VQ-VAE Architecture [Sec 12.1]

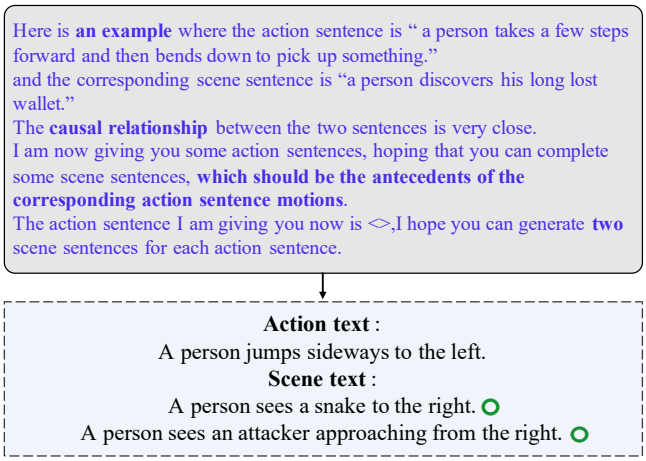
9. Dataset construction

9.1. Comparison of different prompts

In figures 8, 9, 10, 11, and 12 , we present five types of prompts and their corresponding results used during the data construction process. In each example image, the top section displays the prompt used, while the bottom section shows the action text alongside the generated scene text corresponding to the given prompt. Suitable results are marked with a green circle, while unsuitable results are marked with a red cross.

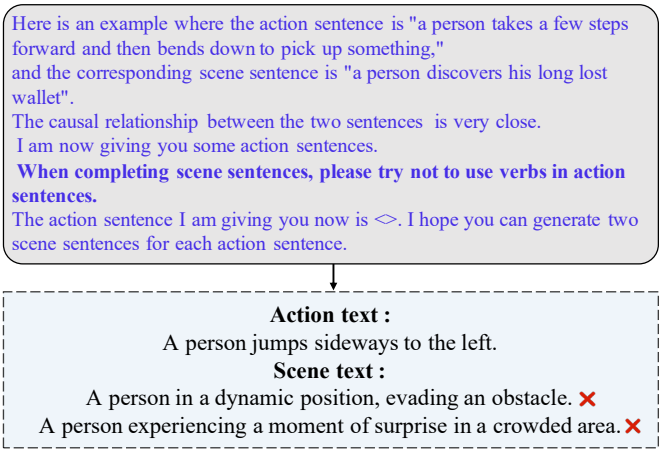
9.2. Data evaluation

Figure 13 shows the results of user evaluations conducted during the first data validation. A total of 20 participants are randomly assigned 100 data samples each and classify the samples into two categories: "reasonable" and "unreasonable." Overall, the number of samples classified as "reasonable" consistently remains high (ranging from 89 to 100), while the number of "unreasonable" samples remains low (ranging from 2 to 11). This indicates that the majority of the data is deemed reasonable by the participants, supporting the quality and reliability of the data processing. The number of "reasonable" samples across participants is relatively concentrated, fluctuating within the range of 89 to 98, which suggests a high degree of consistency in the evaluation standards applied by different participants.



(a) The final selected prompt

Figure 8. (a) represents the final prompt we adopted, which incorporates both contextual example guidance and control over the quantity of generated results. This prompt effectively balances semantic consistency and the desired output quantity.



(b) Limit the use of verbs in prompts

Figure 9. (b) illustrates the results when the prompt includes restrictions on the use of verbs contained within the action text. During the generation process, we observed that this prompt often restricts the use of more verbs (not limited to those within the action text provided), thereby limiting the expressiveness of the scene text and adversely affecting the generated results.

9.3. Data Examples

We adopt the same annotation structure as HUMANML3D, where each line represents a distinct textual annotation con-

Here are some events, and I hope you can summarize in one sentence what happened that could have caused such a reaction.
For example, the action sentence is "a person takes a few steps forward and then bends down to pick up something", and the corresponding scene sentence is "a person discovers his long lost wallet".
<miss causal relationship description>
I am now giving you some action sentences, hoping that you can complete some scene sentences.
The action sentence I am giving you now is <>. I hope you can generate two scene sentences for each action sentence.

Action text :
A person jumps sideways to the left.
Scene text :
A person dodges an incoming object. ✗
A person narrowly avoids an obstacle. ✗

(c) Prompt without causal relationship description

Figure 10. (c) presents the results of our prompt under different descriptions of the relationship between scene text and action text sentences. We observed that when the prompt does not explicitly describe a causal relationship between the two, the quality of the generated results tends to be imprecise, and unsuitable.

Here is an example where the action sentence is "a person takes a few steps forward and then bends down to pick up something."
and the corresponding scene sentence is "a person discovers his long lost wallet."
The causal relationship between the two sentences is very close. I am now giving you some action sentences, **hoping that you can complete some scene sentences**, which should be the antecedents of the corresponding action sentence actions.
The action sentence I am giving you now is <>

Action text :
A person jumps sideways to the left.
Scene text :
A person sees a snake to the right. ○
A person hears a loud crazy noise on their right. ○
A person sees a child running across their path. ✗
A person notices a dog running towards them . ✗

(d) Prompt with no limit on quantity

Figure 11. (d) demonstrates the results when the prompt does not specify the number of scene text to be generated. We observed that the absence of a limit on the number of scene texts generated for each action text can result in some scene texts being misaligned with the corresponding motion in the generated results. Furthermore, when the number of scene text sentences exceeds four, the generated results often include undesirable outputs.

sisting of four parts: the original description (in lowercase), the processed sentence, the start time (in seconds), and the end time (in seconds), all separated by "#". For the processed sentence, we use Spacy for tokenization and POS tagging. Some examples of our HUMANML3D++ are shown in Figure 14.

<miss examples>
The causal relationship between action sentence and scene sentence is very close.
I am now giving you some action sentences, hoping that you can complete some scene sentences, which should be the antecedents of the corresponding action sentence actions.
The action sentence I am giving you now is <>,I hope you can generate two sentences for each action sentence.

Action text :
A person jumps sideways to the left.
Scene text :
The ground beneath him is covered in soft grass, providing a perfect landing for his next move. ✗
A group of friends nearby are playing a game, cheering as he prepares for his jump. ✗

(e) Prompt without examples

Figure 12. (e) illustrates the results when the prompt does not include action text and corresponding scene text examples. We found that without specific contextual example guidance, it is challenging to generate scene text that aligns with the action text based solely on linguistic relationship descriptions.

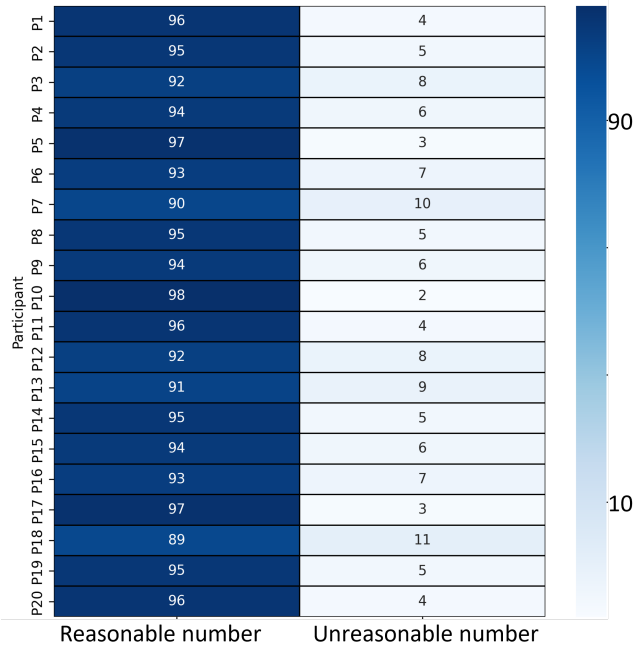
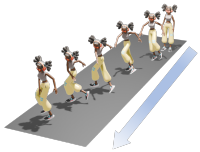


Figure 13. The results of user evaluation of data validation. The majority of scene texts are perceived by users as being well-aligned with the corresponding motions.

Scene Text:
a man sets a coin on the ground.#a/DET man/NOUN sets/VERB a/DET coin/NOUN on/ADP the/DET ground/NOUN #0.0#0.0
A person is searching for something they lost.#A/DET person/NOUN is/AUX searching/VERB for/ADP something/PRON they/PRON lost/VERB #0.0#0.0
Action Text:
a person standing raises their arm and then bends over.#a/DET person/NOUN stand/VERB raise/VERB their/DET arm/NOUN and/CCONJ then/ADV bend/VERB over/ADP#0.0#0.0
someone is holding their right arm next to their head. then bends over and touches the ground.#someone/PRON is/AUX hold/VERB their/DET right/ADJ arm/NOUN next/ADV to/ADP their/DET head/NOUN then/ADV bend/VERB over/ADP and/CCONJ touch/VERB the/DET ground/NOUN#0.0#0.0



Scene Text:
A man is walking through a crowded street when someone unexpectedly bumps into him from right.#A/DET man/NOUN is/AUX walking/VERB through/ADP a/DET crowded/ADJ street/NOUN when/SCONJ someone/PRON unexpectedly/ADV bumps/VERB into/ADP him/PRON from/ADP right/NOUN#0.0#0.0
A man is in a busy store, and someone rushes past him in a hurry.#A/DET man/NOUN is/AUX in/ADP a/DET busy/ADJ store/NOUN and/CCONJ someone/PRON rushes/VERB past/ADP him/PRON in/ADP a/DET hurry/NOUN#0.0#0.0

Action Text:
a man stumbles to his right. the motion seems surprised so he was probably pushed.#a/DET man/NOUN stumble/VERB to/ADP his/DET right/NOUN the/DET motion/NOUN seem/VERB surprised/ADJ so/SCONJ he/PRON was/AUX probably/ADV push/VERB#0.0#0.0
a person stumbles to the right and recovers their balance.#a/DET person/NOUN stumble/VERB to/ADP the/DET right/NOUN and/CCONJ recover/NOUN their/DET balance/NOUN#0.0#0.0

Scene Text:
When a person is running in a hurry to save time, there is a package on the ground to the right that blocks the way. #When/SCONJ a/DET person/NOUN is/AUX running/VERB in/ADP a/DET hurry/NOUN to/ADP save/VERB time/NOUN,/PUNCT there/PRON is/AUX a/DET package/NOUN on/ADP the/DET ground/NOUN to/ADP the/DET right/NOUN that/PRON blocks/VERB the/DET way/NOUN #0.0#0.0
a person is frustrated with something or someone.#a/DET person/NOUN is/AUX frustrated/ADJ with/ADP something/PRON or/CCONJ someone/PRON #0.0#0.0

Action Text:
a person kicks something with their right foot.#a/DET person/NOUN kick/VERB something/PRON with/ADP their/DET right/ADJ foot/NOUN#0.0#0.0
a man kicks something from the ground with his right leg.#a/DET man/NOUN kick/VERB something/PRON from/ADP the/DET ground/NOUN with/ADP his/DET right/ADJ leg/NOUN#0.0#0.0

Figure 14. Dataset examples of HUMANML3D++.

10. More Implementation Details

10.1. More Implementation Details

The codebook is sized at 512×512, with a downsampling rate l of 4. Training is conducted in two phases: the first 200K iterations use a learning rate of 2×10^{-4}, followed by 100K iterations at 1×10^{-5}. The VQ loss $\mathcal{L}_{vq}$ and reconstruction loss $\mathcal{L}_{re}$ are weighted with $\beta = 1$ and $\alpha = 0.5$, respectively. Following the approach of Guo et al., the maximum motion length is set to 196 for both HUMANML3D++ and HUMANML3D [9] in ACT model.

10.2. Prompts in inference

We employ the following prompt to infer motions from Scene Text in the Scene Text to Motion task.

	Description
⟨Task Introduction⟩	I will provide you with a **Scene Text**, which describes a scene or an event. Your task is to generate the corresponding **Action Text**.
⟨Relationship Explanation⟩	The causal relationship between the two sentences should be strong, where the **Scene Text** serves as the antecedent of the corresponding **Action Text**.
⟨Example⟩	• **Scene Text**: A man sees a coin on the ground. • **Corresponding Action Text**: The man walks forward and bends down to pick something up.
⟨Task Requirements⟩	Your task is to comprehensively consider factors such as the character and location to output all possible reasonable results. Instead of outputting an **Action Text**, extract the **Action Instruction** from it.
⟨Output Format⟩	• (1) Only output **Action Instructions** without any additional reasoning process. • (2) Each possible result should be on a separate line, following the format like this: `(walk forward, bend down).`

11. Experiment

11.1. Details of Action Text to Motion Metrics

We provide the commonly used metric calculation metrics for the Action Texts to Motion domain. Specifically, FID evaluates the distribution distance between generated motion and ground truth, R-Precision and MM-Dist measure the consistency between Action Texts and generated motion, Diversity assesses the diversity of the entire set of generated motions, and MModality examines the diversity of motions generated from the same action text. Detailed as follows:

- **R-Precision**: Given one motion sequence and 32 text descriptions (1 ground-truth and 31 randomly selected mismatched descriptions), we rank the Euclidean distances between the motion and text embeddings. Top-1, Top-2 and Top-3 accuracy of motion-to-text retrieval are reported.
- **Frechet Inception Distance (FID)**: We calculate the distribution distance between the generated and real motion using FID on the extracted motion features.
- **Multimodal Distance (MM-Dist)**: The average Euclidean distances between each text feature and the generated motion feature from this text.
- **Diversity**: From a set of motions, we randomly sample 300 pairs of motion. We extract motion features and compute the average Euclidean distances of the pairs to measure motion diversity in the set.
- **Multimodality (MModality)**: For one text description, we generate 20 motion sequences forming ten pairs of motion. We extract motion features and compute the average Euclidean distances of the pairs. We finally report the average over all the text descriptions.

11.2. Analysis on our new Metrics

As shown in Figure 15, Arbitrary Text to Motion is more challenging and flexible compared to Action Text to Motion (Action Label to Motion). Action Text (Label) to Motion relies on explicit action labels or specific action-descriptive texts and is deterministic in nature, focusing on generating a single action pattern from a given action text. In contrast, our task expands to more general scene text inputs, which may not contain explicit action labels. This is a multi-solution task, where multiple plausible motion patterns can be generated from a single scene text. Due to this fundamental distinction, previous metrics that required generated results to strictly match ground truth are incompatible with our task. Our new metrics allow the evaluation of multiple generated results rather than focusing solely on a "single correct" answer. They are well-suited to the characteristics of multi-solution tasks and address the limitations of the original evaluation framework.

Figure 16 illustrates the impact of different numbers N of generated results on Hit Accuracy. As N increases, the model is able to generate a greater variety of potential motions, which are more likely to include results similar to the motion in the dataset, naturally increasing the probability of hitting. The upward trend further demonstrates that the metric effectively reflects the model's ability to capture multiple plausible solutions.

Evaluation Consistency. In the supplementary materials, we provide additional visualization results and user studies to further validate the consistency between our metric evaluations and subjective assessments. For example, the evaluation results indicate that TAAT outperforms T2M-GPT [52], which is consistent with the visualization results and user study findings. Additionally, the evaluation results show

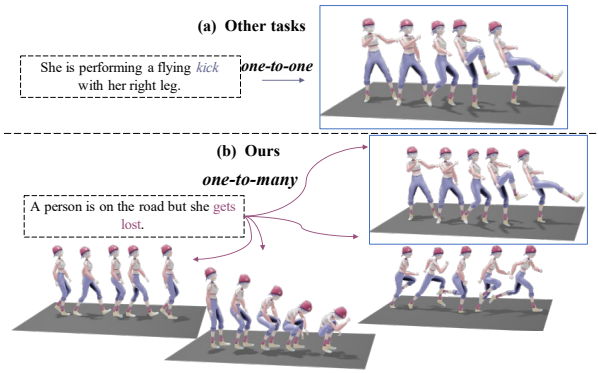


Figure 15. Our task is fundamentally different from previous tasks. We are a multi-solution task, and even if the generated results are inconsistent with the GT in the dataset, as long as they match the scene text, they are reasonable.

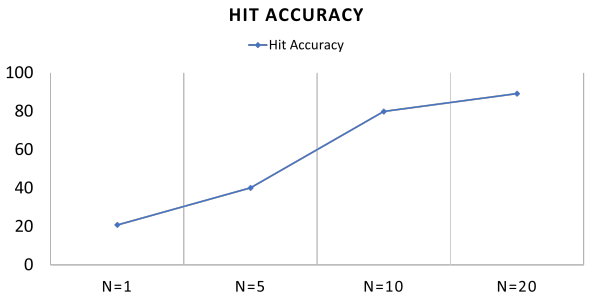


Figure 16. The impact of different numbers of generated results on Hit Accuracy.

that the performance ranking of the models is TAAT, T2M-GPT [52], MDM [43], and MLD [3], which aligns with the visualization results.

11.3. User study results

Figure 17 presents the results of our user study. We visually assessed each method's performance on 100 Scene Texts and 100 Action Texts, with independent evaluations provided by 30 participants across 5 groups. Participants rated the motion results based on their appropriateness and consistency with the provided textual information, rating them as suitable, acceptable, or unsuitable. The results reveal that our model's visual effects are superior to those of alternative methods, consistently generating realistic motions that align well with human perceptual cognition.

11.4. Generate multiple motion results under Scene Text

Our model effectively understands Scene Text and can generate multiple reasonable results for the same Scene Text. Figure 18 illustrates multiple motion results generated under the same Scene Text condition.

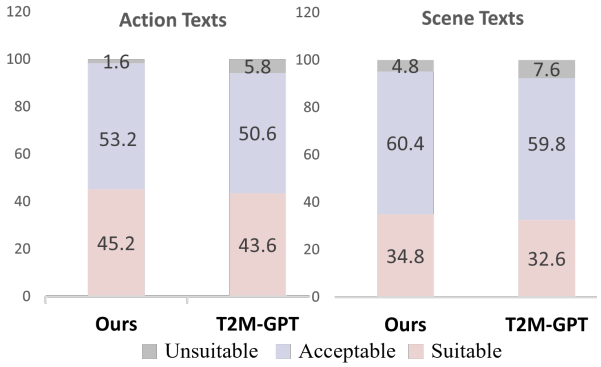


Figure 17. The user study of our model and T2M-GPT [52] on Action Texts and Scene Texts. Our model generates motions that are both realistic and aligned with human cognition.

11.5. Action Texts Results

In experiments conducted on Action Texts (in figure 19 and figure 20), our model demonstrates the capability to sequentially generate all actions as stipulated by the text. In contrast, alternative models fail to generate all actions and exhibit inaccuracies in the sequencing of actions. The Action Texts: "A person jumps first, then walks forward, then sits down, and finally starts running." contains four actions (in figure 19). MDM [43] only performs walk and run actions, while MLD [3] executes run, sit, and walk, however, the order of actions is disordered. MotionDiffuse [53] completes the walk and jump actions but omits two actions, and the sequence is disordered. T2M-GPT [52] only generates walk and jump actions, with the sequence being disordered. Only our approach successfully performs all four actions in sequence.

The Action Texts "A person bends down first, then walks, and finally jumps up." contains three actions (in figure 20). MDM [43] generates actions that all involve jumping forward, which do not meet our requirements. MLD [3] produces bend-down and walk actions, but the sequence is disordered. MotionDiffuse [54] completes the walk action but omits two actions, and the sequence is disordered. T2M-GPT [52] only generates the jump action. Only our approach successfully performs all three actions in sequence.

11.6. Scene Texts Results

In experiments conducted on Scene Texts (in figure 21 and figure 22), our model demonstrates strong capability in comprehending the textual context and generating corresponding actions, whereas alternative models display poor performance in this regard. In the Scene Text "The cleaner saw someone throwing garbage in the park" (in figure 21) , when faced with the textual description, MDM [43], MotionDiffuse [53], and T2M-GPT [52] all generate the action "throw"as described in the text, while MLD [3] generates the action of looking around in place. Our model demon-

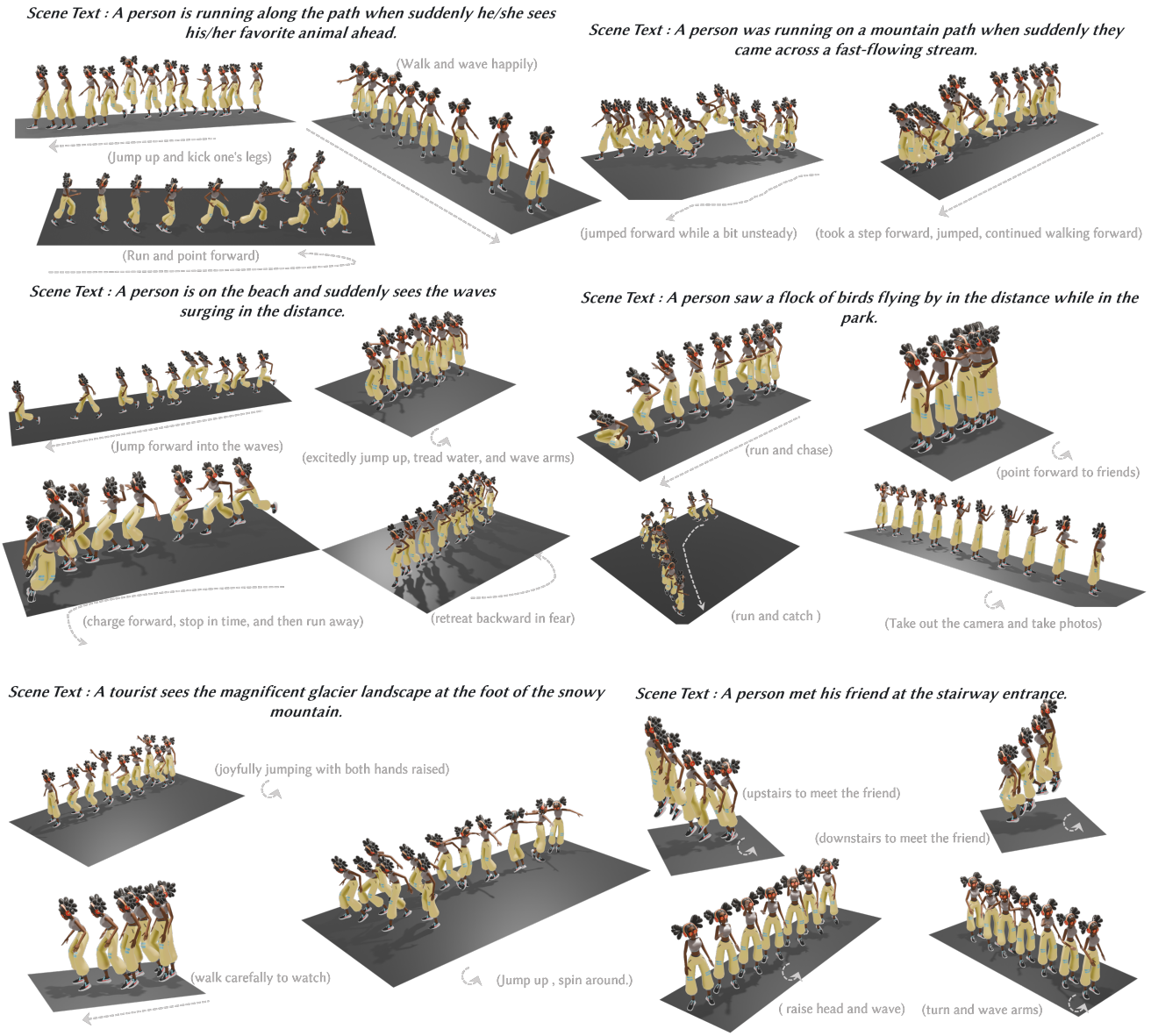


Figure 18. Our TAAT is capable of generating multiple reasonable motions from a single scene text. In addition to that, we have provided additional video visualization results.

strates a comprehensive understanding of the text, generating the appropriate response action for a cleaner in this scenario - running forward, bending over, and cleaning up the garbage.

In the Scene Text "A person on the road with a motorcycle approaching in the distance." (in figure 22) , when presented with the textual description, MLD [3] primarily consist of repetitive hand movements near the face, while the posture of other parts of the body lacks a credible response to danger. MDM [43] depicts squatting movements, which do not correspond to the scenario of an approaching vehicle. T2M-GPT [52] consist of repetitive forward arm extensions, appearing mechanical. Each model's response actions are unreasonable for the given Scene Texts. Our model(TAAT) demonstrates dynamic body coordination during running, including arm swings and leg propulsion, which align with the natural logic of human locomotion. It exhibits behavior consistent with "escaping" or "avoiding" with the natural arm movements and forward-leaning posture further enhancing the impression of "rapid movement," reflecting a reasonable response to perceived danger.

11.7. Hyper-parameters evaluation of Think Model

For the two key hyperparameters of LoRA, r and α, a smaller r indicates fewer parameters, while α controls the scaling factor of the output from LoRA's fully connected layer. We adopt the same parameter setting as [55], ensuring a scaling factor of $\alpha/r = 2$ and setting $r = 8$. This configuration aligns with the commonly used parameters in [55], and has been shown to be effective in subsequent small-scale tasks related to [55], while also reducing resource consumption.

We further examined the impact of batch size on model performance in Table 8. The results indicate that as the batch size increases, Hit Accuracy gradually improves, FID decreases, and Diversity exhibits slight fluctuations across different batch sizes. Specifically, when the batch size is set to 16, Hit Accuracy reaches 79.9%, FID decreases to 0.379, and Diversity is 8.950, demonstrating superior performance compared to other batch size settings.

Batch Size	Hit Accuracy↑	FID↓	Diversity↑
4	70.8	0.498	8.867
8	73.5	0.437	9.106
16	79.9	0.379	8.950

Table 8. Assessment of Results under Different Batch Sizes.

11.8. Evaluation of Think Model prompt

We present the ablation results of different parts of the prompt, where the Prompt Struct on the left is consistent with Sec 10.2. Under the condition of a standardized output format, the results are generally acceptable in most cases. However, in a few instances, unreasonable action instructions are observed (w/o Example, that is, Prompt Struct 2), or incomplete outputs are generated (w/o Task Requirements, that is, Prompt Struct 4). Moreover, when the output format is not standardized, a significant amount of invalid information is produced (w/o Output Format, that is, Prompt Struct 3). Overall, the prompt structure that effectively mitigates these issues (Prompt Struct 1) demonstrates better performance.

Prompt Struct 1:	**Example Output:**
Task Introduction	(upstairs to meet the friend)
Relationship Explanation	(downstairs to meet the friend)
Example	(jump, wave hands)
Task Requirements	(raise head, wave)
Output Format(1)	(turn, wave arms)
Output Format(2)	
Prompt Struct 2 :	**Example Output:**
Task Introduction	(Smile)
Relationship Explanation	(Wave, Extend hand)
Task Requirements	(met his friend)
Output Format(1)	
Output Format(2)	
Prompt Struct 3 (w/o :	**Example Output:**
Task Introduction	The output is:
Relationship Explanation	(upstairs to meet the friend)
Example	(turn, wave arms)
Task Requirements	These actions usually vary based on personal habits and their relationship.
Output Format(2)	
Prompt Struct 4:	**Example Output:**
Task Introduction	(raise head, wave)
Relationship Explanation	(upstairs to meet the friend)
Example	(turn, wave arms)
Output Format(1)	
Output Format(2)	
Prompt Struct 5:	**Example Output:**
Task Introduction	(raise head, wave)
Example	(turn, wave arms)
Task Requirements	walk
Output Format(1)	(upstairs)
Output Format(2)	

Action Text: A person jumps first, then walks forward, then sits down, and finally starts running

MLD

(a) MLD [3] does not strictly follow the action sequence described in the text. While the character performs a walking action after jumping, it does not sit down as instructed, but instead transitions directly into running. The sitting action is entirely missing. The character's running motion appears somewhat rigid, lacking the expected fluidity and sense of speed.

MDM

(b) MDM [43] does not fully follow the action sequence described in the text. After jumping, the character does not perform the walking action, and the sitting motion in the subsequent steps is unclear, followed by a direct transition into running.

MotionDiffuse

(c) MotionDiffuse [53] does not fully follow the action sequence described in the text. While the character performs the walking action after jumping, it skips the subsequent sitting action entirely, and the running motion also lacks realism.

T2M-GPT

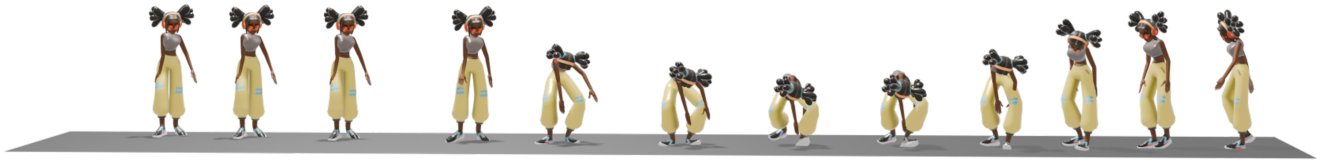
(d) T2M-GPT [52] does perform the walking action after jumping, but it does not accurately complete the sitting action.

Ours

(e) Our TAAT generates actions that accurately follow the sequence described in the text. The character first performs a jumping action, then walks forward, sits down, and finally starts running. Each action aligns with the textual requirements, demonstrating a high level of correspondence.

Figure 19. Comparison of motions generated by different methods for action text "A person jumps first, then walks forward, then sits down, and finally starts running."

Action Text: A person bends down first, then walks, and finally jumps up



MLD

(a) MLD [3] fails to execute the jumping action correctly, and the sequence of bend down and walk actions is incorrectly ordered.

MDM

(b) MDM [43] does not accurately follow the described sequence of actions. The character begins to jump immediately after bending down, rather than performing the walking action before jumping.

MotionDiffuse

(c) MotionDiffuse [53] does not fully adhere to the action sequence described in the text. While the character does perform the bending and walking actions, it fails to complete the jumping action. Additionally, the bending motion appears somewhat rigid.

T2M-GPT

(d) T2M-GPT [52] does not strictly follow the action sequence described in the text, transitioning directly from standing to jumping, and then into walking without performing the bending action.

Ours

(e) Our TAAT accurately executes the actions in the sequence described by the text. The character first bends down, then begins to walk and finally jumps. The transitions between each action are natural.

Figure 20. Comparison of different methods for executing the action text "A person bends down first, then walks, and finally jumps up."

MLD

(a) The actions generated by MLD [3] appear to be simple standing and walking motions, lacking overall coherence, and displaying disjointed movements, such as bending, turning, and walking.

MDM

(b) The actions generated by MDM [43] resemble throwing something away, directly reflecting the text content, but they lack an appropriate response to the scenario, resulting in a low alignment with the Scene Texts.

MotionDiffuse

(c) The actions generated by MotionDiffuse [53] appear very bland, primarily consisting of standing and slight hand movements, resulting in a low alignment with the Scene Texts.

T2M-GPT

(d) The actions generated by T2M-GPT [52] primarily consist of stepping forward and throwing, which may not be entirely realistic as reactive motions.

Ours

(e) Our method generates actions that depict the process of standing, observing, and gradually bending down to pick up the garbage, aligning well with the scenario described in the Scene Texts.

Figure 21. Comparison of different methods for generating actions based on Scene Texts "The cleaner saw someone throwing garbage in the park."

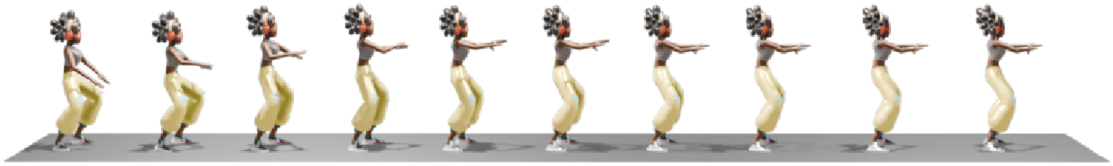
Scene Text: A person on the road with a motorcycle approaching in the distance.

MLD

(a) The actions generated by MLD [3] primarily consist of repetitive hand movements near the face, while the posture of other parts of the body lacks a credible response to danger. This does not fully align with the context implied by the scene text.

MDM

(b) The actions generated by MDM [43] primarily depict squatting movements, which do not correspond to the scenario of an approaching vehicle.



T2M-GPT

(c) The actions generated by T2M-GPT [52] consist of repetitive forward arm extensions, appearing mechanical and lacking an appropriate response to the context implied by the scene text.

Ours

(d) Our TAAT demonstrates dynamic coordination of the body during running, including arm swings and leg propulsion, aligning with the natural logic of human locomotion. It exhibits behavior consistent with "escaping" or "avoiding," with the natural arm movements and forward-leaning posture further enhancing the impression of "rapid movement," reflecting a reasonable response to perceived danger.

Figure 22. Comparison of motions generated by different methods in response to Scene text "a person on the road with a motorcycle approaching in the distance."

12. Network Architecture

12.1. Architecture of VQ-VAE

Figure 23 illustrates the architecture of our motion VQ-VAE [44]. The left side illustrates the overall structure of the VQ-VAE, which comprises an encoder (E), a quantizer, and a decoder (D). The quantizer employs a codebook to map the input data to discrete vector representations. The right side provides a more detailed view of the encoder and decoder structures. The encoder consists of multiple 1D convolutional layers (Conv1D) and residual blocks (ResBlock), processed initially through a ReLU activation function, followed by downsampling in certain layers using convolution operations with a stride of 2. The decoder mirrors the encoder's structure but with operations in reverse order.

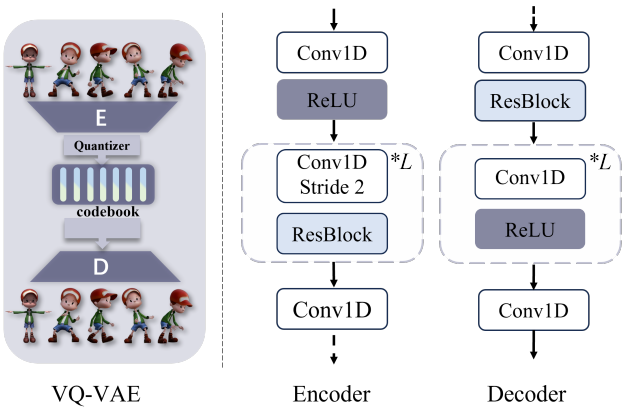


Figure 23. Architecture of our motion VQ-VAE.