Open the Motion Door: Atomic Motion Decomposition and Recomposition for Open-Vocabulary Motion Generation (Supplementary)

Ke Fan[1] Jiangning Zhang[2,3] Ran Yi[1†] Jingyu Gong[4] Yabiao Wang[2,3]
Yating Wang[1] Xin Tan[4,5] Chengjie Wang[3,1] Lizhuang Ma[1,4,6†]
[1]Shanghai Jiao Tong University [2]Zhejiang University
[3]Tencent Youtu Lab [4]East China Normal University
[5]Shanghai AI Laboratory [6]MoE Key Lab of Artificial Intelligence, SJTU

1. Overview

In this section, we systematically present the following:

- Additional Implementation Details
- Details of the Textual Decomposition Algorithm
- Further t-SNE Analysis
- More Qualitative Comparisons
- Qualitative Ablation
- Visualization Explanation on Atomic Text and Generation Results

2. Additional Implementation Details

Implementation Details. All experiments were conducted on a setup consisting of four Tesla V100 GPUs. During the *first training* stage, we utilized a residual VQ-VAE architecture with one base layer and five residual layers. Further implementation details regarding Residual VQ-VAE can be found in MoMask [2]. The model was trained for 100 epochs, with a learning rate of 2×10^{-4} and a batch size of 512 per GPU. The codebook size and downsampling ratio were set to 512 and 4, respectively. Generative models were trained for both the base and residual layers, with shared parameters across all residual layers, differing only in the respective layer IDs. Both generative models were trained for 500 epochs, with a learning rate of 2×10^{-4} and a batch size of 64 per GPU.

3. Details of the Textual Decomposition Algorithm

In this section, we elaborate on the specific steps of the textual decomposition algorithm employed for both training and evaluation. During training, we apply a manually designed algorithm to decompose the HumanML3D training dataset into atomic actions. These decompositions are represented as triplets `<raw_texts, atomic_texts, motions>`, which serve as the training input for our model. For evaluation, we derive additional cases from the training dataset, guiding the large language model (LLM) to partition the input raw texts into atomic texts. Subsequently, the raw and atomic texts are used to generate corresponding motions. The generated motions are then compared to the ground truth motions using evaluators to compute various performance metrics.

The textual decomposition process consists of two primary components: (1) the Fine-grained Description Conversion Algorithm and (2) LLM Summarization. Specifically, the Fine-grained Description Conversion Algorithm includes the following sub-processes: (i) Pose Extraction, (ii) Pose Aggregation, (iii) Clip Aggregation, and (iv) Description Conversion. Below, we describe each of these processes in detail, along with the role of the LLM in generating atomic motion texts during evaluation.

3.1. Pose Extraction

For each frame x_i in a 3D motion sequence, we extract a set of pose descriptors (PD_i), which include joint angles, orientations, absolute positions, and pairwise distances between joints.

Categories of Pose Descriptors:

- **Angle:** Describes the bending or extension of specific body parts.
- **Orientations:** Describes the changes in orientation of the root joint along the X, Y, and Z axes.
- **Absolute Positions:** Describes the position of the root joint in 3D space, which is subsequently used in the Clip Aggregation phase to calculate the body's movement range.
- **Pairwise Distances:** Describes the distance between two body parts at a specific time, which is later used to assess changes in distances between body parts during Clip Aggregation.

Involved Body Parts or Joints:

† Corresponding author

Body Part	Joint
Left Knee	Left Hip, Left Knee, Left Ankle
Right Knee	Right Hip, Right Knee, Right Ankle
Left Elbow	Left Shoulder, Left Elbow, Left Wrist
Right Elbow	Right Shoulder, Right Elbow, Right Wrist

Table 1S. Involved body parts and their corresponding joints of Angle.

Joint1	Joint2
Left Shoulder	Right Shoulder
Left Elbow	Right Elbow
Left Hand	Right Hand
Left Knee	Right Knee
Left Foot	Right Foot
Neck	Pelvis
Left Ankle	Neck
Right Ankle	Neck
Left Hip	Left Knee
Right Hip	Right Knee
Left Hand	Left Shoulder
Right Hand	Right Shoulder
Left Foot	Left Hip
Right Foot	Right Hip
Left Wrist	Neck
Right Wrist	Neck
Left Hand	Left Hip
Right Hand	Right Hip
Left Hand	Torso
Right Hand	Torso
Left Foot	Torso
Right Foot	Torso

Table 2S. Involved joints of Pairwise Distances.

- **Angle:** The involved body parts and joints are listed in Tab. 1S.
- **Orientations:** Root joint.
- **Absolute Positions:** Root joint.
- **Pairwise Distances:** The involved body parts are listed in Tab. 2S.

3.2. Pose Aggregation and Description Conversion

As discussed in the main text, we aggregate pose descriptors PD_i from consecutive frames into a clip, denoted as S_{PD_i}, if the sign of their differences remains consistent (either positive or negative). We then compute the velocity V_{PD_i} for each descriptor.

Categorization of Descriptions: We categorize the pose descriptors into directional descriptions based on the sign of S_{PD_i} and classify them according to their absolute values. Additionally, we categorize motion speed based on the absolute magnitude of velocity. The descriptions for various pose descriptors are as follows:

- **Sign:**
  - **Angle:** Bending, Extending.
  - **Absolute Position:** Upward, Downward, Southward, Westward, Eastward, Northward.
  - **Orientation:** Leaning Backward, Leaning Forward, Counterclockwise, Clockwise, Leaning Right, Leaning Left.
  - **Pairwise Distances Between Joints:** Left-to-Right, Right-to-Left, Above-to-Below, Below-to-Above, Behind-to-Front, Front-to-Behind.
- **Intensity:** Significant, Slight, Moderate, Stationary.
- **Speed:** Very Slow, Slow, Moderate, Fast, Very Fast.

To manage the large number of possible combinations, we only include descriptions that occur infrequently in the dataset, as these provide valuable insights into action style and semantics. These less common behaviors are incorporated with low probability, as illustrated in Fig. 1S, which presents statistical results for different attributes.

Conversion Results: As shown in Fig. 2S, we present some conversion results in the blue section.

3.3. Clip Aggregation

As we mentioned in the main part, we partition the motion sequence into P uniformly spaced temporal bins. Here, we set the P as 10. We explain the reason to choose the number 10. Since we distribute the motion descriptors into bins according to their start times, some long motion descriptors may span multiple bins, resulting in a single motion generating less than P bins. As shown in the Fig. 3S, we set the maximum bin number range from 20 to 25, and find that the bin number with the highest occurrences is always around 10. Besides, the longest motion in the HumanML3D dataset is 10s, which means an atomic motion is at most 1s, and this is reasonable. Therefore, we set the bin num as 10.

3.4. LLM Summarization

As depicted in the blue section of Fig. 2S, our algorithm employs a template-based description system that includes detailed information about various body parts. While the generated descriptions align well with the motion patterns and exhibit high accuracy, two key issues remain: (1) insufficient diversity and (2) redundancy in the descriptions. To address these problems, we utilize a pre-trained large language model (LLM). By providing the conversion results and corresponding raw text to the LLM, we guide it to extract key atomic patterns that reflect the semantic behavior of the raw text, thereby generating more diverse summaries. The LLM's performance is shown in the green section of Fig. 2S, where it successfully extracts critical information.

The prompt used for LLM summarization is displayed in Fig. 4S.

3.5. LLM for Atomic Motion Texts During Evaluations

During inference, we use in-context learning to guide the LLM in decomposing the raw text according to the provided examples, generating atomic motion texts. The examples consist of 15 converted results from the training set. We instruct the LLM to partition the raw text into several time periods, with each period representing one of the six atomic motion categories (spine, left/right upper/lower limbs, and trajectory). The specific prompts used for LLM inference are presented in Tab. 3S.

4. Further t-SNE Analysis

We present the t-SNE results using BERT features [1] in Fig. 5S, where we observe similar outcomes to those discussed previously.

5. More Qualitative Comparisons

As shown in Fig. 6S, our method significantly outperforms existing state-of-the-art approaches. For example, in the task of "Standing to Kneeling Down," all other methods fail to correctly interpret the temporal sequence of the two motions (standing and kneeling), whereas our method accurately respects the temporal constraints. Furthermore, while "Getting Rocked By A Big UpperCut" can only be achieved by MotionMillion, which was trained on a millionaire dataset with abundant semantics. However, our method, trained on HumanML3D, generates semantically correct actions, highlighting the effectiveness of our Textual Decomposition and Atomic Recomposition approach.

6. Qualitative Ablation

As shown in Fig. 7S, compared with directly concate all atomic texts together (CFF*), using our proposed compositional feature fusion (CFF) could significantly enhance the generation performance, which further indicates the effectiveness of our proposed CFF module.

7. Visualization Explanation on Atomic Text and Generation Results

As shown in Fig. 8S, Akimbo is a rare motion in the dataset, by leveraging the atomic motion decomposition capability of large language models, we can decompose the motion of placing hands on the hips into distinct phases of upper limb flexion and positioning near the hip. This enables the generation of actions that are semantically consistent with the concept of Akimbo. This demonstrates the effectiveness of our proposed *Textual Decomposition and Atomic Recomposition* approach.

References

[1] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 3

[2] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1900–1910, 2024. 1

system prompt: I would like you to play the role of a kinesiology expert to assist me in accurately describing an motion.

 # CONTEXT #

I will provide you with a description of an individual's motion. Each description encompasses information regarding the actions of the person.
The actions might be described too abstractly or coarsely. I require you to furnish me with a detailed account of the motion based on your kinesiology expertise and the subsequent instructions.
I expect you to:
(1) Segregate this action into several distinct stages.
(2) For each stage, provide a detailed description of the following body parts for each individual.
The body parts should include ["spine", "left_upper_limb", "right_upper_limb", "left_lower_limb", "right_lower_limb", "trajectory"].
(3) The rules and output requirements are listed below. Please adhere to them to accomplish the task.
(4) I have provided you with some examples to facilitate your comprehension of the task. Kindly review them before commencing the task.
The output method should be strictly in the form as in the example, and for the description methods of different stage body parts, please refer to the example.

 # RULES #

(1) Avoid using uncertain words like "may" in the split statement. Also, refrain from using words such as "also", "too" in the split statement.
(2) The output description should be physically plausible,
The behavior of each body part must be capable of reflecting the comprehensive.

 # OUTPUT REQUIREMENTS #

(1) Return Format: JSON
(2) Please follow the format of the example below to return the output, don't output other information.

 # Examples #

Example 1:
<input:>he stomps his left feet</input>
<output:>{
"0": {
"spine": "remains relatively stable as the motion initiates",
"left_upper_limb": "left arm moves down slightly",
"right_upper_limb": "no significant movement",
"left_lower_limb": "left hip shifts preparatory to stomp, ankle begins to flex",
"right_lower_limb": "stationary",
"trajectory": "preparing for stomping action"
},
"1": { ...(×6)
}...
...
Example 15:
...

Table 3S. The prompts used in the LLM for obtaining the atomic motion texts during evaluation.

Figure 1S. Statistical analysis of the frequency of different motion descriptors.

Raw Text

a person raises his right arm and moves his left arm and left hand up and down in a regular fashion

Fine-grained Description Coversion Algorithm

<Atomic descriptions>
"0": The left hand is slightly moving from behind to the front relative to the the left shoulder extremely slow with the left hand moving to the right slightly relative to the right hand at a very slow pace, the left knee is moderately bending extremely slow with the right hand slightly moving from behind to the front relative to the the neck at a very slow pace, and the right hand is a bit moving from behind to the front relative to the the right hip extremely slow. The right hand is slightly spread away from the left foot extremely slow with the right hand slightly getting away from the right knee extremely slow, the left hand is moving to the right a bit relative to the torso very slow, while the right hand is moving to the above quite a little bit relative to the right hip extremely slow, while the left hand is moving to the right a bit relative to the neck at a very slow pace. The left hand is moving to the front a bit relative to the torso extremely slow, while the left hand is moving to the front a bit relative to the left hip very slow, while the left hand is a bit closing to the the right shoulder at a very slow pace. The right hand is a bit getting away from the right ankle at a very slow pace with the left hand slightly get far from the left leg very slow.
"1": The left hand is slightly spread away from the right foot very slow.
"2": Their right hand is a bit get cloesd to the their right foot very slow, and their left hand is slightly moving from right to the left relative to the their torso very slow, while their left hand is moving to the behind slightly relative to their left shoulder at a very slow pace with their left hand slightly closing to the their left ankle extremely slow. Their right hand is moving to the above quite a little bit relative to their right hip at a very slow pace, while their left knee is moderately getting extended very slow, their right hand is moving to the behind a bit relative to their neck very slow, their right hand is slightly get cloesd to the their right leg at a very slow pace with their left hand slightly closing to the their right ankle very slow. Their right hand is a bit moving from front to the behind relative to the their right hip at a very slow pace, their right hand and their left foot are a bit closing at a very slow pace, while their right hand is moving to the behind a bit relative to their torso at a very slow pace, and their left hand is moving to the left slightly relative to their left hip extremely slow with their left hand a bit distance from their right shoulder at a very slow pace.
"3": Her right hand is quite a little bit moving from below to the above relative to the her right hip at a very slow pace.
</Atomic descriptions>

LLM Sumarization


```
{
"0": {
"left_upper_limb": "moves from behind the shoulder towards the front and slightly right",
"right_upper_limb": "moves from behind the hip towards the front, slightly spreading away from the left foot",
"spine": "slight adjustment to accommodate arm movements",
"left_lower_limb": "left knee bending",
"right_lower_limb": "stabilizing with minor adjustments",
"trajectory": "stationary with upper limb motion"
},
"1": {
"left_upper_limb": "moves away from right foot",
"right_upper_limb": "remains raised, possibly starting to lower",
"spine": "neutral, adjusting for left arm's wider movement",
"left_lower_limb": "stable",
"right_lower_limb": "stable",
"trajectory": "upper limbs' positions changing"
},
"2": {
"left_upper_limb": "moves in a downward towards left, away from torso",
"right_upper_limb": "may be returning towards the hip, moving behind relative to neck",
"spine": "possibly straightening",
"left_lower_limb": "left knee begins to straighten",
"right_lower_limb": "minor adjustments, possibly returning to neutral",
"trajectory": "left arm descending"
},
"3": {
"left_upper_limb": "continues regular up-down motion, now ascending",
"right_upper_limb": "raising again from below",
"spine": "remains neutral",
"left_lower_limb": "stable, preparing for next cycle",
"right_lower_limb": "stabilizes",
"trajectory": "repetitive motion established"
}
}
```


Figure 2S. The process of Textual Decomposition Algorithm.

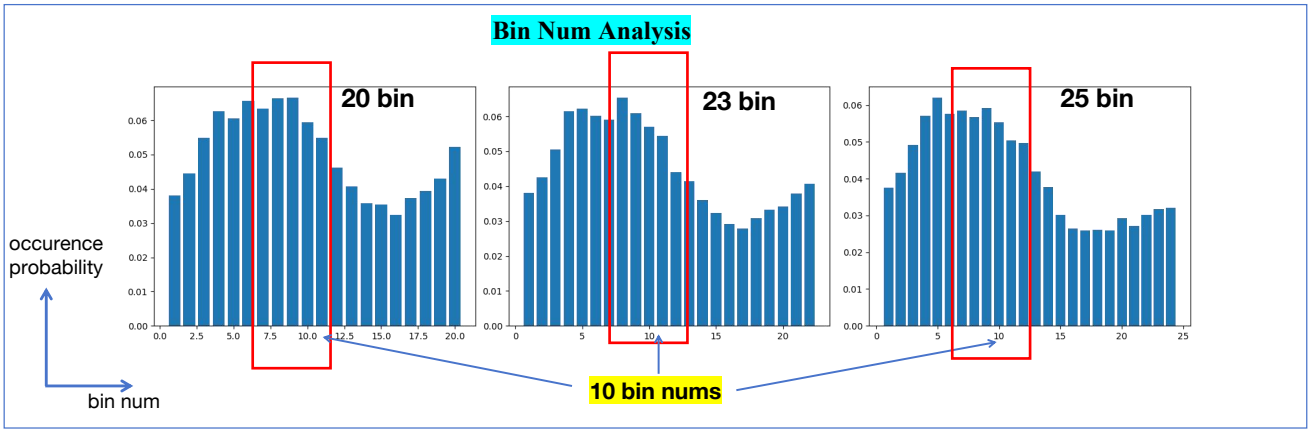

Figure 3S. Anaylsis of bin number selection.

Prompt of LLM Sumarization

```
I would like you to play the role of a kinesiology expert to assist me in accurately describing an motion.
# CONTEXT #
 I will give you a human motion description and the stages descriptions of the various joints in each stage. Now
each stage may contain a very large number of joints and corresponding redundant motion information, and some
stages may contain only some of the joints. I hope you:
(1) Combined with the semantics of movement, summarize the behavior of each joint in each stage. If the stage is
already over, discard the subsequent stage.
(2) Summarize the movements of the following joints in each stage. The joints should contain ["spine",
"left_upper_limb", "right_upper_limb", "left_lower_limb", "right_lower_limb", "trajectory"].
# OUTPUT REQUIREMENTS #
 - Return Language: English
 - Return Format: JSON
# EXAMPLE #
<human motion description> a person squats down then jumps </human motion description>
<human motion description> a person squats down and puts their hands above their head </human motion description>
<stage descriptions>
"0": "The left hand moderately get far from the left foot extremely slow with the right hand quite getting away
from the right foot very slow, the right hand quite distance from the left foot very slow.",
"1": "The left hand is quite moving from behind to the front relative to the the torso at a very slow pace, the
right hand is moderately moving from behind to the front relative to the the torso very slow with the right hand
quite moving from behind to the front relative to the the right hip extremely slow with the right hand moving to
the front quite relative to the right shoulder extremely slow.",
</stage description>
<output>
{
 "0": {
 "spine": "remains neutral while arms start moving away from the body",
 "left_upper_limb": "left arm moves away from left foot",
 "right_upper_limb": "right arm moves away from right foot",
 "left_lower_limb": "no significant movement",
 "right_lower_limb": "no significant movement",
 "trajectory": "stationary"
 },
 "1": {
 "spine": "beginning to flex forward",
 "left_upper_limb": "left hand moves forward towards the torso",
 "right_upper_limb": "right hand moves forward, bending at the elbow",
 "left_lower_limb": "preparation for squat with slight anterior movement",
 "right_lower_limb": "similar to left, slight preparation for squat",
 "trajectory": "forward motion of arms"
 },
}
</output>
```

Figure 4S. The prompt used for LLM Summerization.

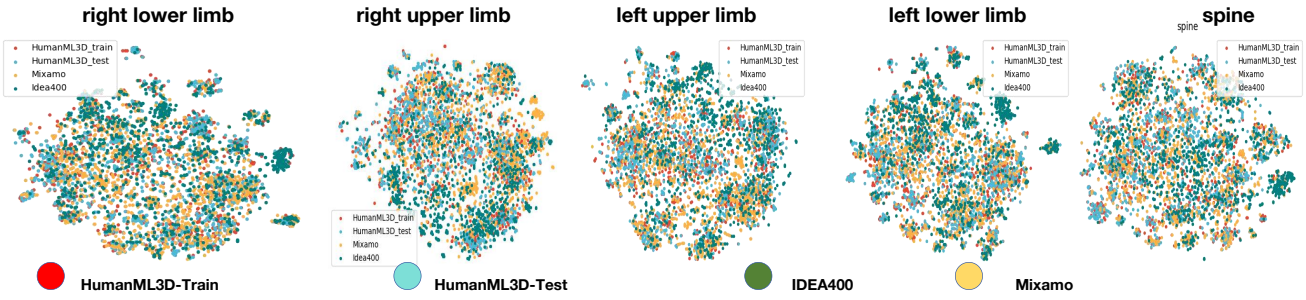

Figure 5S. t-SNE analysis on atomic texts via BERT model.

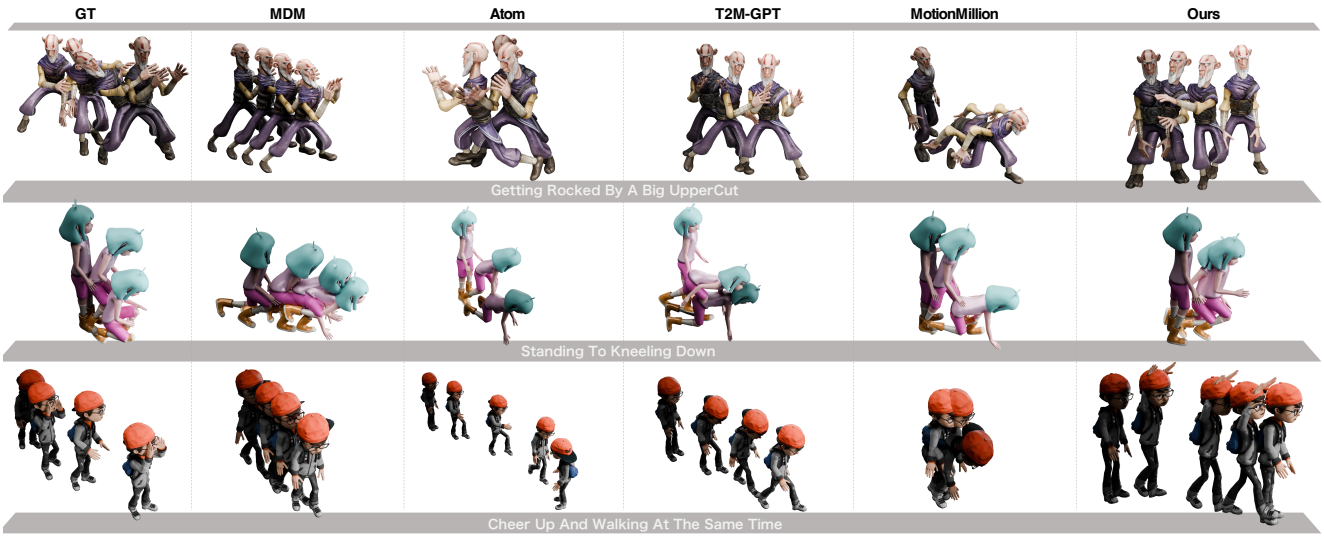

Figure 6S. Comparison with several state-of-the-art on open vocabulary texts.

Figure 7S. Ablation on CFF* and our proposed compositional feature fusion (CFF) module.

GT

Ours

Raw Text: Akimbo And Stand At The Same Time

Atomic Text:


```
{
"0": {
"spine": "remains straight, displaying erect posture",
"left_upper_limb": "left arm moves outward, hand settling on left hip, elbow bent",
"right_upper_limb": "right arm mirrors the left, positioning hand on right hip",
"left_lower_limb": "left leg bears part of the weight, slight outward stance for balance",
"right_lower_limb": "right leg equally distributes weight, stance slightly widened",
"trajectory": "arms moving to hip position"
 },
"1": {
"spine": "preserves an upright position, spine aligned",
"left_upper_limb": "left hand now rests firmly on hip, forearm perpendicular to the body",
"right_upper_limb": "right hand similarly placed, creating an akimbo stance",
"left_lower_limb": "ankle adjusts for slight redistribution of weight",
"right_lower_limb": "ankle mirrors the left, stabilizing the wider stance",
"trajectory": "established akimbo posture"
 },
"2": {
"spine": "subtle engagement of core muscles to maintain stability",
"left_upper_limb": "left hand stays firm on hip, displaying confidence",
"right_upper_limb": "right hand maintains its position, reinforcing the stance",
"left_lower_limb": "left foot maintains contact, distributing weight evenly",
"right_lower_limb": "right foot does the same, ensuring stability",
"trajectory": "static stance with active core"
 },
"3": {
"spine": "may exhibit a minute forward lean to project assertiveness",
"left_upper_limb": "left elbow slightly flexes, adding to the relaxed akimbo posture",
"right_upper_limb": "right elbow follows suit, relaxing yet firm",
"left_lower_limb": "left leg muscles slightly tense to counteract lean",
"right_lower_limb": "right leg mirrors the left, maintaining balance",
"trajectory": "minimal forward body lean"
 },
"4": {
"spine": "returns to a strict vertical alignment, displaying poise",
"left_upper_limb": "left hand and arm retain position, conveying confidence",
"right_upper_limb": "right hand stays akimbo, emphasizing posture",
"left_lower_limb": "left foot adjusts for any micro-movements, grounded",
"right_lower_limb": "right foot remains firm, completing the stable standing pose",
"trajectory": "settled and balanced stance"
 }
}
```


Figure 8S. Visualization on atomic text and generation results.