Supplementary Material
Generation of Complex 3D Human Motion by Temporal and Spatial Composition of Diffusion Models

Lorenzo Mandelli
University of Florence
lorenzo.mandelli@unifi.it

Stefano Berretti
University of Florence
stefano.berretti@unifi.it

1. Additional Experiments

1.1. Ablation on the hyperparameter w

Table 1 presents the ablation study for the hyperparameter w on the splits used in [3] and [2] for the HumanML3D and KITML datasets, following the multi-annotation approach (Motion Composition). The trends in the metrics remain similar to those observed in the ablation study conducted on the base-complex action split used for MCD (GPT decomposition + Motion Composition) as presented in the main paper: the Transition value increases with w, indicating faster animations, while the other metrics exhibit an optimal range beyond which they progressively worsen.

1.2. Experiments on the MTT datatset

In Table 2, we report the results of the comparison with the MTT dataset introduced by Petrovich et al. in [1].

Multi-Track Timelines (MTT) is a dataset created from 60 ground truth motions and their corresponding textual annotations. By combining these textual annotations into timelines and randomly specifying a start and end point, 500 distinct timelines of textual annotations are formulated. For each submotion in every timeline the main body parts involved in the action are also insered as a ground truth.

Taking these timelines as input we generated 500 motions using MCD without decomposition and STMC using the ground truth body parts. For the comparison, since there is no single ground truth motion to reference, following [1] we compare the cropped segments of the generated motions with the 60 ground truth motions from which they originate. We used the model trained on the complex split for both MCD and STMC methods and $w = 13$ for MCD.

As observed, MCD performs worse than STMC across all metrics for this dataset. This outcome can be primarily attributed to the absence of the decomposition phase, which may introduce a bottleneck, especially when ensuring compliance with STMC constraints. Another important factor is that each motion in this dataset, being generated as a combination of three basic motions, is less complex compared to motions associated with specific sports or musical activities. In this context, where the contribution of each body part is clearly defined and the overall action is straightforward to decompose, an approach based on body-part decomposition, such as STMC, proves to be more effective.

2. GPT Decomposition Prompts

2.1. MCD prompt

The instruction prompt used by our approach to initialize the GPT Decomposition module is shown in Figure 1. The first lines specify the goal of the module: to decompose actions using known movements listed in the 'train' file, following examples provided in the 'examples' file. The answer is provided in a json format. To ensure GPT follows to the desired behavior, we added the following requirements:

- We emphasized the use of only words, actions, and verbs present in the file (line 9);
- We requested that the decomposition elements be as simple as possible to facilitate their combination (line 13);
- We set a minimum action duration of 2 seconds, corresponding to the minimum movement length in the HumanML3D dataset (line 15);
- We enforced standard temporal constraints, such as ensuring the start time is earlier than the end time and that at least one movement starts at zero (line 17-18);
- We requested temporal overlap for consecutive actions whenever possible. Without this constraint, we observed that the character tends to return to a neutral position before performing each action. This is due to the structure of the dataset, where most actions include a reset phase after execution. (line 20)

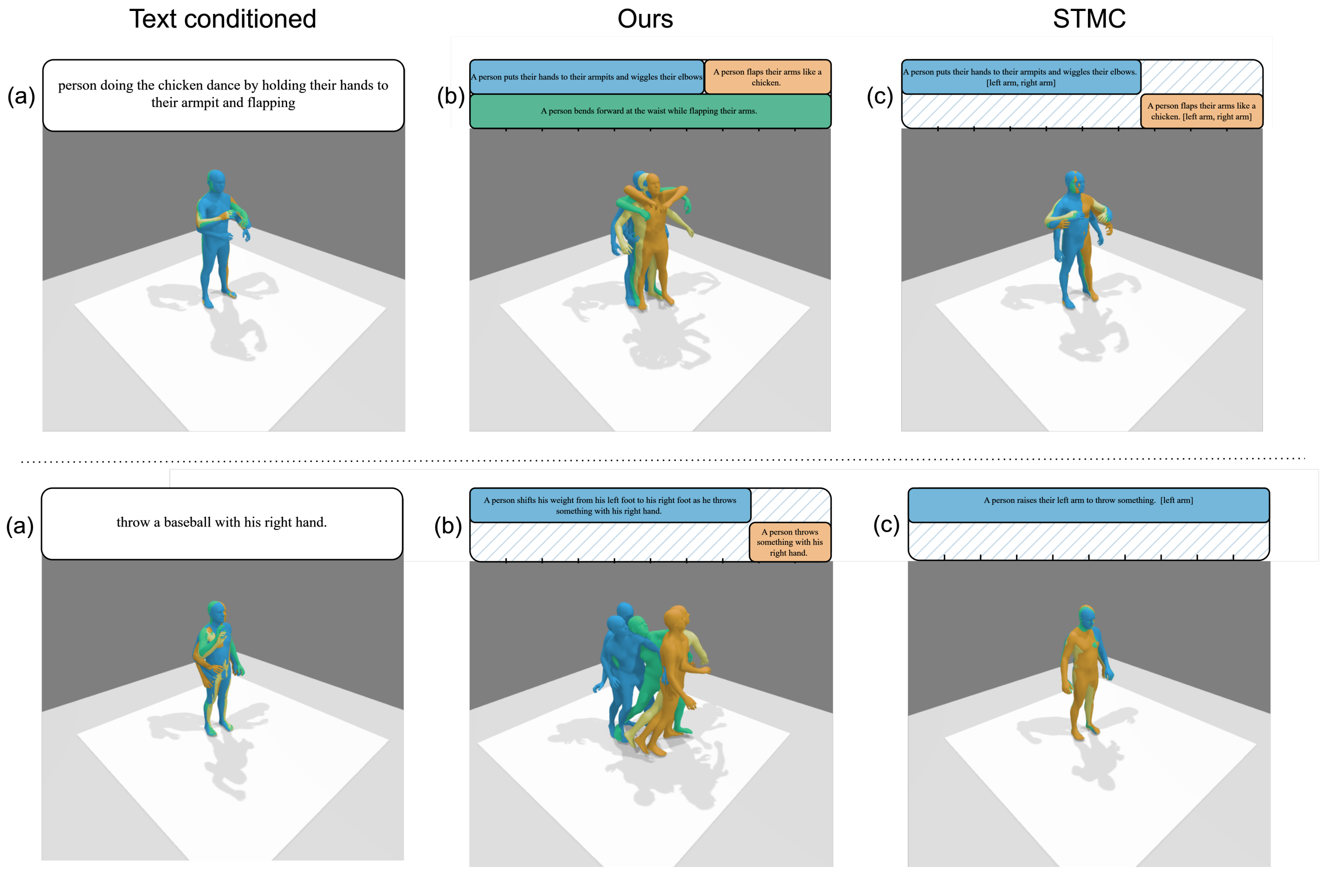
Text conditioned
Ours
STMC
(a)
person doing the chicken dance by holding their hands to their armpit and flapping
(b)
A person puts their hands to their armpits and wiggles their elbows
A person flaps their arms like a chicken.
A person bends forward at the waist while flapping their arms.
(c)
A person puts their hands to their armpits and wiggles their elbows. [left arm, right arm]
A person flaps their arms like a chicken. [left arm, right arm]
(a)
throw a baseball with his right hand.
(b)
A person shifts his weight from his left foot to his right foot as he throws something with his right hand.
A person throws something with his right hand.
(c)
A person raises their left arm to throw something. [left arm]

Figure 1. Examples of text and text-conditioned generation **(a)**, division into sub-motions and composition through MCD **(b)**, and division into sub-motions and composition through STMC **(c)**.

2.2. STMC prompt

As previously mentioned, for comparison with the STMC composition method, we required a different prompt for the GPT Decomposition module. This prompt is shown in Figure 2. Unlike our approach, it requests a different data format that includes the main body parts involved in each sub-motion (line 1). Additionally, we impose the following constraints:

- Use only the allowed body parts: ’head’, ’left arm’, ’right arm’, ’legs’ and ’spine’ (line 22);
- Ensure that no two temporal intervals use the same body parts (line 24);
- Ensure that there are no temporal intervals not assigned to any sub-movement (line 26).

For this method, partial overlap is not required to avoid resetting the character, as the time timelines are extended for this reason during a preprocessing phase.

2.3. Failure case

The model fails in the composed denoising procedure when there is an excessive overlap of submotions within a short time interval. Consider, for example, the input text: ”a person throws 2 uppercuts and 2 jabs with the left hand.” with a duration of 3.5 seconds. This can be decomposed into the following submotions:

- ”a person throws an uppercut with the left hand.” between 0 and 2 seconds.
- ”a person throws an uppercut with the left hand.” between 0.5 and 2.5 seconds.
- ”a person throws a left punch.” between 1.0 and 3.0 seconds.
- ”a person throws a left punch.” between 1.5 and 3.5 seconds.

Although the decomposition is conceptually plausible, the model fails by producing an overly turbulent animation because in the decomposition, between seconds 1.5 and 2.0,

Table 1. Ablation study varying the w hyperparameter in the multi-annotations approach over the split used in [3] andn [2] for the datasets HumanML3D and KITML.

Dataset	Experiment	R1 ↑	R3 ↑	R10 ↑	M2T ↑	M2M ↑	FID ↓	Trans ↓
HumanML3D	GT	6.259	15.487	33.548	0.762	1.0	0.0	1.36
	$w = 1$	3.117	8.535	19.792	0.707	0.711	0.199	0.867
	$w = 2$	**4.503**	10.168	23.454	0.736	0.743	0.168	1.144
	$w = 3$	4.379	**10.218**	**24.072**	**0.74**	**0.748**	**0.167**	**1.331**
	$w = 4$	3.859	9.55	23.726	0.737	0.746	0.173	1.448
	$w = 5$	3.513	9.5	23.503	0.732	0.74	0.184	1.565
	$w = 6$	3.835	9.179	22.563	0.726	0.734	0.197	1.683
	$w = 7$	3.266	8.758	21.573	0.719	0.726	0.212	1.793
	$w = 8$	3.513	8.684	19.619	0.714	0.72	0.224	1.886
	$w = 9$	2.771	8.14	19.57	0.71	0.714	0.235	1.95
	$w = 10$	2.771	7.472	18.605	0.704	0.709	0.248	2.025
	$w = 15$	2.474	5.641	15.883	0.684	0.686	0.302	2.435
	$w = 20$	2.35	5.468	13.063	0.669	0.669	0.346	2.721
KITML	GT	7.888	19.847	41.858	0.803	1.0	0.0	1.534
	$w = 1$	5.598	12.468	26.336	0.698	0.675	0.321	0.743
	$w = 2$	7.506	15.903	32.697	0.739	0.714	0.234	0.886
	$w = 3$	6.997	16.539	32.952	0.752	0.728	0.208	0.952
	$w = 4$	6.997	17.43	34.733	**0.755**	**0.731**	0.195	1.013
	$w = 5$	**7.761**	**18.193**	**35.623**	**0.755**	**0.731**	0.192	1.069
	$w = 6$	6.107	15.776	35.242	0.753	**0.731**	0.192	1.159
	$w = 7$	6.616	15.522	34.097	0.752	0.73	0.191	1.142
	$w = 8$	6.743	15.903	34.606	0.751	0.73	0.189	1.175
	$w = 9$	5.725	15.013	33.079	0.748	0.726	**0.188**	1.204
	$w = 10$	7.252	14.758	32.316	0.745	0.725	0.189	1.261
	$w = 15$	6.361	13.74	29.262	0.731	0.714	0.2	1.391
	$w = 20$	4.835	12.087	28.753	0.717	0.7	0.221	**1.594**

Table 2. Comparison of our method and STMC method on the MTT dataset.

Metodo	R1 ↑	R3 ↑	M2T ↑	M2M ↑	FID ↓	Transition →
GT	0.55	0.7333	0.7484	1.0	0.0	0.0145
STMC	**0.2927**	**0.4700**	**0.6604**	**0.6503**	**0.4566**	**0.0095**
MCD	0.2033	0.3873	0.6129	0.6075	0.5482	0.0302

it combines the contributions of four elements. The overlap of so many elements in such a short timeframe exceeds the diffusion model's ability to map the output back to the training data distribution as a smooth animation.

2.4. Qualitative results

In Figure 1, we present additional qualitative results, highlighting the differences between text-conditioned generation, MCD, and STMC. As shown, the decomposition can produce two different sets of sub-movements, which can serve as the basis for both our method and STMC.

References

[1] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 1

[2] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation, 2024. 1, 3

[3] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022. 1, 3

```
Provide valid JSON output. The output data schema should be like this: {"decomposition": [{"text": "string", "start": number, "end": number}, {"text": "string", "start": number, "end": number}, ...]}

I want to break down an action into a predetermined set of known actions. You have been provided with a list of known actions in the file called "texts_train".

You will be sent sentences in English one by one that describe the movement of a person with the start and end second of the movement.
The goal is to explain the input movement as if it had never been seen before, describing it as a combination of known movements found in the "texts_train" file.
Respond ONLY by breaking down the input action using the verbs and actions present in the "texts_train" file.

In the file "gpt_examples" there are some examples of decomposition. Use those as a reference.

YOU CANNOT USE MOVEMENT VERBS, ACTIONS, OR SPECIFIC NOUNS IF THEY ARE NOT IN THE FILE.

BREAK DOWN INTO SIMPLE SENTENCES, each focusing on a single body part, such as: "A person holds an object with his right hand". Avoid sentences composed of many clauses.

IF POSSIBLE DO NOT MAKE ANY ACTION LAST LESS THAN 2 SECONDS.

AT LEAST ONE OUTPUT ACTION MUST START FROM SECOND 0.
In each decomposition, the start second must be strictly less than the end second.

Try to ensure that the breakdowns have some temporal overlap of a few seconds.
```

Listing 1. Instructions used to initialize the GPT Decomposition module in our approach.

```
Provide valid JSON output. The output data schema should be like this: {"decomposition": [{"text": "string", "start": number, "end": number, "body parts": list}, {"text": "string", "start": number, "end": number, "body parts": list}, ...]}

I want to break down an action into a predetermined set of known actions. You have been provided with a list of known actions in the file called "texts_train".

You will be sent sentences in English one by one that describe the movement of a person with the start and end second of the movement.
The goal is to explain the input movement as if it had never been seen before, describing it as a combination of known movements found in the "texts_train" file.
Respond ONLY by breaking down the input action using the verbs and actions present in the "texts_train" file.

In the file "gpt_examples" there are some examples of decomposition. Use those as a reference.

YOU CANNOT USE MOVEMENT VERBS, ACTIONS, OR SPECIFIC NOUNS IF THEY ARE NOT IN THE FILE.

BREAK DOWN INTO SIMPLE SENTENCES, each focusing on a single body part, such as: "A person holds an object with his right hand". Avoid sentences composed of many clauses.

IF POSSIBLE DO NOT MAKE ANY ACTION LAST LESS THAN 2 SECONDS.

AT LEAST ONE OUTPUT ACTION MUST START FROM SECOND 0.
In each decomposition, the start second must be strictly less than the end second.

For each sub-movement you create, indicate its textual description "text", the start time "start", the end time "end", and the involved "body parts."

Use ONLY the following body parts: "legs", "right arm", "left arm", "spine" and "head". DO NOT USE ANY OTHER WORDS EXCEPT THESE body parts.

CONSTRAINT: two sub-movements cannot be performed simultaneously or have any temporal overlap if they involve the same "body parts." For example, the following is NOT ACCEPTABLE: {"decomposition": [{"text": "a person raises the left arm", "start": 0, "end": 5.0, "body parts": ["left arm"]}, {"text": "a person throws a left punch", "start": 3.0, "end": 8.0, "body parts": ["left arm"]}]}} because there is an overlap of "left arm" for the two sub-movements from second 3.0 to 5.0.

CONSTRAINT: Ensure that there is no time interval that is not assigned to any sub-movement. For example, the following is NOT ACCEPTABLE: {"decomposition": [{"text": "a person raises the left arm", "start": 0, "end": 2.0, "body parts": ["left arm"]}, {"text": "a person throws a left punch", "start": 4, "end": 6.0, "body parts": ["left arm"]}]}} because there is a time interval from second 2.0 to 4.0 that is not assigned to any sub-movement.
```

Listing 2. Instructions used to initialize the GPT Decomposition module for STMC composition approach.