Multi-Modal Controlled Coherent Motion Generation
— Supplementary Material —

A Body-Part Motion Statistics

A core design principle of MOCO is the spatial decomposition of multi-modal conditioning: text descriptions govern lower-body motion, while speech audio drives upper-body dynamics. This appendix provides a quantitative characterization of the motion distributions associated with each conditioning modality within the scope of our target task, offering an empirical basis for understanding why this decomposition is well-suited to the concurrent text-and-speech control scenario.

Task-Scoped Text Conditions and Motivation for Generated Samples

Our target application involves concurrent control by both speech audio and text descriptions. In this setting, text conditions are expected to specify body-level actions that are *spatially complementary* to co-speech gestures—that is, actions primarily activating the lower body without conflicting with concurrent upper-body gesture generation. We therefore focus our analysis on a curated set of 40 text descriptions centered on locomotion and postural transitions (see Appendix H), which represent the intended operating domain of the text conditioning signal in MOCO. These descriptions are representative of the text conditions under which MOCO is designed to operate, and do not aim to characterize the full diversity of text-driven motion.

To obtain the motion distribution for these curated descriptions, we generate 1,000 sequences using a pretrained text-to-motion model (MDM [13]) trained on HumanML3D [5], as these specific lower-body–oriented descriptions do not have corresponding ground-truth annotations in existing datasets. This approach allows us to directly estimate the body-part activation patterns associated with the text conditions employed in our multi-modal benchmark.

Setup

Both data sources are represented using the unified 205-dimensional SMPL-X body parameterization. Body parts are defined via fixed joint indices, grouped into five regions: left arm, right arm, head, spine, and legs. For each motion sequence, we compute the temporal variance of each body-part region as a proxy for motion activity, and normalize across all five parts so that they sum to 100%

per sequence. We further aggregate upper-body parts (head, left arm, right arm) and lower-body parts (spine, legs), and define the lower-body ratio as:

$$r_{\text{lower}} = \frac{\text{Var}_{\text{lower}}}{\text{Var}_{\text{upper}} + \text{Var}_{\text{lower}}}, \tag{1}$$

where $\text{Var}_{\text{upper}}$ and $\text{Var}_{\text{lower}}$ denote the aggregated temporal variance of upper- and lower-body joints, respectively. This normalization eliminates confounds from sequence length and global motion scale, enabling a fair comparison across datasets.

Statistics are computed over two sources:

- **Generated (text-driven):** 1,000 sequences produced by a pretrained text-to-motion diffusion model conditioned on our curated set of 40 locomotion- and posture-oriented descriptions.
- **BEAT2 (speech-driven):** 5,000 sequences sampled from the BEAT2 training set [8].

Results

Quantitative results are reported in Tab. 1. Under our task-scoped text conditions, the generated motions exhibit a lower-body variance share of 62.0%, indicating that locomotion- and posture-oriented descriptions predominantly activate the legs and spine. In contrast, BEAT2 speech-gesture sequences are strongly upper-body dominant, with 85.1% of motion variance concentrated in the head and arms—consistent with the expressive, gesture-rich nature of co-speech motion. These complementary distributions provide empirical support for our proposed spatial decomposition.

Table 1: Body-part motion statistics for text-driven generated sequences and BEAT2 speech-gesture data. Temporal variance is computed per body-part region and normalized within each sequence.

Source	Upper MeanVar	Lower MeanVar	Lower / Upper	Lower-Body Ratio
Generated (text-driven)	0.00387	0.00663	1.71×	62.0%
BEAT2 (speech-driven)	0.00587	0.00164	0.28×	14.9%

B Theoretical Analysis for Decoupled Denoising

Our proposed MOCO relies on the assumption that the joint conditional probability $p(b_{t-1} \mid c_{\text{text}}, c_{\text{audio}}, b_t)$ can be approximated by $p(b_{t-1,\text{lower}} \mid c_{\text{text}}, b_t) \cdot$

$p(b_{t-1,\text{upper}} \mid c_{\text{audio}}, b_t)$, expressed as:

$$\begin{aligned} &p(b_{t-1} \mid c_{\text{text}}, c_{\text{audio}}, b_t) \\ &\approx p(b_{t-1,\text{lower}} \mid c_{\text{text}}, b_t) \cdot p(b_{t-1,\text{upper}} \mid c_{\text{audio}}, b_t), \end{aligned} \tag{2}$$

where b_t denotes the motion at denoising step t, composed of upper-body motion $b_{t,\text{upper}}$ and lower-body motion $b_{t,\text{lower}}$.

$$\begin{aligned} &p(b_{t-1} \mid c_{\text{text}}, c_{\text{audio}}, b_t) \\ &= p(b_{t-1,\text{lower}}, b_{t-1,\text{upper}} \mid c_{\text{all}}), \text{ where } c_{\text{all}} = \{c_{\text{text}}, c_{\text{audio}}, b_t\} \\ &= p(b_{t-1,\text{lower}} \mid c_{\text{all}}) \cdot p(b_{t-1,\text{upper}} \mid c_{\text{all}}, b_{t-1,\text{lower}}) && (3) \\ &\approx p(b_{t-1,\text{lower}} \mid c_{\text{all}}) \cdot p(b_{t-1,\text{upper}} \mid c_{\text{all}}) && (4) \\ &\approx p(b_{t-1,\text{lower}} \mid c_{\text{all}} \setminus \{c_{\text{audio}}\}) \cdot p(b_{t-1,\text{upper}} \mid c_{\text{all}} \setminus \{c_{\text{text}}\}) && (5) \\ &= p(b_{t-1,\text{lower}} \mid c_{\text{text}}, b_t) \cdot p(b_{t-1,\text{upper}} \mid c_{\text{audio}}, b_t). \end{aligned}$$

The first approximation occurs in the transition from Eq. (3) to Eq. (4). Here, we approximate:

$$\begin{aligned} &p(b_{t-1,\text{upper}} \mid c_{\text{all}}, b_{t-1,\text{lower}}) \\ &= p(b_{t-1,\text{upper}} \mid c_{\text{text}}, c_{\text{audio}}, b_{t,\text{upper}}, b_{t,\text{lower}}, b_{t-1,\text{lower}}) \\ &\approx p(b_{t-1,\text{upper}} \mid c_{\text{text}}, c_{\text{audio}}, b_{t,\text{upper}}, b_{t,\text{lower}}) \\ &= p(b_{t-1,\text{upper}} \mid c_{\text{all}}). \end{aligned}$$

This approximation assumes that b_t already encapsulates sufficient information about b_{t-1}, allowing us to neglect the influence of $b_{t-1,\text{lower}}$ when estimating $b_{t-1,\text{upper}}$. This simplification is justified by the proximity of the diffusion steps and the strong correlation between the states at steps t and $t-1$.

The second approximation occurs in the transition from Eq. (4) to Eq. (5), where we decouple modality-specific influences:

$$\begin{aligned} p(b_{t-1,\text{lower}} \mid c_{\text{all}}) &\approx p(b_{t-1,\text{lower}} \mid c_{\text{all}} \setminus \{c_{\text{audio}}\}), \\ p(b_{t-1,\text{upper}} \mid c_{\text{all}}) &\approx p(b_{t-1,\text{upper}} \mid c_{\text{all}} \setminus \{c_{\text{text}}\}). \end{aligned}$$

This approximation leverages the observation that text input (c_{text}) primarily influences lower-body movements (*e.g.*, walking or shifting stance), while audio input (c_{audio}) predominantly affects upper-body movements (*e.g.*, gestures or facial expressions). By excluding c_{audio} from the conditioning set for $b_{t-1,\text{lower}}$ and c_{text} for $b_{t-1,\text{upper}}$, we ensure the conditioning focuses on the most relevant modality for each body part.

C Leveraging LLM for Motion Planning

In summary, our prompt instructing the LLM to generate a motion timeline consists of several key components:

1. Purpose: Clearly state the objective of generating a motion timeline based on text and audio conditions.
2. Output Formats: Specify the formats that the LLM should adhere to when producing output.
3. Guidelines for Generating Timelines: Outline several essential rules that the LLM must follow.
4. Examples of Timelines: Provide a pair of example timelines, showcasing both an effective and a less effective version.

Additionally, to enhance the precision and completeness of the LLM's reasoning, we append the phrase "please reason step by step" at the end of the dialogue.

To illustrate this process more clearly, we provide a complete dialogue record with GPT-4o mini in `motion-planning-dialogue.pdf`. In this dialogue, the input conditions are processed by the LLM as follows:

Condition	Start Time	End Time	Body Part
sit down on a chair	0.0	5.0	legs, spine
stand up from chair	5.0	7.0	legs, spine
turn around	7.0	11.5	legs, spine
walk forward	11.5	17.5	legs, spine
5_stewart_0_1_1,$785440$911328	5.0	12.85	head, arms
5_stewart_0_1_1,$918560$1050080	17.5	24.3	head, arms

Explanation:

- Text Description: "sit down on a chair"
  - Start and End Time in Overall Motion: # 0.0 # 5.0
  - Controlled Body Parts: # legs # spine
- Audio Input File: speak:5_stewart_0_1_1
  - Start and End Frames in Audio File: $785440$911328
  - Start and End Time in Overall Motion: # 5.0 # 12.85
  - Controlled Body Parts: # head # arms

Notable, the LLM may occasionally produce calculation errors, such as miscalculating the duration of the second audio segment. However, these issues can be easily resolved through code refinement. The motion generated by our MOCO based on the refined timeline is provided in `motion-planning-result.mp4`.

D Conflict Resolution and Flexibility in Multi-Modal Control

MOCO 's default configuration assigns upper-body control to audio signals and lower-body control to text descriptions. However, there are scenarios where users

Condition	Start Time	End Time	Body Part
a man walks then waves hands	0.0	8.0	legs, spine
speak:1_wayne_0_103_103,$9248$99808	2.578	8.238	left arm, right arm, head
speak:1_wayne_0_103_103,$107552$196064	9.722	15.254	left arm, right arm, head

may wish to explicitly control specific upper-body motions, such as waving hands while talking. To clarify how our method handles this case, we provide the following example:

In this example, a conflict arises between the upper-body motion cues from the text description "waves hands" and the speech audio during the overlapping period from 2.578 seconds to 8.0 seconds. To resolve such conflicts, our method relies on a predefined timeline that gives precedence to the speech audio in governing upper-body motion during this interval. Consequently, the 'waves hands" motion is suppressed between 2.578 and 8.0 seconds.

While this illustrates a limitation of our default setup, the flexibility of MOCO offers a way to address such conflicts by enabling finer-grained user control. If users wish to control upper-body motions using text during speech, we can simply adjust the timeline as follows:

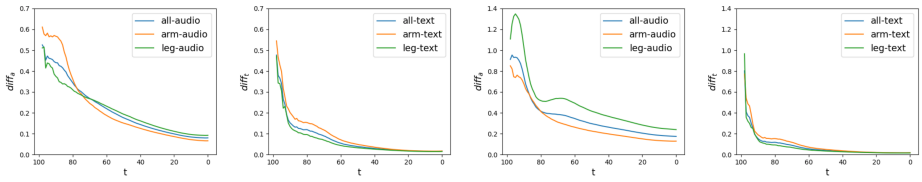
Condition	Start Time	End Time	Body Part
a man walks	0.0	8.0	legs, spine
wave hands	7.0	11.0	left arm, right arm
speak:1_wayne_0_103_103,$9248$99808	2.578	8.238	left arm, right arm, head
speak:1_wayne_0_103_103,$107552$196064	9.722	15.254	left arm, right arm, head

After modifying the timeline, the text command "wave hands" controls the arms from 7.0 to 11.0 seconds. Since the speech audio spans from 2.578 to 15.254 seconds, there is an overlap between 7.0 and 11.0 seconds during which both audio and text attempt to control the arms. According to our conflict resolution rules, **the condition with fewer controlled body parts takes precedence**. In this case, the text command "wave hands" governs the arms during the overlapping period.

This example demonstrates the flexibility of our method. Although our default configuration prioritizes audio for the upper body and text for the lower body, users can easily reconfigure these priorities through timeline adjustments to achieve their desired behavior. The corresponding visualization is provided in the supplementary materials as `flexibility.mp4`.

E Limitations of Weighted Sum in Multi-Modal Motion Generation

To understand why the *Weighted Sum* method achieves good results in speech-to-gesture metrics but poor performance in text-to-motion metrics, we conducted the following experiments.



(a) T, text ="standing" (b) A, text="standing" (c) T, text="sitting" (d) A, text="sitting"

Fig. 1: Comparison of differences calculated by the speech-to-gesture and text-to-motion models during motion updates. "**T**" denotes using a text-to-motion model to update motion, while "**A**" denotes using a speech-to-gesture model to update motion. The results show that the speech-to-gesture model computes larger differences than the text-to-motion model, indicating it adjusts the motion more aggressively based on the conditions. This explains why the weighted sum method attains strong performance on speech-to-gesture metrics but poor performance on text-to-motion metrics. Additionally, when the text condition is "sitting," the speech-to-gesture model calculates larger differences in the legs than in the arms. This is counterintuitive and could be attributed to data bias in the speech-to-motion dataset. Best viewed in color.

Given both speech and text inputs, we used only the text-to-motion model to update the motion. At each denoising step t, we computed the difference diff_t between the speech-to-gesture model's prediction—based on the speech input and the current motion from the text-to-motion model—and the current motion from the text-to-motion model. This difference quantifies how much the speech-to-gesture model perceives a mismatch between the speech condition and the current motion. Conversely, when we used only the speech-to-gesture model to update the motion, the calculated difference indicated how much the text-to-motion model perceived a mismatch between the text condition and the current motion. A larger difference suggests a greater mismatch and that the model will update the motion more aggressively.

We recorded these differences in both scenarios and divided them into whole body, arms, and legs for clearer illustration. Comparing Fig. 1 (a) and (b), as well as Fig. 1 (c) and (d), we found that the differences calculated by the speech-to-gesture model are larger than those by the text-to-motion model. This indicates that the speech-to-gesture model adjusts the motion more aggressively based on its conditions than the text-to-motion model does. This explains why, when using the weighted sum method, the generated result closely resembles that produced entirely by the speech-to-gesture model.

Furthermore, by comparing Fig. 1 (a) and (c), which have different text conditions, we observe that when the text condition is "sitting," the differences calculated by the speech-to-gesture model in the legs are larger than in the arms. This is counterintuitive since speech is typically associated with upper-body gestures rather than lower-body movements. Conversely, when the text condition is "standing," the differences in the legs are smaller than in the arms, aligning with expectations. This phenomenon may be attributed to data bias in

the speech-to-motion dataset, where most motions are performed in standing positions.

These observations reveal the limitations of weighted sum in multi-modal motion generation and suggest the validity of our proposed decoupled denoising process.

F Computational Complexity

Table 2: Complexity of each denoiser of MOCO.

	Parameters (M)	Model Size (MB)	FLOPs (G)	Inference Time (ms/frame)
G_{T2M}	27.01	103.02	5.19	2.26
G_{S2G}	36.86	140.62	6.72	4.30
G_{T2V}	0.34	1.31	0.06	6.20
G_{S2D}	36.94	140.94	6.74	4.37

Our framework, MOCO, comprises four transformer-based denoisers: G_{T2M} for text-to-motion (T2M), G_{S2G} for speech-to-gesture (S2G), G_{T2V} for trajectory-to-velocity (T2V), and G_{S2D} for speech-to-details (S2D), which manages facial expressions and hand poses. To clearly illustrate the computational complexity of MOCO, we present various metrics, including the number of parameters, model size, FLOPs, and inference time on a single NVIDIA 4090 GPU, as shown in Tab. 2.

As indicated in the table, our framework is overall lightweight and sufficiently fast. Specifically, the speech-to-gesture denoiser G_{S2G} and the speech-to-details denoiser G_{S2D} are relatively larger than the other denoisers due to additional cross-attention parameters. In contrast, the trajectory-to-velocity denoiser G_{T2V} is the most lightweight module, featuring fewer hidden state dimensions and transformer layers because the task it handles involves low-dimensional data. However, the introduction of a guidance mechanism for more accurate predictions results in G_{T2V} having the longest inference time.

Finally, to generate the motion sequences for a 35-second demo video consisting of nine clips under different conditions and with a total duration of 54 seconds, our method completed the body motion generation task in only 3.72 seconds. This fast generation time highlights the potential of our approach for real-time applications.

For a like-for-like comparison against competing methods, we additionally report end-to-end generation speed (seconds per sequence) under the concurrent multi-modal setting in Tab. 3 (Appendix G.1): MOCO is the fastest at 0.69 s/seq, ahead of STMC (0.75 s/seq) and the StableMoFusion-backbone variant (0.86 s/seq), confirming that MOCO attains its naturalness and synchronization advantages without incurring additional inference cost.

G Additional Experiments

Table 3: Comparison with per-step composition (STMC) and a stronger diffusion backbone (StableMoFusion). Best among generative methods is in **bold**.

	Naturalness		Text2Motion		Speech2Gesture		Speed
Method	MC↑	$\text{Jitter}_{\text{spine}}$ ↓	FID+↓	R1↑	FID-A↓	BC↑	s/seq↓
GT	−1.95	0.59	0.000	40.0	–	–	–
MOCO (Ours)	**−3.39**	**0.60**	0.862	24.6	**3.83**	2.72	**0.69**
STMC [12]	−3.84	0.63	**0.684**	**27.5**	4.44	2.22	0.75
StableMoFusion [6]	−3.91	0.69	0.752	27.2	4.83	**2.92**	0.86

G.1 Comparison with STMC and Stronger Backbones

This section presents experiments comparing MOCO against two additional baselines, with results reported in Tab. 3. We include STMC [12] because its single-modality-per-step composition inspired our multi-modal decoupled denoising, and StableMoFusion [6] to examine whether a backbone that is stronger on text-to-motion transfers into improved motion quality in our concurrent control setting. To adapt STMC to the multi-modal setting while faithfully preserving its original design, we use self-attention to process both text and audio conditions, share parameters across denoisers, and rely on an LLM to automatically annotate the assignment of body parts. In addition to the text-to-motion (FID+, R1) and speech-to-gesture (FID-A, BC) metrics used in the main paper, we further report two naturalness-oriented metrics: *MotionCritic* (MC) [14], a learned motion-quality score aligned with human perception, and $\text{Jitter}_{\text{spine}}$, the jerk of the spine joints at the boundary between the audio-controlled upper body and the text-controlled lower body, which quantifies how smoothly the two independently denoised regions are stitched together.

We report the results transparently. On pure text-to-motion fidelity, STMC is in fact slightly stronger than MOCO, and StableMoFusion is likewise competitive on these metrics. However, MOCO achieves better speech-to-gesture fidelity, boundary smoothness, and MotionCritic score. These results suggest that, rather than being uniformly best on every isolated single-modality metric, MOCO delivers the most natural and best-synchronized motion under joint text-and-speech control, as further corroborated by the pairwise user study in Appendix G.5. We note that MC should be interpreted only as an auxiliary indicator, because it is trained predominantly on text-driven motion, whereas MOCO produces dense, expressive co-speech upper-body gestures that may lie partly outside its training distribution.

Table 4: Evaluation of trajectory control under classifier-free guidance (CFG), OmniControl (OC) [15], and post-hoc L-BFGS optimization, on location (m) and orientation (rad). OmniControl is a dedicated spatial-control method used to replace our trajectory-to-velocity module G_{T2V}; the first three rows therefore use G_{T2V}.

Method			Location		Orientation	
CFG	OC	L-BFGS	Average Difference	Goal Difference	Average Difference	Goal Difference
✗	✗	✗	0.56	1.21	0.71	1.26
✓	✗	✗	0.57	1.32	0.83	1.51
✓	✗	✓	0.11	0.20	0.71	1.28
✗	✓	✗	0.44	0.79	0.53	0.97
✗	✗	✓	**0.08**	**0.13**	0.59	1.10
✗	✓	✓	0.14	0.20	**0.48**	**0.87**

G.2 Evaluation of Trajectory Control

Tab. 4 evaluates trajectory control methodologies by assessing the effects of classifier-free guidance (CFG), OmniControl (OC) [15], and L-BFGS optimization on both location (meters) and orientation (radians). Here, OmniControl is a dedicated spatial-control method that replaces our lightweight trajectory-to-velocity module G_{T2V}, whereas the other configurations build on G_{T2V}. For each category, two metrics are reported: Average Difference, the mean deviation between the generated trajectory and the ground truth (GT), and Goal Difference, the discrepancy at the final point relative to the GT.

The post-hoc L-BFGS optimization is the single most effective component: applied on top of G_{T2V}, it drastically reduces the location error while also modestly improving orientation. In contrast, CFG does not help and can even interfere with the optimization, slightly worsening both location and orientation. Replacing G_{T2V} with OmniControl improves orientation, and OC+L-BFGS achieves the overall best orientation—but at the cost of a larger location error and additional inference time. We therefore retain G_{T2V}+L-BFGS as our default for its better location accuracy and efficiency, and note that adopting a dedicated spatial controller is a worthwhile direction when orientation fidelity is prioritized.

We further emphasize, in the interest of transparency, that the absolute orientation accuracy of MOCO is still not fully satisfactory. We attribute this to a deliberate design trade-off: MOCO adopts a motion representation chosen to facilitate temporal stitching, which is prone to accumulated error along the trajectory—hence the orientation error remains comparatively large even after optimization.

G.3 Single Modality Performance

To demonstrate MOCO's performance in single-modality scenarios, we trained it from scratch on HumanML3D for text-to-motion and on BEAT2 for speech-to-gesture, respectively, ensuring a fair comparison. The results, presented in

Table 5: Quantitative results of text-to-motion generation on the HumanML3D test set.

Method	R-Precision			FID↓	MM Dist↓	Diversity↑	MM↑
	Top 1	Top 2	Top 3				
Ground Truth	0.511	0.703	0.797	0.002	2.974	9.503	-
T2M-GPT [20]	0.491	0.680	0.775	0.116	3.118	9.761	1.856
MDM [13]	-	-	0.611	0.544	5.566	9.559	2.799
FineMoGen [22]	0.504	0.690	0.784	0.151	2.998	9.263	2.696
MoMask [4]	0.521	0.713	0.807	0.045	2.958	-	1.241
LMM-Tiny [21]	0.496	0.685	0.785	0.415	3.087	9.176	1.465
LMM-Large [21]	0.525	0.719	0.811	0.040	2.943	9.814	2.683
SynTalker [1]	0.375	0.564	0.681	4.385	4.499	9.374	
MOCO (Ours)	0.434	0.618	0.720	0.530	3.563	9.856	2.663

Table 6: Quantitative results of speech-to-gesture generation on the BEATX test set.

Method	FGD↓	BC	Diversity↑	MSE↓	LVD↓
FaceFormer [2]	-	-	-	7.787	7.593
CodeTalker [16]	-	-	-	8.026	7.766
S2G [3]	28.15	4.683	5.971	-	-
Trimodal [19]	12.41	5.933	7.724	-	-
HA2G [10]	12.32	6.779	8.626	-	-
DisCo [7]	9.417	6.439	9.912	-	-
CaMN [9]	6.644	6.769	10.86	-	-
DiffStyleGesture [17]	8.811	7.241	11.49	-	-
TalkShow [18]	6.209	6.947	13.47	7.791	7.771
EMAGE [8]	5.512	7.724	13.06	7.680	7.556
ProbTalk [11]	6.170	8.099	10.43	8.990	8.385
SynTalker [1]	6.413	7.971	12.72	-	-
MOCO (Ours)	5.543	7.089	14.05	7.285	7.573

Tab. 5 and Tab. 6, show that in the HumanML3D text-to-motion benchmark (Tab. 5), our model achieves performance comparable to the widely-used MDM. This outcome is expected since our text-to-motion denoiser, G_{T2M}, is based on MDM. In the BEAT2 speech-to-gesture benchmark (Tab. 6), MOCO attains competitive performance compared to state-of-the-art methods.

G.4 Details of User Study

To construct the questionnaire for the user study, we first generated 20 videos using each method, including MOCO (Ours), Pseudo-Text, Weighted Sum, and Combine Once. The videos generated by MOCO were then concatenated with the other videos, resulting in 60 pairs of comparison videos. The user study involved 10 participants, each of whom evaluated all 60 pairs of comparison videos.

Participants were asked to select their preferred video or indicate that neither was better based on one of the following criteria: text following, beat synchronization, body coherence, and temporal fluidity. The first 40 video comparisons involved MOCO versus Pseudo-Text and MOCO versus Weighted Sum, focusing on text following and beat synchronization. The last 20 video comparisons involved MOCO versus Combine Once, focusing on body coherence and temporal fluidity. The statistical results are presented in the main paper.

G.5 Extended Pairwise-Preference Study against STMC.

The study above compares MOCO against the weighted-sum, pseudo-text, and combine-once baselines. To further corroborate the metric-level comparison with STMC reported in Appendix G.1 through human perception, we additionally conducted a larger-scale pairwise-preference study against STMC [12] (the same adapted STMC baseline as in Appendix G.1). This study involved **51 participants** drawn from a bachelor's-level (undergraduate) background, none of whom were involved in the project; each evaluated 60 comparison videos. We report participant background explicitly here for transparency, as the demographic composition of evaluators can influence perceptual judgments. For each video pair, participants chose the preferred clip—or "no preference"—along three criteria: text alignment, audio synchronization, and naturalness.

The results are summarized in Tab. 7. MOCO is preferred over STMC on all three criteria, with the margin reaching statistical significance for *audio synchronization* (47.1% vs. 39.4%, $p = 0.02$) and *naturalness* (46.1% vs. 39.8%, $p = 0.03$), using a two-sided binomial test with ties excluded. For *text alignment* the preference favors MOCO but does not reach significance (45.9% vs. 42.0%, $p = 0.11$), which is consistent with the metric-level observation in Appendix G.1 that STMC is competitive on text-to-motion metrics. Taken together, these perceptual results reinforce our central claim: under concurrent text-and-speech control, MOCO produces motion that human viewers judge to be better synchronized and more natural, even where it does not dominate on isolated text-to-motion metrics.

Table 7: Extended pairwise-preference user study comparing MOCO against STMC [12] under concurrent text-and-speech control. Each cell reports the fraction of comparisons in which participants preferred MOCO, expressed no preference, or preferred STMC. The p-value is from a two-sided binomial test on the preference between the two methods (ties excluded); * denotes statistical significance at $p < 0.05$.

Criterion	Prefer MOCO	No Preference	Prefer STMC	p-value
Text Alignment	**45.9%**	12.1%	42.0%	0.11
Audio Synchronization	**47.1%**	13.5%	39.4%	0.02*
Naturalness	**46.1%**	14.1%	39.8%	0.03*

H Details of Multi-Modal Benchmark

To effectively evaluate our proposed task, we developed a multi-modal benchmark comprising 1,000 test clips, following the methodology outlined in [12]. Each test clip is automatically generated and includes two text descriptions and two audio clips.

For the text descriptions, we manually curated a set of 40 texts focusing on lower-body movements commonly associated with speech delivery or conversation. These descriptions provide the necessary context for evaluating the corresponding movements within the benchmark. Regarding the audio clips, we selected recordings from the BEAT2 dataset, specifically choosing eight speakers with speaker IDs below 10. These audio files were segmented into clips using a Voice Activity Detector (VAD), resulting in 694 audio clips with an average duration of 9.14 seconds.

The 1,000 test clips were generated through an automated process that utilizes the curated text descriptions and audio clips. For each test clip, two text descriptions are randomly selected and assigned random durations. Subsequently, two neighboring audio clips are randomly chosen. The start times for both the text and audio intervals are determined randomly, allowing the sequence to commence with either text or audio. This process results in the creation of four intervals that correspond to the selected text descriptions and audio clips.

Optional text descriptions are listed here:

```
walk in a circle clockwise
walk in a circle counterclockwise
walk in a quarter circle to the left
walk in a quarter circle to the right
turn 180 degrees to the left on the left foot
turn 180 degrees to the left on the right foot
turn left
turn right
walk forwards
walk backwards
slowly walk forwards
slowly walk backwards
quickly walk forwards
quickly walk backwards
run
jogs forwards
jogs backwards
slowly walk in a circle
perform a squat
sit down
turn around then sit down in a chair
sit down then get back up and walk back
sit down for a moment
step back and sit down
sit down indian style
```

```
take a step to their right and sit down
sit criss cross
sit down on the ground and cross their legs
squat down
sit on a high object
sit on a barstool and rest their legs on the stool
take a large step and sits on a stool
get down on their knees
sit on the ground with his legs extended in front of him
walk up to a backwards chair and sit down on it with legs
  outstretched
sit down and adjust themselves
sit down and swap their legs crossing back and forth
sit and lie down on a lounge chair
sit down and lean on the chair
sits very still in the chair
```

References

1. Chen, B., Li, Y., Ding, Y.X., Shao, T., Zhou, K.: Enabling synergistic full-body control in prompt-based co-speech motion generation. In: Proceedings of the ACM International Conference on Multimedia. pp. 6774–6783 (2024)
2. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: FaceFormer: Speech-driven 3D facial animation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18770–18780 (2022)
3. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3497–3506 (2019)
4. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1900–1910 (2024)
5. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (2022)
6. Huang, Y., Yang, H., Luo, C., Wang, Y., Xu, S., Zhang, Z., Zhang, M., Peng, J.: Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. In: Proceedings of the ACM International Conference on Multimedia. pp. 224–232 (2024)
7. Liu, H., Iwamoto, N., Zhu, Z., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In: Proceedings of the ACM International Conference on Multimedia. pp. 3764–3773 (2022)
8. Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., Iwamoto, N., Zheng, B., Black, M.J.: Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling. arXiv preprint arXiv:2401.00374 (2023)
9. Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: European Conference on Computer Vision (ECCV). pp. 612–630. Springer (2022)

10. Liu, X., Wu, Q., Zhou, H., Xu, Y., Qian, R., Lin, X., Zhou, X., Wu, W., Dai, B., Zhou, B.: Learning hierarchical cross-modal association for co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10462–10472 (2022)
11. Liu, Y., Cao, Q., Wen, Y., Jiang, H., Ding, C.: Towards variable and coordinated holistic co-speech motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1566–1576 (2024)
12. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1911–1921 (2024)
13. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
14. Wang, H., Zhu, W., Miao, L., Xu, Y., Gao, F., Tian, Q., Wang, Y.: Aligning human motion generation with human perceptions. In: International Conference on Learning Representations (ICLR). vol. 2025, pp. 77626–77656 (2025)
15. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023)
16. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12780–12790 (2023)
17. Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Cheng, M., Xiao, L.: Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-23). pp. 5860–5868 (2023)
18. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 469–480 (2023)
19. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics **39**(6), 1–16 (2020)
20. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14730–14740 (2023)
21. Zhang, M., Jin, D., Gu, C., Hong, F., Cai, Z., Huang, J., Zhang, C., Guo, X., Yang, L., He, Y., et al.: Large motion model for unified multi-modal motion generation. In: European Conference on Computer Vision (ECCV). pp. 397–421. Springer (2024)
22. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. Advances in Neural Information Processing Systems (NeurIPS) **36**, 13981–13992 (2023)