PERSONAACT: A PERSONA-CENTRIC DATASET FOR CONVERSATIONAL MOTION GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: Examples from our PersonaAct dataset. Asset credits are provided in Appendix E.

ABSTRACT

For 3D digital characters to perform expressively, their motion should reflect not only what they say and do, but also their personalities. However, datasets that link persona descriptions to conversational motion remain limited, making it difficult to learn and evaluate how characters express their personalities through motion. To address this issue, we first present **PersonaAct**, a multimodal 3D motion dataset with 7,542 clips covering 302 personas. It pairs persona profiles and motion descriptions with synchronized speech and motion across different dialogue scenarios. Second, we evaluate motion quality and alignment with persona profiles, speech, and motion descriptions. To assess whether the motion fits the intended character, we introduce a persona–motion matching metric. Third, experiments show that current methods struggle to balance multiple conditions, particularly persona profiles and motion descriptions. These findings highlight key challenge in persona-conditioned motion generation. PersonaAct provides data and evaluation tools for developing and testing methods that address these challenges. Dataset visualizations and examples are available at `https://anonymous.4open.science/r/PersonaAct-ICLR/`.

1 INTRODUCTION

Recent text-to-motion methods have advanced motion generation from text descriptions (Guo et al., 2024; Zhang et al., 2023), and speech-driven methods generate body gestures and facial expressions coordinated with speech (Liu et al., 2024; Jin et al., 2026). While these advances enable motion generation from text and speech, bringing different characters to life also requires motion that reflects their personalities. Even a simple greeting can be performed in different ways. For example, a confident character may greet someone with a broad wave, while a shy character may give a small

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

wave with their head lowered. Such differences help bring digital characters to life in games, animation, and virtual conversations. However, generating conversational motion that reflects a specified persona while following speech and motion descriptions remains challenging.

Learning such performances faces two main challenges. First, data that link persona to character performances remain limited. Datasets such as BEAT (Liu et al., 2022) provide synchronized speech and motion with speaker and emotion labels, but lack detailed persona profiles. Other work represents personal styles through reference motions, as in PersonaBooth (Kim et al., 2025), or uses character descriptions for locomotion, as in MotionPersona (Shi et al., 2025). SuSuInterActs (Jin et al., 2026) provides conversational motion but focuses on a single character. Collecting motion across many roles is costly and time-consuming. Each role requires actors to study its profile and perform multiple scripts, adding to both preparation time and studio costs. Second, evaluating persona-conditioned conversational motion generation requires more than existing motion metrics. These metrics assess motion quality and alignment with speech and text descriptions, but do not measure whether the motion reflects the specified persona. A benchmark should therefore assess both alignment with the motion description and consistency with the specified persona.

To address these challenges, we introduce PERSONAACT, a Chinese multimodal 3D motion dataset for multi-character performance. It covers 302 personas, each with a profile describing its background, personality, and motion habits. We create dialogue scripts for each persona across different scenarios. Professional actors study each profile and perform the scripts following motion and expression cues. We capture synchronized speech, body and hand motion, and facial expressions in a motion-capture studio. The 1,467 dialogues yield 7,542 clips, each pairing a motion description with synchronized speech and motion. Building on PERSONAACT, we establish a body motion generation benchmark conditioned on persona profiles, speech, and motion descriptions. It evaluates motion quality and alignment with each input condition. We also propose a persona–motion matching metric to assess whether generated motion fits the given character. Baseline comparisons and ablations show that current methods still struggle to satisfy multiple conditions simultaneously. Stronger persona matching does not ensure alignment with motion descriptions. Joint adherence to multiple conditions remains a key challenge in persona-conditioned motion generation.

Our contributions are three-fold: **First**, we construct **PersonaAct**, a 3D motion dataset for multi-character performance, pairing persona profiles and dialogue scripts with motion, speech, and facial expressions. **Second**, we establish a motion generation benchmark with comprehensive metrics to assess motion quality and alignment with persona profiles, motion descriptions, and speech. **Third**, our experiments reveal that existing methods struggle to satisfy multiple conditions simultaneously, particularly persona profiles and motion descriptions. The findings highlight the demand for methods that jointly respect multiple conditions in motion generation.

2 RELATED WORK

2.1 CONVERSATIONAL MOTION DATASETS

Co-speech motion datasets pair speech with human motion to support conversational motion generation. Trinity (Ferstl & McDonnell, 2018) records a single speaker's gestures in a motion-capture studio, while Talking With Hands 16.2M (Lee et al., 2019) captures synchronized speech, body, and finger motion during spontaneous two-person conversations. Other datasets include information about who is speaking or how they move. PATS (Ahuja et al., 2020) records gestures from multiple speakers, ZeroEGGS (Ghorbani et al., 2023) provides speech and motion performed in different styles, and BEAT (Liu et al., 2022) combines multilingual speech with body, hand, and facial motion and emotion labels. SHOW (Yi et al., 2023) reconstructs body, hand, and facial motion from talking videos. SuSuInterActs (Jin et al., 2026) further pairs Chinese dialogue with synchronized motion and action descriptions for a single virtual character. These datasets support methods such as TalkSHOW (Yi et al., 2023) and EMAGE (Liu et al., 2024), which generate body, hand, and facial motion from speech. Persona-Gestor (Zhang et al., 2024) also learns speaker-specific gesture styles from speech audio. However, speaker identities, emotion labels, and action descriptions do not fully describe a character's personality. Generating conversational motion for different characters calls for data that connect such character descriptions with speech and motion.

Table 1: Comparison with representative motion datasets. Text denotes motion or behavior descriptions. B, H, and F denote body, hands, and facial motion.

Dataset	Source	Conditions	Motion	Hours	Persona Info.
Trinity (2018)	MoCap	Speech	B+H	4.1	–
PATS (2020)	Video	Speech	Upper-body+H (2D)	251	Speaker ID
BEAT (2022)	MoCap	Speech	B+H+F	76	Speaker/emotion labels
ZeroEGGS (2023)	MoCap	Speech	B+H	2.3	Style labels
SHOW (2023)	Video	Speech	B+H+F	26.9	–
SuSuInterActs (2026)	MoCap	Speech+Text	B+H+F	37	Single-character profile
PerMo (2025)	MoCap	Text	B	–	Actor/style labels
PersonaAct(Ours)	MoCap	Speech+Text	B+H+F	15.3	Text profile

2.2 Persona in Motion Datasets

Existing motion datasets describe style and personality in several ways. 100STYLE (Mason et al., 2022) records locomotion in 100 styles, while PsyMo (Cosma & Radoi, 2024) pairs gait recordings with personality traits measured through questionnaires. Inter-X (Xu et al., 2024) provides interaction motions with relationship annotations, Big-Five scores, and part-level descriptions. These datasets offer information about movement styles and personality, but do not include synchronized conversational speech. Other work represents personas through performers. PerMo (Kim et al., 2025) defines personas through actor–style combinations. Its accompanying method, PersonaBooth, learns personalized motion from reference sequences. Similarly, PersonaAnimator (Qian et al., 2025) transfers a person's motion style from reference videos while preserving the actions in a separate input video. MotionPersona (Shi et al., 2025) pairs motion with textual descriptions of character traits, but focuses on locomotion. RoleMotion (Peng et al., 2025), which pairs body and hand motion with text descriptions and scene-specific roles. These roles describe what a person does in a scene, rather than their personality.

Taken together, these datasets describe characters through style labels, performer identities, scene roles, or textual attributes, but few combine natural-language persona profiles with synchronized speech and motion. As shown in Table 1, PersonaAct pairs synchronized speech and motion with character-level persona profiles and clip-level motion descriptions, enabling models to learn not only what actions to perform, but also how to perform them in a way that reflects the character's persona.

3 Dataset

We construct PersonaAct, a Chinese multimodal 3D motion dataset, with two goals: to cover diverse personas and to capture conversational motion that reflects their personalities. The dataset contains 302 personas, each with a profile and dialogue scripts for different scenarios. We record performances of these scripts across 1,467 dialogues and divide the recordings into 7,542 clips. As shown in Figure 3(b), each clip contains synchronized speech, body and hand motion, and facial expressions, together with a speech transcript, a motion description, and an expression label.

3.1 Dataset Construction

As shown in Figure 2, we construct PersonaAct by designing personas from a range of categories and writing profiles and dialogue scripts that reflect their personalities. Actors study the profiles to understand their characters before performing the scripts. We record their speech, body and hand motion, and facial expressions synchronously in a motion-capture studio.

Persona Design. To cover different personalities, we select personas from broad role categories, as shown in Figure 3(a), including everyday occupations, historical figures, fictional characters, and etc. Each profile describes the character's background, personality, speaking style, and behavior in different situations. The profile also includes performance guidance, such as on posture style and habitual gestures. For example, Zhu Bajie is a lazy, heavyset character from *Journey to the West*. His profile guides actors to walk with heavy steps, and rub their belly. We review each profile for

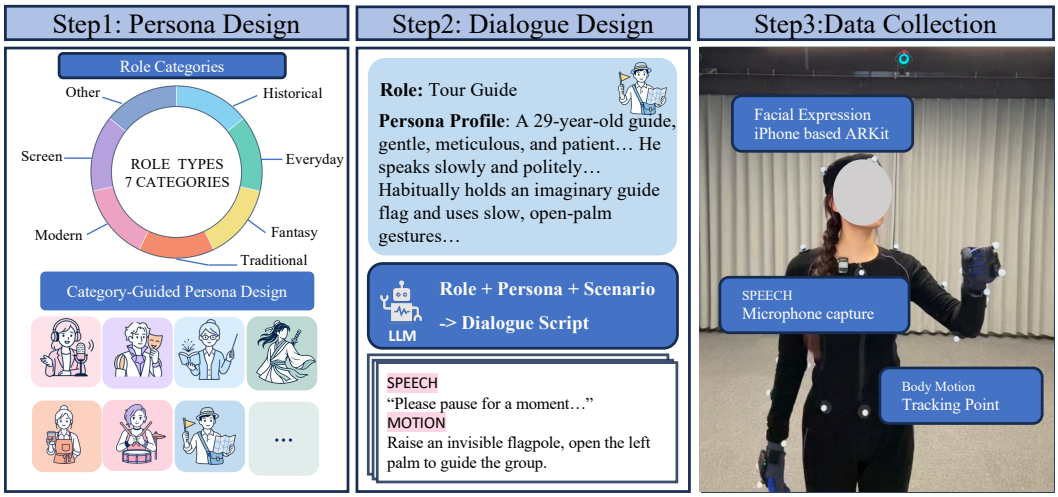


Figure 2: Overview of our dataset construction pipeline, from persona and dialogue design to synchronized recording of motion, facial expressions, and speech.

feasibility, checking that the required movement range and action difficulty are suitable for motion capture. We also provide a short version of each profile as input to motion generation models.

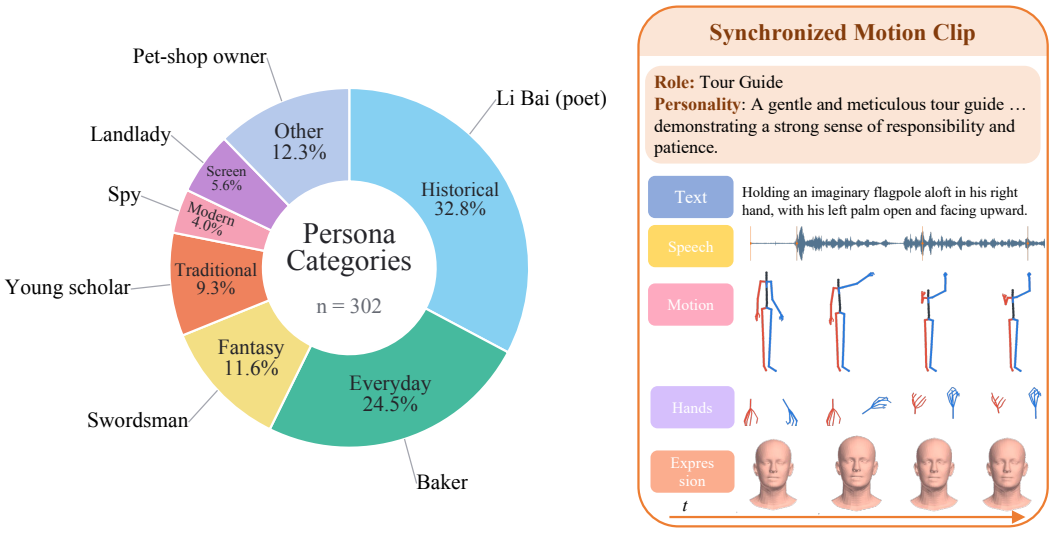
Dialogue Script Generation. For each persona, we design five dialogue scenarios with different situations, conversation partners, and goals. Given the persona and scenario, we generate a multi-turn Chinese dialogue. Each turn specifies the speaker, spoken text, a motion description, and an expression label. These motion descriptions and expression label use concrete, performable actions and guide actors before capture. Human reviewers check the scripts for consistency with the persona, and feasible actions. They revise or discard scripts that do not meet these requirements. Each script is a two-person dialogue, but we capture only the target character's performance. The other speaker's lines help the actor understand the situation and respond in a way that fits the character.

Motion and Facial Expression Capture. Professional actors study each character's profile and script before playing the role. We record full-body motion, including hand gestures, with optical motion capture at 30 FPS. Facial expressions are captured with iPhone ARKit as 52 blendshape (Apple, n.d.) coefficients at 30 FPS. We also record speech as 24-kHz mono audio. We send simultaneous start and stop signals to all recording devices to keep the captured streams temporally aligned.

Data Representation and Quality Control. We use a common feature format for model training and evaluation. Each frame contains 393 motion features (153 for the body and 120 per hand) and 52 facial blendshape coefficients. To ensure complete recordings and accurate annotations, we combine automatic checks with manual review. Automatic checks remove clips with missing modalities, corrupted files, invalid dimensions, or mismatched motion and facial frame counts. Manual review checks motion quality and whether annotations match the performance. After filtering, we retain 7,542 valid clips. Further details are provided in Appendix A.

3.2 Dataset Statistics

PersonaAct provides comparable amounts of recorded motion across most personas. The median duration is 2.92 minutes per persona, with the middle 80% ranging from 2.0 to 4.3 minutes (Figure 4(a)) for a persona. The dataset contains short performance clips with concise text annotations. Clips have a median duration of 6.85 seconds, and 90% last 3.9–12.2 seconds (Figure 4(b)). The median text lengths are 23 Chinese characters for spoken text and 16 for motion descriptions (Figure 4(c)). The recorded performances also reflect differences in character personality. We manually label personas as expressive or reserved, and as organized or easygoing. We compare body expansiveness and upper-body velocity. The former measures hand reach and span relative to shoulder width, while the latter measures average upper-body joint speed. For each persona, we take the median of each measure across its clips. As shown in Figure 5, expressive personas tend to show greater body expansiveness and faster upper-body motion than reserved personas. Organized and



(a) Distribution of persona categories.

(b) Multimodal content of one clip.

Figure 3: Overview of PERSONAACT. (b) Each clip contains a persistent persona profile, a motion description, synchronized speech, and aligned body, hand, and facial motion. (a) The dataset covers 302 personas across seven broad role types; representative personas are annotated outside the ring.

easygoing personas also differ on these measures. These differences highlight the variety of character performances captured in PersonaAct.

4 TASK AND EVALUATION

Building on our dataset, we propose a task of persona-conditioned motion generation and define evaluation metrics to assess both motion quality and consistency with the input conditions.

4.1 TASK DEFINITION

Given speech audio $\mathbf{A}$, a motion description $\mathbf{D}$, and a persona profile $\mathbf{P}$, we generate a body motion sequence:

$$\hat{\mathbf{M}} = G(\mathbf{A}, \mathbf{D}, \mathbf{P}).$$

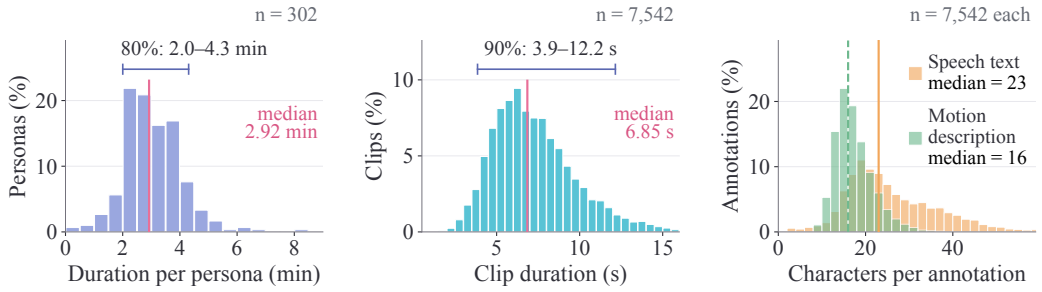
Speech provides the spoken content and prosody. The motion description specifies the intended actions in the clip, while the persona profile guides how the character performs them. We represent each motion frame with 3D root translation and continuous 6D rotations of 25 body joints, giving 153 dimensions per frame. We provide synchronized hand and facial motion in the dataset but do not include them in the generation target.

4.2 EVALUATION METRICS AND PROTOCOLS

To provide a comprehensive evaluation, we assess both motion quality and how well the generated motion matches each input condition: speech, motion descriptions, and persona profiles. Implementation details are provided in Appendix B.

Motion Quality. We use *Fréchet Inception Distance (FID)* and *Diversity* following prior work (Guo et al., 2024; Jin et al., 2026). FID compares the feature distributions of generated and real motion; lower is better. Diversity measures motion variation through the average distance between sampled motion features. We report real-motion Diversity as a reference. To assess temporal smoothness, we additionally report *Jerk Ratio*, the ratio of mean jerk in generated and real motion. A value close to 1.00 indicates a jerk magnitude similar to real motion.

Motion–Description Alignment. We evaluate motion–description alignment through text-to-motion retrieval using a trained evaluator. Each description queries its corresponding generated



(a) Persona duration distribution (b) Clip duration distribution (c) Text length distribution

Figure 4: Statistics of PersonaAct. We show the distributions of (a) total motion duration per persona, (b) clip duration, and (c) speech transcript and motion description lengths, measured in Chinese characters.

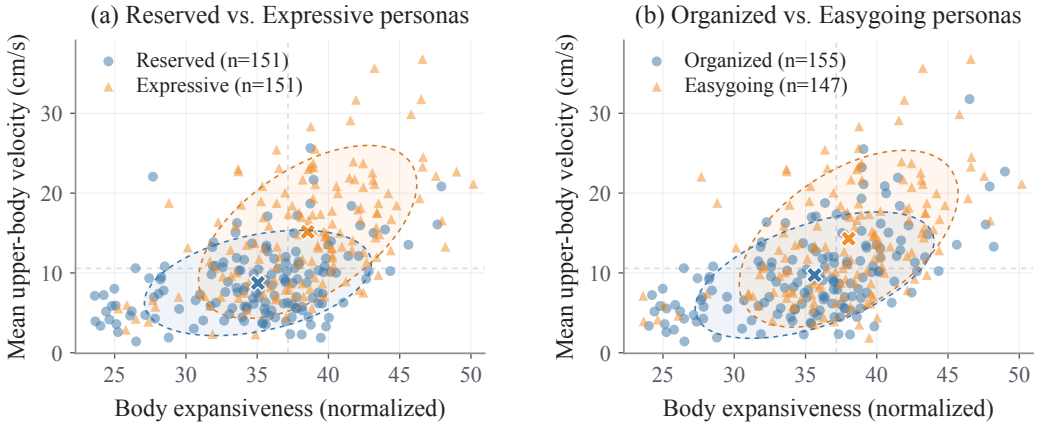


Figure 5: Persona–motion statistics for 302 personas. Both panels use the same motion features with different persona groupings; each point represents one persona, and crosses mark group means.

motion among 32 candidates. We report R@1, R@2, and R@3, the proportions of queries whose corresponding motion ranks in the top 1, 2, and 3, respectively; higher is better.

Persona–Motion Alignment. We use a persona–motion evaluator to assess alignment at the clip level. We encode body motion with a Transformer-based motion encoder and short persona profiles with a frozen Chinese BERT encoder. A matching MLP then predicts a compatibility score from the two embeddings. We train the evaluator only on the training split. At evaluation time, we compare each clip's target persona profile with 10 negative profiles sampled from the other personas. Persona–Motion Accuracy (P–M Acc.) measures how often the target profile receives a higher score than a negative profile for the same motion clip. We report it as a percentage; higher is better. Appendix B provides architecture and training details.

Speech–Motion Coordination. Event Sync Distance (ESD) measures the symmetric nearest-neighbor distance between audio-beat times and motion velocity peak times. We report the mean clip-level distance in seconds, lower is better.

5 Experiments

5.1 Experimental Setup

Baselines. We select representative methods from three groups: EMAGE (Liu et al., 2024) and DiffSHEG (Chen et al., 2024) for audio-conditioned generation; HY-Motion (Wen et al., 2025), T2M-GPT (Zhang et al., 2023), and MoMask (Guo et al., 2024) for text-conditioned generation; and SentiAvatar (Jin et al., 2026) for joint audio–text generation. Audio-conditioned methods generate motion from speech, while text-conditioned methods receive persona profiles and motion descrip-

Table 2: Quantitative comparison on PersonaAct. The best results are in **bold** and the second-best are underlined.

Method	R@1↑	R@2↑	R@3↑	FID↓	ESD↓	Diversity	Jerk Ratio	P-M Acc.↑
Real Motion	47.62	62.09	69.38	0.000	0.4561	22.560	1.000	83.85
Audio-only methods								
EMAGE (Liu et al., 2024)	4.00	8.12	12.27	482.431	0.5941	2.270	**1.091**	51.27
DiffSHEG (Chen et al., 2024)	8.91	16.53	22.15	53.716	0.3047	21.063	3.482	74.17
Text-only methods								
HY-Motion (Wen et al., 2025)	5.91	10.87	14.86	219.228	0.3674	16.231	1.198	53.68
T2M-GPT (Zhang et al., 2023)	25.59	37.49	45.54	**21.013**	0.5236	**22.451**	7.023	84.59
MoMask (Guo et al., 2024)	24.21	35.83	43.49	24.568	**0.2414**	22.086	7.760	**89.51**
Audio + Text methods								
SentiAvatar (Jin et al., 2026)	**39.30**	**53.66**	**61.21**	25.846	0.4505	22.213	2.047	80.94

tions. SentiAvatar uses all three conditions, allowing us to examine how models combine speech, persona information, and the requested actions.

Implementation Details. We divide PersonaAct into training, validation, and test sets with an approximate 70/15/15 ratio, keeping all clips from the same dialogue in one split. We train or fine-tune all baselines on the training set except HY-Motion, which we evaluate zero-shot. We use fine-tuned SentiAvatar for the conditioning ablations and evaluate all methods using the metrics in Section 4.2. Appendix C provides baseline adaptations and fine-tuning settings.

5.2 Main Results

Table 2 reports the quantitative results on PersonaAct. SentiAvatar, which accepts all three input conditions, achieves relatively balanced results across motion quality and condition alignment. It obtains the highest motion-description alignment, with an R@1 of 39.30%, while maintaining a competitive FID and a substantially lower jerk ratio than MoMask and T2M-GPT. Audio-only methods show different strengths: EMAGE achieves the jerk ratio closest to real motion but produces limited motion diversity and a high FID, while DiffSHEG achieves a low ESD but performs less well in motion-description alignment and persona matching. Among text-only methods, MoMask achieves the highest P-M accuracy and the lowest ESD, yet also has the highest jerk ratio. T2M-GPT obtains the lowest FID but falls behind SentiAvatar in motion-description alignment. These results suggest that existing methods, when adapted to PersonaAct, leave room for improvement in jointly following persona profiles, motion descriptions, and speech.

The challenge extends beyond combining different input modalities: models also struggle to follow persona profiles and motion descriptions together, even though both are provided as text. Table 2 shows that MoMask and T2M-GPT exceed real motion in P-M accuracy (89.51% and 84.59% versus 83.85%), yet achieve only 24.21% and 25.59% R@1. Figure 6 illustrates this contrast. In (a), MoMask raises an arm into a theatrical pose rather than lowering both arms as described. In (b), T2M-GPT holds both arms near the shoulders, resembling the persona's equipment-carrying posture, instead of slapping the thigh and pointing toward the ground. Both models produce motions consistent with the persona profiles but not aligned with the motion description. These examples show that, under the evaluated setups, motions can reflect the persona profiles while missing the description, motivating more effective methods for joint persona and motion-description control.

5.3 Conditioning Ablations

We conduct conditioning ablations on SentiAvatar to assess the contribution of each input condition. We remove speech, motion descriptions, and persona profiles one at a time during both fine-tuning and inference, keeping the evaluators and reference data fixed. Table 3 reports the results.

As Table 3 shows, removing persona profiles lowers P-M accuracy from 80.94% to 67.57%, showing that persona profiles help the model generate character-consistent motion. Removing motion

Table 3: Conditioning ablations of fine-tuned SentiAvatar. The best results are in **bold** and the second-best are underlined.

Variant	R@1↑	R@2↑	R@3↑	FID↓	ESD↓	Diversity	Jerk Ratio	P-M Acc.↑
Real Motion	47.62	62.09	69.38	0.000	0.4561	22.560	1.000	83.85
Full	39.30	53.66	61.21	25.846	0.4505	22.213	2.047	80.94
w/o persona	**41.08**	**55.87**	**64.13**	29.401	**0.4434**	22.080	2.013	67.57
w/o speech audio	39.72	53.66	61.74	**18.308**	0.6024	**22.214**	**1.758**	**83.24**
w/o motion description	9.09	16.55	22.83	27.325	0.4608	22.086	1.909	81.47

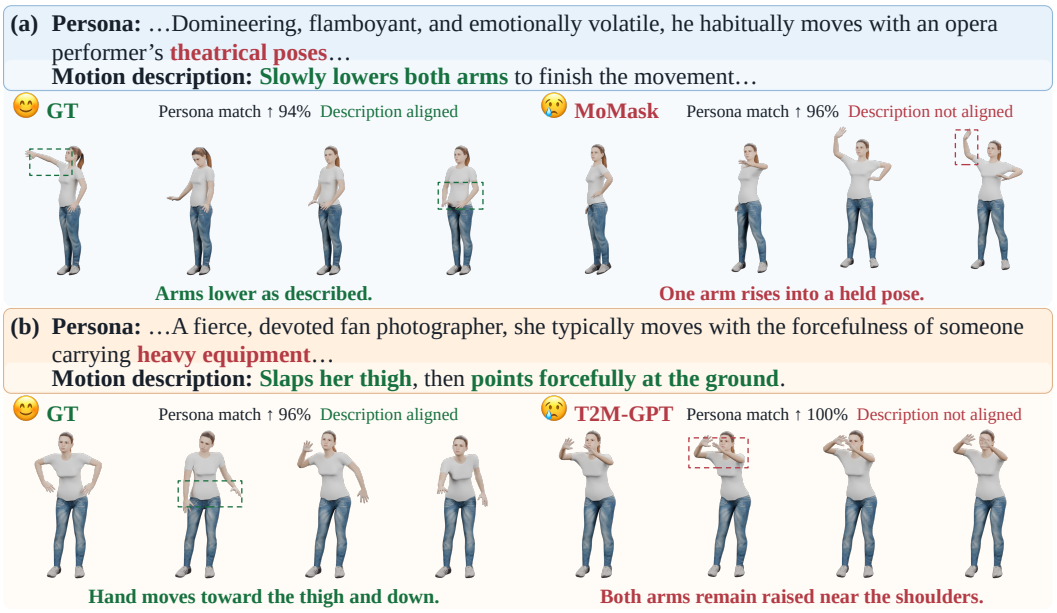


Figure 6: MoMask and T2M-GPT generate motions that match the personas but not aligned with the descriptions, illustrating the challenge of following both conditions.

descriptions reduces R@1 from 39.30% to 9.09%, confirming their importance for specifying the intended actions. Removing speech increases ESD from 0.4505 to 0.6024 seconds, supporting its role in coordinating motion with speech. These results demonstrate the value of all three conditions for our task.

The key challenge, however, lies in satisfying these conditions simultaneously. Additional MoMask ablations show that removing persona profiles improves R@1 but reduces P–M accuracy (Appendix D). Thus, the trade-offs observed across baselines also appear within a fixed architecture. Together with the SentiAvatar ablations, these results highlight the difficulty of jointly following multiple conditions, particularly persona profiles and motion descriptions.

5.4 User Study

We conduct user studies to verify the quality of PersonaAct and the reliability of our P-M evaluator. As shown in Figure 7.

For dataset persona consistency, five participants view 50 clips from dataset with full-body motion, and speech, then choose between the target persona and a sampled negative persona, with an uncertain option. The 50 clips and each negative persona are randomly sampled from PersonaAct. Participants select the target persona in 94.1% of judgments excluding uncertain responses. This suggests that the clips convey the intended personas, supporting the quality of the dataset.

For evaluator–human agreement, we conduct separate experiments on real motions from PersonaAct and motions generated by baseline methods. In both experiments, five participants view only the motion and choose the better-matching persona from the target profile and one sampled negative,

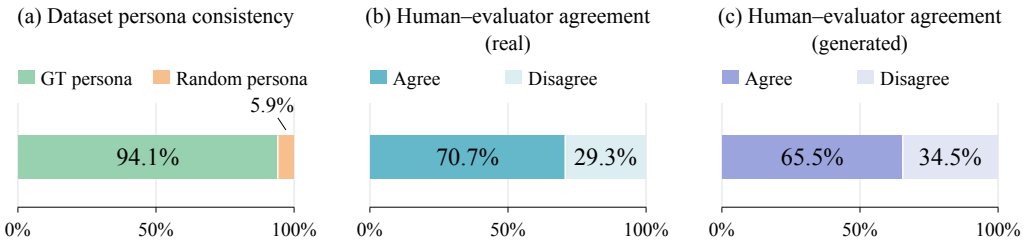


Figure 7: Human evaluation of dataset quality and P-M reliability. Participants judge whether clips match their persona profiles (a), and we compare their choices with P-M predictions on real (b) and generated (c) motions.

with an uncertain option. The clips and each negative persona are randomly sampled from PersonaAct. On real motions, participants evaluate randomly 50 ground-truth clips, and P-M agrees with the human majority on 70.7% of resolved cases. On generated motions, we select 20 test outputs from each of three representative methods—MoMask, SentiAvatar, and EMAGE—for a total of 60 clips. P-M achieves 65.5% agreement with the human majority. These results support the use of P-M to evaluate persona–motion alignment on both real and generated motions.

6 Conclusion

We introduced PersonaAct, a multimodal 3D motion dataset and benchmark for multi-character performance. It pairs persona profiles and motion descriptions with synchronized speech, body and hand motion, and facial expressions. We also introduced a persona–motion matching metric to assess whether generated motion fits the intended character. Baseline comparisons and ablations show that current methods still struggle to follow persona profiles and motion descriptions together. Motion can reflect a persona while missing actions specified in the description. Future work could develop methods that preserve the requested actions while adapting motion to each character. These methods should also keep motion coordinated with speech. Another direction is joint body, hand, and facial motion generation using the synchronized recordings in PersonaAct.

AI use statement

We used Large Language Model (LLM) to assist in drafting persona profiles and dialogue scripts, which human reviewers checked and revised. We also used LLM to assist with translation, manuscript drafting and editing. The authors reviewed and revised outputs and take responsibility for the final content, experimental results, and conclusions of this work.

Ethics Statement

Our human studies evaluate persona consistency in dataset clips and agreement between the P-M evaluator and human judgments. All performers agreed to the release of their processed recordings. The processed dataset will be released for academic research only.

Reproducibility Statement

We describe the dataset construction, generation task, and evaluation protocols in the main paper. The appendix provides details of temporal processing and data integrity checks, evaluator architectures and training objectives, and baseline adaptations and fine-tuning settings.

References

Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*, pp. 248–265. Springer, 2020.

Apple. ARFaceAnchor.BlendShapeLocation. Apple Developer Documentation, n.d. URL https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation. Accessed: September 24, 2026.

Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7352–7361. IEEE, 2024.

Adrian Cosma and Emilian Radoi. PsyMo: A dataset for estimating self-reported psychological traits from gait. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4591–4601. IEEE, 2024.

Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th international conference on intelligent virtual agents*, pp. 93–98, 2018.

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, volume 42, pp. 206–216. Wiley Online Library, 2023.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1910. IEEE, 2024.

Pengcheng Guo and Shixing Liu. chinese_speech_pretrain, 2022. URL https://github.com/TencentGameMate/chinese_speech_pretrain.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Chuhao Jin, Rui Zhang, Qingzhe Gao, Haoyu Shi, Dayu Wu, Yichen Jiang, Yihan Wu, and Ruihua Song. SentiAvatar: Towards expressive and interactive digital humans. *arXiv preprint arXiv:2604.02908*, 2026.

Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, and Hyung Jin Chang. PersonaBooth: Personalized text-to-motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22756–22765, 2025.

Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. Talking with Hands 16.2M: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 763–772, 2019.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, 2022.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1144–1154, 2024.

Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–18, 2022.

Junran Peng, Yiheng Huang, Silei Shen, Zeji Wei, Jingwei Yang, Baojie Wang, Yonghao He, Chuanchen Luo, Man Zhang, Xucheng Yin, and Wei Sui. RoleMotion: A large-scale dataset towards robust scene-specific role-playing motion synthesis with fine-grained descriptions. *arXiv preprint arXiv:2512.01582*, 2025.

Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9454–9463. IEEE, 2023.

Ziyun Qian, Runyu Xiao, Shuyuan Tu, Wei Xue, Dingkang Yang, Mingcheng Li, Dongliang Kou, Minghao Han, Zizhi Chen, and Lihua Zhang. Personaanimator: Personalized motion transfer from unconstrained videos. *arXiv preprint arXiv:2508.19895*, 2025.

Mingyi Shi, Wei Liu, Jidong Mei, Wangpok Tse, Rui Chen, Xuelin Chen, and Taku Komura. MotionPersona: Characteristics-aware locomotion control. *arXiv preprint arXiv:2506.00173*, 2025.

Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, Ningxin Jiao, Changhai Chen, Weijie Chen, Yiran Wang, et al. Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-X: Towards versatile human-human interaction analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22260–22271. IEEE, 2024.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese CLIP: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 469–480. IEEE, 2023.

Fan Zhang, Zhaohan Wang, Xin Lyu, Siyuan Zhao, Mengjian Li, Weidong Geng, Naye Ji, Hui Du, Fuxing Gao, Hao Wu, et al. Speech-driven personalized gesture synthetics: Harnessing automatic fuzzy feature inference. *IEEE Transactions on Visualization and Computer Graphics*, 30(10): 6984–6996, 2024.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14730–14740, 2023.

Appendix

A Temporal Alignment and Data Integrity

The final release contains 7,542 clips from 1,467 dialogues and 302 personas, with 15.303 hours of motion and speech audio.

A.1 Temporal Processing

We match speech, motion, and facial recordings from the same recording take. We trim the initial recording cue when detected and crop all modalities to their shared time range. We resample joint rotations with spherical linear interpolation and interpolate translations, joint positions, and facial coefficients linearly. We use facial timestamps to resample the facial recordings.

We sample motion and facial data on a uniform 20-FPS grid and crop the 24-kHz audio to the same interval. For a clip with T motion frames, we retain $1{,}200T$ audio samples, giving both modalities a duration of $T/20$ seconds. Table 4 reports checks on the processed files; equal durations do not measure sensor timing offsets.

A.2 Release Integrity Checks

We check every released clip for readable, nonempty files, complete annotations, valid array dimensions, finite values, and matching frame counts across body, hands, joint positions, and facial motion. All clips pass these checks. We find no entirely static body sequences, all-zero audio clips, or exact duplicate body arrays or audio waveforms. No clip or dialogue appears in multiple splits, and unseen personas are disjoint from training personas and from the other unseen split.

Table 4: Integrity and duration checks on the final 7,542 clips.

Check	**Result**
Nonempty motion, face, and audio files	7,542 / 7,542
Nonempty speech, motion description, and expression	7,542 / 7,542
Clips linked to persona profiles	7,542 / 7,542
Expected array shapes and finite values	7,542 / 7,542
Equal frame counts across all five output arrays	7,542 / 7,542
Readable, nonempty 24-kHz mono 16-bit audio	7,542 / 7,542
Total motion / audio duration	15.303 / 15.303 h
Clips with matching audio and motion durations	7,542 / 7,542

B Evaluation Metrics

This section describes the evaluators and metrics used in our benchmark. Following previous work (Petrovich et al., 2023; Heusel et al., 2017), we train a motion encoder to extract clip-level features for FID, Diversity, and motion–description retrieval. We also use it to initialize the persona–motion evaluator. We train the evaluators on the training set, select them on the validation set, and keep them fixed when comparing generation methods.

Motion encoder. Following TMR (Petrovich et al., 2023), we encode each motion clip into a 256-dimensional feature vector. Each frame contains 153 body-motion features at 20 fps. We standardize these features using training-set statistics and linearly project them to 256 dimensions. We prepend a learnable token and add sinusoidal positional encodings:

$$\mathbf{H} = \mathrm{Transformer}([\mathbf{q}; W\widetilde{\mathbf{x}}_1 + \mathbf{b}; \ldots; W\widetilde{\mathbf{x}}_T + \mathbf{b}] + \mathbf{E}_{\mathrm{pos}}), \qquad \mathbf{z}_M = \mathbf{H}_0, \tag{1}$$

where $\widetilde{\mathbf{x}}_t$ is the standardized feature vector at frame t, $\mathbf{q}$ is the learnable token, and $\mathbf{E}_{\mathrm{pos}}$ denotes positional encodings. The output at the prepended token, $\mathbf{z}_M$, represents the clip. The Transformer has six layers, four attention heads, a feed-forward dimension of 1,024, and dropout of 0.1.

Motion–description alignment. We jointly train the motion encoder, a Chinese BERT text encoder with a projection head, and a motion decoder to learn aligned text and motion representations. Given a batch of B paired motion clips and descriptions, we use a symmetric contrastive loss:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(c_{ii}/\tau)}{\sum_{j=1}^{B}\exp(c_{ij}/\tau)} + \log\frac{\exp(c_{ii}/\tau)}{\sum_{j=1}^{B}\exp(c_{ji}/\tau)}\right], \tag{2}$$

where c_{ij} is the cosine similarity between the motion embedding $\mathbf{z}_{M_i}$ and text embedding $\mathbf{z}_{D_j}$, and $\tau = 0.1$. We also reconstruct the motion from both embeddings and encourage paired embeddings to be close:

$$\mathcal{L}_{\text{rec}} = \ell(G(\mathbf{z}_M), M) + \ell(G(\mathbf{z}_D), M), \tag{3}$$

$$\mathcal{L}_{\text{align}} = \ell(\mathbf{z}_M, \mathbf{z}_D), \tag{4}$$

where G is the motion decoder and ℓ denotes the mean Smooth L1 loss. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{rec}} + 10^{-5}\mathcal{L}_{\text{align}}. \tag{5}$$

At evaluation time, each description queries its paired motion among 32 candidates using cosine similarity between L2-normalized embeddings.

FID and Diversity. We extract motion features using the learned motion encoder above and standardize both real and generated features using the real-motion feature mean and standard deviation. Let (μ_r, Σ_r) and (μ_g, Σ_g) denote the means and covariances of the standardized real and generated features. We compute

$$\text{FID} = \|\mu_g - \mu_r\|_2^2 + \text{Tr}\Big(\Sigma_g + \Sigma_r - 2(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2})^{1/2}\Big). \tag{6}$$

Lower FID indicates closer feature distributions. Diversity measures the average distance between $L = 300$ randomly sampled feature pairs:

$$\text{Diversity} = \frac{1}{L}\sum_{\ell=1}^{L}\|\mathbf{z}_{a_\ell} - \mathbf{z}_{b_\ell}\|_2, \tag{7}$$

where $\mathbf{z}$ denotes a standardized motion feature. We report real-motion Diversity as a reference. For these metrics, the encoder uses the first 224 frames of each sequence, with masked padding for shorter sequences.

Persona–motion accuracy. We initialize the motion encoder from the motion–description alignment model and fine-tune it for persona matching. A frozen Chinese BERT encodes each persona profile, and a projection head maps its mean-pooled features to 256 dimensions. Given the L2-normalized motion and persona embeddings $\widehat{\mathbf{z}}_M$ and $\widehat{\mathbf{z}}_P$, we compute their compatibility score as

$$s(M, P) = \text{MLP}([\widehat{\mathbf{z}}_M; \widehat{\mathbf{z}}_P; |\widehat{\mathbf{z}}_M - \widehat{\mathbf{z}}_P|; \widehat{\mathbf{z}}_M \odot \widehat{\mathbf{z}}_P]), \tag{8}$$

where $[\cdot;\cdot]$ denotes concatenation and $\odot$ denotes elementwise multiplication. The MLP has dimensions $1024 \to 256 \to 1$. We train the motion encoder, text projection head, and matching MLP using balanced binary cross-entropy and a pairwise margin-ranking loss.

At evaluation time, we compare each clip's target persona against $K = 10$ distinct negative profiles sampled uniformly from the remaining personas in the 302-persona pool. P-M accuracy measures how often the target profile receives a higher score:

$$\text{P-M Acc.} = \frac{100}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K} h\big(s(M_i, P_i) - s(M_i, P_{ij}^{-})\big), \tag{9}$$

where N is the number of clips, and $h(x)$ equals 1, 0.5, or 0 when x is positive, zero, or negative, respectively. We average the results over five fixed negative-sampling sets shared across methods. Higher scores indicate better persona matching.

Jerk Ratio. Let $\mathbf{p}_{itj}$ denote the root-relative position of body joint j at frame t of a clip with T_i frames. We exclude the pelvis and use the remaining $J = 24$ joints. With $\Delta t = 1/20$ seconds and the third finite difference $\Delta^3 \mathbf{p}_{itj} = \mathbf{p}_{i,t+3,j} - 3\mathbf{p}_{i,t+2,j} + 3\mathbf{p}_{i,t+1,j} - \mathbf{p}_{itj}$, we compute

$$q_i = \frac{1}{J(T_i - 3)} \sum_{j=1}^{J} \sum_{t=1}^{T_i-3} \left\| \frac{\Delta^3 \mathbf{p}_{itj}}{\Delta t^3} \right\|_2, \qquad \text{Jerk Ratio} = \frac{\text{mean}_i \, q_i^{\text{gen}}}{\text{mean}_i \, q_i^{\text{real}}}. \tag{10}$$

A ratio close to one indicates a jerk magnitude similar to real motion.

Event Sync Distance. Let $\mathcal{A}_i$ contain audio-beat times and $\mathcal{B}_i$ contain motion-activity peak times, both in seconds. We compute

$$\text{ESD}_i = \frac{1}{2} \left(\frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} \min_{b \in \mathcal{B}_i} |a - b| + \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} \min_{a \in \mathcal{A}_i} |b - a| \right). \tag{11}$$

We extract audio beats with librosa and motion activity from the norm of consecutive differences in the body-motion representation. We retain activity peaks above the mean plus 0.2 standard deviations and average ESD_i across clips. If both event sets are empty, we assign zero; if only one is empty, we assign two seconds. Lower ESD indicates closer event timing; real-motion ESD provides a reference.

C Baseline Implementation Details

SentiAvatar fine-tuning. We initialize from the released SentiAvatar checkpoint and fine-tune all parameters of its language-model motion planner, including token embeddings and the output head. The audio encoder and quantizer, motion infilling Transformer, and RVQ motion tokenizer/decoder remain frozen. Chinese short rolecards and motion descriptions are combined with speech tokens extracted using the released Chinese-HuBERT-base (Guo & Liu, 2022) encoder and quantizer. Supervised fine-tuning predicts sparse motion tokens, with condition tokens excluded from the loss. We use the final 20-epoch checkpoint; the main settings are listed in Table 5.

Table 5: SentiAvatar fine-tuning configuration.

Setting	Value
Trainable module	Language-model motion planner
Optimizer	AdamW
Learning rate	5×10^{-5}, cosine decay
Batch size	128
Training epochs	20
Maximum sequence length	4,096 tokens

Text-conditioned baselines. We adapt MoMask and T2M-GPT to the 153-dimensional body representation at 20 fps. To encode Chinese persona and motion descriptions, we replace the original CLIP text encoder with a frozen Chinese-CLIP (Yang et al., 2022) encoder (`OFA-Sys/chinese-clip-vit-base-patch16`). The concatenated short rolecard and motion description are encoded into a 512-dimensional feature vector that conditions the motion generator.

Audio-conditioned baselines. DiffSHEG uses its gesture-only diffusion branch with a 153-dimensional body output, conditioned on frozen Chinese-HuBERT (Guo & Liu, 2022) features (`TencentGameMate/chinese-hubert-large`) and mel-spectrograms. It receives no persona or motion-description text. For EMAGE, we adapt the upper-body, lower-body, and global-motion codecs to the body skeleton, using forward-kinematics joint supervision in place of SMPL-X mesh supervision.

Zero-shot HY-Motion. We retain the released model and its text encoders, using the Chinese short rolecard and motion description as input. Native outputs are retargeted to the benchmark skeleton. Inference uses 50 Euler steps, guidance scale 5, and the model's 12-second duration limit.

D Additional Conditioning Ablations

Table 6 reports conditioning ablations on MoMask. Removing persona profiles increases R@1 from 24.21% to 37.87% but reduces P-M accuracy from 89.51% to 69.66%. Removing motion descriptions lowers R@1 to 11.85%, while P-M accuracy rises to 91.30%. These results show that each text condition helps the model follow its corresponding input, but combining them remains challenging even within a single input modality.

Table 6: Conditioning ablations of MoMask. The best results are in **bold** and the second-best are underlined.

Variant	R@1↑	R@2↑	R@3↑	FID↓	ESD↓	Diversity	Jerk Ratio	P-M Acc.↑
Full	24.21	35.83	43.49	24.568	0.2414	22.086	7.760	89.51
w/o persona	**37.87**	**53.35**	**62.15**	**23.721**	**0.2331**	21.939	8.773	69.66
w/o motion description	11.85	20.57	27.49	30.562	0.2444	**22.100**	**7.178**	**91.30**

E Asset Credits

The drum-set model in Fig. 1 is adapted from "drum set" by zephcrazy, licensed under CC BY 3.0. Materials and placement were modified.

F QUALITATIVE EXAMPLES

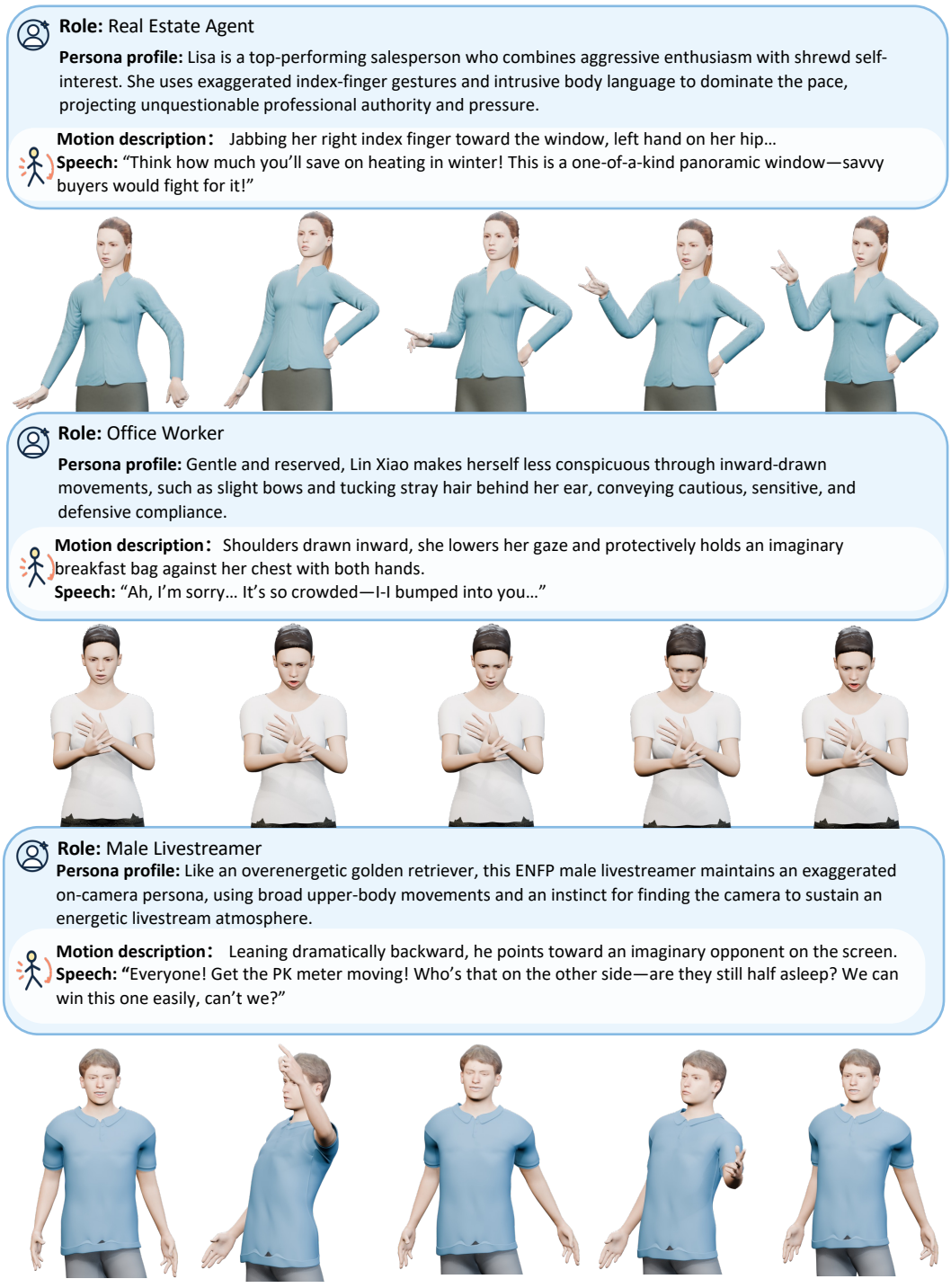


Figure 8: Examples of recorded performances with persona traits, motion descriptions, and spoken content. Successive poses illustrate the motion within each clip.