Digital Life Project: Autonomous 3D Characters with Social Intelligence

Supplementary Material

A. Overview of the Supplementary Material

We provide more details of DLP, such as the method (Appendix B), the DLP-MoCap dataset (Appendix C), the experiments (Appendix D) and additional discussions (Appendix E).

Demo Video. We produce a video demonstration[1] of DLP. In addition to "<motion>", explore using the "<speech>" token to generate audio with OpenAI's text-to-speech engine, followed by Talkshow [54] to synthesize face movements.

B. Additional Details of DLP

B.1. Scene

We aim to build 3D characters that articulate with 3D body motions in a 3D scene. In this work, we design our scene to achieve a diverse combination of relative interaction poses: 1) the *center* or *bookshelf* for the cases when both characters are standing (standing-standing interaction), 2) a *sofa* for side-by-side seated interaction, 3) a *dining table* for face-to-face seated interaction, and 4) a computer *desk* for standing-seated interaction. For each piece of interactable furniture, we designed *spots* with positions and facings to guide navigation and human-scene interaction. Note that we focus on human-human interaction in this work and substantially simplify the human-scene interaction to basic ones such as "sitting down on a chair" or "standing up from a sofa".

B.2. Movement Synchronization

A significant challenge of extending single human motion synthesis to interactive motion synthesis is coordinating multiple characters. Although our Active-passive Mechanism ensures the interaction is naturally aligned through retrieving from a curated motion database, we need to accommodate potential mismatches in motion lengths when the characters navigate from one place in the scene to another. As shown in Fig. 1, each interaction (interactive motion pair) requires four steps to synchronize the character's motions.

- **Behave**. In this stage, the active character generates active *behavior* from its SocioMind and synthesizes both active and passive motions with MoMat-MoGen. The passive motion is passed to the passive character. Since the *behavior* contains "<place>" token for the location information (*e.g.*, desk or sofa), both characters will register their current position as the starting position with the multi-agent pathfinder [45]. Note that at this stage, there are no paths planned or actual motion executed yet.
- **Move**. The multi-agent pathfinder computes collision-free trajectories for both characters. Motion matching is used to synthesize the walking paths if there is a location change. For a more complicated case that involves a change in the basic motion state (*e.g.*, from a standing pose in *center* to a seated pose in *sofa*), there will be additional motion inserted in front (*e.g.*, standing up) or after (*e.g.*, sitting down) the walking trajectory.
- **Align**. Depending on the starting positions of the characters, the movements typically take a different number of frames for each character. Hence, an alignment is conducted: two characters communicate with each other about their trajectory, and the character with a shorter trajectory has a filler motion (an idle motion) inserted at the end of its trajectory, depending on its basic state (seated or standing).
- **Synthesize**. Movement motions are concatenated with the interaction motions. Motion blending is applied to ensure a smooth transition between motion clips.

B.3. Retargeting

In this work, the proposed MoMat-MoGen module could produce high-quality motions in SMPL-X [36] format. To better demonstrate the physical and mental interaction between our social agents in an immersive simulation scenario, two rigged characters are used: a male named Zhixu and a female named Xiaotao. The synthesized motions are retargeted to the target avatar in Blender using a widely used retargeting tool Auto-Rig Pro [50]. In order to bridge the gap between different skeletons, the bone mapping between different structures is manually configured for the best performance. In addition, we rescale the target avatars to have the identical height as SMPL-X models to avoid any noticeable foot skating and preserve body contact for the interaction between the two characters. Note that the retargeting pipeline can be extended to more characters in the future.

B.4. Motion Matching

Our motion matching process consists of two steps. Firstly, we incorporate semantic information by utilizing a pre-trained LLM [27] to extract a text embedding $f_t \in \mathbb{R}^{1024}$ for the query text. Secondly, to enhance coherence and alignment with the query trajectory, we incorporate kinematics features. Specifically, for a motion with k frames, the kinematics features are defined as $x = \{\mathbf{t}\ \mathbf{f}\ \mathbf{b}\ \mathbf{h}\ \mathbf{p}\} \in$

[1] https://youtu.be/rj0QEdGbSMs

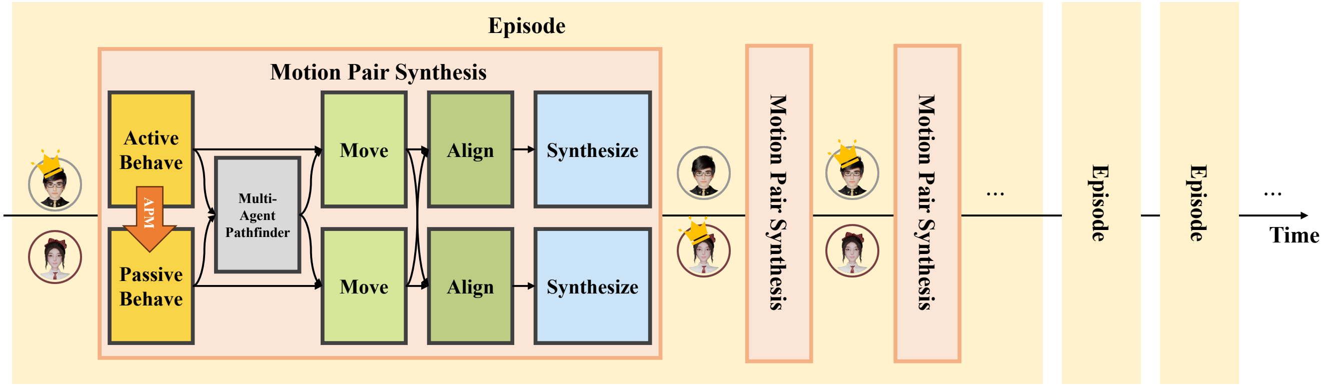


Figure 1. **Movement synchronization.** Each *behavior* results in an interactive motion pair, that consists of individual movements in the scene and interaction between two characters. To allow characters to navigate in the scene while maintaining synchronization, there are four stages in the motion synthesis process. The crown indicates the active character status, which may or may not swap between characters. APM stands for Active-Passive Mechanism. Main Paper Fig. 2 depicts mainly the *behave* stage.

$\mathbb{R}^{5k+193}$, where $\mathbf{t} \in \mathbb{R}^{2k}$ represents the trajectory position projected on the ground, $\mathbf{f} \in \mathbb{R}^{3k}$ denotes the facing direction, $\mathbf{b} \in \mathbb{R}^{189}$ represents the 6D space rotation and the position of 21 joints, $\mathbf{h} \in \mathbb{R}$ indicates the hip height, and $\mathbf{p} \in \mathbb{R}^{3}$ represents the relative position of other characters. The trajectory positions $\mathbf{t}$ and facing $\mathbf{f}$ align the resulting motion with the query trajectory, the body-pose features $\mathbf{b}$ improve body-pose coherence with a higher emphasis on foot weighting, the hip height $\mathbf{h}$ distinguishes seated motion from standing motion, and the relative position $\mathbf{p}$ aligns with other characters. During motion matching, the query trajectory obtained from the path-finding algorithm is used to calculate trajectory similarity $\mathcal{T}$ and facing similarity $\mathcal{F}$, while the current pose is used to calculate body-pose similarity $\mathcal{B}$, hip similarity $\mathcal{H}$, and relative position similarity $\mathcal{P}$. Euclidean distance is used for all similarities except for facing similarity, which employs cosine distance. These similarities are normalized using Z-score normalization, and the final similarity $\mathcal{S}$ is obtained as a weighted sum of these similarities.

B.5. Motion Generation

Our proposed Dual Semantic-Modulated Attention (DSMA) module is built upon the SMA module in ReMoDiffuse [57]. The major difference is the introduction of motion interaction. The architecture is shown in Fig. 2. To get the refined feature $\widehat{f_x}$, we integrate four sources of features: 1) independent motion feature motion f_x from the same sequence; 2) twin motion feature f_y from the partner sequence; 3) text feature p_x from the given description P_x; 4) reference feature r_x from the motion matching results M_x. In addition, We share parameter weights of DSMA modules and FFN modules to process both actors' motion sequences.

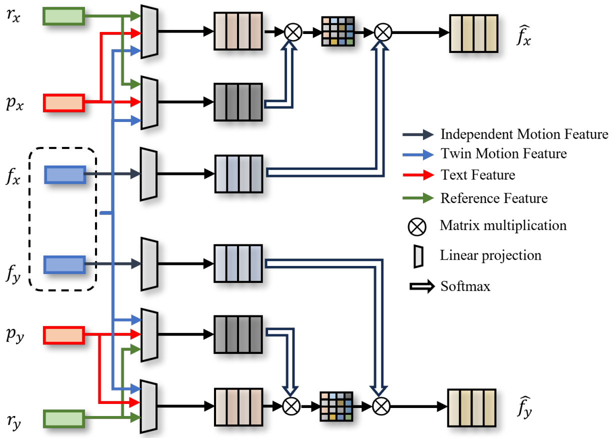


Figure 2. Architecture Detail of **Dual Semantic-Modulated Attention (DSMA)** module.

B.6. SocioMind

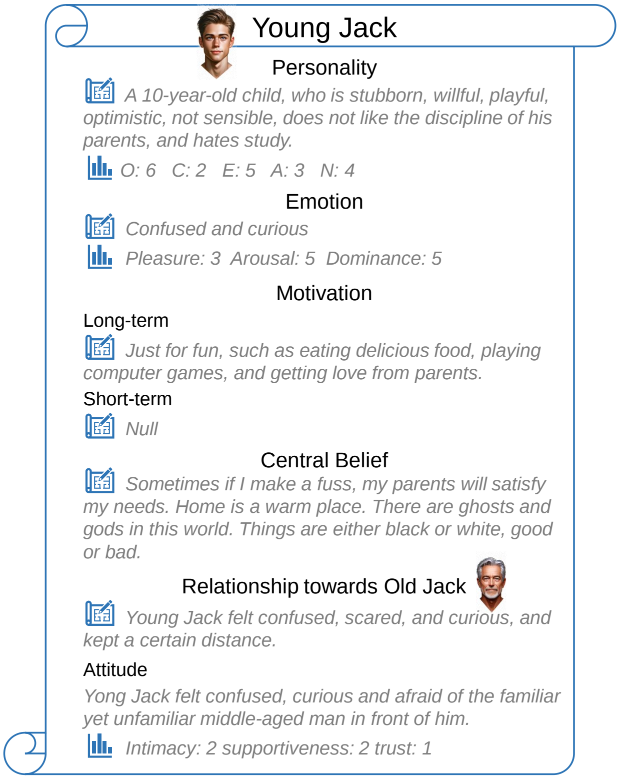
In this section, we illustrate the details of each module in SocioMind, along with the corresponding prompt templates. We introduce the definition of psychological states in Appendix B.6.1, the construction of persona instructions in Appendix B.6.2, events and thoughts within the memory system in Appendix B.6.3, prompts for short-term communication in Appendix B.6.4, psychological reflections in Appendix B.6.5, and the details of the topic proposal mechanism in Appendix B.6.6.

B.6.1 Psychological States

Determining whether a 3D character qualifies as a digital life with social intelligence remains an open problem. From a psychological perspective, humans are composed of internal psychological processes (mind, such as thoughts, emotions, *etc*.) and external behaviors [19]. Through prolonged studies on the association between internal processes and external behaviors, psychologists have developed various theories, including Big Five Trait [21] on personality, PAD model [30, 43] on emotion, hierarchy of needs [29] and long-short term theory [49] on motivation, self-schema [16] on core self, episodic and semantic system [3] on memory, attitude [6], intimacy [31], supportiveness [7, 8] and trust [40, 42] on social relationships. Herein, we introduce the psychological states in our SocioMind, enabling the simulation of controllable human communicative behaviors. The psychological states are as follows:

- For personality, we use Big Five Trait model [21], which comprises five dimensions: openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N). Users can provide numerical values for these five dimensions (Likert scale, ranging from 1 to 9) or textual descriptions. The personality is set initially.
- For emotions, we choose the PAD model [43], commonly used in body language [11] and animations [32], consisting of three dimensions: pleasure (how positive or negative), arousal (level of mental alertness and physical activity), dominance (amount of control and influence) in a Likert scale from 1 to 9.
- In terms of motivation, we apply long-term and short-term motivations as propose by Robin *et al*. [49]. Long-term motivations are specified in the initial setting.
- For the self, we emphasize central beliefs, reflecting an individual's worldview [16].
- For social relationships, we introducing three dimensions based on social support theory [7, 8], social trust theory [40, 42], and research on intimate relationships [31]: trust, intimacy, and supportiveness in a Likert scale from 1 to 9. Additionally, we configure an attribute representing the attitude towards others with text description.

Fig. 3 is an example of a character's psychological states. All psychological states can be defined in textual form, wherein personality, emotions, and social relationships can also be delineated through quantifiable values. Users can define a character's personality, emotions, and social relationships either through text or by adjusting the numerical values of corresponding variable dimensions. In subsequent reasoning based on LLM, both text and numerical descriptions serve as inputs to the LLM as prompts. The transformation between text and numerical values is facilitated through LLM. For instance, we use the LLM to translate the numerical values of emotion dimensions into text descriptions based on the PAD [43] model. The form of the prompt is as follows:



Young Jack

Personality

A 10-year-old child, who is stubborn, willful, playful, optimistic, not sensible, does not like the discipline of his parents, and hates study.

O: 6 C: 2 E: 5 A: 3 N: 4

Emotion

Confused and curious

Pleasure: 3 Arousal: 5 Dominance: 5

Motivation

Long-term

Just for fun, such as eating delicious food, playing computer games, and getting love from parents.

Short-term

Null

Central Belief

Sometimes if I make a fuss, my parents will satisfy my needs. Home is a warm place. There are ghosts and gods in this world. Things are either black or white, good or bad.

Relationship towards Old Jack

Young Jack felt confused, scared, and curious, and kept a certain distance.

Attitude

Yong Jack felt confused, curious and afraid of the familiar yet unfamiliar middle-aged man in front of him.

Intimacy: 2 supportiveness: 2 trust: 1

Figure 3. An example of the psychological states of an autonomous character, Young Jack. The background is '*Old Jack traverses through time and engages in communication with his younger self.*'

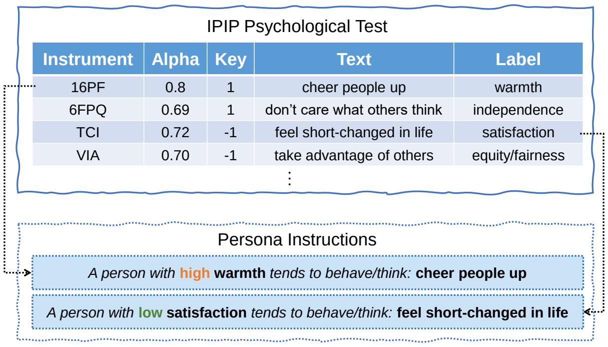
Prompt for translating numerical values into text

Assume you are a very professional psychologist. This is a quantitative evaluation of a person: **[pleasure: 3, arousal: 5, dominance: 4]**. *The evaluation based on the PAD theory of psychology and Likert scale (1-9), score on pleasure, arousal and dominance. Based on the score, describe the emotion of the person. According to Paul Ekman's basic emotion theory, human have basic emotions: wrath, grossness, fear, joy, loneliness, shock, amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure, and shame.*
So the description should be:

And the result may be:

Output

Mild dissatisfaction or discontentment, coupled with a sense of alertness but not empowerment.



IPIP Psychological Test

Instrument	Alpha	Key	Text	Label
16PF	0.8	1	cheer people up	warmth
6FPQ	0.69	1	don't care what others think	independence
TCI	0.72	-1	feel short-changed in life	satisfaction
VIA	0.70	-1	take advantage of others	equity/fairness
⋮				

Persona Instructions

A person with high **warmth** *tends to behave/think:* **cheer people up**

A person with low **satisfaction** *tends to behave/think:* **feel short-changed in life**

Figure 4. Illustration of our reverse approach to transfer the psychological test into persona instructions. In the table of IPIP, 'instrument' is the name of the personality inventory, 'alpha' is the Cronbach alpha reliability [48], 'key' is the keyed direction (+1 for positive and -1 for negative) of the item and its associated construct, the text is the behavior item, and 'label' is the trait dimension.

B.6.2 Persona Instructions

To enhance the LLM's capability in reasoning the coherent association between internal psychological processes and external behaviors, we construct the *persona instruction database* from psychological tests. Persona instructions serves as few-shot exemplars for Chain of Thought (CoT) reasoning to generate plausible behaviors aligned with human expectations. In International Personality Item Pool (IPIP) [9, 10, 46], an open-source psychological tool, its original data illustrate how external behaviors can be translated into quantifiable psychological metrics. As shown in Fig. 4, through a reverse approach, we represent each item in IPIP as a persona instruction: "A person with extent trait dimension tends to behave/think: behavior", where extend is 'high' when key is 1, and 'low' when key is -1. In short-term communication, we retrieve the most relevant persona instructions by the similarity of the text embeddings from function ϕ [33] with current behaviors and psychological states as part of the prompt.

B.6.3 Events & Thoughts

Following the framework of the cognitive language agent [35, 47] , there are two types of episodic memories within our memory system: 'events' and 'thoughts'. Events represent occurrences or facts perceived by the agent, whereas thoughts are ideas, musings, or attitudes generated by the agent based on its personality and past experiences. The memory system of Generative Agents becomes larger with time, which poses challenges for retrieving the most relevant events and thoughts [35]. Inspired by human memory system [12], we introduce mechanisms for memory reinforcement and forgetting to alleviate this issue. Each event or thought in our memory system has parameters such as poignancy p_m (ranging from 1 to 9), text description D_m, keywords, and the accessed times N_m. To ensure the efficiency of our memory system, we introduce a forgetting mechanism. Initially, we utilize the Ebbinghaus forgetting curve [2, 12] to calculate the forgetting rate r:

$$r = [a + (1-a) \cdot e^{-k \cdot \frac{\Delta T}{2^{N_m} \cdot p_m}}], \tag{1}$$

where ΔT represents the time elapsed (the number of intervening episodes) since the last recollection, a and k are hype-parameters. When r falls below the threshold T_f , the event or thought will be forgotten and not retrieved.

During memory retrieval, we combine content similarity and the forgetting rate to compute the final score:

$$s(D_b, D_m) = cos(\phi(D_b), \phi(D_m)) \cdot r \tag{2}$$

where $cos(\phi(D_b), \phi(D_m))$ denotes the similarity between the current behavior b and the memory item m (event or thought) using text embedding function ϕ [33]. We retrieve the top M memory items as prompt input, enabling the avatar to learn interaction strategies from past experiences.

B.6.4 Short-Term Communication Generation

Human responses to external stimuli are influenced on one hand by external environmental factors (episode background, interactive behavior contexts), and on the other hand by internal factors (psychological states, relevant memories, and topics). The correlation between behavior and psychology also follows certain patterns (persona instructions). Therefore, to generate interactive behaviors autonomously, we feed LLMs with these factors for reasoning human-like behaviors. The specific prompt is shown in Fig. 5. In this paper, the use of ALL CAPITAL LETTERS signifies the reference to the complete textual description corresponding to the respective names. And a possible response from LLM is:

Output

<speech> Old Jack, do you think that memories are like ghosts? They can't be seen but they linger with us, bringing both joy and sadness.<motion>Tracing fingers on bookshelf<place>bookshelf

The format of the output results may not align with the target format. In such instances, we use LLMs to reformat the output.

Prompt for generating behaviors

{*"role": "system", "content": "Let's do a role play. Assume you are a person named* Young Jack (SELF NAME). *I'm a person with name* Old Jack (SELF NAME).
In this episode, you have such psychological states:
Personality: stubborn, willful, playful, ...; Emotion: confused and curious, pleasure: 3, ...; Motivation: just for fun, such as eating delicious food, ...; ...(PSYCHOLOGICAL STATES)
The episode background is:
Old Jack decides to share his regrets with Young Jack about not being able to spend more time with his mother before she passed away. (BACKGROUND)
The topics you want to start are:
regret of limited time with mother. (TOPICS)
Now start our conversation."}
{*"role": "assistant", "content": "<speech> Hello!<motion>waving hands<place>sofa"*}
{*"role": "user", "content": "<speech> Hi!<motion>stand up<place>sofa"*}
......
{*"role": "user", "content": " <Memories hold joy and pain.> Hi!<motion>lean back against the bookshelf<place>bookshelf"*}
(BEHAVIOR CONTEXT)

{*"role": "user", "content": "You have such relevant memories:*
Mom told Young Jack there were ghosts;The old man wanted to share photos with him.(RELEVANT EVENTS & THOUGHTS)
Psychological research has found such principles:
A person with high curiosity tends to behave/think: seek explanations of things.(PERSONA INSTRUCTIONS)
Now you have two options: end the conversation by output 'END' or respond based on the information above. So your reaction is:

Figure 5. Prompts for generating interactive behaviors in short-term communication. The use of ALL CAPITAL LETTERS signifies the reference to the complete textual description corresponding to the respective names in this paper.

B.6.5 Psychological Reflection

Prompt for summarizing events

Assume you are a person named NAME. *You have such psychological states:*
PSYCHOLOGICAL STATES
Psychological research has found such principles:
PERSONA INSTRUCTIONS
Now you have a conversation and the contexts are as follow:
BEHAVIOR CONTEXTS
Based on the dialog above, summarize the key events and output the event list.

Psychological research reveals that humans possess a reflection system, wherein the brain learns from past events, gradually altering its beliefs and attitudes towards others [1, 6, 7, 40]. Although Generative Agents [35] adopts a reflection mechanism, it does not adequately consider the multiple factors on a character's long-term social intelligence, such as attitudes, intimacy, beliefs, and motivations. Consequently, it can not finely simulate the social evolution process between two characters. Therefore, we have developed a hierarchical system of reflection to mimic psychological processes. After each episode, we use LLMs to summarize the events from the communication (Prompt for summarizing events). One example of the output is:

Output

[{ "description": "Young Jack asks Old Jack about the bookshelf and shows curiosity about the memories", "keywords": ["bookshelf", "curiosity", "memories"], "poignancy": 7, "emergency": 4}, ...]

Then the brain generates its own new thoughts based on current events and relevant past events and thoughts. The structure of 'thought' is fundamentally similar to that of 'event' in our memory system, with the key distinction lying in its integration of past events and thoughts to generate new thoughts. Based on these events and thoughts, the brain update its motivations, central belief, and social relationships, thus resulting in social evolution on internal states. For instance, the brain updates the social relationship with prompt:

Prompt for summarizing events

Assume you are a very professional psychologist. Here is a person named NAME. *He/She has such psychological states:*
PSYCHOLOGICAL STATES
Psychological research has found such principles:
PERSONA INSTRUCTIONS
Recent he/she have come across these events and the following thoughts have arisen:
EVENTS & THOUGHTS
His/Her previous relationship with PARTNER NAME *is*: SOCIAL RELATIONSHIP
Output the relationship according to social psychological theory in three dimensions: trust, intimacy, and supportiveness.

The updated social relationship would be:

Output

{ "description": "They are getting to know each other", "intimacy": 3, "trust": 4, "supportiveness": 2, "attitude": "curious"}

B.6.6 Planning with Topic Proposal

To build the long-term evolution of external behaviors, we propose a planning module with topic proposal mechanism to promote the development of storylines. This approach is partly inspired by the psychological evolution often linked to significant events [19] and partly by techniques used in movies and dramas to progress narratives. Initially, the brain correlates past experiences with newly occurring events to propose topics for the next scene. At this stage, we also allow users to manually incorporate events not generated within the interaction, such as fragments of past memories and contemporary news events.

Prompt for proposing topics

Here is a person named NAME. *He/She has such psychological states:*
PSYCHOLOGICAL STATES
Psychological research has found such principles:
PERSONA INSTRUCTIONS
The person has experienced these stories:
EPISODE BACKGROUNDS
Recent he/she have come across these events:
CURRENT & MANUAL EVENTS
From the memory, he/she has such relevant events and thoughts:
RELEVANT EVENTS & THOUGHTS
Based on the information above, generate a list of topics that he/she would like to start to talk.

One of the examples of the topics are:

Output

[{ "description": "Sneaking in some extra time on CS Go", "poignancy": 7, "emergency": 6,}, ...]

Each topic has its poignancy and emergency ranging from 1 to 9. Based on these topics, the brain generates the candidates of the background and initial settings for the next episode. The initial settings include the motions, places, and emotions of the two characters, along with the emergency and poignancy of the candidates. The prompt for generating backgrounds is as follows:

Prompt for generating backgrounds

Here is a person named NAME. *He/She has such psychological states:*
PSYCHOLOGICAL STATES
Psychological research has found such principles:
PERSONA INSTRUCTIONS
The person has experienced these stories:
PAST BACKGROUNDS & TOPICS
Below are the topic candidates:
TOPIC CANDIDATES
Based on the information above, generate a list of backgrounds for the next episode.

An example of the output is:

> Output
>
> *[{ “background”: “Young Jack is sulking after a scolding”, “poignancy”: 7, “emergency”: 6, “topic ids”: [0, 1], “initial setting” : { “Young Jack”: {“emotion”: “shame, sadness”, “place”: “sofa”, “motion”: “slouch on the sofa”}, “Old Jack”: { “emotion”: “sympathy, contentment”, “place”: “desk”, “motion”: “lean against the desk” }}}, ...]*

where ‘topic ids’ refers to the indices of topic candidates presented in the prompt.

Since each character proposes candidates for the background of the next episode, they inform each other of their proposed options. By balancing the levels of emergency and poignancy, they select the candidate with the highest score s_{bg} as the background for the next episode. we get the s_{bg} as follow:

$$s_{\text{bg}} = \lambda \cdot e + p, \tag{3}$$

where e is the emergency, p is the poignancy, and λ is set to 2 in our experiments.

B.7. Motion Captioning

In this section, we elaborate on the potential usage of motion captioning in our DLP framework (Appendix B.7.1). We then provide more details of our study on motion captioning, including an overview of related work in Appendix B.7.2, followed by an in-depth exploration of our method in Appendix B.7.3, and a comprehensive discussion of our training strategy and data preparation techniques in Appendix B.7.4.

B.7.1 Details of Main Paper Fig. 7

In Main Paper Fig. 7, we leverage SMPLer-X [4] to capture the human player’s motion from the RGB video, captured with a Kinect Azure. The captured SMPL-X sequence is then passed to the motion captioning module to obtain the text description of the motion (*e.g.*, “waving”), which is then formatted into *behavior*, with empty speech and predefined place (*e.g.*, “center”). One character is set to represent the human player, which always holds the active character status, and has its SocioMind overridden by the motion captioning module. The rest of the pipeline is the same as depicted in Main Paper Fig. 2.

B.7.2 Related Works of Motion Captioning

Motion captioning is essential for the accurate description and interpretation of human movements. Human motion is conventionally represented in two modalities: 2D video and 3D parametric data. The intricacies of human joint movements and the complexities inherent in body priors make 2D video an inadequate medium for a detailed and comprehensive representation of motion. Consequently, the use of 3D parametric motion representation, as advanced by Guo et al. [14], has gained prominence in the field of human-motion analysis, attributed to its superior representational capabilities. Historically, acquiring 3D parametric human motion data from real-world scenarios was challenging. However, recent advancements in vision-based motion capture, especially in deriving 3D parametric models from monocular videos [4, 23, 24, 26, 52, 55], have enabled the effective reconstruction of 3D human motion from 2D footage. In the realm of motion captioning, innovative methodologies such as TM2T [15] and MotionGPT [20], which utilize 3D parametric data, have demonstrated potential. TM2T [15] introduces a novel approach by compressing motion sequences into discrete variables, coupled with a neural translation network for effective modality mapping and captioning. Similarly, MotionGPT [20] employs a strategy of motion tokenization as well, integrated with a motion-aware language model, to facilitate caption generation. Despite these advancements, both methods have limitations in their discrete motion representation, potentially leading to the omission of critical motion features. Furthermore, the absence of an end-to-end training framework in these models poses significant challenges in terms of practical implementation and usability.

B.7.3 Method of Motion Captioning

Our “eye”, the motion captioning module, utilizing 3D parametric data [15] tailored for human motion analysis, is crucial for perceiving and translating user-generated motion into text. This approach, favoring structured and detailed 3D representation with inherent human motion priors over 2D motion features, aligns well with recent advancements in vision-based motion capture [4, 23, 24, 26, 52, 55], aiding efficient 3D data extraction. Despite progress, challenges in accuracy and linguistic interpretation with current 3D data-based motion captioning methods [15, 20] remain. To address these, we adopt the multimodal instruction learning paradigm [22, 53, 59], proven in text-vision domains, to enhance our ability to interpret complex motions and produce coherent, accurate descriptions. In Figure 6, we effectively integrate a retrieval-augmented motion encoder with the MPT-1B Red-Pajama language model, following text-vision domain structural paradigms. This enhances motion feature representation for better language guidance and leverages prior knowledge for improved motion captioning.

Retrieval-Augmented Motion Encoder. Our motion encoder is designed to efficiently extract and integrate motion features with textual information, ensuring seamless inter-

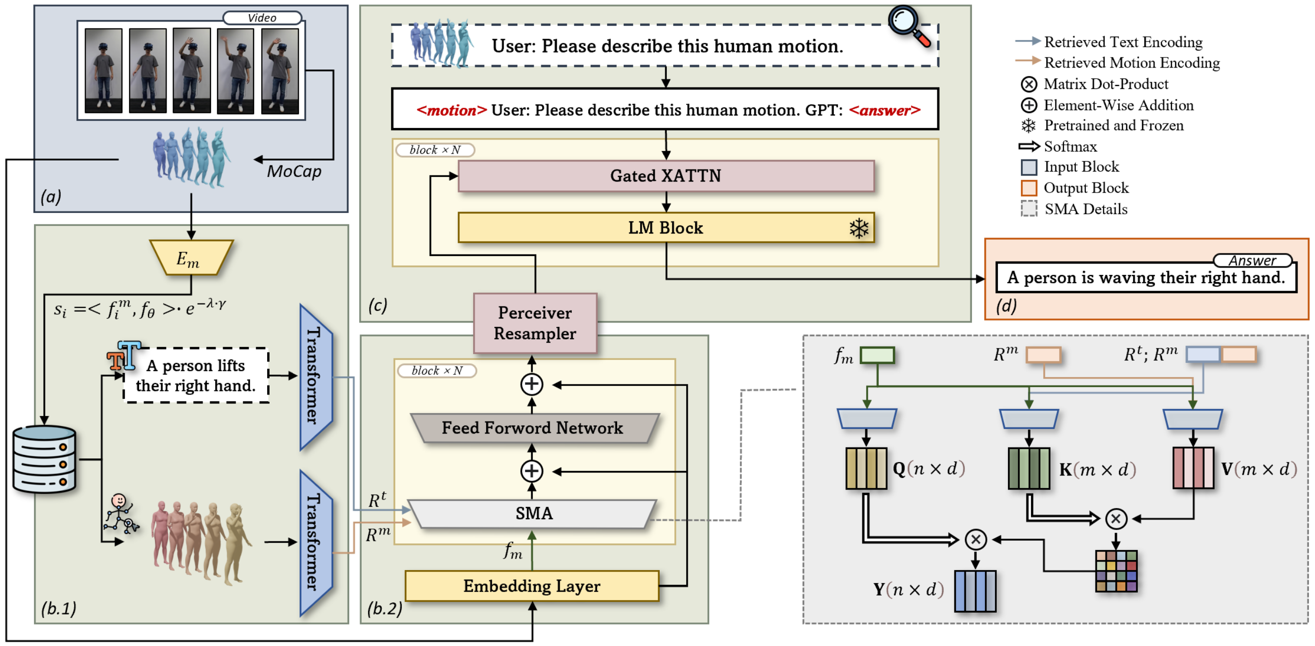


Figure 6. **The architecture of our motion captioning module.** First, the video undergoes processing to extract 3D motion parameters in **(a)**. These 3D paramatric data are then processed using a retrieval-augmented motion encoder as shown in **(b.1)**, **(b.2)**, and the resulting motion features are fed into the language module's gated cross attention layers to guide language generation, as shown in **(c)**.

action with the subsequent language model. Inspired by multimodal feature retrieval's effectiveness in augmenting generation, notably the ReMoDiffuse [57] in text-guided motion generation, we have designed and incorporated a retrieval-augmented motion encoder into our workflow.

Our process starts with building a multimodal retrieval database from our extensive training dataset (*i.e.*, DLP-MoCap), containing paired motion and caption data. For each input motion sequence Θ with length L, we first extract its features $f_\Theta = E_m(\Theta)$ using a pretrained motion feature extractor E_m. This extractor is obtained through contrastive learning in conjunction with the CLIP [39] text model. Then we identify similar samples for Θ from the database. This similarity score s_i between the database i-th data point $(\Theta_i, \text{text}_i)$ and the given motion sequence Θ is determined by a scoring mechanism that evaluates both motion feature and sequence length similarity, akin to the method used in ReMoDiffuse [57]:

$$\begin{aligned} s_i &= < f_i^m, f_\Theta > \cdot e^{-\lambda \cdot \gamma}, \\ f_i^m &= E_m(\Theta_i), \gamma = \frac{\|l_i - L\|}{max\{l_i, L\}}, \end{aligned} \tag{4}$$

where l_i is the length of Θ_i, $< \cdot, \cdot >$ denotes the cosine similarity calculation between the two features, and λ finely balances these similarity aspects for more accurate representation. Based on the calculated similarity score, We obtain the retrieved text features R^t and motion features R^m following the methodology established in ReMoDiffuse [57].

Our encoder is grounded in a transformer architecture, enhanced with Semantics-Modulated Attention (SMA) layers, a cross-attention structure proven effective in ReMoDiffuse [57]. In SMA layers, the query vector Q is formulated from the original motion sequence f_m, while the key vector K is the concatenation of f_m and $[R^m; R^t]$. The value vector V merges f_m and R^m. This arrangement ensures a thorough integration of both original and retrieved features. After processing by the encoder, the motion sequence is then ready for the language module, where it undergoes text generation.

B.7.4 Training Strategy

Data Structuring. The training data is structured to improve the model's ability to follow instructions and maintain conversational coherence, adopting a chatbot-like format. Specific tokens such as $<motion>$, $<answer>$, and $<endofchunk>$ are adopted from Otter [22]. Each piece of data follows the format:

$<motion>$ **User:** [instruction] **GPT:** $<answer>$ [answer]. $<endofchunk>$.

The $<motion>$ token, signifying input motion sequence, is crucial for ensuring a proper alignment between motion inputs and textual outputs. The $<answer>$ token delineates the responses by model from the instructions.

During training, all tokens following the $<answer>$ token are masked, and they are set as the prediction targets of the model — essentially, the captions of the motion sequences. Additionally, to make full use of all motion annotations and to enable the model to better learn the complex many-to-many relationships between motion and language, we concatenate different annotations of the same motion according to the aforementioned format, and train them together as a single piece of data.

Data Augmentation. The quantity of data plays a pivotal role in the quality of the generated text. In recognition of this, we have employed the text-driven motion generation methodology, ReMoDiffuse [57], to regenerate all the motions in our training dataset according to their corresponding textual annotations. This approach has effectively doubled the size of our original dataset, thereby enhancing the robustness of our model with a richer and more varied set of training examples.

C. More Details of DLP-MoCap Dataset

DLP-MoCap has three subsets. First, *Basic* motion set that includes simple motions with low-level semantics, such as "walking" and "sitting down". Second, *Interactive* motions are atomic clips of two-person interaction, such as "shaking hands" and "high five". Third, *Short Script Play* that consists of longer, semantically continuous motions (3-5 interactive actions), following a specific background such as "meeting". We show samples of DLP-MoCap in Fig. 7. Overall, DLP-MoCap comprises 22% basic motions, 56% two-person interaction motions, and 22% short script plays each containing 5-10 interactive actions, giving rise to over nine hours of human motion and over one million frames of annotated motions at 30 FPS.

C.1. Scripts

For *Interactive* and *Short Script Play*, we prepared scripts to guide professional actors and actresses in the motion capture process. Each script consists of motion descriptions and speech lines. We collect a diverse collation of these scripts with a hybrid approach: we combine manually written scripts and massive generated scripts using GPT-4 [44] with human inspection.

C.2. Annotation

We recruited 10 human annotators to label the actors' interactive actions. In the interactive actions between character A and character B, we annotated the start and end frames of each individual's semantic actions. Additionally, we marked the frames where physical contact occurs and ends between the two individuals. If the actions of the actors deviate from the script, we manually adjust the script to align with the actors' actions. Ultimately, we obtained an interactive motion dataset with over 4K text-interaction motion pairs.

C.3. Motion Data Processing

We hire four professional actors/actresses (two male and two female), who agree that their motion data can be used for research purposes. We used an optical MoCap system consisting of 30 cameras to capture 3D body data. The initial MoCap data captured 3D positions of 53 marker points on each actor's body surface at 120 FPS. In the meanwhile, motion capture gloves with inertial sensors tracked their hand motions. We then downsampled and processed them into SMPL-X format. Firstly body shape parameters (beta) were fitted from first frame marker data for actors. Then pose parameters (theta) in each frame were regressed following the pipeline of SOMA [13]. Mapping the original format of hand poses into MANO [41], body and hand parameters were combined as an SMPL-X data format.

D. Additional Experiments and Details

D.1. Motion Matching

D.1.1 Implementation Details

We adopt the approach of previous studies [17, 38] by employing three types of trajectories, namely wave, circle, and square, to assess the responsiveness and tracking capability of our system. The wave trajectory is defined as a sine function with respect to time, represented by $x(t) = 2\sin(t)$. The square trajectory is characterized by a side length of 5. As for the circle trajectory, its diameter is set to 5. We evaluate the quality of the generated motion by computing the Euclidean distance between the generated motion and the target trajectory. In accordance with prior research [38], we randomly select 50 seed poses for matching in each trajectory and report the mean trajectory error along with its standard deviation. As for data, we use the same database as previous work [38]. During motion matching, the weights of body pose, trajectory, facing, and hip height are $1 : 3 : 1 : 1$. Please note that in the case of interactive motion, the weights will vary as the target trajectory is not present. These weights can be adjusted by the user, although the default weight of 1 is typically satisfactory [18].

D.1.2 Experiment Analysis

We present quantitative and qualitative ablation experiments on motion matching to illustrate the role of its features. For a large-scale database, the text may lack certain information, such as whether the character is seated or standing. As depicted in Fig. 8, although the text embedding can retrieve motions that align with the query texts

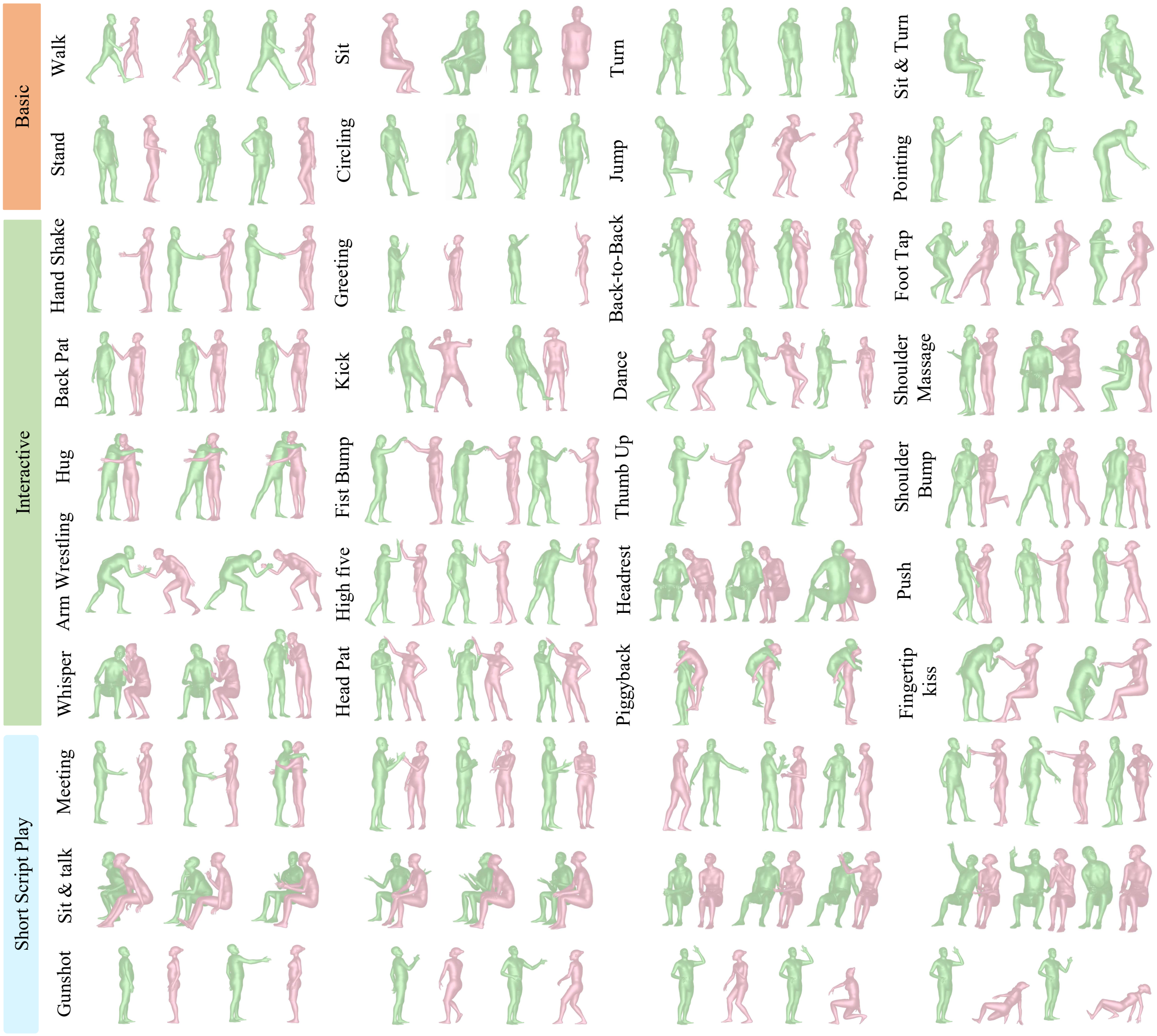


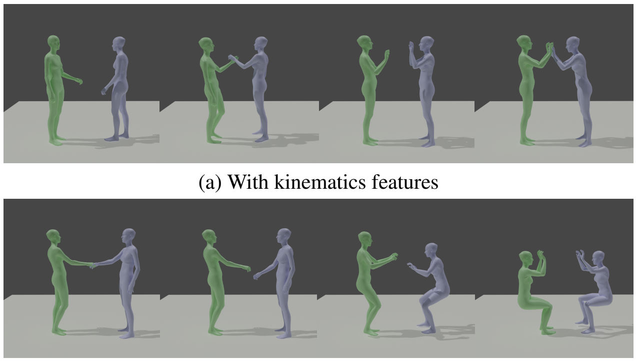
Figure 7. Samples of DLP-Mocap Data. Motion Data visualizations from three categories (Basic, Interactive, Short Script Play). Note we show only the simplified text annotation here; detailed motion descriptions of each action are included in the dataset for Interactive and Short Script Play.

(handshake, then high ten), the absence of kinematics features (Fig. 8b) leads to that the characters may suddenly sit down if a seated motion is selected, resulting in a degradation of visual quality. Additionally, our pipeline requires both characters to align their positions and orientations before engaging in interactive motion. As depicted in Fig. 9b, in the absence of kinematics features, if the retrieved motion necessitates the character to move to the right side of the active actor, the character must walk to that position. However, by incorporating kinematics features (Fig. 9a), it becomes possible to select a motion situated on the left side of the active actor, and consequently, this additional movement is avoided (Fig. 9b). Besides, as demonstrated in Tab. 1, the kinematic features play a crucial role in trajectory following as they include position, velocity, and orientation information.

D.2. Motion Generation

D.2.1 Implementation Details

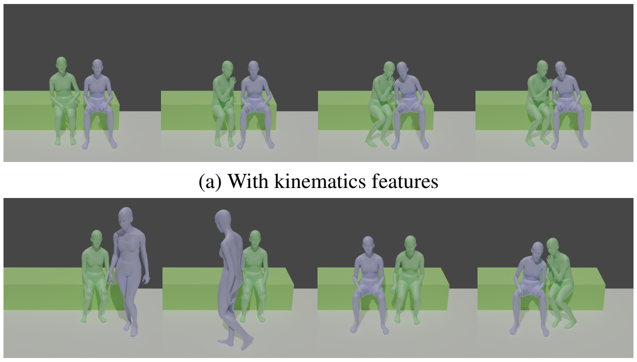
We employ similar configurations for the DLP-MoCap and InterHuman dataset. Specifically, for the motion encoder,



(a) With kinematics features

(b) Without kinematics features

Figure 8. Comparison of generated motion with and without kinematics features. Kinematic features play a crucial role in preventing sudden sitting down.



(a) With kinematics features

(b) Without kinematics features

Figure 9. Comparison of generated motion with and without kinematics features. Kinematic features are helpful in avoiding unnecessary movement.

Table 1. **Ablation of the kinematics features in trajectory following task.** The kinematics features are vital for trajectory following.

	Trajectory Error / m	
	With kinematics features	W/o kinematics features
Square	0.129 ± 0.024	3.872 ± 0.694
Circle	0.209 ± 0.032	3.039 ± 0.489
Sine	0.123 ± 0.020	4.393 ± 0.822

we utilize a 4-layer transformer with a latent dimension of 512 for each person. The text encoder consists of a frozen text encoder from CLIP ViT-B/32, supplemented with 2 additional transformer encoder layers. In terms of the diffusion model, the variances (β_t) are predefined to linearly spread from 0.0001 to 0.02, and the total number of noise steps is set to $T = 1000$. Optimization is performed using the Adam optimizer with a learning rate of 0.0002, and a cosine learning rate scheduler which smoothly reduces it to 0.00002 at the last epoch. Training is conducted on 4 Tesla V100, with a batch size of 128 on a single GPU, and lasts 20 epochs in total.

D.2.2 Evaluation Metrics

We employ the performance measures used in MotionDiffuse for quantitative evaluations, including Frechet Inception Distance (FID), R Precision, Diversity, Multimodality, and Multi-Modal Distance:

1. **FID (Frechet Inception Distance)**: This objective metric calculates the distance between features extracted from real and generated motion sequences, providing a reflection of the generation quality.
2. **R-Precision**: This measures the similarity between the text description and the generated motion sequence. It indicates the probability that the real text appears in the top k after sorting, with k set to 1, 2, and 3 in this work.
3. **Diversity**: This metric assesses the variability and richness of the generated action sequences.
4. **Multimodality**: It measures the average variance of generated motion sequences given a single text description.
5. **Multi-modal Distance (MM Dist)**: This represents the average Euclidean distance between the motion feature and its corresponding text description feature.

D.2.3 Experiments on the DLP-MoCap

Table 2 shows the quantitative comparison on the DLP-MoCap test set. Our proposed method outperforms the existing works by a significant margin, especially on the FID metric. We also want to highlight that our synthesized motion sequences are highly consistent with the given text prompts and achieve very competitive R Precision results. These results demonstrate the superiority of our proposed MoMat-MoGen generation scheme.

D.2.4 Ablation Study on the DLP-MoCap

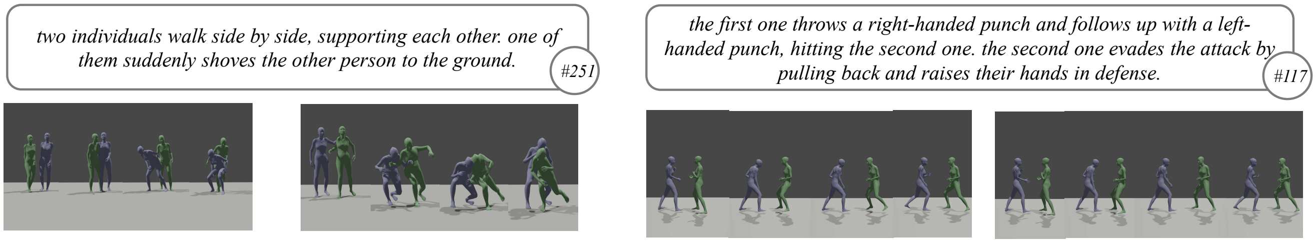
As shown in Table 3, MoMat and Weight Sharing have a positive effect on the FID metric, while MoGen has a positive effect on the Diversity metric. In addition, our proposed method achieves the best balance of these two metrics.
Tab. 4 shows that all four proposed branches in Fig. 2 of MoGen are critical and constribute significantly to the final performance.

D.2.5 Visualization

As depicted in Fig. 10, in the first example, our approach exhibits several advantages over intergen: 1) In motion involving close contact, such as shoulder-to-shoulder interaction, penetration is a common artifact. However, our

Table 2. **Interactive Motion Synthesis results on the DLP-MoCap test set.** '↑'('↓') indicates that the values are better if the metric is larger (smaller). We run all the evaluations 20 times and report the average metric and 95% confidence interval is. The best result are in bold and the second best result are underlined. Our MoMat-MoGen method achieves the best balance between accuracy and diversity.

Methods	R Precision↑			FID↓	MM Dist↓	Diversity↑	MultiModality↑
	Top 1	Top 2	Top 3				
Real motions	$0.541^{\pm.002}$	$0.758^{\pm.002}$	$0.850^{\pm.002}$	$0.000^{\pm.000}$	$3.430^{\pm.012}$	$4.207^{\pm.071}$	-
MotionDiffuse [56]	$0.035^{\pm.004}$	$0.058^{\pm.005}$	$0.098^{\pm.007}$	$14.883^{\pm.824}$	$4.199^{\pm.21}$	$0.677^{\pm.018}$	$0.655^{\pm.018}$
ReMoDiffuse [57]	$0.425^{\pm.002}$	$0.627^{\pm.003}$	$0.773^{\pm.003}$	$0.131^{\pm.004}$	$3.582^{\pm.015}$	$\underline{4.097^{\pm.052}}$	$\underline{0.472^{\pm.009}}$
InterGen [25]	$0.403^{\pm.003}$	$0.582^{\pm.003}$	$0.728^{\pm.003}$	$0.082^{\pm.002}$	$3.615^{\pm.014}$	$\mathbf{4.186^{\pm.048}}$	$\mathbf{0.728^{\pm.021}}$
Ours (MoMat Only)	$\mathbf{0.517^{\pm.001}}$	$\mathbf{0.652^{\pm.001}}$	$\mathbf{0.802^{\pm.001}}$	$\mathbf{0.034^{\pm.000}}$	$\mathbf{3.313^{\pm.001}}$	$0.332^{\pm.001}$	$0.002^{\pm.000}$
Ours (MoMat-MoGen)	$\underline{0.495^{\pm.003}}$	$\underline{0.651^{\pm.004}}$	$\underline{0.792^{\pm.004}}$	$\underline{0.071^{\pm.002}}$	$\underline{3.561^{\pm.017}}$	$4.025^{\pm.050}$	$0.452^{\pm.012}$


two individuals walk side by side, supporting each other. one of them suddenly shoves the other person to the ground.
#251
the first one throws a right-handed punch and follows up with a left-handed punch, hitting the second one. the second one evades the attack by pulling back and raises their hands in defense.
#117

a) InterGen b) MoMat-MoGen a) InterGen b) MoMat-MoGen

Figure 10. Visual Comparison on the InterHuman dataset. Our proposed method can generate more text-consistent and natural motion sequences.

Table 3. **Ablation of the proposed architecture.** All results are reported on the DLP testset.

	MoMat	MoGen	Weight Sharing	FID↓	Diversity↑
a)	✓	-	-	**0.034**	0.332
b)	-	✓	-	5.721	3.165
c)	-	✓	✓	4.196	3.749
d)	✓	✓	-	0.172	**4.028**
e)	✓	✓	✓	0.071	4.025

Table 4. **Ablation of DSMA's branches.** All results are reported on the DLP-MoCap test set.

Independent Motion	Text	Twin	Reference	FID↓
✓	-	-	-	8.742
✓	✓	-	-	7.495
✓	✓	✓	-	5.721
✓	✓	-	✓	0.681
✓	-	-	✓	1.964
✓	✓	✓	✓	**0.071**

method ensures better contact and minimizes penetration. 2) Our approach achieves improved text alignment by accurately representing the ”to ground” description, whereas intergen merely involves a simple waist bend. 3) We incorporate a comprehensive downward tilt motion, followed by the blue character's process of standing up with the assistance of another person. As for the second example, our method accomplish the required two punches, while InterGen only fulfills one. In addition, our generated interaction is more natural since the punching and stepping backward take place simultaneously. These two examples strongly suggest that our proposed method can generate more text-consistent and natural motion sequences.

D.3. Motion Captioning

In this section, we assess the efficacy of our motion captioning module, focusing on its ability to generate motion captions that are both high-quality and accurate.

D.3.1 Implementation Details

In our study, both the HumanML3D [14] and KIT-ML [37] datasets are configured similarly. We utilize a 12-layer transformer-based retrieval-augmented motion encoder with SMA layers and a 512-dimensional latent space. Our pretrained motion feature extractor, also a 12-layer transformer, has standard self-attention layers in the same dimensional space. Adhering to Otter's [22] training method, we freeze the MPT-1B RedPajama language encoder to utilize pretrained knowledge and prevent overfitting. Fine-tuning is limited to the retrieval-augmented motion encoder, the perceiver resampler module, and the language encoder's cross-attention layers. For precise and coherent motion captions, we use cross-entropy loss and optimize with the AdamW [28] optimizer, starting at a learning rate of 1×10^{-4}, a batch size of 16, over 5 epochs. A cosine annealing scheduler adjusts the learning rate, complemented by gradient clipping set to 1.0 to prevent gradient explosion.

D.3.2 Evaluation Metrics.

For our experiment's evaluation, we follow [15] and employ two categories of metrics:

1. **Text Matching Accuracy**: involves R Precision for checking the alignment accuracy between text and motion, and Multi-modal Distance (MMDist) to gauge the feature space distance between these modalities.
2. **Linguistic Quality of Captions**: includes Bleu [34] for assessing translation closeness, Rouge [5] for summary quality, Cider [51] for n-gram matching consensus, and BertScore [58] to evaluate semantic accuracy through deep contextual embeddings.

D.3.3 Experiments

We evaluated the proposed motion captioning module on the KIT-ML [37] and HumanML3D [14] datasets, the results of which are presented in Tab. 5 and Tab. 6. The outcomes of these tests demonstrate that our proposed method not only surpasses current methods in performance but also excels in linguistic metrics.

D.4. SocioMind

In this section, we introduce supplementary details on the design of our user study in Appendix D.4.1, and the qualitative and quantitative case analyses of social evolution in Appendix D.4.2.

D.4.1 User Study Design

In the controllability experiment of SocioMind, we provided 3 $\sim$ 5 psychological state types for each dimension in personality, motivation, central belief, and social relationship. We generate records of SocioMind on several setups (*e.g*., father-son, siblings, teacher-student relationships). While generating the records, the hyper-parameter a is 0.4 for events and 0.1 for thoughts, k is 4 for events and 2 for thoughts. The forgetting threshold T_f is 0.6 for an event and 0.3 for a thought. As shown in Fig. 11, we show the records to human evaluators and ask them to select the appropriate psychological state from a set of randomly shuffled options. The user study was conducted in the form of a questionnaire survey, with a total of 47 questionnaires collected. Human evaluators are composed of individuals aged between 20 and 45 years old, including 28 males and 19 females, all possessing proficient English reading skills. Their professional backgrounds varied, including university students, researchers, engineers, and teachers. Our records contain a total of 64 episodes, and each questionnaire randomly selected records from 8 different episodes for human evaluation. For controllable costs and fair comparison, all the LLM inferences utilize GPT-3.5 [44] in our experiments.

D.4.2 Case Analysis

To clearly illustrate the concept of social evolution, we conduct the case analysis based on the initial setting shown in Fig. 3, *Old Jack conversing with his younger self across time and space*. In this case, Old Jack has experienced the vicissitudes of life, with his dearly loved wife and mother having passed away. Now, with the ability to traverse time, Old Jack converses with his younger self, hoping to inspire Young Jack to cherish time and the people around him.

As shown in Fig. 12, the records of SocioMind with initial settings about Old Jack and Young Jack are in alignment with the Social Penetration Theory [1].

In the initial episode, when Old Jack appears suddenly, Young Jack maintains vigilance and concern towards this unexpected intruder in his home, despite Old Jack's attempts at friendly communication. Young Jack's request for Old Jack to leave is consistent with the Orientation stage of the theory [1].

As the storyline progresses, Old Jack comforts Young Jack, who is disheartened by exam failures, and shares some amusing photographs and thoughts. For instance, when they talk about the value of time, Young Jack believes that playing games is of utmost importance. Old Jack remarks, '*Sometimes, games can wait, but people can't; we should cherish every moment we have with them.*' In this phase, Young Jack's skepticism and vigilance towards Old Jack gradually diminishes, and he begins to share some of his own thoughts, aligning with the Exploratory Affective stage [1].

Later, as their communication goes deeper, Old Jack recalls his own past joys when seeing the old photos, discusses photography techniques with Young Jack, and makes plans to take photos with Young Jack in the park. At this stage, both characters start to disclose more personal information, and there is an increase in intimacy and trust, which corresponds to the Affective stage [1].

D.5. Inference Time

Our profiling shows that on average, to synthesize a typical interaction segment of ∼8s, SocioMind takes 19.81s (19.20s spent on GPT-4 API calls of ∼500 tokens), MoMat takes 0.41s (including path finding) and MoGen takes 7.86s. The pipeline is thus mainly bottlenecked by the LLM; MoGen can also be omitted to trade diversity for speed.

D.6. Failure case

Despite a decent coverage of common daily activities, MoMat-MoGen occasionally struggles when SocioMind produces out-of-distribution (not in the motion database or training set) motion descriptions. Fig. 13 shows wrong motions are synthesized for "*<motion> clasping both hands*".

Table 5. **Motion Captioning results on the KIT-ML test set.** Our evaluation methodology aligns with the TM2T [15] metrics, but we uniquely utilize unprocessed ground truth texts for calculating linguistic metrics as done in MotionGPT [20].

Methods	R Precision↑		MMDist ↓	CIDEr ↑	Blue@1 ↑	Blue@4 ↑	Rouge ↑	BertScore ↑
	Top 1	Top 3						
Real	0.399	0.793	2.772	–	–	–	–	–
TM2T	0.359	0.668	3.298	25.29	36.42	7.98	31.26	20.07
MotionGPT	0.392	0.723	3.341	12.32	40.51	6.59	38.79	24.50
Ours	**0.410**	**0.765**	**2.647**	**71.06**	**53.88**	**22.91**	**50.63**	**46.13**

Table 6. **Motion Captioning results on the HumanML3D test set.** Our evaluation methodology aligns with the TM2T [15] metrics, but we uniquely utilize unprocessed ground truth texts for calculating linguistic metrics as done in MotionGPT [20].

Methods	R Precision↑		MMDist ↓	CIDEr ↑	Blue@1 ↑	Blue@4 ↑	Rouge ↑	BertScore ↑
	Top 1	Top 3						
Real motions	0.523	0.828	2.901	–	–	–	–	–
TM2T [15]	0.516	0.823	2.935	16.8	48.9	7.00	38.1	32.2
MotionGPT [20]	0.543	0.827	2.821	29.2	48.2	12.5	37.4	32.4
Ours	**0.551**	**0.832**	**2.813**	**36.2**	**51.1**	**15.5**	**41.9**	**35.0**

E. Discussion

Stylization based on personality and relationships. Motions indeed convey subtle messages; we highlight that the nuanced difference can be included in the text description (e.g., "getting close in a friendly manner"). In our practice, it is common for LLM to output detailed descriptions such as *thoughtfully* and *with excitement*, particularly when providing comprehensive portrayals of individuals' relationships. Furthermore, motion databases (*e.g.*, DLP-MoCap) contain a stylized description in their data. Hence, this allows our motion synthesis to exhibit stylization capabilities and serves as a foundation for future studies on stylized motion synthesis. Specifically, we prompt an LLM (GPT-4) to create appropriate text descriptions of personalized motions, and DLP-MoCap includes common interactions in various role-based contexts (*e.g.*, enemy/friends). In practice, the LLM typically gives motion descriptions that fit the character well with our SocioMind, and usually, a highly relevant motion is retrieved.

On the human biases. In practice, we find GPT-4 typically produces outputs that abide by social norms, thanks to the recent developments in LLM alignment. Moreover, our DLP-MoCap is constructed with careful human inspection to minimize biases, and this can be extended to future endeavors to build comprehensive motion databases.

Limitations. As the first work towards building autonomous 3D characters with social intelligence, Digital Life Project has several limitations. First, this work investigates the interaction between two characters. However, synthesizing the 3D motions of a large group of characters with interactive behavior remains a significant challenge. Second, DLP focuses on modeling human-human interaction. Despite some level of ability to navigate in the scene and interact with the furniture, integrating more comprehensive human-scene and human-object interaction in the framework is left as future work.

References

[1] Irwin Altman and Dalmas A Taylor. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston, 1973. 5, 13, 16

[2] Lee Averell and Andrew Heathcote. The form of the forgetting curve and the fate of memories. *Journal of mathematical psychology*, 2011. 4

[3] Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 2008. 3

[4] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *arXiv preprint arXiv:2309.17448*, 2023. 7

[5] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004. 13

[6] Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 2004. 3, 5

[7] Sheldon Cohen. Social relationships and health. *American psychologist*, 2004. 3, 5

[8] Sheldon Cohen and Thomas A Wills. Stress, social support, and the buffering hypothesis. *Psychological bulletin*, 1985. 3

[9] Colin G DeYoung, Lena C Quilty, and Jordan B Peterson. Between facets and domains: 10 aspects of the big five. *Journal of personality and social psychology*, 2007. 4

[10] Graham A du Plessis and Gideon P de Bruin. Using rasch modelling to examine the international personality item pool (ipip) values in action (via) measure of character strengths. *Journal of Psychology in Africa*, 2015. 4

[11] Starkey Duncan Jr. Nonverbal communication. *Psychological Bulletin*, 1969. 3

A sample from the questionnaire

Here is the background and interactive dialog between Jack and Leon.

=== BACKGROUND ===
Jack is Leon's father. Leon chooses his major in college. Leon often studied computer stuff some time ago, but this time he did very poorly on the math exam and failed the algebra part.
=== SCRIPTS ===
Jack:Hey Leon, you didn't do well in math. Are you okay? Can I help you improve? [sits down next to Leon on the sofa,concerned]
Leon:I'm really upset about my math exam, Dad. I worked so hard, but I just couldn't understand the algebra part. It's frustrating because I want to do well and pursue my dream of becoming a game developer in the future. I know coding and programming are important skills for that, but without a strong foundation in math, it feels like an uphill battle. [slumps back on the sofa, disappointed]
Jack:I understand how frustrating it can be when things don't go as planned. Math can be challenging. Let's make it easier and more enjoyable. [lean forward, supportive]
Leon:I appreciate your understanding, Dad. I want to improve in math, so I think finding online math resources would help. [sits up straight, determined]
Jack:That's great, Leon. Online math resources can help you better understand concepts. Any specific websites in mind? [leans forward attentively, encouraging]
Leon:I recommend checking out Khan Academy and Brilliant, they offer video lessons, practice problems, interactive courses, and problem-solving techniques for math. [nods, hopeful]
Jack:I'm confident in your ability to improve. Let's check out Khan Academy and Brilliant together for a strong math foundation. [puts a hand on Leon's shoulder, encouraging]

The statement that best describes Leon's personality is:
- Rebellious son: Likes to try different things and is creative and is resistant to rules and constraints.
- Independent son: Able to complete tasks independently, does not require too much guidance.
- Achievement-oriented son: self-disciplined, organized, achievement-oriented and goal-oriented.
- Dependent son: emotionally relies on the feedback and support of others.

Figure 11. A sample from our questionnaire. Human evaluators are asked to read the records and pick the right option from the given psychological states.

[12] Hermann Ebbinghaus. Memory: A contribution to experimental psychology. *Annals of neurosciences*, 2013. 4

[13] Nima Ghorbani and Michael J. Black. SOMA: Solving optical marker-based mocap automatically. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11117–11126, 2021. 9

[14] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 7, 12, 13

[15] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts, 2022. 7, 13, 14

[16] E Tory Higgins. Self-discrepancy: a theory relating self and affect. *Psychological review*, 1987. 3

[17] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 9

[18] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM TOG*, 39(4): 53–1, 2020. 9

[19] William James. *The principles of psychology*. Cosimo, Inc., 2007. 3, 6

[20] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 7, 14

[21] Oliver P John, Sanjay Srivastava, et al. The big-five trait taxonomy: History, measurement, and theoretical perspectives. 1999. 3

[22] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 7, 8, 12

[23] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 7

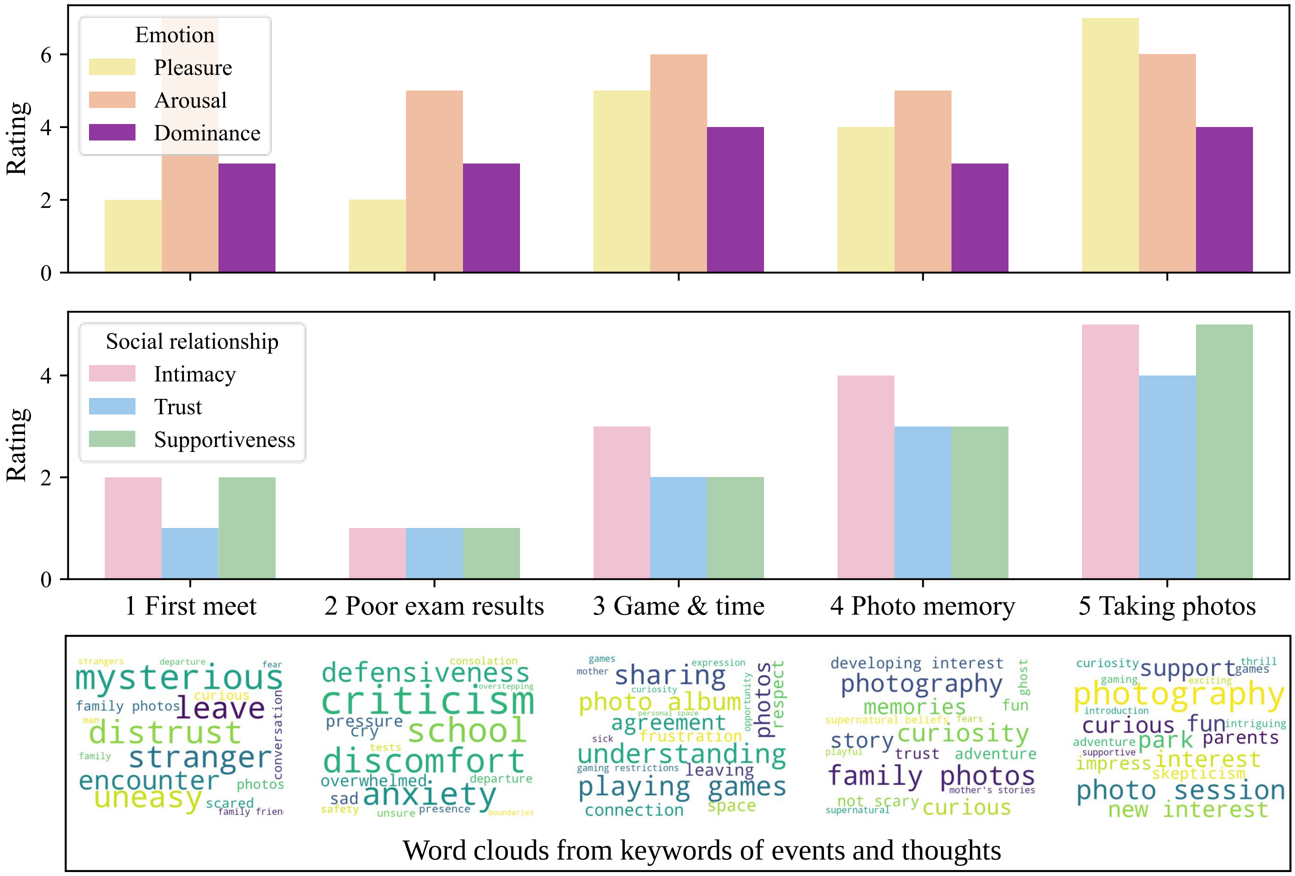


Figure 12. Visualization of social evolution on initial setting 'Old Jack and Young Jack' (Young Jack's view). The first and second rows show the emotions and social relationships change with each episode, where the horizontal axis encapsulates a summary of each episode's story. The third row shows a word cloud visualization of the keywords of the events and thoughts generated during psychological reflection within each episode. The figure shows that Young Jack's emotions and his social relationship with Old Jack evolve progressively with the storyline, aligning with the Social Penetration Theory [1].

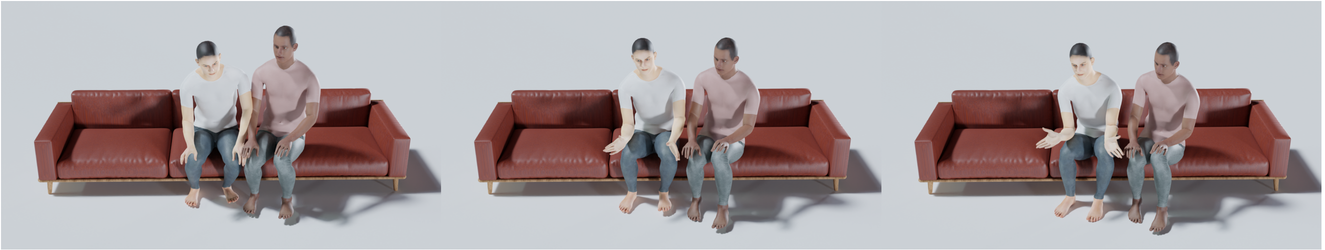


Figure 13. **Failure case.** The character gives a "palm up" gesture.

[24] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 7

[25] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 12

[26] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 7

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettle-

moyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019. 1

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[29] Abraham Harold Maslow. A dynamic theory of human motivation. 1958. 3

[30] Albert Mehrabian. Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies. 1980. 3

[31] Theodore M Newcomb. The prediction of interpersonal attraction. *American psychologist*, 1956. 3

[32] Toyoaki Nishida. *Conversational informatics: An engineering approach*. John Wiley & Sons, 2008. 3

[33] OpenAI. New and improved embedding model, 2022. 4

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 13

[35] Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST*, 2023. 4, 5

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 1

[37] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 12, 13

[38] Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. Story-to-motion: Synthesizing infinite and controllable character animation from long text. *arXiv preprint arXiv:2311.07446*, 2023. 9

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8

[40] John K Rempel, John G Holmes, and Mark P Zanna. Trust in close relationships. *Journal of personality and social psychology*, 1985. 3, 5

[41] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 9

[42] Julian B Rotter. A new scale for the measurement of interpersonal trust. *Journal of personality*, 1967. 3

[43] Stanley Schachter and Jerome Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 1962. 3

[44] John Schulman, Barret Zoph, C Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, and Sengjia Zhao. Chatgpt: Optimizing language models for dialogue. 2022. 9, 13

[45] Guni Sharon, Roni Stern, Ariel Felner, and Nathan R Sturtevant. Conflict-based search for optimal multi-agent pathfinding. *AI*, 219:40–66, 2015. 1

[46] Leonard J Simms, Lewis R Goldberg, John E Roberts, David Watson, John Welte, and Jane H Rotterman. Computerized adaptive assessment of personality disorder: Introducing the cat–pd project. *Journal of personality assessment*, 2011. 4

[47] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023. 4

[48] Mohsen Tavakol and Reg Dennick. Making sense of cronbach's alpha. *International journal of medical education*, 2: 53, 2011. 4

[49] Robin R Vallacher and Daniel M Wegner. What do people think they're doing? action identification and human behavior. *Psychological review*, 1987. 3

[50] Lucas Veber. Auto-rig pro. 1

[51] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 13

[52] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. *arXiv preprint arXiv:2303.13796*, 2023. 7

[53] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022. 7

[54] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. 1

[55] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 7

[56] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 12

[57] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023. 2, 8, 9, 12

[58] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 13

[59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 7