Digital Life Project: Autonomous 3D Characters with Social Intelligence

Zhongang Cai[*,1,2,3], Jianping Jiang[*,2], Zhongfei Qing[*,2], Xinying Guo[*,1], Mingyuan Zhang[*,1], Zhengyu Lin[2], Haiyi Mei[2], Chen Wei[2], Ruisi Wang[1,2], Wanqi Yin[2], Liang Pan[1,3], Xiangyu Fan[2], Han Du[2], Peng Gao[2], Zhitao Yang[2], Yang Gao[2], Jiaqi Li[2], Tianxiang Ren[2], Yukun Wei[2], Xiaogang Wang[2], Chen Change Loy[1], Lei Yang[†,2,3], Ziwei Liu[†,1]
[1]S-Lab, Nanyang Technological University, [2]SenseTime Research, [3]Shanghai AI Laboratory
[*] Equal Contributions, [†] Corresponding Author

https://digital-life-project.com

Figure 1. **Digital Life Project** empowers virtual characters to interact with each other using articulated body motions. We demonstrate the interaction of two characters across four occasions (*episodes*) that leads to evolving relationship.

Abstract

*In this work, we present **Digital Life Project**, a framework utilizing language as the universal medium to build autonomous 3D characters, who are capable of engaging in social interactions and expressing with articulated body motions, thereby simulating life in a digital environment. Our framework comprises two primary components: **1) SocioMind:** a meticulously crafted digital brain that models personalities with systematic few-shot exemplars, incorporates a reflection process based on psychology principles, and emulates autonomy by initiating dialogue topics; **2) MoMat-MoGen:** a text-driven motion synthesis paradigm for controlling the character's digital body. It integrates motion matching, a proven industry technique to ensure motion quality, with cutting-edge advancements in motion generation for diversity. Extensive experiments demonstrate that each module achieves state-of-the-art performance in its respective domain. Collectively, they enable virtual characters to initiate and sustain dialogues autonomously, while evolving their socio-psychological states. Concurrently, these characters can perform contextually relevant*

bodily movements. Additionally, an extension of DLP enables a virtual character to recognize and appropriately respond to human players' actions.

1. Introduction

Recent advancements in Large Language Models (LLMs) [53, 62] have transformed the landscape of human-computer interaction, catalyzing the emergence of innovative applications across various domains. Remarkably, many once far-fetched fantasies have gradually become tangible realities. In this work, the term *Digital Life Project* (DLP), as envisioned in the recent science fiction blockbuster *The Wandering Earth II*, is adopted to frame our endeavor. What qualifies as a digital life? From the psychological perspective, humans are composed of internal psychological processes (mind, such as thoughts) and external behaviors [32]. In this light, our objective is to harness the sophisticated capabilities of LLM to craft virtual 3D characters, that emulate the full spectrum of human psychological processes, and engage in diverse interactions with synthesized 3D body motions.

Recently, Park *et al*. introduced Generative Agents [42] to advance AI agents capable of simulating human-like behavior. Despite the encouraging progress, this pioneering work is built upon many simplifications of interaction: the agents are represented by pixelated 2D figures. Co-LLM-Agents [73] aims to build collaborative embodied AI and includes 3D agents. However, the 3D agents are still constrained by a small set of actions and do not exhibit the capability to socialize. Existing works thus overlook the importance of sophisticated human body language, through which a crucial amount of information is conveyed [7, 25, 26]. Moreover, there is a notable deficiency in the current modeling of social intelligence. This aspect is critical for the creation of characters that not only mimic human actions but also possess human-like thinking and emotional responses, even the ability to foster long-term relationships.

To achieve the aspirations of DLP, we introduce a framework consisting of two essential components. **First**, the SocioMind which is a carefully designed "digital brain", anchoring its design in rigorously applied psychological principles. Utilizing emergent abilities of LLMs [40, 53, 66], the brain generates high-level instructions and plans the character's behaviors. Notably, SocioMind introduces few-shot exemplars from psychological tests to form guiding instructions for personality modeling, utilizes social cognitive psychology theories in the memory reflection process, and designs a negotiation mechanism between characters for story progression. **Second**, the "digital body" that introduces the MoMat-MoGen paradigm to address interactive motion synthesis, which exploits the complementary nature of motion matching [12] and motion generation [76]. Here, motion matching is a foundational technique in modern-day industry-level character animation that retrieves high-quality motion clips from a database to ensure motion quality, whereas motion generation is a line of works that rapidly gained popularity recently for their excellent ability to produce diverse human motions.

Experiment results demonstrate that SocioMind and MoMat-MoGen outperform existing arts in their respective domains. Specifically, SocioMind demonstrates outstanding alignment between character behavior and psychological states (*e.g*., personality and relationship); MoMat-MoGen is able to achieve a balance between motion quality and diversity. Equipped with both modules, we further show DLP's controllability as manual editing of character attributes can result in semantically accurate and aesthetically realistic interactive motions. Moreover, we explore human-character interaction by developing a motion captioning module as an extension of DLP, that translates monocular human video to motion description, thus enabling virtual characters to understand and appropriately respond to human players.

In summary, we contribute DLP, a framework to build autonomous 3D characters with social traits. It features SocioMind: a controllable psychology-based "brain" to enable short-term interactive communication and long-term social evolution, and MoMat-MoGen: a "body" that synthesizes high-quality and diverse interactive motions through synergizing motion matching and motion generation.

2. Related Works

2.1. Motion Synthesis

Motion matching is widely employed in the industry to generate long-lasting, high-quality motion. The classic motion matching [12] retrieves the segment that best matches the current pose and target trajectory. Learned motion matching [30] employs an auto-regressive neural network to predict the next motion state based on a given control signal. The Story-to-motion [47] further incorporates semantic control through LLM and enhances transition using transformer models. Recently, significant strides have been made in motion generative models for text-driven motion generation. Early works aimed to establish a unified latent space for natural language and motion sequences [3, 21, 45, 60]. Guo et al. [23], TM2T [24], and T2M-GPT [74] employ an auto-regressive scheme to generate lengthy motion sequences. Diffusion-based generative models have demonstrated remarkable performance in leading benchmarks for the text-to-motion task. MotionDiffuse [75], MDM [61], and FLAME [34] represent early attempts to apply the diffusion model to the text-driven motion generation field. Subsequent models such as MLD [9], ReMoDiffuse [76], Fg-T2M [65], FineMoGen [77], InsActor [49], and PhysDiff [72] have further advanced this idea, achieving im-

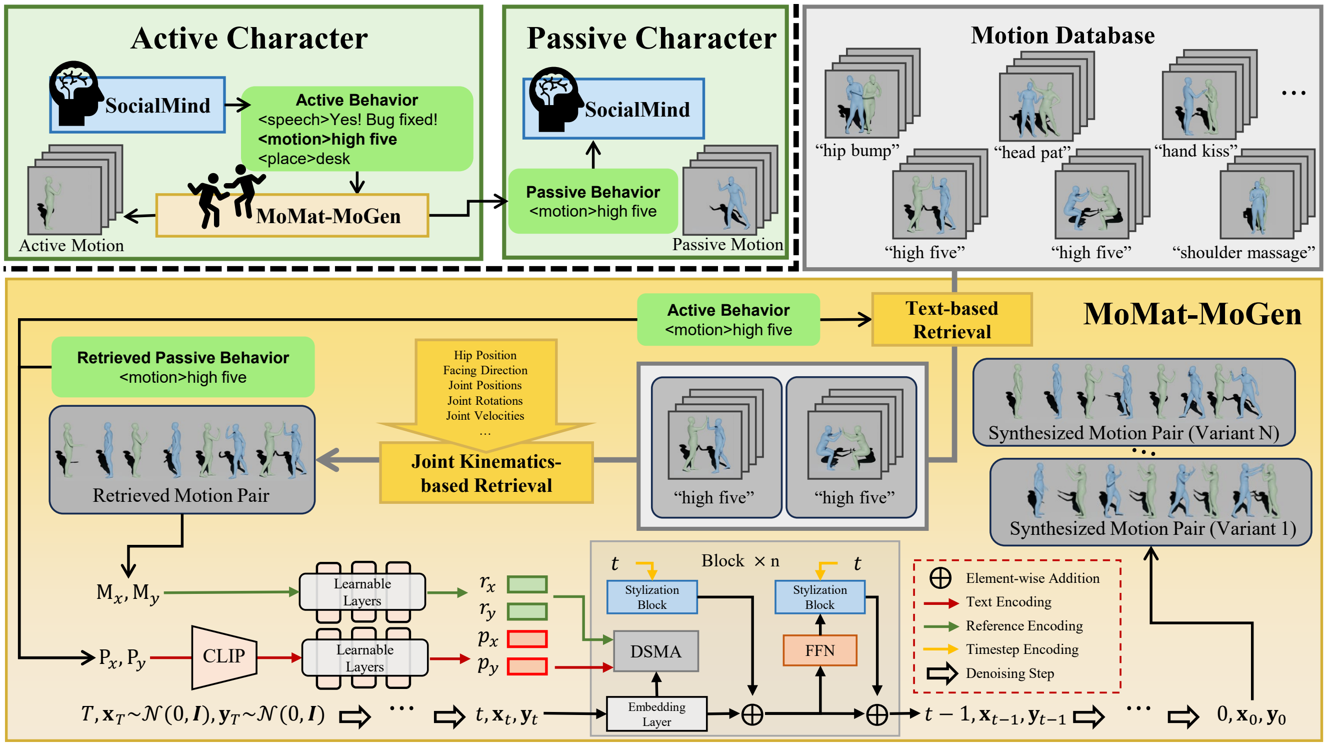

Figure 2. **Digital Life Project** framework for interactive autonomous characters. The top left part depicts the Active-Passive Mechanism, and the rest of the figure illustrates MoMat-MoGen. SocioMind is shown in details in Fig. 3.

proved text-motion consistency, motion quality, and physical plausibility. Recently, PriorMDM [54] propose a fine-tuning strategy to extend the MDM to human interaction generation. Inter-X [68] and InterGen [35] propose two large-scale datasets for human interaction generation with textual description. In addition, InterGen also proposes a two-stream diffusion architecture, serving as a significant baseline in this field. ReMoS [22] focuses on plausible hand interaction and decomposes the whole generation process into full-body and hand motion generation.

2.2. LLM Agents

With the emergent abilities of large language models (LLM) in reasoning, planning, and learning [17, 40, 62, 66], LLMs swiftly evolve through three phases: the standalone primitive LLM, language agents [2, 31] that directly interact with the environment via text, and cognitive language agents [43, 52, 64, 70, 71] with internal cognitive structures [59]. Under the prime framework of cognitive language agents, the system design hinges on the intended application and objectives: reward systems for game agents [64, 69, 79], chains of API calls for tool agents [44, 52, 56], and so forth. Moreover, the emergence of human-like behaviors in LLMs has prompted researchers to investigate controllable mental behaviors in LLMs, such as a stable personality [51] and human simulation in political science [5] and social psychology [1]. Recently, Social Simulacra [41] and S^3 [20] build agent systems with autonomous posting and reposting skills in internet community space. Generative Agents [43] facilitates the formation of social relationships and information diffusion by daily schedules and brief communications within a 2D sandbox gaming space.

3. Methodology

3.1. Text as the Universal Medium

We define *behavior*, a dictionary-like structured text message to bridge the "brain" (Sec. 3.4) and the "body" (Sec. 3.3). For example, <***speech***>*Hello!* <***motion***>*waves right hand* <***place***>*table* contains pre-set *keys* encapsulated by pointy brackets, followed by the respective *values*, also in natural language. Behaviors are thus interpretable by the LLM and the regular-expression parser. In this work, we focus on <*motion*>, but we discuss the use of other tokens in the Supplementary Material: <*place*> triggers navigation and basic state transfer (*e.g.*, sit down), <*speech*> may be used for face control.

3.2. Active-Passive Mechanism

There exists an intrinsic order in human interaction. For example, "shaking hands" may appear to be a simultaneous action by two subjects, it typically initiates with one person extending a hand first. Moreover, the other person's action is largely predictable: it is socially appropriate for that person to reciprocate the handshake as a basic courtesy. Another example of real-life collaborative activ-

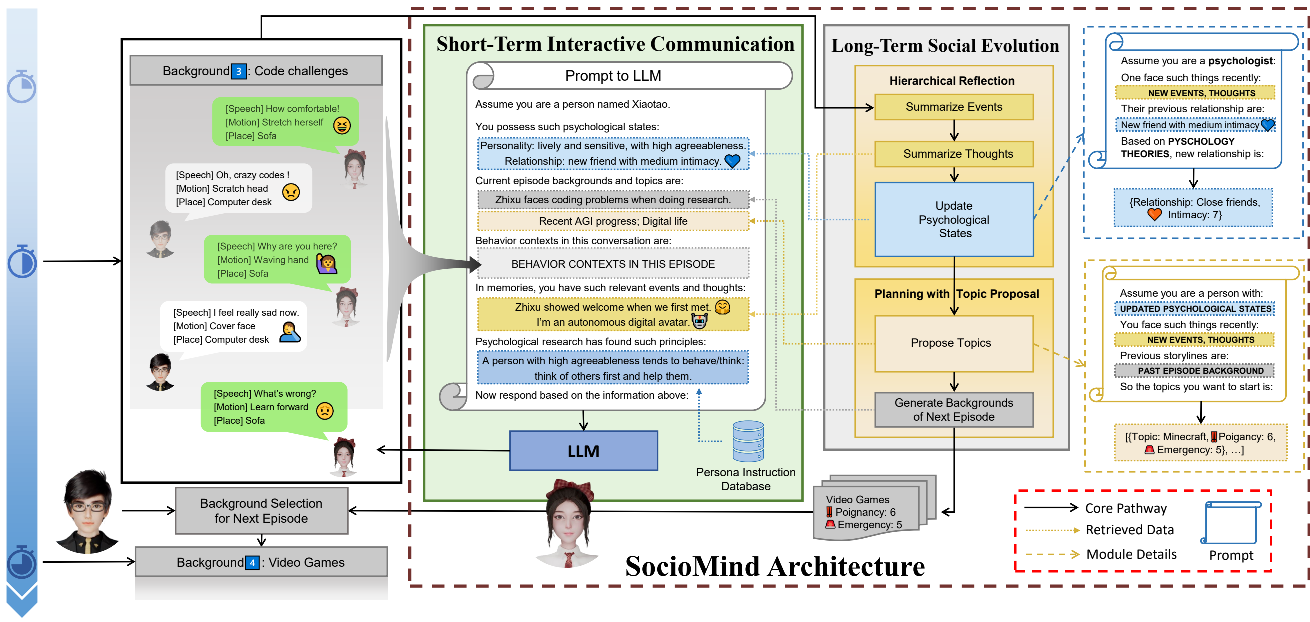

Figure 3. Overview of **SocioMind**. To enable 3D characters with social intelligence, our brain utilizes psychological principles to emulate controllable behaviors for short-term interactive communication. For long-term social evolution, our brain assures the consistency of psychological states and plots towards initial settings through psychological reflection and planning with topic proposal.

ities is the partner dance, where the leader/follower roles alternate [58]. Drawing from these observations, we design the Active-Passive Mechanism shown in Fig. 2, where the subject to whom the *behavior* is assigned becomes the "active" character, whereas the partner becomes the "passive" character. The active character generates a motion pair for both characters engaged in the interaction. Both passive *behavior* and the corresponding motion are then passed to the passive character. However, the passive character can still retain discretion: it only executes the passive motion if its brain "approves" the passive *behavior* (potentially by prompting the LLM with the suggested behavior and behavior context in its memory). Note that the "active" and "passive" roles constantly swap between characters as the interaction progresses.

3.3. Interactive Motion Synthesis

In our application scenario, the generated actions need to fulfill two main requirements: 1) They must be highly accurate to ensure natural interaction between characters, such as having sufficient contact when shaking hands. 2) They should generate diverse actions to adapt to different plots. In this paper, we propose a new paradigm called MoMat-MoGen to generate dual-person actions that are both diverse and accurate. As shown in Fig. 2, MoMat-MoGen leverages motion matching (Sec. 3.3.1) to achieve a relevant motion from a small database as a prior, and motion generation (Sec. 3.3.2) afterward to diversify the motion with text input while retaining interactive relations between two characters.

3.3.1 Motion Matching for High-Quality Motion Prior

The motion matching algorithms retrieve motion segments from a database in an auto-regressive manner based on predefined features. The basic motion matching [12] relies on state-based features (e.g., joint position) along with trajectory. The Story-to-Motion [47] further incorporates text-based features to enable semantic control. However, both methods are designed for single-person scenarios.

In this work, we extend the Text-based Motion Matching [47] to accommodate interactive scenarios. Our objective is to find a motion pair for both characters that aligns with the query text and trajectory while maintaining a consistent body pose to ensure coherence with the previous motion. In this light, we use a coarse-to-fine motion search strategy, leveraging the text for a high-level semantic understanding of the desired motion, and kinematic features for the low-level control. **First**, we incorporate semantic control by employing a pre-trained sentence encoder [36] to extract text embedding from the query text. Then top-K_1 candidates are selected using cosine similarity for subsequent matching. **Second**, trajectory and coherence constraints are incorporated through joint kinematics features. For the trajectory constraint, the features include the position of the hip joint and the facing direction. For the coherence constraint, the features include positions, velocities, and rotation in 6D space [78] of the body joints. For the two-person scenario, a new challenge arises: the interaction between the two characters requires that their relative positions and orientations align with the intended motions. Therefore, the relative position of the other character is taken into account to minimize blending artifacts caused by long-distance movements. To expedite retrieval, the aforementioned features are pre-calculated and Z-Score normalization is applied to account for magnitude differences. During retrieval, query features are calculated based on the current pose and target trajectory, and the Top-K_2 motions are selected using the Euclidean distance. Random selection is used if multiple suitable candidates exist.

Moreover, motion matching is used for single-person motions. This includes 1) navigation in the scene, where multi-agent path finder [55] is used to plan a collision-free trajectory, follow which walking motions are matched from AMASS, and 2) basic character-object interaction such as "sit down on the chair". More details are included in the Supplementary Material.

The neural motion blending model is used [47] to generate the transition motion. Hence, the short motion clips are blended into long motions. Notably, the blending model provides smooth transitions to let the character move to the correct place and turn in the correct direction to interact with the other character.

3.3.2 Motion Generation for Diversity

The MoMat-MoGen structure shares many similarities with ReMoDiffuse [76], incorporating retrieval techniques to enhance generation quality. However, applying ReMoDiffuse to interaction generation is not trivial. **Firstly**, it lacks a mechanism for interaction modeling, resulting in a poor correlation between the two generated sequences. **Secondly**, achieving physical naturalness is challenging if we solely rely on data-driven generation. To address these challenges, we 1) design a Dual-path Semantic-Modulated Attention module (DSMA) to model the interaction between two individuals. 2) During the inference stage, we adaptively extract interaction information from the referenced motion and use it as a constraint for the denoising process, providing additional supervisory signals.

Motion Diffusion Model. In the diffusion process, it repeatedly adds Gaussian noises to the clean motion sequence pair $(\mathbf{x_0}, \mathbf{y_0})$ to noised sequence pair $(\mathbf{x}_T, \mathbf{y}_T)$.

$$\begin{aligned} q(\mathbf{x}_T, \mathbf{y}_T|\mathbf{x}_0, \mathbf{y_0}) &:= \prod_{t=1}^{T} q(\mathbf{x}_t, \mathbf{y}_t|\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), \\ q(\mathbf{x}_t, \mathbf{y}_t|\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) &:= \mathcal{N}(\sqrt{1-\beta_t}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), \beta_t \mathbf{I}), \end{aligned} \tag{1}$$

where T is the total diffusion steps. $\beta_1, \cdots, \beta_T$ is a series of pre-defined variance scales for different timesteps. In the reverse process, given the text prompt P, the motion matching result $\bar{\Theta}$ and the timestep t, the initial sequence pair is estimated by a network $S_\theta(\mathbf{x}_t, \mathbf{y}_t, t, \bar{\Theta}, P)$.

Network Architecture. Similar to ReMoDiffuse, our network is built upon transformer layers. We modify the design of the attention module in ReMoDiffuse to better capture the interaction. Specifically, in our DSMA module, the input includes motion feature sequences, f_x and f_y, feature sequences extracted from the motion matching results, r_x and r_y, and text feature sequences p_x and p_y. When refining f_x, we utilize the generated global attention from f_x, f_y, r_x, p_x. The process is similar when refining f_y. This approach ensures a more comprehensive fusion of text information, interaction states, and prior information from motion matching.

Training and Inference. In the training stage, we only use the reconstruction loss as the target:

$$\mathcal{L} = \mathrm{MSE}((\mathbf{x}_0, \mathbf{y}_0), S_\theta(\mathbf{x}_t, \mathbf{y}_t, t, \bar{\Theta}, P)). \tag{2}$$

In the inference stage, we introduce a contact loss to make the interaction part more natural.

$$\bar{S} = S + \lambda \cdot \nabla(\sum_{i,j_1,j_2} \|\bar{D}_{i,j_1,j_2} - D_{i,j_1,j_2}\| \cdot [\bar{D}_{i,j_1,j_2} < \gamma]), \tag{3}$$

where $\bar{D}_{i,j_1,j_2}$ indicate the distance between the j_1-th joint and the j_2-th joint in the i-th frame from the motion matching results. D_{i,j_1,j_2} is the distance from the motion generation results. $[\cdot]$ is the Iverson bracket whose value is 1 if and only if the expression inside the parentheses is true. Otherwise the value will be 0. This auxiliary loss enforces the generated results to imitate the interaction pattern from the prior information and will yield more natural motions.

3.4. Controllable Emulation of Human Psychology

We aim to harness the advancements in large language models (LLMs) in building realistic social intelligence. From a social psychology perspective, human social intelligence is characterized by 1) various and patterned interactive behaviors during short-term communication [7], and 2) the evolution of emotions, attitudes, and relationships *etc*. over long-term interactions [11, 38, 48]. Hence, we propose SocioMind, a text-centric cognitive framework derived from the idea of "from strings to symbolic AGI" [39, 59]. As shown in Fig. 3, when avatars are engaged in communication, SocioMind prompts the LLM with psychological states, persona instructions, relevant memories, and context behaviors, to output *behavior* to manipulate the 3D character. Moreover, SocioMind autonomously reflects on psychological states at the end of each interaction session, where several rounds of *behaviors* are generated between characters. We refer to such a session as a episode. It also determines the background for the next episode through planning with topic proposal. We include more implementation details in the Supplementary Material.

3.4.1 Short-Term Interactive Communication

Interactive behaviors are strongly influenced by internal psychological states. Here we adopt the most critical dimensions with psychological theories: Big Five Trait model [33] for personality, long-term and short-term motivations [63], central beliefs [29], and trust [48, 50], intimacy [38], and supportiveness [13, 14] in social relationships. However, the safe alignment restricts current LLMs to a friendly and cooperative personality [40, 51, 62]. We introduce persona instructions to enhance the controllability of psychological states on behaviors below.

Persona instructions. In CoT [67], constructing accurate few-shot exemplars can effectively enhance the reasoning capability of LLMs. When prompting LLMs to infer behavior based on human psychological states, crafting precise and reliable exemplars presents a challenging task due to the lack of an exemplar database with high quality. Considering that lots of psychological tests [15, 19, 27, 37] measure psychological traits through observable behavior, we build a database of trait-to-behavior relationships from psychological tests. For psychological tests, we choose International Personality Item Pool (IPIP) [16, 18, 57], an open-sourced tool with over 3,000 items and 250 scales for creating advanced measures of personality, motivations, and *etc*. Each item, called *persona instruction*, in this database follows the format: "A person with {extent} {trait dimension} tends to behave/think: {behavior}", where {extent} are "high" or "low" according to the test questionnaire setup. For interactive behavior generation, we retrieve the most similar persona instructions by text embeddings to obtain few-shot exemplars, and include it in the prompt.

3.4.2 Long-Term Social Evolution

Long-term social intelligence requires consistency in two aspects with the initial character setup: 1) the evolution of psychological states such as emotions, relationships, and motivations *etc*. towards others [11, 13, 38, 48]; 2) the progression of overall plots or events [4]. SocioMind achieves the former aspect through psychological reflection and the latter aspect through planning with the topic proposal.

Psychological Reflection. Theories in social cognitive psychology [6, 10, 28] suggest that humans learn, attribute, and form judgments about others from past experiences. Therefore, we introduce a reflection mechanism based on psychological principles. Within each episode, agents introspect on their emotions periodically. At the end of each episode, agents summarize events and their thoughts into a memory system based on the behavior contexts. Events represent occurrences or facts perceived by the agent, whereas thoughts are ideas, musings, or attitudes generated by the agent based on their personality and past experiences. Leveraging current events and thoughts, agents retrieve past relevant events and thoughts, and reflect on their motives, central beliefs, and social relationships. For instance, after *'knowing they share the same interests'*, it is observed that the *intimacy* of two characters typically increases with psychological reflection.

Planning with Topic Proposal. We create a planning module with a topic proposal mechanism for diverse and plausible story progression. After each psychological reflection, each agent independently proposes new topics for the next episode based on past memories and character settings, followed by the background and initial states of both agents for the upcoming episode. The two agents collect the topics proposed by them and select the most important one for the next episode. Through this mechanism, the two agents can continuously interact with each other from one episode to the next. For example, after the topic proposal, the character wants to start several topics (such as the movie *'Mountains may depart'* with the highest *emergency* and *poignancy*) and generate the background *'Weekend Plan'* for next episode. The two characters, based on the proposals offered by each, will select an option that holds both high priority and significance, forming the background for the subsequent episode.

4. Experiments

To the best of our knowledge, Digital Life Project is the first comprehensive framework to enable autonomous social characters with articulated 3D bodies. In addition to MoMat-MoGen and SocioMind, we also evaluate a motion captioning module as an extension of DLP, on the KIT-ML [46] and HumanML3D [23] datasets in the Supplementary Material.

4.1. Interactive Motion Synthesis

We evaluated the proposed MoMat-MoGen module on two datasets: the public InterHuman dataset [35] and DLP-MoCap, an optical motion capture dataset for interactive motion generation. Due to space constraints, the test results on the DLP-MoCap are included in the Supplementary Material. Tab. 1 presents a comparative analysis of our proposed interactive motion generation method against three existing approaches: ReMoDiffuse [76], MotionDiffuse [75], and InterGen [35]. Our method exhibits significant improvements on the InterGen dataset, especially in R precision, FID, MM Dist, and Diversity metrics. It is noteworthy that we achieve an impressive balance between precision and diversity, which is essential for our application, ensuring that the generated motions closely resemble

Table 1. **Interactive Motion Synthesis results on the InterHuman 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 results are in bold and the second best results are underlined.

Methods	R Precision↑ Top 1	Top 2	Top 3	FID↓	MM Dist↓	Diversity↑	MultiModality↑
Real motions	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
TEMOS [45]	$0.224^{\pm.010}$	$0.316^{\pm.013}$	$0.450^{\pm.018}$	$17.375^{\pm.043}$	$6.342^{\pm.015}$	$6.939^{\pm.071}$	$0.535^{\pm.014}$
T2M [23]	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$5.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
MDM [61]	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$7.125^{\pm.018}$	$7.602^{\pm.045}$	$\mathbf{2.355}^{\pm.080}$
ComMDM [54]	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.010}$	$7.069^{\pm.054}$	$6.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
MotionDiffuse [75]	$0.401^{\pm.004}$	$0.541^{\pm.004}$	$0.622^{\pm.005}$	$12.663^{\pm.083}$	$3.805^{\pm.001}$	$7.639^{\pm.035}$	$1.176^{\pm.027}$
ReMoDiffuse [76]	$\underline{0.442^{\pm.004}}$	$\underline{0.589^{\pm.005}}$	$\mathbf{0.666}^{\pm.003}$	$6.366^{\pm.102}$	$\underline{3.802^{\pm.001}}$	$\underline{7.956^{\pm.030}}$	$1.226^{\pm.044}$
InterGen [35]	$0.371^{\pm.010}$	$0.515^{\pm.012}$	$\underline{0.624^{\pm.010}}$	$\underline{5.918^{\pm.079}}$	$5.108^{\pm.014}$	$7.387^{\pm.029}$	$\underline{2.141^{\pm.063}}$
Ours (MoMat-MoGen)	$\mathbf{0.449}^{\pm.004}$	$\mathbf{0.591}^{\pm.003}$	$\mathbf{0.666}^{\pm.004}$	$\mathbf{5.674}^{\pm.085}$	$\mathbf{3.790}^{\pm.001}$	$\mathbf{8.021}^{\pm.035}$	$1.295^{\pm.023}$

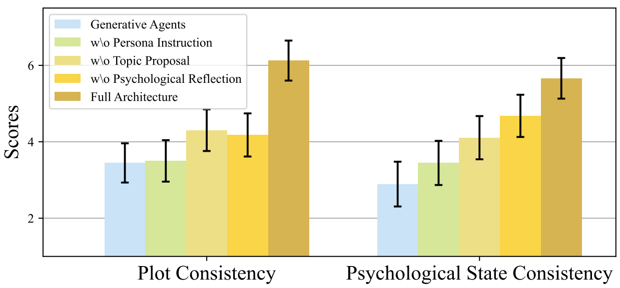

Figure 4. Ablation results on consistency with 95% confidence.

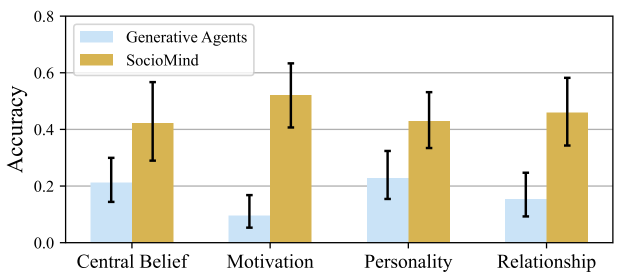

Figure 5. Results on controllibility with 95% confidence.

the high-quality motion references with strong priors yet exhibit a broad range of variety.

4.2. Social Intelligence

To evaluate the social intelligence of SocioMind, we measure the controllability of behaviors in short-term interactive communication and the consistency of psychological states and plots in long-term social evolution. Following the previous evaluation approach [43], we engage 47 human evaluators to review the behavioral records of the agents. More details are included in the Supplementary Material.

4.2.1 Controllability

Controllability is measured by whether altering psychological traits can cause noticeable different behaviors in short-term communication. We show evaluators the behavioral records of 64 episodes, ask them to select the corresponding psychological traits from multiple options, and subsequently calculate the accuracy. Results in Fig. 5 show that SocioMind significantly outperforms Generative Agents [43] in key attributes: central belief, motivation, personality, and relationship, demonstrating the effective guidance of persona instructions for the LLM in simulating interactive human behavior.

Table 2. **User study** on the integrated performance.

"Brain"	"Body"	Script↑	Motion↑	Overall↑
GA [42]	InterGen [35]	5.57	5.03	4.93
GA [42]	MoMat-MoGen	5.88	6.12	6.07
SocioMind	InterGen [35]	6.28	4.60	4.78
SocioMind	MoMat-MoGen	**7.17**	**6.77**	**6.88**

4.2.2 Consistency

Long-term social evolution consistency implies that the plot development and internal state changes are coherent with initial settings. To measure this, we use four different types of initial settings (family, crime, romance, and military) to generate records with multiple episodes. Human evaluators use the records to rate the degrees of consistency on plots and psychological states on a scale of 1 to 9. Thus we evaluate the effectiveness of modules in the SocioMind for social evolution. Results in Fig. 4 show that SocioMind demonstrates superior performance over Generative Agents [43] on consistency over plots and psychological states, and ablating results show that persona instruction, psychological reflection, and planning with topic proposal are crucial for long-term social evolution.

4.3. Integrated Evaluation

We further conduct a user study with 30 human participants to evaluate the entire pipeline. We use SocioMind and Generative Agents (GA) [42] as the "brain" to generate full episode scripts given various contexts (*e.g.*, "Xiaotao is sad lately"), and MoMat-MoGen and InterGen [35] as the

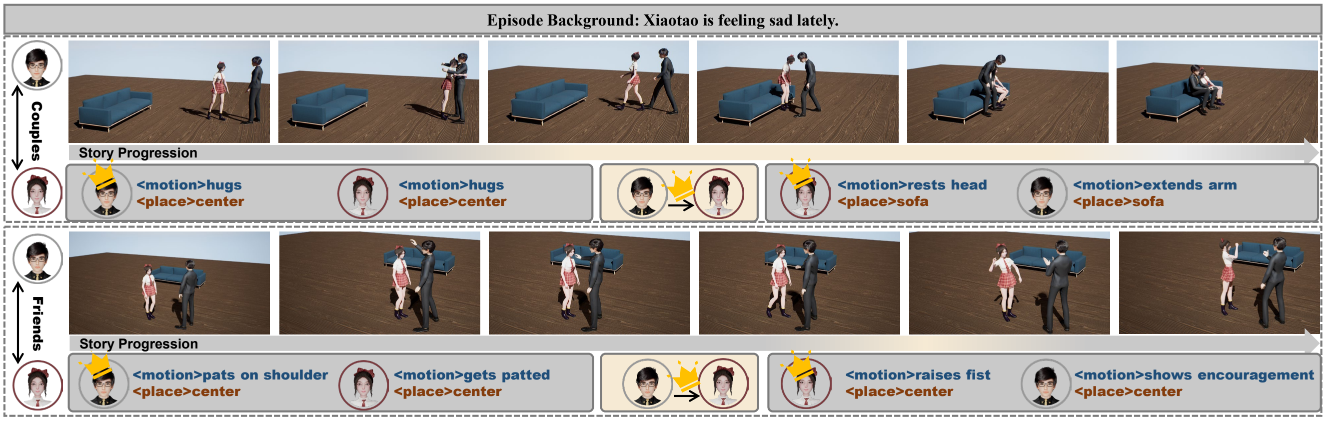

Figure 6. We explore the **controllability** of DLP. Given the same background, manually editing the relationship state between characters, results in different social behaviors. Interestingly, "couples" tend to have more intimate interactions than "friends". The crown indicates the active player. The story progression bar is color-coded in accordance with the stages represented by boxes: gray boxes represent *behaviors*, whereas yellow boxes represent active-passive swapping in between *behaviors*.

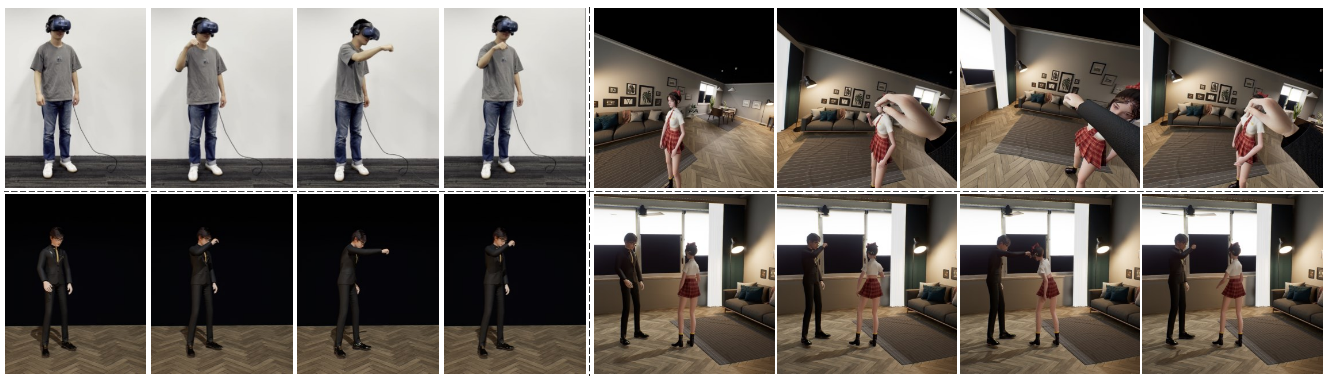

Figure 7. Our motion captioning module translates human motion into text description, allowing a virtual character to respond to the human player's "fist bump". Top Left: RGB video of the human player; Bottom Left: motion capture [8] result; Top right: first-person view of the human-driven character; Bottom right: third-person view of the interaction. More details are included in the Supplementary Material.

"body" to synthesize character motions based on the motion descriptions. We then render videos of the characters and ask evaluators to rate the script quality, motion quality, and overall quality from 1 to 9. in Tab. 2 shows our SocioMind and MoMat-MoGen deliver better results with convincing margins.

4.4. Visualization

As shown in Fig. 6, our framework possesses a rational correlation between psychological states and physical behaviors. In addition, our system has the potential to add human players in the virtual world to interact with the digital avatars (Fig. 7, elaborated in the Supplementary Material).

5. Conclusion

In this paper, we introduce Digital Life Project, an innovative and comprehensive system that harnesses the latest advancements in generative models to create autonomous 3D characters. DLP integrates SocioMind, a text-centric cognitive framework that simulates sophisticated internal psychological processes, and MoMat-MoGen, a text-driven motion synthesis pipeline that replicates diverse external physical behaviors. Both modules achieve state-of-the-art performance in the respective domains, enabling the entire system to engage in natural interactions with social intelligence.

Acknowledgement

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). We extend our sincere gratitude to Fei Wang, Jingyi Jiang, Jinyun Lyu, Wanying Zu, Bo Li, Yuanhan Zhang, Jiawei Ren, and Cunjun Yu for their insights and helpful discussions.

References

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023. 3

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 3

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 2

[4] Irwin Altman and Dalmas A Taylor. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston, 1973. 6

[5] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 2023. 3

[6] Albert Bandura and Richard H Walters. *Social learning theory*. Englewood cliffs Prentice Hall, 1977. 6

[7] Charles R Berger, Michael E Roloff, and David R Ewoldsen. *The handbook of communication science*. Sage, 2010. 2, 5

[8] 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. 8

[9] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2

[10] Chi-yue Chiu, Ying-yi Hong, and Carol S Dweck. Lay dispositionism and implicit theories of personality. *Journal of personality and social psychology*, 73(1):19, 1997. 6

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

[12] Simon Clavet. Motion matching and the road to next-gen animation. In *Proc. of GDC*, 2016. 2, 4

[13] Sheldon Cohen. Social relationships and health. *American psychologist*, 2004. 6

[14] Sheldon Cohen and Thomas A Wills. Stress, social support, and the buffering hypothesis. *Psychological bulletin*, 1985. 6

[15] Paul T Costa and Robert R McCrae. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, pages 179–198, 2008. 6

[16] 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. 6

[17] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021. 3

[18] 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. 6

[19] Hans Jurgen Eysenck and Sybil Bianca Giuletta Eysenck. *Manual of the Eysenck Personality Questionnaire (junior & adult).* Hodder and Stoughton Educational, 1975. 6

[20] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S^3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023. 3

[21] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1396–1406, 2021. 2

[22] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: Reactive 3d motion synthesis for two-person interactions. In *arXiv*, 2023. 3

[23] 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. 2, 6, 7

[24] 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. 2

[25] Edward Twitchell Hall. *The hidden dimension*. Anchor, 1966. 2

[26] Edward T Hall. *The silent language*. Anchor, 1973. 2

[27] Starke R Hathaway and John C McKinley. A multiphasic personality schedule (minnesota): I. construction of the schedule. *The Journal of Psychology*, 1940. 6

[28] Fritz Heider. *The psychology of interpersonal relations*. Psychology Press, 2013. 6

[29] E Tory Higgins. Self-discrepancy: a theory relating self and affect. *Psychological review*, 1987. 6

[30] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM TOG*, 39(4): 53–1, 2020. 2

[31] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. 3

[32] William James. *The principles of psychology*. Cosimo, Inc., 2007. 2

[33] Oliver P John, Sanjay Srivastava, et al. The big-five trait taxonomy: History, measurement, and theoretical perspectives. 1999. 6

[34] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 2

[35] 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. 3, 6, 7

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019. 4

[37] Isabel Briggs Myers, Mary H McCaulley, and Robert Most. Manual: A guide to the development and use of the myers-briggs type indicator. *(No Title)*, 1985. 6

[38] Theodore M Newcomb. The prediction of interpersonal attraction. *American psychologist*, 1956. 5, 6

[39] Allen Newell, Paul S Rosenbloom, and John E Laird. Symbolic architectures for cognition. *Foundations of cognitive science*, 1989. 5

[40] OpenAI. Gpt-4 technical report, 2023. 2, 3, 6

[41] Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *UIST*, 2022. 3

[42] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, New York, NY, USA, 2023. Association for Computing Machinery. 2, 7

[43] 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. 3, 7

[44] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023. 3

[45] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2, 7

[46] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 6

[47] 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. 2, 4, 5

[48] John K Rempel, John G Holmes, and Mark P Zanna. Trust in close relationships. *Journal of personality and social psychology*, 1985. 5, 6

[49] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, Xiao Ma, Liang Pan, and Ziwei Liu. Insactor: Instruction-driven physics-based characters. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[50] Julian B Rotter. A new scale for the measurement of interpersonal trust. *Journal of personality*, 1967. 6

[51] Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023. 3, 6

[52] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. 3

[53] 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. 2

[54] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3, 7

[55] Guni Sharon, Roni Stern, Ariel Felner, and Nathan R Sturtevant. Conflict-based search for optimal multi-agent pathfinding. *AI*, 219:40–66, 2015. 5

[56] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 3

[57] 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. 6

[58] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. 2024. 4

[59] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023. 3, 5

[60] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2

[61] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 7

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 3, 6

[63] Robin R Vallacher and Daniel M Wegner. What do people think they're doing? action identification and human behavior. *Psychological review*, 1987. 6

[64] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. 3

[65] Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22035–22044, 2023. 2

[66] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 2, 3

[67] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 6

[68] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-x: Towards versatile human-human interaction analysis. *arXiv preprint arXiv:2312.16051*, 2023. 3

[69] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*, 2023. 3

[70] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. 3

[71] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023. 3

[72] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16010–16021, 2023. 2

[73] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023. 2

[74] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[75] 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. 2, 6, 7

[76] 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, 5, 6, 7

[77] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[78] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 4

[79] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023. 3