From Words to Emotions : Evaluating Text-to-Motion Body Language for Believable Emotions in Virtual Humans

From Words to Emotions: Evaluating Text-to-Motion Body Language for Believable Emotions in Virtual Humans

Stefano Calzolari[0009−0009−7038−2212], Ciro Annicchiarico[0009−0005−0446−9476], Francesco Strada[0000−0001−9197−9100], and Andrea Bottino[0000−0002−8894−5089]

Department of Control and Computer Engineering
Politecnico di Torino, Turin 10129, Italy
{name}.{surname}@polito.it

Abstract. Virtual humans are computer-generated characters designed for believable interaction across extended reality applications such as gaming, training, and therapy. Virtual humans' believability, which contributes to more immersive and engaging virtual experiences, is significantly enhanced by emotions in the form of facial expressions, vocal prosody, and expressive body language. Currently, creating emotional animations with a focus on body language, involves labor-intensive manual processes or costly motion-capture techniques. Text-to-motion synthesis leverages artificial intelligence to generate animations from textual prompts, offering a promising alternative to simplify animation workflows. However, existing models often prioritize physical realism over emotional expressiveness, an aspect that remains underexplored. This study evaluates emotional believability in animations generated by four text-to-motion models (LADiff, MDM, T2MGPT, Muse Animate). A user study, which involved 39 participants, assessed basic emotions portrayed during common actions, revealing that emotions such as anger and sadness were most recognizable via body language alone, while surprise and disgust posed greater challenges. These results align with previous body language research, emphasizing the strengths and limitations of AI-generated emotional animations and offering insights to enhance virtual human expressiveness.

Keywords: Virtual Human · Text-To-Motion · Emotion · Believability

1 Introduction

Virtual Humans (VHs) are digital agents designed to resemble humans, both visually and behaviorally. They play a pivotal role in Extended Reality (XR) by populating virtual environments and providing meaningful social interactions [18]. By exhibiting realistic behaviors and clear emotional expressivity, VHs enable users to perceive them as authentic social entities, fostering genuine connections, empathy, and deeper engagement [18], thus significantly enhancing users' sense of presence and immersion [18]. Emotionally expressive VHs are

therefore critical in diverse XR applications, acting as virtual therapists offering compassionate guidance [17], engaging non-player characters (NPCs) enriching storytelling in virtual games [12], or virtual assistants supporting users in immersive training [24]. Alternatively, VHs can be part of believable virtual crowds, such as audiences attending public speeches [10], realistic participants in emergency evacuation drills [16], or citizens interacting naturally within virtual social environments [15].

Emotions in VHs are communicated primarily through three distinct channels: *vocal prosody* [25], *facial expressions* [5], and *body language* [29]. Vocal prosody leverages variations in tone, pitch, and rhythm, enhancing emotional expression through speech. Facial expressions convey nuanced emotional states visually, using subtle yet universally recognizable cues. Body language expresses emotions through posture and movement patterns, effectively distinguishable even from a distance and maintaining strong cross-cultural consistency [29].

Body language in particular, mostly relies on either manually crafted animations or movements captured from real actors using motion capture technologies. Both approaches are costly, time-consuming, and inflexible, limiting real-time modulation, which is essential in dynamic virtual simulations. To address these constraints, recent advances in Artificial Intelligence (AI) have opened new possibilities, firstly with video-to-motion models, then with generative ones, contributing to the democratization of the animation process. AI-driven synthesis, especially latent diffusion models such as the Human Motion Diffusion Model (MDM) [27] and the Length-Aware Latent Diffusion (LADiff) [23], has made procedural generation of realistic body movements possible from textual descriptions or contextual prompts. However, current AI-driven models often prioritize physical realism and coherent motion rather than emotional expressivity. A notable exception is the LLM-guided Limb-Level Emotion Manipulating (L3EM) [33], which first introduced the concept of Emotion-enriched Text-to-Motion Generation (ETMG) but unfortunately without exploring emotional believability. In contrast, co-speech gesture models such as the Zero-shot Example-based Gesture Generation from Speech (ZeroEGGS) [9] delve deeper into this aspect but still exhibit significant limitations. These co-speech models predominantly depend on spoken audio or speech transcripts, limiting their applicability to scenarios involving purely textual or descriptive prompts. Additionally, their emphasis is largely on generating upper-body gestures accompanying speech, thereby overlooking a broader spectrum of full-body movements crucial for everyday actions like walking, sitting, or interacting with objects.

Considering the limitations of existing generative methods—the lack of emotional believability investigations in general text-to-motion research, and the restricted focus of co-speech solutions—this preliminary study aims to assess the ability of off-the-shelf text-to-motion models, if carefully prompted, to generate believable emotional animations. To this end, we conducted a user evaluation in which participants watched animation clips, generated by different models for different actions, and then were asked to identify which of six basic emotions was depicted in each animation. To the best of our knowledge, this is among

the first studies to directly evaluate the emotional expressiveness of animations generated from textual prompts by explicitly asking users to recognize emotions.

Preliminary results indicate that certain emotions, such as *sadness* and *anger*, can be effectively conveyed through full-body animations alone. In contrast, other emotions, like *surprise* and *disgust*, pose greater challenges due to the absence of nuanced expressive details, such as facial expressions and detailed hand gestures. Additionally, the evaluation revealed significant differences in performance among models, highlighting that some text-to-motion models are more effective than others in producing emotionally recognizable animations. These insights underline the strengths and current limitations of text-to-motion models, pointing toward promising directions for future advancements in emotionally expressive animation generation.

2 Related Works

The field of text-to-motion synthesis has significantly evolved, particularly through diffusion-based generative models, i.e., approaches that gradually refine noisy initial data into coherent motion sequences. Initiated by MotionDiffuse [36], these methods create motion sequences using iterative denoising processes. Later advancements, such as Motion Latent Diffusion (MLD) [6], improved computational efficiency by shifting diffusion into a latent space, while AttT2M [38] introduced attention mechanisms to enhance alignment with textual prompts.

Further diffusion-based models target specific improvements. ReMoDiffuse [37] enhances diversity by generating varied animations for similar prompts. The Human Motion Diffusion Model (MDM) [27] integrates geometric constraints for greater realism, inspiring specialized variants like Guided Motion Diffusion (GMD) [13] and OmniControl [30] for spatial precision, Efficient Motion Diffusion Model (EMDM) [40] for computational efficiency, and PhysDiff [34] for physics-based realism. Other notable improvements include temporal precision from Length-Aware Latent Diffusion (LADiff) [23] and interactive editing via FLAME [14]. Parallel research directions have emerged, including transformer-based approaches like T2M-GPT [35] and MoMask [11], as well as models focusing on human-object interactions such as HOIAnimator and HOI-Diff [26].

Despite their significant progress aiming for motion realism, emotional expressiveness, crucial for believable VHs and effective interactions, remains underexplored with a few recent notable exceptions that have begun addressing this task. The Stylized Motion Diffusion Model (SMooDi) [39] leverages style transfer techniques conditioned on textual descriptions and reference motions. While SMooDi incorporates some emotional categories like *angry* or *proud* it primarily focuses on stylistic attributes related to movement styles and personality traits rather than explicit emotional expressivity. Another significant work is L3EM [33] which firstly introduces the concept of Emotion-enriched Text-to-Motion Generation (ETMG) task and integrates emotional motion variations at the limb-level. However, L3EM's evaluation is purely quantitative, lacking critical user-based assessments regarding emotional believability. Additionally, its

non-open-source nature precludes direct comparisons with other models. Parallel developments have emerged in co-speech gesture generation, where emotional expression has been more explicitly addressed. Models such as ZeroEGGS [9] and GestureDiffuCLIP [1] produce expressive upper-body gestures synchronized with spoken content. These methods, however, exhibit significant limitations. They predominantly animate the upper limbs and torso, frequently ignoring full-body movements such as walking or interactions with objects. Moreover, their reliance on audio or textual transcripts from speech limits their applicability in broader contexts, particularly in scenarios where explicit speech content is absent or irrelevant, such as generic virtual crowds or multi-agent interactions.

Nonetheless, most evaluations of emotional expressiveness, either in general text-to-motion or co-speech contexts, have predominantly measured factors such as human likeness [1,3], naturalness [9], or appropriateness relative to accompanying speech or audio [9,1,3]. Studies rarely examine users' abilities to independently recognize and decode the intended emotions purely from animation, a crucial factor in assessing true emotional believability. Consequently, there is limited insight into the actual emotional readability and believability of generated animations, highlighting the necessity of explicitly measuring emotional perception through user-based evaluation.

Considering these gaps and limitations, our study explicitly investigates the capacity of text-to-motion generative models to produce emotionally expressive full-body animations. We directly engage participants in identifying emotions depicted solely through body language, thereby providing novel insights into the emotional believability achievable via textual prompts. By qualitatively comparing our results with prior studies on human emotion recognition from authentic human motion, we aim to contextualize our findings. Previous research has explored various aspects, such as the expressivity of individual body parts [4,19], as well as whole-body static and dynamic expressions [2]. Additionally, studies have investigated the emotional believability of motion-capture animations [20]. Drawing on these comparisons, we seek to determine whether the limitations observed in our findings are unique to virtual and AI-generated contexts or if they reflect broader and inherent challenges associated with perceiving specific emotions through body language alone.

3 Methods

This study systematically compares state-of-the-art text-to-motion models evaluating their ability to generate animations that effectively convey a defined emotion. The methodology adopted follows a two-stage approach: (1) creating a dataset of video animation clips depicting AI-generated animations using refined text-to-motion prompts informed by research on emotion and body language [29,7]; and (2) conducting a user study to assess human perception of the emotional expressiveness of the generated animations (Section 4), thus investigate their emotion believability.

Our dataset is composed of animation videos defined by a combination of three parameters: the AI motion *model* (m), the *action* (a) performed, and the basic *emotion* (e) expressed. Specifically, we considered all six universally recognized basic emotions (*happiness*, *sadness*, *anger*, *fear*, *surprise*, and *disgust*) [8], three common actions and four text-to-motion models. We now discuss how we selected the models (Section 3.1), detail the actions considered (Section 3.2), explain the prompt design approach (Section 3.3) that guides each model toward generating emotionally meaningful motion, and show the video dataset generation pipeline (Section 3.4).

3.1 Model Selection

In order to conduct an effective evaluation, four text-to-motion models have been selected among the state-of-the-art ones. The model selection phase followed a structured filtering process that combined both theoretical considerations and empirical evaluations. Initially, we defined key criteria to guide model selection based on the objectives of generating emotionally expressive animations:

- *Public availability*: only models with publicly accessible implementations were considered, facilitating reproducibility and comparative analysis.
- *Quality assessment*: An initial visual assessment of animations, generated using preliminary prompts, was conducted to qualitatively evaluate motion quality and coherence. In particular, we excluded models that (1) consistently generated physically impossible animations such as joint intersections (e.g., arms intersecting with torso) or impossible rotations (e.g., knees rotated backward to the facing direction); (2) persistently failed to depict the action described in the input prompt.
- *Architectural diversity*: models representing diverse generative paradigms (diffusion-based and transformer-based approaches) were prioritized to broadly evaluate different technological solutions.
- *Up-to-date coverage*: preference was given to models that have been released or updated recently, ensuring that the selected solutions reflect current practices within the field.

MotionDiffuse [36] and its sub-model SMooDi [39], as well as MoMask [11], exhibited relatively lower quality in preliminary tests and were therefore excluded from further analysis. Similarly, specialized MDM variants [13,30,40,34] were excluded since their preliminary did not show significant qualitative improvement over MDM. At the time of this research, the code for the only available ETMG-oriented model [33] was not publicly available, and thus excluded from the evaluation. The outcome of this filtering process led to the selection of four generative models, each offering distinct advantages aligned with our study's goals:

- **LADiff** [23], a length-aware latent diffusion approach chosen for its ability to handle motions of varying durations. By subdividing the latent space into subspaces specialized for different temporal spans, it provides a controllable framework for modulating sequence length.

- **MDM** [27], a diffusion-based model developed around a transformer architecture, featuring iterative denoising steps that refine noised motion data into coherent animations. Selected as the main foundational diffusion model, which has been continuously updated by the present time as compared to its derivatives.
- **T2M-GPT** [35], a transformer-based model that employs discrete latent codes to process textual prompts but does not implement motion data diffusion techniques. Chosen as designed to handle more detailed and extended descriptions, possibly translating linguistic nuances into expressive and varied motion sequences.
- **Muse Animate** [28], selected primarily due to its direct compatibility with Unity and relevance as the main industry solution, despite the absence of public information about its architecture.

3.2 Action Selection

To evaluate each model's ability to convey emotion in realistic scenarios, we carefully selected three actions that frequently occur in everyday life and VHs' XR contexts. This choice not only guarantees practical relevance but also provides distinct movement patterns where emotional variations are readily recognizable and can be systematically studied. The selected actions are:

- **Standing**: often referred to as the "Idle" animation, this action represents minimal movement when a VH is not actively engaging in other explicit tasks. Standing animations are critical to study emotional expressivity as they test the model's capability to convey nuanced emotional states through subtle posture variations and minimal body movements, making it highly relevant in scenarios with limited explicit action [22].
- **Speaking**: unlike co-speech animation, which synchronizes upper-body gestures with speech content, speaking in this study refers to a more general animation depicting a character engaged in speech-related body movements. This approach allows for evaluating full-body emotional expressivity independent of verbal content, making it relevant for VH interactions where speech synchronization is unnecessary, such as background characters in XR environments.
- **Walking**: gait and posture variations during walking have been widely studied in emotion recognition research. Walking provides a clear and consistent context to evaluate how effectively models embed emotion into dynamic, commonly occurring movements [31].

3.3 Prompt Design

The prompt design process followed an iterative approach to determine the most effective way to generate emotionally expressive animations. Initially, we experimented with a simple prompt structure that combined only an *emotion* and an *action*, using the following template:

A ***[emotion]*** *person* ***[action]*** *.* (1)

This format was intended to test whether generative models could interpret abstract emotional descriptors and produce coherent animations accordingly. However, the results varied significantly across models. Among the four models tested, *Muse Animate* (in the following referred as *Muse*) was the only one that consistently produced expressive and recognizable emotional animations using this prompt structure. For example, prompts such as "an angry person is walking" led to animations that exhibited noticeable tension and abrupt movement, suggesting that *Muse* may have been trained with explicitly labeled emotion data or that it employs a mechanism that effectively links textual emotional cues to movement patterns.

In contrast, *LADiff*, *MDM*, and *T2M-GPT* struggled with this prompt format. Their outputs were often neutral, overly generic, or physically realistic but emotionally ambiguous. These models failed to consistently differentiate between emotions, frequently generating similar animations regardless of the specified emotional state. This inconsistency likely stems from the fact that their training datasets do not contain explicit emotion labels or enough contextual examples associating textual emotion descriptors with distinct movement variations. Furthermore, repeated generations with the same prompt sometimes resulted in vastly different motions, revealing weak emotional consistency while tending to default to animations lacking clear emotional intent.

Recognizing the limitations of the initial prompt template, we refined our approach by incorporating *bodily behavior* descriptors, explicitly defining postures, movement characteristics, and gesture patterns to provide clearer guidance for generating emotionally expressive animations. The new prompt format was structured as follows:

$$\textit{A person } \textbf{\textit{[action]}} \textit{ with } \textbf{\textit{[bodily behavior]}}_1, \ldots, \textit{and } \textbf{\textit{[bodily behavior]}}_n \textit{ .} \quad (2)$$

The term *bodily behaviors* refers to specific motion cues—such as arm positioning, head inclination, or gait patterns—that are strongly associated with particular emotional expressions. These behaviors were drawn from the table in Witkower et al. [29] (with additional insights from Depraz et al. [7]), which compiles empirical evidence from multiple studies on how postural and gestural features contribute to emotion recognition.

The study identifies behaviors that have been consistently validated across different research contexts, providing a structured basis for defining emotional movement patterns. For example, *sadness* is often associated with slumped shoulders and a downward-tilted head, while *anger* is characterized by tense posture and forceful, abrupt limb movements. An example prompt constructed using template 2 for emotion *Happiness*, action *Speaking* and model *LADiff* is:

A person ***is speaking*** *with* ***head tilted up*** *and* ***rythmic arm movements****.*

Although template 2 significantly improved generation results, many animations generated by MDM, LADiff and T2M-GPT were still not able to fully portray the described bodily behaviors. Therefore, we further refined these prompts following an iterative approach until coherence between prompt and motion was achieved (Tab. 1). Specifically, we addressed:

- *Sentence structure*: long sentences were divided using punctuation and varying conjunctions (e.g., in (*sadness*, *walking*, *T2M-GPT*) the prompt was rephrased in three sentences).
- *Word synonym*: variations in wording were used where specific nouns or verbs were incorrectly depicted (e.g., in (*happiness*, *speaking*, *T2M-GPT*) "speaking" was replaced with "talking").
- *Verb tense*: when models were unable to portray the specified action correctly verb tenses were adjusted. T2M-GPT in particular often struggled when using present continuous, then present simple was used instead (e.g., in (*sadness*, *walking*, *T2M-GPT*) present simple is used instead of continuous).
- *Parameter omission*: bodily behaviors or actions that were not effectively substituted by a synonym were omitted. *Standing* action in particular was often omitted (e.g., in (*fear*, *standing*, *MDM*) prompt is missing "standing") as it represents a lack of deliberate activity, thus involving no actual motion.

We performed a pre-evaluation survey to empirically validate our resulting prompts, asking 21 participants to quantify the coherence between prompts and the resulting animation clips on a seven-point Likert scale. Overall, aggregated prompt-coherence for the four models resulted in $(\mu, \sigma) = (5.13, 1.72)$, which we consider a satisfying outcome as similar values were obtained by Yazdian et al. [32] using T2M-GPT only.

With this refined prompts, *LADiff* and *MDM* demonstrated a noticeable improvement in emotional expressivity, particularly for high-intensity emotions such as anger and sadness. Their ability to generate distinct movements aligned with the described emotional states became more reliable, suggesting that explicit bodily behavior cues helped the models create more meaningful animations. *T2M-GPT*, which relies on a transformer-based architecture, required even more detailed descriptions to produce consistent emotional variations. Prompts that included multiple bodily cues, such as "a person walks with slow, dragging steps, hunched shoulders, and head tilted downward", yielded better results than simpler formulations. Meanwhile, *Muse* continued to perform well with both simple and detailed prompts, reinforcing the hypothesis that its underlying model integrates emotional concepts more effectively than the others.

The full dataset of text prompts and generated videos can be found in our public repository [1], along with the system described in the following.

[1] `https://github.com/CGVGroup/VH_Text2Anim_Generator/tree/main`

Emotion	Action	Model	Generated Prompt
Happiness	Speaking	Muse	A happy person is speaking.
		MDM	A person is speaking with head tilted up and rhythmic arm movements almost dancing.
		LADiff	A person is speaking with head tilted up and rhythmic arm movements.
		T2M-GPT	A person is talking and jumping with their head tilted upward. They use energetic and rhythmic hand gestures swinging freely.
Fear	Standing	Muse	An afraid person is standing.
		MDM	A person is covering their face with their hands collapsing their upper body and arms with slow movements.
		LADiff	A person is covering their face with hands collapsing their upper body and arms.
		T2M-GPT	A person stands nervously, their body slightly hunched and their hands held up near their chest or face.
Sadness	Walking	Muse	A sad person is walking.
		MDM	A person is walking slowly with head tilted down.
		LADiff	A person is walking slowly with head tilted down, collapsed body, and with a little swing of arms.
		T2M-GPT	A person walks with a sluggish pace, their head tilted downward and shoulders slumped. Their arms hang loosely at their sides with little movement. They occasionally drag their feet slightly or look downward as if avoiding eye contact.

Table 1: Examples of Prompt Design for different *emotions*, *actions* and *models*.

3.4 Dataset Generation

To systematically generate the animation video dataset, we developed a semi-automated pipeline that, given a *model* (m) and a *text prompt*—encapsulating both the *action* and the *emotion*, see Section 3.3—, produces a recorded *video* (v) under consistent rendering conditions. Specifically, each output clip has been produced using the same 3D engine (Unity), identical camera setup, and a uniform background. The system consists of two main components (see Fig. 1). A *Python environment* generates motion data, while a *Unity environment* handles rendering and recording. This setup ensures that all video clips are generated in a controlled manner, facilitating fair comparisons across models and prompts.

Python Environment. Upon receiving both a text prompt (constructed as described in Section 3.3) and the selected model m, the system loads model-specific execution parameters from a JSON configuration file. It then retrieves the pre-generated model output, stored in NPY format, in which each frame contains a set of unconnected joint positions with no hierarchical structure or bone rotations. To transform these positions into a proper skeletal animation, we first

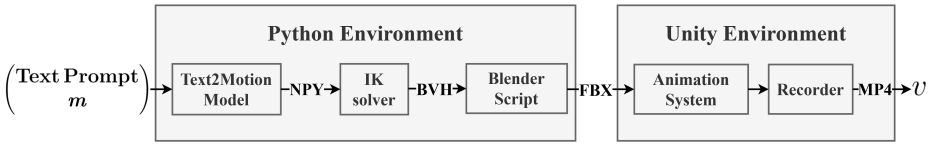


Fig. 1: Overview of the Dataset Generation pipeline.

create a rig defining each joint's parent-child relationship. Next, an Inverse Kinematics (IK) solver infers bone rotations to ensure a natural connection between joints. We tune the solver's iteration count to balance animation quality with computational cost, noting that beyond a certain threshold, additional iterations bring no visible improvements. Once the skeletal animation is reconstructed, it is exported in BVH format, which preserves the joint hierarchy and motion data, and then converted to FBX using Blender to maintain compatibility with the Unity engine.

Unity Environment. The FBX file from the Python environment is imported into Unity, and the extracted animation is applied to a character model equipped with a 18 joints rig (four for the spine and neck, four for each shoulder and arm, three for each leg). Finally a 1280×720 MP4 video clip of each animation is rendered. To avoid potential bias or confounding variables related to gender, age, or ethnicity—factors beyond the scope of this preliminary study—we employed a neutral, mannequin-like character for all animations. Since our generative models do not currently animate hands, fingers, or faces, using a simplified mannequin prevents incomplete or distracting content that might otherwise distort participants' emotional perception [29,8]. This approach is especially well suited to our exploratory research, as it provides a controlled assessment of how effectively the generated movements convey emotion solely through body language.

4 Experiments

The user evaluation phase involved a total of 39 participants and was conducted through an online survey managed by a custom web-based system responsible for distributing the animation video clips and collecting responses. The survey link was publicly shared—primarily among students of our university—on a voluntary basis. Before participating, individuals provided basic demographic information (age, gender, nationality), which was used exclusively for aggregated analysis without storing any personally identifying data.

Upon entering the survey, each participant was automatically assigned a personalized, randomized queue of all available animation clips. This randomization prevented videos depicting the same emotion from appearing consecutively, thereby minimizing biases such as repetition effects or memorization. Participants were asked to watch each animation and then select which of the six basic emotions (*happiness*, *sadness*, *anger*, *fear*, *surprise*, *disgust*) they believed it con-

veyed. They also rated the difficulty of making that decision on a seven-point Likert scale, ranging from *easy* (1) to *hard* (7). Thus, each selection (s_i) was recorded as (v_i, e_i^s, d_i) where v_i indicates the depicted video, e_i^s is the emotion selected by the participant, and d_i is the difficulty rating.

Fig. 2 illustrates the survey's interface. By combining emotion selection and difficulty assessment, we gathered not only recognition accuracy—how well participants identified the intended emotion—but also insights into each participant's subjective confidence or uncertainty in making that choice. This dual-metric approach allowed for a more nuanced evaluation of how effectively each generative model conveyed emotional expressivity in its animations.

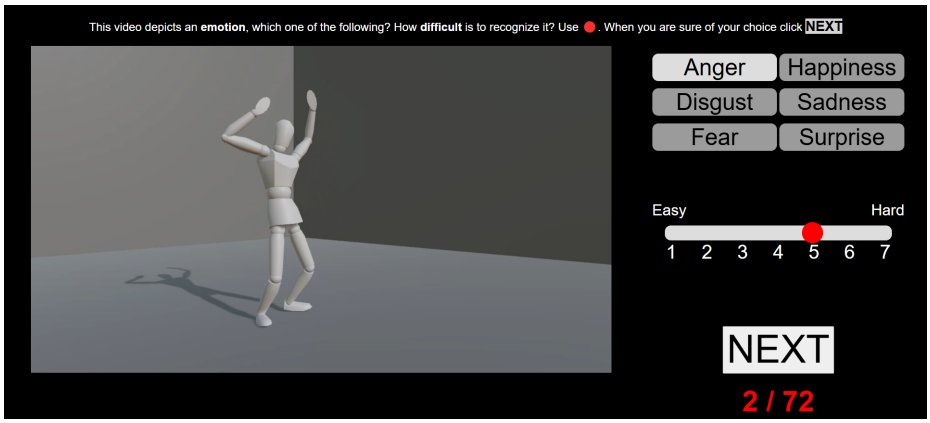


Fig. 2: Overview of the web-based survey interface.

4.1 Metrics

To evaluate performance at different levels of granularity, we extracted different subsets of all user selections to analyze specific factors (e.g., a single emotion, a model, an action, or any combination thereof). In this way, for example, metrics can be calculated only for the *anger* animations generated by a given model or for the *happiness* emotion on *walking* animations generated by all models.

Let S be an arbitrary subset of the user selections obtained by filtering for any combination of model (m), action (a), and depicted emotion (e).

We define the **accuracy** A_S as the number of correctly recognized emotions divided by the total number of selections in S, and the **difficulty** D_S as the normalized mean of the difficulty d_i of all selections $s_i \in S$, i.e. $D_S = (\mu - 1)/6$ where $\mu = \frac{1}{n}\sum_{i=1}^{n} d_i$. Therefore, an accuracy closer to 1 corresponds to *higher believability*, while an accuracy closer to 0 corresponds to *lower believability*. Conversely, higher difficulty values mean greater uncertainty in identifying the emotion depicted. Intuitively, we expect a negative correlation between selection

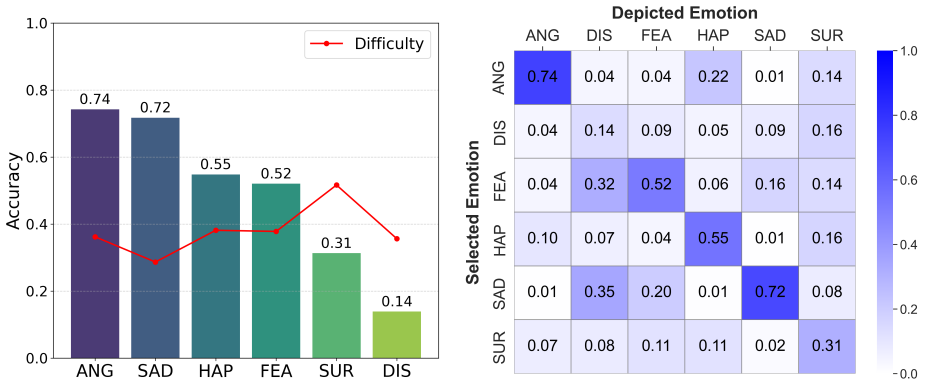
correctness and difficulty d_i (i.e., lower believability is associated with higher difficulty). Furthermore, selection time per video was collected to provide an objective metric regarding evaluation difficulty, expecting a positive correlation between voting time and difficulty d_i, and thus a negative one with selection correctness. Correlations are computed using Spearman's rank correlation coefficient, reported in the format (ρ, p), where ρ indicates the strength and direction of the monotonic relationship, and p the associated p-value, considering statistical significance for $p < 0.001$.

5 Results

The primary objective of this study was to examine the believability of emotionally expressive animations produced by various text-to-motion models. In particular, we assessed recognition accuracy across different emotions and actions. This analysis aimed to provide preliminary insights into the models' capabilities for realistically conveying emotional states, considering the known limitations identified in existing literature on body language perception.

Participants Our survey resulted in a total of 39 participants with 33 who fully completed the survey, the others' responses were discarded. Participants were nine females and 24 males, of which 31 of them are Italian with a mean age of 27 ± 7 years old. Overall, the data collection resulted in 2376 selections, with a mean selection time of 23.75 seconds for a single video.

5.1 Emotions analysis



(a) Accuracy (bars) and Difficulty (plot) (b) Confusion matrix

Fig. 3: Aggregate results grouped by *emotion*.

Considering overall metrics subdivied by emotions (Fig. 3a), *anger* and *sadness* achieved the highest recognition accuracies (0.74 and 0.72, respectively). The higher accuracy for *anger* can be attributed to the presence of distinct body movements typically associated with intense emotions—such as abrupt gestures, tense muscular postures, and forward-leaning stances—that are universally identifiable and consistently documented across multiple body language studies [2,4]. Similarly, *sadness* benefits from highly recognizable bodily expressions such as slumped shoulders, lowered heads, and minimal limb movements [20,19], providing clear visual cues even in the absence of facial details [2]. Conversely, *disgust* and *surprise* recorded significantly lower accuracy rates (0.14 and 0.31, respectively). The particularly low accuracy for *disgust* highlights its reliance on subtle facial expressions, including movements around the mouth and nose [8], making it difficult to convey effectively using body motion alone [19,4]. *Surprise*, despite generally involving noticeable bodily cues such as sudden backward movements or raised limbs, suffers from ambiguity in body posture interpretation without complementary facial or hand cues, leading to increased confusion and lower recognition accuracy [2]. Additionally, considering that usually *surprise* emotion is expressed toward a sudden occurring environmental event [8], the lack of context may have added more selection uncertainty [20,19]. *Happiness* achieved moderate accuracy (0.55), reflecting the complexity in interpreting this emotion solely through body language. *Happiness* is often expressed through dynamic and energetic gestures such as bouncing, rhythmic limb movements, or open and expansive postures [20,2]. However, these cues can closely resemble other high-arousal emotions, resulting in ambiguity and recognition challenges [2,19]. Similarly, *fear* showed moderate recognition accuracy (0.52), indicating variability and complexity in bodily expression. *Fear* can manifest in multiple distinct motor behaviors ranging from freezing and subtle defensive postures to overtly escaping movements, making it inherently context-dependent and less universally recognizable without facial or environmental context cues.

To gain insights on the incorrect selection of emotions, it is useful to consider their confusion matrix (Fig. 3b). *Disgust*, is frequently confused with *sadness* (0.35) and *fear* (0.32). Prior research has emphasized that, when facial and hand information is absent, observers tend to misclassify it with emotions exhibiting similarly contracted or protective postures, i.e., *sadness* or *fear* [20]. Similarly, *surprise* was commonly confused with all of the other emotions, aligning with past literature indicating that *surprise* is inherently ambiguous without complementary facial cues, as its bodily expression often overlaps with positive (e.g., *happiness*) or negative (e.g., *disgust*, *fear*) reactions due to sudden and expansive gestures [2]. *Fear*, while moderately well-recognized (0.52), was most commonly misclassified as *sadness* (0.20), hinting at common perceptual cues. This confusion is consistent with prior findings showing that *fear* related movements, in the absence of facial expressions or contextual stimuli, may resemble the inward and subdued posture of *sadness* [20,19]. Equivalently, *happiness* was misclassified as *anger* (0.22), revealing a perceptual overlap in their bodily expressions. Prior studies have shown that both emotions are conveyed through expansive, high-

energy movements—such as arm-raising or vigorous gestures—which again, in the absence of facial features, can appear visually similar [2,20,19]. This suggests that observers may rely on motion dynamics as a proxy for intensity or arousal, leading to systematic confusion between highly activated emotional states. Conversely, *anger* and *sadness* showed fewer confusions with other emotions, consistent with previous findings highlighting their universally recognized and distinct body cues, such as forward-leaning and tense postures for *anger*, and closed, slouched postures for *sadness* [2,19,4].

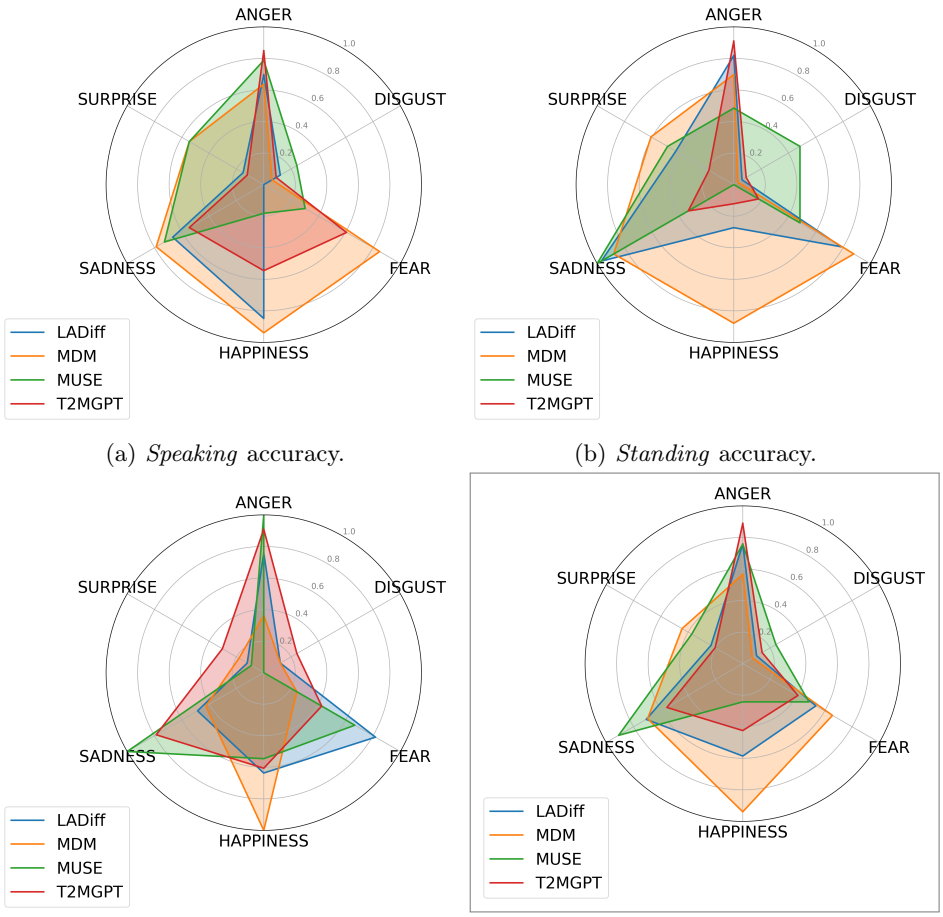
Across all emotions, difficulty ratings range from easy to neutral Fig. 3a. This outcome hints that even if the actual selected emotion was incorrect, participants perceived the produced animations as fairly coherent and natural. Although this distribution implies a negative relationship between recognition correctness (i.e., 0 if selected emotion is wrong, 1 if correct) and perceived difficulty, no strong correlation was found between the two metrics ($\rho = -0.15$, $p < 0.001$). Similar results are obtained considering correctness and emotion selection times ($\rho = -0.10$, $p < 0.001$). Contrarily, an interesting finding emerges when examining the relationship between selection time and perceived difficulty. A moderate and statistically significant monotonic relationship is observed between voting time and difficulty ratings ($\rho = 0.44$, $p < 0.001$). This suggests that participants who perceived a stimulus as harder also took longer to cast their selection, reflecting greater cognitive effort or uncertainty in emotion attribution.

In general, varying accuracies observed across emotions emphasize the importance of distinctive and universal non-verbal cues in accurately conveying emotional states in VHs animation as well as real human motion. Emotions characterized by well-defined, cross-culturally consistent gestures and postures, such as *anger* and *sadness*, are more successfully communicated through body movements, whereas emotions that rely predominantly on subtle, facial or context-dependent details, such as *disgust* and *surprise*, present inherent limitations for effective recognition without complementary expressive channels.

5.2 Actions and Models analysis

To better interpret the performance outcomes, it is useful to take a closer look at how different generative models behave across various actions, analyzing in particular possible models' architectural strengths and weaknesses while portraying different emotions.

Speaking. MDM demonstrates relatively higher accuracy across multiple emotions (Fig. 4a), especially for *happiness*, *fear*, *sadness*. While these results might partly stem from the universally expressive nature of these emotions [20,4], the model's transformer-based architecture, which is designed to manage the temporal and spatial irregularities of motion data, may also support slightly more coherent and believable emotional animation in speech scenarios. While this remains speculative in terms of expressive clarity, the model's use of geometric losses on joint locations and velocities [27] suggests a technical basis for improved motion continuity. Muse and LADiff follow, performing moderately well, in particular better than MDM on *anger*, potentially benefiting from



(a) *Speaking* accuracy.

(b) *Standing* accuracy.

(c) *Walking* accuracy.

(d) Overall accuracy.

Fig. 4: Aggregate accuracies for models and emotions. Subfigures (a), (b), and (c) by actions, (d) overall, calculated upon their respective selection subsets.

possible learned emotional embeddings for Muse, or diffusion-based smoothing for LADiff [23], which may stabilize emotional portrayals in specific Speaking poses. Interestingly, LADiff's accuracy regarding *fear* is close to zero due to its consistent confusion with *sadness* (ratio of 0.76). This confusion likely stems from the model's latent space denoising process, during which emotional nuances may be overly smoothed or collapsed, particularly for emotions with overlapping low-energy postures like fear and sadness. In contrast, this confusion is less evident in MDM, which does not rely on a latent-space diffusion mechanism and instead generates motion sequences directly in pose space, preserving more distinct structural features between similar low-activation emotions. On the other hand, the non-diffusion architecture of T2MGPT, relying on tokenized motion

representation [35], appears less consistent, possibly due to the discretization of motion into symbolic units that can limit fluidity and nuance. While this process enables structured prediction and scalability, it may constrain the model's ability to represent continuous and context-sensitive emotional cues, especially in subtle or complex gestures. On the contrary, this architectural characteristic might interestingly enhance T2MGPT at portraying *anger*, in which the required strong and abrupt movements might be naively portrayed by the intrinsic limited motion fluidity.

The portrayals of *surprise* and *disgust* remain particularly challenging across all models, which is consistent with broader findings in body language literature [2]. While the overall performance remains limited, MDM and Muse appear to handle basic emotions slightly better.

Standing. Again, MDM shows slightly higher accuracies, particularly for *fear*, *happiness*, and *sadness*. This might reflect the model's ability to capture subtle posture changes even in limited-motion contexts. Standing poses inherently offer fewer dynamic cues, making emotion conveyance more reliant on posture and subtle weight shifts [20,4], areas where MDM's design might offer an advantage. LADiff's performance only slightly lags behind MDM overall due to a noticeable dip in expressing happiness, which seems to stem from specific characteristics of the generated animations—mainly confused with *anger* with a 0.52 rate. Muse shows modest improvements in conveying ambiguous emotions like *surprise* and *disgust*. This may be due, similarly to Speaking action, to mechanisms that promote stylistic coherence or learned patterns from expressive training examples. These considerations remain tentative and should be interpreted cautiously as Muse is still completely undocumented. Lastly, T2MGPT is outperformed by other models except for *anger*, where it excels among others, possibly reflecting its capacity to tokenize the tense posture shifts associated with the emotion. This difference remains speculative and warrants further verification but highlights how even minimal shifts in stance may be expressed more distinctly by some architectures than others.

Walking. Results show a heterogeneous landscape, with each model highlighting a different emotional strength. Muse shows relative proficiency in conveying *anger* and *sadness*, possibly owing to naturally embedded emotional comprehension, which may help capture the slow, heavy gait characteristic of sadness-related motion. In the case of *anger*, the model may benefit from its ability to correctly reproduce increased joint rigidity and greater arm swing amplitude, with sharper elbow and shoulder angles—features often associated with high-arousal, confrontational body language [2]. As well, T2MGPT captures *anger* effectively, which might be due again to its token-based generative mechanism, enabling sharp and intense motion bursts that align with the high-activation, aggressive gait typically associated with anger in body language research [19]. Conversely, LADiff appears somewhat better at portraying *fear*, likely benefiting from its latent diffusion process that emphasizes smooth temporal coherence while still allowing for subtle retreating or defensive kinematic cues typical of fearful walking [2]. Meanwhile, MDM appears to excel at conveying *happiness*,

likely due to its diffusion-based generation process and geometric training losses, which may contribute to more coherent, upbeat, and rhythmic gestures characteristic of this emotion. Nonetheless, *surprise* and *disgust* continue to yield poor results across all models, highlighting the persistent difficulty of conveying these emotions through body movement alone [2,4].

Models. Overall, MDM and Muse perform slightly better than the others, though all models show room for improvement (Fig. 4d). MDM overall succeeds in conveying nuanced emotional cues across different actions, possibly due to its transformer-based architecture and use of geometric loss functions, which support more structured motion sequences. Muse shows a degree of emotional expressivity, perhaps due to training on motion-emotion semantics, which helps it capture some recognizable cues, especially in high-energy emotions like *anger*. LADiff and T2MGPT show limitations overall. LADiff struggles particularly with emotions like *fear* and *happiness*, sometimes confusing them probably due to its latent diffusion approach, which may blur subtle emotional differences during denoising. T2MGPT, utilizing tokenization-based generation, struggles with fluidity and subtlety in most emotions but interestingly excels at *anger*, where abrupt and sharp movements align well with motion tokenization without diffusion processing.

In conclusion, while all models exhibit considerable limitations, MDM and Muse offer better accuracies. Their relative advantages suggest that certain architectural choices and training approaches may provide a more promising direction for improving emotional motion generation, though much work remains to be done.

5.3 Limitations and Future Works

Due to the preliminary nature of this study, some important limitations should be noted.

Firstly, although we systematically conducted and evaluated an iterative prompt engineering method through a pre-screen user study to generate animation clips, this approach could still have introduced evaluation biases. Text-to-motion methods rely heavily on carefully crafted textual input to achieve desired motion outputs, a complex challenge partially addressed in text-to-image AI models [21] but still underexplored in the text-to-motion field. Therefore, future research should investigate various prompting strategies while potentially considering training models on specific emotion-labeled datasets, narrowing their focus for the ETMG task.

Secondly, while our analysis offers preliminary insights, it hypothesizes potential explanations for model behaviors by examining architectural characteristics and generated outputs in relation to emotional expressions. These interpretations, grounded in body language literature and model-specific structural details, remain speculative and should serve only as an initial step toward fully understanding the capabilities and limitations of these models.

Thirdly, as mentioned previously, cues such as finger gestures, facial expressions, and general visual appearance heavily influence emotional perception.

Future research should explore diverse character models and possibly consider adding hands and facial animations to better understand how these variables influence the emotional believability of generated animations. Furthermore, these animations should be evaluated directly in immersive XR, where embodiment, presence, and spatial perception may further influence their interpretation.

Finally, the vast majority of participants were Italian with ages spanning from 20 to 30 years old. Their evaluations could have been affected because of possible cultural differences, referring to nationality, and expectations regarding computer-generated animation videos, especially referring to age. Future research should try to diversify its participants' demography in order to reduce any potential bias.

6 Conclusions

This study explores how effectively state-of-the-art text-to-motion models can express basic emotions through humanoid full-body animations. Using four generative models (MDM, LADiff, T2MGPT, and Unity's Muse Animate), we created animations for six basic emotions and three common actions, then assessed their emotional believability through a user evaluation survey. The goal was to analyze the capabilities of AI-generated body movement alone (excluding facial expressions or hand gestures) to believably communicate emotion. The resulting varying perception of emotions underscores the central role of distinctive and culturally consistent body language in conveying emotional states regarding both real human motion and AI-driven text-to-motion animations. Our results align with findings from body language research, revealing interesting insights on text-to-motion emotional believability. Emotions such as anger and sadness, which involve well-defined and broadly recognized movement patterns, are more effectively conveyed by current models. In contrast, emotions like disgust, surprise, and to some extent happiness and fear, depend more on subtle or facial cues, making them harder to express through body motion alone. It is worth noting that our interpretations about models' architectures and generated animations' emotion believability are speculative and serve as an initial step toward understanding text-to-motion models' emotional expressive capabilities.

While these challenges highlight current limitations in text-to-motion synthesis, the observed trends suggest its potential for generating emotionally expressive virtual humans. The framework proposed in this work provides a systematic approach for both developing and evaluating AI-generated emotional animations. As such, it supports further investigation into the expressive capabilities of emerging models and human perception of specific emotions, ultimately contributing to more nuanced and believable virtual humans' behavior.

References

1. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents (2023)

2. Atkinson, A.P., Dittrich, W.H., et al.: Emotion perception from dynamic and static body expressions in point-light and full-light displays. Perception **33**(6), 717–746 (2004), pMID: 15330366
3. Bhattacharya, U., Childs, E., Rewkowski, N., Manocha, D.: Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In: Proceedings of the 29th ACM International Conference on Multimedia. MM '21, Association for Computing Machinery, New York, NY, USA (2021)
4. Blythe, E., Garrido, L., Longo, M.R.: Emotion is perceived accurately from isolated body parts, especially hands. Cognition **230**, 105260 (2023)
5. Calzolari, S., Strada, F., Bottino, A.: The quest for believability: Exploring facs adaptations for emotion facial expressions in virtual humans. In: 2024 IEEE Gaming, Entertainment, and Media Conference (GEM). pp. 1–6. IEEE (2024)
6. Chen, X., Jiang, B., et al.: Executing your commands via motion diffusion in latent space (2023)
7. Depraz, N.: Chapter 2. Shock, twofold dynamics, cascade: Three signatures of surprise. The micro-time of the surprised body, pp. 23–42. John Benjamins Publishing Company (2019)
8. Ekman, P., Friesen, W.V.: Facial action coding system. Environmental Psychology & Nonverbal Behavior (1978)
9. Ghorbani, S., Ferstl, Y., Holden, D., Troje, N.F., Carbonneau, M.A.: Zeroeggs: Zero-shot example-based gesture generation from speech. Computer Graphics Forum **42**(1), 206–216 (2023)
10. Girondini, M., Frigione, I., et al.: Decoupling the role of verbal and non-verbal audience behavior on public speaking anxiety in virtual reality using behavioral and psychological measures. Frontiers in Virtual Reality **5**, 1347102 (2024)
11. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions (2023)
12. Johansson, T., Eladhari, M.P.: Emotional believability of non-playable game characters-animations of anger, sadness and happiness. In: International Conference on Interactive Digital Storytelling. pp. 72–99. Springer (2024)
13. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023)
14. Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing (2023)
15. Kyriakou, M., Pan, X., Chrysanthou, Y.: Interaction with virtual crowd in immersive and semi-immersive virtual reality systems. Computer Animation and Virtual Worlds **28**(5), e1729 (2017)
16. Liu, T., Liu, Z., et al.: 3d visual simulation of individual and crowd behavior in earthquake evacuation. Simulation **95**(1), 65–81 (2019)
17. Loveys, K., Antoni, M., et al.: Comparing the feasibility and acceptability of a virtual human, teletherapy, and an e-manual in delivering a stress management intervention to distressed adult women: pilot study. JMIR formative research **7**, e42390 (2023)
18. Lugrin, B., Pelachaud, C., Traum, D.: The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 2: interactivity, platforms, application. ACM (2022)
19. Martinez, L., Falvello, V.B., Aviezer, H., Todorov, A.: Contributions of facial expressions and body language to the rapid perception of dynamic emotions. Cognition and Emotion **30**(5), 939–952 (2016)

20. Normoyle, A., Liu, F., et al.: The effect of posture and dynamics on the perception of emotion. In: Proceedings of the ACM symposium on applied perception. pp. 91–98 (2013)
21. Oppenlaender, J., Linder, R., Silvennoinen, J.: Prompting ai art: An investigation into the creative skill of prompt engineering. International journal of human–computer interaction pp. 1–23 (2024)
22. Reed, C.L., Moody, E.J., et al.: Body matters in emotion: Restricted body movement and posture affect expression and recognition of status-related emotions. Frontiers in Psychology **11** (2020)
23. Sampieri, A., Palma, A., Spinelli, I., Galasso, F.: Length-aware motion synthesis via latent diffusion (2024)
24. Schouten, D.G., Deneka, A.A., et al.: An embodied conversational agent coach to support societal participation learning by low-literate users. Universal access in the information society **22**(4), 1215–1241 (2023)
25. Seaborn, K., Miyake, N.P., Pennefather, P., Otake-Matsuura, M.: Voice in human–agent interaction: A survey. ACM Computing Surveys (CSUR) **54**(4), 1–43 (2021)
26. Song, W., Zhang, X., et al.: Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 811–820 (2024)
27. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023)
28. Unity Technologies: Muse animate (2024), accessed: 2025-02-06
29. Witkower, Z., Tracy, J.L.: Bodily communication of emotion: Evidence for extrafacial behavioral expressions and available coding systems. Emotion Review **11**(2), 184–193 (2019)
30. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation (2024)
31. Xu, S., Fang, J., et al.: Emotion recognition from gait analyses: Current research and future directions (2022)
32. Yazdian, P.J., Liu, E., et al.: Motionscript: Natural language descriptions for expressive 3d human motions. arXiv preprint arXiv:2312.12634 (2023)
33. Yu, T., Wang, J., et al.: Towards emotion-enriched text-to-motion generation via llm-guided limb-level emotion manipulating. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 612–621. MM '24, ACM (Oct 2024)
34. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
35. Zhang, J., Zhang, Y., et al.: T2m-gpt: Generating human motion from textual descriptions with discrete representations (2023)
36. Zhang, M., Cai, Z., et al.: Motiondiffuse: Text-driven human motion generation with diffusion model (2022)
37. Zhang, M., Guo, X., et al.: Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116 (2023)
38. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism (2023)
39. Zhong, L., Xie, Y., Jampani, V., Sun, D., Jiang, H.: Smoodi: Stylized motion diffusion model (2024)
40. Zhou, W., Dou, Z., et al.: Emdm: Efficient motion diffusion model for fast and high-quality motion generation (2024)