Craft Knowledge and Embodied Labelling in ML-Assisted Motion Capture

Theodore Koterwas
University of Edinburgh
Edinburgh, United Kingdom
tkoterwa@ed.ac.uk

Jack Morrow
RETìníZE
Belfast United Kingdom
jack@retinize.com

Stefano Braghetto
RETìníZE
Belfast United Kingdom
stefano@retinize.com

Phillip Vaughan
Abertay University
Dundee, United Kingdom
p.vaughan@abertay.ac.uk

The StyleCap project developed a Machine Learning (ML) model for styling full body animation from sparse motion capture and text. The development contributes an approach to Artificial Intelligence (AI) in HCI research prioritising subjective expertise and embodied knowledge. Interviews with animators documented their values, expertise and use of language in animation. The interviews included a live motion capture session and were coded to produce a database of words describing style. Inverting the traditional process of using language to label data, data was created by expressing language through bodily movement. An ML model was trained on this dataset to produce stylised animations which were evaluated by professional animators. These evaluations were then used to measure the effectiveness of GPT 4.1 as a translator of the evaluators' language to the specific labels utilised in training. The accuracy of these translations of the craft knowledge and vocabulary used by individual practitioners and enacted through embodied performance demonstrates the potential of human-centred approaches to HCI research with AI.

Machine Learning. Animation. Language. Embodied Knowledge.

1. INTRODUCTION

Professional character animation for film, television and games is a laborious process and requires a range of expensive technologies. Motion capture with live actors is regularly used as a starting point, but the data generated requires tedious cleaning and painstaking editing to stylise it into expressive movement. Reducing the effort required could enable smaller teams to achieve higher quality results and compete with larger production houses. It has the potential to increase creative freedom by reducing the cost of experimentation, and if the process could be closer to real-time, then animators would have more direct, responsive control.

Funded through the University of York as an Embedded R&D subaward of the "XR Network+" grant from the Engineering and Physical Sciences Research Council (EPSRC) researchers from the University of Edinburgh, Abertay University, and industrial partner RETìníZE investigated the use of supervised Machine Learning (ML) to train a model for stylising motion capture animation in real-time with direction such as "move in a more exaggerated way" or "imagine you're scared" or "happy" or "proud".

The project was undertaken against a backdrop of increasing concern among creative professionals over the use of Artificial Intelligence in creative work. Even with the good intentions to increase access and creativity by reducing the dull, labour-intensive processes of animation, the pool of potential participants was small due to widespread AI resistance. Sharing these concerns as creative practitioners, the project team crafted a methodology that places the subjective expertise and values of animators at the centre of the process.

Animators were interviewed about what they valued in animation, its practice, and the tools they use. They were asked to direct a motion capture session and describe improvements to the results. Transcripts were analysed for frequency of words and common themes, and words were categorised according to the broader body of expert knowledge. This bottom-up approach, starting from the individual, deliberately contrasts with the top-down approach of scraping billions of texts to construct statistical consensus in a Large Language Model (LLM). Then, rather than adopting the standard approach to painstakinglylabelling a large corpus of existing data, new data was generated through motion capture of actors using their bodies to interpret labels drawn from words used frequently by the interviewees. After an ML model was trained to produce animations wth this dataset, a group of animation practitioners evaluated the animations. Their evaluations formed the basis for determining the feasibility of using an LLM for translating their

© Koterwas et al. Published by BCS Learning and Development Ltd.

rich knowledge into the specific labels used in training. After detailing this method we discuss the results and broader implications of this approach for responsible HCI research with AI.

2. BACKGROUND

2.1 Industry need for more efficient styling of motion capture-based animation

3D character animation is expensive, representing 20-50% of a film's budget, with studios charging $5000-$25,000 per minute (*Animation Market Size, Statistics Report, Industry Growth, 2032*, 2024; *Estimating Your Animation Budget: Hourly vs. Flat Rate Projects*, 2025). With long workflows and quotas, it is hard to experiment and iterate in this space. For television production, typical quotas for animators range between 5 and 10 seconds per day, and for feature films they are 3-5 seconds per week. (*How Slow Should You Go? | Into Animation*, 2023) Utilising motion capture has the potential to address this problem, but motion capture data isn't as expressive as keyframed animation. Painstaking manual work is required to clean and style the motion capture data. Based on its own pilot projects for the BBC and Moonbug Entertainment, RETìníZE estimates that making this process more efficient would significantly reduce costs and improve focus on creative elements of production. It could enable junior animators to spend less time on laborious keyframing and more on developing their expressive skills. It would likewise allow individual creators and small studios to compete in a traditionally exclusive market where large studios and large budgets dominate. The global animation market was estimated to be worth $436 Billion in 2024 and is projected to grow by 7.46% annually to reach $896 by 2034, with 3D animation making up $193 Billion, or 44.27% of the market in 2024 (*Animation Market Size to Hit over USD 895.71 Bn by 2034*, 2025). Making the process of 3D animation more natural and less costly removes barriers to entry into this market.

As a step toward these goals, RETìníZE has developed the Animotive platform to make 3D animation feel more like a live action set. In a VR headset animators embody characters and interact with props to give more natural, direct control over performances than keyframe-based tools. Adding style in real time through text-based prompts would extend expressive options, allowing for more experimentation and less work manually stylising the captured data later. Understanding the language animators use in practice is critical to successfully supporting their work.

2.2 Language and animation style

Disney's 12 Principles of Animation, first presented in 1981 by Frank Thomas and Ollie Johnston in their book *The Illusion of Life: Disney Animation*, remain foundational to the way animators speak about style (Thesen, 2020). The principles are encapsulated in easy to remember words and phrases: "squash and stretch", "exaggeration" and "anticipation" are shorthand for characteristics of animated movement that add expressivity to characters (Thomas & Johnston, 1981). A word like "stretch" is easy to visualise, but some of the words, such as "appeal" are more abstract, with the potential to describe a wide variety of movement. For creating a lexicon useful in the development of an ML model, the more concrete words from the 12 Principles appear promising. However, while these words featured frequently in informal discussions and formal interviews with animators, animators also use more specific language in practice.

Specific formalisations of language have been theorised and implemented in the design of HCI systems and robotics. In *A Human Movement Language for Computer Animation*, Herbson-Evans (1980) reviews seven methods he categorises as "ideographic notation, animation chart, specialised dance languages, natural language, numerical description, demonstration and muscle tension." Many of these are too mechanical, functional or reductive to capture the richness of style. Conversational gestures offer an avenue for exploration; however, attempts at developing ML models to generate meaningful gestures have demonstrated the difficulty in translating mechanical movement into language (Alexanderson et al., 2020; Ali et al., 2020; Yoon et al., 2022). This is primarily due to the lack of clear correspondence between specific movements and meanings, except in the case of specialised gestures such as Sign Language. Understanding the situated nature of bodily communication, animators have turned towards principles for live acting employing empathy, negotiation and objective seeking to create more expressive animation (Hooks, 2017; Kennedy, 2021; Sultana et al., 2013). Guides for directing acting for motion capture provide practical insight, drawing upon the art of directing live actors more generally to highlight the importance of motivation, story and social interactions. (Bancroft & Minkoff, 2019; Kines, 1998; Ronfard, 2021; Weston, 2021). This turn toward an actors' ability to embody a character points to the richness of embodied knowledge and away from attempts to over-rationalise human movement.

2.3 Embodied knowledge

As computational models of cognition have developed so have theories emphasising the role the body plays in shaping human experience and thought (Chemero, 2009; Clark, 2009; Pfeifer et al., 2007). In addition to theories of embodied cognition,

empirical studies underscore the centrality of bodily sensation to our experience; for example, mapping abstract concepts and emotions to sensations within the body (Daikoku et al., 2024; Nummenmaa et al., 2018), and demonstrating how they affect decision-making (Dunn et al., 2010). These considerations take on added significance in the critical discourse around disembodied "neural" networks succeeding at several tasks, particularly Large Language Models (LLM) mimicking language (Bender et al., 2021; Vaswani et al., 2017). Improving the capabilities of these models indiscriminately through increasing the size of training data and neural network connections risks reducing their relevance to human experience, including false, problematic and harmful information, and promoting meaning-making where there is none (Bender et al., 2021). The algorithmic processing of words reconstructs semantic relationships within a training corpus, but meaning also exists in the space between words, not just their arrangement, and these spaces are filled in from lived experience (Bisk et al., 2020; Merleau-Ponty, 1978). Meaning relies on the perception of agency and intention behind language, which are missing in an LLM, despite the layers of fine-tuning and reinforcement learning designed to make it seem human-like (Barandiaran & Almendros, 2025). Embodying AI systems in avatars and robots impacts on trust, empathy and rapport-building, producing a complex interplay between anthropomorphism and the acknowledgement of artificiality (Gratch et al., 2013; Pecune & Marsella, 2020; Schlesener et al., 2025; Zhao et al., 2014). This interplay is also at the heart of virtual character animation. Generating meaningful movement with Machine Learning requires extending language processing to include embodied enactment of language.

2.4 AI for generating stylised animation

Förger and Takala (2016) turned to enactment as a source for synthesising style in animation, but through a highly reductive approach. They instructed an actor to combine basic styles like "'sad+fast' and 'sad+aggressive' in addition to plain 'sad'" (Förger & Takala, 2016, p.195), and then computationally attempted to isolate the features essential to each style through a subtractive comparison of the differences. This required the actor to "consciously analyze different aspects of individual styles and devise a way to combine them" (Förger & Takala, 2016, p.196) rather than drawing more holistically on intuitive movement and experience. While the accuracy of this approach in producing recognisable basic styles was statistically validated through crowdsourcing, the expressiveness of the demonstrations accompanying the paper is rudimentary. As an alternative method, Reinforcement Learning (RL) holds promise as a method to imitate behaviour. However, RL models learn by being rewarded when moving in ways similar to an example, executing sequences of movements rather than learning underlying expressive qualities (Merel et al., 2017; Peng et al., 2018, 2022; Wang et al., 2017). While such automated mimicry might come visually close, it misses the mark in emotional expressivity.

There are many other attempts to use AI in simulating human motion (Aberman et al., 2020; Karg et al., 2013; Mason et al., 2018; Mourot et al., 2022; Zhu et al., 2024). The overarching approach is to collect a large number of samples, make generalisable decisions about what is important to quantify, and then apply an algorithm to train a model on these features. Rigorously reducing something so richly qualitative and ambiguous as bodily expression into quantitative material for training algorithms necessitates loss and may be impossible (Birhane, 2021). When we attempt to model it, we should acknowledge this limitation and align the values guiding our method with those of the people we involve and impact.

2.5 Animators' concerns over AI

The theories and principles of animation discussed so far were created by animators, as were some of the approaches to generate style with algorithms. Attempting to understand one's craft through rational systemisation is natural, but with the proliferation of generative AI tools there is a concern among creative practitioners about agency and experience being devalued (*AI and Hollywood: 5 Questions for SAG-AFTRA's Chief Negotiator | World Economic Forum*, 2024; *How Labor Strikes and AI Are Shaping the Future of Entertainment*, 2024; *Statement on AI Training*, n.d.). Animation for film, television, and gaming is predominantly digital, so animators are particularly worried. The promises of boosted efficiency, amplified visual capabilities, and democratization are offset by the potential for job displacement, loss of creative control and homogenisation, alongside a number of other ethical considerations (Izani et al., 2024). The antipathy towards AI within our own network was such that we struggled to identify animators to approach for interview. The current negative temperature in the industry underscores how vital it is to develop an approach to designing with AI that places their skills, knowledge and values at the centre of the process.

3. A HUMAN-CENTRED APPROACH TO DESIGNING HUMAN-AI INTERACTIONS.

Our overall objective was to investigate the use of Machine Learning and natural language to style full body animations. This was broken into the following:

- To better understand animators' perceptions of animation style, how they

currently attempt to create it, and what natural language descriptors they use.

- To determine what supervised learning methods and data labelling strategies are effective in producing adjustments to motion capture data that convey qualitative characteristics.
- To evaluate the results of these methods by how they are perceived by animators and how they might be improved.
- To discover how these models might be implemented in a system for real time animation to increase experimentation and produce a more satisfying creative experience

Approaching these objectives in a way that centres on subjective expertise and embodied knowledge involved a multistage approach to the ML pipeline of dataset generation, training, and evaluation:

(i) Generate a vocabulary based on subjective expertise: conduct semi-structured interviews to identify words animators use in their animation practice and in expressing their values, creating a rich database from which to select stylistic labels
(ii) Generate a labelled motion capture dataset from embodied knowledge: capture actors interpreting the stylistic labels created in the previous step
(iii) Train an ML model on the labelled dataset
(iv) Evaluate the animations generated by the trained model with subjective expertise: ask animation professionals to describe the generated animations and how they would improve them in their own words
(v) Evaluate the feasibility of using subjective expertise to generate stylised animations with the model: investigate the accuracy with which an LLM translates the words used by the animation professionals into the labels used by the model in generating the animations they evaluated

3.1 Interviews to better understand animators' values and language

RETìníZE was founded by experienced creative professionals with first-hand knowledge of animation and its practice. The academic researchers are also creative practitioners. This combined experience guided the research at a high level, but to investigate in detail, a three-part semi-structured interview was carried out with four animators. While this sample size was small due to challenges previously discussed, the interviews provided a depth of subjective, situated information to augment the broader theory in the literature and check assumptions about what animators value, how they use language in practice, and their views of tools and automation.

3.1.1 Interview Part 1: Subjective values regarding animation quality

The following questions probed how animators use language to express their subjective values and practice:

> What makes an animation "good"? *Follow up questions: What specific things do you look for? Can you describe an example you particularly like? What things stand out to you in bad animation?*
>
> Are there particular styles of animation you find especially effective? *Follow up questions: What makes them effective? How would you go about emulating them?*
>
> Can you describe some of the techniques you use often? *Follow up questions: Do you have any rules of thumb? If you have to work quickly, what do you prioritise? How would you describe your own individual style?* (Interview Script)

The four animators referenced the 12 Principles of Animation, with "exaggeration", "squash and stretch", and "anticipation" being most frequently used, but they emphasised the need for animation to be "expressive" by whatever means achieved it. For one participant good animation depended on the intention and "vibe" of the storytelling. To describe a current project, they used the 12 Principles word "exaggerated", but more specifically "cheeky", "irreverent", and "colourful." When asked how these might be achieved, they responded:

> Depends on the character. If they can have a certain sass about their face, or even their walk cycle could convey a lot. Or they could just flick someone the bird, you know, or just have a sort of dismissive sort of wave of the hand. (Participant 1)

Another animator valued animation that was "pleasing," and not "janky" or "unclean." For the third animator it was the quality of the key poses the character moved between they valued most, and the fourth animator underscored the importance of planning, storyboarding and staging to "get a good feel for the three-dimensional space" (Participant 4). While the 12 Principles provided a shared vocabulary to describe animation in the abstract, the language they used to describe their subjective values and practice was more varied.

3.1.2 Interview Part 2: Language used to direct animation and describe improvements

For the next part of the interview, interviewees were asked to direct a motion capture session to record their choice of language when describing movement in the moment. They were provided a scenario and asked to instruct an actor in a motion capture studio

to perform multiple takes. To simulate the practice of styling animation after capture, the interviewee was asked to choose one of the takes and describe improvements they would make as they watched it back. The scenario, drawn from a scene in the video game *The Last of Us,* was as follows:

> An exhausted zombie apocalypse survivor is exploring a ruined pet shop looking for medication to save her injured friend. She hears a sound and then throws a can at a window to create a diversion. The glass shatters as she ducks behind an empty display stand. She hears the sounds of a zombie approaching and draws her rifle. The zombie enters. The survivor moves from behind the display stand and shoots the zombie. She hides again. (Interview script)

An example from one of the interviewees demonstrates the immediacy of their language and the use of evocative verbs like “stumble” and “crawl” and concrete adverbs “quickly” in addition to abstract adjectives “dynamic”:

> …instead of just walking over, I kind of want you to kind of stumble on your knees and crawl behind the display. Just to give it more of a dynamic scene. So here, throw the can and kind of stop, like, quickly stop. But yeah, that’s … like that. Yes. Yes. (Participant 2)

When asked what improvements they would make watching it back, they used vocabulary from the 12 Principles, along with more specific language describing the mechanics of movement:

> I would more like, have more, like pulling back and as it throws I would like more stretch. So like if it's like…as an animator we would stretch his arm further. Before coming back...mm hmm to make it like that illusion and even the timing when you're throwing, I would probably have that a lot quicker. (Participant 2)

When directing animation, another interviewee conveyed the character’s emotional state and motivation straight away:

> You're tired. Before we do anything, right? The motivation for you is you're tired and you're fed up. You're hungry as well. So you're looking for…I know it's a pet shop, but you're, you're looking for anything that you could use to, to eat, survive. (Participant 1)

During playback they suggested increasing the scale and energy of the movement:

> Pump that up to 11, a little bit like running away. That's…That's quite nice. Well, perhaps exaggerate the arms and the legs just to, to give it a bit more, to give the silhouette a bit more energy. (Participant 1)

A third interviewee described an emotional state to set the scene during directing but took into consideration how a viewer might distinguish between characters:

> You're quite tense, like you have an injured friend. Maybe you're hurt a little bit as well. I wouldn't say you're hurt on your feet because if you...if you walk weird, you might look more like the zombie, so maybe you're hurt a little bit on your shoulder. (Participant 4)

They continued to describe how situational factors should impact the movement during playback:

> ...if there's like really the necessity to come back with medicine quickly, I think to like kick open the door a little bit, could be really interesting to just make and feel it more like, make it feel more rushed, but I like the take. (Participant 4)

The language used when directing animation focused on specific qualities of motion, emotional state and motivation, and other situational aspects of the scene. When describing improvements during playback, the animators used the language of the 12 Principles as a guide to mechanical improvements but with more specific language while also continuing to reference the character’s situation.

3.1.3 Interview Part 3: Tools and automation

The final part of the interview was intended to check assumptions about how AI might be deployed to aid animators, and to better understand how they derived agency and enjoyment from the craft:

> What do you enjoy about the process of animation? *Follow up questions: What’s the “essence” of the craft for you? What things do you feel particularly good at?*
>
> Are there parts of the process you dislike? *Follow up questions: Why do you dislike them? Would you rather they be done for you?*
>
> What tools do you use in your work? *Follow up questions: How do they affect the quality of your work? Would you use these tools if you weren’t under time pressure? Can you describe a particular improvement you’d like to see in one of your tools?*
>
> If you were to design a tool to make you a better animator, how would it work? (Interview script)

The four animators described enjoyment in defining the overall stylistic direction, flow and timing of animation. Three specifically mentioned defining “breakdowns” or key poses as essential to the craft and would want automated tools just for the tedious work of “tweening,” or filling the gaps. For one animator, it was also tedious to think systematically through the details of a single action according to the 12 Principles:

> It's more the tedious part is whenever you have to think, right, I need to go faster and then slow. I

> need to go in an arc. I need to go before and an after. That's tedious: parts where you have to think. I think so many different things just to put out one little punch. So that's what I don't like about animation, is just trying to combine all the smaller details just to make one animation look good. (Participant 2)

The 12 Principles were referenced by another interviewee when describing tools, alongside the importance of understanding the emotional motivation of a character: “Key for that kind of tool [is] just to understand things from an animation principle, but then also to understand the emotional reasoning behind a movement” (Participant 1). Two of the interviewees emphasised the need to be able to fine tune animations further after any automated process. All four interviewees referred to the need to stay in control, but also the desire to experiment: “And then be able to tweak it and play around, maybe with what works, to give it …see how it goes and see how much else you need to add” (Participant 3). The first interviewee was particularly eloquent in describing the essence of animation as a process of agency and creative discovery:

> Going back to the very start, it's bringing things to life really. Whether it be your own creation, that's always better, when you see your own brought to life. But it is really the process of breathing life into something and discovering what its character is. (Participant 1)

The interviews were transcribed and analysed to develop a database of words. Phrases describing animation style or character movement were highlighted and categorised according to the quality of movement they described, for example: “scale of motion”, “expressiveness”, “pacing” et al. Words within the highlighted phrases were sorted by frequency to determine the words used most. To evaluate the broader relevance of these subjectively generated words and to connect them to a shared conceptual framework, they were further annotated according to the 12 Principles of Animation.

Principle	Category	Word	Count
exaggeration	scale of motion	**exaggerated**	16
squash and stretch	scale of motion	**stretch**	11
squash and stretch	scale of motion	**squash**	10
	expressiveness	**breathing**	5
	expressiveness	**frantic**	4
	expressiveness	**kick**	4
	scale of motion	**little**	4
appeal	appeal	**pleasing**	4
	expressiveness	**quiet**	4
timing	pacing	**slowly**	4
anticipation	pacing	**anticipate**	3

Figure 1: *Analysis of most frequent words used by animators in interviews to describe animation*

3.2 Motion capture session to generate a dataset from embodied knowledge

To generate a labelled dataset from the words used by interviewees an industry standard Vicon motion capture system was utilised to capture actors performing the words. These words then became the labels for the motion capture data collected. Three experienced student actors participated with informed consent. The objective was to capture the actors performing the same actions in the style of each of the labels. Walking, running and sidestepping were chosen as common to animation in film, television and gaming. Three criteria determined which of the most frequent words to perform: whether the word might describe a “style” of walking, running or sidestepping, whether it was relevant beyond Zombie-type scenarios, and whether the range of selected words represented a variety of the quality of movement categories and the principles of animation. To ensure the words were expressible from embodied knowledge, the session started with a trial of the top 20 to choose the ones the actors and researchers felt led to the most expressive performances.

	A	B	C	D
1	**Animation Principle**	**Label Category**	**Label**	**Filename_P1**
2	exaggeration	quality of motion	**exaggerated**	exaggerated_001
3	squash and stretch	quality of motion	**stretch**	stretch_001
4	squash and stretch	quality of motion	**squash**	squash_001
5		expressiveness	**thinking - contemplating**	
6	appeal	expressiveness	**attitude, confidence**	attitude_001
7		expressiveness	**shaky-traumatized**	shaky_001
8		expressiveness	**tense**	
9		emotion	**terrified**	
10		emotion	**joyful?**	
11		expressiveness	**exhausted**	**exhaustion_001**
12		quality of motion	**weight/ heavy**	**heavy_001**
13		quality of motion	**weight/light ?**	**light_001**
14		type of walk	**hiding/ creeping**	**stealthy_001**
15		type of walk	**stumble**	**stumble_001**
16		type of walk	**limp**	
17	anticipation	quality of motion	**anticipation?**	
18		type of walk	**shuffling**	
19		emotion	**miserable?**	
20		quality of motion	**energetic**	
21		expressiveness	**quiet**	
22		emotion	**frantic**	
23	timing	pacing	**pause**	
24	timing	pacing	**quick**	
25		pacing, scale of morion	**dynamic**	
26	appeal	expressiveness	**confident**	
27		expressiveness	**crouch**	

Figure 2: *Motion capture label selection*

After trialling the labels, the actors performed the five actions, walking and running forward, walking and running backward, and sidestepping for two minutes in the style of each label. They found the duration tiring, so after the first label, the time was reduced from two minutes to one and running was removed from the set of actions. The performance of nine labels was captured by the end of the session, with “exhaustion” as the final label to enact, ensuring a level of authenticity. Each take was saved out to a file named after the label, thus “pre-labelling” the dataset.

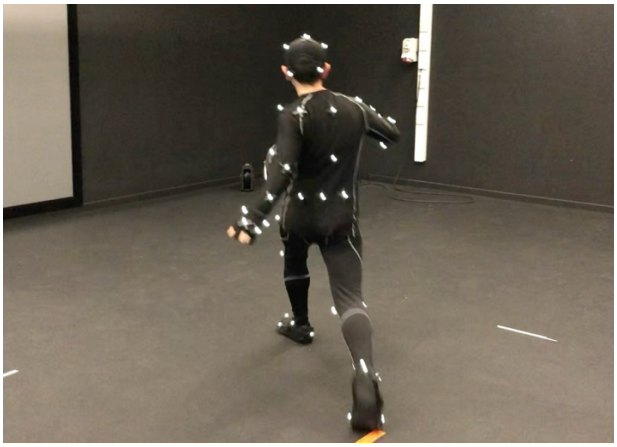


Figure 3: *Actor performing the style label "stretch"*

After the dataset was created, three labels were chosen for training the model on the basis that their motion capture data didn't require extensive cleaning and they were expressive and different: "attitude", "stealthy", and "stumble". While the definitions of these words were deliberately left to the actors, "attitude" was broadly understood as projecting confidence, pride or swagger; "stealthy" as attempting to go unnoticed out of fear or trying to sneak; and "stumble" as lacking coordination.

3.3 Training the model to include style

To achieve full body motion capture from sparse data (three points: VR headset and two handheld controllers) RETìníZE implemented Categorical Codebook Matching (Starke et al., 2024). This approach produces two "codebooks", one from sparse data and one from full body data, and by reducing the probability distributions between them, can predict full body motion when presented with new sparse data. This approach was modified by applying the style label to the training features of the full body dataset and then supplying a label as part of the sparse input. After training, three animations were generated from the model, one for each label.

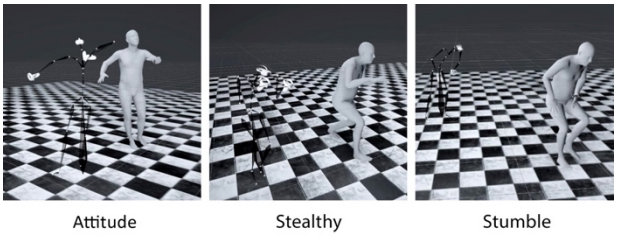


Figure 4: *Animations generated from the model*

3.4 Evaluating the model with subjective expertise

To evaluate the model, a group (N=13) of animation and film professionals viewed each animation and answered the following questions:

- What is the character doing?
- Describe the style of the character's movement in one word.
- Describe the style of the character's movement in one sentence.
- What three words best describe what the character might be feeling?
- How might the expressiveness of this character's movement be improved?

The survey responses indicated that the animations were expressive in ways relevant to the intended style label:

- For the "attitude" animation, 5 out of 13 respondents used words that were synonymous such as "strut", "proud", "prancing" or "swagger", while 7 referred to "dancing", and 12 referred to positive emotions such as "joyful."
- For "stumble" 4 used "stumbling" with all 9 other responses employing synonymous or closely related words like "tripping", "unstable," "chaotic," "drunk," "dodging", or "disoriented."
- For "stealthy" all respondents used a derivation of the words "sneak", "creep", "stalk" or "secretive."

3.5 Exploring the feasibility of an LLM to translate subjective expertise

The anticipated approach to implementing the model as a software feature is to utilise an LLM to translate natural speech to the style labels used to train the model. To evaluate the suitability of the model for this purpose, the University of Edinburgh's implementation of GPT 4.1 API "ELM" (University of Edinburgh, n.d.) was used to carry out three investigations:

3.5.1 Prompting GPT with the survey questions to evaluate the videos

For comparison with the language used by the human evaluators, prompts were created from the survey questions and provided with the videos to GPT 4.1. The results of repeating the prompts 10 times for each video were mixed and unsurprisingly formulaic and verbose, but they were generally relevant, and in some cases similar in content to the responses given by the professional animators. For the "attitude" video GPT 4.1 used the words "expressive", "flamboyant", "dramatic" and "theatrical," and mentioned "dance or "dancing" all ten times. 7 of 13 animators used that word. For the "stumble" video it used the words "cautious", "uncertain", "anxious" and "reactive," but did not use language more descriptive of stumbling or moving clumsily, whereas all 13 animators used such language. For "stealthy" the results from GPT 4.1 were much closer to the animators and more

accurate overall. All ten responses used the word "sneaky" or "sneaking" along with words like "creeping", "stealthy", and "cautious."

3.5.2 Prompting GPT to choose labels
To assess whether an LLM could identify the labels used by the model to produce the animations, GPT 4.1 was prompted to select which of the three training labels best described each animation video. For the "stealthy" animation, GPT chose "stealthy" 10 out of 10 times. For "stumble," it chose "stumble" 8 out of 10 times. For "attitude," it chose "attitude" 10 out of 10 times.

3.5.3 Prompting GPT to summarise survey responses
To investigate whether an LLM could translate the subjective language used by professional animators into the specific labels used by the model, GPT 4.1 was prompted to summarise the professional responses from our survey and to translate them into one of the three labels. Each of the translations were scored based on 1) the respondent's one word style description and 2) the overall accuracy of the translations for the one-word style description, the description of what the character is doing in one word, what they are doing in a sentence, and three-word description of what they are feeling. The results over the 13 survey responses showed moderate overall accuracy for all three words with a median of 0.75, and good one-word accuracy for "Stealthy" and "Attitude." The accuracy for translating the one-word translation for "Stumble" was less than 50%.

Table 1: *Accuracy of GPT 4.1 translating animator's descriptions into style labels*

Label	One word Accuracy	Overall Accuracy
Attitude	*9/13*	*Median: 0.75 (Avg 0.75)*
Stumble	*6/13*	*Median: 0.75 (Avg. 0.63)*
Stealthy	*11/13*	*Median: 0.75 (Avg 0.67)*

4 DISCUSSION

The project provides an example of a Machine Learning development pipeline that prioritises subjective expertise and embodied knowledge in the dataset generation stages and subjective expertise in the evaluation stage to enable animators to use their own words to style animation. Specifically the project makes the following key contributions to HCI research in AI for animation: an advance in the knowledge of the subjective values and vocabulary used by animators in practice and theory, an embodied process for creating datasets for training AI/ML models, and a method for using an LLM to translate between subjective expertise and the specific style labels used in training a model. The success of the model in producing animations that could be effectively mapped to the labels used in training, while moderate, achieves proof of concept for our approach and a foundation for future work to build upon and improve.

4.1 Knowledge of values and language used by animators in practice and theory

Even though the sample size was small owing to resistance to AI, a depth of information was gathered from four animators about how they use language. We created a database of words these animators use frequently in practice. To situate these words more broadly, the database provides metadata categorising them by the type of description and their alignment to Disney's *12 Principles of Animation.* Some of the most frequently used words were "exaggerated", "stretch", "squash", "thinking" "attitude", "confidence", "shaky/traumatized", "tense", "weight", "creeping/stealthy." This list includes words used to direct a scenario from the zombie apocalypse videogame, so words such as "shaky/traumatized" should be viewed as examples of describing specific emotions rather than representing the most frequently described emotions. The categorisation of the words by types of description such as quality/scale of motion ("tense", "weight"), emotional states ("attitude", "confidence", "traumatized"), pacing ("quick"), and expressivity ("creeping", "stealthy") makes a step toward a broader framework for understanding practical language. Mapping these practical words to the 12 Principles highlights the degree to which the principles guided the interviewees and places them within a frequently referenced conceptual framework. The representation of a range of types of description and mapping to the relevant principles lend credence to the collection of words beyond its use by the small number of interview participants.

4.2 Embodied process for creating datasets

Prior to the project RETìníZE had captured 20 hours of high-quality full body motion capture. We initially created a web-based tool for viewing and labelling this data with textual descriptions. Upon review, we rejected this approach for perpetuating troubling processes in AI development. Asking people to undertake the tedious, dehumanising work of labelling large amounts of data to produce AI that could potentially replace them in performing more fulfilling jobs is exploitative and contravenes their best interests (Crawford, 2021). While it may be difficult to argue that labelling twenty hours of motion capture is dehumanising in comparison to the millions of hours worked by poorly paid labourers to create ChatGPT (Perrigo, 2023), it is the principle that matters, particularly should our approach be replicated at scale. Furthermore, when working with

data taken from the human body, the approach of observing an avatar and applying labels is overly reductive and detached from the rich embodied source of the movement. Other human-centred approaches to AI and movement have utilised unsupervised or interactive Machine Learning methods to reduce the constraints of fixed labels in exploratory or freely expressive use cases (Fiebrink & Gillies, 2018; Gillies, 2019; Lamberti et al., 2018; Tomás et al., 2021). From these examples, one can imagine a system enabling animators to perform the motion they want to animate, and indeed Animotive does exactly this, using a VR headset and handheld controllers. An ideal system would be multimodal, enabling animators to act out motions they cannot describe in words or describe a motion in words that they struggle to act out. Often animators receive verbal or written direction, so using language to augment motion capture is appropriate to our use case, and because animators have a vocabularly in the 12 Principles, supervised learning with labels provided a starting point. The challenge was to produce labelled data in a more responsible and expressive way than the traditional approach of manual labelling.

Instead of manual labelling, we engaged the skill and creativity of actors to create a labelled dataset. While physically tiring due to the amount of repetitive movement required to create enough data for training an ML model, the actors were visibly engaged, positive and proactive in interpreting the labels with their bodies throughout the session. They clearly enjoyed the challenge. The labelling process itself was achieved simply by naming nine data files. This approach can be applied beyond the generation of motion capture data. “Pre-labelling” data by employing skilled humans to apply their craft values human effort, skill, experience and development. As AI automates the routine aspects of production, humans have an opportunity to extend their practice rather than retroactively labelling past work. Issues remain over who gets to develop skills and who gets stuck labelling, and the very act of taking data of any sort for training AI is fundamentally extractive. Adopting human-centred HCI research approaches to every step in the pipeline can help to align with the values and interests of those impacted. If the results are also of a higher quality, there is an incentive for others to adopt similar methods and the practice to spread.

4.3 Method for using an LLM to evaluate a model based on subjective human expertise

We explored the feasibility of utilising an LLM to translate subjective expert descriptions of animation to specific labels. Utilising OpenAI's GPT 4.1 API we tested three potential aspects of this translation: comparison of responses to the same survey questions, accuracy of identifying the label for each animation, and summarisation and translation of the animators' survey responses to the training labels. The accuracy of these translations had a median value of 75%. While this accuracy is not good enough for a production-level system, it is encouraging for an initial proof of concept. The method of evaluation is also promising. The translations captured the semantics in the human expert interpretations while reflecting their uncertainty; for example, one expert described the “attitude” animation as “dancing” in their one-word description but added words like “drunkenly” and “awkward” in the longer descriptions, leading GPT 4.1 to translate “dancing” to “attitude”, but the longer descriptions to “stumble.” Actors were given very broad latitude in how they enacted the word, and the interpretations by our experts were correspondingly broad. They would often embellish their actions in ways inspired by an imagined scenario like winning a trophy and holding it aloft for “attitude.” While the intention was to draw on the actors’ embodied knowledge and experience, an experienced director could help focus the performances to enhance expressiveness, leading to much higher accuracy in future.

Refinement can be achieved through cleaning the remaining motion capture files and reviewing the entire set for actions that may not match their labels or be too ambiguous, for example some of the actions in the "attitude" data could be confused with dancing. Additional sessions with an expanded selection of labels would greatly improve the existing dataset.

With such improvements, our approach would enable animators to use a subjective vocabulary based on their individual expertise rather than being limited to invoking the specific training labels defined by others, expressing in their own words what they want to achieve. As a feature of the Animotive platform, which enables animators to animate characters by performing movements physically, the ability to use natural language to adapt and refine their work has the potential to support increased creativity through multimodal experimentation. Creative agency could be enhanced further by training bespoke text to motion mappings.

5 CONCLUSION

The StyleCap project developed a human-centred HCI research approach to AI in animation. It was a collaboration between Academia and industry and as such had to balance the priorities and concerns of both. Our industrial partner has a duty to investors and shareholders to spend its time on research that advances its commercial value. Academic research has a duty to perform research ethically and in a way that advances knowledge. While not incompatible these duties can present a challenge, particularly when working to shorter timescales and budgets.

Following an HCI research approach to AI that embeds theoretical and ethical concerns in the process of achieving practical outputs can address this challenge and have impact. Our industrial partner RETìníZE describes the impact as follows

> The project fostered a genuine two-way exchange of expertise, which was essential in navigating such a rapidly evolving field and staying aligned with the cutting edge. Together, these contributions not only advanced our immediate goals but also laid the foundation for future innovation at the intersection of research and industry. (Jack Morrow, RETìníZE)

The pace of AI research creates pressure on both Industry and Academia to act quickly and adopt expedient methods. This project demonstrates that by placing subjective expertise and embodied knowledge at the centre of the HCI research process we can approach AI more responsibly while still achieving promising ML models, thus demonstrating a path forward for HCI research with AI.

6. ACKNOWLEDGEMENTS

The authors would like to thank the CoSTAR Realtime Lab for providing motion capture facilities and technical support. The work received funding from the University of York as a subaward of a grant from the Engineering and Physical Sciences Research Council (EPSRC) for the programme entitled “XR Network+: Virtual Production in the Digital Economy”

7. REFERENCES

Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., & Chen, B. (2020). Unpaired motion style transfer from video to animation. *ACM Trans. Graph.*, *39*(4), 64:64:1-64:64:12. https://doi.org/10.1145/3386569.3392469

AI and Hollywood: 5 questions for SAG-AFTRA’s chief negotiator | World Economic Forum. (2024. Retrieved 14 November 2025, from https://www.weforum.org/stories/2024/03/ai-hollywood-strike-sag-aftra-technology/

Alexanderson, S., Henter, G. E., Kucherenko, T., & Beskow, J. (2020). Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum*, *39*(2), 487–496. https://doi.org/10.1111/cgf.13946

Ali, G., Lee, M., & Hwang, J. I. (2020). Automatic text-to-gesture rule generation for embodied conversational agents. *Computer Animation and Virtual Worlds*, *31*(4–5). https://doi.org/10.1002/cav.1944

Animation Market Size, Statistics Report, Industry Growth, 2032. (2024). Retrieved 14 November 2025, from https://www.zionmarketresearch.com/report/animation-market

Animation Market Size to Hit over USD 895.71 Bn by 2034. (2025). Retrieved 14 November 2025, from https://www.precedenceresearch.com/animation-market

Bancroft, T., & Minkoff, R. (2019). *Directing for animation: Everything you didn’t learn in art school*. Routledge. https://learning.oreilly.com/library/view/directing-for-animation/9780240818023/xhtml/15_Chapter05.xhtml

Barandiaran, X. E., & Almendros, L. S. (2025). Transforming agency: On the mode of existence of large language models. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-025-10094-3

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, 610–623. https://doi.org/10.1145/3442188.3445922

Birhane, A. (2021). The Impossibility of Automating Ambiguity. *Artificial Life*, *27*(1), 44–61. https://doi.org/10.1162/ARTL_A_00336

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience Grounds Language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Chemero, A. (2009). *Radical embodied cognitive science*. MIT Press.

Clark, A. (2009). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Daikoku, T., Tanaka, M., & Yamawaki, S. (2024). Bodily maps of uncertainty and surprise in musical chord progression and the underlying emotional response. *iScience*, 109498. https://doi.org/10.1016/j.isci.2024.109498

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., Cusack, R., Lawrence, A. D., & Dalgleish, T. (2010). Listening to your heart:

How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, *21*(12), 1835–1844. https://doi.org/10.1177/0956797610389191

Estimating Your Animation Budget: Hourly vs. Flat Rate Projects. (2025). Retrieved 14 November 2025, from https://videoigniter.com/animation-budget/

Fiebrink, R., & Gillies, M. (2018). Introduction to the Special Issue on Human-Centered Machine Learning. *ACM Transactions on Interactive Intelligent Systems*, *8*(2), 1–7. https://doi.org/10.1145/3205942

Förger, K., & Takala, T. (2016). Animating with style: Defining expressive semantics of motion. *The Visual Computer*, *32*(2), 191–203. https://doi.org/10.1007/s00371-015-1064-4

Gillies, M. (2019). Understanding the Role of Interactive Machine Learning in Movement Interaction Design. *ACM Transactions on Computer-Human Interaction*, *26*(1), 1–34. https://doi.org/10.1145/3287307

Gratch, J., Kang, S.-H., & Wang, N. (2013). Using Social Agents to Explore Theories of Rapport and Emotional Resonance. In J. Gratch & S. Marsella (Eds), *Social Emotions in Nature and Artifact* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387643.003.0012

Herbison-Evans, D. (1980). A human movement language for computer animation. In J. M. Tobias (Ed.), *Language Design and Programming Methodology* (Vol. 79, pp. 117–128). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-09745-7_9

Hooks, E. (2017). *Acting for Animators* (4th edn). Routledge. https://doi.org/10.4324/9781315618272-4

How labor strikes and AI are shaping the future of entertainment. (2024). Retrieved 14 November 2025, from https://www.redsharknews.com/how-labor-strikes-and-ai-are-shaping-the-future-of-entertainment

How slow should you go? | Into Animation. (2023). Retrieved 14 November 2025, from https://www.intoanimation.com/articles/how-slow-should-you-go

Izani, M., Razak, A., Rehad, D., & Rosli, M. (2024). The Impact of Artificial Intelligence on Animation Filmmaking: Tools, Trends, and Future Implications. *2024 International Visualization, Informatics and Technology Conference (IVIT)*, 57–62. https://doi.org/10.1109/IVIT62102.2024.10692804

Karg, M., Samadani, A.-A., Gorbet, R., Kühnlenz, K., Hoey, J., & Kulić, D. (2013). Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation. *IEEE Transactions on Affective Computing*, *4*(4), 341–359. IEEE Transactions on Affective Computing. https://doi.org/10.1109/T-AFFC.2013.29

Kennedy, J. (2021). The animator's sensorium: The impact of acting and animation experience on creating reference performances. *Animation Practice, Process & Production*, *10*(1), 95–116. https://doi.org/10.1386/ap3_000028_1

Kines, M. (1998, October). Directing a Motion Capture Shoot. *Game Developer*, *5*(10), 48–53.

Lamberti, F., Paravati, G., Gatteschi, V., Cannavò, A., & Montuschi, P. (2018). Virtual Character Animation Based on Affordable Motion Capture and Reconfigurable Tangible Interfaces. *IEEE Transactions on Visualization and Computer Graphics*, *24*(5), 1742–1755. https://doi.org/10.1109/TVCG.2017.2690433

Mason, I., Starke, S., Zhang, H., Bilen, H., & Komura, T. (2018). Few-shot Learning of Homogeneous Human Locomotion Styles. *Computer Graphics Forum*, *37*(7), 143–153. https://doi.org/10.1111/cgf.13555

Merel, J., Tassa, Y., TB, D., Srinivasan, S., Lemmon, J., Wang, Z., Wayne, G., & Heess, N. (2017). *Learning human behaviors from motion capture by adversarial imitation* (arXiv:1707.02201). arXiv. https://doi.org/10.48550/arXiv.1707.02201

Merleau-Ponty, M. (1978). *Signs* (6. print). Northwestern Univ. Pr.

Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., & Hellier, P. (2022). A Survey on Deep Learning for Skeleton-Based Human Animation. *Computer Graphics Forum*, *41*(1), 122–157. https://doi.org/10.1111/cgf.14426

Nummenmaa, L., Hari, R., Hietanen, J. K., & Glerean, E. (2018). Maps of subjective feelings. *Proceedings of the National Academy of Sciences*, *115*(37), 9198–9203. https://doi.org/10.1073/pnas.1807390115

Pecune, F., & Marsella, S. (2020, October 20). A framework to co-optimize task and social dialogue policies using Reinforcement Learning. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020*. https://doi.org/10.1145/3383652.3423877

Peng, X. B., Abbeel, P., Levine, S., & van de Panne, M. (2018). DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.*, *37*(4), 143:1-143:14. https://doi.org/10.1145/3197517.3201311

Peng, X. B., Ma, Z., Abbeel, P., Levine, S., & Kanazawa, A. (2022). *AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control* (arXiv:2104.02180). arXiv. https://doi.org/10.48550/arXiv.2104.02180

Perrigo, B. (2023, January 18). *OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive*. TIME USA, LLC. Time. https://time.com/6247678/openai-chatgpt-kenya-workers/

Pfeifer, R., Bongard, J., & Grand, S. (2007). *How the body shapes the way we think: A new view of intelligence*. MIT Press.

Ronfard, R. (2021). Film Directing for Computer Games and Animation. *Computer Graphics Forum*, *40*(2), 713–730. https://doi.org/10.1111/cgf.142663

Schlesener, E. A., Ziolkowski, M., Wong, S. K., Westmoreland, B., & Babu, S. V. (2025). 'Am I Understood?': How the Interplay Between Embodiment and Theory of Mind Behavior Affects LLM-based Conversational Agents on Perceived Trust, Anthropomorphism, Presence, Usability, and User Experience. *ACM Transactions on Interactive Intelligent Systems*, 3774779. https://doi.org/10.1145/3774779

Starke, S., Starke, P., He, N., Komura, T., & Ye, Y. (2024). Categorical Codebook Matching for Embodied Character Controllers. *ACM Trans. Graph.*, *43*(4), 142:1-142:14. https://doi.org/10.1145/3658209

Statement on AI training. (n.d.). Retrieved 14 November 2025, from https://www.aitrainingstatement.org/

Sultana, N., Peng, L. Y., & Meissner, N. (2013). Exploring Believable Character Animation Based on Principles of Animation and Acting. *2013 International Conference on Informatics and Creative Multimedia*, 321–324. https://doi.org/10.1109/ICICM.2013.69

Thesen, T. P. (2020). Reviewing and Updating the 12 Principles of Animation. *Animation*, *15*(3), 276–296. https://doi.org/10.1177/1746847720969919

Thomas, F., & Johnston, O. (1981). *Disney animation: The illusion of life* (First edition.). Abbeville Press.

Tomás, E., Gorbach, T., Tellioğlu, H., & Kaltenbrunner, M. (2021, June 1). Embodied Gestures: Sculpting Energy-Motion Models into Musical Interfaces. *NIME 2021*. NIME 2021. https://doi.org/10.21428/92fbeb44.ce8139a8

University of Edinburgh. (n.d.). *ELM*. Retrieved 15 February 2026, from https://elm.edina.ac.uk/elm/elm

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999–6009. https://doi.org/10.48550/arxiv.1706.03762

Wang, Z., Merel, J. S., & Reed, S. E. (2017). Robust Imitation of Diverse Behaviors. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Weston, J. (2021). *Directing actors: Creating memorable performances for film and television*. Michael Wiese Productions.

Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2022). The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. *ACM International Conference Proceeding Series*, 736–747. https://doi.org/10.1145/3536221.3558058

Zhao, R., Papangelis, A., & Cassell, J. (2014). Towards a Dyadic Computational Model of Rapport Management for Human-Virtual Agent Interaction. In T. Bickmore, S. Marsella, & C. Sidner (Eds), *Intelligent Virtual Agents* (Vol. 8637, pp. 514–527). Springer International Publishing. https://doi.org/10.1007/978-3-319-09767-1_62

Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., & Wang, Y. (2024). Human Motion Generation: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(4), 2430–2449. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2023.3330935