EMo-Mask: Emotional Controllable Motion Generation

Yunong Liu

I. Abstract

This project introduces EMo-Mask, an extension of the motion generation model MoMask, which incorporates emotion understanding to generate diverse and expressive human motions. By integrating an EmotionEmbedder module, EMo-Mask learns meaningful representations of emotions and guides the motion generation process. Experiments demonstrate that incorporating emotion embeddings at the residual transformer input of MoMask yields the best performance. The project identifies limitations, such as the sensitivity of average Mean Squared Error loss, and proposes future directions for enhancing the EmotionEmbedder architecture and exploring perceptually relevant loss functions. EMo-Mask aims to generate emotionally expressive motions, contributing to the creation of more engaging and believable animated characters. Check Visualizations on Project Web page

II. Introduction

Motion generation models, such as MoMask [5], have made significant strides in generating realistic human motions from textual descriptions. However, these models often struggle to capture and convey the emotional expressiveness inherent in human movements [1]. Emotions play a crucial role in creating believable and engaging animated characters, as they add depth, personality, and relatability to their actions [12] Despite the importance of emotional expressiveness, MoMask and other state-of-the-art motion generation models primarily focus on generating motions based on the described actions, without explicitly considering the emotional context [7]. This limitation hinders the creation of truly expressive and diverse motions that can evoke the desired emotional response from the audience. In this project, we tried to incorporate emotion understanding into the motion generation process, which improve MoMask's ability to generate diverse and expressive motions, enhancing the overall quality and realism of animated characters [2].

III. Methodology

A. Emotion Embedding

To incorporate emotion understanding into MoMask, we propose the EmotionEmbedder module, which learns compact and meaningful representations of emotions. We focus on four key emotions derived from Plutchik's wheel of emotions [11]: Joy, Sadness, Fear, and Anger.

Fig. 1: Plutchik's wheel of emotions[11]

The EmotionEmbedder consists of an embedding layer that maps emotion labels to dense vectors, followed by fully connected layers and activation functions [13]. The module is trained to capture the semantic relationships between emotions and their associated motion characteristics.

B. Data collection process

To train the EmotionEmbedder, we require a dataset of emotionally expressive motions. However, existing motion capture datasets often lack detailed emotional annotations [9]. To overcome this challenge, we propose decomposing emotionally expressive motions into textual descriptions using a Large Language Model (LLM) [3]. The LLM is prompted to describe the body language and movements associated with each emotion, providing a rich and detailed representation of emotionally expressive motions. These descriptions are then used as training data for the EmotionEmbedder.

C. EmotionEmbedder with MoMask

We explore three different integration points for incorporating the EmotionEmbedder into MoMask: the Residual Transformer, Mask Transformer, and VQ-VAE [6]. The Residual

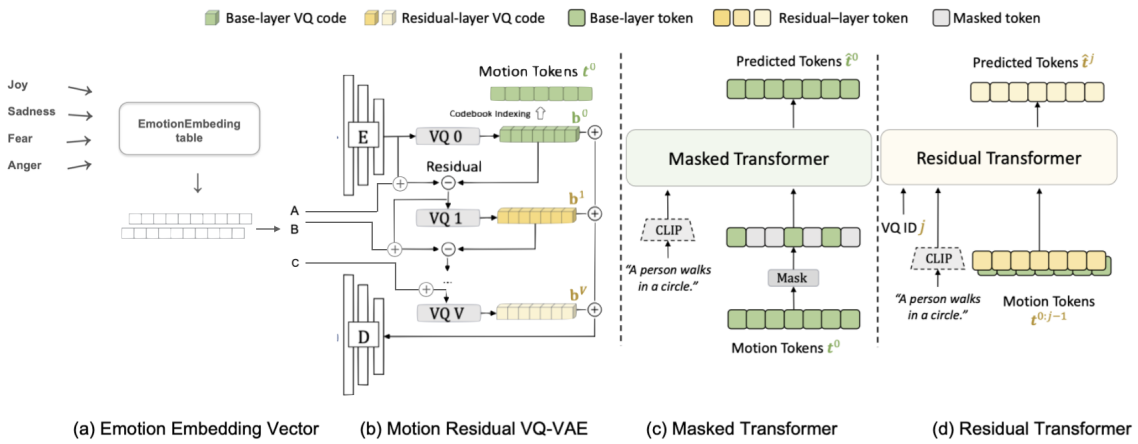

Fig. 2: **EMo-Mask Approach overview.** (a) Emotion Embedding and Integration: The EmotionEmbedder learns emotion representations, which are integrated into the motion generation process at different points, including the R-transformer, M-transformer, and VQ-VAE inputs. (b) Motion Residual VQ-VAE: The base quantization layer employs vector quantization (VQ) to tokenize motion sequences, while the residual quantization layers utilize hierarchical residual quantization to capture fine-grained motion details. (c) Masked Transformer: The Masked Transformer performs parallel prediction by masking and predicting base layer tokens, conditioned on text input and emotion embeddings. (d) Residual Transformer: The Residual Transformer conducts progressive prediction, predicting residual tokens in a layer-by-layer manner, conditioned on previous layer tokens and emotion embeddings.

Transformer is responsible for refining the generated motion tokens based on the previous layers' tokens, while the Mask Transformer generates the initial base motion tokens from the input text description. The VQ-VAE encodes the motion sequences into a latent space and decodes them back into motion tokens. By integrating the EmotionEmbedder at these different points, we aim to investigate the most effective way to incorporate emotion understanding into the motion generation process.

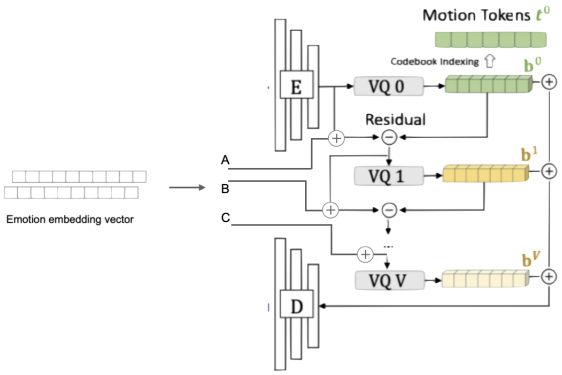

Fig. 3: 3 different points for incorporating the EmotionEmbedder into MoMask: (A) Mask Transformer. (B) Residual Transformer. (C) VQ-VAE.

D. Training the EmotionEmbedder

To train the EmotionEmbedder, we collected a dataset consisting of 320 motions, encompassing 4 emotions (Joy, Sadness, Fear, Anger) and 2 motion types (Walking and Running). The dataset contains 40 motion sequences for each emotion-motion pair (e.g., "Joy-Walking", "Anger-Running"), resulting in a total of 8 emotion-motion combinations. These motion sequences serve as the ground truth for learning the emotional representations.

The EmotionEmbedder is trained using a combination of loss functions to effectively capture the relationship between emotions and motions. The primary loss function employed is the Mean Squared Error (MSE) loss, which measures the average squared difference between the generated motion and the average motion of the 40 collected motion sequences for each emotion pair. The MSE loss is calculated as follows:

$$\mathcal{L}_{MSE} = (m_g - \frac{1}{N}\sum_{i=1}^{N} m_i)^2 \tag{1}$$

where N is the total number of motion sequences for the the corresponding emotion pair (i.e. 40 in this project), m_g is the generated motion sequence. .

By minimizing the MSE loss, the EmotionEmbedder learns to generate motions that closely resemble the average motion of each emotion pair, effectively mapping emotions to their associated motion representations. However, relying solely on the MSE loss may lead to generated motions that lack diversity and fail to capture the nuances of individual motion sequences within each emotion pair.

To address this limitation, we incorporate a contrastive loss alongside the MSE loss during training. The contrastive loss encourages the emotion embeddings to be distinct and well-separated in the embedding space. It is calculated as follows:

$$\mathcal{L}_{contrastive} = \sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}d_{ij} + (1-y_{ij})\max(0, m-d_{ij}) \quad (2)$$

where N is the total number of emotion pairs, d_{ij} is the Euclidean distance between the embeddings of the i-th and j-th emotion pairs, y_{ij} is a binary label indicating whether the i-th and j-th emotion pairs belong to the same emotion category, and m is a margin parameter.

The contrastive loss penalizes the model when the embeddings of emotion pairs from the same category are far apart and when the embeddings of emotion pairs from different categories are close together. By minimizing the contrastive loss, the EmotionEmbedder learns to generate emotionally distinct and discriminative motion representations, enabling the model to differentiate between different emotional states effectively.

The total loss for training the EmotionEmbedder is a weighted combination of the MSE loss and the contrastive loss:

$$\mathcal{L}_{total} = \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_{contrastive}\mathcal{L}_{contrastive} \quad (3)$$

where λ_{MSE} and $\lambda_{contrastive}$ are hyperparameters that control the relative importance of each loss term.

By jointly optimizing the MSE loss and the contrastive loss, the EmotionEmbedder learns to generate emotionally expressive motions that closely resemble the average motion of each emotion pair while maintaining distinct and well-separated emotion representations in the embedding space.

IV. Results and Analysis

When the EmotionEmbedder is integrated at the Residual Transformer input, EMo-Mask achieves the best performance, generating motions that exhibit the desired emotional expressiveness. The Residual Transformer, responsible for refining the generated motion tokens based on the previous layers' tokens [6], benefits the most from the emotion embeddings. The additional emotional context provided by the EmotionEmbedder guides the refinement process, resulting in motions that accurately convey the intended emotions. In contrast, integrating the EmotionEmbedder at the Mask Transformer input yields suboptimal results. The generated motions tend to be short and less realistic compared to the ground truth motions. This can be attributed to the fact that the Mask Transformer generates the initial base motion tokens from the input text description [6]. The emotion embeddings, when introduced at this early stage, may conflict with the content of the text description, leading to inconsistencies in the generated motions.

Integrating the EmotionEmbedder at the VQ-VAE input also produces unsatisfactory results. The generated motions appear unrealistic and not human-like, lacking the fluidity and coherence expected from natural human movements. The VQ-VAE, responsible for encoding motion sequences into a latent space and decoding them back into motion tokens [6], may struggle to effectively incorporate the emotion embeddings during the encoding and decoding process. The emotion information may be lost or distorted, resulting in motions that fail to capture the desired emotional expressiveness.

Further analysis of the impact of each integration point on the generation process reveals insights into the interplay between emotion understanding and motion generation. The Residual Transformer, with its ability to refine and adjust the generated motions based on the provided emotion embeddings, proves to be the most suitable point for integrating emotional context. The Mask Transformer and VQ-VAE, on the other hand, may not have the necessary capacity or flexibility to effectively incorporate emotion information without compromising the quality and realism of the generated motions.

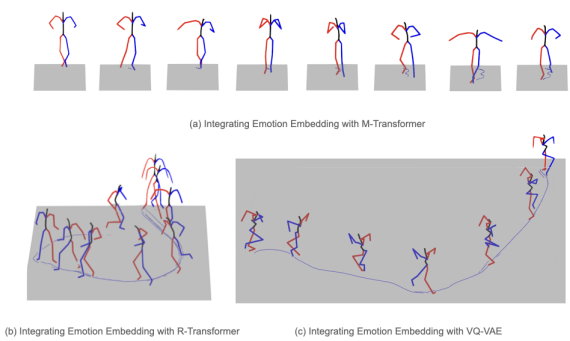

Fig. 4: Results for integrating emotion embedding to different points. (a) Integrating Emotion Embedding with M-Transformer, corresponding to point A in Fig 3. (b) Integrating Emotion Embedding with R-Transformer, corresponding to point B in Fig 3.(c) Integrating Emotion Embedding with VQ-VAE, corresponding to point C in Fig 3

These findings highlight the importance of carefully selecting the integration point for incorporating emotion understanding into the motion generation pipeline. The results suggest that integrating the EmotionEmbedder at the Mask Transformer input strikes the right balance between preserving the content of the text description and infusing the generated motions with the desired emotional expressiveness [2]

V. Lessons Learned

The EMo-Mask project has yielded valuable insights into the process of incorporating emotion understanding into motion generation models. One of the key lessons learned is the importance of selecting the appropriate integration point for emotion embeddings. The experiments conducted with different integration points (Residual Transformer, Mask Transformer, and VQ-VAE) demonstrate that the choice of where to introduce emotion information significantly impacts the

quality and expressiveness of the generated motions [6]. The Mask Transformer emerged as the most effective integration point, highlighting the need for careful consideration of the model architecture when incorporating additional context.

Another important lesson pertains to the limitations of using average Mean Squared Error (MSE) loss during the training process. The MSE loss, while commonly used in motion generation tasks [10], has several drawbacks. It is sensitive to outliers, meaning that a few highly dissimilar or noisy motion samples can disproportionately affect the overall loss, leading to suboptimal training. Additionally, the MSE loss lacks perceptual relevance, as it treats all motion differences equally without considering their perceptual significance [2]. This can result in generated motions that may be numerically similar to the ground truth but fail to capture the nuances and expressiveness of human movements.

Furthermore, the averaging effect introduced by the MSE loss can lead to generated motions that lack the specificity and diversity of individual reference motions. This highlights the need for exploring alternative loss functions and training strategies that can better capture the perceptual qualities of motions and encourage the generation of diverse and expressive movements.

The EMo-Mask project also reveals the potential for incorporating a wider range of emotions and motion types. While the current implementation focuses on four key emotions (Joy, Sadness, Fear, Anger), there is an opportunity to expand the emotional repertoire to capture more nuanced and complex emotional states [11]. Additionally, exploring the integration of emotion embeddings into different motion types, such as gestures, facial expressions, and full-body movements, can further enhance the expressiveness and realism of the generated motions [2].

VI. Future Directions

The EMo-Mask project has demonstrated the potential of incorporating emotion understanding into motion generation models, but there are several exciting avenues for future research and development. One promising direction is to enhance the EmotionEmbedder architecture and explore advanced learning techniques. This could involve investigating more sophisticated architectures, such as transformer-based embedders [13] or graph neural networks [8], to capture the complex relationships and dynamics between emotions. Additionally, incorporating unsupervised or self-supervised learning techniques [4] could help in learning more nuanced and robust emotion representations from unlabeled motion data.

Another important area for future research is the investigation of perceptually relevant loss functions and evaluation metrics. Moving beyond the limitations of average MSE loss, exploring loss functions that prioritize perceptual similarity and capture the subjective quality of motions could lead to more realistic and expressive generated outputs [15]. Furthermore, developing evaluation metrics that assess the emotional expressiveness and perceptual fidelity of the generated motions would provide a more comprehensive understanding of the model's performance.

Extending EMo-Mask to handle a broader spectrum of emotions and motion styles is another exciting direction. Incorporating a more diverse set of emotions, including subtle and complex emotional states, would enhance the model's ability to generate nuanced and realistic motions [11]. Additionally, exploring the integration of emotion embeddings into different motion styles, such as dance, sports, or character-specific movements, could expand the applicability of the EMo-Mask framework to various domains.

Finally, exploring the application of EMo-Mask in domains such as gaming and virtual reality presents immense potential. Emotionally expressive motion generation can greatly enhance the immersion and engagement of interactive experiences [14]. By incorporating EMo-Mask into game engines or virtual reality platforms, developers can create more lifelike and relatable characters that respond dynamically to user interactions and emotional cues. This could revolutionize the way stories are told and experiences are crafted in these domains.

VII. Conclusion

The EMo-Mask project was motivated by the need to incorporate emotion understanding into motion generation models, specifically MoMask, to create more expressive and diverse human motions. Through the development of the EmotionEmbedder and its integration into MoMask at different points, the project successfully demonstrated the impact of emotion embeddings on the quality and expressiveness of the generated motions.

The key findings of the project highlight the importance of selecting the appropriate integration point, with the Mask Transformer input yielding the best results. The project also identified the limitations of average MSE loss and the need for exploring alternative loss functions and training strategies. The potential for incorporating a wider range of emotions and motion types was also recognized as a valuable direction for future research.

The successful integration of emotion understanding in MoMask through EMo-Mask has significant implications for creating more expressive and engaging animated characters. By generating motions that accurately convey the intended emotions, EMo-Mask enables animators and developers to bring characters to life in a more meaningful and relatable way. This has the potential to enhance the storytelling and immersion in various applications, from animated films to video games and virtual reality experiences.

Looking ahead, the future of emotionally expressive motion generation is bright. With continued research and development in this field, we can expect to see more sophisticated and nuanced motion generation models that capture the full spectrum of human emotions. As these technologies advance, they will transform the way we create and interact with virtual characters, leading to more compelling and emotionally resonant experiences.

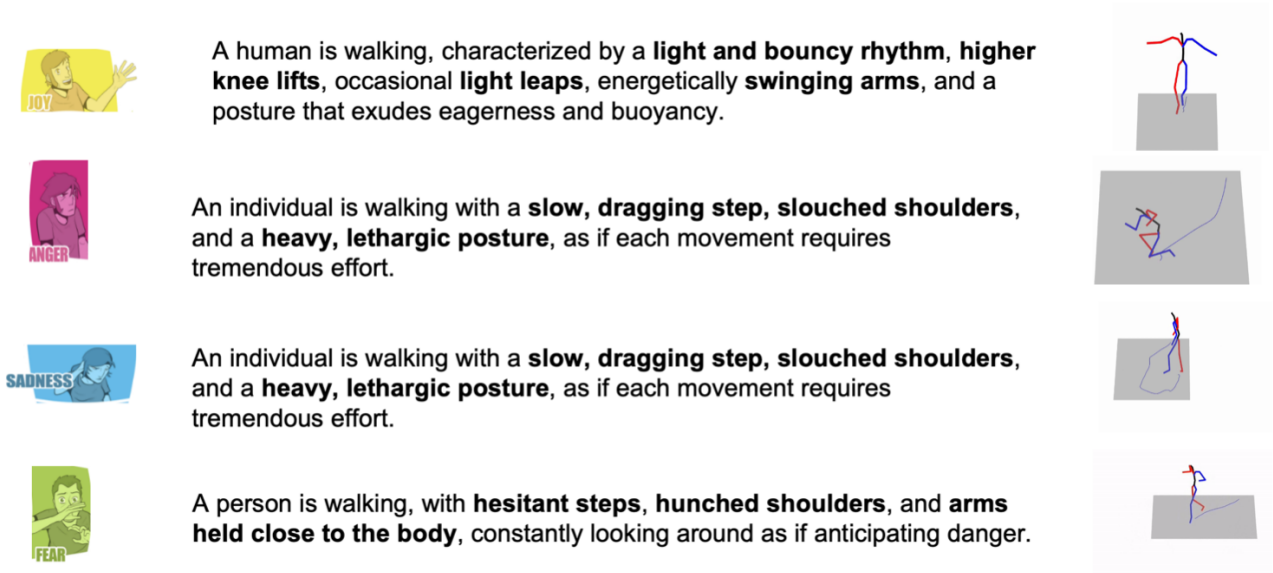

Fig. 5: **Data Collection Process:** (a) Identify required emotion. (b) Collect emotionally expressive motion descriptions from LLM. (c) Collect emotionally expressive motion descriptions from MoMask using prompt from LLM

References

[1] Aristidou, A., Cohen-Or, D., Hodgins, J.K., Chrysanthou, Y., Shamir, A.: Deep motifs and motion signatures. ACM Transactions on Graphics (TOG) **37**(6), 1–13 (2018)

[2] Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In: 2021 IEEE virtual reality and 3D user interfaces (VR). pp. 1–10. IEEE (2021)

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[4] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)

[5] Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. arXiv preprint arXiv:2312.00063 (2023)

[6] Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)

[7] Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)

[8] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

[9] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)

[10] Pavllo, D., Feichtenhofer, C., Auli, M., Grangier, D.: Modeling human motion with quaternion-based neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7118–7128 (2019)

[11] Plutchik, R.: A psychoevolutionary theory of emotions (1982)

[12] Tinwell, A., Grimshaw, M., Williams, A.: The uncanny wall. International journal of arts and technology **4**(3), 326–341 (2011)

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[14] Vinayagamoorthy, V., Gillies, M., Steed, A., Tanguy, E., Pan, X., Loscos, C., Slater, M.: Building expression into virtual characters. In: Eurographics (STARs). pp. 21–61. Citeseer (2006)

[15] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)