FEEL: Emotion-Enriched Text-to-Motion Generation Using Small-Scale Emotion Motion Data

A. Overview

In this supplementary material, we provide additional content to support the main paper. First, we present experimental results conducted on the synthetic emotion dataset, EmotionT2M, followed by detailed explanations of the evaluation metrics used in our experiments. Next, we provide additional qualitative results that visually illustrate the generated motions. We also offer comprehensive quantitative comparisons with other models under various emotional conditions. Lastly, we include further analysis of our experimental findings.

B. Quantitative Comparison on Emotion-Synthetic Data

B.1. Dataset

In this study, we evaluate the performance of emotion-based text-to-motion generation using the EmotionT2M [4] dataset. This dataset is constructed based on HumanML3D and IDEA400, where a large language model (LLM) is employed to automatically generate emotion-rich textual descriptions. The motion data itself is synthesized by modifying existing samples through LLM-guided retargeting. Emotion labels are annotated either manually or automatically. The dataset includes both sentence-level emotion annotations (EmotionalT2M) and fine-grained limb-level annotations (Limb-ET2M), serving as a benchmark to simulate realistic emotional expressions. In our work, we evaluate on a subset of the test set, focusing only on overlapping emotions: joy, sad, and anger.

Motion Representation A motion sequence x consists of a series of poses, represented as $x = \{\dot{r_a}, \dot{r_x}, \dot{r_z}, r_y, j_p, j_v, j_r, c_f\}$, where root features include the angular velocity $\dot{r_a} \in \mathbb{R}$, linear velocities $(\dot{r_x}, \dot{r_z}) \in \mathbb{R}^2$, and root height $r_y \in \mathbb{R}$. Joint features are composed of positions $j_p \in \mathbb{R}^{3N_j}$, velocities $j_v \in \mathbb{R}^{3N_j}$, and rotations $j_r \in \mathbb{R}^{N_j}$, where N_j denotes the number of joints. Foot contact states are encoded as $c_f \in \mathbb{R}^4$, capturing ground interaction information. The total dimensionality of this feature vector is 263, effectively representing both joint dynamics and root movements for motion generation. For emotion-aware text-to-motion synthesis, we leverage HumanML3D for modeling general text-motion alignment and EMILYA for providing explicit emotion conditioning.

Implementation Detail HumanML3D [2] contains 14,616 motion sequences and 44,970 text annotations, split into train (12,102 motions), validation (1,441 motions), and test (1,472 motions) sets following the official protocol.

C. Additional Experiments

C.1. Evaluation Metrics Detail

To evaluate the effectiveness of our approach, we utilize several established evaluation metrics that assess the quality and diversity of generated motions. Following prior works, we extract motion and text features using a pretrained network and measure the alignment between text and motion. However, our model differs from existing methods in that it considers both textual descriptions and emotional attributes during motion generation. As a result, while standard evaluation metrics can capture how well generated motions match their corresponding text descriptions, these metrics are designed to evaluate motions based on text descriptions without emotional attributes, motions generated with emotional considerations may not achieve high performance under standard evaluation criteria. Consequently, when comparing emotion-aware generation models with text-only models, natural differences arise due to the added complexity of incorporating emotions.

R-Precision R-Precision evaluates the accuracy of motion-to-text retrieval by ranking a given motion among a set of text descriptions. Each motion is compared against 32 candidate text descriptions, including one ground-truth and 31 randomly sampled mismatched descriptions. The ranking is based on Euclidean distances in the feature space, and we report the accuracy at Top-1, Top-2, and Top-3 levels. Since our model generates motions conditioned on both text and emotion, the retrieved text may align well with the motion but may not fully capture the emotional nuances encoded in the motion. This inherent

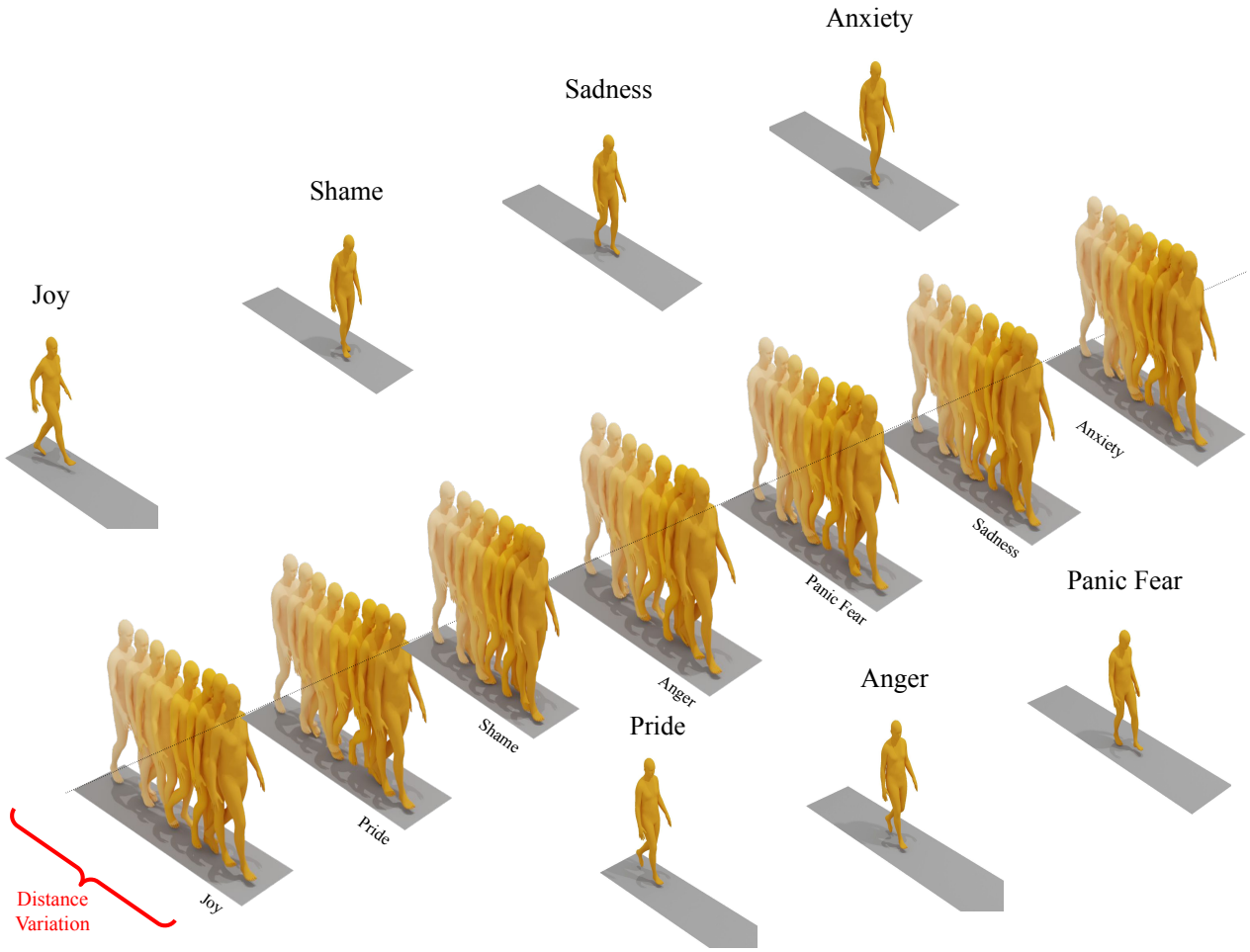

Figure A. Qualitative comparison of emotion-conditioned motion generation using our model. The figure visualizes changes across seven emotional states based on the input text prompt "man walks forward moving hands and neck".

difference makes direct comparison with text-only models less straightforward.

Fréchet Inception Distance (FID) FID measures the overall quality of generated motions by computing the distributional distance between real and generated motion features:

$$\text{FID} = \|\mu_{gt} - \mu_{pred}\|^2 + \text{Tr}(\Sigma_{gt} + \Sigma_{pred} - 2(\Sigma_{gt}\Sigma_{pred})^{\frac{1}{2}}). \quad (1)$$

Lower FID scores indicate that the generated motions closely resemble real motion data. While FID is widely used to assess motion quality, it does not differentiate between emotionally expressive and neutral motions. Consequently, FID comparisons between emotion-conditioned motions and standard text-to-motion models are inherently different due to the impact of emotional conditioning, leading to inevitable discrepancies.

Diversity Diversity measures the variance among generated motion sequences within the dataset. It is computed by randomly selecting $S_{dis} = 300$ pairs of motions and averaging their feature-space Euclidean distances:

$$\text{Diversity} = \frac{1}{S_{dis}} \sum_{i=1}^{S_{dis}} \|f_{pred,i} - f'_{pred,i}\|. \quad (2)$$

A high Diversity score suggests that the model generates a wide range of motion variations. Since our model incorporates emotion as an additional control signal, it inherently increases motion variation compared to text-only baselines. Therefore, our Diversity scores may be higher, reflecting not only general motion variability but also the diversity of emotional expressions.

Multimodality (MModality) MModality evaluates the diversity of motions generated from the same text description. Given a single text prompt, we generate 20 differ-

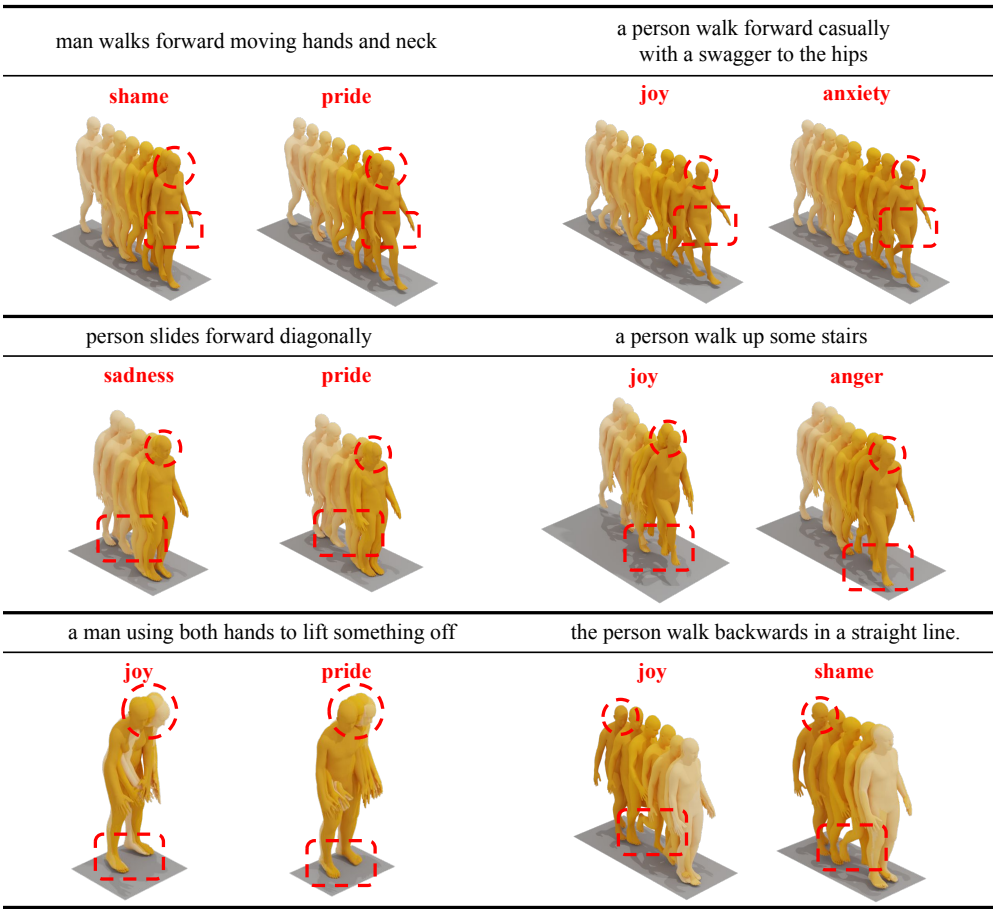

Figure B. Qualitative comparison of emotion-conditioned motion generation using our model. The figure visualizes changes across seven emotional states based on various text.

ent motion sequences and form 10 random motion pairs. The average Euclidean distance between these paired motion features is computed as:

$$\text{MModality} = \frac{1}{10N} \sum_{i=1}^{N} \sum_{j=1}^{10} \| f_{pred,i,j} - f'_{pred,i,j} \|, \quad (3)$$

where N is the total number of text descriptions. Unlike standard models that rely solely on textual inputs, our model introduces an additional emotional dimension that influences motion variability. This leads to higher MModality scores, as the same text prompt can yield significantly different motions depending on the emotional context. Consequently, comparing MModality scores between our model and a text-only baseline must consider the inherent increase in variation due to emotional conditioning.

Overall, while these metrics provide valuable insights into motion generation quality, they do not fully capture the expressive variations introduced by emotion conditioning. This intrinsic difference must be considered when comparing our model with conventional text-to-motion generation models.

C.2. Additional Qualitative Results

Figure A illustrates qualitative differences in motion generation when the same text input is conditioned on different emotions. Variation in displacement across emotional states aligns with findings from gait analysis [3], which suggest that walking speed reflects emotional context. These observations support the model's ability to capture emotion-dependent motion characteristics. Moreover, intermediate frames reveal distinct emotional cues, indicating that our method generates realistic and affectively expressive motions.

For instance, when the prompt "man walks forward moving hands and neck" is conditioned on different emotions, clear variations emerge in stride length, posture, and arm movement. Joy and pride lead to longer, more dynamic strides, while shame and sadness produce shorter strides and a more withdrawn posture. Anxiety and panic fear result in tightened arm motion and reduced forward momentum, reflecting tension and urgency.

Table A. Comparison of ERA-2 scores across different models The highest score for each emotion is highlighted in red, while the second highest score is highlighted in blue. Higher ERA-2 scores indicate greater emotional consistency in the generated motions.

Emotion	Ours	MoMask	ParCo	TM2T	BAMM	MMM
Joy	$55.8944^{\pm 0.1095}$	$34.713^{\pm 0.243}$	$52.6437^{\pm 2.575}$	$44.4612^{\pm 0.6565}$	$44.914^{\pm 0.282}$	$39.7701^{\pm 3.143}$
Pride	$50.8226^{\pm 0.4283}$	$32.399^{\pm 0.230}$	$35.1724^{\pm 1.912}$	$34.3175^{\pm 0.7795}$	$27.601^{\pm 0.768}$	$29.1954^{\pm 0.736}$
Shame	$20.6214^{\pm 0.2987}$	$11.774^{\pm 0.377}$	$17.7011^{\pm 3.841}$	$12.5647^{\pm 0.5713}$	$10.905^{\pm 0.313}$	$10.1149^{\pm 1.326}$
Anger	$20.7507^{\pm 0.2202}$	$8.786^{\pm 0.477}$	$11.0345^{\pm 3.372}$	$11.6379^{\pm 0.2823}$	$8.139^{\pm 0.265}$	$10.1149^{\pm 3.143}$
Panic Fear	$42.9562^{\pm 0.2661}$	$33.484^{\pm 0.433}$	$43.2184^{\pm 5.419}$	$29.7917^{\pm 0.7387}$	$29.353^{\pm 0.060}$	$50.1149^{\pm 1.947}$
Sadness	$40.9591^{\pm 0.3022}$	$39.770^{\pm 0.376}$	$25.2874^{\pm 2.873}$	$39.2170^{\pm 0.4490}$	$37.723^{\pm 0.677}$	$32.6437^{\pm 6.813}$
Anxiety	$36.6092^{\pm 0.6267}$	$28.218^{\pm 0.611}$	$20.2299^{\pm 3.143}$	$29.4109^{\pm 0.6856}$	$23.513^{\pm 0.573}$	$30.5747^{\pm 2.048}$

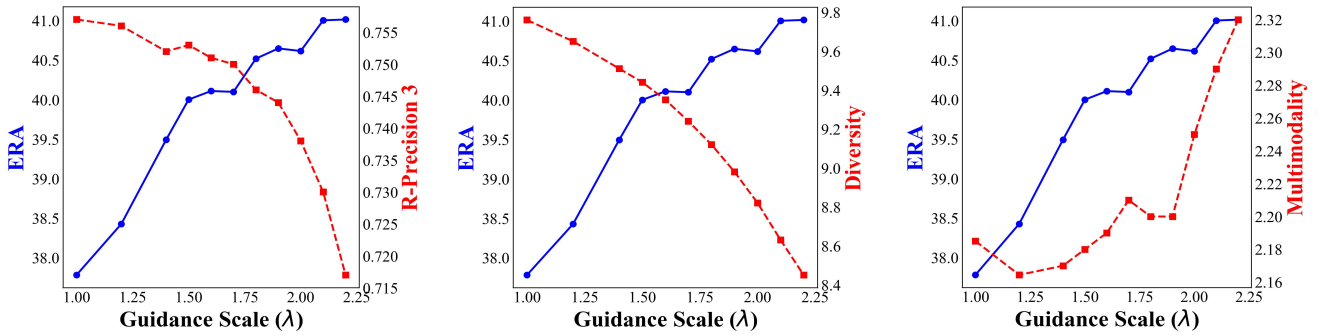

Figure C. Effect of guidance scale on emotion recognition and content preservation. Blue lines indicate ERA (ERA); red lines show (a) R-Precision 3, (b) Diversity, and (c) Multimodality. Higher guidance improves ERA but reduces motion fidelity and diversity, revealing trade-offs across metrics.

Figure B provides a more detailed qualitative comparison across emotional conditions. Red boxes highlight differences in stride length or arm movement, while red circles denote variations in head posture. Compared to baseline models that often produce generic and less expressive outputs, our method captures subtle, emotion-specific details with greater coherence and realism. For example, when "shame" is applied to the prompt "man walks forward moving hands and neck," the stride shortens, arm movement becomes constrained, and the overall posture appears hunched with the head lowered—effectively conveying withdrawal. Conversely, "pride" results in longer strides, increased arm dynamics, and an upright posture, with the head lifted and chest projected forward to express confidence.

In another example, the emotion "joy" induces energetic and rhythmic motion with natural arm swings and steady head movement, while "anxiety" causes irregular stride patterns, limited arm movement, and a slight head tilt that communicates tension. Similar patterns are observed across other prompts such as walking diagonally, climbing stairs, and lifting with both hands. In each case, emotions such as "sadness," "anger," and "pride" modulate stride, arm articulation, and head position in ways that are consistent with their expressive intent.

These results confirm that emotions influence not only stride length and limb movement but also head posture and overall motion dynamics. Unlike previous approaches that struggle to produce emotionally coherent or expressive motion, our method captures these nuanced variations while maintaining the structural integrity and fluidity of the movement.

C.3. Additional Quantitative Results

Table A presents a comparison of ERA-2 scores across various models. Our approach consistently outperforms baseline methods, particularly in emotions like Joy (55.89) and Pride (50.82), where emotional expressiveness and motion realism are critical. In contrast, Shame and Anger remain challenging across all models, though our method still leads in both categories.

For Panic Fear, MMM slightly outperformed our model in ERA, but our method achieved a better FID score, in-

dicating stronger visual quality despite marginally lower emotional alignment. In Sadness and Anxiety, our model achieved top scores, demonstrating stable emotional modeling across a range of affective states.

These results highlight the strength of our emotion-aware motion synthesis framework in generating expressive, realistic, and emotionally coherent sequences across diverse affective conditions.

Ablation Study on Scale Guidance Figure C illustrates the effect of guidance scale variation on emotion recognition and motion preservation. As the guidance scale increases, emotion alignment measured by ERA steadily improves. However, this comes at the cost of motion fidelity and diversity, suggesting a clear trade-off between emotional fidelity and content preservation.

In subplot (a), R-Precision 3 declines as ERA increases, indicating that stronger emotion conditioning reduces similarity to the original motion. In (b), Diversity gradually decreases, implying that increased emotional influence limits its variation in motion expression. Interestingly, (c) reveals a different trend: Multimodality remains stable at low guidance levels but increases beyond a certain threshold, indicating enhanced intra-class variation within the same emotional category.

These results underscore the importance of carefully tuning the guidance scale to strike a balance between emotion expressiveness and motion consistency. Beyond a certain point, stronger guidance can simultaneously improve emotional expressiveness and promote motion variability highlighting its dual impact on both fidelity and diversity.

Effect of Classifier Guidance on Emotion Recognition To control the strength of emotion conditioning, we introduce a scaling factor λ. This factor adjusts the contribution of the emotion difference term ΔG^i, allowing the model to modulate the intensity of the emotion while maintaining the consistency of the text. Figure D shows the impact of λ on classifier-guided emotion recognition. As λ increases, the emotion recognition performance (ERA, blue line) improves. However, beyond a certain threshold, motion quality deteriorates, leading to higher FID scores (red line). These results highlight the strength of the guidance to balance emotion recognition accuracy and motion quality.

C.4. Additional Analysis

Confusion Matrix Analysis Figure E presents the confusion matrix, illustrating how the emotion recognition model classifies generated motions based on input emotions. Similar emotions tend to be misclassified more frequently, appearing in lighter colors, while distinct emotions are more clearly separated in darker shades. For instance, motions generated with joy are often confused with pride or panic

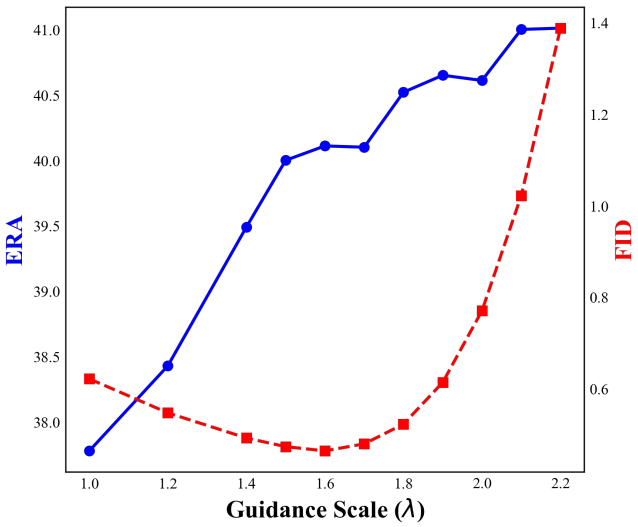

Figure D. The graph investigates the change in the emotion recognition matrix (ERA) for the emotion *Sadness* as the guidance scale varies.

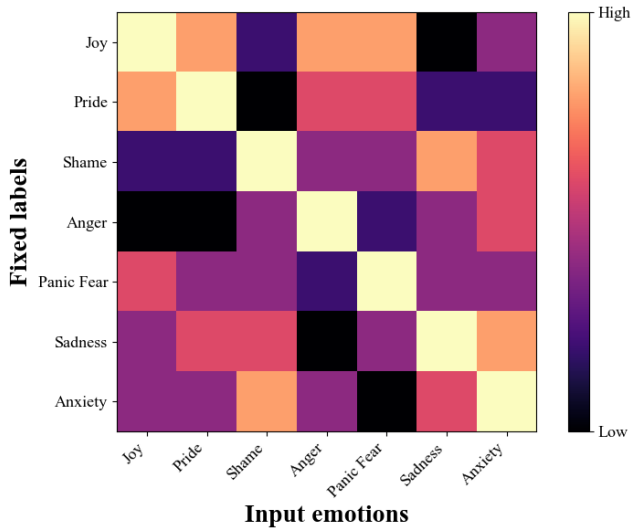

Figure E. Confusion matrix of the emotion recognition model. The y-axis shows fixed emotions, and the x-axis represents input emotions.

fear, but remain distinguishable from shame and anger. These patterns indicate that the generated motions effectively capture emotional nuances, aligning with expectations that emotions like joy and anger should be distinguishable. This suggests FEEL maintains a reasonable and consistent emotional representation in motion synthesis.

Effect of Emotion on Motion Distribution Figure F presents a PCA visualization of generated motions. We generate 10 motion samples for the same text caption under

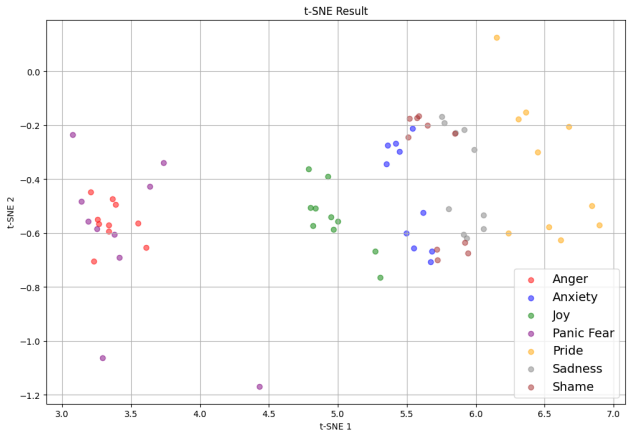

Figure F. PCA visualization of motion sequences generated from 10 semantically equivalent texts across 7 emotions.

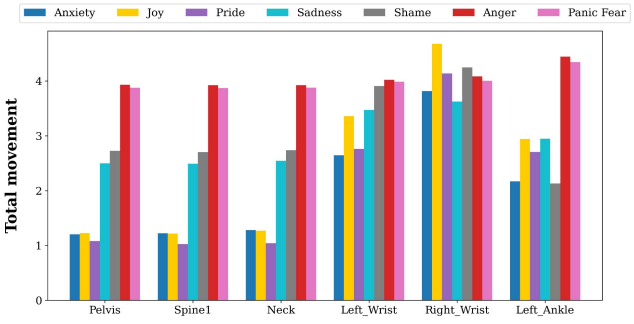

Figure G. Joint movement visualization.

7 different emotion inputs (70 motions in total) and apply PCA to project them into the latent space.

The PCA results show that motions generated with anger, panic, and fear are clustered together, while those conditioned on shame, sadness, and anxiety form a dense group. This distribution aligns with our observations that motions generated with shame, sadness, and anxiety exhibit similar characteristics.

Upon further analysis, we observe that motions generated with shame, sadness, and anxiety share similar movement patterns, often displaying more constrained and reserved motions. In contrast, motions generated with anger, panic, and fear tend to exhibit more dynamic and intense movements. This supports the validity of our experimental results, demonstrating that emotional conditioning leads to meaningful variations in motion characteristics.

Effect of Joint-Wise Movement We analyze the effect of emotion conditioning on joint-wise total movement under identical base motion. Rather than uniformly increasing overall scale, emotions lead to different distributions of movement across joint groups. As seen in Figure G, anger shows the largest overall movement, with noticeable activity in the core (Pelvis, Spine1, Neck), the wrists (Left_Wrist, Right_Wrist), and the lower extremities such as the Left_Ankle. Panic Fear exhibits a similar but slightly weaker pattern, whereas Joy, Pride, and Anxiety remain more restrained, showing less movement, often concentrating movement in the wrists. These results align with known behavioral tendencies for each emotion and indicate that the generated motions reasonably capture affect-specific variations.

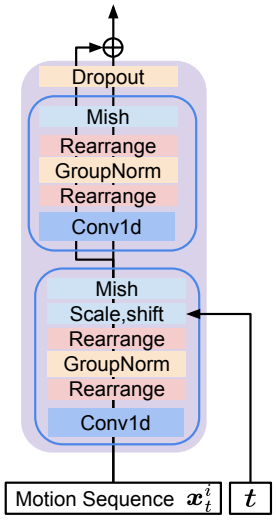

Figure H. Detailed architecture of the Conv1D block used in our model.

Computational Cost Analysis Adopting a shared-parameter dual-network architecture inevitably increases model complexity. To quantify this cost, we report that our model requires approximately **23GB GPU memory** and **1.2 days of training time**, compared to **16.7GB** and **1 day** for the baseline model (StableMofusion). This corresponds to about **6GB additional memory consumption** and a **20% longer training time**, while maintaining comparable inference efficiency due to parameter sharing.

D. Architecture Detail

Architecture Overview Our model builds upon StableMoFusion, utilizing the Conv1D block from the Conv1D UNet architecture with AdaGN [1]. StableMoFusion is a diffusion-based text-to-motion framework that conditions motion generation on CLIP text embeddings via Feature-wise Linear Modulation (FiLM), projecting both the text and timestep embeddings into per-frame scale and shift parameters. The Conv1D UNet framework is well-suited for motion generation as it preserves spatial and temporal details through skip connections and hierarchical feature re-

construction.

The Conv1D block follows a typical structure, where input features pass through multiple 1D convolutional layers with kernel size k, stride s, and padding. These layers are followed by activation functions and normalization layers to ensure stable feature transformation. Additionally, feature-wise linear modulation (FiLM) is applied by projecting the Clip-based sentence-level text embedding along with the timestep embedding into scale and shift parameters. This enables a consistent condition-based linear mapping to be applied to each frame's pose embedding. However, since the same timestep-level noise perturbation is applied across the entire motion sequence, this approach may not fully capture the distinct semantics present in individual frames.

Furthermore, to mitigate the sensitivity of Group Normalization to sequence padding, motion embeddings are structured to enable feature-wise normalization, improving performance on datasets with diverse motion lengths.

EABPA Weight Selection EABPA maintains two learnable weight matrices $\mathbf{W}_{\text{pose}}, \mathbf{W}_{\text{vel}} \in \mathbb{R}^{m \times d_j}$, where m is the number of emotion classes and d_j is the joint feature dimension. Given an emotion label $y^i \in \{1, \ldots, m\}$, EABPA selects the corresponding row via index lookup:

$$\mathbf{w}_{\text{pose}}[y^i],\ \mathbf{w}_{\text{vel}}[y^i] \in \mathbb{R}^{1 \times d_j}, \tag{4}$$

which is then passed through a fully connected layer followed by a sigmoid activation to produce the spatial attention vectors $\mathbf{A}_{\text{pose}}$ and $\mathbf{A}_{\text{vel}}$. Each emotion class independently learns its own body-part attention weights through standard gradient descent, enabling fine-grained emotion-specific modulation of pose and velocity features.

Distinction Between $\mathcal{L}_{\text{clus}}$ and $\mathcal{L}_{\text{emo}}$ Although both losses encourage emotion-discriminative representations, they operate at different levels and serve complementary roles. $\mathcal{L}_{\text{emo}}$ is an output-level classifier-guidance loss applied to the generated motion $G_e(\cdot)$, providing global emotion supervision via a pre-trained classifier $\mathcal{R}$. $\mathcal{L}_{\text{clus}}$ is an intermediate-level contrastive metric loss applied to the EABPA features $\tilde{\mathbf{F}}^i_*$, enforcing compact and well-separated clusters in the latent space for each emotion class. As shown in the ablation study (Table 4 of the main paper), removing either loss independently degrades ERA, confirming that the two losses address distinct aspects of emotion-aware learning.

References

[1] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 6

[2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1

[3] Shihao Xu, Jing Fang, Xiping Hu, Edith Ngai, Wei Wang, Yi Guo, and Victor C. M. Leung. Emotion recognition from gait analyses: Current research and future directions. *IEEE Transactions on Computational Social Systems*, 11(1):363–377, 2024. 3

[4] Tan Yu, Jingjing Wang, Jiawen Wang, Jiamin Luo, and Guodong Zhou. Towards emotion-enriched text-to-motion generation via LLM-guided limb-level emotion manipulating. In *ACM Multimedia 2024*, 2024. 1