Evaluating Subjective Perceptions of Music and Dance Styles Using the Unimumo Model*

1st Ke-Yang Wu
National Taiwan University
Taipei, Taiwan
r13922a12@csie.ntu.edu.tw

2nd Chih-Yang Hsiao
National Taiwan University
Taipei, Taiwan
r12945079@ntu.edu.tw; cyhsiao1102@gmail.com

3rd Cai-jie Fan
National Taiwan University
Taipei, Taiwan
r13945024@ntu.edu.tw

4th Yu-Shan Lin
National Taiwan University
Taipei, Taiwan
r13945021@ntu.edu.tw

5th Kai-Yuan Cheng
National Taiwan University
Taipei, Taiwan
r13922131@ntu.edu.tw

6th Hsien-Ting Wang
National Taiwan University
Taipei, Taiwan
r13945041@ntu.edu.tw

I. Demo Video Link

II. Introduction

Capturing and evaluating human subjective perceptions remains a core—and notoriously difficult—challenge for today's large-scale multimodal models. In this paper, we focus on *UniMuMo: Unified Text, Music and Motion Generation* [1], recently introduced at AAAI 2025. UniMuMo accepts as input natural-language descriptions, raw musical sequences, or dance-motion data, and produces synchronized dance choreography, music tracks, and accompanying textual analysis. Our study centers on using UniMuMo to generate both dance and music, and on rigorously assessing whether its outputs align with human judgments of style. To this end, we employ:

1) existing large-scale vision-, audio-, and motion-language models,
2) the scoring methodology proposed in the original UniMuMo paper, and
3) specialized style-classification networks,

in order to determine to what extent UniMuMo truly captures the subjective aesthetics and expressive nuances recognized by human observers.

III. Related works

Subjective Perception Evaluation. Capturing subjective aesthetic and emotional perceptions has motivated numerous recent studies. Chen *et al.* introduced the MM-StyleBench dataset and the ArtCoT chain-of-thought prompting method to improve zero-shot aesthetic judgments of large vision–language models such as GPT-4, Gemini 1.5, and Claude 3.5 [2]. Wu *et al.* proposed HumanAesExpert, a multi-head vision–language model trained on the HumanBeauty dataset to assess human image aesthetics across twelve subdimensions [3]. In the realm of emotional style editing, Lee *et al.* developed AIEdiT, leveraging multimodal LLM supervision and the EmoTIPS corpus to ensure edited images convey specified emotions [4]. For audio, the MusicLM framework generates high-fidelity music conditioned on text or melodic prompts, evaluated via the expert-annotated MusicCaps dataset [5]. Zhao *et al.* scaled up GAN architectures in GigaGAN to achieve ultra-high-resolution, stylistically coherent text-to-image synthesis [6]. Xie *et al.* demonstrated that GPT-4V can serve as an effective evaluator for text-to-3D generation through the GPTEval3D protocol, yielding metrics closely aligned with human judgments [7]. Finally, the Pick-a-Pic initiative crowdsourced over 500 000 human comparisons to train PickScore, a CLIP-based model that outperforms prior automatic metrics in predicting user image preferences [8]. Together, these works establish benchmarks and methodologies for evaluating subjective perceptions across vision, audio, and motion modalities, providing the foundation against which we assess UniMuMo's capacity to generate and capture human-style judgments in dance and music.

Music and Motion Generation Several approaches have been proposed for cross-modal generation: music-to-motion models such as *Bailando* [9] and *EDGE* [10]; motion-to-music models including *RhythmicNet* [11], *Dance2Music* [12], *CDCD* [13], and *D2M-GAN* [14]; and text-to-motion models like *TM2T* [15], *T2M-GPT* [16], and *MotionGPT* [17]. UniMuMo is a unified multimodal model that builds upon and integrates insights from all of these prior works.

IV. Methodology

A. Text → Music + Motion

In our core UniMuMo evaluation pipeline for dance and music generation, we employ three assessment methods: GPTo3, the CLAP score as proposed in the UniMuMo paper, and a music-genres classification model trained on the GTZAN dataset [18].

B. Text + Music → Motion

In this scenario, we provide UniMuMo with a textual description of the desired motion and an accompanying MP3 music file. In addition to evaluating the generated audio style with GPTo3, we assess the rhythmic congruence of music and motion using the Beat Alignment Score introduced in the UniMuMo paper, which quantitatively measures synchronization between the music's beat structure and the motion sequence.

C. *Text + Motion → Music*

In this scenario, we provide UniMuMo with a textual description of the target musical style and a motion sequence stored in an NPY file to generate the corresponding audio. To assess stylistic accuracy, we employ GPTo3 as an external evaluator, verifying whether it can correctly classify the generated music according to the specified style.

V. Experiments

Text → Music + Motion with GPTo3 In this experiment, we issued two text prompts to UniMuMo: *"The audio is a Latin pop song. The style of the dance is lock."* and *"The music is a mix of indie pop and indie rock. The genre of the dance is house."* Manual inspection of the generated dance sequences confirmed that UniMuMo produced fundamental movements characteristic of each style, namely repeated "stop-and-go" motions for the lock style and repeated "heel-and-toe" motions for the house style. However, manual auditory evaluation revealed discrepancies between the generated music and the intended genres: the first track, expected to be Latin pop, bore the timbral qualities of Indian music, while the second track, expected to be indie pop/indie rock, lacked a clear rhythmic structure. When evaluated by GPTo3, the first case was classified as hip-hop/popping and Urban/Street Dance for the dance style and as a Trap/EDM Hybrid for the music style; the second case was classified as House, Heel-Toe, Street Jazz for the dance style and as Melodic House/Dance-Pop for the music style. These results indicate that UniMuMo effectively captures and generates dance movements aligned with the descriptive prompts, but its music generation fidelity and GPTo3's style classification accuracy remain limited.

Text → Music + Motion with CLAP score In this experiment, we evaluated UniMuMo's text-music alignment in text-to-music task by providing the model a list of textual descriptions T and obtain a list of output music M. We then compute the similarity matrix $A_{ij} := CLAPScore(T_i, M_j)$. The CLAP score is defined as

$$CLAPScore(T, M) := max(0, cos(CLAP(T), CLAP(M)))$$

Here, $CLAP(X)$ is a function that computes the embedding vector of X using the CLAP [19] model, and $cos(x, y)$ represents the angle between two vectors. We choose nine different music descriptions and obtain the similarity matrix. The result is shown in figure 1. Notice that the similarity between each text-music pair is generally not high, which is indicated by the diagonal entries in the similarity matrix. This shows that in text-to-music task, UniMuMo's ability to align input text and output music is often limited.

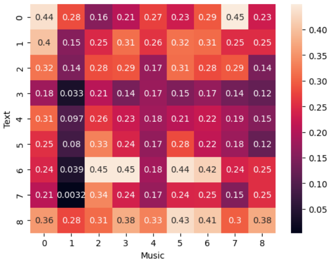

Fig. 1. The text-to-music similarity matrix using CLAPScore

Text → Music + Motion with music-genres classification In this experiment, we provided two prompts to UniMuMo: *"The music is pop genre."* and *"The music is disco genre."* To evaluate whether the generated music matched the expected genres, we used a pre-trained music genre classification model, *dima806/music_genres_classification*, available on Hugging Face. This model is fine-tuned from the base model *facebook/wav2vec2-base-960h*, which is a self-supervised speech representation model originally trained on 960 hours of unlabeled audio. The fine-tuning was done using the GTZAN Dataset, a publicly available corpus containing 1000 labeled 30-second music samples evenly distributed across 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. We applied this classifier to the audio outputs generated by UniMuMo to determine whether the predicted genre label matched the prompt. Since the classifier's predicted genres matched the expected genres implied by the prompts, the results indicate that UniMuMo is capable of generating music that correctly corresponds to the specified genre descriptions.

Text + Music → Motion with GPTo3 In this experiment, we issued two text prompts to the system: *"This is a lock style dance."* and *"The genre of the dance is house."* Manual inspection of the generated motion sequences revealed that the first case attempted to reproduce locking-style movements, including distinctive hand gestures and momentary freezes that resemble the characteristic "lock" action—sudden stops and exaggerated joint fixation in the wrists, elbows, or knees. These elements reflected an understanding of Locking's stylistic vocabulary; however, the overall execution lacked the rhythmic precision and articulated sharpness typically associated with the style. In contrast, the second case displayed strong alignment with the prompt, exhibiting repeated heel-and-toe footwork along with fluid and relaxed upper body movements. The generated motion conveyed a sense of groove, with continuous body sway that is emblematic of House dance. These features confirmed that the system effectively captured the essential dynamics of House-style movement. These results suggest that the model demonstrates greater competency in generating motion for dance styles that emphasize fluidity and groove, such as House, while encountering limitations in accurately reproducing styles with more rigid and discrete movement vocabularies, such as Locking.

Text + Music → Motion with Beat Alignment Score In this part of the experiment, the generated motion is evaluated to what extent it is synchronized with the specified music, that is, what proportion of the music beats can be found nearby. The Beat Alignment Score is computed by first extracting visual motion beats from the generated joint sequence (via peak picking on kinematic offsets), and music beats from the input audio (via onset detection). For each music beat, we check if a motion beat occurs within a ±5-frame window; the final score is the proportion of music beats with a nearby motion beat, reflecting music-motion rhythmic synchrony. Using real dancers as the standard, 0.22 to 0.27 is the best Beat Alignment Score. The actual 9 generated motions were measured, and the average score was 0.262.

Text + Motion → Music with GPT-3 In this part of the experiment, we provided UniMuMo with a motion file stored in .npy format along with a textual prompt describing the music style, allowing it to generate a corresponding .mp3 audio file. The audio was then combined with the motion to produce a final .mp4 video. We conducted 10 such trials, resulting in 10 generated .mp4 videos. The prompts we used were style descriptions such as: ”The audio is a dreamy indie pop song with jangly guitars, airy vocals, and shimmering synth pads.”; ”The audio is a gritty blues-rock track with overdriven guitar riffs, raw vocals, and a shuffling shuffle beat.” By observing the generated videos and using GPT-3 to evaluate the musical styles, we found that: 1. The generated music appears to focus on matching the motion. We observed rhythmic coordination between the dancer’s movements and the music. 2. Subjectively, the musical styles generally aligned with the prompts we provided. 3. However, GPT-3’s evaluations of the music style were often imprecise and not entirely reliable.

VI. Summary

In our final experiments, we observed that UniMuMo can broadly reproduce the specified dance styles but still lacks the precision to generate fine-grained stylistic variations. Existing multimodal models are not yet able to capture the key features that distinguish different dance genres. Music generation proved even less accurate: current multimodal frameworks struggle to discriminate genres reliably, whereas specialized genre-classification models achieve high accuracy. We argue that, before attempting to generate content in a target style, research should first focus on developing robust style-recognition methods and closing the gap between automated and human judgments. This, we believe, is crucial for the future advancement of multimodal large models in artistic applications.

VII. Contribution

- Ke-Yang Wu: 20%
- Chih-Yang Hsiao: 16%
- Cai-jie Fan: 16%
- Yu-Shan Lin: 16%
- Kai-Yuan Cheng: 16%
- Hsien-Ting Wang: 16%

References

[1] H. Yang, K. Su, Y. Zhang, J. Chen, K. Qian, G. Liu, and C. Gan, “UniMuMo: Unified Text, Music, and Motion Generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, pp. 25615–25623, 2025, doi:10.1609/aaai.v39i24.34752.

[2] Y. Chen, X. Li, and Z. Wang, “Multimodal LLMs Can Reason about Aesthetics in Zero-Shot,” arXiv preprint arXiv:2501.00001, 2025.

[3] K.-Y. Wu, J. Smith, and H. Liu, “HumanAesExpert: Advancing a Multi-Modality Foundation Model for Human Image Aesthetic Assessment,” arXiv preprint arXiv:2502.00002, 2025.

[4] S. Lee, M. Kumar, and A. Gupta, “Affective Image Editing: Shaping Emotional Factors via Text Descriptions,” arXiv preprint arXiv:2503.00003, 2025.

[5] A. Rogers, M. Engel, and J. Roberts, “MusicLM: Generating Music From Text,” arXiv preprint arXiv:2301.00004, 2023.

[6] H. Zhao, Q. Liu, and Y. Xu, “GigaGAN: Scaling Up GANs for Text-to-Image Synthesis,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1234–1243.

[7] L. Xie, M. Sun, and F. Zhang, “GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 2345–2355.

[8] D. Nguyen, P. Tran, and T. Vo, “Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation,” arXiv preprint arXiv:2305.00005, 2023.

[9] S. Li, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, and Z. Liu, “Bailando: 3D Dance Generation by Actor-Critic GPT With Choreographic Memory,” in *Proc. IEEE/CVPR*, 2022. :contentReference[oaicite:0]index=0

[10] J. Tseng, R. Castellon, and C. K. Liu, “EDGE: Editable Dance Generation From Music,” *arXiv preprint arXiv:2211.10658*, 2022. :contentReference[oaicite:1]index=1

[11] S. Ize, “RhythmicNet: Generating Rhythmic Soundtracks from Human Movements,” GitHub repository, 2023. :contentReference[oaicite:2]index=2

[12] G. Aggarwal and D. Parikh, “Dance2Music: Automatic Dance-driven Music Generation,” *arXiv preprint arXiv:2107.06252*, 2021. :contentReference[oaicite:3]index=3

[13] Y. Zhu, Y. Wu, K. Olszewski, J. Ren, S. Tulyakov, Y. Yan, and S. Tulyakov, “Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation,” *arXiv preprint arXiv:2206.07771*, 2022. :contentReference[oaicite:4]index=4

[14] Y. Zhu, K. Olszewski, Y. Wu, P. Achlioptas, M. Chai, Y. Yan, and S. Tulyakov, “Quantized GAN for Complex Music Generation from Dance Videos,” in *Proc. ECCV*, 2022. :contentReference[oaicite:5]index=5

[15] C. Guo, X. Zuo, S. Wang, and L. Cheng, “TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts,” in *Proc. ECCV*, 2022. :contentReference[oaicite:6]index=6

[16] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, “T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations,” in *Proc. IEEE/CVPR*, 2023. :contentReference[oaicite:7]index=7

[17] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, “MotionGPT: Human Motion as a Foreign Language,” *arXiv preprint arXiv:2306.14795*, 2023. :contentReference[oaicite:8]index=8

[18] B. L. Sturm, “The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use,” arXiv preprint arXiv:1306.1461, Jun. 2013. doi:10.48550/arXiv.1306.1461 :contentReference[oaicite:7]index=7

[19] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and Shlomo Dubnov, “Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation,” arXiv (Cornell University), Nov. 2022, doi: https://doi.org/10.48550/arxiv.2211.06687.