

MusicInfuser: Making Video Diffusion Listen and Dance

Supplementary Material

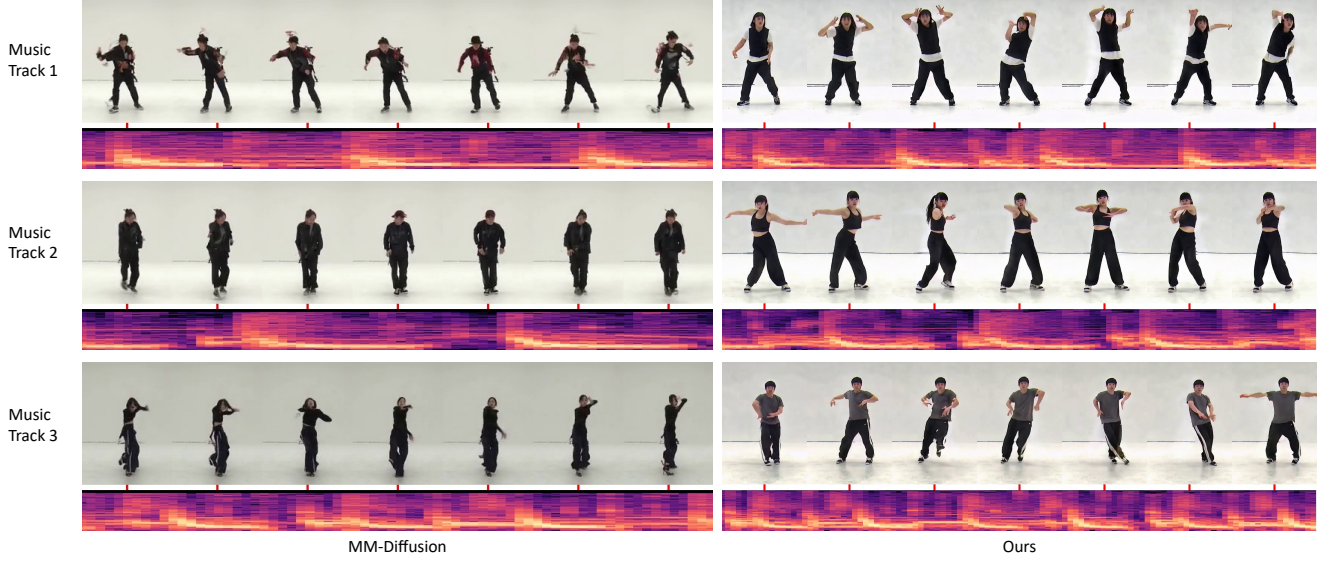


Figure 10. Comparison of audio-driven generation with MM-Diffusion [39]. Our method produces fewer artifacts (shown in the first and third rows), while generating more realistic dance videos with more natural movements (first row) and more dynamic motion (second and third rows). Note that we use the same music track for each row, and the spectrogram is stretched for MM-Diffusion since we generate longer videos. For our method, we use the fixed caption “a professional dancer dancing ...” across all music tracks.

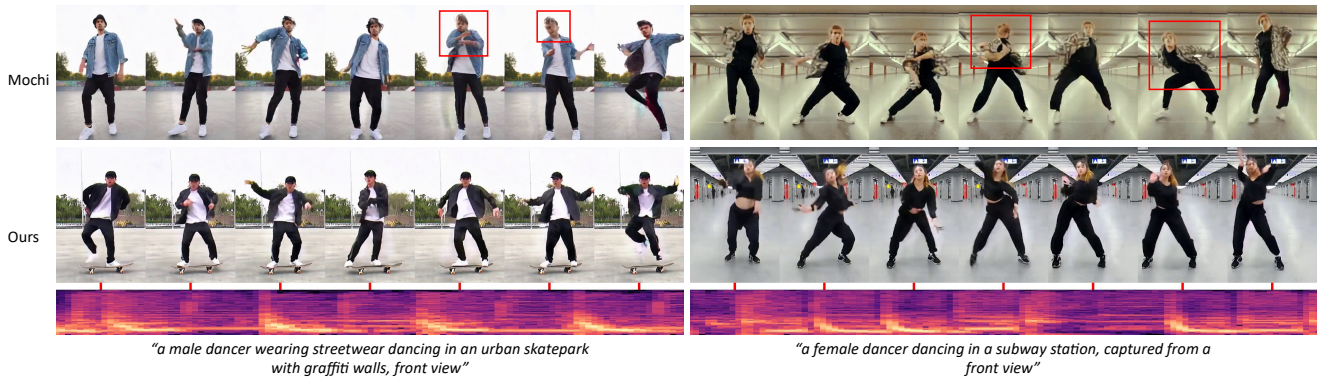


Figure 11. MusicInfuser infuses listening capability into the text-to-video model (Mochi [46]), while preserving prompt adherence and improving overall consistency and realism.

A. Video Results

We present the flattened video results along the time axis and the corresponding spectrograms in the main paper. However, our frame sampling rate does not exceed the Nyquist frequency for the general musical beat, causing the movement to appear slower. Therefore, we encourage readers to view the supplementary video.

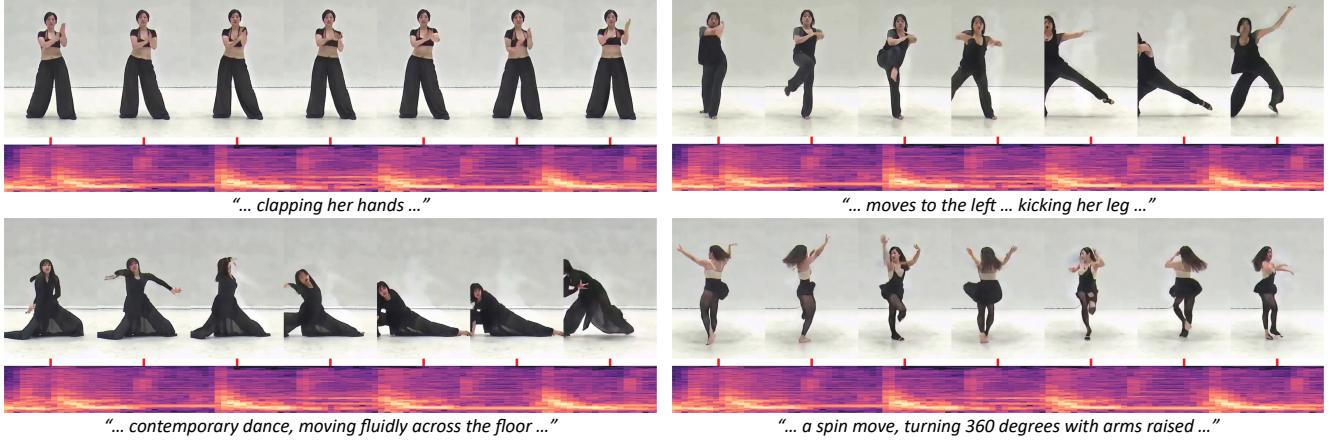


Figure 12. Text-based dance control.

Layer Selection Strategy	Style Alignment	Beat Alignment	Body Representation	Movement Realism	Choreography Complexity	Dance Quality Average	Imaging Quality	Aesthetic Quality	Overall Consistency	Video Quality Average	Overall
Layer Adaptability	7.56	8.89	7.16	8.24	7.90	7.95	9.60	7.87	9.39	8.95	8.33
Evenly Distributed Layers	7.31	8.81	7.28	7.70	7.96	7.81	9.33	7.78	9.04	8.72	8.15
First Layers	7.25	8.82	6.86	7.37	8.05	7.67	9.66	7.91	9.27	8.95	8.15
Middle Layers	7.91	8.87	6.74	7.83	7.98	7.86	9.21	7.97	9.20	8.79	8.21
Last Layers	7.52	8.81	7.01	7.47	8.00	7.76	9.45	7.73	9.14	8.77	8.14
All Layers	7.49	8.53	6.72	8.16	7.85	7.75	9.33	7.99	9.11	8.81	8.15

Table 5. Evaluation of layer selection strategies using VideoLLaMA 2 [14].

Model	Style Alignment	Beat Alignment	Body Representation	Movement Realism	Choreography Complexity	Dance Quality Average	Imaging Quality	Aesthetic Quality	Overall Consistency	Video Quality Average	Overall
Full (Ours)	7.56	8.89	7.16	8.24	7.90	7.95	9.60	7.87	9.39	8.95	8.33
No ZICA Layer Selection	7.31	8.81	7.28	7.70	7.96	7.81	9.33	7.78	9.04	8.72	8.15
No Higher Rank	7.37	8.76	6.86	7.75	7.98	7.74	9.55	7.94	9.49	8.99	8.21
No LoRA	7.48	8.62	7.02	7.53	7.95	7.72	9.43	8.08	9.36	8.96	8.18
No Beta-Uniform Schedule	8.04	9.07	6.35	7.88	7.91	7.85	9.17	7.85	9.37	8.80	8.21
Feature Addition	7.62	8.90	6.78	7.97	7.88	7.83	9.44	7.88	9.31	8.88	8.22

Table 6. Ablation study. Feature addition denotes that we spatially expand the audio feature and add it to the corresponding frame. We use VideoLLaMA 2 [14] for the evaluation.

B. Text-Driven Choreography

To demonstrate text-based control of dance sequences, we conducted additional experiments in which text prompts guide the dance dynamics (clapping, kicking, spinning, etc.). The results are shown in Fig. 12. This demonstrates that MusicInfuser can also generate diverse music-synchronized videos that follow the motion directions specified in the text prompt.

C. Dance Difficulty Control

We demonstrate difficulty control of the choreography in Fig. 13, which is achieved using the same seed and music but with prompts of varying specificity. For basic dance, we use the general prompt “a professional dancing in a studio with a white backdrop.” For styled dance, we additionally specify the dance genre but use “basic dance setting,” and for advanced, we change it to “advanced dance setting.”

D. Human Evaluation Protocol

For each test music track [33], we conducted fully anonymized A/B testing. We asked 33 participants to evaluate video quality, music-dance alignment, motion realism, and choreography complexity. The following are examples of the questionnaire items:

1. Which video has higher visual quality?
2. Which video’s dance aligns better with the music?

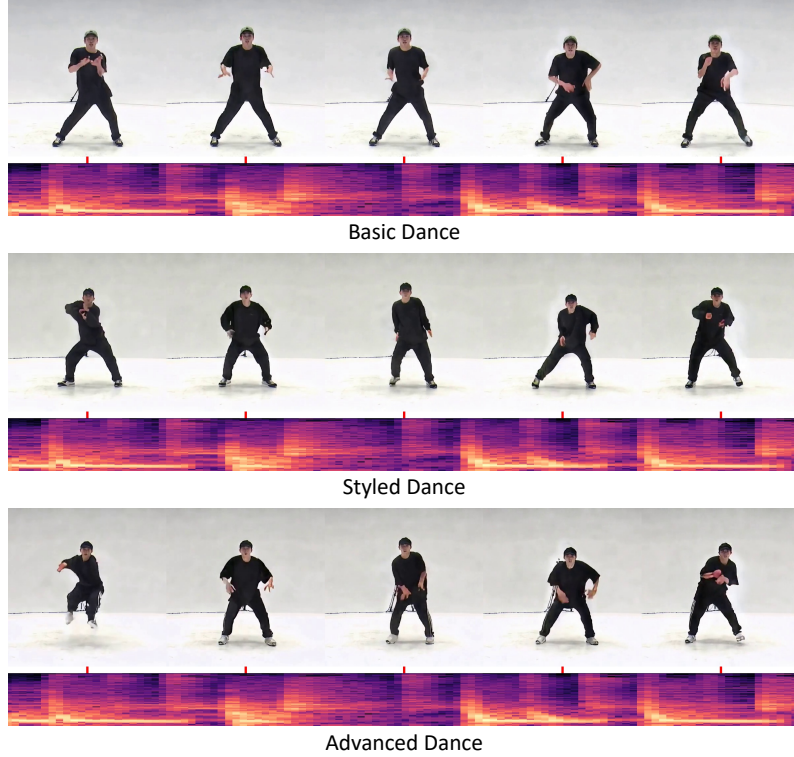


Figure 13. Changes in the complexity of choreography.

3. Which video’s motion is more realistic?
4. Which video’s dance is more complex?

E. Limitations

Although our method adds listening capability to text-to-video models and improves dance generation, some properties such as style capture from the prompt and imaging quality are bounded by the capabilities of the underlying models. Additionally, our method inherits some problems from text-to-video models. Sometimes, fine details such as fingers and faces fail to be generated properly, especially when our model synthesizes dance videos with fast movements. Furthermore, our model is easily misled by the silhouette of the dancers, meaning that under the same silhouette, it may merge or swap the positions of body parts, which is also a problem in the base model. We include some examples of failure cases in Fig. 19.

F. Additional Qualitative Analysis

We show more music-and-text-to-video generation examples in Fig. 20. Fig. 10 presents a side-by-side comparison with MM-Diffusion [39]. Unlike MM-Diffusion, which generates shorter videos with limited style control, MusicInfuser produces longer sequences with both musical synchronization and prompt-based style control, while improving the overall consistency of the video and reducing artifacts. We show a comparison with Mochi [46] in Fig. 11. Note that Mochi is not able to hear the music. Compared to Mochi, MusicInfuser produces more consistent human forms, fewer visual artifacts, and more fluid, realistic movements. Our method adds music responsiveness while maintaining or improving video consistency. We also compare with EDGE [47], a state-of-the-art skeleton-based dance generation model. EDGE performs 3D skeleton generation requiring 3D pose reconstruction, whereas our method performs direct video synthesis. Fig. 16 shows a visual comparison under the same music track using our method and EDGE. This demonstrates that our approach captures nuances that skeleton-based methods cannot represent, such as hair dynamics, clothing motion, a flexible backbone, and hand articulation.

We present qualitative results of our ablation study in Fig. 14 and Fig. 15. Our full model successfully generates consistent body shapes that align with the music while preserving prior knowledge without introducing significant artifacts.

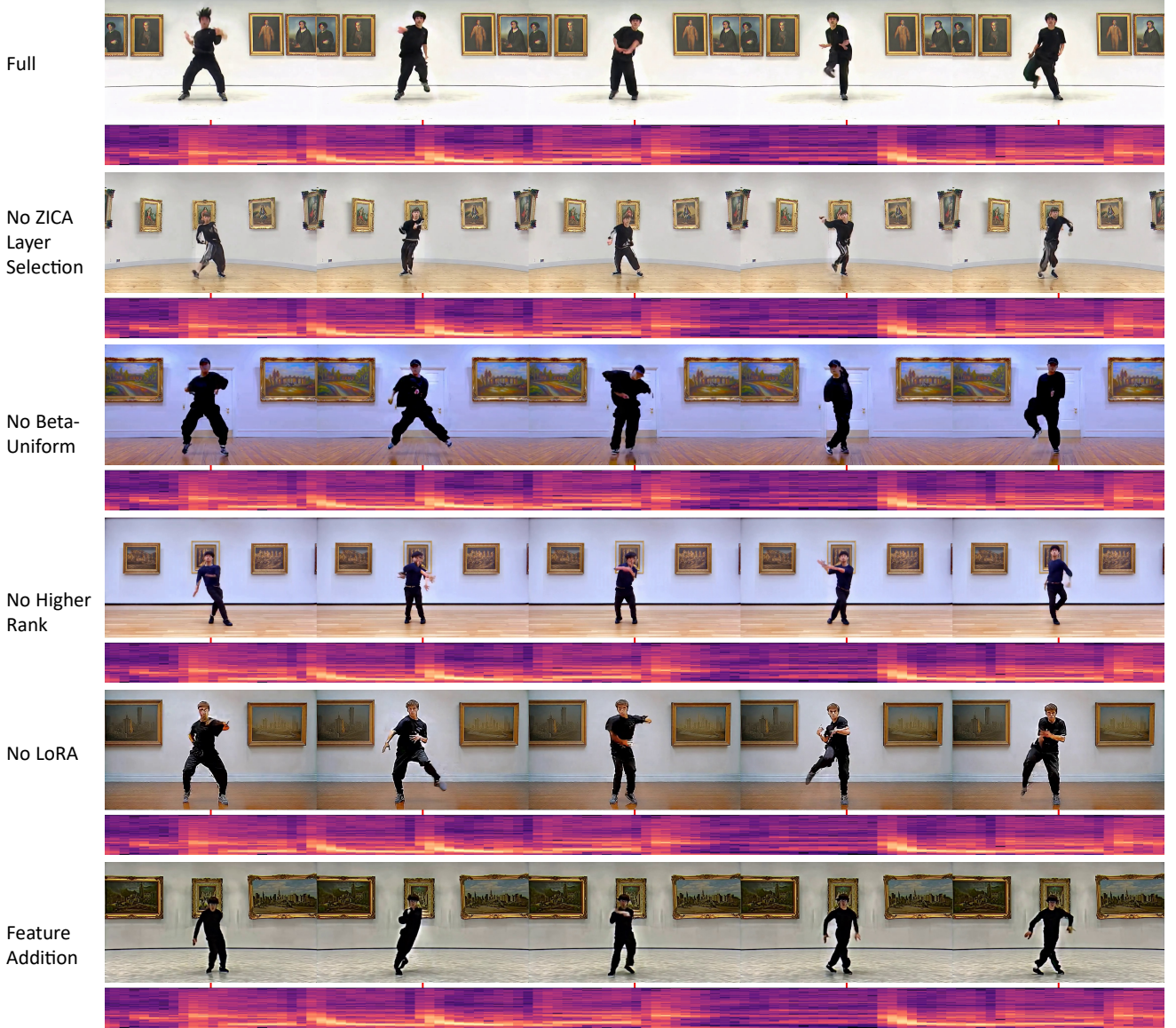


Figure 14. Ablation study. The prompt is set to “a male dancer dancing in an art gallery with some paintings, captured from a front view”. The seed and music are set the same across all methods.

G. Additional Quantitative Analysis

Similar to the layer selection baselines and ablation studies in the main paper using Qwen3-Omni [50], we show evaluation using VideoLLaMA 2 [14] in Tables 5 and 6. The full model achieves the highest score. Using a higher rank for LoRA contributes substantially to movement realism, while our Beta-Uniform scheduling improves body representation. The naive feature addition baseline, where instead of using the ZICA adapter we simply spatially expand the audio feature and add it to the corresponding frame, performs worse than our approach on most metrics, confirming the effectiveness of our ZICA strategy. In Table 7, we present comparisons between MM-Diffusion and our method, both trained on the identical AIST++ training dataset without in-the-wild data. This shows that our model trained on the AIST dataset alone already surpasses MM-Diffusion, while in-the-wild data further enhances generalization capability.

In Table 4 in the main paper, we show the trade-off between style capture and creative interpretation of the prompt depending on the base prompt ratio, meaning how frequently we replaced the prompt with the basic prompt.

Additionally, we present commonly used intrinsic metrics from related work, BeatAlign and kinetic diversity [33, 41],

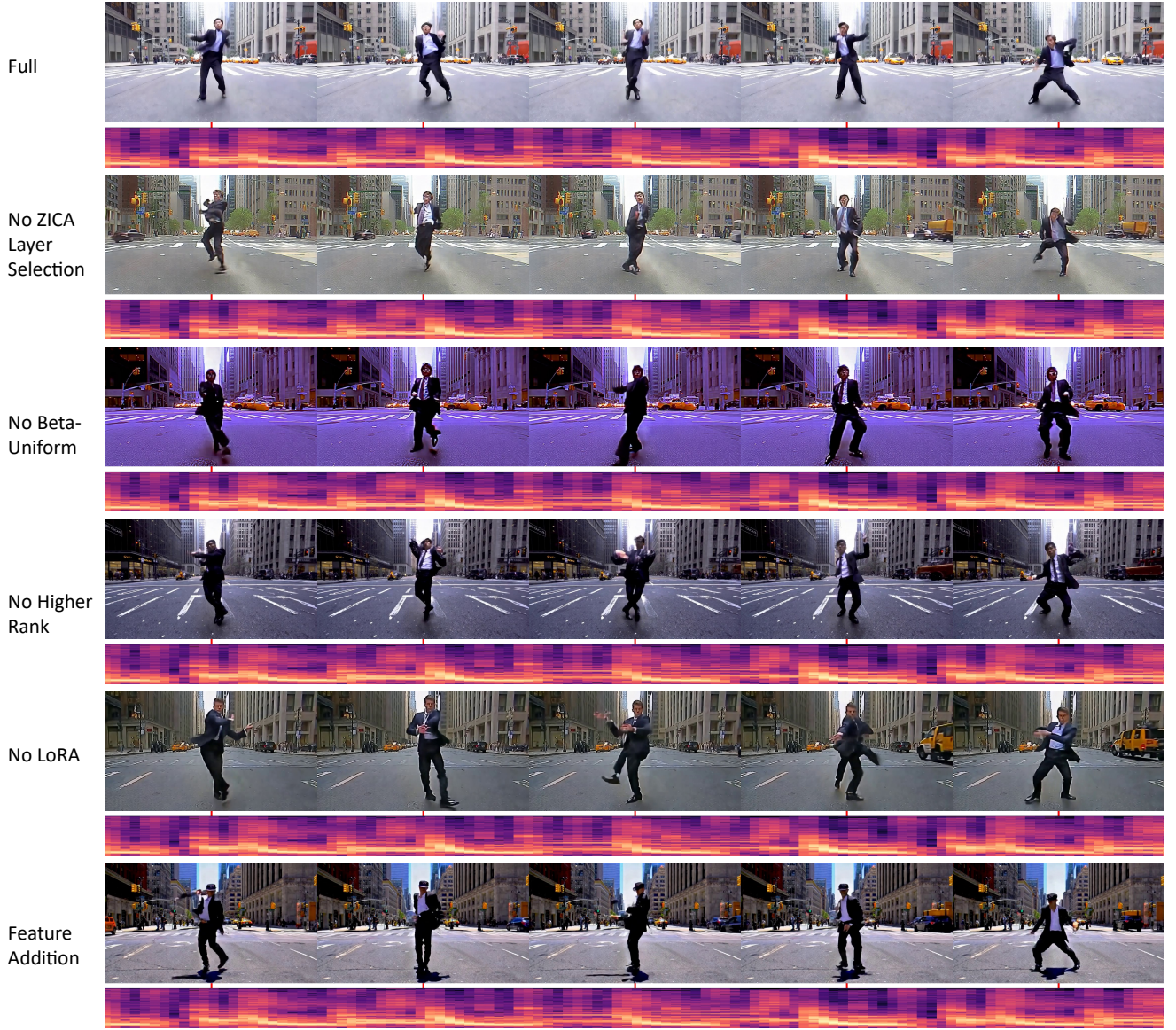


Figure 15. Ablation study. The prompt is set to “a male dancer wearing a suit dancing in the middle of a New York City, captured from a front view”. The seed and music are set the same across all methods.

measured after extracting 2D pose sequences from generated videos. Table 8 shows these metrics, demonstrating a comparable score to the AIST test set and superior scores compared to the baselines.

H. Layer Adaptability

The imaging and aesthetic quality of the base model [46] is presented in Fig. 17. This is analyzed with STG [26], an inference-time technique, and the score is calculated with VBench [25]. Based on the imaging quality, which is highly related to the structure and noisiness of the video samples, we select the top 16 out of 48 layers in terms of imaging quality.

I. Beta-Uniform Scheduling

The visualization of the Beta-Uniform scheduling strategy is shown in Fig. 18.

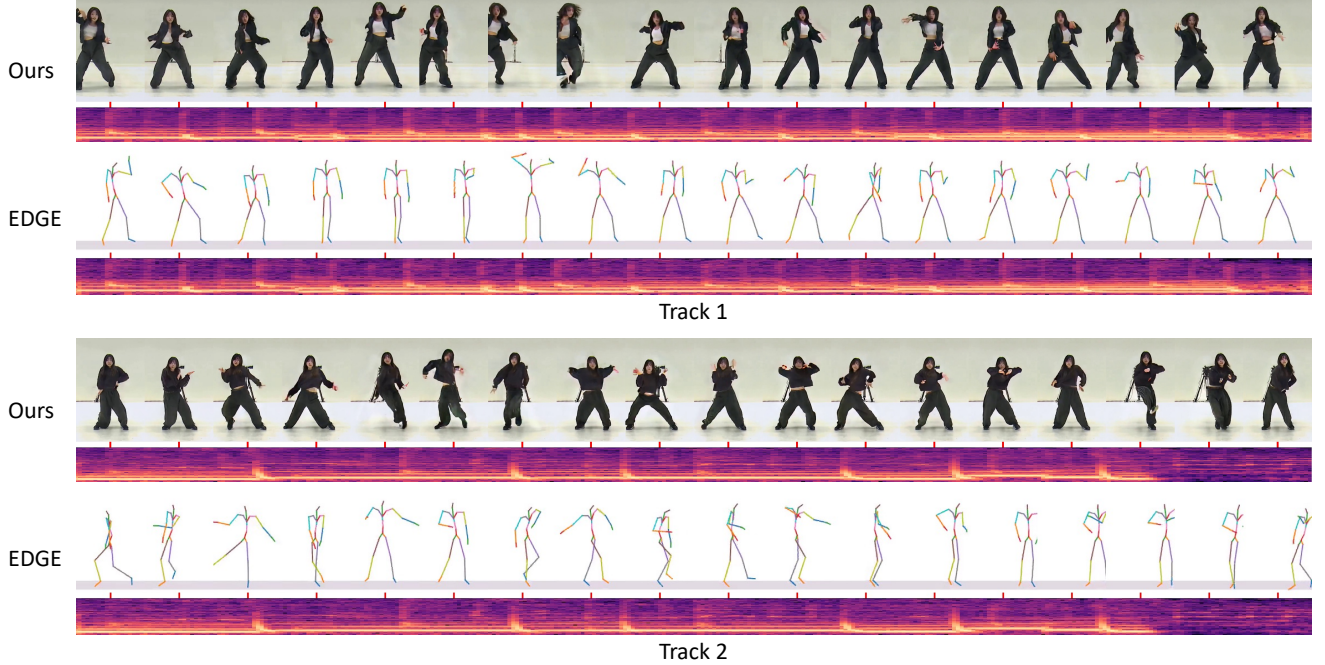


Figure 16. Comparison with EDGE [47]. Note that the poses of our method and EDGE need not align.

Model	Style Alignment	Beat Alignment	Body Representation	Movement Realism	Choreography Complexity	Imaging Quality	Aesthetic Quality	Overall Consistency
MM-Diffusion [39]	7.16	8.56	5.52	7.05	7.53	8.94	6.52	8.38
Ours (Only AIST)	7.83	9.10	6.89	8.58	7.96	9.55	8.02	9.75

Table 7. Comparisons between MM-Diffusion [39] and our method, both trained on the AIST++ training dataset.

Method	BeatAlign \uparrow	Dist $_k\uparrow$
AIST Dataset (GT)	0.2448	9.027
MM-Diffusion [39]	0.1553	2.126
Mochi [46]	0.1976	8.886
MusicInfuser (Ours)	0.2432	9.849

Table 8. BeatAlign and kinetic diversity metrics based on 2D poses.

J. Test Music Tracks

For evaluating our method, we use music tracks that are set aside from the training set [48], following AIST++ [33]. The full list of test music codes is provided in Table 9.

K. Prompts

As mentioned in our main paper, we use a proper prompt format and base prompt for AIST [48]. The full list is shown in Table 10. Note that since we use VideoChat2 [32] to label YouTube videos, we have only the base prompt for that dataset. We also provide a predefined set of prompts in Table 11 that is used to generate samples for the evaluation, ultimately resulting in $10 \times 10 = 100$ videos per model configuration. The system prompts for VideoLLaMA 2 [14] and Qwen3-Omni [50] used for evaluation are provided in Table 12.

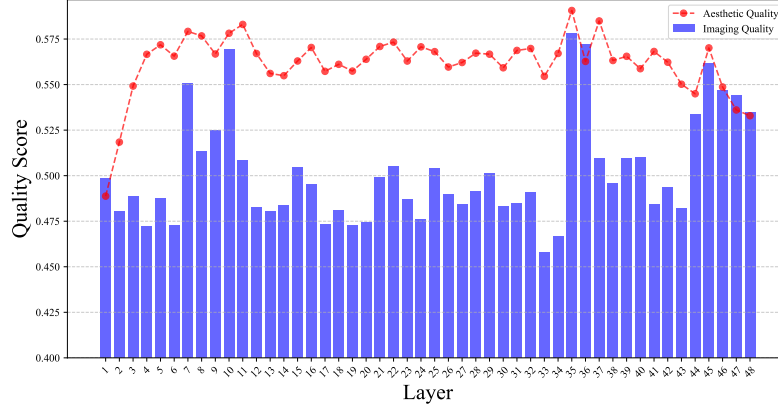


Figure 17. Layer adaptability graph from [26], showing imaging and aesthetic quality.

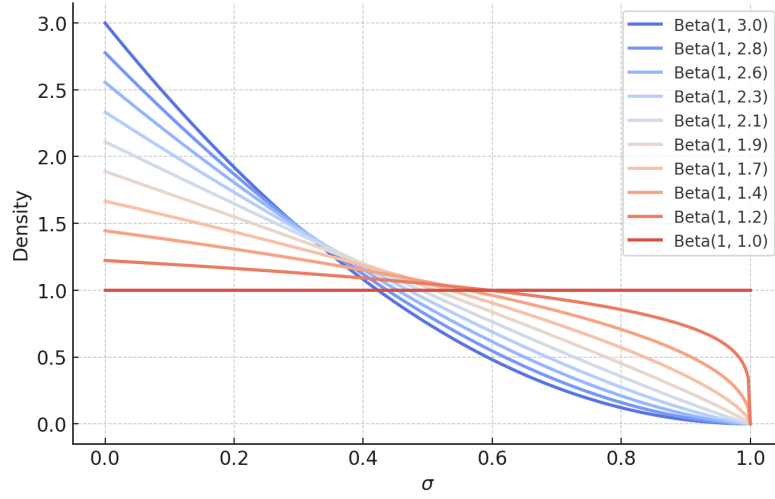


Figure 18. Beta distributions.

Test Music Code	Genre
mLH4	LA style Hip-hop
mKR2	Krump
mBR0	Break
mLO2	Lock
mJB5	Ballet Jazz
mWA0	Waack
mJS3	Street Jazz
mMH3	Middle Hip-hop
mHO5	House
mPO1	Pop

Table 9. List of test music codes with corresponding dance genres.

L. Concurrent Work

Several concurrent approaches have emerged alongside our research that address related challenges. Notable among these is VideoJAM [11], which enhances motion generation by jointly denoising both the motion maps and the video, an approach

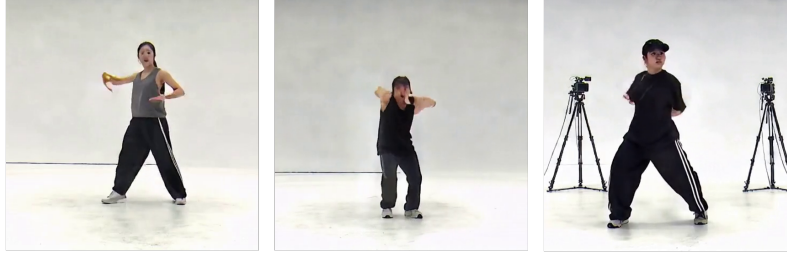


Figure 19. Failure cases. Our model inherits some issues from the base model, such as failing to generate fine details (e.g., fingers and faces) and being fooled by the silhouette of the dancers.

Category	Dataset	Prompt Template
Prompt Format	AIST	{dancers.text} dancing {genre_name} in a {situation_name} setting in a studio with a white backdrop, captured from a {camera_view}
Prompt Format	AIST	a {camera_view} video of {dancers.text} performing {genre_name} choreography against a white background in a {situation_name} scene
Prompt Format	AIST	{dancers.text} executing {genre_name} movements in a minimalist studio space in a {situation_name} setting, shot from a {camera_view}
Prompt Format	AIST	a {genre_name} dance performance by {dancers.text} in a pristine white studio, {camera_view}, {situation_name}
Base Prompt	AIST	a professional dancer dancing in a studio with a white backdrop
Base Prompt	YouTube	a dance video

Table 10. Dance prompt templates categorized by type and dataset, including parameterized formats and simple base prompts.

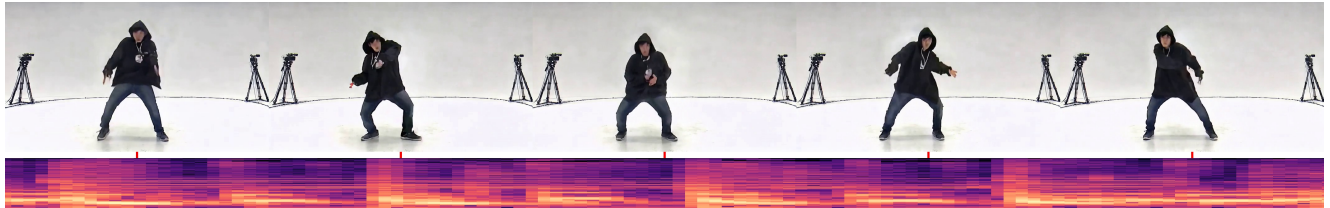
Prompts
a male dancer dancing on a rooftop at sunset, captured from a front view
a female dancer dancing in a subway station, captured from a front view
a male dancer dancing in an art gallery with some paintings, captured from a front view
a female dancer wearing a leather jacket dancing in a studio with a white backdrop, captured from a front view
a male dancer wearing a hoodie dancing in a studio with a white backdrop, captured from a front view
a female dancer wearing a denim vest dancing in a studio with a white backdrop, captured from a front view
a female dancer wearing a Hawaiian dress dancing on Waikiki Beach at sunset with Diamond Head in the background, captured from a front view
a male dancer wearing a suit dancing in the middle of a New York City, captured from a front view
a male dancer wearing a chef’s uniform dancing in a busy restaurant kitchen with flames from the grill behind him, captured from a front view
a female dancer wearing a Renaissance gown dancing in a Venetian masquerade ball with ornate chandeliers overhead, captured from a front view

Table 11. Collection of dance scene prompts with various subjects, attire, and settings.

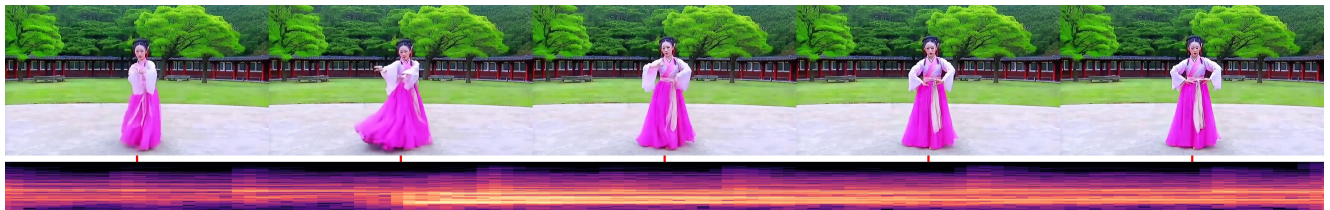
that is orthogonal to ours. Another related line of research is OmniHuman-1 [34], which integrates audio and pose inputs into diffusion models. The application of OmniHuman-1 remains primarily confined to scenarios that do not require much creative movement, relies on a private model, and necessitates full fine-tuning procedures, which distinguishes it from our approach.

Metric	Prompt
Dance Quality	
Style Alignment	Rate the style alignment of the dance to music where: 0 means poor style alignment of the dance to music, 5 means moderate style alignment of the dance to music, and 10 means perfect style alignment of the dance to music. Output only the number.
Beat Alignment	Rate the beat alignment of the dance to music where: 0 means poor beat alignment of the dance to music, 5 means moderate beat alignment of the dance to music, and 10 means perfect beat alignment of the dance to music. Output only the number.
Body Representation	Rate the body representation of the dancer where: 0 means unrealistic/distorted proportions of the dancer, 5 means minor anatomical issues of the dancer, and 10 means anatomically perfect representation of the dancer. Output only the number.
Movement Realism	Rate the movement realism of the dancer where: 0 means poor movement realism of the dancer, 5 means moderate movement realism of the dancer, and 10 means perfect movement realism of the dancer. Output only the number.
Choreography Complexity	Rate the complexity of the choreography where: 0 means extremely basic choreography, 5 means intermediate choreography, and 10 means extremely complex/advanced choreography. Output only the number.
Video Quality	
Imaging Quality	Rate the imaging quality where: 0 means poor imaging quality, 5 means moderate imaging quality, and 10 means perfect imaging quality. Output only the number.
Aesthetic Quality	Rate the aesthetic quality where: 0 means poor aesthetic quality, 5 means moderate aesthetic quality, and 10 means perfect aesthetic quality. Output only the number.
Overall Consistency	Rate the overall consistency where: 0 means poor consistency, 5 means moderate consistency, and 10 means perfect consistency. Output only the number.
Prompt Alignment	
Style Capture	How well does the dance video capture the specific style mentioned in the prompt: '{prompt}'? Rate 0-10 where: 0 means completely missed the style, 5 means some elements of the style are present, and 10 means perfectly captures the style. Output only the number.
Creative Interpretation	Based on the prompt '{prompt}', rate the creativity in interpreting the prompt 0-10 where: 0 means generic/standard interpretation, 5 means moderate creativity, and 10 means highly creative and unique interpretation. Output only the number.
Overall Prompt Satisfaction	Rate the overall prompt satisfaction 0-10 where: 0 means the video fails to satisfy the prompt '{prompt}', 5 means it partially satisfies the prompt, and 10 means it fully satisfies all aspects of the prompt. Output only the number.

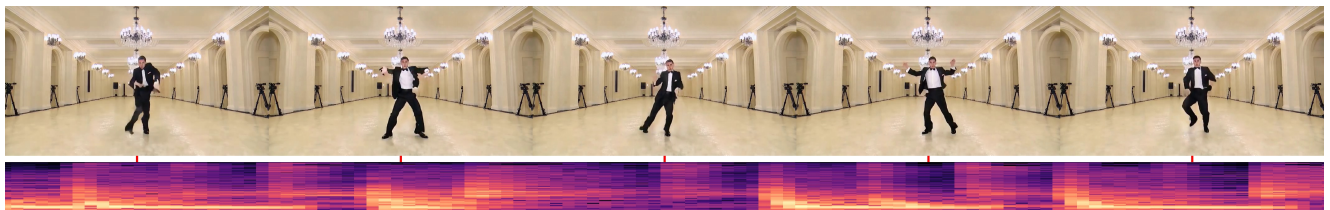
Table 12. System prompts for evaluation



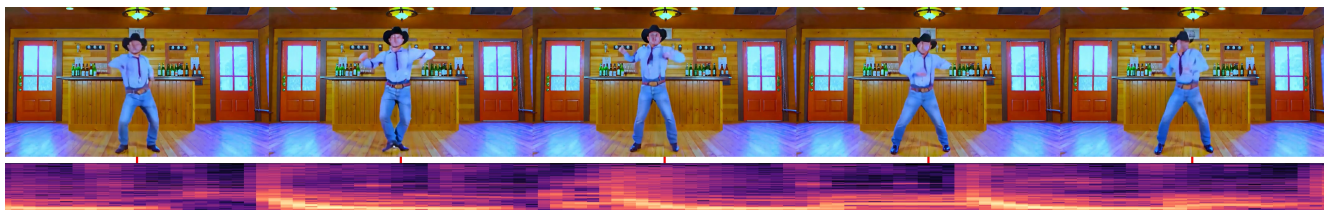
"a male dancer dancing on a rooftop at sunset, captured from a front view"



"a female dancer dancing in a Korean palace garden, front view"



"a male dancer wearing a tuxedo dancing in an elegant ballroom with crystal chandeliers, front view"



"a male dancer wearing a cowboy outfit dancing in a Western saloon with wooden bar and swinging doors, front view"

Figure 20. More music-and-text-to-video generation results.