OmniDance: Multimodal Driven Dance Video Generation with Large-scale Internet Data

Kaixing Yang[1], Jiashu Zhu[2‡], Xulong Tang[5], Ziqiao Peng[1], Xiangyue Zhang[4], Chubin Chen[3], Puwei Wang[1†], Jiahong Wu[2†], Xiangxiang Chu[2], Hongyan Liu[3†], and Jun He[1†]

[1] Renmin University of China, Beijing, China
{yangkaixing,pengziqiao,wangpuwei,hejun}@ruc.edu.cn
[2] AMAP, Alibaba Group, Beijing, China
{zhujiashu.zjs,hongxi.wjh}@alibaba-inc.com, cxxgtxy@gmail.com
[3] Tsinghua University, Beijing, China
liuhy@sem.tsinghua.edu.cn, trubeeen@gmail.com
[4] Wuhan University, Wuhan, China
xiangyuezhang@whu.edu.cn
[5] Malou Tech Inc, Plano, Texas, USA
xulong.tang@maloutech.com

Abstract. Music-driven dance video generation aims to synthesize expressive human motion that is temporally aligned with music while maintaining high visual fidelity. Despite recent progress, existing methods still struggle to generate dance videos that simultaneously exhibit expressive motion and high visual quality. This limitation primarily arises from two factors: *(1) Dataset.* The lack of large-scale and high-quality datasets and effective data collection pipelines specifically tailored for dance video generation; and *(2) Method.* The absence of principled framework-level solutions for effectively integrating music as a complementary conditioning signal into the Video Generation Foundation Models. To address the dataset limitation, we introduce **CIPE-Dance**, a large-scale Internet-sourced dance video dataset, equipped with **C**horeograph **I**nformed text annotations and constructed via a **P**rogressive **E**xpert pipeline. To the best of our knowledge, **CIPE-Dance** is the largest dataset for dance video generation to date, comprising 300k high-quality clips (over 400 hours) and covering diverse dancers, environments, and dance genres. To overcome the method limitation, we propose **OmniDance**, a framework-level recipe for integrating music into a TI2V foundation model without sacrificing its original controllability or visual fidelity. Motivated by the complementary roles of text (low-frequency semantics) and music (high-frequency temporal dynamics), OmniDance co-designs a depth-aware specialization model architecture, an anchored easy-to-hard curriculum learning strategy, and modality-specialized time-dependent CFG strategy, achieving unified TI2V/MI2V/MTI2V generation. Extensive experiments on the **CIPE-Dance** dataset demonstrate that **OmniDance** achieves state-of-the-art performance across TI2V, MI2V, and MTI2V

‡Project leader.
†Corresponding authors.

(a) Music-Text-Image-to-Video (MTI2V) (b) Text-Image-to-Video (TI2V) (c) Music-Image-to-Video (MI2V)

Fig. 1: Given a reference image, OmniDance supports multimodal driven dance video generation, including music only (MI2V, right), text only (TI2V, middle), and both (MTI2V, left). OmniDance not only preserves identity and visual fidelity, but also produces artistically expressive and music-aligned motion.

tasks, while exhibiting robust multimodal integration capability. Project is available at https://github.com/AMAP-ML/OmniDance.

Keywords: Dance Video Generation · Digital Human · AI for Art

1 Introduction

Dance is a vital part of human culture: through choreographed movement, dancers communicate emotion and narrative intent while showcasing the beauty of human motion [3, 41]. In the internet era, dance videos have become a dominant form of visual content on platforms such as YouTube and TikTok, and recent advances in AI-generated content (AIGC) have made high-fidelity video synthesis increasingly accessible [4–6,11,16,19,20,24]. Crucially, dance is jointly governed by *semantic intent* (e.g., style and choreographic goals) and *musical structure* (e.g., rhythm, tempo, and energy evolution) [45], motivating multimodal-driven dance video generation under text, music, and their combination.

In recent years, substantial progress has been made in several dance video related research directions, including talking head generation [27, 29, 30], pose-driven image animation [15, 32], and music-driven 3D dance generation [31, 34]. However, talking head generation does not model the tight correspondence between audio and full-body motion [9,28], pose-driven image animation overlooks the central role of music in dance composition [12, 39]s, and music-driven 3D dance generation is typically trained on small-scale datasets, ultimately constraining their motion generation capacity and generalization to diverse in-the-wild settings [17,42,43]. Despite growing interest, progress in music-driven dance video generation remains limited. Early works [7,26,33] first predict 2D keypoints from music and subsequently drive image-based animation, while others [36, 46] instead infer 3D SMPL parameters and render image animations accordingly. Moreover, [10, 37] introduce end-to-end diffusion-based generation pipelines.

However, existing methods still struggle to generate dance videos that simultaneously exhibit expressive motion and high visual quality, primarily due to two factors. *(1) Dataset.* The lack of large-scale open datasets and effective data collection pipelines specifically tailored for dance video generation. *(2) Method.* The absence of principled framework-level solutions for effectively integrating music as a complementary conditioning signal into Video Generation Foundation Models, which are trained without audio supervision.

To tackle the dataset limitation, we firstly develop a *Progressive Expert-Based Data Collection Pipeline*. To improve computational efficiency, we adopt a progressive easy-to-hard strategy. For simple task, a lightweight verification expert (Qwen3-VL-2B [2]) is sufficient to directly predict positive cases. For more complex scenarios, we employ multiple heavier filtering experts (Qwen3-VL-8B [2]) to inversely identify and eliminate common failure modes (i.e., negative cases). Specifically, the pipeline consists of the following six stages: (1) Popular Creator Mining, (2) Visual quality verification, (3) Reference clarity verification, (4) Dance video verification, (5) Single-dancer filtering, and (6) Scene stability filtering. Secondly, we adopt *Choreograph Informed Text Annotations*, under which each video is annotated from five complementary choreography-informed aspects using Qwen3-VL-8B [2]: (1) Body Dynamics, (2) Choreographic Content, (3) Expressiveness, (4) Camera Presentation, (5) Overall Look. Finally, we propose **CIPE-Dance**, the large-scale Internet-source dataset, equipped with the **C**horeography **I**nformed annotations and constructed by the **P**rogressive **E**xpert pipeline. To the best of our knowledge, the **CIPE-Dance** dataset is currently the largest open-source dataset for dance video generation, comprising approximately 300k clips (over 400 hours) and covering diverse dancer demographics, performance environments, and over 30 dance genres.

To address the methodological limitation, we propose **OmniDance**, a unified framework that supports multimodal driven dance video generation (TI2V, MI2V, and MTI2V), as shown in Fig. 1. **OmniDance** effectively injects music conditioning into a Video Generation Foundation Models, without compromising its original capabilities, through a three-level design. **(1) Model Architecture.** Built upon WAN2.2-TI2V-5B [35], we augment each DiT block with an additional *music* cross-attention layer. Motivated by the depth hierarchy of DiT models—where shallow layers capture low-frequency global structure and deeper layers refine high-frequency details—we design a *Music-Text Progressive Specialization* scheme: shallow layers emphasize text conditioning to establish semantic layout, while deeper layers progressively strengthen music conditioning to refine rhythm-aware motion dynamics via the audio residual. **(2) Training Strategy.** We adopt an *Easy-to-Hard Curriculum Learning Strategy* to extend the pretrained TI2V model toward unified TI2V, MI2V, and MTI2V generation *while preserving* its original text/image controllability and visual fidelity. In Stage I, **OmniDance** is trained exclusively on TI2V to adapt the model to dance-specific textual descriptions. In Stage II, a music branch is introduced, and **OmniDance** jointly performs TI2V and MTI2V, enabling music understanding under text guidance. In Stage III, **OmniDance** is trained on all three

tasks—TI2V, MI2V, and MTI2V—allowing the model to generate dance videos solely from music when text is absent. **(3) Inference Strategy.** For classifier-free guidance (CFG), we propose a *Modality-Specialized CFG* that progressively *shifts the guidance emphasis* from text to music along the sampling trajectory: text dominates early steps to establish global semantics, while music becomes increasingly influential at later steps to refine rhythm-aware motion dynamics.

In conclusion, our contributions are as follows: (1) We introduce **CIPE-Dance**, the largest dance video generation dataset to date, equipped with choreography informed text annotations and constructed via a progressive expert pipeline. (2) We propose **OmniDance**, a three-level co-designed framework (architecture, curriculum learning, and modality-specialized inference) for injecting music conditioning into TI2V foundation models while preserving controllability and visual fidelity. (3) We unify **TI2V**, **MI2V**, and **MTI2V** in a single model (**OmniDance**), enabling seamless modality switching at inference time without training or maintaining separate generators. (4) We conduct extensive experiments on the **CIPE-Dance** dataset, demonstrating the state-of-the-art (SOTA) performance of **OmniDance** on TI2V, MI2V, and MTI2V generation, together with its robust multi-modal integration capability.

2 Related Work

2.1 Human Motion Generation

Human motion generation has advanced rapidly in recent years and is closely related to dance video generation, particularly in pose-driven image animation, talking head generation, and music-driven 3D dance generation. **(1) Pose-driven image animation.** Pose-driven image animation utilizes 2D keypoints to generate motion videos, achieving notable advances [8,15,32]. However, these methods overlook the fact that music constitutes the structural backbone of dance. Generating dance motions by choreographing pose sequences from music—rather than animating pre-specified poses—represents a substantially more valuable and challenging problem. **(2) Talking head generation.** Talking head generation employs audio features to generate vivid and lip-synced videos, also achieving significant breakthroughs [28–30]. However, these approaches fail to establish a principled correspondence between audio signals and full-body motion, even though such alignment is essential for the aesthetic expression of dance. **(3) Music-driven 3D dance generation.** Music-driven 3D dance generation aims to synthesize full-body motion from music [13,22,40,44,48]. Beyond dance-specific studies, speech-driven 3D gesture generation has also made notable progress in recent years [49–54]. However, most existing 3D dance methods are developed on relatively small-scale datasets (e.g., FineDance—the largest publicly available 3D dance dataset to date—contains only approximately 8 hours of motion), which ultimately constrains their motion generation capacity and generalization to diverse in-the-wild settings.

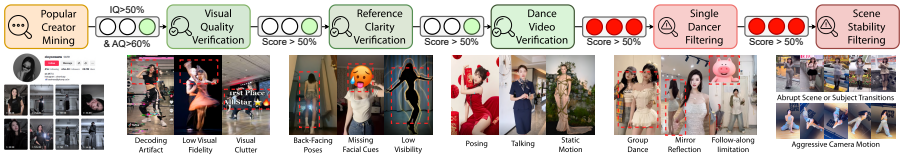

Fig. 2: Overview of the Progressive Expert-Based Data Collection Pipeline.

2.2 Music-Driven Dance Video Generation

Finally, research on music-driven dance video generation remains limited and can be broadly categorized into three paradigms: 2D keypoint–based generation, 3D SMPL-based generation, and end-to-end video generation. **(1) 2D keypoint-based generation.** X-Dancer [7], M2PE-Diff [26], and STG-Mamba [33] first predict 2D keypoints from music and then drive image-based animation, but they remain challenged by limb occlusions and complex full-body locomotion in dance videos. **(2) 3D SMPL-based generation.** ChoreoMuse [36] and MACE-Dance [46] instead infer 3D SMPL parameters and render image animation accordingly. However, the high cost of acquiring 3D supervision keeps existing 3D dance datasets small in scale, limiting the diversity and fidelity of the learned motion patterns. **(3) End-to-end generation.** DabFusion [37], MusicInfuser [14] and MuseDance [10] introduce end-to-end generation pipelines; however, they do not effectively leverage Video Generation Foundation Models, resulting in in-the-wild outputs with limited motion expressiveness and degraded visual appearance.

In conclusion, existing methods still struggle to generate dance videos that simultaneously exhibit expressive motion and high visual quality. This limitation primarily arises from two factors: *(1) Dataset.* The lack of large-scale and high-quality datasets and effective data collection pipelines specifically tailored for dance video generation; and *(2) Method.* The absence of principled framework-level solutions for effectively integrating music as a complementary conditioning signal into Video Generation Foundation Models, which are trained without audio supervision.

3 Dataset

3.1 Data Collection

To address data quality challenges in large-scale web-crawled dance video collection, we develop a *Progressive Expert-Based Data Collection Pipeline*, as shown in Fig. 2. To improve computational efficiency, the pipeline adopts a progressive easy-to-hard data filtering strategy. For simple cases, a lightweight verification expert is sufficient to directly identify positive samples. For hard scenarios, we employ multiple moderately heavier filter experts to inversely identify and eliminate common failure modes (i.e., negative samples). Specifically, the pipeline

consists of the following stages. **(1) Popular Creator Mining.** We collect all publicly available videos from approximately 500 dance content creators on Douyin or TikTok. Creators are required to have more than 50K followers, serving as a proxy for content quality and production standards. To ensure popularity, we restrict the dataset to videos published after 2018. The collected videos are segmented into 5-second clips at 16 FPS for downstream processing, resulting in a total of 630k clips. **(2) Visual quality verification.** To ensure visual fidelity, we apply VBench [21] for quality filtering. We retain only clips with image quality (IQ) scores above 60.00 and aesthetic quality (AQ) scores above 50.00, removing samples with low visual fidelity, decoding artifacts and visual clutter. This filtering rule preserves 96.76% of the collected clips. **(3) Reference clarity verification.** The first frame of each clip is treated as the reference image. We apply Qwen3-VL-2B [2] to automatically assess human subject visibility. The prompt explicitly specifies low visibility, back-facing poses, and missing facial cues as negative cases to filter out ambiguous or unclear references. After filtering, 81.45% of the clips remain. **(4) Dance video verification.** We apply Qwen3-VL-2B [2] to perform automatic dance content classification. The prompt defines talking, posing, and static motions as negative conditions to filter out ambiguous clips. This filtering step retains 89.24% of the clips. **(5) Single-dancer filtering.** Directly detecting single-dancer videos is non-trivial. Instead, we identify and remove three representative negative scenarios using Qwen3-VL-2B [2]: (i) group dance videos (e.g., duet or multi-person performances), (ii) follow-along imitation videos where a dancer replicates movements from an overlaid reference clip, and (iii) mirror-reflection artifacts commonly observed in studio settings, where reflections create the appearance of multiple performers. These three filtering criteria retain 80.35%, 97.02%, and 95.57% of the clips, respectively, resulting in an overall retention rate of 74.45%. **(6) Scene stability filtering.** To address complex scene dynamics, we apply two Qwen3-VL-8B [2] models to remove representative negative scenarios: (i) abrupt scene or subject transitions (e.g., sudden changes from one scene or subject to another), (ii) aggressive camera motion characterized by discontinuous or jittery movements. These two filtering criteria retain 93.95% and 97.55% of the clips, respectively, resulting in an overall retention rate of 91.65%.

3.2 Data Annotation

To better adapt the dataset to dance video generation, we adopt *Choreography-Informed Text Annotations*. Fig. 3 shows an example. Under this scheme, each video is annotated by Qwen3-VL-8B [2] from five independent yet complementary choreography-oriented aspects: **(1) Body Dynamics.** Describes the concrete, moment-to-moment actions and body mechanics, including which body parts move and how (e.g., arm swings, hand waves, and hip isolations). **(2) Choreographic Content.** Summarizes the *dance-level semantics* of the choreography, including the dance genre/style (e.g., popping, hip-hop, jazz funk) and characteristic movement vocabulary or techniques (e.g., tutting, waving,

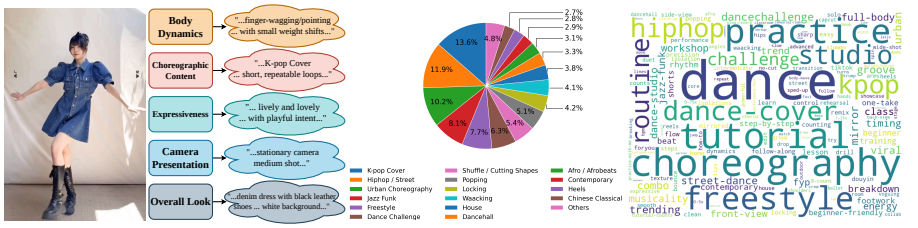

Fig. 3: Dataset analysis. (Left) Display of typical annotations; (Middle) Dance genre distribution from a random sample of 400 entries; (Right) Word cloud summarizing the caption content.

Thomas flair). **(3) Expressiveness.** Characterizes the overall emotional expression and performance intent conveyed by the dancer (e.g., joyful, playful, confident, melancholic), reflected through facial expression, posture, energy, and attitude. **(4) Camera Presentation.** Describes shot composition, camera movement, and framing strategies that influence visual perception of the dance. **(5) Overall Look.** Summarizes appearance attributes (e.g., clothing, styling) and environmental context that contribute to visual semantics.

3.3 Data Statistic

Finally, we present **CIPE-Dance**, a large-scale Internet-sourced dataset constructed via the Progressive Expert pipeline and equipped with Choreography-Informed annotations. To the best of our knowledge, **CIPE-Dance** is currently the largest open-source dataset for dance video generation, containing approximately 300K clips (over 400 hours of video content). See Fig. 3 for dataset analysis. Beyond scale, CIPE-Dance provides comprehensive coverage across multiple orthogonal dimensions: **(1) Genre Diversity.** The dataset spans over 30 dance genres, exhibiting a naturally long-tailed distribution that reflects real-world choreography trends. **(2) Environmental Diversity.** Videos are collected from diverse real-world settings, including studios, stages, streets, and home environments, covering substantial variation in lighting and background complexity. **(3) Performer Diversity.** CIPE-Dance features performers with diverse appearance attributes and performance styles, spanning professional stage recordings and casual social-media videos. **(4) Motion Complexity.** Choreography ranges from rhythm-driven groove patterns to highly dynamic sequences involving large body displacement, jumps, and rapid limb articulation. **(5) Camera Variability.** Clips include diverse viewpoints and framing styles (e.g., static, handheld, mild panning), while stability filtering ensures consistent identity and minimal abrupt transitions. For evaluation, we randomly sample 100 clips from **CIPE-Dance** that are excluded from the training set to form a test split.

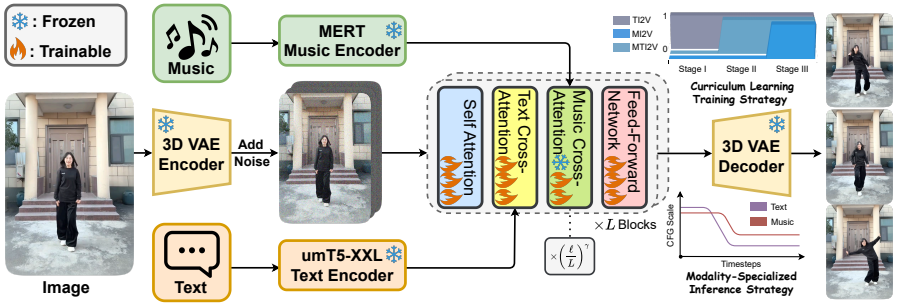

Fig. 4: Overview of OmniDance: music–text progressive specialization model architecture, curriculum learning training strategy, and modality-specialized inference strategy.

4 Methodology

4.1 Overview

To effectively integrate music conditioning into a Video Generation Foundation Model without degrading its original generative capabilities, **OmniDance** introduces a three-level design: (1) a music–text progressive specialization model architecture, (2) an easy-to-hard curriculum training strategy, and (3) a modality-specialized inference strategy. Moreover, **OmniDance** unifies TI2V, MI2V, and MTI2V generation within a single multimodal framework, as shown in Fig. 4.

4.2 Model Architecture

Wan2.2-TI2V-5B Our backbone is Wan2.2-TI2V-5B, a Diffusion Transformer (DiT) pretrained foundation model for text/image-to-video generation. Each DiT block consists of a self-attention layer, a text-conditioned cross-attention layer, and a feed-forward network. Wan2.2-TI2V-5B is equipped with a spatio-temporal VAE that downsamples the input by a factor of $(4, 16, 16)$ along the (t, h, w) dimensions. This higher compression ratio reduces the latent resolution and provides approximately $4\times$ computational savings compared to prior configurations at the same output resolution. Wan2.2-TI2V-5B is trained under the Flow Matching paradigm. Given the original latent representation $\mathbf{x}_0$ (data) and pure noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, We define the linear optimal transport path $\mathbf{x}_t$ and its corresponding velocity field $\mathbf{v}_t$ as:

$$\mathbf{x}_t = (1-t)\cdot\mathbf{x}_0 + t\cdot\epsilon, \quad \mathbf{v}_t = \epsilon - \mathbf{x}_0 \tag{1}$$

where the timestep t is randomly sampled from a uniform distribution $\mathcal{U}(0, 1)$. The velocity field predictor $\mathbf{v}_\theta$, instantiated as WAN2.2-TI2V-5B, is trained to minimize:

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|\mathbf{v}_\theta - \mathbf{v}_t\|_2^2\right]. \tag{2}$$

Music-Text Progressive Specialization DiT-based foundation models naturally exhibit a coarse-to-fine learning hierarchy: early layers capture high-level and low-frequency global structures, while deeper layers progressively refine low-level and high-frequency details. In multimodal-driven dance video generation, music and text play complementary roles. Music primarily governs high-frequency temporal structures (e.g., rhythm, tempo, and energy evolution), whereas text specifies global semantic intent (e.g., motion style, choreography intent, and key pose constraints). Directly injecting both modalities into all layers may lead to representational competition rather than synergy.

To promote complementary specialization, we introduce **Music-Text Progressive Specialization**. Built upon WAN2.2-TI2V-5B, we augment each DiT block at layer $\ell \in \{1, \ldots, L\}$ with an additional audio cross-attention module. The residual contribution of the audio branch is scaled in a depth-aware manner to progressively amplify music influence in deeper layers:

$$\mathbf{x}_\ell \leftarrow \mathbf{x}_\ell + \left(\frac{\ell}{L}\right)^\gamma \cdot \mathbf{r}^\ell_{\text{audio}}. \tag{3}$$

Here, $\mathbf{x}_\ell$ denotes the latent token features at layer ℓ, and $\mathbf{r}^\ell_{\text{music}}$ is the residual update produced by the added music cross-attention branch. In early layers ($\ell \ll L$), text conditioning dominates, establishing the overall action layout and choreography structure. As generation proceeds to deeper layers ($\ell \to L$), the audio contribution gradually increases, enabling fine-grained refinement of motion dynamics, rhythmic alignment, and temporally localized movement details.

4.3 Training Strategy

Current mainstream video generation foundation models released in open-source form are primarily designed for text/image-to-video (TI2V) tasks and lack audio-related supervision signals. To effectively extend the TI2V capability of WAN2.2-5B toward unified TI2V, MI2V, and MTI2V generation, we adopt a *Curriculum Training Strategy*, consisting of three easy-to-hard progressive stages.

Stage I: Text-Adaptation Warm-up. In the first stage, **OmniDance** is trained exclusively on TI2V to adapt the pretrained foundation model to dance-specific textual descriptions. This phase stabilizes text-to-video alignment in the dance domain and ensures reliable semantic grounding before introducing additional modalities. The music branch remains frozen to prevent premature interference with the pretrained backbone. Formally, we optimize:

$$\min_{\theta_{\text{text}}} \mathcal{L}_{\text{TI2V}}, \quad \theta_{\text{music}} \text{ frozen}. \tag{4}$$

Stage II: Anchored Music Integration. In the second stage, a music branch is introduced while preserving the TI2V objective. The model is trained on both TI2V and MTI2V tasks, allowing music conditioning to be progressively

learned under text guidance. This design mitigates abrupt degradation of the pretrained TI2V capability while establishing music–motion alignment in a controlled manner. To stabilize optimization when introducing the music branch, we zero-initialize its residual projection and apply a linear learning-rate warm-up to the newly introduced parameters, reducing the risk of disrupting the pretrained backbone during early training. Formally, we train on:

$$\min_{\theta_{\text{text}},\theta_{\text{music}}} \mathcal{L} \in \{\mathcal{L}_{\text{TI2V}}, \mathcal{L}_{\text{MTI2V}}\}. \tag{5}$$

Stage III: Modality Decoupling and Specialization. In the final stage, **OmniDance** is trained on all three tasks—TI2V, MTI2V, and MI2V. Having acquired stable music–motion alignment in earlier stages, the model is encouraged to generate dance videos solely from music when textual input is absent. This stage promotes genuine music-driven specialization and reduces over-reliance on textual cues, enabling the model to capture intrinsic music–dance correlations. All parameters are optimized during this phase. Formally, we train on:

$$\min_{\theta_{\text{text}},\theta_{\text{music}}} \mathcal{L} \in \{\mathcal{L}_{\text{TI2V}}, \mathcal{L}_{\text{MTI2V}}, \mathcal{L}_{\text{MI2V}}\}. \tag{6}$$

4.4 Inference Strategy

Classifier-Free Guidance Under the Flow Matching paradigm, earlier timesteps ($\tau \to 0$) primarily model coarse, low-frequency global structures, while later timesteps ($\tau \to 1$) refine fine-grained, high-frequency details. Text conditioning typically conveys global semantic intent (e.g., choreography layout and motion style), whereas music provides temporally localized motion cues (e.g., rhythm, tempo, and energy variations). Motivated by this frequency alignment, we propose an *Modality-Specialized CFG Strategy* that assigns text and music guidance to the timesteps where they are most effective. Specifically, text guidance is emphasized at early stages and gradually decays, while music guidance progressively increases toward later stages. The unified guidance formulation is:

$$\begin{aligned}
\mathbf{v}_\theta^{\text{TI2V}}(x_\tau) &= \mathbf{v}_\theta(x_\tau, \varnothing, \varnothing) + \lambda_t(\tau)\Big(\mathbf{v}_\theta(x_\tau, text, \varnothing) - \mathbf{v}_\theta(x_\tau, \varnothing, \varnothing)\Big), \\
\mathbf{v}_\theta^{\text{MI2V}}(x_\tau) &= \mathbf{v}_\theta(x_\tau, \varnothing, \varnothing) + \lambda_m(\tau)\Big(\mathbf{v}_\theta(x_\tau, \varnothing, music) - \mathbf{v}_\theta(x_\tau, \varnothing, \varnothing)\Big), \\
\mathbf{v}_\theta^{\text{MTI2V}}(x_\tau) &= \mathbf{v}_\theta(x_\tau, \varnothing, \varnothing) + \lambda_t(\tau)\Big(\mathbf{v}_\theta(x_\tau, text, \varnothing) - \mathbf{v}_\theta(x_\tau, \varnothing, \varnothing)\Big) \\
&\quad + \lambda_m(\tau)\Big(\mathbf{v}_\theta(x_\tau, text, music) - \mathbf{v}_\theta(x_\tau, text, \varnothing)\Big).
\end{aligned} \tag{7}$$

Over the generation trajectory, we use time-dependent guidance scales that both decay (text: $\lambda_t(\tau)$ from $5 \to 1$, music: $\lambda_m(\tau)$ from $4 \to 2$); since λ_t decays more aggressively, the relative guidance gradually shifts from text-dominant to more music-influenced refinement.

Long-Sequene Generation Following the WAN-style [35] autoregressive extension, we adopt a sliding-window generation scheme: the last frame of the

Table 1: Quantitative comparison on the CIPE-Dance dataset for dance video generation task under different conditioning modalities, including Music-Image-to-Video (MI2V), Text-Image-to-Video (TI2V) and Music-Text-Image-to-Video(MTI2V).

Task	Method	Video Quality						Motion Quality				Alignment	
		IQ↑	AQ↑	SC↑	BC↑	MS↑	TF↑	FID_k↓	FID_g↓	DIV_k↑	DIV_g↑	BAS↑	OC↑
–	Ground Truth	67.11	54.16	91.51	93.62	97.23	95.36	–	–	8.24	4.71	0.288	12.86
MI2V	Hallo2 [9] [ICLR'25]	64.09	52.37	90.77	92.94	97.05	95.72	16.49	1.81	10.30	5.45	0.244	11.85
	WAN-S2V [12] [Tongyi'25]	65.33	55.81	89.97	91.92	97.34	95.47	21.47	2.83	10.25	5.39	0.283	12.10
	Echomimic-V3 [25] [AAAI'26]	62.45	53.38	89.92	92.26	97.66	95.71	21.13	1.99	10.38	5.09	0.274	12.02
	OmniDance (MI2V)	**65.61**	**54.96**	**92.63**	**93.80**	**98.09**	**96.65**	**17.55**	**1.64**	**10.55**	**5.60**	**0.293**	**12.40**
TI2V	CogVideoX1.5-5B [47] [ICLR'25]	55.77	51.58	87.32	91.50	97.05	95.66	15.51	2.82	10.36	5.55	0.222	**12.53**
	HunyuanVideo [18] [Hunyuan'25]	63.92	54.11	**90.70**	92.53	97.87	96.15	17.89	2.44	8.82	5.28	0.243	11.79
	WAN2.2-TI2V-5B [35] [Tongyi'25]	64.66	55.02	90.10	91.78	98.08	96.12	19.15	2.16	10.62	5.25	0.252	12.11
	OmniDance (TI2V)	**67.74**	**55.17**	90.16	**92.70**	**99.24**	**98.46**	**13.55**	**2.04**	**11.15**	**5.79**	**0.259**	**12.53**
	OmniDance (MTI2V)	67.77	55.76	94.34	94.54	99.22	98.49	13.97	1.56	10.33	6.41	0.287	13.08

current clip is reused as the reference (starting) frame for the next window, enabling long-sequence synthesis by iteratively chaining short clips.

5 Experiment

5.1 Experimental Setup

Implementation Details We implement OmniDance using PyTorch on NVIDIA H20 GPUs. The backbone is Wan2.2-TI2V-5B with 30 DiT blocks. The audio encoder is MERT and the text encoder is umT5-XXL. Stage 1 trains for 5k steps, Stage 2 for 5k steps and Stage 3 for 10k, all with a batch size of 32, learning rate 1×10^{-5}, and the AdamW optimizer. For Music-Text Progressive Specialization, we set $\gamma = 1.5$ in Eq. 3. We train OmniDance on 77-frame videos (5s at 16 FPS) at 704×1280 resolution using flow-matching sampling with 40 steps. The full training takes approximately 70 hours on 32 NVIDIA H20 GPUs, with about 80GB memory usage per GPU. During inference, generating a 5-second video on a single NVIDIA H20 GPU takes about 4.5 minutes for TI2V and MI2V, and about 7.2 minutes for MTI2V, with approximately 40GB GPU memory usage.

Evaluation Metrics To comprehensively evaluate the performance of dance video generation task, we adopt metrics inspired by both VBench [21] and 3D Dance Generation [23], assessing the results from three aspects: Video Quality, Motion Quality, and Alignment. **(1) Video Quality.** Considering the dynamic and temporal nature of dance videos, we select several metrics from VBench [21], including imaging quality (IQ), aesthetic quality (AQ), subject consistency (SC), background consistency (BC), motion smoothness (MS), and temporal flickering (TF). **(2) Motion Quality.** We evaluate the diversity and fidelity of generated dance motions using DIV and FID computed on motion features. Specifically, we first extract 2D keypoint sequences using ViTPose [38], and then compute

two types of motion features: (i) Kinetic Features [23] (k), which capture motion dynamics, and (ii) Geometric Features [23] (g), which encode spatial joint relationships. **(3) Alignment.** To measure music–motion synchronization, we adopt the Beat Alignment Score (BAS) [23]. To assess text–video consistency, we use Overall Consistency (OC) from VBench [21]. Additionally, BAS can also be applied to the TI2V task, as videos that align well with textual descriptions often exhibit rhythmic patterns similar to music. Conversely, OC can be used for the MI2V task, where the text prompt is simplified as "a person is dancing."

5.2 Comparison

Text-Image-to-Video We compare OmniDane (TI2V) against the representative strong TI2V foundation models, including CogVideoX1.5-5B [47] (ICLR'25), HunyuanVideo [18] (Hunyuan'25), and WAN2.2-TI2V-5B [35] (Tongyi'25).

Quantitatively (Tab. 1), **OmniDance** demonstrates overall state-of-the-art performance in the TI2V task. Specifically, it achieves the best *video quality* among TI2V baselines, with the highest IQ (67.74), AQ (55.17), BC (92.70), MS (99.24), and TF (98.46). It also yields better *motion quality*, achieving the best fidelity (FID_k: 13.55, FID_g: 2.04) with strong diversity (DIV_k: 11.15, DIV_g: 5.79). Alignment is also improved, with BAS increasing to 0.259 (vs. 0.252 for WAN2.2) while maintaining competitive OC (12.53).

Qualitatively (Fig. 5), OmniDance generates more diverse hand articulations and maintains a natural smile, producing dance-like motion with subtle weight shifts and hip sways.

We attribute these gains to two factors: (i) the Stage-I *text-adaptation warm-up* bridges the domain gap between generic TI2V pretraining and dance-specific semantics. (ii) CIPE-Dance's high-quality, single-dancer, and scene-stable clips provide cleaner supervision for appearance fidelity and motion consistency.

Music-Image-to-Video Due to limited open-source implementations specifically tailored for music-driven dance video generation, we compare OmniDance (MI2V) against two representative categories of audio-driven image animation baselines. (i) Fine-tuning-based. We adapt Hallo2 [9] (ICLR'25) from talking-head animation to full-body dance by replacing its facial mask with a full-body mask. (ii) Inference-only. We directly evaluate models advertised as general-purpose human motion generation systems without task-specific training, including WAN-S2V [12] (Tongyi'25) and Echomimic-V3 [25] (AAAI'26).

Quantitatively (Tab. 1), **OmniDance** exhibits the state-of-the-art performance in the MI2V task. Specifically, it achieves the best overall *video quality* in terms of IQ (65.61), SC (92.63), BC (93.80), MS (98.09), and TF (96.65), while maintaining competitive AQ (54.96). It also improves *motion quality*, achieving the best geometric fidelity (FID_g: 1.64) and the highest diversity (DIV_k: 10.55, DIV_g: 5.60). For *alignment*, OmniDance reaches the best BAS (0.293) and OC (12.40), indicating stronger music-synchronized motion and overall consistency.

Qualitatively (Fig. 5), OmniDance produces rhythm-aware and intensity-progressive choreography with coherent gesture transitions that follow the drum

Fig. 5: Comparison with SOTAs on Text-Image-to-Video (TI2V, left) and Music-Image-to-Video (MI2V, right) tasks on the CIPE-Dance dataset.

beats, while maintaining stable identity and framing. In contrast, Hallo2 suffers from degraded visual quality with noticeable artifacts and blur; Echomimic-V3 often yields oversimplified and occasionally distorted motions; WAN-S2V tends to lose dancer identity over time (e.g., facial/appearance drift).

We attribute these gains to two factors: (i) the music branch is learned in a controlled curriculum (Stage II/III), enabling genuine MI2V generation without sacrificing stability; and (ii) the music-text progressive specialization together with modality-specialized CFG allocates music guidance to later refinement stages, improving beat-level synchronization.

5.3 Multimodal Integration

Quantitative Analysis Tab. 1 shows that OmniDance performs strongly under both *single-modality* settings. Under TI2V and MI2V, it achieves SOTA-level video and motion quality (e.g., TI2V: IQ 67.74 / MS 99.24; MI2V: SC 92.63 / BAS 0.293), indicating that enabling the music branch does not compromise the backbone's core generation capability. When both modalities are available, **OmniDance (MTI2V)** further strengthens overall multimodal consistency and joint control: SC and BC improve to 94.34 and 94.54, and OC increases to 13.08. Although a few metrics slightly decrease compared to the best single-modality counterpart (e.g., BAS 0.287 vs. 0.293 in MI2V, and DIV_k 10.33 vs. 11.15 in TI2V), MTI2V yields a better overall trade-off by jointly preserving identity/scene stability and maintaining near ground-truth beat synchronization (0.287 vs. 0.288), demonstrating effective multimodal synergy.

Qualitative Analysis As shown in Fig. 5, OmniDance benefits from clear cross-modal complementarity: adding music to TI2V enriches rhythmic structure and fine-grained motion dynamics, while adding text to MI2V strengthens semantic

Table 2: Ablation study on the model architecture on MTI2V task.

	AQ↑	SC↑	MS↑	BAS↑	OC↑
Ground Truth	54.16	91.51	97.23	0.288	12.86
w/o. MTPS	54.18	90.61	93.74	0.257	12.53
MTPS (R)	52.31	89.89	95.98	0.234	12.37
w/o. MERT	**56.10**	**95.72**	98.64	0.241	12.78
Full Model	55.76	94.34	**99.22**	**0.287**	**13.08**

Table 3: Ablation study on the generative strategy on MTI2V task.

	AQ↑	SC↑	MS↑	BAS↑	OC↑
Ground Truth	54.16	91.51	97.23	0.288	12.86
w/o. Stage I	52.92	92.74	94.88	0.239	11.72
w/o. Stage II	51.38	**94.61**	98.11	**0.293**	12.71
w/o. MS CFG	53.27	93.96	96.84	0.279	**13.22**
Full Model	**55.76**	94.34	**99.22**	0.287	13.08

understanding. Overall, MTI2V leverages complementary cues from both modalities to generate more diverse and temporally coherent choreography, with richer expressions and more varied gesture sequences.

We attribute the strong multimodal integration to: (i) *progressive specialization* that decouples text-guided semantics (shallow layers) from music-driven dynamics (deep layers); (ii) an *easy-to-hard curriculum* that learns music conditioning without degrading the pretrained TI2V capability; and (iii) *modality-specialized CFG* that emphasizes text early and music late along the generation trajectory.

5.4 Ablation Study

Model Architecture We ablate key architectural components of OmniDance on the **MTI2V** task and report AQ/SC/MS/BAS/OC (Tab. 3). We evaluate: (i) **w/o. MTPS**, disabling the depth-aware audio residual scaling in Eq. (3); (ii) **MTPS (R)**, reversing the schedule with stronger audio weights in shallow layers and weaker weights in deeper layers; (iii) **w/o. MERT**, replacing MERT with Wav2Vec2.0 [1]; and (iv) the **Full Model**.

Specifically, **(1) Music-Text Progressive Specification.** Removing MTPS degrades temporal smoothness and music-motion synchronization (MS: 99.22 → 93.74; BAS: 0.287 → 0.257; OC: 13.08 → 12.53). Reversing the schedule further hurts performance (BAS: 0.234; SC: 89.89), supporting progressive strengthening of music in deeper layers; **(2) Music Representaion.** Replacing MERT with Wav2Vec2.0 improves appearance (AQ: 56.10; SC: 95.72) but reduces alignment and consistency (BAS: 0.241 vs. 0.287; OC: 12.78 vs. 13.08), indicating MERT provides more informative music cues for motion refinement.

Generative Strategy We ablate the proposed training and inference strategy of OmniDance on the **MTI2V** task and report AQ/SC/MS/BAS/OC (Tab. 3). We evaluate: (i) **w/o. Stage I**, removing the Stage-I text-adaptation warm-up while keeping the remaining stages unchanged; (ii) **w/o. Stage II**, skipping the anchored music-integration stage and directly training the remaining stages; (iii) **w/o. MS CFG**, replacing our modality-specialized CFG with a unified scheme where *text* and *music* share the same guidance weight that decays from 5 to 3 along the flow; and (iv) the **Full Model**.

Specifically, **(1) Curriculum stages.** Removing Stage I causes a clear performance drop across quality and alignment (AQ: 55.76 → 52.92; MS: 99.22 → 94.88; BAS: 0.287 → 0.239; OC: 13.08 → 11.72), indicating that text warm-up is critical for stabilizing dance-domain semantics before multimodal learning. Skipping Stage II yields strong beat alignment (BAS: 0.293) and competitive SC (94.61) but notably degrades appearance and consistency (AQ: 51.38; OC: 12.71), suggesting that abruptly introducing music can bias optimization toward synchronization at the expense of visual/text-consistent generation; **(2) Multimodality-Specific CFG.** Removing MS-CFG reduces smoothness and beat synchronization (MS: 99.22 → 96.84; BAS: 0.287 → 0.279), while slightly increasing OC (13.08 → 13.22), showing that modality-specification CFG strategy improves rhythm-driven refinement with a better overall trade-off.

6 Conclusion

In this work, we introduce **CIPE-Dance**, the largest dance video generation dataset to date, constructed via a progressive expert-based pipeline and equipped with choreography-informed text annotations. Next, we propose **OmniDance**, a unified framework that supports multimodal driven dance video generation, enabled by a music-text progressive specialization model architecture, an easy-to-hard curriculum training strategy, and a modality specialized inference strategy. Extensive experiments on **CIPE-Dance** demonstrate that **OmniDance** achieves state-of-the-art performance, while exhibiting strong multimodal integration capability. Although OmniDance does not yet support real-time generation, real-time generation plays an important role for this field. We believe that combining *Long-Video Generation via Distillation* offers a promising pathway toward efficient dance video synthesis.

7 Acknowledge

This work was supported in part by the National Natural Science Foundation of China under Grants 62436010, 72572090, 62572474 and 62172421.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631

3. Butterworth*, J.: Teaching choreography in higher education: A process continuum model. Research in dance education **5**(1), 45–67 (2004)
4. Chen, C., Hu, S., Zhu, J., Wu, M., Chen, J., Li, Y., Huang, N., Fang, C., Wu, J., Chu, X., et al.: Taming preference mode collapse via directional decoupling alignment in diffusion reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12775–12786 (2026)
5. Chen, C., Zhu, J., Feng, X., Huang, N., Wu, M., Mao, F., Wu, J., Chu, X., Li, X.: S^2-Guidance: Stochastic self guidance for training-free enhancement of diffusion models. arXiv preprint arXiv:2508.12880 (2025)
6. Chen, R., Sun, L., Tang, J., Li, G., Chu, X.: Finger: Content aware fine-grained evaluation with reasoning for ai-generated videos. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3517–3526 (2025)
7. Chen, Z., Xu, H., Song, G., Xie, Y., Zhang, C., Chen, X., Wang, C., Chang, D., Luo, L.: X-dancer: Expressive music to human dance video generation. arXiv preprint arXiv:2502.17414 (2025)
8. Cheng, G., Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Li, J., Meng, D., Qi, J., Qiao, P., et al.: Wan-animate: Unified character animation and replacement with holistic replication. arXiv preprint arXiv:2509.14055 (2025)
9. Cui, J., Li, H., Yao, Y., Zhu, H., Shang, H., Cheng, K., Zhou, H., Zhu, S., Wang, J.: Hallo2: Long-duration and high-resolution audio-driven portrait image animation. arXiv preprint arXiv:2410.07718 (2024)
10. Dong, Z., Hao, W., Wang, J.C., Zhang, P., Polak, P.: Every image listens, every image dances: Music-driven image animation. arXiv preprint arXiv:2501.18801 (2025)
11. Feng, X., Yu, H., Wu, M., Hu, S., Chen, J., Zhu, C., Wu, J., Chu, X., Huang, K.: Narrlv: Towards a comprehensive narrative-centric evaluation for long video generation models. arXiv e-prints pp. arXiv–2507 (2025)
12. Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Meng, D., Qi, J., Qiao, P., Shen, Z., Song, Y., et al.: Wan-s2v: Audio-driven cinematic video generation. arXiv preprint arXiv:2508.18621 (2025)
13. Gong, K., Lian, D., Chang, H., Guo, C., Jiang, Z., Zuo, X., Mi, M.B., Wang, X.: Tm2d: Bimodality driven 3d dance generation via music-text integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9942–9952 (2023)
14. Hong, S., Kemelmacher-Shlizerman, I., Curless, B., Seitz, S.M.: Musicinfuser: Making video diffusion listen and dance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 43751–43761 (June 2026)
15. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8153–8163 (2024)
16. Hu, S., Chen, C., Zhu, J., Wu, J., Chu, X., Li, X.: Embedding-perturbed exploration preference optimization for flow models. arXiv preprint arXiv:2605.15803 (2026)
17. Jia, T., Yang, K., Yang, X., Tang, X., Qiu, K., Wei, S., Zhao, Y.: Bitdiff: Fine-grained 3d conducting motion generation via bimamba-transformer diffusion. arXiv preprint arXiv:2604.04395 (2026)
18. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang,

J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models (2025), https://arxiv.org/abs/2412.03603

19. Lei, J., Liu, K., Berner, J., Yu, H., Zheng, H., Wu, J., Chu, X.: There is no vae: End-to-end pixel-space generative modeling via self-supervised pre-training. arXiv preprint arXiv:2510.12586 (2025)
20. Li, H., Wang, Y., Huang, T., Huang, H., Wang, H., Chu, X.: Ld-rps: Zero-shot unified image restoration via latent diffusion recurrent posterior sampling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13684–13694 (2025)
21. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
22. Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., Li, X.: Finedance: A fine-grained choreography dataset for 3d full body dance generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10234–10243 (2023)
23. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021)
24. Ling, X., Zhu, C., Wu, M., Li, H., Feng, X., Yang, C., Hao, A., Zhu, J., Wu, J., Chu, X.: Vmbench: A benchmark for perception-aligned video motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13087–13098 (2025)
25. Meng, R., Wang, Y., Wu, W., Zheng, R., Li, Y., Ma, C.: Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. arXiv preprint arXiv:2507.03905 (2025)
26. Park, N.T.: M2pe-diff: Music-to-pose encoder for dance video generation leveraging latent diffusion framework. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 2419–2428 (2025)
27. Peng, Z., Chen, Y., Ma, Y., Zhang, G., Sun, Z., Zhou, Z., Zhang, Y., Zhou, Z., Fan, Z., Liu, H., et al.: Actavatar: Temporally-aware precise action control for talking avatars. arXiv preprint arXiv:2512.19546 (2025)
28. Peng, Z., Hu, W., Ma, J., Zhu, X., Zhang, X., Zhao, H., Tian, H., He, J., Liu, H., Fan, Z.: Synctalk++: High-fidelity and efficient synchronized talking heads synthesis using gaussian splatting. arXiv preprint arXiv:2506.14742 (2025)
29. Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., Zhao, H., He, J., Liu, H., Fan, Z.: Synctalk: The devil is in the synchronization for talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 666–676 (2024)
30. Peng, Z., Liu, J., Zhang, H., Liu, X., Tang, S., Wan, P., Zhang, D., Liu, H., He, J.: Omnisync: Towards universal lip synchronization via diffusion transformers. arXiv preprint arXiv:2505.21448 (2025)
31. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11050–11059 (2022)

32. Tan, S., Gong, B., Wang, X., Zhang, S., Zheng, D., Zheng, R., Zheng, K., Chen, J., Yang, M.: Animate-x: Universal character image animation with enhanced motion representation. arXiv preprint arXiv:2410.10306 (2024)
33. Tang, H., Shao, L., Zhang, Z., Van Gool, L., Sebe, N.: Spatial-temporal graph mamba for music-guided dance video synthesis. IEEE transactions on pattern analysis and machine intelligence (2025)
34. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 448–458 (2023)
35. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), https://arxiv.org/abs/2503.20314
36. Wang, X., Wang, H., Cai, W.: Choreomuse: Robust music-to-dance video generation with style transfer and beat-adherent motion. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 7912–7921 (2025)
37. Wang, X., Wang, H., Liu, D., Cai, W.: Dance any beat: Blending beats with visuals in dance video generation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5136–5146. IEEE (2025)
38. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in neural information processing systems **35**, 38571–38584 (2022)
39. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1481–1490 (2024)
40. Yang, K., Tang, X., Diao, R., Liu, H., He, J., Fan, Z.: Codancers: Music-driven coherent group dance generation with choreographic unit. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. pp. 675–683 (2024)
41. Yang, K., Tang, X., Hu, Y., Yang, J., Liu, H., Zhang, Q., He, J., Fan, Z.: Matchdance: Collaborative mamba-transformer architecture matching for high-quality 3d dance synthesis. arXiv preprint arXiv:2505.14222 (2025)
42. Yang, K., Tang, X., Peng, Z., Hu, Y., He, J., Liu, H.: Megadance: Mixture-of-experts architecture for genre-aware 3d dance generation. arXiv preprint arXiv:2505.17543 (2025)
43. Yang, K., Tang, X., Peng, Z., Zhang, X., Wang, P., He, J., Liu, H.: Flowerdance: Meanflow for efficient and refined 3d dance generation. arXiv preprint arXiv:2511.21029 (2025)
44. Yang, K., Tang, X., Wu, H., Xue, Q., Qin, B., Liu, H., Fan, Z.: Cohedancers: Enhancing interactive group dance generation through music-driven coherence decomposition. arXiv preprint arXiv:2412.19123 (2024)
45. Yang, K., Zhou, X., Tang, X., Diao, R., Liu, H., He, J., Fan, Z.: Beatdance: A beat-based model-agnostic contrastive learning framework for music-dance retrieval. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. pp. 11–19 (2024)

46. Yang, K., Zhu, J., Tang, X., Peng, Z., Zhang, X., Wang, P., Wu, J., Chu, X., Liu, H., He, J.: Mace-dance: Motion-appearance cascaded experts for music-driven dance video generation. arXiv preprint arXiv:2512.18181 (2025)
47. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
48. Yang, Z., Yang, K., Tang, X.: Tokendance: Token-to-token music-to-dance generation with bidirectional mamba. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5520–5529 (2026)
49. Zhang, X., Cai, Y., Li, K., Yang, K., Zhou, Y., Li, Z., Chu, X., Zhang, J., Liu, H.: Personagesture: Single-reference co-speech gesture personalization for unseen speakers. arXiv preprint arXiv:2605.06064 (2026)
50. Zhang, X., Jia, Y., Zhang, J., Yang, Y., Tu, Z.: Robust 2d skeleton action recognition via decoupling and distilling 3d latent features. IEEE Transactions on Circuits and Systems for Video Technology (2025)
51. Zhang, X., Li, J., Ren, J., Zhang, J.: Mitigating error accumulation in co-speech motion generation via global rotation diffusion and multi-level constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12834–12842 (2026)
52. Zhang, X., Li, J., Zhang, J., Dang, Z., Ren, J., Bo, L., Tu, Z.: Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13761–13771 (2025)
53. Zhang, X., Li, J., Zhang, J., Ren, J., Bo, L., Tu, Z.: Echomask: Speech-queried attention-based mask modeling for holistic co-speech motion generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10827–10836 (2025)
54. Zhou, P., Zhang, X., Shen, X., Hu, Y.: Not all frames are equal: Complexity-aware masked motion generation via motion spectral descriptors. arXiv preprint arXiv:2603.29655 (2026)