APV-Gen: Audio-Pose-Driven Full-Body Video Generation with Weakly Supervised Alignment

Anonymous submission

Abstract

Recent advances in diffusion models have significantly enhanced video generation capabilities, particularly for portrait video synthesis conditioned on diverse inputs such as reference images, audio, and motion data. In real-world applications, users often require joint control over both audio and pose, for example, generating singing and dancing celebrities or speakers making expansive gestures. Therefore, this paper introduces a novel task: **audio-pose-driven full-body video generation**, which synthesizes realistic full-body human videos where full-body motion, including gestures, torso, and leg movements, is coordinated with speech audio. To address this task, we introduce **APV-Gen**, an innovative weakly supervised training framework that includes two key innovations: (1) Facial Data Alignment, an alignment technique that linearly scales facial data to match full-body proportions; (2) A mixed-data training strategy that alternates between aligned facial data and full body data to prevent catastrophic forgetting. Extensive experiments demonstrate that APV-Gen produces vivid and natural video output and achieves a trade-off between motion quality and audio-lip synchronization compared to existing approaches.

Introduction

With the advancement of diffusion models, video generation has gained significant traction within the research community. Taking advantage of their impressive generative capabilities, recent work has focused on enhancing the controllability of video generation models under specific conditioning signals. Among key research directions, portrait video generation has attracted widespread attention and extensive study due to its importance in daily life. It offers promising applications for reducing video production costs and enabling more accessible artistic creation. In real-world scenarios, users often require joint audio and pose control, such as generating singing and dancing celebrities or speakers making expansive gestures, where video content is highly correlated with both audio and pose information. Therefore, we introduce a novel and valuable task: audio-pose-driven full-body video generation, which aims to synthesize realistic full-body human videos with full-body motion (gestures, torso, and legs), synchronized to speech audio at the same time.

Existing work for portrait video generation falls roughly into two categories: audio-driven portrait face generation and pose-driven video generation. On the one hand, influenced by the tradition of talking face generation, audio-driven models focus on generating natural facial expressions and achieving lip-sync alignment but typically fail to generate full-body movements. On the other hand, pose-driven models prioritize natural and coherent full-body motion generation but lack lip-sync and audio alignment. Neither approach adequately addresses audio-pose-driven full-body video generation.

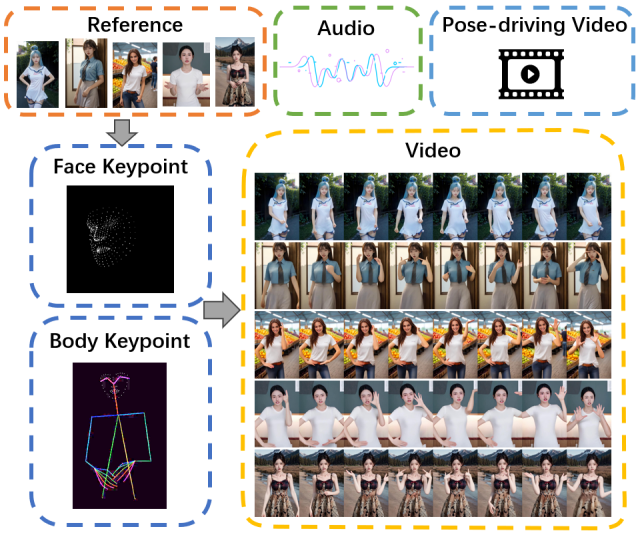

Figure 1: APV-Gen generates vivid and natural videos featuring full-body movements and lip-synchronized facial details, conditioned on a reference image, audio input, and a pose-driving video sequence.

This limitation stems primarily from the scarcity of suitable multimodal datasets containing synchronized audio, pose, facial expressions, and full-body movement. In existing audio-focused datasets (Shen et al. 2023; Wei, Yang, and Wang 2024; Xu et al. 2024; Wang et al. 2024a; Yang et al. 2024; Tian et al. 2025; Liu et al. 2024; Lin et al. 2025; Meng et al. 2024), audio information is typically associated with portrait videos that emphasize fine-grained facial control, where the footage primarily captures faces without significant body movement. In contrast, pose-focused

datasets (Chang et al. 2023; Hu et al. 2023; Zhang et al. 2024; Tong et al. 2024; Peng et al. 2024; Tu et al. 2024; Wang et al. 2024b; Karras et al. 2023; Xu et al. 2023; Shao et al. 2024) provide full-body movement information but lack detailed facial expression control synchronized with audio. This inherent misalignment of audio, facial expression, pose, and body movement information poses a significant challenge for joint audio-pose-driven full-body video generation.

To address this challenge, we propose APV-Gen, a novel weakly supervised alignment approach that achieves vivid and natural audio-pose-driven full-body video generation. APV-Gen leverages a pretrained pose-to-video generation model. To enable the model to generate particular facial expressions aligned with the audio, we first employ FDA (Facial Data Alignment), a linear scaling method to rescale the facial data so that it can align with the faces in the full-body data, where only the facial part provides weak supervision for the generated videos during training. To avoid the static artifacts caused by the only facial weak supervision, we further propose the mixed-data training strategy, where we train the model on the audio-face data and the pose-video data, the faces in which are annotated as the same format as the audio-face data, maintaining pose-generative ability. The weak supervision process enables the pretrained model with audio-to-facial expression generation, and the mixed-data training strategy maintains the ability of the model to generate dynamic videos that follow the pose information. In summary, our contributions are as follows:

- We investigate a novel and challenging task: audio-pose-driven full-body video generation, which holds significant value for both research and practical applications.
- We introduce APV-Gen, an innovative weakly supervised training framework built upon existing models and datasets, capable of generating audio-pose-driven full-body videos.
- We propose FDA, a novel alignment technique that enables seamless integration of audio-correlated facial data into the pose-driven foundation model, facilitating effective learning of facial dynamics.
- Extensive experiments demonstrate that APV-Gen produces vivid and natural video output, achieving an optimal balance between motion quality and audio-lip synchronization compared to existing approaches, effectively addressing the core challenges of this task.

Related Works

Audio-Driven Video Generation

Audio-driven video generation has a long history as a research task. For many years, the focus has primarily been on talking face generation, which naturally aligns with audio inputs. Recent advances have introduced various innovative approaches. DiffTalk (Shen et al. 2023) utilizes LDM to improve the quality of generation. AniPortrait (Wei, Yang, and Wang 2024) predicts a sequence of 3D meshes and projects them into 2D to obtain pose frames. VASA-1 (Xu et al. 2024) emphasizes controllability and supports real-time generation. V-Express (Wang et al. 2024a) balances multiple control signals through a series of progressive dropout operations. MegActor-Σ (Yang et al. 2024) introduces DiT to improve scalability and performance. EMO2 (Tian et al. 2025) identifies correlations between audio and hand movements, proposing a two-stage framework for audio-driven gesture generation. TANGO (Liu et al. 2024) generates co-speech body gesture videos using a CLIP-like embedding space. CyberHost (Lin et al. 2025) proposes a single-stage audio-driven video framework that bypasses intermediate representations, addressing common synthesis degradations. EchoMimic2 (Meng et al. 2024) advances half-body video generation with hand gestures, but it still lacks support for large-scale movements.

Pose-Driven Video generation

Pose-driven video generation has long been a mainstream approach in human animation research. These studies typically utilize various pose representations, including skeleton pose, dense pose, depth maps, mesh models, and optical flow, in conjunction with text and speech inputs as guiding modalities. Recently, conditional models based on Latent Diffusion Models(LDM) have gained significant attention from researchers. For example, MagicDance (Chang et al. 2023) incorporates ControlNet (Zhang, Rao, and Agrawala 2023) to tackle this task. Disco (Wang et al. 2023) performs pretraining on human appearance features to enhance the model's generalization across multiple identities and angles. Several advanced methods, such as AnimateAnyone (Hu et al. 2023), MimicMotion (Zhang et al. 2024), MusePose (Tong et al. 2024), ControlNeXt (Peng et al. 2024), MotionFollower (Tu et al. 2024), and UniAnimate (Wang et al. 2024b), employ DWPose (Yang et al. 2023) or OpenPose (Cao et al. 2019) for skeleton pose extraction. These approaches utilize lightweight neural networks with minimal convolutional layers, replacing the zero-convolution mechanism to provide aligned pose guidance during the denoising process. Meanwhile, DreamPose (Karras et al. 2023) and MagicAnimate (Xu et al. 2023) adopt DensePose (Güler, Neverova, and Kokkinos 2018) to extract detailed pose information, which is then concatenated with noise and fed into the denoising model through ControlNet. In a distinctive approach, Human4DiT (Shao et al. 2024) extracts 3D mesh maps and implements the Diffusion Transformer(DiT) architecture for video generation.

Methodology

Task Definition

We investigate a novel task: audio-pose-driven full-body video generation. Formally, given a sequence of full-body pose frames $\boldsymbol{p_f} = \{p_{f_1}, p_{f_2}, \ldots, p_{f_n}\}$, a corresponding audio signal $\boldsymbol{w}$, and a reference image r, the goal is to generate a realistic video $\boldsymbol{v} = \{v_1, v_2, \ldots, v_n\}$ that satisfies:

$$\boldsymbol{v} = \mathcal{M}(\boldsymbol{p_f}, \boldsymbol{w}, r) \tag{1}$$

where $\mathcal{M}$ denotes the generative model. The generated video $\boldsymbol{v}$ maintains temporal coherence with the input pose

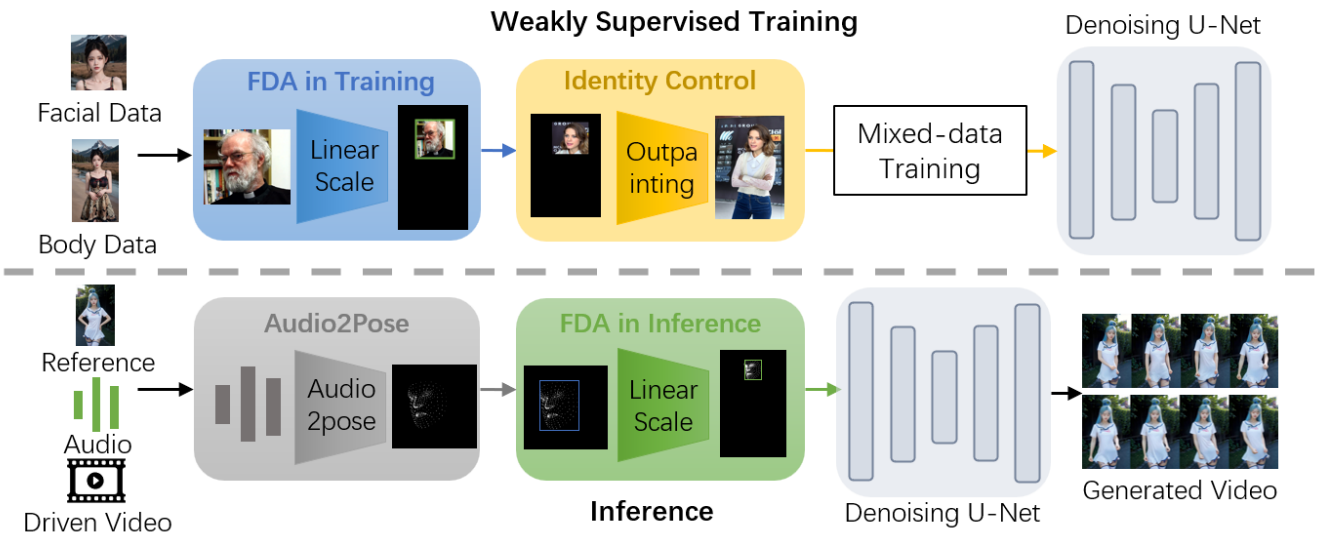

Figure 2: Overview of the APV-Gen framework: Leverages FDA to bridge facial data and full-body data for weakly supervised training. Incorporates an identity-consistent control module to resolve FDA-induced boundary artifacts of references, and employs mixed-data training to jointly learn facial dynamics while preserving pose-generation capabilities. During inference, audio-driven facial keypoints are predicted and adaptively scaled through FDA.

sequence $\boldsymbol{p_f}$, ensuring that each frame v_i aligns with the corresponding pose frame p_{f_i}, while simultaneously synchronizing lip movements with the audio signal $\boldsymbol{w}$ and preserving the visual identity of the reference image r.

Overview

The general architecture of APV-Gen is shown in Figure 2. Our work begins by collecting facial data $\mathcal{D}_f$ and full-body data $\mathcal{D}_b$ associated with audio. Inspired by the Cross Normalization technique proposed in ControlNeXt(Peng et al. 2024), we propose a novel Facial Data Alignment (FDA) technique to ensure compatibility between these two datasets, which is used in our weakly supervised alignment process. The primary objective of FDA is to align $\mathcal{D}_f$ with $\mathcal{D}_b$, allowing pose-driven models initially trained on data similar to $\mathcal{D}_b$ to effectively acquire facial generation capabilities during the fine-tuning phase.

Moreover, we introduce an identity-consistent control module to resolve boundary discontinuity artifacts introduced by the FDA technique.

Subsequently, we alternately use aligned facial data $\hat{\mathcal{D}}_f$ and processed full-body data $\hat{\mathcal{D}}_b$ to fine-tune the model. This mixed-data training strategy not only enables the model to learn generate facial details but also mitigates the risk of catastrophic forgetting regarding pose generation abilities.

Facial Data Alignment

Audio-driven models have achieved remarkable success in predicting facial poses $\boldsymbol{p}$, where $\boldsymbol{p} \in \mathbb{R}^{n \times h \times w}$ includes expressions and lip controls. Therefore, the key challenge becomes allowing the pose-driven model to learn the relationship between facial poses $\boldsymbol{p}$ and realistic frames $\boldsymbol{v}$.

To address this, we employ FDA technology to align facial poses $\boldsymbol{p}$ with poses from full-body data. The aligned poses then serve as conditional controls to guide the generation process. This approach proves to be more effective than Cross Normalization when dealing with structured conditional controls like facial keypoints.

FDA in Training Specifically, we detect facial regions in full-body videos $\boldsymbol{v}_b$ from $\mathcal{D}_b$ and extract an empirical distribution of scaling parameters based on the ratio of facial regions to the entire image. The empirical distribution is defined as follows:

$$a_x = \bar{a}_x + \mathcal{U}(-\delta, \delta) \tag{2}$$

$$a_y = \bar{a}_y + \mathcal{U}(-\delta, \delta) \tag{3}$$

$$b_x = \frac{1 - a_x}{2} + \mathcal{U}\left(-\frac{1 - ax}{2}, \frac{1 - ax}{2}\right) \tag{4}$$

$$b_y = \bar{b}_y + \mathcal{U}(-\delta, \delta) \tag{5}$$

where $\bar{a}_x$ and $\bar{a}_y$ represent the average ratios of the scaled facial regions to the complete image in horizontal and vertical dimensions, $\bar{b}_x$ and $\bar{b}_y$ represent the average horizontal and vertical offsets of the facial regions, while δ serves as a small hyperparameter for random perturbation of scaling parameters.

As shown in Figure 3, we then scale facial video frames $\boldsymbol{v}_f$ with corresponding pose frames $\boldsymbol{p}_f$ into $\hat{\boldsymbol{v}}_f$ and $\hat{\boldsymbol{p}}_f$ frame by frame using sampled scaling parameters:

$$\hat{\boldsymbol{v}}_f = (a_x, a_y) \cdot \boldsymbol{v}_f + (b_x, b_y) \tag{6}$$

$$\hat{\boldsymbol{p}}_f = (a_x, a_y) \cdot \boldsymbol{p}_f + (b_x, b_y) \tag{7}$$

where (a_x, b_x) and (a_y, b_y) are scaling parameters for horizontal and vertical dimensions, respectively, with blank areas filled in black.

For facial data training, we utilize aligned facial video-pose frame pairs $(\hat{\boldsymbol{p}}_f, \hat{\boldsymbol{v}}_f)$ to construct the dataset $\hat{\mathcal{D}}_f$ to fine-tune the base model $\mathcal{M}_p$. The loss function for this training phase is defined as follows:

$$\mathcal{L}_{face} = w \cdot \mathbb{E}_{t,\boldsymbol{c}_t,\epsilon\sim\mathcal{N}(0,1)} \left[\|\epsilon - \hat{\epsilon}(\boldsymbol{x}_t, t, \boldsymbol{c}_t)\|\right], \quad (8)$$

where w is a mask initialized at 1, belonging to $\mathbb{R}^{h\times w}$, then scaled along with $\boldsymbol{p}_f$ and $\boldsymbol{v}_f$, with the remaining areas filled with 0. This mask ensures that the model focuses on learning relevant facial regions. $\boldsymbol{x}_t$ represents noisy images sampled from aligned facial videos, and $\boldsymbol{c}_t$ denotes conditional controls obtained by encoding pose frame sequences $\hat{\boldsymbol{p}}_f$ corresponding to sampled video frames. More details of the training pipeline can be found in Figure 5.

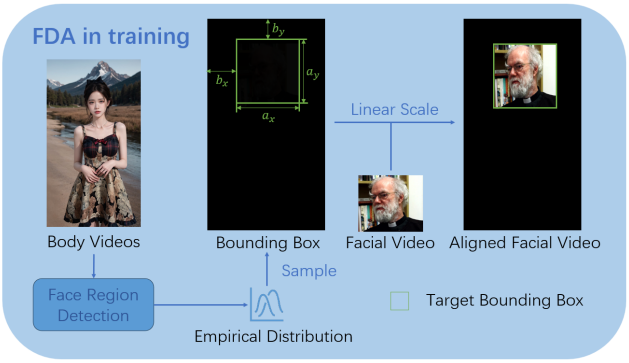

Figure 3: Diagram of FDA in training: Extracts an empirical distribution from body videos, samples scaling parameters from it, and linearly scales facial videos.

FDA in Inference During inference, we use the Audio2Pose module $\mathcal{M}_a$ from AniPortrait(Wei, Yang, and Wang 2024) to predict facial poses $\boldsymbol{p}_f$ based on given audio information $\boldsymbol{w}$ and facial portions r_f cropped from the reference image r. The final video generation process can be formalized as follows:

$$\boldsymbol{p}_f = \mathcal{M}_a\left(\boldsymbol{w}, r_f\right) \quad (9)$$

$$\boldsymbol{v} = \mathcal{M}_p\left(\mathrm{FDA}\left(\boldsymbol{p}_b, \boldsymbol{p}_f\right), r\right), \quad (10)$$

where $\mathrm{FDA}\left(\boldsymbol{p}_b, \boldsymbol{p}_f\right)$ represents the replacement of facial keypoints in pose frame sequences $\boldsymbol{p}_b$ detected in driving videos with linearly scaled facial keypoints from $\boldsymbol{p}_f$ using FDA technology.

Since the process needs to be controlled by driving videos, the scaling parameters for FDA here are not randomly sampled but determined based on detected facial regions in driving videos.

However, directly obtaining scaling parameters through facial region detection as in training data processing would cause misalignment between scaled facial keypoints and full-body keypoints, resulting in mismatched pose guidance.

Therefore, we employ a facial detection module(Wei, Yang, and Wang 2024) to extract facial keypoint coordinate sequences $\boldsymbol{k_b}$ from driving videos, while facial poses $\boldsymbol{p}_f$ correspond to facial keypoint coordinate sequences $\boldsymbol{k_f}$. We calculate minimum bounding boxes for both sets of keypoints using:

$$(x_{f,min}, x_{f,max}) = (\min_{\boldsymbol{k}_f} k_{f,x}, \max_{\boldsymbol{k}_f} k_{f,x}) \quad (11)$$

$$(y_{f,min}, y_{f,max}) = (\min_{\boldsymbol{k}_f} k_{f,x}, \max_{\boldsymbol{k}_f} k_{f,y}) \quad (12)$$

$$(x_{b,min}, x_{b,mbx}) = (\min_{\boldsymbol{k}_b} k_{b,x}, \max_{\boldsymbol{k}_b} k_{b,x}) \quad (13)$$

$$(y_{b,min}, y_{b,max}) = (\min_{\boldsymbol{k}_b} k_{b,y}, \max_{\boldsymbol{k}_b} k_{b,y}) \quad (14)$$

Subsequently, we perform alignment based on the calculated minimum bounding boxes, as shown in Figure 4, aligning the facial keypoint bounding boxes predicted by the Audio2Pose module frame-by-frame with those detected in driving videos. The scaling parameters (a_x, a_y) and the offset parameters (b_x, b_y) are calculated as:

$$a_x = \frac{x_{b,max} - x_{b,min}}{x_{f,max} - x_{f,min}} \quad (15)$$

$$a_y = \frac{y_{b,max} - y_{b,min}}{y_{f,max} - y_{f,min}} \quad (16)$$

$$b_x = x_{b,min} - a_x \cdot x_{f,min} \quad (17)$$

$$b_y = y_{b,min} - a_y \cdot y_{f,min} \quad (18)$$

Finally, replacing facial keypoints in original pose frame sequences $\boldsymbol{p}_b$ from driving videos with scaled facial keypoints yields full-body motion representation for model inference.

In general, the FDA ensures seamless integration of facial information from audio and full-body pose information from driving videos, enabling coherent video generation.

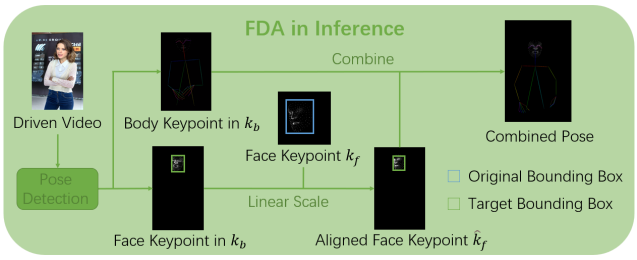

Figure 4: Diagram of FDA in inference: Detects pose keypoints from driven video, scales facial keypoints predicted by Audio2Pose module to match detected facial proportions, and combines them with body keypoints to form full-body motion representation.

Identity-consistent Control

Directly using reference images r sampled from $\hat{\boldsymbol{v}}_f$ during training may lead to training collapse, as discontinuous boundaries between scaled foreground content and filled areas could hinder the model's semantic understanding of reference identity features.

Thus, based on the open source model Diffusers Image Outpaint, we apply outpainting technology to generate naturally extended complete images $\hat{r}$ from reference images r, which are then used as reference images during training.

Simultaneously, to maintain the model's ability to generate identity-consistent full-body features when only facial keypoints are provided, we introduce reference training, in which each facial data training iteration involves predicting noise based on facial keypoints p_f corresponding to reference image r, according to it to approximate the denoising process of complete image $\hat{r}$. The loss function for this part is defined as follows:

$$\mathcal{L}_{refer} = \mathbb{E}_{t,c_t,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \hat{\epsilon}(\hat{x}_t, t, c_t)\|\right], \tag{19}$$

where x_t represents noisy images corresponding to the complete image $\hat{r}$, and c_t denotes conditional controls obtained by encoding facial keypoints p_f corresponding to the reference image r.

In implementation, as shown in Figure 5, the reference training process is combined with facial data training, where the outpainted full-width image $\hat{r}$ is treated as a single frame and combined with sampled video frames to participate in diffusion model noise addition and denoising processes.

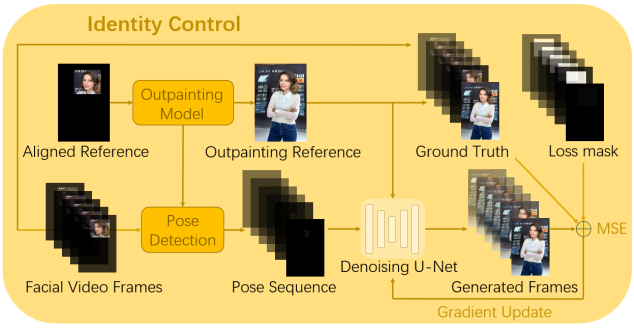

Figure 5: Diagram of Identity-consistent Control: Extends reference image based on the outpainting model, then integrates the reference training process with facial data training.

Mixed-Data Training Strategy

When fine-tuning the model solely using the aligned facial dataset $\hat{\mathcal{D}}_f$, as depicted in Figure 8, the model faces limitations in representing nonfacial features such as hands, background, and other body parts. This limitation arises from the structure of the specified loss functions, which prioritize generating facial features, leading to the neglect of other elements that remain identical to the reference images.

To alleviate this issue, we introduce a mixed-data training strategy that alternately fine-tunes the model on the facial dataset $\hat{\mathcal{D}}_f$ and full-body dataset $\hat{\mathcal{D}}_b$. The loss function for this part is defined as follows:

$$\mathcal{L}_{body} = \mathbb{E}_{t,\boldsymbol{c}_t,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \hat{\epsilon}(\boldsymbol{x}_t, t, \boldsymbol{c}_t)\|\right], \tag{20}$$

where $\boldsymbol{x}_t$ represents noisy images sampled from processed full-body videos, and $\boldsymbol{c}_t$ denotes conditional controls obtained by encoding pose frame sequences $\boldsymbol{p}_b$ corresponding to sampled video frames.

Facial keypoints in pose frames $\boldsymbol{p}_b$ corresponding to sampled frames from video $\boldsymbol{v}_b$ are obtained using the same pose detection module(Wei, Yang, and Wang 2024) as for facial data. Although full-body data lack sufficient facial details, maintaining the same format as $\hat{\mathcal{D}}_f$ ensures alignment in conditional feature space with facial data.

Since $\hat{\mathcal{D}}_f$ remains in the training process, this method minimally impacts the model's facial generation capability while significantly enhancing its ability to generate realistic outputs for other body parts and backgrounds.

In implementation, we adopt an iterative training scheme where each complete iteration includes: (1) an optimization step in $\hat{\mathcal{D}}_f$; (2) another optimization step in $\hat{\mathcal{D}}_b$. This approach ensures training stability while enabling balanced feature learning across both datasets. Generally, the complete training loss function can be expressed as

$$\mathcal{L} = \mathcal{L}_{face} + \beta \times \mathcal{L}_{refer} + \alpha \times \mathcal{L}_{body} \tag{21}$$

where $\mathcal{L}_{face}$, $\mathcal{L}_{refer}$, and $\mathcal{L}_{body}$ represent loss functions customized for facial data training, reference training, and full-body data training, respectively. Hyperparameters α and β play a crucial role in explicitly regulating the relative importance of full-body data in contrast to facial data within the optimization process, as well as determining the significance of reference training compared to facial data training.

Experiments

Implementation Details

The model was trained on video data with a resolution of 768×512 pixels using NVIDIA A100 GPU with 80 GB of memory. We used ControlNeXt(Peng et al. 2024) as the backbone architecture for our model. The training process consisted of 10 epochs with a batch size of 21 frames per iteration. For the loss function, the hyperparameters were set to $\alpha = 0.1$ and $\beta = 1.0$. For the FDA technique, the scaling parameters follow an empirical distribution with $\bar{a}_x = 0.5$, $\bar{a}_y = 0.35$, $\bar{b}_y = 0.1$, and $\delta = 0.05$.

The training dataset consists of: (1) Approximately 10,000 aligned facial portrait videos processed from the VFHQ dataset(Xie et al. 2022) (a high-quality, high-resolution facial video dataset); (2) Around 700 processed full-body portrait videos from the TikTokDataset(Jafarian and Park 2021) (a dataset containing diverse human body movements). For video frames for which corresponding poses cannot be detected, solid black images are substituted to maintain temporal continuity in training videos. This carefully constructed dataset ensures precise control of facial details and the diversity of full-body movements in the generated videos.

Evaluation Metrics

Given the lack of ground-truth data for audio-pose-driven full-body portrait videos and the weakly supervised nature of our method, we employ a series of ground-truth-free evaluation metrics to assess video quality. For video quality assessment, we use FID (Fréchet Inception Distance)(Heusel et al. 2018) and FVD (Fréchet Video Distance)(Unterthiner et al. 2018). Specifically, we calculate FID and FVD between feature distributions extracted from generated videos

and the full-body dataset $\mathcal{D}_b$ using image and video feature extraction networks. To evaluate audio-lip synchronization, SyncNet(Chung and Zisserman 2016) is used to calculate Sync-C and Sync-D metrics. For identity consistency, we use Deepface(Serengil and Ozpinar 2024) to compute CSIM (cosine similarity) between facial features extracted from reference images and generated video frames. To assess full-body pose and hand movements, we introduce AKD (Average Keypoint Distance), HKC (Hand Keypoint Confidence), and HKV (Hand Keypoint Variance) derived from pose estimation(Yang et al. 2023).

Qualitative Results

As the first effective solution for audio-pose-driven full-body portrait video generation, we compare APV-Gen with ControlNeXt(Peng et al. 2024) (state-of-the-art pose-driven method) and EchoMimicV2(Meng et al. 2024) (state-of-the-art audio-driven half-body method). Figure 6 shows qualitative comparisons under identical driving conditions.

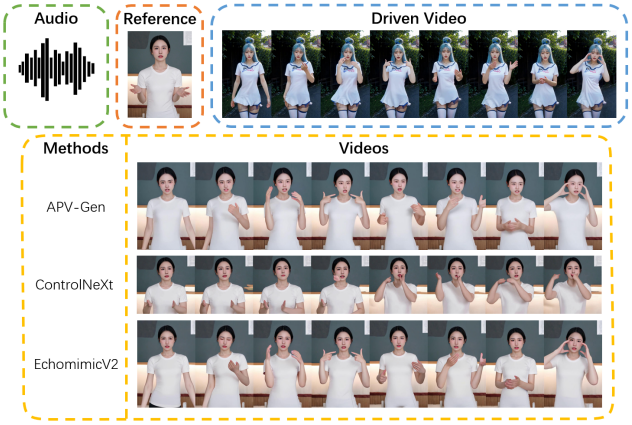

Figure 6: Qualitative comparison between APV-Gen and baseline methods: Identical driving conditions employed for all visual results.

Comparison with Pose-Driven Methods Compared to ControlNeXt's audio-independent facial generation, APV-Gen produces more natural lip movements synchronized with audio and richer emotional facial details, while maintaining comparable full-body motion quality.

Comparison with Audio-Driven Methods Compared to EchoMimicV2, APV-Gen generates higher quality hand and full-body movements. EchoMimicV2 exhibits artifacts in hand generation, especially around mouth-hand interactions, and fails to maintain full-body motion consistency due to limited training data.

Additional APV-Gen Results Figure 1 shows more APV-Gen results, demonstrating identity preservation, pose-consistent movements, and audio-aligned lip synchronization.

Quantitative Results

Table 1 presents quantitative comparisons, with bold highlighting the best performance and underline indicating the second best. The quantitative results illustrate that APV-Gen strikes a favorable balance between motion quality and audio-lip synchronization.

Compared to ControlNeXt, APV-Gen exhibits significant advantages in lip sync metrics (Sync-C / D), alongside improvements in video quality (FID/FVD), identity preservation (CSIM), and hand accuracy (HKC/HKV). In contrast to EchoMimicV2, APV-Gen demonstrates substantial improvements in video quality and pose accuracy metrics.

Ablation Studies

We analyze the effectiveness of key components through quantitative and qualitative results.

Facial Data Alignment Analysis From the results in Table 2, the FDA effectively bridges the granularity gap between facial data ($\hat{\mathcal{D}}_f$) and full-body data ($\hat{\mathcal{D}}_b$), improving most metrics. The slightly better FVD without FDA may be because FDA focuses more on facial details. Figure 7 shows the results generated without the FDA technique. It can be observed that the model struggles to learn unified generative capabilities from multi-modal data inputs, resulting in generated videos with blurred facial details, noticeable distortions, and degraded visual quality.

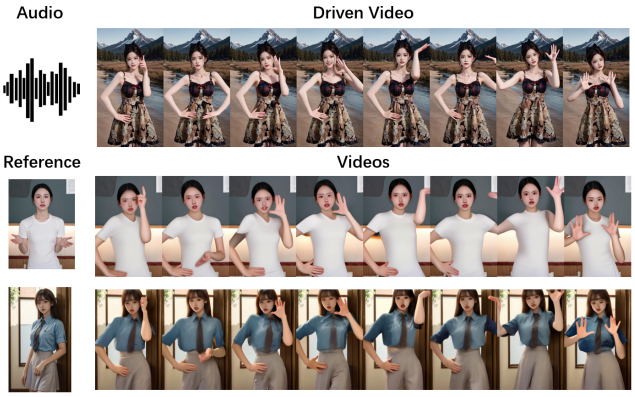

Figure 7: Results without FDA technique: Exhibits failure in unified multi-modal learning, leading to blurred facial details, noticeable distortions, and degraded video quality.

Mixed-Data Training Analysis The quantitative results in Table 2 demonstrate that the mixed-data training strategy successfully mitigates catastrophic forgetting in the backbone network. The strategy significantly outperforms models without mixed-data training in video quality metrics and pose accuracy metrics. Furthermore, improvements are observed in lip sync metrics and identity consistency metrics. This enhancement may stem from the fact that training without full-body portrait data not only compromises the model's ability to preserve pose generation capabilities but also impairs its original facial generation abilities, indicating potential overfitting to limited data patterns. Figure 8 provides qualitative results without the mixed-data training strategy. As shown, the generated portraits exhibit nearly static poses with noticeable motion artifacts in hand movements. In contrast, the implementation of the mixed-data training strategy

Metrics	Video quality		Lip sync		Identity consistency	Pose accuracy		
Method	FID↓	FVD↓	Sync-D↓	Sync-C↑	CSIM↑	AKD↓	HKC↑	HKV↑
ControlNeXt	746.80	228.88	11.91	0.70	0.58	**0.11**	0.64	7.70
Echomimicv2	987.30	393.54	**9.39**	**4.31**	**0.79**	0.21	0.55	7.73
APV-Gen	**632.79**	**219.54**	11.55	1.39	0.62	0.12	**0.71**	**8.14**

Table 1: Quantitative comparison of APV-Gen versus baseline methods: Bold highlighting the best performance and underline indicating the second best. ↑ represents that the higher the metric, the better, while ↓ represents the opposite.

Metrics	Video quality		Lip sync		Identity consistency	Pose accuracy		
Method	FID↓	FVD↓	Sync-D↓	Sync-C↑	CSIM↑	AKD↓	HKC↑	HKV↑
w/o FDA	706.25	**216.35**	12.20	0.87	0.54	0.13	0.67	8.12
w/o mixed-data training	1042.37	526.71	12.31	1.16	0.59	0.15	0.62	7.59
w/o reference training	712.28	231.25	11.95	1.18	0.52	0.13	0.65	7.98
APV-Gen	**632.79**	219.54	**11.55**	**1.39**	**0.62**	**0.12**	**0.71**	**8.14**

Table 2: Quantitative comparison of APV-Gen versus ablation variants: Bold highlighting the best performance and underline indicating the second best. ↑ represents that the higher the metric, the better, while ↓ represents the opposite.

effectively solves these issues.

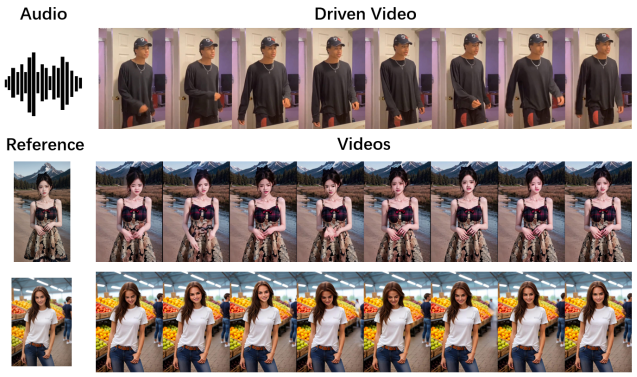

Figure 8: Results without mixed-data training: Generated portraits exhibit static poses and hand motion artifacts.

Reference Training Analysis Table 2 presents quantitative results for models trained without external inpainting techniques or reference images, demonstrating the degradation of identity consistency metrics and video quality metrics. Currently, performance declines are observed across lip sync metrics and pose accuracy metrics. This degradation likely stems from the model receiving semantically inconsistent reference images during training, which impacts the model's generative capability. Figure 9 visually confirms these limitations that both identity consistency and video realism exhibit noticeable deterioration when reference training is not used. Facial scaling inconsistencies and noticeable artifacts can be observed in the output videos.

Conclusion

In this work, we present APV-Gen, a novel weakly supervised training framework for generating realistic, full-body videos jointly driven by audio and pose signals building upon existing models and datasets. Our approach introduces FDA, a key innovation that effectively integrates audio-synchronized facial motions into pose-guided generation. To mitigate catastrophic forgetting during training, we propose a mixed-data learning strategy to enhance the model's stability. Extensive experiments have demonstrated that our method produces highly natural and expressive animations, achieving an optimal balance between motion fidelity and audio-lip synchronization compared to state-of-the-art methods.

Figure 9: Results without reference training: Both identity consistency and video realism exhibit noticeable deterioration when reference training is not used.

References

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chang, D.; Shi, Y.; Gao, Q.; Fu, J.; Xu, H.; Song, G.; Yan, Q.; Yang, X.; and Soleymani, M. 2023. MagicDance: Realistic Human Dance Video Generation with Motions & Facial Expressions Transfer. *arXiv preprint arXiv:2311.12052*.

Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lipreading, ACCV*.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. DensePose: Dense Human Pose Estimation In The Wild.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500.

Hu, L.; Gao, X.; Zhang, P.; Sun, K.; Zhang, B.; and Bo, L. 2023. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. *arXiv preprint arXiv:2311.17117*.

Jafarian, Y.; and Park, H. S. 2021. Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12753–12762.

Karras, J.; Holynski, A.; Wang, T.-C.; and Kemelmacher-Shlizerman, I. 2023. DreamPose: Fashion Image-to-Video Synthesis via Stable Diffusion. arXiv:2304.06025.

Lin, G.; Jiang, J.; Liang, C.; Zhong, T.; Yang, J.; Zheng, Z.; and Zheng, Y. 2025. CyberHost: A One-stage Diffusion Framework for Audio-driven Talking Body Generation. In *The Thirteenth International Conference on Learning Representations*.

Liu, H.; Yang, X.; Akiyama, T.; Huang, Y.; Li, Q.; Kuriyama, S.; and Taketomi, T. 2024. TANGO: Co-Speech Gesture Video Reenactment with Hierarchical Audio Motion Embedding and Diffusion Interpolation. arXiv:2410.04221.

Meng, R.; Zhang, X.; Li, Y.; and Ma, C. 2024. EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation. arXiv:2411.10061.

Peng, B.; Wang, J.; Zhang, Y.; Li, W.; Yang, M.-C.; and Jia, J. 2024. ControlNeXt: Powerful and Efficient Control for Image and Video Generation. *arXiv preprint arXiv:2408.06070*.

Serengil, S.; and Ozpinar, A. 2024. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Journal of Information Technologies*, 17(2): 95–107.

Shao, R.; Pang, Y.; Zheng, Z.; Sun, J.; and Liu, Y. 2024. Human4DiT: 360-degree Human Video Generation with 4D Diffusion Transformer. *ACM Transactions on Graphics (TOG)*, 43(6).

Shen, S.; Zhao, W.; Meng, Z.; Li, W.; Zhu, Z.; Zhou, J.; and Lu, J. 2023. DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation. In *CVPR*.

Tian, L.; Hu, S.; Wang, Q.; Zhang, B.; and Bo, L. 2025. EMO2: End-Effector Guided Audio-Driven Avatar Video Generation. arXiv:2501.10687.

Tong, Z.; Li, C.; Chen, Z.; Wu, B.; and Zhou, W. 2024. MusePose: a Pose-Driven Image-to-Video Framework for Virtual Human Generation. *arxiv*.

Tu, S.; Dai, Q.; Zhang, Z.; Xie, S.; Cheng, Z.-Q.; Luo, C.; Han, X.; Wu, Z.; and Jiang, Y.-G. 2024. MotionFollower: Editing Video Motion via Lightweight Score-Guided Diffusion. *arXiv preprint arXiv:2405.20325*.

Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*.

Wang, C.; Tian, K.; Zhang, J.; Guan, Y.; Luo, F.; Shen, F.; Jiang, Z.; Gu, Q.; Han, X.; and Yang, W. 2024a. V-Express: Conditional Dropout for Progressive Training of Portrait Video Generation.

Wang, T.; Li, L.; Lin, K.; Zhai, Y.; Lin, C.-C.; Yang, Z.; Zhang, H.; Liu, Z.; and Wang, L. 2023. Disco: Disentangled control for realistic human dance generation. *arXiv preprint arXiv:2307.00040*.

Wang, X.; Zhang, S.; Gao, C.; Wang, J.; Zhou, X.; Zhang, Y.; Yan, L.; and Sang, N. 2024b. UniAnimate: Taming Unified Video Diffusion Models for Consistent Human Image Animation. *arXiv preprint arXiv:2406.01188*.

Wei, H.; Yang, Z.; and Wang, Z. 2024. AniPortrait: Audio-Driven Synthesis of Photorealistic Portrait Animations. arXiv:2403.17694.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. VFHQ: A High-Quality Dataset and Benchmark for Video Face Super-Resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Xu, S.; Chen, G.; Guo, Y.-X.; Yang, J.; Li, C.; Zang, Z.; Zhang, Y.; Tong, X.; and Guo, B. 2024. VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time. arXiv:2404.10667.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.-W.; Zhang, C.; Feng, J.; and Shou, M. Z. 2023. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. In *arXiv*.

Yang, S.; Li, H.; Wu, J.; Jing, M.; Li, L.; Ji, R.; Liang, J.; Fan, H.; and Wang, J. 2024. MegActor-Σ: Unlocking Flexible Mixed-Modal Control in Portrait Animation with Diffusion Transformer. arXiv:2408.14975.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4210–4220.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

Zhang, Y.; Gu, J.; Wang, L.-W.; Wang, H.; Cheng, J.; Zhu, Y.; and Zou, F. 2024. MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance. *arXiv preprint arXiv:2406.19680*.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no)

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no)

2.4. Proofs of all novel claims are included (yes/partial/no)

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no)

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no)

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA)

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA)

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) partial

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyz-

ing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes