

FantasyTalking: Realistic Talking Portrait Generation via Coherent Motion Synthesis

Mengchao Wang*

AMAP, Alibaba Group

wangmengchao.wmc@alibaba-inc.com

Qiang Wang*

AMAP, Alibaba Group

yijing.wq@alibaba-inc.com

Fan Jiang†

AMAP, Alibaba Group

frank.jf@alibaba-inc.com

Yaqi Fan

Beijing University of Posts and
Telecommunications

yqfan@bupt.edu.cn

Yunpeng Zhang

AMAP, Alibaba Group

daoshi.zyp@alibaba-inc.com

Yonggang Qi‡

Beijing University of Posts and
Telecommunications

qiyg@bupt.edu.cn

Kun Zhao

AMAP, Alibaba Group

kunkun.zk@alibaba-inc.com

Mu Xu

AMAP, Alibaba Group

xumu.xm@alibaba-inc.com

ABSTRACT

Creating a realistic animatable avatar from a single static portrait remains challenging. Existing approaches often struggle to capture subtle facial expressions, the associated global body movements, and the dynamic background. To address these limitations, we propose a novel framework that leverages a pretrained video diffusion transformer model to generate high-fidelity, coherent talking portraits with controllable motion dynamics. At the core of our work is a dual-stage audio-visual alignment strategy. In the first stage, we employ a clip-level training scheme to establish coherent global motion by aligning audio-driven dynamics across the entire scene, including the reference portrait, contextual objects, and background. In the second stage, we refine lip movements at the frame level using a lip-tracing mask, ensuring precise synchronization with audio signals. To preserve identity without compromising motion flexibility, we replace the commonly used reference network with a facial-focused cross-attention module that effectively maintains facial consistency throughout the video. Furthermore, we integrate a motion intensity modulation module that explicitly controls expression and body motion intensity, enabling controllable manipulation of portrait movements beyond mere lip motion. Extensive experimental results show that our proposed approach achieves higher quality with better realism, coherence, motion intensity, and identity preservation. Ours project page: <https://fantasy-amap.github.io/fantasy-talking/>.

KEYWORDS

Diffusion Models, Video Generation, Talking Head

1 INTRODUCTION

Generating an animatable avatar from a single static portrait image has long been a fundamental challenge in computer vision and graphics. In particular, the ability to synthesize a realistic talking avatar given a reference image unlocks a wide range of applications

in gaming, filmmaking, and virtual reality. It is crucial that the avatar can be seamlessly controlled using audio signals, enabling intuitive and flexible manipulation of expressions, lip movements, and gestures to align with the desired content.

Early attempts [3, 14, 30, 37, 44, 50] to tackle this task mainly resort to 3D intermediate representations, such as 3D Morphable Models (3DMM) [41] or FLAME [27]. However, these approaches typically face challenges in accurately capturing subtle expressions and realistic motions, which significantly limits the quality of the generated portrait animations. Recent research [4, 7, 21, 40, 46] has increasingly focused on creating talking head videos using diffusion models, which show great promise in generating visually compelling content that adheres to multi-modal conditions, such as reference images, text prompts, and audio signals. However, the realism of the generated videos remains unsatisfactory. Existing methods typically focus on tame talking head scenarios, achieving precise audio-aligned lip movements while neglecting other related motions, such as facial expressions and body movements, both of which are essential for producing smooth and coherent portrait animations. Moreover, the background and contextual objects usually remain static throughout the animation, which makes the scene less natural.

In this work, we leverage pretrained video diffusion transformer models to generate highly realistic and visually coherent talking portraits. In essence, we propose a multi-modal alignment framework built on the DiT-based video generation model to encourage unified dynamics across the whole scene, encompassing the reference portrait, associated contextual objects, and the background. Technically, we propose a dual-stage audio-visual alignment strategy to facilitate portrait video generation. In the first stage, leveraging the powerful temporospatial modeling capabilities of the DiT-based model, we devise a clip-level training to capture diverse implicit connections between the audio and visual dynamics across the entire clip. This enables an overall coherent generation of global motion. Lip movements are critical for enhancing the quality of the portrait video. However, the lip typically only occupies a small region in a frame, so it is challenging to precisely align lip movements with the audio signals on the entire frame. Therefore, in the second

*Equal contribution

†Project leader

‡Corresponding author

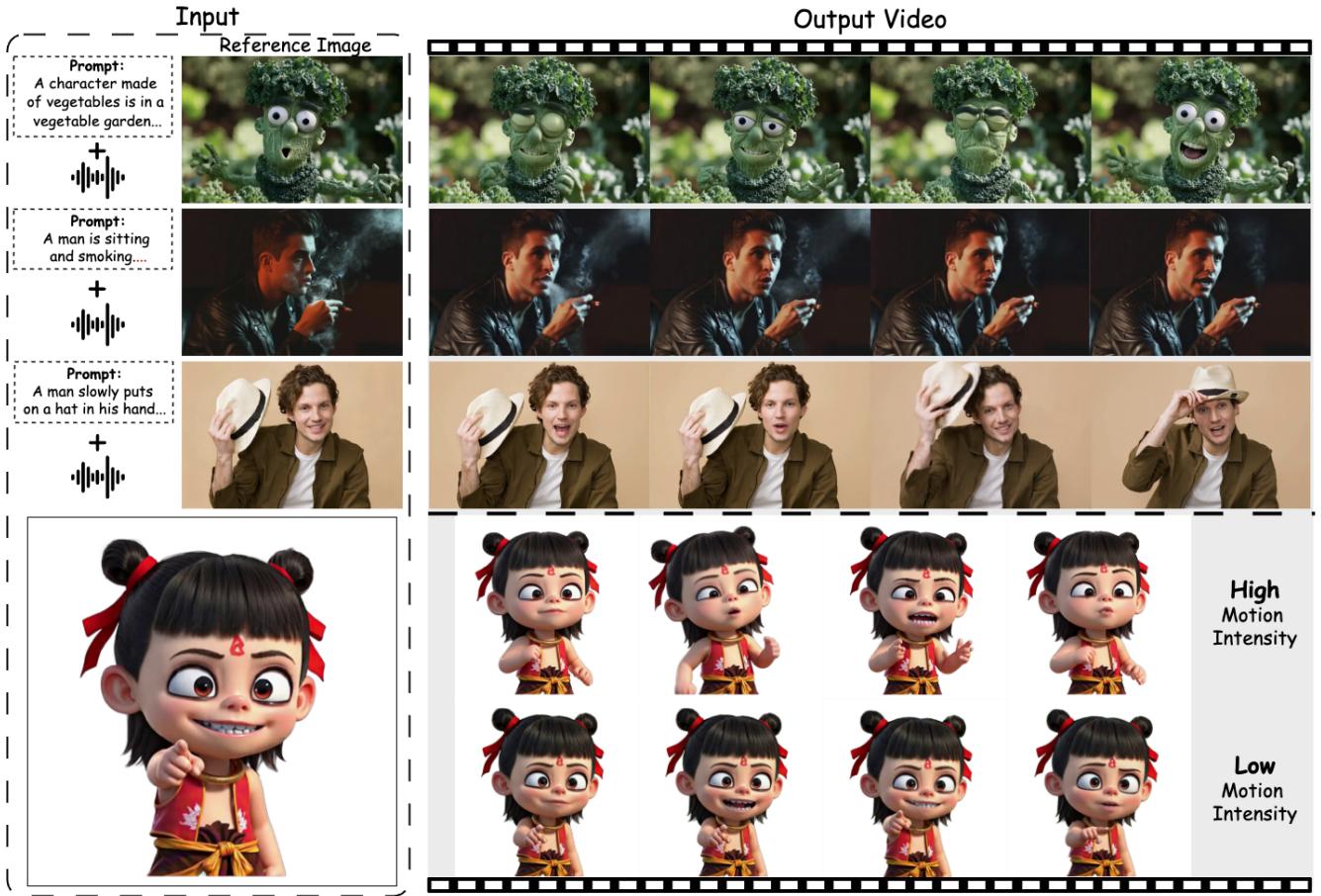


Figure 1: Given a portrait image, voice and text, FantasyTalking can generate animated portraits with rich expressions, natural body movements, and identity features. In addition, FantasyTalking can control the motion intensity of animated portraits. Please refer to our supplementary materials for the video results.

stage, we learn the attention of visual tokens mapped from audio tokens and employ a mask that enforces the refinement of lip movements, ensuring they adhere more closely to the audio content at the frame level. Moreover, we avoid using the commonly adopted reference network for identity preservation. We found out that such an approach typically references the entire image and severely restricts the dynamic effects of the portrait. Instead, we reveal that a cross-attention module focusing on facial modeling effectively ensures identity consistency throughout the video. Lastly, we introduce a motion intensity conditioning module that decouples the character’s expressions and body movements, thereby enabling the manipulation of motion intensity in the generated dynamic portrait.

In summary, our contributions are as follows:

- We devise a dual-stage audio-visual alignment training strategy to adapt a pretrained video generation model to first establish coherent global motions involving background and contextual objects other than the portrait itself, corresponding to input audio at clip level, then construct precisely

aligned lip movements to further improve the quality of the generated video.

- Instead of adopting the conventional reference network for identity preservation, we streamline the process by devising a facial-focused cross-attention module that concentrates on modeling facial regions and guides the video generation with consistent identity.
- We integrate a motion intensity modulation module that explicitly controls facial expression and body motion intensity, enabling controllable manipulation of portrait movements beyond mere lip motion.
- Extensive experiments demonstrate that our proposed approach achieves new SOTA in terms of video quality, temporal consistency, and motion diversity.

2 RELATED WORK

2.1 Diffusion-Based Video Generation

The remarkable achievements of diffusion models in image generation [12, 13, 33] have inspired extensive research into video

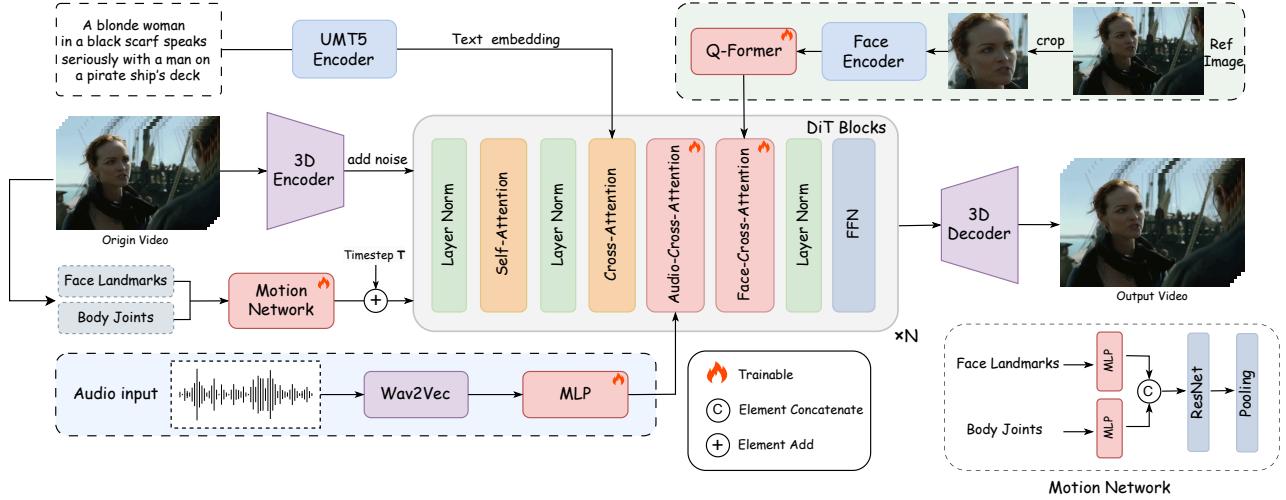


Figure 2: Overview of FantasyTalking.

generation [19, 24, 36]. Early methods employing diffusion models predominantly relied on the UNet architecture[34], with notable examples being AnimateDiff [15] and Stable Video Diffusion [1]. These approaches, leveraging pretrained image generation models, harness their robust spatial generation capabilities and incorporate specifically designed temporal layers to acquire motion-related understanding. More recently, models based on the DiT architecture [31] have significantly propelled the advancement of video generation technology [24, 38, 45, 47]. These models employ 3D VAE [22] as the encoder and decoder, coupled with the Transformer’s formidable sequence modeling prowess, showcasing substantial potential in tackling intricate video generation tasks. They have demonstrated impressive capabilities in maintaining human identity [49, 51], controlling expressions [32], and virtual try-on [53] applications, among others.

2.2 Audio-driven Talking head Generation

The task of synthesizing realistic talking face videos from input audio has remained a persistent research focus. Early approaches [30, 44, 50] employed 3D intermediate representations, utilizing facial animation parameters derived from 3D Morphable Models (3DMM) as guidance for video generation. However, the limited expressiveness of 3DMM in capturing intricate facial expressions and head movements significantly constrained the authenticity and naturalness of synthesized videos. In contrast, emerging end-to-end audio-to-video synthesis methods [4, 8, 21, 40] demonstrate enhanced potential, yet still face two critical challenges. Firstly, existing approaches typically employ reference networks initialized from backbone architectures to preserve speaker identity, and the input of the reference network is the whole image rather than focusing on the face, which inadvertently restricts the model’s capacity to generate videos with broader motion ranges. Secondly, although prior methods have emphasized precise audio-lip synchronization, the inherent weak correlations between audio signals and other

facial expressions and body movements remain largely underexplored. Despite Hallo3 initially progress in the wild talking head task, the areas of facial-focused identity preservation and complex scene interaction are yet to be thoroughly explored.

3 METHOD

Given a single reference image, a driving audio and a prompt, FantasyTalking is designed to generate the video synchronized with the audio while ensuring that the identity characteristics of the person are maintained during their actions. An overview of FantasyTalking is illustrated in Figure 2. We investigate a Dual-Stage method to maintain audio-to-visual alignment when injecting audio signals (Sec. 3.2). Additionally, we employ an identity learning method to preserve the identity characteristics in the video (Sec. 3.3) and a motion network to control the expressions and the motion intensity (Sec. 3.4). The following section (Sec. 3.1) elaborates on the preliminaries of our method.

3.1 Preliminaries

Latent Diffusion Model. Our method is built upon the Latent Diffusion Model (LDM), which is a framework that learns in the latent space rather than the pixel space. During training, we use a pre-trained VAE encoder E to compress video data x from the pixel space into latent tokens $z = E(x)$. During training, the Gaussian noise ϵ is progressively added to z to create $z_t = \sqrt{\alpha_t}z + \sqrt{1 - \alpha_t}\epsilon$ at t timestep. Here, α_t represents the noise scheduler. The training objective of the LDM focuses on a reconstruction loss that aims to minimize the difference between the added noise and the noise predicted by the network ϵ_θ :

$$L = \mathbb{E}_{t, z_t, c, \epsilon \sim N(0, 1)} [\|\epsilon_\theta(z_t, t, c) - \epsilon\|_2^2] \quad (1)$$

where c denotes the conditions like audio, text or images. In the inference phase, the model iteratively denoises latent sampled from a Gaussian distribution. Subsequently, the denoised latent

representations are decoded back into videos using the VAE decoder D .

Diffusion Transformer. The Diffusion Transformer (DiT) [31] is a diffusion model designed based on the Transformer architecture [43], showcasing significant potential in the field of video generation. Specifically, we adopt Wan2.1 [38] as the foundational architecture. This model employs a causal 3D VAE to compress videos both temporally and spatially, while utilizing UMT5 [5] to encode textual information, yielding the text-conditioned input c_{text} . The text embeddings are then integrated into the DiT through cross-attention mechanisms. In addition, the embeddings of the timestep t are injected into the model by predicting six modulation parameters individually.

3.2 Dual-Stage Audio-Visual Alignment

Audio-Visual Alignment. We utilize Wav2Vec [35] to extract audio tokens containing multi-scale rich acoustic features. As shown in Figure 3, the audio tokens length l differs from that of the video tokens length ($f \times h \times w$), where f , h and w are the frame numbers, height and width of latent videos. There exists a one-to-one mapping relationship between these two token sequences. The task of tame talking head video generation typically focuses on the frame-level alignment of lip movements. However, wild talking head generation requires attention not only to the lip movements that are directly correlated with the audio but also to the movements of other facial components and body parts that are weakly correlated with the audio features, such as eyebrows, eyes, and shoulders. These movements are not strictly temporally aligned with the audio. To address this, we propose a Dual-Stage Audio-Vision Alignment approach. In the first training stage, we learn visual features related to the audio at the clip level. In the second training stage, we focus on the visual features that are highly correlated with the audio at the frame level.

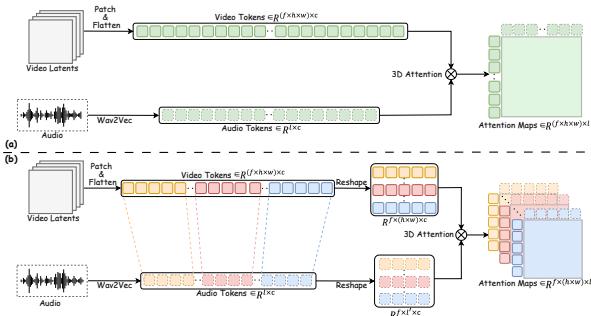


Figure 3: Dual-Stage Audio-Visual Alignment.

Clip-Level Training. As illustrated in Figure 3(a), the first training stage computes 3D full attention correlations across full-length audio-visual token sequences at the clip level, establishing global audiovisual dependencies while enabling holistic feature fusion. While this stage enables joint learning of both weakly audio-correlated non-verbal cues (e.g., eyebrow movements, shoulder motions) and strongly audio-synchronized lip dynamics, but the model struggles to learn precise lip movements. This is due to the fact that the lips

occupy only a small portion of the entire visual field, while the video sequence is highly correlated with the audio in each frame.

Frame-Level Training. In the second training stage, as depicted in Figure 3(b), we focus exclusively on lip-centric motion refinement through frame-exact audio-visual alignment. We segment the audio and videos according to a one-to-one mapping relationship, reshape the video tokens into the shape of $f \times (h \times w) \times c$ and the audio tokens into the shape of $f \times l' \times c$, where c represents the number of channels. Subsequently, we compute the 3D full attention between these tokens, ensuring that the visual features attend only to their corresponding audio features.

Additionally, in order to focus the attention on the lip area, we leverage MediaPipe [29] to extract precise lip masks in pixel space, which are then projected into the latent space via trilinear interpolation, forming our lip-focused constraint mask M . The frame-level loss in Eq. 1 is thus reweighted as:

$$L_c = M \odot L \quad (2)$$

where \odot denotes element-wise multiplication. However, exclusive reliance on lip-specific constraints risks over-regularization, suppressing natural head movements and background dynamics. To mitigate this issue, we employ a probability η to control the application of the constraint, allowing the model to balance between focusing on lip movements and maintaining the naturalness of overall movements.

$$L' = \begin{cases} L_c, & \text{if } p > \eta \\ L, & \text{otherwise} \end{cases} \quad (3)$$

3.3 Identity Preservation

While audio conditioning effectively establishes correlations between acoustic inputs and character motions, prolonged video sequences and intensified movements often lead to rapid identity degradation in synthesized results. Previous methods [4, 8, 21, 40] typically employ reference networks initialized from the backbone model to preserve identity characteristics, yet these methods exhibit two critical limitations. Firstly, the reference network processes full-frame images rather than facial regions of interest, biasing the model towards generating static backgrounds and motions with constrained expressiveness. Secondly, the reference network model typically has a network structure similar to that of the backbone model, resulting in a high degree of redundancy in their feature representation capabilities, and increases the computational load and complexity of the model.

To address this issue, we propose an identity preservation method to maintain consistency of facial features. Specifically, we first crop the facial region from the reference image [11] to ensure that the model only focuses on identity related facial regions. Subsequently, we utilize ArcFace [10] to extract the facial feature and then employ Q-Former [26] for alignment, resulting in the ID embedding F_{id} . Similar to audio conditioning, these identity features interact with each pretrained DiT attention block through dedicated cross-attention layers. Formally, the hidden state Z_i of each DiT block is reformulated as:

$$Z'_i = Z_i + \lambda_1 * \text{Attention}(Q_i, K_i^a, V_i^a) + \lambda_2 * \text{Attention}(Q_i, K_i^{id}, V_i^{id}) \quad (4)$$

where i represents the layer number of the attention block, Q_i is query matrices, K_i^a and K_i^{id} are the audio and identity key matrices, V_i^a and V_i^{id} are the audio and identity values matrices of the attention operation. The hyperparameters λ_1 and λ_2 control the relative contributions of audio and identity conditioning.

3.4 Motion Intensity Modulation Network

Individual speaking styles exhibit significant variations in facial expressions and body movement amplitudes, which cannot be explicitly controlled solely through audio and identity conditioning. Particularly in the context of wild talking head scenarios, the character’s expressions and body movements are more varied and dynamic compared to tame talking head scenarios. Therefore, we introduce a motion intensity modulation network to govern these dynamics.

Specifically, we utilize Mediapipe [29] to extract the variance of facial landmark keypoint sequences, denoted as facial expression movement coefficient ω_l , and DWPose [48] to compute the variance of body joint sequences, denoted as body movement coefficient ω_b . Both ω_l and ω_b are normalized to the range $[0, 1]$, representing the intensity of facial expressions and body movements, respectively. As illustrated in Figure 2, motion intensity modulation network consists of MLP layers, a ResNet layer [16], and an average pooling layer. The resulting motion embeddings are added with the timesteps. During inference stage, users are allowed to customize the input coefficient ω_l and ω_b to control the amplitude of facial and body motion intensity.

4 EXPERIMENTS

4.1 Setups

Implementation Details. We adopt Wan2.1-I2V-14B [38] as the foundational model. During the clip-level training stage, we train for approximately 80,000 steps, and during the frame-level training stage, we train for approximately 20,000 steps. Throughout all training phases, both the identity network and the motion network are incorporated into end-to-end training. We employ Flow Matching [28] to train the model, with the entire training conducted on 64 A100 GPUs. The learning rate is set to 1e-4. λ_1 is set to 1, λ_2 is set to 0.5, and η is set to 0.2. To enhance video generation variability, the reference image, guiding audio and prompt are each set to be independently discarded with a probability of 0.1. In the inference stage, we employ the sampling steps of 30, the motion intensity parameter ω_l and ω_b are set to neutral value of 0.5, and the CFG [18] of audio is set to 4.5.

Datasets. The training dataset we use consists of three parts: Hallo3 [8], CelebV-HQ [54], and data collected from the internet. We utilize InsightFace [9, 10] to exclude videos with a facial confidence score below 0.9 and remove clips [6] where the speech and mouth motion are not synchronized. This filtering process results in approximately 150,000 clips. We use 50 clips from the HDTF [52] for evaluating the tame talking head generation. Additionally, we evaluate our model on the collected wild talking dataset containing 80 different individuals.

Evaluation Metric and Baselines. We employ eight metrics for evaluation. Frechet Inception Distance (**FID**) [17] and Fréchet Video Distance (**FVD**) [42] are used to assess the quality of the generated data. **Sync-C**[6] and **Sync-D**[6] is utilized to measure the synchronization between audio and lip movements. The Expression Similarity (**ES**) method extracts facial features between video frames [11] and calculates the similarity between these features to evaluate the preservation of identity characteristics. ID consistency (**IDC**) is achieved by extracting the facial region and computing the DINO [2] similarity metric between frames to measure the consistency of the character’s identity features. We utilize SAM [23] to segment the frame into foreground and background, and separately measure the optical flow scores [39] for the foreground and background to evaluate Subject Dynamics (**SD**) and Background Dynamics (**BD**), respectively. **Aesthetic** quality is evaluated using the LAION aesthetic predictor [25] to assess the artistic and aesthetic value of videos.

We have selected several state-of-the-art methods to evaluate our approach, all of which have publicly available code or implementations. These methods include the UNet-based approaches Aniportrait [44], EchoMimic [4] and Sonic [20], as well as the DiT-based method Hallo3 [8]. For fair comparison, our method sets the prompt to empty during inference.

4.2 Results and Analysis

Comparison on Tame Dataset. The tame talking head dataset features limited variability in background and character poses, with a primary focus on lip synchronization and facial expression accuracy. Table 1 and Figure 4 present the evaluation results. Our method achieves the best scores in FID, FVD, IDC, ES, and Aesthetic score. This success is mainly attributed to our model’s ability to generate videos with the most natural and expressive facial expressions, resulting in the highest quality and aesthetically pleasing video outcomes. Additionally, our method achieves the best or second-best results in Sync-C and Sync-D, indicating that our DAVA approach enables the model to learn accurate audio synchronization.

Comparison on Wild Dataset. Table 1 and Figure 5 present the evaluation results on the wild talking head dataset, which includes significant variations in both foreground and background elements. Previous methods heavily rely on reference images, which limits the naturalness of the generated facial expressions, head movements, and background dynamics. In contrast, our method achieves the best results across all metrics, producing outputs with more natural variations in both foreground and background, improved lip synchronization, and higher overall video quality. This performance is primarily due to our DAVA approach and the identity preservation method focused on facial features. These methods enable our model to better understand the input audio, thereby generating more complex and natural head and background movements while preserving the character’s identity features. As a result, our approach better meets the demands of practical application scenarios.

Comparison of Motion Intensity Controller with Sonic. In our comparative study, Sonic exhibits a similar ability to control motion intensity, allowing users to regulate the expressiveness and head movement through an input parameter β . We conducted comparative experiments by categorizing the motion intensity into three

Dataset	Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	ES↑	IDC↑	SD↑	BD↑	Aesthetic↑
Tame Talking	Aniportrait	37.672	397.114	1.095	12.461	0.9508	0.9372	4.639	-	0.5129
	EchoMimic	33.765	471.452	2.514	10.743	0.9527	0.9419	5.783	-	0.5108
	Sonic	30.396	358.023	4.197	9.103	0.9595	0.9885	8.832	-	0.5312
	Hallo3	32.617	347.358	4.060	9.371	0.9566	0.9774	8.415	-	0.5247
	FantasyTalking	27.695	301.173	4.226	9.251	0.9612	0.9892	11.745	-	0.5362
Wild Talking	Aniportrait	63.574	841.962	0.996	12.084	0.9318	0.9031	2.252	1.9287	0.5357
	EchoMimic	59.746	590.373	1.949	10.754	0.9463	0.9202	3.201	1.9508	0.5311
	Sonic	45.400	489.985	2.689	10.194	0.9539	0.9607	10.484	3.9019	0.5913
	Hallo3	47.403	488.499	2.673	10.292	0.9420	0.9538	11.411	5.2840	0.5842
	Ours	43.137	483.108	3.154	9.689	0.9589	0.9754	13.783	7.9624	0.6183

Table 1: Comparison of different methods on tame and wild talking head datasets. The best results are highlighted in bold.

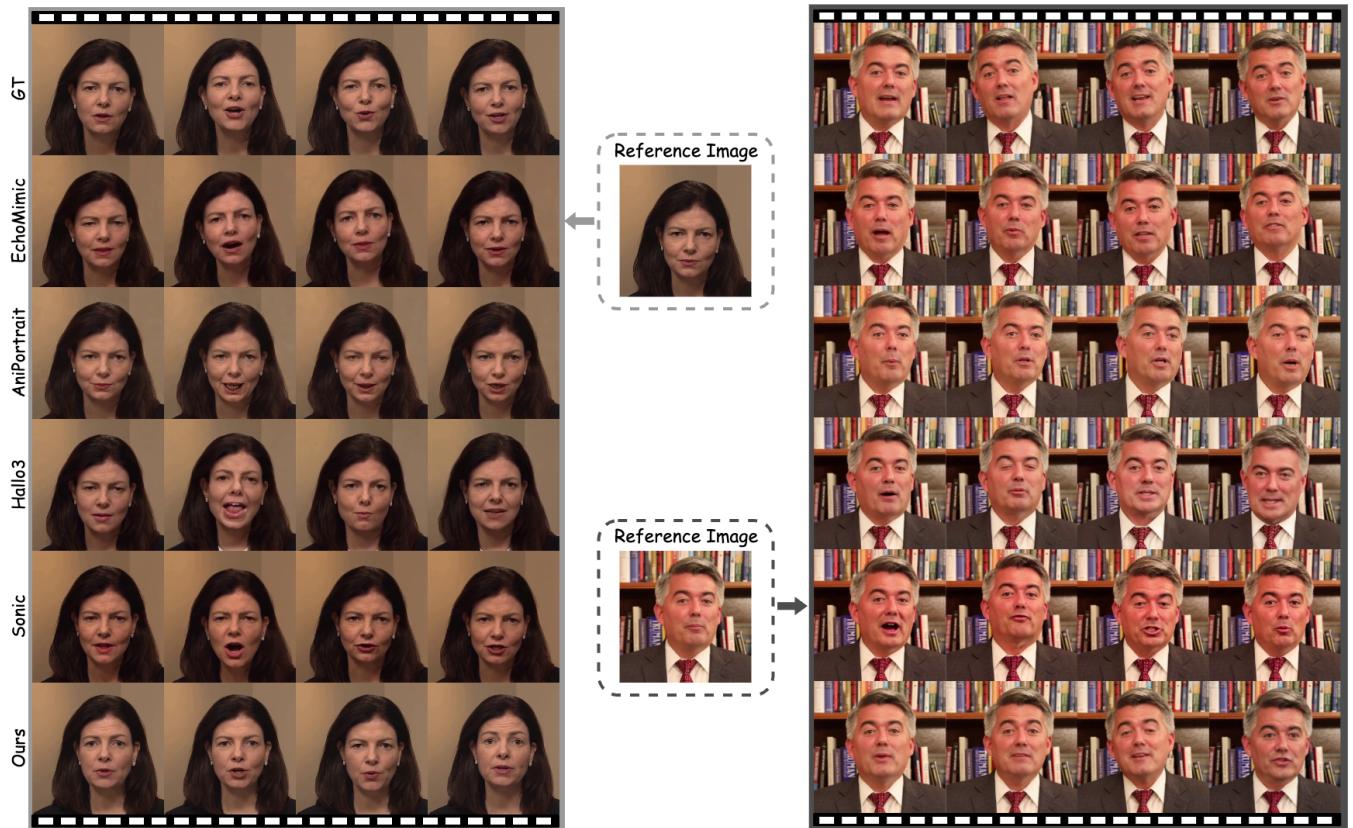


Figure 4: Qualitative comparison on tame talking head dataset (HDTF).

levels: subtle ($\beta=0.5, \omega_l=0.1, \omega_b=0.1$), natural ($\beta=1.0, \omega_l=0.5, \omega_b=0.5$) and intense ($\beta=2.0, \omega_l=1.0, \omega_b=1.0$). The experimental results are presented in Table 2 and Figure 6. At the natural and subtle levels, both our method and Sonic demonstrate excellent control over motion intensity while maintaining lip synchronization. However, in scenarios involving intense movements, our method achieves superior results. This is because our limb control approach focuses on the entire body movement, including the head, whereas Sonic only considers head movements. Consequently, our method exhibits a

more competitive ability in representing the full range of human motion.

Comparison of Visualization Results with Hallo3. We present additional visualization comparisons with Hallo3 in Figure 7, which is a DiT-based method for generating wild talking head videos. Our approach demonstrates more realistic results. For instance, the outputs of Hallo3 exhibit noticeable distortions and artifacts on the person's face and lips, as well as unrealistic background movements in the top row of 7, and relatively stiff head movements



Figure 5: Qualitative comparison on wild talking head dataset.



Figure 6: Comparison of Motion Intensity Controller with Sonic.

in the bottom row of 7. In contrast, our results showcase more authentic expressions, head movements, and background dynamics. These improvements can be attributed to our focus on facial knowledge learning, which enhances the identity features of the person, and the DAVA method, which strengthens the learning of lip synchronization.

User Studies. To further validate the effectiveness of our proposed method, we conducted a subjective evaluation on the Wild Talking Head dataset. Each participant assessed four critical dimensions:

Level	Method	FVD↓	Sync-C↑	Sync-D↓	IDC↑	SD↑
subtle	Sonic	508.66	2.64	11.23	0.978	8.32
	Ours	496.22	3.11	10.04	0.982	8.12
natural	Sonic	489.99	2.69	10.19	0.988	10.48
	Ours	483.11	3.15	9.69	0.989	13.78
intense	Sonic	522.78	2.06	12.59	0.971	12.32
	Ours	501.67	3.09	9.81	0.980	18.14

Table 2: Comparison of Motion Intensity Controller with Sonic.

Lip Synchronization (**LS**), Video Quality (**VQ**), Identity Preservation (**IP**), and Motion Diversity (**MD**). A total of 24 participants rated each aspect on a scale from 0 to 10. As shown in Table 3, the scores demonstrate that FantasyTalking outperforms baseline methods across all evaluated dimensions, exhibiting particularly notable improvements in motion diversity. This comprehensive evaluation highlights the superiority of our approach in generating realistic and diverse talking head animations while maintaining consistent identity representation and high visual fidelity.

Method	LS	VQ	IP	MD
Aniportrait	8.18	6.78	7.82	5.28
EchoMimic	8.22	6.31	7.05	4.40
Sonic	9.07	8.17	8.13	6.25
Haloo3	8.93	7.89	7.82	6.44
FantasyTalking	9.45	9.18	8.44	9.81

Table 3: User Study results.



Figure 7: Comparison of Visualization Results with Hallo3.



Figure 8: Ablation on DAVA.

5 ABLATION STUDIES AND DISCUSSION

Method	FVD↓	Sync-C↑	Sync-D↓	IDC↑	SD↑
Clip-Level	492.85	1.98	11.21	0.986	13.66
Frame-Level	534.39	3.54	9.02	0.987	8.22
w/o Identity	510.62	3.06	10.15	0.945	12.96
FantasyTalking	483.11	3.15	9.69	0.989	13.78

Table 4: Ablation studies on DAVA and Identity Preservation in Wild Dataset.

Ablation on DAVA. To validate the effectiveness of our DAVA method, we performed experiments using audio-visual alignment at clip level and only at frame level for training. The results, as presented in Table 4 and illustrated in Figure 8. Training with only clip-level alignment leads to a significant decline in the Sync-C metric. This indicates that relying solely on clip-level alignment is insufficient to learn the precise correspondence between audio and lip movements. However, training with only frame-level alignment, while demonstrating strong lip-sync capabilities, noticeably limits the dynamic nature of facial expressions and subject movements. In contrast, our proposed DAVA method effectively combines the

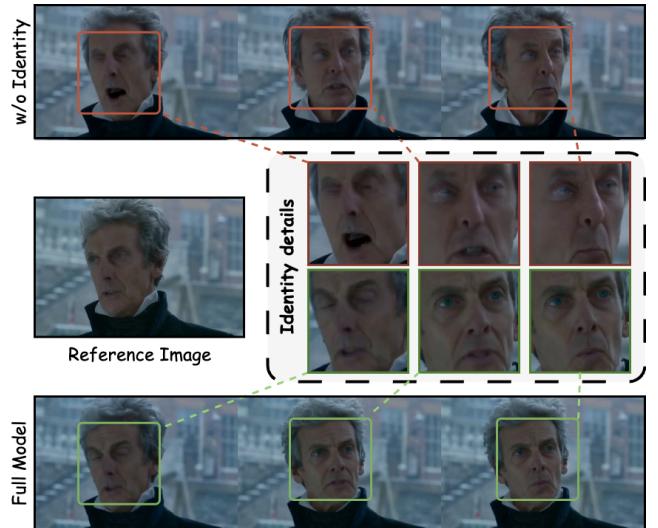


Figure 9: Ablation on Identity Preservation.

advantages of both clip-level and frame-level alignments, which achieves precise audio-to-lip synchronization while enhancing the vividness of character animations and background dynamics.

Ablation on Identity Preservation. The results presented in Table 4 underscore the importance of identity preservation in our model. Without identity preservation, IDC significantly decreases, which implies that the model’s ability to maintain the character’s identity features is greatly reduced, leading to a decline in video quality. As shown in Figure 9, the absence of identity preservation lead to artifacts and distortions in the facial features. In contrast, our proposed identity preservation method, which incorporates focused facial knowledge learning, enhances the model’s ability to maintain the character’s identity while preserving lip synchronization and rich motion capabilities. This leads to improved identity retention and overall video quality.

Ablation on Motion Intensity Modulation Network. Figure 10 illustrates the quantitative results of adjusting the motion intensity coefficient ω_l and ω_b on FVD and SD. When one parameter is varied, the other is fixed at a neutral value of 0.5. As shown in Figure 10 (a), the results with natural motion intensity ($\omega_l = 0.5$, $\omega_b = 0.5$) achieve the best FVD scores. This suggests that facial and body motion intensities that are either too high or too low tend to produce visual representations that deviate from realistic scenarios, which result in less authentic visual representations. Figure 10(b) demonstrates that as the ω_l or ω_b parameters increase, the subject dynamic score becomes significantly more pronounced. This highlights the effectiveness of our motion control mechanism, which provides users with a tool for explicitly controlling the speaking motion intensity.

Limitations and Future Works. Despite the significant progress achieved by our method, especially in the scenario of wild talking head video generation, due to the iterative sampling process required by the diffusion model during inference to achieve optimal results, the overall runtime can be relatively slow. Investigating acceleration strategies would facilitate its use in scenarios with

higher real-time requirements, such as live streaming and interactive real-time applications. Furthermore, investigating interactive portrait dialogue solutions with real-time feedback based on audio-driven talking head generation can broaden applications in realistic digital human avatar scenarios.

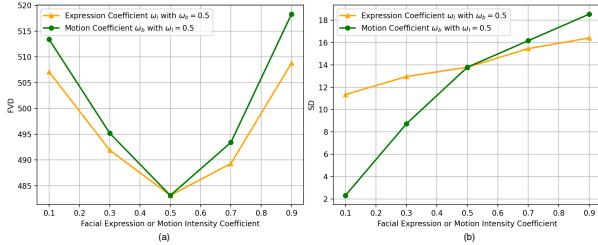


Figure 10: Ablation on Motion Intensity Modulation Network.

6 CONCLUSIONS

In this paper, we introduce FantasyTalking, a novel audio-driven portrait animation technique. By employing a dual-stage audio-visual alignment training process, our method effectively captures the relationship between audio signals and lip movements, facial expressions, as well as body motions. To enhance identity consistency within the generated videos, we propose a facial-focused approach to retain facial features accurately. Additionally, a motion network is utilized to control the magnitude of facial expressions and body movements, ensuring natural and varied animations. Both qualitative and quantitative experiments demonstrate that FantasyTalking outperforms existing SOTA methods in several key aspects, including video quality, motion diversity, and identity consistency.

REFERENCES

- [1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.
- [3] Aggelia Chatzagiapi, ShahRukh Athar, Abhinav Jain, Rohith MV, Vimal Bhat, and Dimitris Samaras. 2023. Lipnerf: What is the right feature space to lip-sync a nerf?. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.
- [4] Zhiyuan Chen, Jiajiong Cao, Zhiqian Chen, Yuming Li, and Chenguang Ma. 2024. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136* (2024).
- [5] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151* (2023).
- [6] Joon Son Chung and Andrew Zisserman. 2017. Out of time: automated lip sync in the wild. In *Computer Vision-ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part II* 13. Springer, 251–263.
- [7] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. 2024. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718* (2024).
- [8] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. 2024. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv preprint arXiv:2412.00733* (2024).
- [9] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaceface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5203–5212.
- [10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.
- [11] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.
- [12] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- [13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- [14] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. 2023. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1505–1515.
- [15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- [18] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- [19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.
- [20] Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. 2024. Sonic: Shifting Focus to Global Audio Perception in Portrait Animation. *arXiv preprint arXiv:2411.16331* (2024).
- [21] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. 2024. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. In *The Thirteenth International Conference on Learning Representations*.

- [22] Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.
- [23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026.
- [24] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanyvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).
- [25] LAION-AI. 2023. LAION-AI/aesthetic-predictor: A linear estimator on top of CLIP to predict the aesthetic quality of pictures. <https://github.com/LAION-AI/aesthetic-predictor>
- [26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [27] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.
- [28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).
- [29] Camillo Lugaressi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
- [30] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. 2023. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767* 2, 3 (2023).
- [31] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
- [32] Di Qiu, Zhengcong Fei, Rui Wang, Jialin Bai, Changqian Yu, Mingyuan Fan, Guibin Chen, and Xiang Wen. 2025. SkyReels-A1: Expressive Portrait Animation in Video Diffusion Transformers. *arXiv preprint arXiv:2502.10841* (2025).
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer, 234–241.
- [35] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* (2019).
- [36] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
- [37] Linsen Song, Wayne Wu, Chaoyou Fu, Chen Change Loy, and Ran He. 2022. Audio-driven dubbing for user generated contents via style-aware semi-parametric synthesis. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 3 (2022), 1247–1261.
- [38] Wan Team. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. (2025).
- [39] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II* 16. Springer, 402–419.
- [40] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. 2024. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*. Springer, 244–260.
- [41] Luan Tran and Xiaoming Liu. 2018. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7346–7355.
- [42] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation. (2019).
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [44] Huawei Wei, Zejun Yang, and Zhisheng Wang. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694* (2024).
- [45] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. 2024. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991* (2024).
- [46] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. 2024. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801* (2024).
- [47] Zhuoyi Yang, Jiany Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).
- [48] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4210–4220.
- [49] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhuan Chen, Jiebo Luo, and Li Yuan. 2024. Identity-Preserving Text-to-Video Generation by Frequency Decomposition. *arXiv preprint arXiv:2411.17440* (2024).
- [50] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8652–8661.
- [51] Yunpeng Zhang, Qiang Wang, Fan Jiang, Yaqi Fan, Mu Xu, and Yonggang Qi. 2025. Fantasyid: Face knowledge enhanced id-preserving video generation. *arXiv preprint arXiv:2502.13995* (2025).
- [52] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3661–3670.
- [53] Jun Zheng, Jing Wang, Fuwei Zhao, Xujie Zhang, and Xiaodan Liang. 2024. Dynamic Try-On: Taming Video Virtual Try-on with Dynamic Attention Mechanism. *arXiv preprint arXiv:2412.09822* (2024).
- [54] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. 2022. CelebV-HQ: A large-scale video facial attributes dataset. In *European conference on computer vision*. Springer, 650–667.