Vivid-Animator:Expressive and Controllable Human Animation with Facial-Body Hybrid Guidance

Wei Duan[1,2,3], Haopan Ren[1,2,3], Deqi Li[4], Shi-Sheng Huang[1,2,3] and Hua Huang[1,2,3,*]
[1]School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
[2]Beijing Key Laboratory of Artificial Intelligence for Education, Beijing 100875, China
[3]Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing 100875, China
[4]Department of Computer Science and Technology, Tsinghua University, Beijing, China

***Abstract*—Realistic human animation remains a crucial but challenging task. Although current methods have made significant progress in controlling body poses, they still suffer from identity drift, motion jitter, and inadequate modeling of facial expressions. In this paper, we propose a novel Vivid-Animator framework that leverages dual temporal cues from facial expressions and body poses to achieve realistic and controllable human animation. Specifically, we first introduce a Dynamic Reference Strategy to generate expression-disentangled dynamic identity sequences, enhancing identity preservation during animation. Central to our approach, the effective combination of extracted facial expressions with diverse 3D control sequences jointly model expressive appearance and motion dynamics in the latent space. Furthermore, we employ a Spatio-Temporal Fusion Module to perform spatial fusion and temporal refinement of facial and pose features, reinforcing full-body motion coherence and expression stability. Extensive qualitative and quantitative experiments demonstrate that Vivid-Animator consistently outperforms previous methods in animation quality, producing expressive human videos while better preserving identity.**

***Index Terms*—human animation, expressive video generation, motion transfer**

I. Introduction

Human animation aims to synthesize a realistic video from a reference image by transferring the poses and facial expressions captured in a driving video. It has been widely applied in video production and VR/AR applications. With the advancement of latent diffusion models, current methods have been proposed to extract control signals from driving videos for character animation, leveraging representations such as skeletal keypoints [1]–[3], DensePose [4], and 3D models [5], [6]. However, these methods for full-body animation encounter issues such as identity drift, unnatural facial expressions, inaccurate facial control, and structural distortions, which greatly reduce the realism of animations.

Recently, various methods have been proposed to address identity drift. Certain methods [7] leverage custom subject inversion techniques, which constrain generalization across identities, whereas others utilize additional adapters [8] that remain ineffective in achieving precise facial control in full-body animation pipelines. In the latest research, StableAnimator [9] relies on ArcFace [10] embeddings from the reference image to improve identity consistency. However, these embeddings couple identity with head poses and expressions, leading to identity drift and rigid facial motion. For facial expression modeling, X-Dyna [11] incorporates Face-Vid2Vid [12], yet its lack of explicit facial control causes artifacts and instability. How to achieve realistic and controllable full-body human animation by jointly modeling body poses and facial expressions remains a significant challenge.

To address the above challenges, we propose Vivid-Animator, a framework for generating expressive and temporally coherent full-body character animations. Our core insight is to enhance the generative capability of diffusion models by jointly incorporating facial expressions and body poses, enabling realistic expressions and controllable animation. Specifically, we use the pre-trained head generation model [13] to extract sequential features from the reference image as temporal cues to guide animation generation. Furthermore, by modeling sequential facial expressions (pseudo portrait, FLAME [14] and landmarks) and body poses (MANO [15], SMPLX-CS [6] and skeleton) from sequential features, we design a spatio-temporal feature fusion mechanism that improves the model's understanding of motions and expressions in the driving sequence. This design helps alleviate control signal noise, such as detection errors and temporal jitter, thereby improving the robustness of motion conditioning during the diffusion process. For better feature transmission within the model, we design two parallel injection branches to effectively incorporate various control signals into the diffusion model. These effective designs provide more robust facial and pose feature references, allowing the model to focus more on the spatio-temporal consistency of facial expressions and pose motions. This guides the model to generate animations that accurately follow the motions and expressions of the driving video while preserving the identity of the reference image.

Extensive comparisons were conducted across multiple benchmarks on several datasets. The experimental results demonstrate that the proposed method produces fewer artifacts, clearer facial appearances, and better identity preservation compared to other approaches, allowing the generation of realistic and controllable animations. In conclusion, our contributions are as follows:

Supported by the National Key Research and Development Program of China (No. 2024YFB2808804).

*Corresponding author: Hua Huang (e-mail: huahuang@bnu.edu.cn).

- We jointly model facial expressions and body motions by incorporating multiple control sequences, enabling the generation of realistic facial appearances and controllable body poses.
- We propose a spatio-temporal fusion module that effectively combines spatial structures with temporal-frequency control signals to guide the optimization of the diffusion model.
- An end-to-end diffusion model named Vivid-Animator is proposed to generate full-body human animations with preserved identity.

II. PROPOSED METHOD

Given a single reference image I_r and a driving video V_d, our core idea is to extract a dynamic reference sequence $V_f d$ alongside diverse control conditions to effectively guide the diffusion model in generating vivid and temporally coherent full-body human animations V_f. As shown in Fig. 1, we first introduce a dynamic reference strategy to ensure consistent appearance modeling. Then a joint facial-body conditioning strategy is proposed to provide richer feature priors to improve generation quality. Subsequently, a spatio-temporal feature fusion module is introduced to guide the integration of spatial structures and temporal frequency features of facial expressions and body poses. Finally, we present the training details.

A. Dynamic Reference Strategy

Most pose transfer methods rely on a single reference image, using VAE Encoder and ReferenceNet to extract identity information and inject it into the denoising network. However, these methods primarily focus on torso generation while neglecting facial identity consistency, resulting in identity drift in facial regions. To mitigate this limitation, we utilize a high-efficiency pre-trained model [13] to provide dynamic reference sequences, which helps effective disentanglement expression and identity.

Specifically, we first detect and crop the head region in I_r and V_d via facial landmarks. Then, the cropped head from the driving video is used to animate the cropped reference portrait via [13], producing a dynamic head sequence. Using an affine transformation, this sequence is aligned with the reference image, resulting in a dynamic sequence that preserves identity consistency while conveying dynamic facial information. The generated dynamic reference sequence are then encoded by a VAE encoder and ReferenceNet, and injected into the denoising network through spatial attention mechanisms [1] at each layer. This process introduces rich and disentangled identity information into the diffusion model, significantly enhancing the quality of character animation. This process can be formally expressed as

$$\begin{aligned} V_{fd} &= G_f(I_r, V_d), \\ V_r &= \text{Stitching}(I_r, V_{fd}). \end{aligned} \tag{1}$$

Where I_r and V_d denote the reference image and the driving video, respectively. G_f represents a pretrained head generation model, and V_{fd} is the generated dynamic head sequence. V_r denotes the final dynamic reference sequence.

B. Hybrid Feature Conditions Mechanism

Most existing methods utilize skeletons, SMPL-X [16], or FLAME as motion feature conditions to drive digital human animation. However, parametric model-based methods are limited in their expressiveness, making it difficult to model complex body poses and vivid facial expressions, whereas body and facial keypoint-based approaches are prone to identity leakage issues. To address these limitations, we propose a hybrid facial-body conditioning scheme that delivers more comprehensive full-body control signals, placing special emphasis on detailed head region modeling. To model pose motion, inspired by RealisDance, we employ DWPose [17] to extract skeletons and use OSX [18] to obtain SMPL-X parameters. The SMPL-X parameters are then rendered as SMPLX-CS to incorporate 3D geometry, depth information, and continuous semantic cues. HaMeR [19] is used to predict MANO 3D hand parameters, with the left and right hands rendered using distinct materials.

Additionally, we create dedicated facial control sequences to represent expressions, head pose and position. We first stitch the generated head sequence onto driving frames and mask non-facial regions to form 'pseudo portraits.' Unlike facial landmarks, the pseudo portraits implicitly capture pose, location, and fine details (e.g., nose, ears, gaze), which help improve facial realism. For precise control, we extract facial landmarks (DWPose) and FLAME parameters (DECA) [14] to explicitly model head pose/position. These facial-body feature conditions construct 3D/2D motion and facial representations while enhancing facial and hand details. They are then processed through two convolutional encoders to produce embeddings compatible with the denoising U-Net's input dimensions, which are utilized in the model process.

C. Hybrid Feature Fusion Guidance

After extracting the hybrid feature conditions, we propose a robust hybrid feature fusion model in face and body. This model fully leverages the diversity of facial and body conditions while capturing both spatial and temporal feature correlations, thereby guiding the diffusion model to generate realistic human animations. As illustrated in Fig. 2, we design a Spatio-Temporal Feature Fusion module (STFF) to process motion and facial embeddings separately. This module is composed of a series of convolutional layers and temporal attention mechanisms.

Taking facial embeddings as an example, each embedding e_i^{face} corresponds to the i-th frame, where $i \in [0, N-1]$ and N is the total number of frames. Each e_i^{face} consists of three components: aligned pseudo-portrait embedding e_i^{portrait}, FLAME embedding e_i^{FLAME}, and facial landmark embedding $e_i^{\text{landmarks}}$. These embeddings are first passed through convolutional layers to predict spatial fusion weights $W_i^{\text{s-face}}$ for each pixel and channel. This spatial fusion strategy alleviates the

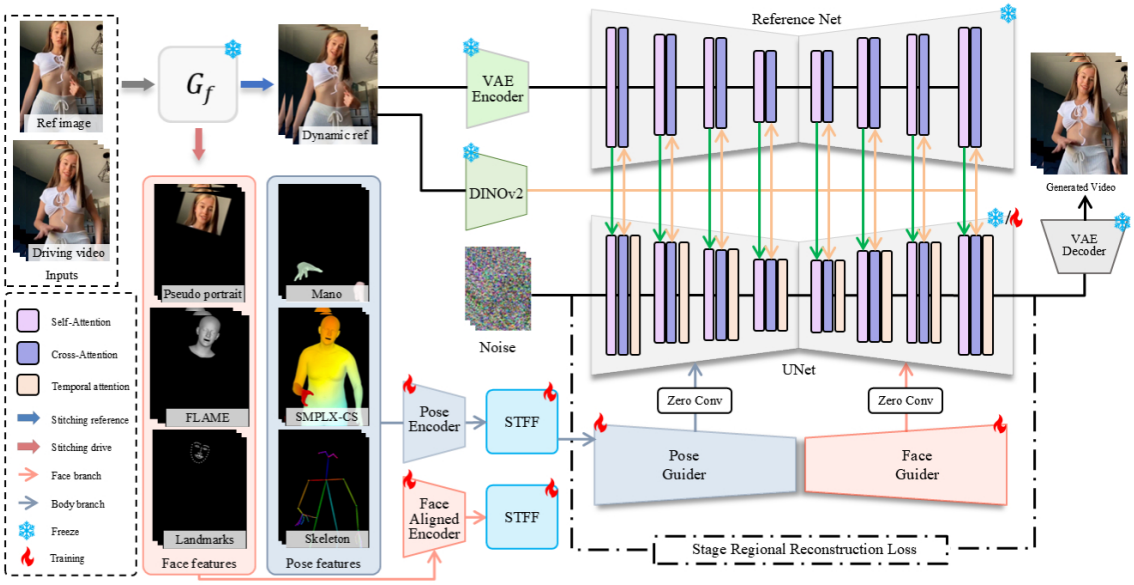

Fig. 1. The pipeline of Vivid Animator. Given a single reference image and a driving video, identity parameters (SMPLX and FLAME) are extracted from the reference image, while pose motion sequences (skeleton, landmarks, SMPLX-CS) and face sequence (MANO and FLAME) are derived from the driving video. A pretrained head model is introduced to create dynamic reference sequence and pseudo portrait labels. The dynamic reference sequence is encoded by the VAE encoder and DINOv2, while other features are encoded and fused via the STFF module. The unified facial-body guidance is then injected into the denoising U-Net through the Pose and Face Guiders. Finally, the VAE decoder generates vivid and coherent full-body animations.

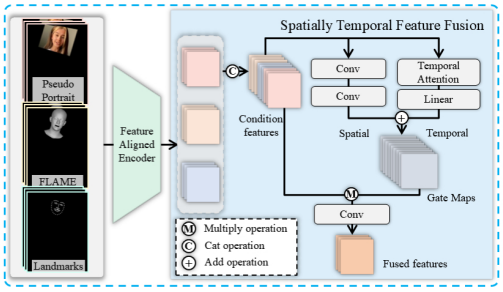

Fig. 2. Framework diagram of STFF in the Face branch.

impact of failure or noise in any single control signal. Furthermore, to mitigate temporal inconsistencies and occasional errors in inter-frame control condition (e.g., jitter or missing detections), we introduce a temporal attention mechanism to predict a weight offset $w_i^{\text{t-face}}$. The final fusion weight is computed by combining $w_i^{\text{s-face}}$ and $w_i^{\text{t-face}}$, producing facial features fused both temporally and spatially, denoted as $e_i^{\text{f-face}}$. The STFF module pipeline is formulated as:

$$\begin{aligned} w_i^{\text{s-face}} &= conv(e_i^{face}), \\ w_i^{\text{t-face}} &= Linear(Att(e_i^{face})), \\ e_i^{\text{f-face}} &= e_i^{face} * (w_i^{\text{s-face}} + w_i^{\text{t-face}}). \end{aligned} \tag{2}$$

D. Training

During training, a two-stage strategy is employed. In the first stage, training is only focusing on images, enabling the model to generate target images based on guidance and the reference image. In this stage, all components are trained except for the temporal attention layers, including the Encoder, STFF, and Guider module. In the second stage, videos are used for training. Only the temporal attention layers of the dual STFF, Guider Module and Denosing U-Net are trained, while the rest of the components remain frozen. Note that the VAE, DINOv2 [20], and ReferenceNet remain frozen throughout the entire training process.

The model is optimized by minimizing the mean squared error (MSE) between the predicted noise and actual noise. To further enhance facial generation, we introduce a Stage Regional Reconstruction loss during training. Specifically, we construct face, eyes and mouth masks based on facial landmarks. During the first 10K iterations of the first stage, the loss weight for the facial region is increased. In the later training phase, the loss weights for the eyes and mouth regions are further increased to encourage finer-grained supervision in these critical areas. It can be formulate as

$$\mathcal{L} = \begin{cases} \|\epsilon - \hat{\epsilon}_\theta\|_2^2 \cdot (1 + M_{\text{f}}), & \text{if t} \leq 10k \\ \|\epsilon - \hat{\epsilon}_\theta\|_2^2 \cdot (1 + M_{\text{f}} + M_{\text{e}} + M_{\text{m}}), & \text{otherwise} \end{cases}$$

Where the ϵ_θ represents the predicted noise and ϵ denotes the real noise. M_{f}, M_{e}, and M_{m} represent face, eyes, and mouth masks based on facial landmarks, respectively.

III. Experiments

A. Experiment Settings

Datasets. The experiments are conducted on three diverse datasets, such as TikTok [2], EMTD [21], and Champ [5]. TikTok contains full-body dance videos with simple dance

Fig. 3. Visual comparisons with SOTA methods.

movements and relatively clear facial features, making it a widely adopted dataset for evaluating motion transfer methods. Champ offers a large collection of dance videos, featuring more complex and dynamic limb movements. EMTD consists of high-quality upper-body speaking videos with clear facial details, diverse head poses, rich lip movements, and expressive facial dynamics. We selected a total of 300 videos from these datasets as our training data. To demonstrate the effectiveness of our method in improving facial quality, we construct a dataset for evaluation that includes facial details and diverse expressions. This set, which contains a total of 92 videos, is curated from EMTD, TikTok, and Champ. These videos are all excluded from the training data. Please refer to the appendix for detailed construction procedures.

Metrics. To evaluate the quality of generated images, we employ L1 loss, PSNR, SSIM, LPIPS, and FVD to measure the differences between generated and real videos. Moreover, CSIM [13] is used to assess identity preservation between two images. APD [22] and AED [22] represent the average pose distance and the average expression distance, respectively, while F-LMD [23] and M-LMD [23] are used to quantify the accuracy of face and lip. Furthermore, we introduce a new metric (DFSR) to measure the success probability of face detection in generated images.

Our model is trained on a single NVIDIA A800 GPU. It is trained for 25k steps on images and 50k steps on 8-frame video clips. Further implementation details are provided in the appendix.

B. Comparison with SOTA Methods

Quantitative Results. We evaluated all methods through self-driven on our curated evaluation dataset and computed quantitative metrics against ground-truth videos, as presented in Table. I. Our method outperforms other methods in conventional metrics and achieves significant improvements in CSIM and FVD, demonstrating its effectiveness in enhancing facial quality and spatio-temporal consistency of the characters. Since our method focuses on facial regions, the facial metrics are calculated. As demonstrated by the AED, F-LMD, and M-LMD metrics, our method achieves optimal facial expression control and provides the highest accuracy in facial reproduction. Additionally, we performed cross-driven on 30 video segments selected from the evaluation dataset, and the results are shown in Table. II. Our method achieves significant superiority in the CSIM and M-LMD metric, maintaining excellent identity consistency and accurate facial driving.

Qualitative Results. The qualitative comparison of different methods is shown in Fig.3. It can be observed that MimicMotion is prone to background artifacts, while Champ, MagicAnimate, and MusePose exhibit inaccuracies in face. RealisDance can accurately model body movements but suffers from unstable backgrounds and an inability to drive facial expressions. StableAnimator can perform simple expression driving but tends to over-beautify the characters. Although X-Dyna can drive facial expressions, the lack of 3D information results in inaccurate facial positioning and expression generation. Our method not only achieves accurate body

TABLE I
THE QUANTITATIVE COMPARISONS WITH EXISTING METHODS.

Method	*L1(E-4)* ↓	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓	*CSIM* ↑	*FVD* ↓	*APD* ↓	*AED* ↓	*F-LMD* ↓	*M-LMD* ↓	*DFSR* ↑
Champ	3.91	28.69	0.556	0.427	0.340	1085.28	0.0563	0.5362	12.848	4.595	0.563
Magic-Animate	2.65	28.47	0.656	0.432	0.290	874.14	0.0401	0.4530	5.843	2.604	0.858
MimicMotion	4.10	28.31	0.484	0.432	0.216	11952.82	0.0485	0.5155	7.070	2.757	0.771
MusePose	2.04	**29.60**	0.696	0.271	0.436	485.38	0.0346	0.4125	5.748	2.737	0.841
X-Dyna	2.70	29.04	0.685	0.310	0.362	640.75	0.0367	0.4176	7.451	2.810	0.789
StableAnimator	2.21	29.47	0.660	0.323	0.205	980.45	0.0358	0.4922	4.928	2.306	0.900
RealisDance	1.91	29.27	0.750	0.234	0.548	424.92	0.0247	0.3179	3.847	2.158	0.914
ours	**1.80**	**29.60**	**0.756**	**0.232**	**0.580**	**370.00**	**0.0239**	**0.3155**	**3.822**	**1.682**	**0.922**

TABLE II
QUANTITATIVE COMPARISONS OF FACIAL METRICS FOR CROSS-DRIVEN WITH EXISTING METHODS

Method	Metrics				
	CSIM ↑	*APD* ↓	*AED* ↓	*F-LMD* ↓	*M-LMD* ↓
Champ	0.343	0.0630	0.6033	15.948	6.291
Magic-Animate	0.243	0.0426	0.5839	6.263	3.097
MimicMotion	0.134	0.0550	0.5933	9.617	3.628
MusePose	0.435	0.0426	0.5819	6.226	3.149
X-Dyna	0.257	0.0402	**0.5614**	7.688	3.149
StableAnimator	0.111	0.0396	0.6004	**4.187**	2.250
RealisDance	0.514	0.0379	0.5774	4.958	2.697
ours	**0.558**	**0.0371**	0.5819	4.781	**2.173**

driving and vivid expression generation, but also maintains video stability. The cross-driven results are also shown in Fig.4. Other methods exhibit some degree of identity leakage. By introducing the dynamic reference strategy and pseudo portraits, our method incorporates richer identity information and implicit expression cues.

C. Ablation Study

According to the scope of influence of the proposed module, the ablation study is divided into four different configuration variants. (1) "only *LM*(landmarks)" indicates the absence of the facial branch, where only the body branch is augmented with facial keypoints. (2) "*w/o* mask" denotes training without the facial mask region. (3) "*w/o* offset" means that temporal fusion is not applied. (4) "*w/o* replace" refers to the lack of dynamic dynamic reference sequence.

Fig. 4. Visual comparisons with SOTA methods in cross-ID setting.

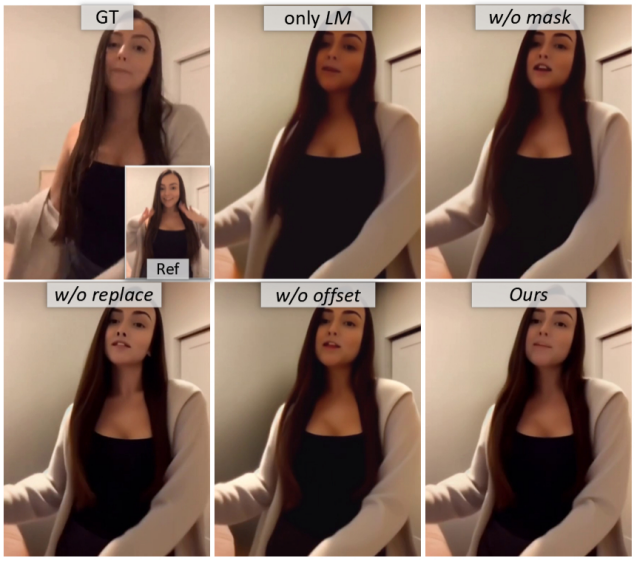

Fig. 5. Visual results of ablation studies.

Ablation Study on the Full-Body Region. We conducted ablation studies on the core components, as shown in Fig. 5 and Table. III. It is evident that relying solely on facial keypoints for guidance leads to stiff and unnatural facial expressions. And the absence of the mask during training results in an overall decrease in facial quality, with all evaluated metrics showing lower performance. Although the model without the temporal fusion module achieves relatively better

TABLE III
FULL-BODY METRICS RESULTS OF THE ABLATION STUDY.

Method	Metrics					
	L1(E-4) ↓	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓	*CSIM* ↑	*FVD* ↓
only LM	2.27	28.71	0.733	0.261	0.522	585.95
w/o FM	2.26	28.74	0.733	0.260	0.512	575.41
w/o offset	1.80	29.55	**0.757**	0.229	0.560	388.77
w/o replace	**1.79**	29.59	0.756	**0.228**	0.555	378.00
ours	1.80	**29.60**	0.756	0.232	**0.580**	**370.00**

TABLE IV
FACIAL METRICS RESULTS OF THE ABLATION STUDY.

Method	Metrics				
	APD ↓	***AED*** ↓	***F-LMD*** ↓	***M-LMD*** ↓	***DFSR*** ↑
only LM	0.0282	0.3923	4.469	1.876	0.874
w/o FM	0.0282	0.3864	4.438	2.010	0.876
w/o offset	0.0250	0.3471	3.992	1.745	0.886
w/o replace	0.0257	0.3302	3.483	1.856	0.900
ours	**0.0239**	**0.3155**	**3.822**	**1.682**	**0.922**

CSIM scores, its outputs tend to exhibit artifacts, leading to a decline in FVD performance. For the "*w/o* replace" variant, the static reference, while enforcing an artificial temporal stability that merely maintains FVD, is the primary cause for the significant drop in CSIM and weaker identity consistency.

Ablation Study on the Face Region. We also computed facial metrics, which reflect expression fidelity and lip-sync accuracy, as shown in Table. IV. When only facial landmarks are used for control, the model shows the worst AED score, reflecting the visibly rigid expressions in the generated results. Removing the facial mask leads to a diminished focus on the facial region, resulting in a significant decline in both the realism and accuracy of facial expression generation. Excluding the temporal fusion module causes a slight drop in all metrics, though the primary impact is observed in video stability. When the dynamic reference sequence is removed, the model tends to treat features such as exposed teeth in the static reference image as inherent identity cues, leading to frequent teeth exposure in the generated results. Consequently, the F-LMD and M-LMD metrics show a significant decrease.

IV. CONCLUSION

In this paper, we propose Vivid-Animator, a novel framework that addresses identity drift, motion jitter, and suboptimal facial expressions by integrating diverse facial and body motion features to guide diffusion model optimization. Additionally, we also propose a spatio-temporal feature fusion mechanism to capture complex correlations across space and time. These mechanisms enhance the model's ability, leading to more coherent, high-fidelity, and controllable human animations. Extensive experiments demonstrate state-of-the-art performance with notable improvements in facial expression accuracy and overall quality. Please refer to the supplementary materials for more details and results.

REFERENCES

[1] L. Hu, "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*, 2024, pp. 8153–8163.

[2] T. Wang, L. Li, K. Lin, Y. Zhai, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, "Disco: Disentangled control for realistic human dance generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9326–9336.

[3] Y. Zhang, J. Gu, L.-W. Wang, H. Wang, J. Cheng, Y. Zhu, and F. Zou, "Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance," *arXiv preprint arXiv:2406.19680*, 2024.

[4] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1481–1490.

[5] S. Zhu, J. L. Chen, Z. Dai, Z. Dong, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu, "Champ: Controllable and consistent human image animation with 3d parametric guidance," in *European Conference on Computer Vision*. Springer, 2024, pp. 145–162.

[6] J. Zhou, B. Wang, W. Chen, J. Bai, D. Li, A. Zhang, H. Xu, M. Yang, and F. Wang, "Realisdance: Equip controllable character animation with realistic hands," *arXiv preprint arXiv:2409.06202*, 2024.

[7] Q. Wang, X. Jia, X. Li, T. Li, L. Ma, Y. Zhuge, and H. Lu, "Stableidentity: Inserting anybody into anywhere at first sight," *arXiv preprint arXiv:2401.15975*, 2024.

[8] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[9] S. Tu, Z. Xing, X. Han, Z.-Q. Cheng, Q. Dai, C. Luo, and Z. Wu, "Stableanimator: High-quality identity-preserving human image animation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 096–21 106.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[11] D. Chang, H. Xu, Y. Xie, Y. Gao, Z. Kuang, S. Cai, C. Zhang, G. Song, C. Wang, Y. Shi *et al.*, "X-dyna: Expressive dynamic human image animation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5499–5509.

[12] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 039–10 049.

[13] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang, "Liveportrait: Efficient portrait animation with stitching and retargeting control," *arXiv preprint arXiv:2407.03168*, 2024.

[14] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[15] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.

[16] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

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

[18] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, "One-stage 3d whole-body mesh recovery with component aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 159–21 168.

[19] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9826–9836.

[20] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[21] R. Meng, X. Zhang, Y. Li, and C. Ma, "Echomimicv2: Towards striking, simplified, and semi-body human animation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5489–5498.

[22] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in neural information processing systems*, vol. 32, 2019.

[23] S. Tan, B. Ji, M. Bi, and Y. Pan, "Edtalk: Efficient disentanglement for emotional talking head synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 398–416.