IDC-Animator: Efficient Identity-Consistent Human Image Animation

1st Mengting Xie
College of Computer Science and Software Engineering
Hohai University
Nanjing, China
mengting@hhu.edu.cn

2nd Yingchi Mao*
College of Computer Science and Software Engineering
Hohai University
Nanjing, China
yingchimao@hhu.edu.cn

***Abstract*—Diffusion models have demonstrated superior performance in the field of human image animation. However, the complexity and diversity of human motion present formidable challenges to effective modeling. Particularly in scenarios involving rapid, large-amplitude movements, existing methods often struggle to maintain structural stability, leading to facial feature collapse or limb motion incoordination. Furthermore, the process of long-form video generation incurs substantial memory overhead. This not only constrains the achievable video duration but also frequently degrades generation quality due to computational resource limitations. To address the aforementioned challenges, we propose IDC-Animator, an efficient video generation network designed to ensure high identity consistency throughout the generation process. First, we introduce the Identity-Adaptive Facial Encoder (IAFE), which significantly enhances feature adaptability through deep parsing of the reference facial region. Second, we construct the Identity-Consistency Calibrator (ICC) to dynamically regulate identity injection via multimodal interaction and alignment. This mechanism ensures high-fidelity synchronization of identity features while maintaining structural stability throughout animation generation. Finally, we incorporate the Mamba architecture as the core for temporal modeling, successfully reducing the computational complexity of long-sequence processing to a linear level, thereby substantially improving the efficiency of long-video generation. Experimental results demonstrate that our proposed method achieves state-of-the-art performance on the TikTok and UBC Fashion public datasets.**

***Index Terms*—human image animation, identity-adaptive facial encoder, identity-consistency calibrator, temporal modeling.**

I. Introduction

With the rapid development of video generation technologies [1], especially the iterative evolution of generative models [2], human image animation [3] has made significant progress and shown broad application potential in areas such as game development, digital human production, and motion analysis. However, despite these advances, practical applications still suffer from instability, where smooth motion can abruptly degrade into noticeable limb distortions.

Currently, many deep learning-based video generation methods [4]–[10] have achieved remarkable performance. These methods can be broadly categorized into two types. The first is GAN-based approaches [11], which generate frames by deforming reference appearances with pose guidance, but suffer from unstable training and limited generalization. The second is diffusion-based methods [12]–[14],which provide better stability and transferability. Representative works such as Disco [15], Animate Anyone [5], and MagicAnimate [16] leverage progressive denoising and strong feature learning to produce detailed and temporally smooth animations. Nevertheless, as illustrated in Fig.1, when capturing large-amplitude and high-dynamic motion sequences, diffusion model-based methods [5], [12]–[17] tend to suffer from severe distortion and loss of identity information, particularly in facial regions.

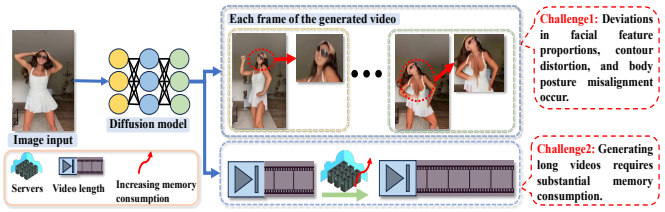

Fig. 1. Two challenges encountered in handling video generation task.

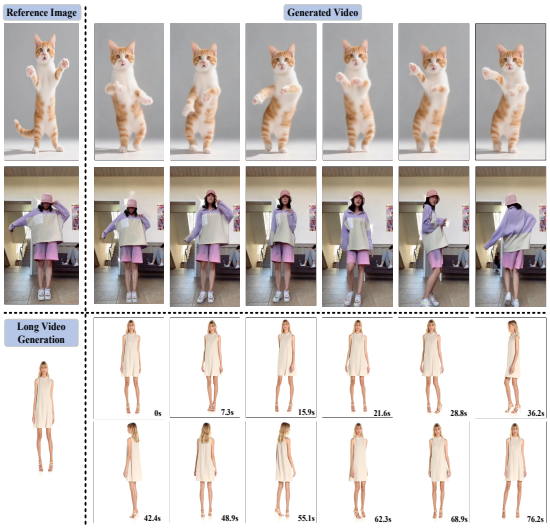

Fig. 2. Video examples synthesized by IDC-Animator. Notably, our method demonstrates exceptional generalization robustness, even without being trained on cartoon character dance data. Furthermore, IDC-Animator is capable of generating high-fidelity videos with a duration of 76.2 seconds, validating the stability of its long-sequence modeling.

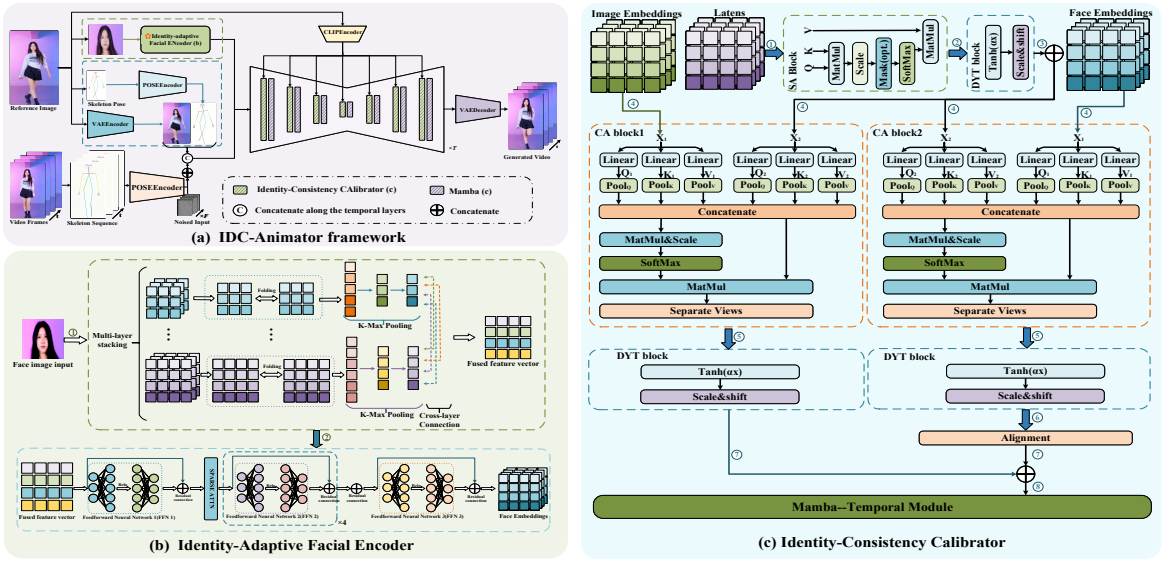

Fig. 3. Overview of the IDC-Animator. (a) Pipeline: The framework synthesizes videos from a reference image and skeleton sequence using a diffusion backbone. (b) Identity-Adaptive Facial Encoder (IAFE): Extracts robust facial embeddings through multi-layer stacking and sparse attention to enhance feature adaptability. (c) Identity-Consistency Calibrator (ICC): Dynamically aligns identity features with diffusion latents via cross-attention and incorporates Mamba for efficient linear-complexity temporal modeling.

Additionally, the quadratic computational and memory complexity of temporal Transformers becomes a major bottleneck. As the number of frames increases, resource consumption grows rapidly, severely limiting video length. Consequently, existing methods struggle to scale efficiently, making long-sequence animation generation impractical under hardware constraints.

To address these challenges, we propose IDC-Animator, an efficient framework for identity-consistent human image animation that overcomes the trade-off between computational efficiency and identity preservation. It adopts a decoupled strategy to separately model identity and motion, enabling independent handling of facial details and global poses. Additionally, we integrate the Mamba [18] architecture for temporal modeling, reducing computational complexity to linear scale while maintaining stable identity consistency in long-sequence generation.

As illustrated in Fig.3, our pipeline begins with the multi-dimensional extraction of facial, motion, and global features. Specifically, we design the Identity-Adaptive Facial Encoder (IAFE) to precisely extract and optimize facial embeddings, ensuring robust identity discriminability. Subsequently, all features are injected into the backbone network for fusion. This backbone innovatively integrates our proposed Identity-Consistency Calibrator (ICC) with the Mamba [18] architecture. The ICC ensures visual realism by dynamically calibrating identity features, whereas the Mamba module significantly boosts the processing efficiency of long-sequence data by restructuring the temporal processing logic.

In general, the contributions of this work are summarized as follows:

- To mitigate character distortion in human image animation, we design an identity-adaptive facial encoder (IAFE) and an identity-consistent calibrator (ICC) to enhance facial features and maintain stable, consistent identity representation.
- To address the issue of high computational complexity in traditional temporal models, we introduce the Mamba model [18] as the core component for temporal modeling to improve the processing efficiency of long-sequence data.
- We design an IDC-Animator framework, and its effectiveness and robustness have been validated by extensive experiments on two public datasets: TikTok and UBC. Experimental results show that while reducing the memory consumption of long sequences by 67.3%, our framework can also generate videos with clearer facial details and stronger motion coherence.

Benefiting from these designs, IDC-Animator generates identity-consistent and smooth long-duration videos (as shown in Fig.2). Extensive qualitative analyses on out-of-domain data, including cartoon characters, demonstrate its strong generalization robustness. Furthermore, successfully generating a continuous 76.2-second video validates its exceptional capability in long-sequence tasks.

II. METHODOLOGY

In this section, we present IDC-Animator, a modular framework designed for high identity consistency (Fig.3a). **Feature Extraction:** Given a reference image and a preprocessed driving skeleton sequence, our IAFE extracts deep facial identity features. Concurrently, latent visual features from the VAEEncoder are concatenated with reference pose features. For motion guidance, the preprocessed skeleton sequence is

encoded and concatenated with the noised input temporally. **Generation & Denoising:** These multi-source features enter the UNet backbone, flowing through the ICC for precise identity injection and the Mamba module for efficient, linear-complexity temporal modeling, supplemented by CLIP global semantic guidance. **Output:** After T denoising iterations, the VAEDecoder reconstructs the final high-fidelity, identity-consistent video sequence.

A. Identity-Adaptive Facial Encoder

As illustrated in Fig.3(b), the IAFE optimizes facial embeddings through a two-stage strategy: enhancing identity discriminability via an angular margin-based loss, followed by deep feature refinement using a feed-forward network and sparse attention. To achieve high discriminability, we normalize inputs x_i and weights W_j to compute their cosine similarity $W_j^T x_i = \cos\theta_j$. To enforce stricter angular decision boundaries, we apply a scale s and an additive angular margin m to the ground-truth class y_i, yielding the transformed logit $s \cdot \cos(\theta_{y_i} + m)$. The final identity loss $\mathcal{L}_{id}$ is formulated as:

$$\mathcal{L}_{id} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{\text{logit}_{y_i}}}{e^{\text{logit}_{y_i}} + \sum_{j\neq y_i}^{n} e^{\text{logit}_j}}\right) \tag{1}$$

where definitions of the logits are:

$$\text{logit}_k = \begin{cases} s\cdot\cos(\theta y_i + m) & \text{if } k = y_i \\ s\cdot\cos(\theta_k) & \text{if } k \neq y_i \end{cases} \tag{2}$$

Where N and n denote the batch size and the number of identity categories. This loss minimizes intra-class and maximizes inter-class variance, enforcing robust identity discrimination on the hypersphere to prevent identity drift.

Subsequently, to fuse multi-modal facial features, the embeddings $F \in \mathbb{R}^{d\times h\times w}$ are projected through a Feed-Forward Network (FFN) for non-linear dimensionality transformation:

$$F_{\text{ffn}} = \text{ReLU}(FW_1 + b_1)W_2 + b_2 \tag{3}$$

Where W_1, W_2 and b_1, b_2 are the projection weights and biases with ReLU activation. This non-linear mapping enhances the representation of complex facial details. Next, to efficiently capture local spatial correlations without the prohibitive cost of global attention, we apply a Sparse Attention mechanism within a defined local window $\mathcal{S}_i$:

$$\text{SparseAttn}(F_{\text{ffn}}, I) = \sum_{i=1}^{n} \text{Softmax}\left(\frac{Q_i K_{\mathcal{S}_i}^T}{\sqrt{d_k}}\right) V_{\mathcal{S}_i} \tag{4}$$

Where Q_i, $K_{\mathcal{S}_i}$, and $V_{\mathcal{S}_i}$ are the standard query, key, and value vectors within window $\mathcal{S}_i$. This scaled mechanism reduces complexity while capturing high-frequency local details. Residual connections are subsequently applied to preserve original features and facilitate deep cross-modal learning, yielding the $(l+1)$-th layer output:

$$F_{\text{out}}^{(l+1)} = F_{\text{ffn}}^{(l)} + \text{SparseAttn}(F_{\text{ffn}}^{(l)}, I) \tag{5}$$

Where l is the layer index. Residual connections mitigate vanishing gradients and preserve identity features. After stacking 4 layers, a final FFN refines the embedding:

$$F_{\text{final}} = \text{FFN}(F_{\text{out}}^{(4)}) \tag{6}$$

This step maps the deeply refined semantic features into the target feature space, generating the final high-fidelity facial embedding F_{final} for use by the video generation module.

B. Identity-Consistent Calibrator

To effectively mitigate character distortion and identity inconsistency in human image animation, we designed the Identity-Consistent Calibrator (Fig.3(c)). This module achieves high-fidelity identity preservation and spatial alignment through the following four progressive steps.

Latent Variable Attention Enhancement. The diffusion latent variable $l_i \in \mathbb{R}^{C\times H\times W}$ contains crucial spatio-temporal structural information. To enhance the perception of local feature relevance within the latent space, we first apply a Multi-Head Spatial Self-Attention mechanism:

$$\begin{aligned} l_i = \text{MultiHeadSelfAtten}(l_i) = \text{Concat}\,(&\text{SelfAtten}_1(l_i), \\ &\text{SelfAtten}_2(l_i), \ldots, \text{SelfAtten}_K(l_i))\cdot \boldsymbol{W}_O. \end{aligned} \tag{7}$$

where l_i denotes the feature map at the current diffusion step, K is the number of attention heads, and W_O is the output projection matrix. This step models long-range spatial dependencies within the latent variables, refining the model's contextual understanding of pose and layout to prepare a structural foundation for subsequent external signal injection.

Cross-Modal Semantic Association Construction. To incorporate identity and appearance information, we introduce image feature embeddings E_{img} and facial feature embeddings E_{face}. After unifying feature dimensions via projection matrices W_{proj}, we construct a cross-attention mechanism to establish semantic mapping:

$$l_i^{\text{img}} = \text{CrossAttn}\left(l_i \cdot W_{\text{proj}}, E_{\text{img}} \cdot W_{\text{proj}}^{\text{img}}\right) \tag{8}$$

$$l_i^{\text{face}} = \text{CrossAttn}\left(l_i \cdot W_{\text{proj}}, E_{\text{face}} \cdot W_{\text{proj}}^{\text{face}}\right) \tag{9}$$

where $W_{\text{proj}}^{\text{img}}$ and $W_{\text{proj}}^{\text{face}}$ are dimension alignment matrices. These equations facilitate semantic injection. Eq. (8) aligns global appearance styles from the reference image to the latents, while Eq. (9) precisely maps high-fidelity facial identity features to the corresponding spatial regions, achieving initial fusion of identity and scene information.

Dynamic Feature Distribution Adaptation. To address distribution discrepancies across different feature modalities, we propose a Dynamic Tanh (DYT) module. Utilizing learnable parameters α, γ, β and a temperature coefficient τ, we construct an adaptive non-linear activation function:

$$\text{DYT}(x) = \alpha \odot \tanh(\gamma \odot (x - \beta)/\tau) \tag{10}$$

where $\odot$ denotes channel-wise element-wise multiplication. α controls scaling, β manages the offset, and γ, τ adjust the

TABLE I
QUANTITATIVE COMPARISONS WITH EXISTING METHODS ON THE TIKTOK DATASET.

Model	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**FVD↓**
MRAA [4]	–	0.672	0.296	284.82
DisCo [15]	29.03	0.668	0.292	292.80
MagicAnimate [16]	29.166	0.714	0.239	171.90
AnimateAnyone [5]	29.56	0.718	0.285	171.90
Champ [6]	29.91	0.802	0.231	160.82
Unianimate [7]	30.77	0.811	0.231	148.06
MimicMotion [8]	–	0.601	0.416	326.57
ControlNeXt [9]	–	0.615	0.416	326.57
StableAnimator [10]	30.81	0.801	0.232	140.62
FOMM [19]	29.01	0.648	0.335	405.22
TPSMM [20]	29.18	0.673	0.299	306.17
DreamPose [21]	28.11	0.511	0.442	551.02
SD-I2V [22]	29.11	0.670	0.295	225.50
Ours	**30.79**	**0.806**	**0.231**	**144.34**

TABLE II
QUANTITATIVE COMPARISONS WITH EXISTING METHODS ON THE UBC FASHION DATASET.

Model	PSNR↑	SSIM↑	LPIPS↓	FVD↓
MRAA [4]	-	0.749	0.212	253.6
TPSMM [20]	-	0.746	0.213	247.5
BDM [23]	24.07	0.918	0.048	148.3
DreamPose [21]	-	0.885	0.068	238.7
DreamPose* [24]	34.75	0.879	0.111	279.6
SD-I2V [22]	36.01	0.894	0.095	175.4
AnimateAnyone [5]	38.49	0.931	0.044	81.6
Unianimate [7]	37.92	0.940	0.031	68.1
DPTN [25]	-	0.907	0.060	215.1
NTED [26]	-	0.890	0.073	278.9
PIDM [27]	-	0.713	0.288	1197.4
Ours	**38.6**	**0.945**	**0.028**	**62.5**

TABLE III
ABLATION STUDY ON THE TIKTOK DATASET. HERE, A DENOTES THE IAFE, AND B REPRESENTS THE ICC.

Model	PSNR↑	SSIM↑	LPIPS↓	FVD↓
baseline	29.56	0.718	0.285	171.90
baseline+A	29.93	0.768	0.247	160.86
baseline+A+B	30.79	0.806	0.231	144.34

slope and sensitivity. Unlike the standard Tanh, DYT allows the model to dynamically scale the feature distribution. This mechanism effectively corrects distribution shifts between image and facial features, preventing optimization issues caused by mismatched value ranges.

Cross-Modal Distribution Alignment. This pivotal step unifies feature distributions via statistical modeling. We first compute the weighted mean μ_{face} for facial features and weighted statistics $(\mu_{\text{img}}, \sigma_{\text{img}})$ for image features using spatial weights $w_{\text{face/img}}$ generated by Sigmoid-activated 1×1 convolutions. Subsequently, we introduce a channel-wise alignment weight vector $\boldsymbol{\lambda} \in \mathbb{R}^C$ to perform a normalization and rescaling transformation. Specifically, $\boldsymbol{\lambda}$ is implemented as a learnable parameter optimized end-to-end, enabling the model to adaptively determine the alignment intensity for each channel based on the generation objective. The transformation is formulated as:

$$\tilde{l}_i^{\text{face\_aligned}}(c,h,w) = \boldsymbol{\lambda}(c) \cdot \frac{\tilde{l}_i^{\text{face}}(c,h,w) - \mu_{\text{face}}(c)}{\sigma_{\text{face}}(c)} \cdot \sigma_{\text{img}}(c) + (1 - \boldsymbol{\lambda}(c)) \cdot \mu_{\text{img}}(c) \tag{11}$$

where $\tilde{l}_i^{\text{face}}$ represents the facial features to be aligned. The formula normalizes these features to a standard distribution (using intrinsic statistics) and then maps them to the image feature distribution space (scaling with σ_{img} and shifting with μ_{img}). $\boldsymbol{\lambda}(c)$ acts as a dynamic balancing factor. This step eliminates the "statistical gap" between facial embeddings and the background image. It ensures that the generated face blends perfectly with the overall video in terms of lighting, texture, and tone, effectively eliminating spatial distortion arising from cross-modal distribution differences.

C. Temporal Feature Fusion Modeling—Mamba

To effectively address the challenge of modeling complex temporal dynamics in human image animation, we integrate the Mamba architecture into the U-Net backbone as a novel temporal modeling paradigm. Unlike traditional attention mechanisms that suffer from quadratic computational complexity, Mamba leverages a selective State Space Model (SSM) architecture. This allows it to capture long-range temporal dependencies efficiently with linear complexity, significantly reducing memory overhead. Specifically, the spatially aligned features derived from the ICC are first augmented with temporal positional information and then processed by the Mamba module [18]:

$$l_i^{\text{out}} = \text{Mamba}\left(\tilde{l}_i + \text{PosEmb}(t)\right) \tag{12}$$

Here, $\tilde{l}_i$ represents the fused feature map from the preceding ICC module. While it encapsulates high-fidelity facial identity and scene structure, it initially lacks temporal context. To address this, $\text{PosEmb}(t) \in \mathbb{R}^C$ (sinusoidal positional encoding) is injected to provide explicit temporal order, enabling the model to distinguish between sequential frames. Consequently, l_i^{out} serves as the final output, integrating long-range temporal dependencies. By processing position-encoded inputs, Mamba ensures temporal coherence and smooth transitions. Crucially, beyond efficiency, Mamba introduces a favorable inductive bias for video generation. Unlike Transformers that treat frames as discrete tokens, Mamba's linear scanning models motion as a continuous evolution of latent states. This alignment with the physical continuity of human movement enables Mamba to capture pose progression more effectively than standard attention, particularly in high-dynamic or long-duration sequences.

III. EXPERIMENTS

A. Experimental setups

Datasets:

We evaluate our model on TikTok (340/100 videos) and UBC Fashion (500/100 videos). Following MagicAnimate

TABLE IV
QUANTITATIVE COMPARISON OF DIFFERENT TEMPORAL MODELING APPROACHES ON THE TIKTOK DATASET.

Method	PSNR↑	SSIM↑	LPIPS↓	FVD↓	Mem↓
Transformer	30.50	0.812	0.240	138.00	8.20
Mamba (ours)	30.79	0.806	0.231	144.34	6.83

[16]'s protocol, we use the same standardized 10-video subset from TikTok to ensure fair and reproducible quantitative comparison.

Implementation Details: We extracted skeleton poses using DWPose. The U-Net backbone and IAFE's spatial cross-attention modules were initialized with Stable Video Diffusion (SVD) weights to preserve identity , while PoseNet and Face Encoder were trained from scratch.

The model was trained distributedly on 4 NVIDIA A100 (80GB) GPUs for 20 epochs. The batch size was set to 1 per GPU, with a learning rate of 1×10^{-5}. This configuration ensures consistency with existing baselines while leveraging pre-trained priors to optimize training efficiency and generalization.

Evaluation Metrics: To evaluate performance, we employ PSNR and SSIM for structural accuracy—where high scores in facial regions, driven by the identity constraints in Eq.(1), reflect precise landmark and texture restoration—LPIPS for perceptual quality, FVD for temporal consistency, and Memory (Mem) for efficiency.

B. Comparisons

1) Quantitative comparisons: We compare our method with state-of-the-art baselines (Disco [15], MagicAnimate [16], Animate Anyone [5], Champ [6]). As shown in Table I, our approach outperforms all competitors across all TikTok dataset metrics, achieving an FVD of 144.34. This demonstrates its superior capability in capturing data distributions to generate highly realistic videos.

Results on the UBC Fashion dataset (Table II) further validate our method's generalization ability. Notably, it achieves a superior SSIM of 0.945, demonstrating exceptional structural preservation and the capacity to synthesize high-fidelity fashion animations.

Furthermore, quantitative analysis (Fig.6a) demonstrates our method reduces memory consumption by 67.3% compared to MagicAnimate. The marginal 1.4 GB overhead against the lightweight MRAA is an acceptable trade-off for substantially higher identity consistency and generation quality. Overall, the Mamba module [18] efficiently lowers the memory barrier, facilitating deployment in resource-constrained scenarios.

2) Qualitative comparisons: Complementing the quantitative metrics, Fig. 4 and Fig. 5 provide qualitative comparisons. While state-of-the-art methods like MagicAnimate [16] and Champ [6] struggle with incomplete limbs and misalignment—and Animate Anyone [5] exhibits artifacts—our model maintains invariant facial geometry and textures even under extreme poses. As shown in Fig.4, this qualitatively validates the ICC module's effectiveness in identity preservation without requiring additional subjective scoring.

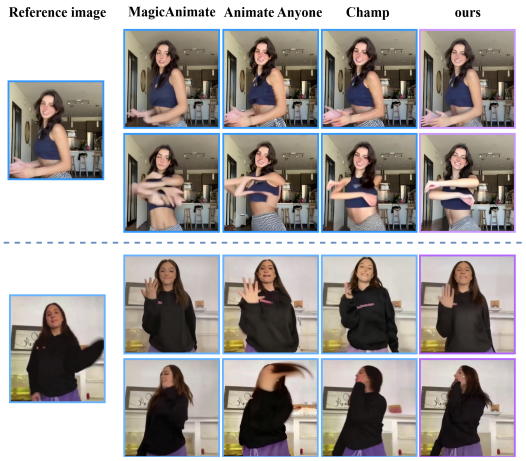

Fig. 4. Qualitative comparisons with existing methods on the TikTok dataset.

In contrast, our method consistently generates high-quality, temporally coherent pose transfers with natural transitions and strict condition adherence. We attribute this to our unified video diffusion framework. By jointly processing reference images and noisy videos, it constructs a shared feature space that resolves the disjoint image-video feature problem of traditional methods. This end-to-end joint training facilitates efficient cross-modal interaction, enhancing the capture of complex motions and details, and significantly surpassing existing solutions in visual quality.

C. Ablation Studies

Ablation studies (Table III) validate our core components: removing IAFE or ICC significantly degrades facial semantic consistency. These results underscore their critical roles in anchoring inter-frame identity stability and ensuring high video fidelity.

Finally, we investigated the choice of the temporal module by conducting comparative experiments between the Temporal Transformer and Mamba. As shown in Table IV, both architectures yield comparable comprehensive performance: they efficiently perform temporal modeling to ensure superior visual quality and temporal coherence, with negligible differences in key metrics such as PSNR and SSIM.

However, the Mamba module demonstrates a decisive memory efficiency advantage. As shown in Fig.6(b), when the sequence length exceeds 256 frames, the Transformer encounters Out-Of-Memory (OOM) errors, whereas Mamba stably generates long videos with significantly lower memory usage. This low memory footprint substantially lowers hardware barriers, enabling broader deployment in resource-constrained scenarios.

IV. CONCLUSION

In this work, we present IDC-Animator, integrating IAFE, ICC, and Mamba for efficient, identity-consistent video gen-

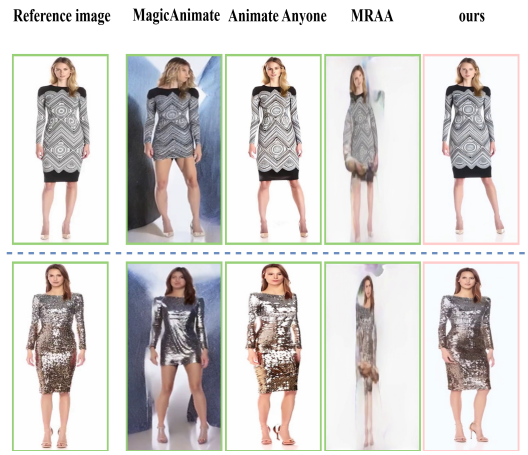

Fig. 5. Qualitative comparisons with existing methods on the UBC Fashion dataset.

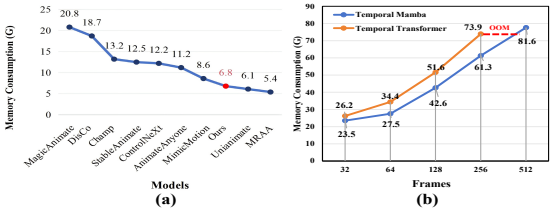

Fig. 6. Analysis of memory consumption. (a) Comparison of inference memory overhead with state-of-the-art methods. The reported values refer to GPU memory consumption when processing 16 frames at a resolution of 576×1024. (b) Memory usage comparison between Mamba and Transformer architectures across varying frame lengths. "OOM" denotes Out Of Memory.

eration. IAFE and ICC mitigate spatial distortions via cross-modal alignment , while Mamba reduces temporal modeling to linear complexity. Experiments on TikTok and UBC Fashion show our framework achieves state-of-the-art performance (FVD 144.34, SSIM 0.945) and reduces GPU memory by over 67.3% without compromising quality.

REFERENCES

[1] Y. Wei, S. Zhang, Z. Qing *et al.*, "Dreamvideo: Composing your dream videos with customized subject and motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6537–6549.

[2] F. Liao, X. Zou, and W. Wong, "Appearance and pose-guided human generation: A survey," *ACM Computing Surveys*, vol. 56, no. 5, pp. 1–35, 2024.

[3] Y. Ma, Y. He, X. Cun *et al.*, "Follow your pose: Pose-guided text-to-video generation using pose-free videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4117–4125.

[4] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 13 653–13 662.

[5] 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 (CVPR)*, 2024, pp. 8153–8163.

[6] S. Zhu, J. L. Chen, Z. Dai *et al.*, "Champ: Controllable and consistent human image animation with 3d parametric guidance," in *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024, pp. 145–162.

[7] X. Wang, S. Zhang, C. Gao *et al.*, "Unianimate: Taming unified video diffusion models for consistent human image animation," *arXiv preprint arXiv:2406.01188*, 2024.

[8] Y. Zhang, J. Gu, L. W. Wang *et al.*, "Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance," *arXiv preprint arXiv:2406.19680*, 2024.

[9] B. Peng, J. Wang, Y. Zhang *et al.*, "Controlnext: Powerful and efficient control for image and video generation," *arXiv preprint arXiv:2408.06070*, 2024.

[10] S. Tu, Z. Xing, X. Han *et al.*, "Stableanimator: High-quality identity-preserving human image animation," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 21 096–21 106.

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[12] X. Zhang, R. Jiang, R. Willett *et al.*, "Nested diffusion models using hierarchical latent priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2502–2512.

[13] P. Shao, J. Feng, J. Lu *et al.*, "Data-driven and knowledge-guided denoising diffusion model for flood forecasting," *Expert Systems with Applications*, vol. 244, p. 122908, 2024.

[14] R. Rombach, A. Blattmann, D. Lorenz *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[15] T. Wang, L. Li, K. Lin *et al.*, "Disco: Disentangled control for referring human dance generation in real world," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[16] Z. Xu, J. Zhang, J. H. Liew *et al.*, "Magicanimate: Temporally consistent human image animation using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 1481–1490.

[17] J. Ho, T. Salimans, A. Gritsenko *et al.*, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First Conference on Language Modeling*, 2024.

[19] A. Siarohin, S. Lathuilière, S. Tulyakov *et al.*, "First order motion model for image animation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[20] J. Zhao and H. Zhang, "Thin-plate spline motion model for image animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3657–3666.

[21] J. Karras, A. Holynski, T. C. Wang *et al.*, "Dreampose: Fashion image-to-video synthesis via stable diffusion," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 22 623–22 633.

[22] X. Guo, M. Zheng, L. Hou *et al.*, "I2v-adapter: A general image-to-video adapter for diffusion models," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.

[23] W. Y. Yu, L. M. Po, R. C. C. Cheung *et al.*, "Bidirectionally deformable motion modulation for video-based human pose transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 7502–7512.

[24] J. Karras, A. Holynski, T. C. Wang *et al.*, "Dreampose: Fashion video synthesis with stable diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22 680–22 690.

[25] P. Zhang, L. Yang, J. H. Lai *et al.*, "Exploring dual-task correlation for pose guided person image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7713–7722.

[26] Y. Ren, X. Fan, G. Li *et al.*, "Neural texture extraction and distribution for controllable person image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 535–13 544.

[27] A. K. Bhunia, S. Khan, H. Cholakkal *et al.*, "Person image synthesis via denoising diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5968–5976.