Animate Anyone 2: High-Fidelity Character Image Animation with Environment Affordance

Supplementary Material

1. Dataset Details

Our video dataset was curated from two major video-sharing platforms, Bilibili and TikTok, with a roughly even split in contribution from each. The collection encompasses a broad spectrum of content, including dancing, daily human activities, sports, and excerpts from films. To ensure diversity, the dataset includes characters depicted in various poses, such as full-body and upper-body, set in both indoor and outdoor environments. It also incorporates a subset of videos with non-human characters (e.g., animations, avatars). The duration of each video clip ranges from 5 to 20 seconds. Furthermore, we applied a preprocessing pipeline to each clip to ensure a consistent character identity is maintained throughout the sequence.

2. Network Details

As shown in Fig.1, we adopt the UNet structure from Stable Diffusion[3] to construct our network architecture. The design of ReferenceNet (depicted in orange) is adapted from Animate Anyone[1]. ReferenceNet extracts appearance information from the character image and temporal information from the last frame of previous video clip. The appearance information is injected into the denoising network via spatial attention, while the temporal information is incorporated through temporal attention. The object information is injected into the denoising network via spatial blending. The merging of these information is applied in the Mid Block and Up Block of the UNet. During video inference, the appearance feature is extracted only once throughout the multi-clip iterative denoising process, contributing negligibly to the overall computational time. Temporal features are extracted once per video clip, which is more computationally efficient compared to methods that employ overlapping inference[1, 5].

3. User Study

Since existing quantitative metrics struggle to fully capture the perceptual quality of the results, we conducted a comprehensive user study for qualitative comparison. For this study, we curated a representative test set of 100 videos, held-out from the training data, which encompass a diverse range of character appearances, motions, and scene interactions. Twenty participants were recruited for a two-alternative forced-choice (2AFC) task. In each trial, they were presented with side-by-side results from our method and a competing approach, and were asked to select the one they perceived as superior. The evaluation was structured around three key dimensions: Character-Motion (C-M), Environment-Integration (E-I), and Object-Interaction (O-I). Tab.1 presents the results of the user study, demonstrating the subjective superiority of our method.

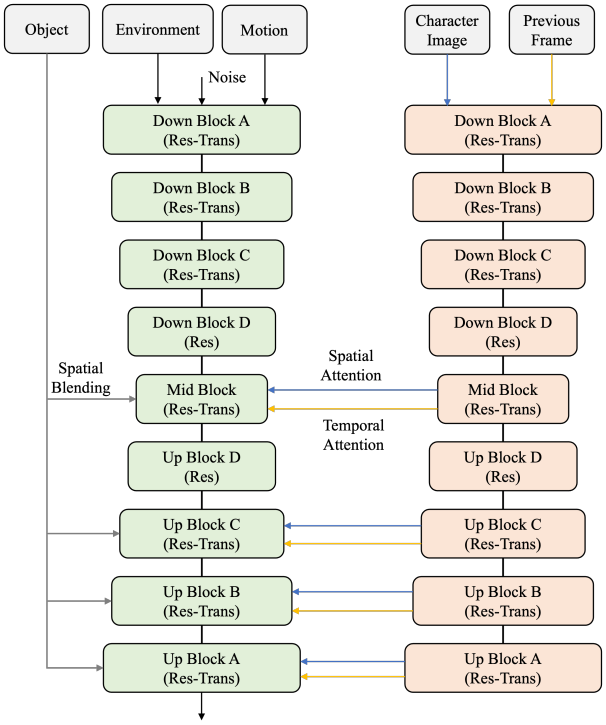

Figure 1. Network details.

Method	C-M	E-I	O-I
MovieCharacter	74.8%	94.8%	97.3%
MIMO	65.4%	81.5%	77.6%

Table 1. User preference rates of our method compared with other competitors across three aspects. C-M / E-I / O-I stand for Character-Motion / Environment-Integration / Object-Interaction.

Method	SSIM ↑	LPIPS ↓	FVD ↓
MRAA	0.749	0.212	253.6
DreamPose	0.885	0.068	238.7
Animate Anyone	0.931	0.044	81.6
UniAnimate	0.940	0.031	68.1
Ours	**0.943**	**0.030**	**63.6**

Table 2. Quantitative comparison on UBC benchmark.

Method	Time(Sec)	GPU memory(GB)
Object modeling	0.05	0.15
Pose modulation	0.07	0.37
Overall model	1.13	22.45

Table 3. Efficiency analysis.

Object Inconsistency Background Inconsistency

Figure 2. Failure cases of our method.

4. Evaluation on UBC

We conduct an evaluation on UBC benchmark[4] as shown in Tab.2. We only compared methods that reported UBC results in their papers.

5. Efficiency Analysis

Tab.3 presents an efficiency analysis of our method. We report the computational time per frame and GPU memory for our approach and its two key modules. The results were obtained on an A100 GPU with a inference resolution of 640×640. The denoising process employs 20 steps, and classifier-free guidance (CFG) is utilized during inference, with a guidance scale of 2.

6. Discussions on Failure Cases

When processing human-object interactions, our method relies on the SAM2[2] to extract the interacted objects. Consequently, the final quality of the interaction is inherently limited by the performance of SAM2. Fig. 2 (left) illustrates a failure case stemming from this dependency. Specifically, when SAM2 fails to robustly track and segment an object in its entirety, the integrity of the object in our final synthesis is compromised, often appearing fragmented or incomplete. Future work can explore integrating object understanding capability into generative model.

During character-environment fusion, our method may encounter challenges in visually complex scenes. This is because our masking strategy, while necessary, can lead to a loss of contextual information from the input environment. Consequently, the model may struggle to faithfully reconstruct all environmental details immediately adjacent to the character, thereby compromising the local consistency between the subject and the background. While employing a perfectly precise mask could mitigate this issue, it would exacerbate the problem of shape leakage from the source video. This reveals a trade-off between environmental consistency and character shape preservation. Fig. 2 (right) illustrates a instance of such consistency artifacts arising from the fusion process.

References

[1] Li 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*, pages 8153–8163, 2024. 1

[2] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[4] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 2

[5] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 1