Efficient Training for Human Video Generation with Entropy-Guided Prioritized Progressive Learning

Supplementary Material

A. Model Architecture for Human Video Generation

We propose a strong baseline model for human video generation based on diffusion transformer.

Transformer-based Denoiser with Temporal Attention. Our baseline model start with a Diffusion Transformer (DiT) designed for image generation. DiT is a latent diffusion model with VAEs to encode the higher level input to lower dimension latent feature. A transformer-based diffusion model learns the distribution of the latent features. We add a temporal attention layer to each block of the diffusion transformer to model the temporal relationship across frames.

Siamese ReferenceNet. ReferenceNet is widely used in controlled video generation as it can capture the detailed appearance of reference image. However, previous works train referencenet as a separate network with its own set of parameters. The optimization of the new set of parameters during training is computationally costly. To reduce GPU memory usage and speedup training, we propose Siamese ReferenceNet, which share exactly the same architecture and parameter of the denoiser network. The Siamese ReferenceNet is connected to the main denoiser network through cross-attention.

Efficient DiT ControlNet. We use an efficient transformer-based ControlNet for pose condition. We use the PixArt with only 1 transformer layer to minimize the memory consumption.

B. Additional Results

Qualitative comparison of ablation study. Fig. A presents the qualitative results of our ablation study on TikTok dataset. Removing any individual component leads to a noticeable degradation in video generation quality. In particular, omitting the adaptive progressive schedule significantly reduces the fidelity of background information. Moreover, eliminating the CEI results in severe deterioration of fine-grained details, highlighting its critical role in maintaining visual quality.

Effect of Siamese ReferenceNet. Given the superior performance of double network structures in generative tasks, we adopt a similar approach inspired by models like Animate Anyone [1] and Champ [2]. Our architecture includes an additional DiT-based reference network alongside 1-layer ControlNet blocks. To evaluate the effectiveness of image-conditioned generation, we conduct experiments on the Bilibili dataset. The results, as presented in Table I, indicate that our original model outperforms the double network structure (Reference Network, RN), achieving competitive results with a simpler architecture and lower memory and computational costs. Notably, our proposed Ent-Prog model achieves the best performance across all metrics, demonstrating enhanced training speed and improved generation quality.

These results qualitatively illustrates the effect of different components on efficient model training. Ent-Prog excels at preserving subject details and accurately synthesizing single-frame images, particularly in retaining fine body details and generating correct images.

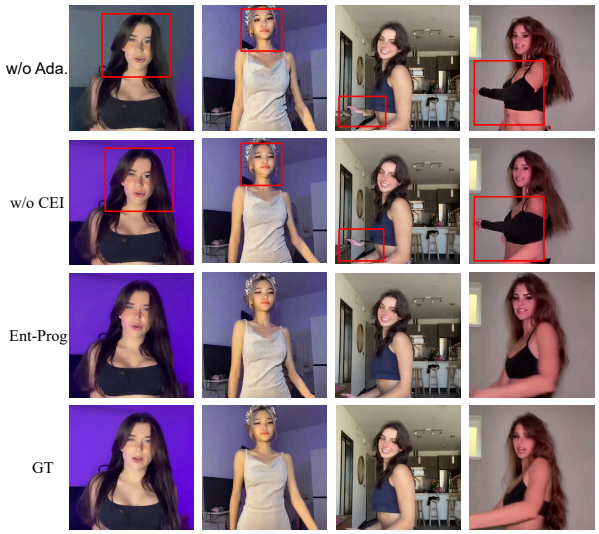

Figure A. Qualitative comparison of the ablation study on TikTok dataset. These results indicate that every component plays a critical role in facilitating efficient training and in maintaining the overall quality of the generated videos. Specifically, removing the adaptive progressive schedule leads to a substantial reduction in background fidelity, while eliminating the CEI module causes severe degradation of fine-grained details. The red boxes mark the most evident artifacts.

Training scheme	L1 ↓	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓
w/o RN	4.15e-5	0.745	26.32	0.330	176.02
Original	3.11e-5	0.718	28.42	0.279	115.42
Ent-Prog	**2.75e-05**	**0.783**	**28.56**	**0.279**	**95.98**

Table I. Comparison with different model structure.

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] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *ECCV*, 2024. 1