IDOL: Unified Dual-Modal Latent Diffusion for Human-Centric Joint Video-Depth Generation — Supplementary Material —

Yuanhao Zhai*[1], Kevin Lin[2], Linjie Li[2], Chung-Ching Lin[2], Jianfeng Wang[2], Zhengyuan Yang[2], David Doermann[1], Junsong Yuan[1], Zicheng Liu[3], and Lijuan Wang[2]

[1] State University of New York at Buffalo
[2] Microsoft
[3] Advanced Micro Devices
https://yhzhai.github.io/idol/

1 Additional experiments

Qualitative results. Please refer to the accompanying video for the qualitative comparison. In the video demonstration, we stitch each generated video snippets together to form a long video, which may lead to unnatural transitions between each snippet. We make the following observations. (1) Compared with DisCo [13], LDM3D [12], and MM-Diffusion [10], our generated videos and depth sequences exhibit more natural and smoother transitions, demonstrating the effectiveness of our method. A notable observation on the NTU120 dataset [6, 11] is LDM3D's limitation in generating diverse frames after fine-tuning the autoencoder, often resulting in repetitive "stuck" video sequences. This issue may stem from the subtle motions and predominant static content in NTU120 videos, suggesting that fine-tuning the autoencoder might necessitate a larger and more varied training dataset. (2) Our IDOL is able to composite different foreground and background, while simultaneously generating video and depth. This feature distinguishes it from concurrent methods [2, 4, 14], offering a unique capability in the area of video-depth synthesis. (3) Our method is able to generalize to different pose conditions, such as OpenPose [1] and DWPose [15].

Cross-attention map consistency. To enhance video-depth alignment in IDOL, we propose a cross-attention map consistency loss $\mathcal{L}_{\text{xattn}}$. This loss function encourages alignment of the video and depth cross-attention maps. We also explore alternate ways to align the cross-attention maps. One straightforward approach is to use a shared cross-attention map for both branches. We test two variations: replacing individual cross-attention maps with their average, and using the video stream's cross-attention map as a substitute for both. Our results in Tab. 1 reveal that sharing a cross-attention map significantly reduces performance, particularly impacting depth L2 accuracy (as seen in rows 2 and 3). These findings highlight the need for each stream to maintain diverse cross-attention maps to produce high-quality outputs. Our implementation of $\mathcal{L}_{\text{xattn}}$

[*] Work done during an internship at Microsoft.

Setting	Video		Depth	Image
	FID-FVD↓	FVD↓	L2↓	FID↓
-	19.28	260.65	0.0360	39.01
Share cross-attention map (avg)	20.82	297.76	0.0706	49.66
Share cross-attention map (video)	20.00	253.73	0.0718	44.58
Apply $\mathcal{L}_{\text{xattn}}$	**19.99**	**244.58**	**0.0351**	**37.89**

Table 1: Ablation study on the cross-attention map operations on the TikTok dataset [5] with HDNet depth [5].

effectively balances the need for consistency with the preservation of each map's unique characteristics, ultimately contributing to superior overall performance.

2 Implementation details

Our code is developed based on diffusers [8]. We follow DisCo [13] to use Stable Diffusion v1.4 [9] as the backbone. For HAOP pre-training, we follow DisCo [13] to freeze the ResBlocks and train the model for 25k steps, with input image size 256×256 and learning rate $1e^{-3}$. For fine-tuning, we adopt a two-stage approach. In the first stage, the temporal layers are removed, and the framework is trained on joint image-depth denoising. In the second stage, the whole framework with temporal modules is trained for the joint video-depth denoising. Both the first and the second stages are trained for 15k steps with a learning rate of $1e^{-4}$. For the second stage, the model is trained on 8-frame sequences. Both the pre-training and fine-tuning are conducted on 32 V100 GPUs. The weight hyper-parameters are set via a grid search: $w_{\text{mo}} = w_{\text{xattn}} = 0.01$. We set the temperature term τ in the motion field computation to $1/\sqrt{D_n}$, where D_n is the number of channels in the n-th layer.

Comparison methods. We use DisCo [13], a recent diffusion-based human dance video generation method, as a strong human-centric video generation baseline. As a pioneering method directly tailed for human-centric joint video-depth generation, we compare our IDOL with the closest multi-modal generation counterparts. We choose MM-Diffusion [10], initially designed for text-to-video-audio synthesis, and LDM3D [12], aimed at text-to-image-depth generation. To facilitate a fair comparison, we align their backbones to the same video LDM baseline, adapting them for the human-centric video-depth task. For MM-Diffusion [10], we replace the audio U-Net with a duplicate of the video U-Net (without sharing parameters, unlike in our IDOL) and retain the rest of the structure unchanged. In the case of LDM3D [12], we inflate the 2D U-Net to a 3D U-Net to accommodate video generation. Both adapted methods employ ControlNet [16] for human pose control and process background and foreground inputs similarly to IDOL, ensuring consistency in our comparative evaluation.

Generalization to different designs. In our main manuscript, we evaluated the generalization ability of our IDOL to different designs, including DWPose [15], AnimateDiff [3], and T2I-Adapter [7]. For the adaptation of AnimateDiff [3], we remove the original temporal convolutional and attention layers in the 3D

U-Net, and insert the AnimateDiff [3] pre-trained motion modules. For the T2I-Adapter [7], we replace the original pose ControlNet with a pre-trained OpenPose T2I-Adapter. Note that the video and depth streams share the same pose T2I-Adapter, similar to the pose ControlNet.

3 Negative societal impact

Our model raises ethical concerns, including the potential for creating deepfake videos, producing biased outputs, and threatening intellectual property rights. To mitigate these risks, we can incorporate invisible watermarks to ensure content authenticity.

References

1. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)
2. Chang, D., Shi, Y., Gao, Q., Fu, J., Xu, H., Song, G., Yan, Q., Yang, X., Soleymani, M.: Magicdance: Realistic human dance video generation with motions & facial expressions transfer. arXiv preprint arXiv:2311.12052 (2023)
3. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
4. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023)
5. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: CVPR. pp. 12753–12762 (2021)
6. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE TPAMI **42**(10), 2684–2701 (2019)
7. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
8. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
10. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: CVPR. pp. 10219–10228 (2023)
11. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: CVPR. pp. 1010–1019 (2016)
12. Stan, G.B.M., Wofk, D., Fox, S., Redden, A., Saxton, W., Yu, J., Aflalo, E., Tseng, S.Y., Nonato, F., Muller, M., et al.: Ldm3d: Latent diffusion model for 3d. arXiv preprint arXiv:2305.10853 (2023)

13. Wang, T., Li, L., Lin, K., Zhai, Y., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. In: CVPR (2024)
14. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. arXiv preprint arXiv:2311.16498 (2023)
15. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: ICCV. pp. 4210–4220 (2023)
16. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)