MOFA-Video: Controllable Image Animation via Generative Motion Field Adaptions in Frozen Image-to-Video Diffusion Model SUPPLEMENTARY

A Implementation Details

A.1 More Architecture Details of MOFA-Adapter

The proposed MOFA-Adapter is composed of three components: 1) Sparse-to-Dense Motion Generation Network (S2D network), 2) Reference Encoder, and 3) Fusion Encoder. We show the detailed architecture for feature merging in Fig. 1. The Fusion Encoder's architecture is identical to that of the SVD [3] Encoder. Forward warping is utilized for spatial warping operations within the feature space.

A.2 More Training Details

The trajectory-based model is trained on the WebVid-10M dataset [2] using the AdamW optimizer with a learning rate of 2×10^{-5}. The batch size is set at 8, and the total number of training iterations is 100,000. The portrait-based model is trained on a self-compiled dataset that includes 5,889 different human portrait videos. The AdamW optimizer is used, with a learning rate of 2×10^{-5}. The batch size is set to 1, and the total training iteration is 200,000.

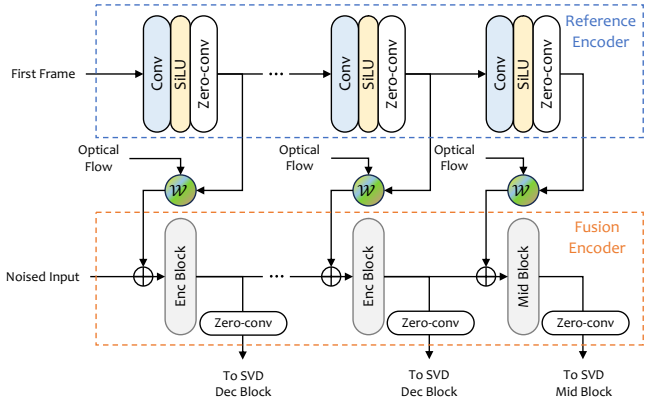

Fig. 1: ***Detailed architecture of MOFA-Adapter.***

A.3 Inference via Multiple MOFA-Adapters

As indicated in the main paper, we can integrate multiple MOFA-Adapters for more sophisticated and complex control using control signals from various modalities. For instance, users can merge the landmark signal with handcrafted trajectories. Specifically, we first route the trajectory control signals and landmark signals through the MOFA-Adapter for each modality separately. Contrary to the original Multi-ControlNet [7] algorithm, we employ a mask-aware strategy to define the control area for each MOFA-Adapter. Specifically, the user can designate the region where the landmark signal is accountable for, such as the human face region. Based on this mask, we extract the deep feature of the multi-scale output of the landmark-based MOFA-Adapters within the mask region, and that of the trajectory-based MOFA-Adapters outside the mask region. Finally, we input the combined features into the frozen SVD to obtain the final output.

B More Visual Results

In this section, we demonstrate more results generated by our methods. For video results, please refer to the video demo provided in the supplementary.

B.1 Trajectory-based Image Animation

The Trajectory-based Image Animation results are demonstrated in Fig. 2.

Camera Motion Control As stated in the main paper, besides handcrafted trajectories, our model is also capable of controlling camera motion via basic optical flow patterns. The corresponding results are illustrated in Fig. 3.

B.2 Portrait Image Animation

More portrait image animation results are demonstrated in Fig. 4.

B.3 Multi-MOFA Adapters

More advanced control results via Multi-MOFA Adapters are demonstrated in Fig. 5.

C Comparison Results

In this section, we demonstrate more comparison results against other methods. For video results, please refer to the video demo provided in the supplementary.

C.1 Trajectory-based

More comparison results with DragNUWA [6] are demonstrated in Fig. 6.

Motion Brush As stated in the main paper, we can employ motion mask brushes to attain detailed control by designating the spatial region of the flow patterns since our method utilizes intermediate optical flow patterns for motion control. Gen-2 [1] also supports motion brushes, but it only supports basic directions (Mask + direction) and is incapable of executing non-linear complex controls (for instance, blinking). Our method combines regional mask and trajectory (mask + trajectory), being able to handle advanced non-linear motions. The comparative results corresponding to this are displayed in Fig. 7.

C.2 Portrait Image Animation

More visual comparison results with StyleHEAT [5] and SadTalker [8] are demonstrated in Fig. 8. We also give more quantitative results with our methods and SadTalker [8] on visual quality (LPIPS). The proposed method shows a much better performance on visual quality (0.2099) than SadTalker (0.2308). Besides, our method also shows comparable results on lip synchronization from the same landmark generated from audio using SadTalker [8]. We give some examples in the supplementary video.

D Ablation Study Results

We also consider the quantitative comparison of the ablation studies in network structure. The same dataset is used for evaluation as the one used for quantitative comparisons with DragNUWA [6] in the main paper. As shown in Tab. 1, the proposed full method achieves the most balanced results in terms of all metrics. Our method w/o tuning of the MOFA-Adapter shows very limited motions

Methods	LPIPS ↓	FID ↓	FVD ↓
w/o warping	0.2619	18.80	184.27
w/o S2D	0.2376	<u>16.87</u>	**81.80**
w/o tuning	**0.2163**	16.97	102.17
Ours	<u>0.2274</u>	**16.82**	<u>86.76</u>

Table 1: Quantitative comparison results for ablation study on trajectory-based image animation.

compared with our full methods. Our method w/o S2D shows second-best results. However, from LPIPS, the generated video is different from the motion guidance. Finally, our method uses explicit warping as motion control, removing the explicit motion warping shows much worse results in all metrics. For video results, please refer to the video demo provided in the supplementary. We also provide the video ablation results for longer video generation and domain-aware MOFA-Adapter in the video demo.

E Video Demo

We provide the video demo in the supplementary, which includes brief introduction, video results, ablation studies, and the limitations of our method.

F Limitations

Unlike SORA [4], our method struggles to control or generate new content that is significantly different from the provided image, as the current video diffusion model is only trained on a limited number of video clips. Additionally, our model may encounter visual artifacts such as blurriness or loss of structure under extensive motion guidance. Visual examples of these issues are provided in Fig. 9.

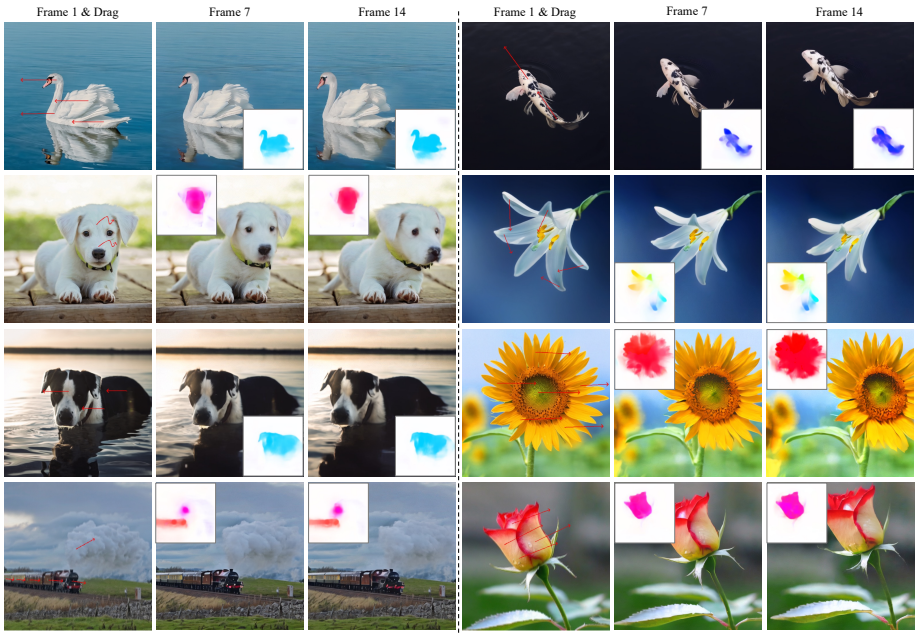

Fig. 2: ***More visual results for trajectory-based image animation.***

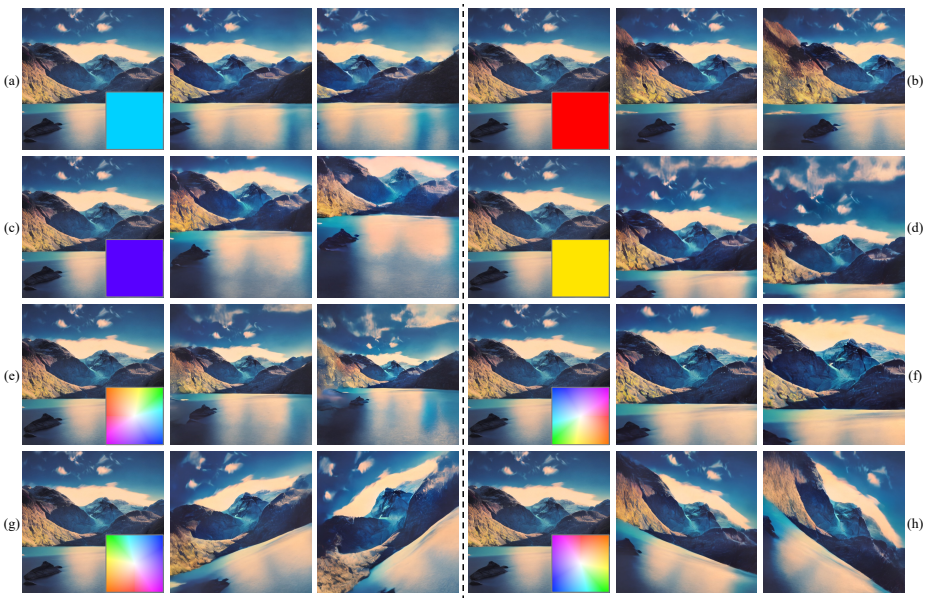

Fig. 3: ***Camera motion control via fixed optical flow patterns.*** (a) Pan Right, (b) Pan Left, (c) Pan Down, (d) Pan Up, (e) Zoom Out, (f) Zoom In, (g) Clockwise, (h) Counter-Clockwise.

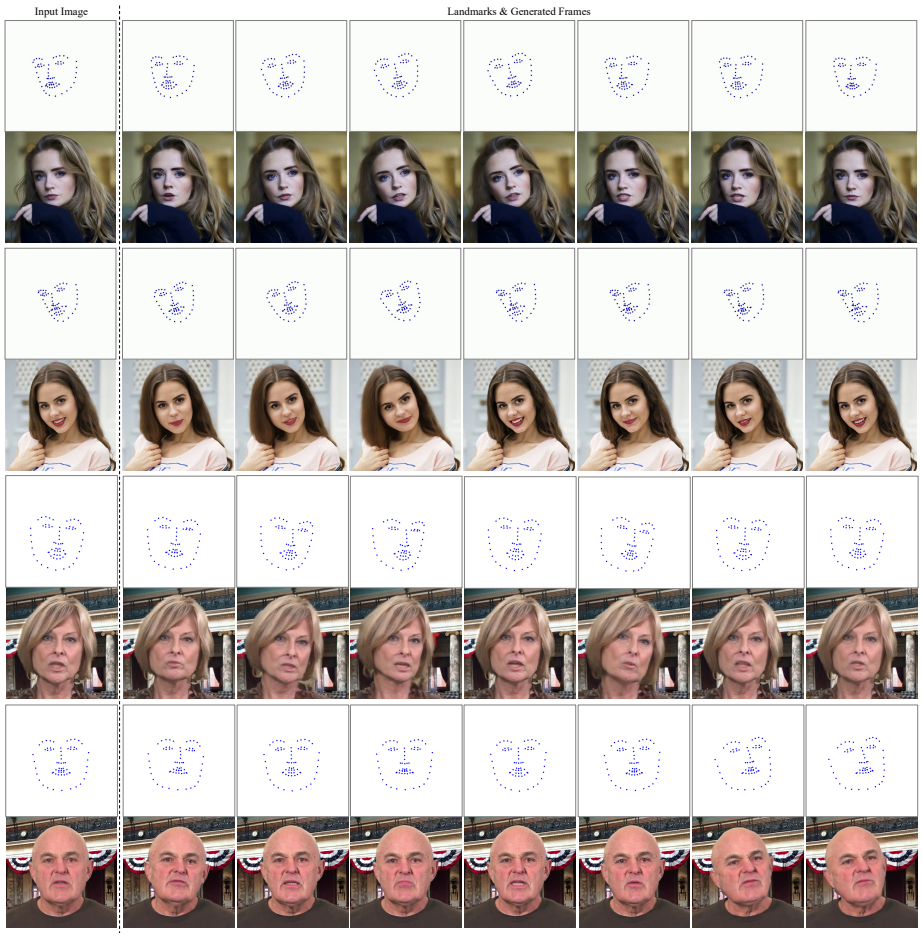

Fig. 4: *More visual results for portrait image animation.*

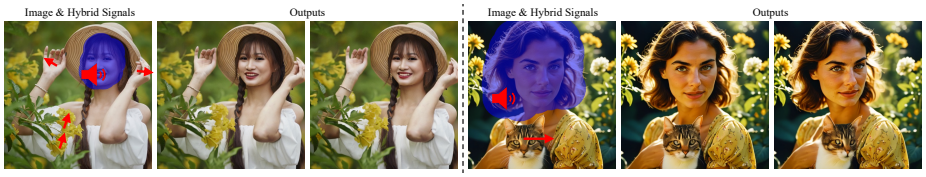

Fig. 5: *Visual results for advanced control with Multi-MOFA Adapters.*

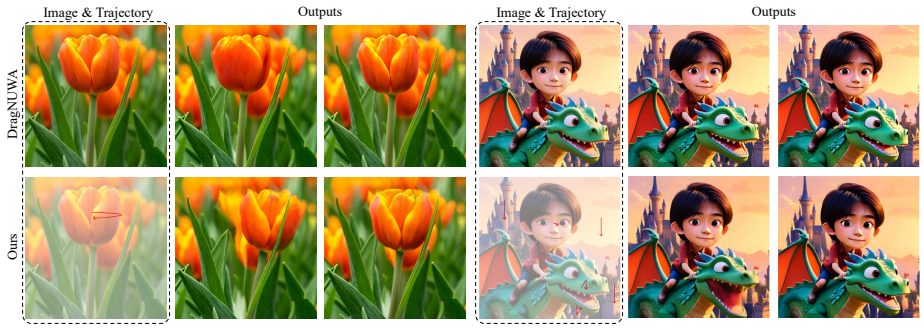

Fig. 6: ***Visual comparisons with DragNUWA [6] for trajectory-based image animation.***

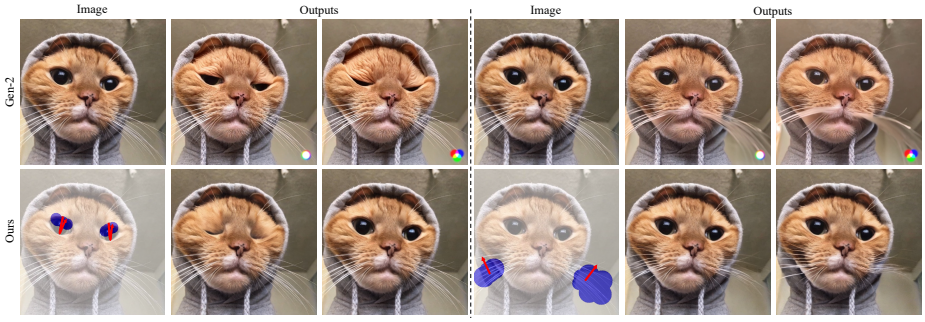

Fig. 7: ***Image animation results from our method and Gen-2 [1].*** Gen-2 employs a mask + direction approach, which is not suitable for managing complex motions. In contrast, our method integrates trajectory control with motion brushes, enabling advanced non-linear control (*e.g.*, blinking) for the target objects.

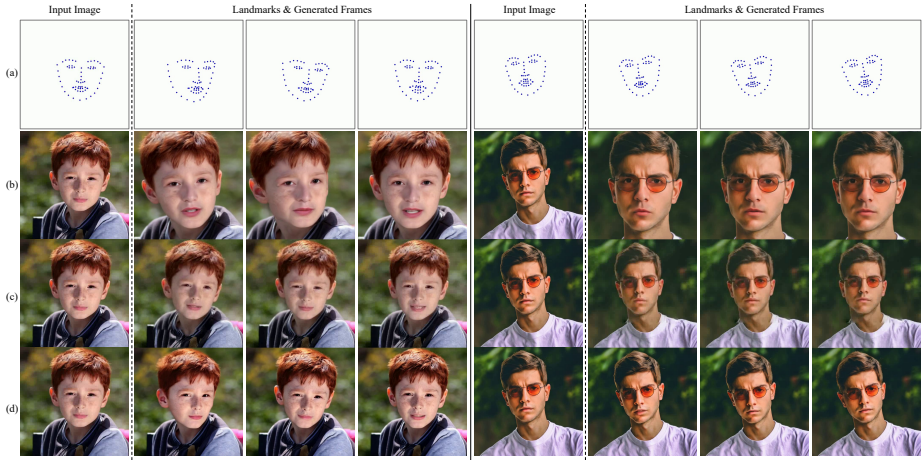

Fig. 8: ***More visual comparisons for portrait image animation.*** (a) StyleHEAT [5], (b) Sadtalker [8], (c) Ours.

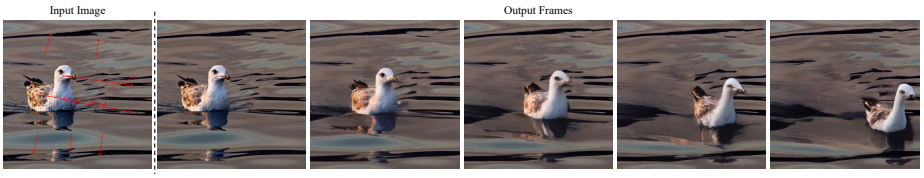

Fig. 9: ***Limitation of our method.*** Our model may encounter visual artifacts such as loss of structure or blurriness under extensive motion guidance.

References

[1] Gen-2. https://runwayml.com/ai-magic-tools/gen-2/ (2023) 3, 7

[2] Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021) 1

[3] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 1

[4] Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators 4

[5] Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: European conference on computer vision. pp. 85–101. Springer (2022) 3, 7

[6] Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023) 3, 7

[7] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023) 2

[8] Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8652–8661 (2023) 3, 7