MotionPro: A Precise Motion Controller for Image-to-Video Generation — CVPR 2025 Supplementary Material*

Zhongwei Zhang[1], Fuchen Long[2], Zhaofan Qiu[2], Yingwei Pan[2]†, Wu Liu[1]†, Ting Yao[2], and Tao Mei[2]
[1]University of Science and Technology of China [2]HiDream.ai Inc.
zhwzhang@mail.ustc.edu.cn, {longfuchen, qiuzhaofan, pandy}@hidream.ai
liuwu@live.cn, {tiyao, tmei}@hidream.ai

The supplementary material contains: 1) the dataset details of MC-Bench; 2) baseline choices and experimental details; 3) the human evaluation of motion control; 4) robustness of motion mask; 5) the application of camera control; 6) runtime comparison; 7) ablation on control signals.

1. Dataset Details of MC-Bench

The proposed MC-Bench consists of 412 high-quality reference images and corresponding $1.1K$ user-annotated trajectories. We collect the reference images with different visual contents, including animal, human, vehicle, etc. There are 72 images sampled from the public DragBench [5] and we further extend it with 340 additional images. Specifically, all the self-collected images about human are automatically generated by DALL·E3 [2] to avoid the potential legal concerns. The remaining self-collected images are real photos which are first crawled on the Pexels platform and then filtered according to the visual quality. For each reference image, the annotator is required to brush the motion region and draw the movement trajectory according to user intention (i.e., fine-grained local part moving or global object moving). During trajectory annotation, all annotators are encouraged to ensure the trajectory diversity, including some complicated trajectories. Finally, the benchmark is annotated with 460 image-trajectory pairs for fine-grained motion control evaluation, and 680 image-trajectory pairs for object-level motion control evaluation, respectively. Figure 3 and Figure 4 further illustrate several visual examples (reference image, trajectory and motion mask) from MC-Bench for the two evaluations.

2. Baseline Choices and Experimental Details

For the evaluation on WebVid-10M [1] of fine-grained motion control, we adopt the commonly-used protocol in recent controllable image-to-video (I2V) advance [3]. Specifically, for each video, we sample the optical flow at the ratio of 15% as the sparse trajectories, which are combined with the first frame as the input condition. Under this experimental setting, we choose DragNUWA [7] and MOFA-Video [3] as baselines for comparison. Notably, DragAnything [6] is deliberately designed for object-level motion control, which only accepts a single trajectory of object, making it inapplicable for fine-grained motion control. Therefore, DragAnything is not involved for comparison in this setting.

For the fine-grained motion control on MC-Bench, we compare our MotionPro with DragDiffusion [5] and MOFA-Video. DragNUWA is not included in this comparison since it only relies on trajectories and lacks the input of motion regions. Thus, DragNUWA usually suffers from the misinterpretation of object and camera movement, making the comparison unfair. The baseline of DragDiffusion is a recent trajectory-guided image editing advance, which also offers convincing results for comparison. To adapt DragDiffusion for video generation, we divide the input trajectories into 15 segments and independently feed each segment into DragDiffusion to generate target frame. All the synthesized frames are concatenated as the final video.

In the evaluation of object-level motion control on MC-Bench, both MOFA-Video and DragAnything are employed as baselines for performance comparison. To facilitate DragAnything in disentangling object and camera moving, we add static points in regions outside the motion mask areas to help DragAnything generate object-level motion instead of camera moving for evaluation. It's worth to noting that MotionPro learns object and camera motion control on "in-the-wild" video data (e.g., WebVid-10M) without applying special data filtering.

3. Human Evaluation

In addition to the evaluation over automatic metrics, we also conduct human evaluation to investigate user preferences from three perspectives (i.e., motion quality, temporal coherence and trajectory alignment) across different controllable I2V approaches. In particular, we randomly sample

*This work was performed at HiDream.ai.
†Co-corresponding author.

Table 1. Human evaluation of user preference ratios (%) over both fine-grained and object-level motion control on MC-Bench.

Evaluation Items	Fine-grained Motion Control			Object-level Motion Control		
	DragDiffusion [5]	MOFA-Video [3]	MotionPro	MOFA-Video [3]	DragAnything [6]	MotionPro
Motion Quality (↑)	3.12	21.88	**75.00**	12.50	18.75	**68.75**
Temporal Coherence (↑)	6.25	40.63	**53.12**	25.00	15.63	**59.37**
Trajectory Alignment (↑)	9.37	18.75	**71.88**	15.62	21.88	**62.50**

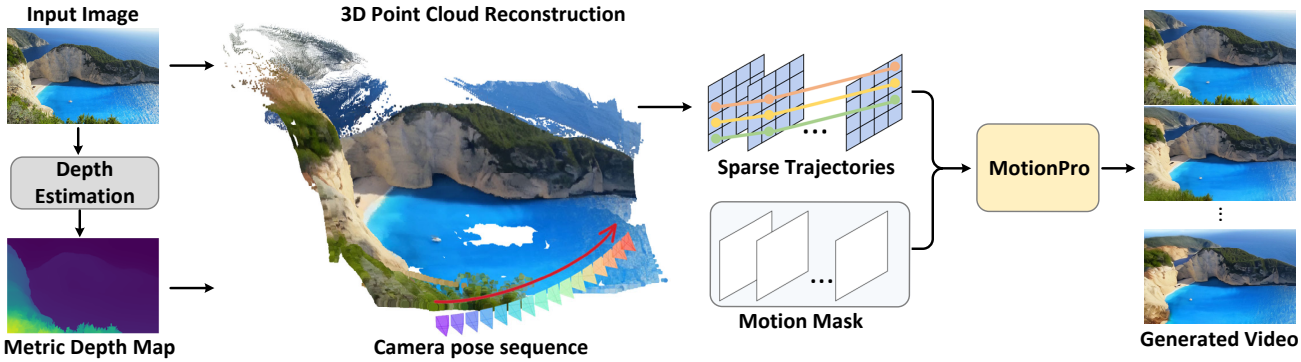

Figure 1. An illustration of I2V camera control using the condition of camera pose sequence in our MotionPro.

200 generated videos from both fine-grained and object-level motion control for evaluation. Through the Amazon MTurk platform, we invite 32 evaluators, and ask each evaluator to choose the best one from the generated videos by all models given the same inputs.

Table 1 shows the user preference ratios across different models on MC-Bench. Overall, our MotionPro clearly outperforms all baselines in terms of the three criteria on both fine-grained and object-level motion control. The results demonstrate the advantage of leveraging complementary region-wise trajectory and motion mask to benefit video synthesis with natural motion, desirable temporal coherence and precise motion-trajectory alignment.

4. Ablation on control signals.

We also include two runs (MotionPro$^{-}_{traj}$: replaces region-wise trajectory with random trajectory, MotionPro$^{-}_{mask}$: disables motion mask with all-one masks). Their FVD (73.7 and 66.2) on WebVid-10M are inferior to our MotionPro (59.88), which validates the effectiveness of our two control signal designs for precise motion formulation.

5. Robustness of motion mask

To be clear, motion mask in our MotionPro refers to the rough dynamic region and does not require precisely-aligned shape at inference. We show I2V results controlled by the same trajectory with various motion masks in Figure 2, which show strong robustness. Such generalization merit is attributed to the use of estimated motion mask (flow map estimated by DOT) at training, rather than ground-truth precise motion mask.

6. Application: Camera Control

Our learnt MotionPro naturally supports two applications of camera control without additional training. The first application is controlling object and camera motion simultaneously with multiple trajectories in I2V generation. Another application is the I2V camera control by exploiting the sequence of camera poses as input condition. To be clear, motion mask in our MotionPro refers to the rough dynamic region and does not require precisely-aligned shape at inference.

Simultaneous object and camera motion control. In this setting, we simply set the input motion mask as all-ones matrix, and feed multiple trajectories that reflect the object and background moving into MotionPro for I2V generation. The video cases are provided in the offline project page.

Camera control with camera poses. Figure 1 illustrates the process of camera control using the condition of camera pose sequence in MotionPro. Concurrently, given an input image and the camera pose sequence, we first estimate the metric depth map of the image using ZoeDepth [4]. Next, we lift the 2D pixels to 3D point cloud using the metric depth map. Through projecting the point cloud into 2D space given the camera pose, we can determine the corresponding 2D positions of the same 3D points under the new view. By calculating the 2D displacement of the pixels projected from the same 3D points in the original and new views, the camera pose sequence is then converted into the

sparse trajectories. Finally, we feed the sparse trajectories and all-ones motion mask into MotionPro for I2V synthesis. The video cases are provided in the offline project page.

7. Runtime Comparison

For 16-frame video generation (resolution: 512×320, on single NVIDIA H100 GPU), the runtime of MotionPro is 17 sec, which is comparable to baselines (DragNUWA: 27, DragDiffusion: 320, MOFA-Video: 15, DragAnything: 32).

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *ICCV*, 2021. 1

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. Improving Image Generation with Better Captions, 2023. 1

[3] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. MOFA-Video: Controllable Image Animation via Generative Motion Field Adaptions in Frozen Image-to-Video Diffusion Model. In *ECCV*, 2024. 1, 2

[4] Diana Wofk Peter Wonka Matthias Müller Shariq Farooq Bhat, Reiner Birkl. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. In *CVPR*, 2023. 2

[5] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. DragDiffusion: Harnessing Diffusion Models for Interactive Point-Based Image Editing. In *CVPR*, 2024. 1, 2

[6] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion Control for Anything using Entity Representation. In *ECCV*, 2024. 1, 2

[7] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. DragNUWA: Fine-Grained Control in Video Generation by Integrating Text, Image, and Trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 1

Figure 2. I2V with various mask shapes (GIF Videos). Please view in **Adobe Reader**.

Figure 3. Visual examples from MC-Bench for fine-grained motion control evaluation. Each reference image is annotated with trajectory and motion mask for image-to-video generation.

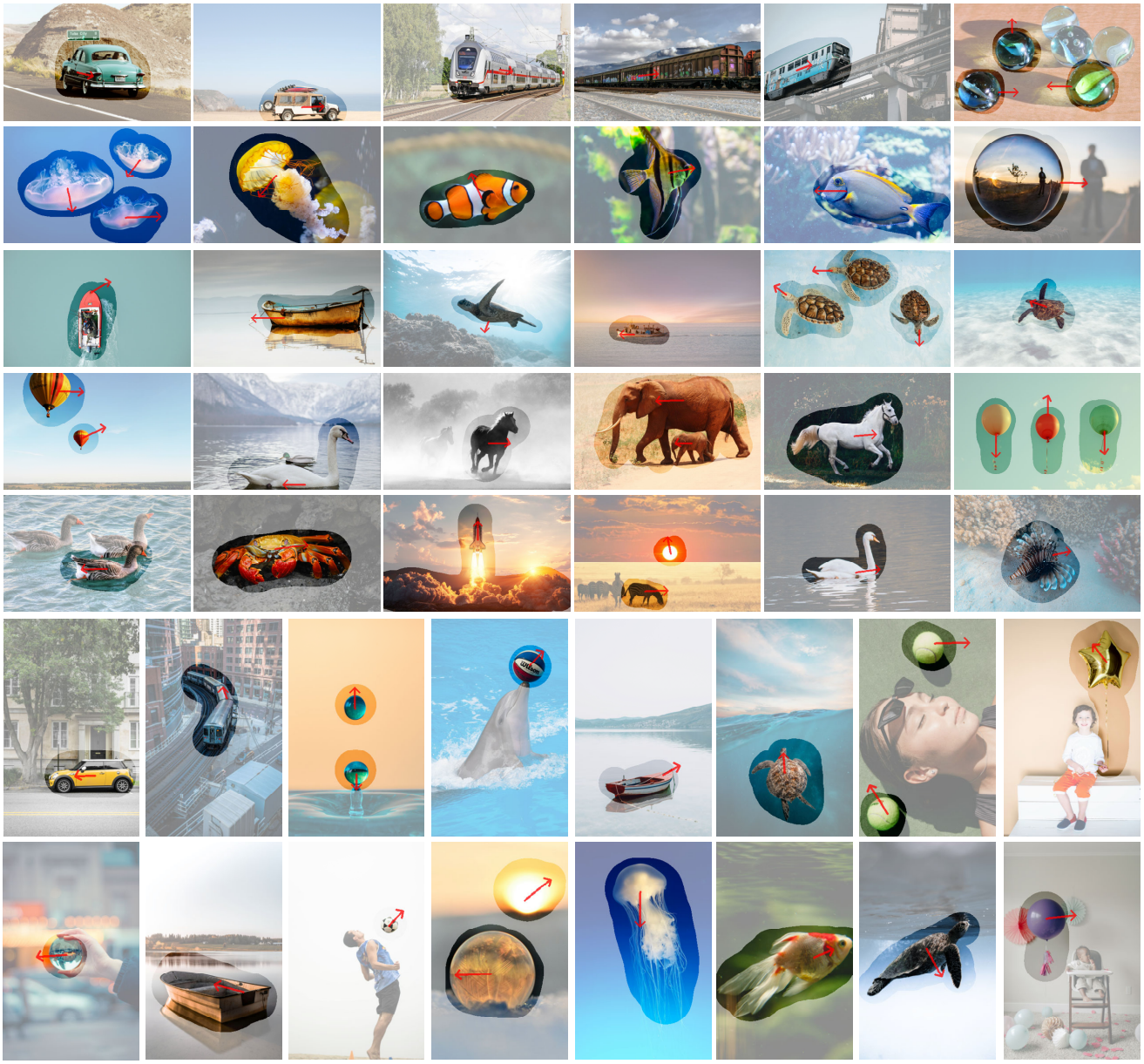

Figure 4. Visual examples from MC-Bench for object-level motion control evaluation. Each reference image is annotated with trajectory and motion mask for image-to-video generation.