Extrapolating and Decoupling Image-to-Video Generation Models: Motion Modeling is Easier Than You Think

Supplementary Material

1. More Relative Research in Video Diffusion Models

To extend Diffusion models (DMs) to video generation, the first video diffusion model (VDM) [11] has been proposed, which utilizes a spacetime factorized U-Net to model low-resolution videos in pixel space. Imagen-Video [10] introduces efficient cascaded DMs with v-prediction for producing high-definition videos. To mitigate training costs, subsequent research [2, 9, 25, 38] has focused on transferring T2I techniques to text-to-video (T2V) [7, 13, 21, 33], as well as on developing VDMs in latent or hybrid pixel-latent spaces. Similar to the addition of controls in text-to-image (T2I) generation [16, 19, 29, 34], the introduction of control signals in text-to-video (T2V) generation, such as structure [6, 26], pose [14, 36] has garnered increasing attention. Nonetheless, visual image conditions in video diffusion models (VDMs) [22, 30], remain under-explored. Recent works, including Seer [8], VideoComposer [24], have investigated image conditions for image-to-video synthesis. However, these approaches either focus on curated domains like indoor objects [24] or struggle to produce temporally coherent frames and realistic motions, often failing to preserve visual details of the input image [35]. Recent proprietary T2V models [3, 15, 21, 23, 32] show potential for extending image-to-video synthesis but often lack adherence to the input image and suffer from unrealistic temporal variations. In this paper, we focus on the image-conditioned video generation task.

2. Dataset Selection

We choose WebVid for fair comparison with prior work CIL [37] and its wide adoption by DynamiCrafter [27], Consisti2v [18], and Motion-i2v [20], due to its uniform resolution and sufficient size. Other datasets like Panda-70M [5] have watermarks and blurriness, OpenVid [17] lacks resolution consistency, Vript [28] is too small for effective training.

3. Generalization of Our Framework.

To demonstrate the generalization, we further perform experiments on SVD [1]. As shown in Table 1, our method also delivers performance improvements on motion degree and motion control.

Table 1. Applying our method in Stable-Video-Diffusion

Model	Video Quality	Motion Degree	Motion Control
SVD	66.38	41.94	20.41
SVD-Ours	**67.24**	**55.16**	**36.53**

4. Model Merging Method

DARE-Pruning [31] DARE employs a parameterized Bernoulli distribution to sample a sparse mask $\boldsymbol{m}^t$, which is then applied to the parameters $\boldsymbol{\delta}$ and rescaled by the mask rate p:

$$\begin{aligned} \boldsymbol{m}^t &\sim \text{Bernoulli}(p), \\ \widetilde{\boldsymbol{\delta}}^t &= \boldsymbol{m}^t \odot \boldsymbol{\delta}^t, \\ \hat{\boldsymbol{\delta}}^t &= \frac{\widetilde{\boldsymbol{\delta}}^t}{1-p}. \end{aligned} \tag{1}$$

Task-Arithmetic [12] Task-Arithmetic introduces the concept of "task vectors." A task vector is obtained by subtracting the weights of a pre-trained model from the weights of the same model after fine-tuning. Performance on multiple tasks can be improved by combining vectors from different tasks. Formally, let $\theta_{\text{pre}} \in \mathbb{R}^d$ be the weights of a pre-trained model, and $\theta^t_{\text{ft}} \in \mathbb{R}^d$ the weights after fine-tuning on task $t \in \{1, \ldots, T\}$. The task vector $\tau_t \in \mathbb{R}^d$ is given by:

$$\tau_t = \theta^t_{\text{ft}} - \theta_{\text{pre}} \tag{2}$$

We can obtain a multi-task version of the model θ_m by summing the task vectors:

$$\theta_m = \theta_{pre} + w \sum_{t=1}^{T} \tau_t \tag{3}$$

where w is a hyperparameter.

5. Details of User Study.

For user study, we randomly select input image and prompt pairs in Vbench and then generate videos by using Ours, SVD [1] and VideoCrafter [4] with DynamiCrafter[27], CIL [37] and DC-FT (fine-tuned DynamiCrafter). In the setup, we conducted pairwise comparisons between our model and other methods, inviting users to evaluate and select the superior one in terms of quality, consistency, dynamism, and

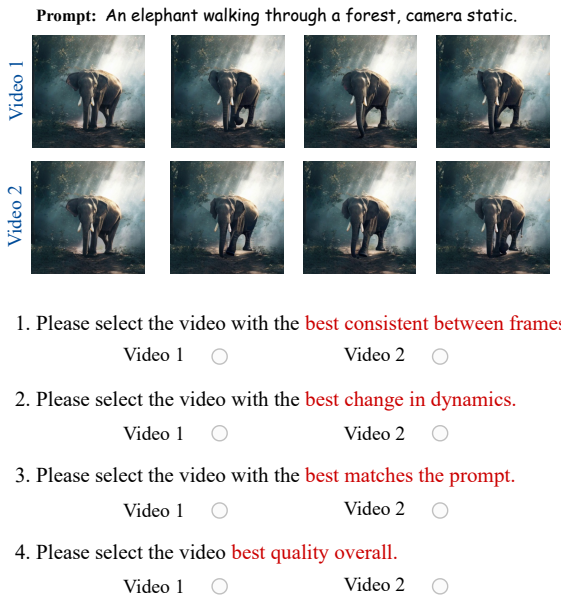

Figure 1. Example of user study questionnaires.

instruction following. We show an illustration of the question cases in Figure 1. There are 40 video comparisons, comprising a total of 160 questions, with the order of the synthetic videos being shuffled. The survey involved a total of 200 participants. Following the approach outlined in [37], we calculated the preference rates, with the results presented in the main paper.

For camera motion, each command was executed 5 times on the 8 images, and the average success rate for each category was then computed. The human evaluation performance is shown in Table 2 and our method improves all categories over DynamiCrafter. Notably, as the training data for "Zoom out" is extremely scarce in the original DynamiCrafter (0.26%), both the DynamiCrafter and our method have relatively low scores on the zoom-out task and are not entirely flawless, although our method can improve it.

Table 2. Camera Movement and Dataset Distribution.

Model	Pan left (6.91%)	Pan right (7.27%)	Tilt up (4.07%)	Tilt down (1.65%)	Zoom in (0.94%)	Zoom out (0.26%)
DynamiCrafter	0.43	0.53	0.53	0.28	0.33	0.08
Ours	**0.45**	**0.55**	**0.60**	**0.35**	**0.53**	**0.33**

6. Examples of Different Scenario.

In this section, we present additional examples of generated videos, as shown in Figure 2. The examples cover various real-world scenes, including human figures, natural phenomena, animals, and car movements. The human figures are generated with complete and fluid actions. In the second row, the generated waves submerge a lighthouse. In the third and fourth rows, the model preserves high consistency.

Figure 2. The figure above illustrates the performance of our model in generating videos across various categories, including humanities, natural phenomena, animals, and modern transportation.

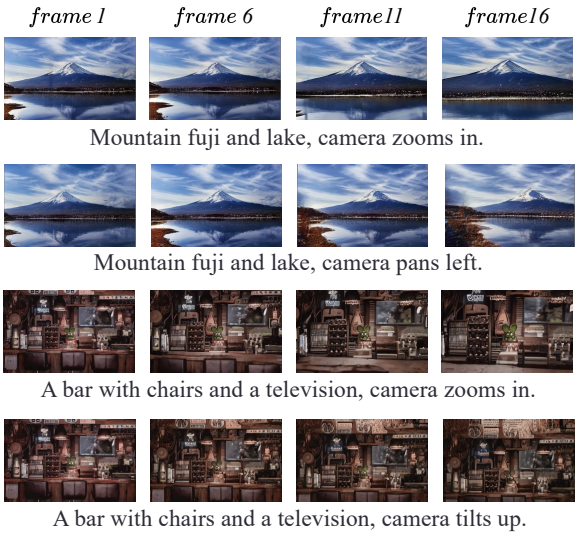

Figure 3. Visualization of Controllable Camera Movement with Different Text Prompts. Examples featuring simple and distinct main subjects are illustrated above, while those showcasing complex and disordered backgrounds are depicted below.

For the same image input, different text instructions are used to control the variations in the actions within the generated videos. In the top two rows of Figure 3, an image of Mount Fuji is fed into the model, with the text instruc-

tions “camera zooms in” and “camera pans left” appended to control the corresponding camera movements in the output. The bottom two rows provide examples of camera control in more complex scenes, utilizing similar instructions such as “camera zooms in” and “camera tilts up.” Despite the high visual similarity, the generated videos respond well to these instructions.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1

[3] Guanjie Chen, Xinyu Zhao, Yucheng Zhou, Tianlong Chen, and Cheng Yu. Accelerating vision diffusion transformers with skip branches. *arXiv preprint arXiv:2411.17616*, 2024. 1

[4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 1

[5] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers, 2024. 1

[6] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 1

[7] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 1

[8] Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023. 1

[9] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2(3):4, 2022. 1

[10] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[11] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1

[12] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. 1

[13] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 1

[14] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 1

[15] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 1

[16] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1

[17] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation, 2024. 1

[18] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 1

[19] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 1

[20] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1

[21] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[22] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[23] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi

Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022. 1

[24] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[25] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1

[26] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[27] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 1

[28] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words, 2024. 1

[29] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1

[30] 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

[31] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*. 1

[32] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 1

[33] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 1

[34] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[35] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 1

[36] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 1

[37] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. *arXiv preprint arXiv:2406.15735*, 2024. 1, 2

[38] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1