Explorative Inbetweening of Time and Space

Supplementary Material

1 Perceptual Study

We conducted a perceptual study to measure human preference between our method and the corresponding baseline. Using Amazon Mechanical Turk (AMT), each participant was presented with 30 pairwise results. The participants were instructed to select the video they found more "realistic, of higher quality, and exhibiting more natural motions and transitions". In each pair, one video was randomly assigned to be from our method, while the other one was the corresponding generation from the closest baseline. The videos presented were randomly selected from either of the three tasks. To ensure the validity of the responses, we included 5 control trials within these comparisons with clearly unnatural videos. From this study, we collected 66 valid responses. The preference rate, indicating the proportion of participants favoring our method over the baseline, was then calculated based on the valid responses.

The results are shown in Table 1. The study shows a clear preference for our method in all three tasks with an overall average preference rate of 83.67%. Particularly, we obtain the higher rate on view-bound results with a 97.79% preference rate. Note that this task (generating camera trajectories from two sparse and unposed views) has traditionally been considered difficult, as also acknowledged by Du *et al.* [1]. While the quality of their method significantly degrades when no camera pose is given, exhibiting blurry and unclear images, our work retains the sharpness and quality of SVD and generates reasonable camera trajectories.

Overall Avg.	View bound	Identical bound	Dynamic bound
83.67%	97.79%	70.28%	82.94%

Table 1: Perceptual study: Preference rates for each of the three subtasks, compared against Du *et al.* [1], Text2Cinemagraph [2] and FILM [3].

2 Further Discussion

Probing I2V models. The bounded generation task along with TRF can offer a unique lens to evaluate SVD's world dynamics understanding. Given two observations, we can analyze how the I2V model connects the motion trajectory,

allowing us to compare the generated and the observed real-world dynamics. For example, the results on Dynamics Bound on the top of Fig. 7 (main paper) indicate the model's ability to understand and generate complex kinematics trajectories of articulated human bodies under different clothing, lighting, or with different image quality. Beyond articulated motion, the results of rows 2 and 4 indicate an ability to synthesize non-rigid motions like expression transitions and hair movements. In addition, the View Bound scenario exhibits 3D consistency across diverse real-world scenes, showcasing the model's generalization ability and 3D understanding of the physical world. The looping videos generated with identical bound indicate how well the model understands the implicit movement tendencies within a static image. These results suggest that applying similar techniques to other I2V models can serve as a way to probe the type and complexity of the dynamics that the model has learned.

The importance of the motion bucket ID. While our Time Reversal Fusion (TRF) method successfully achieves bounded generation without additional training, it does require careful tuning of the temporal conditioning parameters, such as motion bucket ID and frames per second (fps), to produce visually coherent outputs for different inputs. A critical aspect to note is the necessity for a match between the image content and the motion ID. This requirement stems from the underlying principles of Stable Video Diffusion (SVD), where the motion ID influences the intensity of pixel movement in the generated video – higher values result in more dynamic pixel behavior and vice versa. Selecting an appropriate motion ID range is crucial for each input image based on its dynamic contents; otherwise, the generated video may exhibit artifacts. Interestingly, even though bounded generation poses a more complex challenge than straightforward sampling from SVD – requiring the model to generate specific motion trajectories that may not align with its typical motion distribution – our TRF method can effectively alleviate motion incompatibility artifacts. We believe this is due to the fact that the second view acts effectively as a constraint, providing additional guidance for the generation process. Through this we can mitigate the problem of motion ID in SVD, except in cases where the original motion ID is significantly inaccurate. For example, in a static scene, a large motion ID may lead to excessive camera motion or unnatural addition of moving objects into the scene. Conversely, a smaller ID typically results in more subtle camera movements. However, if two wide-baseline views are significantly different, fusing them might inevitably lead to cut or blend effects due to insufficient dynamics that can seamlessly bridge the views.

Limitations. One limitation of our method stems from the stochasticity involved in the generation of the forward and backward passes. For two given images, the distribution of motion paths that SVD can take might vary significantly. This means that the start- and end-frame paths could generate very different videos, resulting in an unrealistically fused video. In the extreme case where start and end frames are completely unrelated, it is generally difficult to obtain good results, due to the constrained generative ability of SVD and the limited video

duration. In addition, our method inherits several limitations of SVD. For example, we observed that in some cases fine-grained color details cannot be well reconstructed. This is mainly due the resolution of the VQ-VAE encoder, and since the starting frame is already encoded with artifacts, the generated video retains them. Further, while SVD's generations suggest strong understanding of the physical world, there is still a lack of understanding regarding "common sense" and causal effect. For example, given an image of the famous moon landing, TRF generates a loop video in which the planted flag moves as if there was wind, which is not possible given the known context of the location. This is not only inaccurate, but could potentially bring ethical issues –e.g. the previous example could be misused as proof that the moon landing never happened. Video examples are shown in our project page.

Interestingly, there are some limitations of SVD that can be mitigated or resolved by our method. For example, SVD usually struggles with complex kinematic motions such as body limbs movement. Here, the generation tends to degrade throughout time, performing worse the further it is from the initial frame. On the other hand, TRF regularizes this through the bi-directional generation process, and can generate good-quality body motion between complex and distinct body poses.

Performance Even though our method requires two diffusion paths, the additional performance cost is not necessarily significant since the two paths are independent and can be run in parallel. The noise re-injection step also does not incur in significant additional cost, since it is only employed during the first few denoising steps. Hence, TRF leads to only marginally slower inference times (11 seconds more on an NVIDIA A100).

3 Additional Details

3.1 Inference

All our results were obtained with the base I2V model 'stabilityai/stable-video-diffusion-img2vid-xt' (model card on huggingface.com) under the scheduler setup of 50 inference steps (not including the noise re-injection / re-denoising steps) with euler EDM sampler. After considering the noise re-injection steps, the required denoising steps varies from 55 to 100 given different samples, with an A100-80GB gpu. Most of the paper results were obtained in under 4 minutes.

3.2 Quantitative evaluations with FVD

Evaluation setup. We compare FVD [4] scores on our self-collected frame interpolation dataset, and the eulerian-validation dataset. On our self-collected frame interpolation dataset, we compare ours with FILM [3]. For a fair comparison with the latter, we generate videos of 33 frames and then sample the video to 25 frames. We compute FVD using those 25 frames against the ground truth

videos (which also consist of 25 frames). On the eulerian-validation dataset, we compare ours against text2cinemagraph [2]. We resize our generated videos and the ground-truth videos to 512×512, to keep consistent with text2cinemagraph's results, and then sample both the ground-truth videos and text2cinemagraph's results to 25 frames.

Sanity baseline. We ran the FVD score for a baseline that merely duplicates the first frame throughout the video. The resulting FVD score is 1559.4, which is more than 3 times higher than our method and 70% higher than the comparison baseline.

4 Acknowledgments

The authors would like to thank Tsvetelina Alexiadis, Taylor McConnell, and Tomasz Niewiadomski for the great help with the perceptual user study. Special thanks are also due to Liang Wendong, Zhen Liu, Weiyang Liu, Zhanghao Sun, Yuliang Xiu, Yao Feng, Yandong Wen for their proofreading and insightful discussions.

Conflict of interest disclosure for Michael J. Black MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon and Meshcapade GmbH. While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

References

1. Du, Y., Smith, C., Tewari, A., Sitzmann, V.: Learning to render novel views from wide-baseline stereo pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4970–4980 (2023)
2. Mahapatra, A., Siarohin, A., Lee, H.Y., Tulyakov, S., Zhu, J.Y.: Text-guided synthesis of eulerian cinemagraphs. ACM Transactions on Graphics (TOG) **42**(6), 1–13 (2023)
3. Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C., Curless, B.: FILM: Frame interpolation for large motion. In: European Conference on Computer Vision. pp. 250–266. Springer (2022)
4. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)