Video Interpolation with Diffusion Models

Supplementary Material

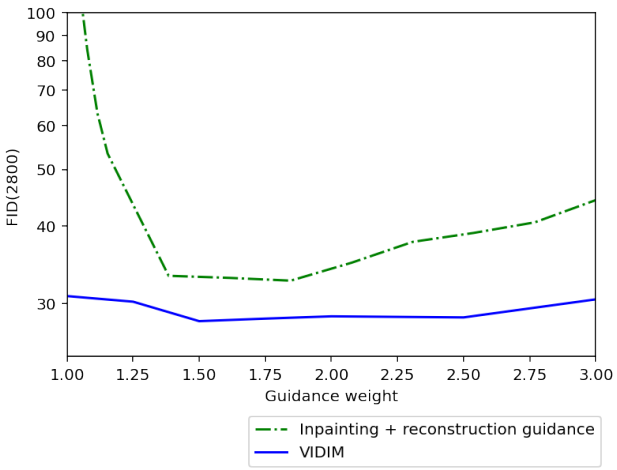


Figure 6. FID scores comparison between VIDIM and an in-painting baseline model at difference guidance and reconstruction guidance weights, respectively. Note that the *reconstruction* guidance weights (x-axis) for the baseline are re-scaled via $f(w) = (w-1)/13 + 1$ to more easily compare scores at the optimal region to VIDIM; the true range for the baseline guidance weights is from 1 to 27.

6. Supplementary website and more samples

Please refer to our supplementary website https://vidim-interpolation.github.io/ for video outputs from VIDIM along with a comparison to the baseline methods studied in this work. We also provide downloadable zip files to the Davis-7 and UCF101-7 datasets we used for benchmarking as described in Sec. 4.2.

We present some more qualitative comparisons against the baselines in Fig. 8. As discussed in Sec. 4.2 these results demonstrate our method's ability to handle large, ambiguous motion between the start and end frames. The baselines, in comparison, tend to generate blurry and unnatural frames. We strongly encourage the reader to visit our Supplementary Website to better appreciate the temporal dynamics present in the generated videos.

7. Additional ablation studies

In Sec. 4.4 we study the importance of explicitly training the super-resolution diffusion model to be conditional on the start and end frames. For the sake of completeness we also demonstrate the impact of this frame conditioning on the base diffusion model. Similar to Sec. 4.4 we create a strong baseline by evaluating inpainting with reconstruction guidance as proposed in [18]. We compare against the "medium" VIDIM model with 441M parameters. As shown in Fig. 6, peak FID scores with reconstruction guidance are still worse than VIDIM despite both models being trained with identical parameter count, data and hyperparameters.

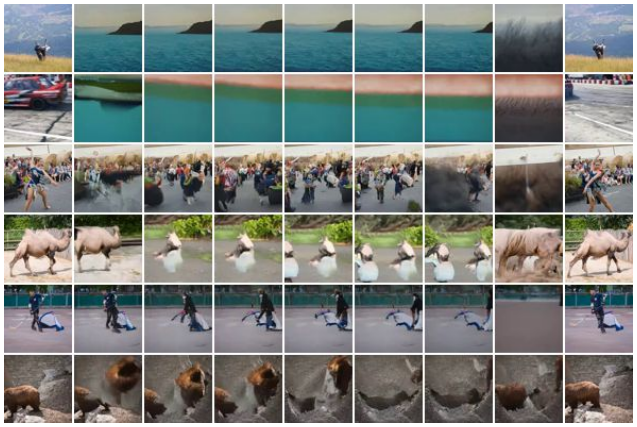
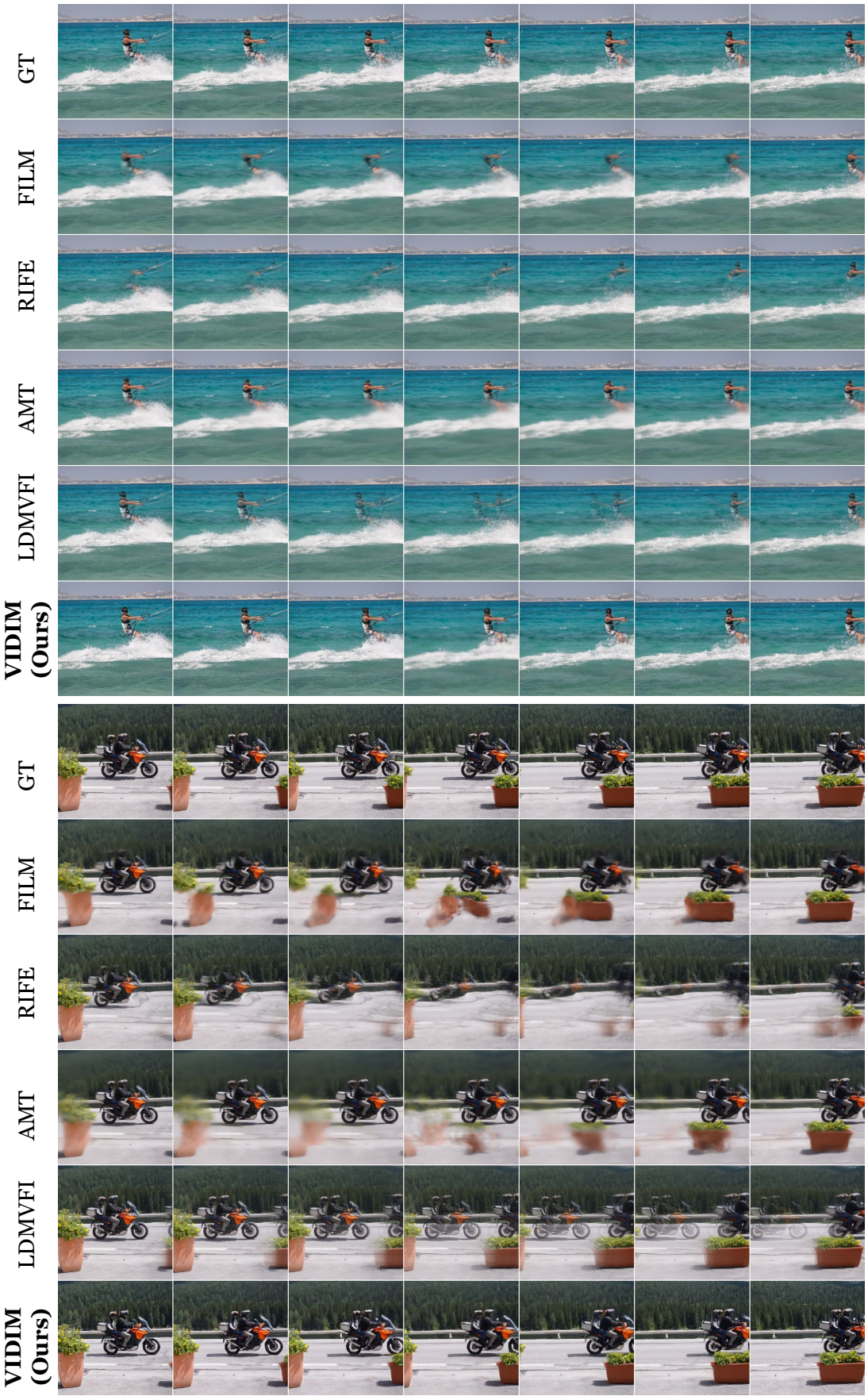


Figure 7. Frames generated via in-painting but with reconstruction guidance weight set to 1.0. The generated frames are not temporally consistent in the absence of reconstruction guidance. We roll out 9 frames (left to right) for six different videos (top to bottom).

In Fig. 7 we show the frames generated when reconstruction guidance weight is set to 1 in the inpainting + reconstruction baseline as described in the preceding paragraph. We found that without enough reconstruction guidance, standard in-painting cannot even produce consistent frames.


GT
FILM
RIFE
AMT
LDMVFI
VIDIM (Ours)
GT
FILM
RIFE
AMT
LDMVFI
VIDIM (Ours)
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5
Frame 6
Frame 7

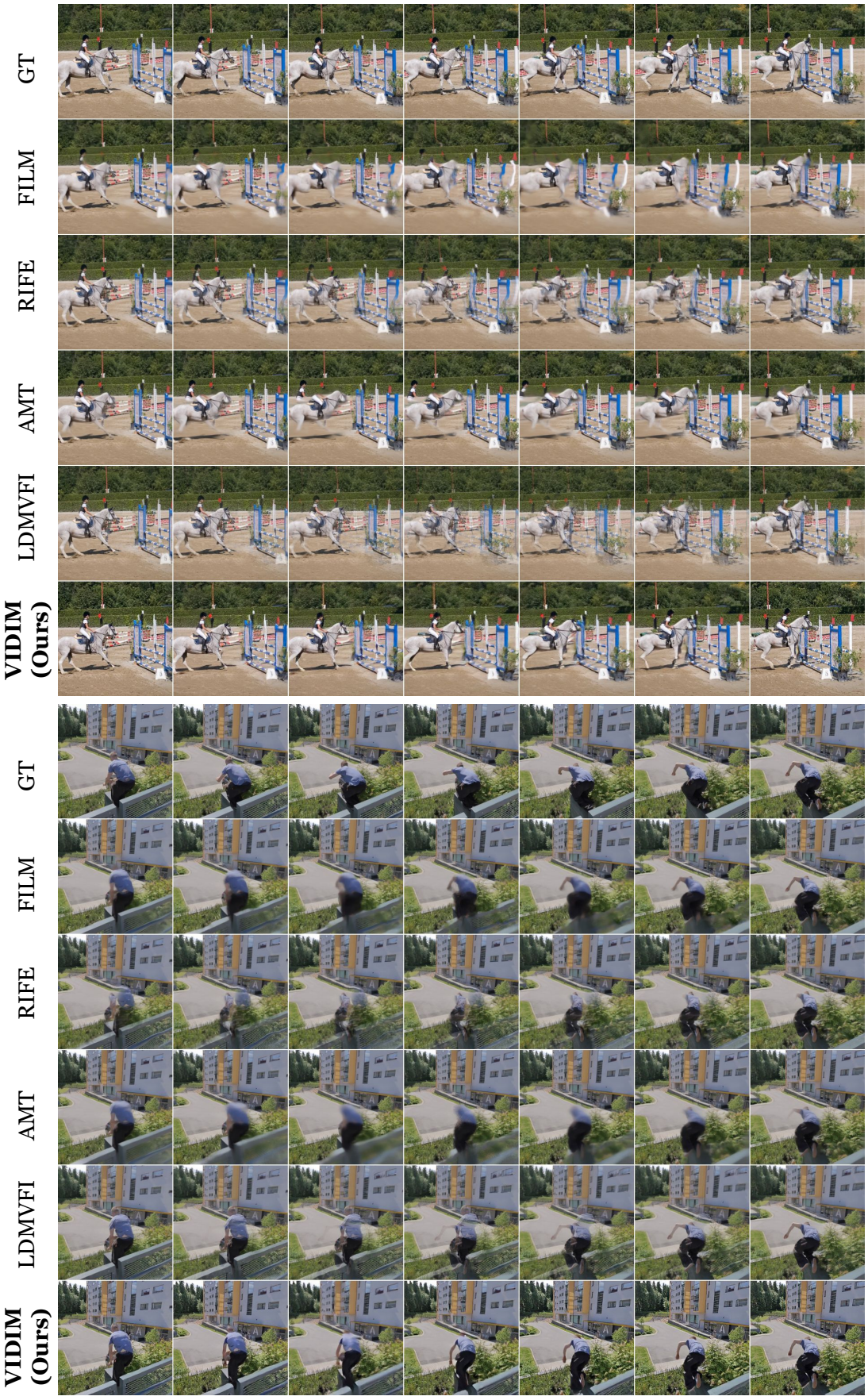


Figure 8. Examples from the DAVIS-9 dataset showing the results from VIDIM compared to the baseline methods.