VideoVAE+: Large Motion Video Autoencoding with Cross-modal Video VAE – Supplementary Materials –

Yazhou Xing* Yang Fei* Yingqing He*† Jingye Chen Jiaxin Xie
Xiaowei Chi Qifeng Chen†
The Hong Kong University of Science and Technology

A. Effectiveness of our image-video joint training

Our models support image-video joint training. We show that with join training, the fine-grained details of the input video can be better reconstructed, such as the definition of the small text region.

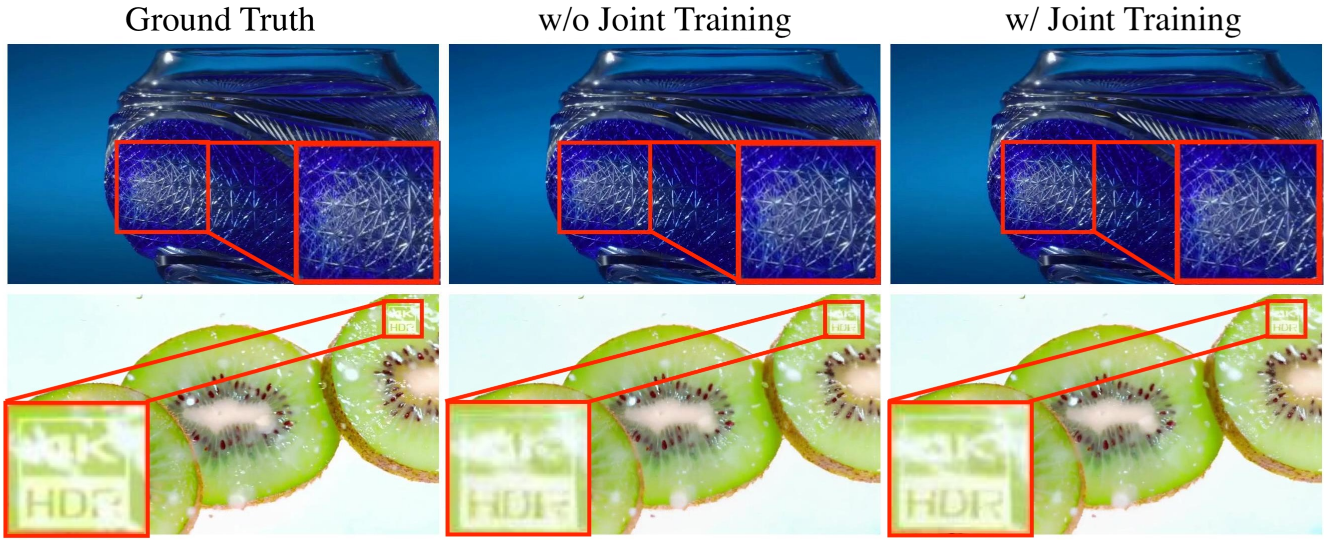


Figure 1. Visual ablation on the effectiveness of image-video joint training.

We show the image reconstruction ablation results of joint training in the Table 1. The image reconstruction comparison is conducted on a set of 500 images with a resolution of 480x864, randomly sampled from a UHD-4K video dataset. For the image reconstruction, our 4-channel latent Video VAE slightly outperforms SD1.4, and also improves on SSIM and LPIPS, indicating better perceptual quality. For the 16-channel VAE, while our model achieves competitive results in terms of PSNR, it falls slightly short of SD3.5. However, our model still demonstrates strong performance in terms of SSIM and LPIPS, suggesting that our joint training approach maintains high perceptual quality despite the slight drop in PSNR.

B. Ablation on the settings of the temporal-aware spatial autoencoder

As shown in Table 2, we deeply investigate the impact of the kernel size in the temporal convolutional layer of temporal-aware spatial autoencoder, as well as the GAN losses. The results are evaluated on the randomly select 98 videos from the MMTrailer dataset for this ablation study.

*equal contribution
†corresponding authors

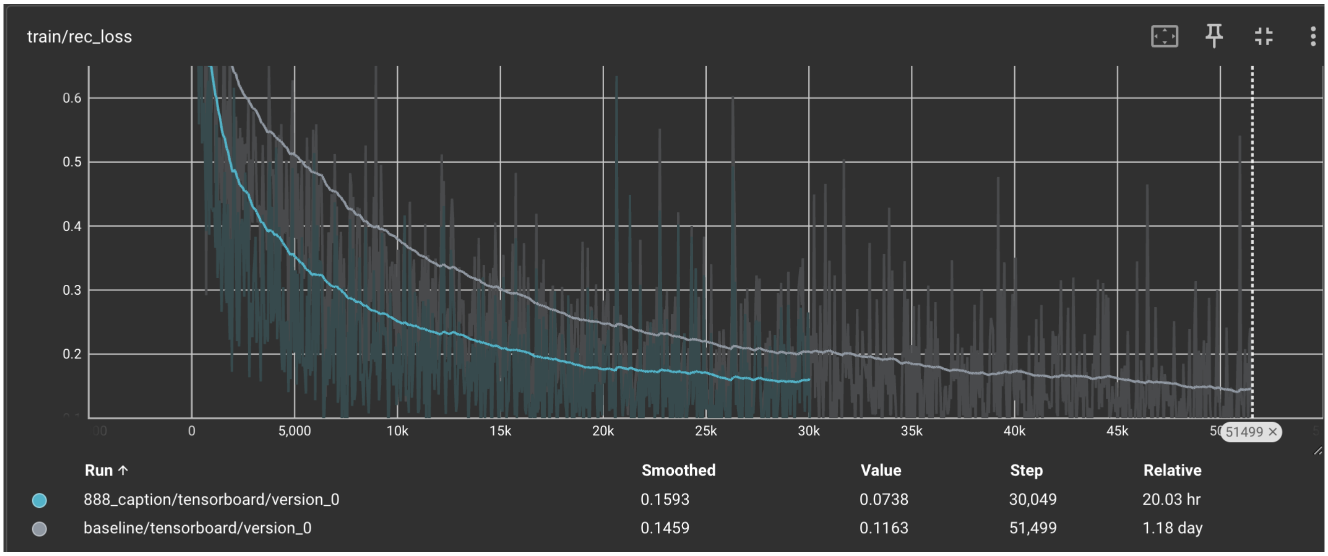
Model	**#ch**	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**
SD1.4	4	30.2199	0.8974	0.0440
Ours w/o JT*	4	15.1001	0.5561	0.4339
Ours	4	**30.8650**	**0.9042**	**0.0397**
SD3.5	16	**36.5208**	**0.9646**	**0.0116**
Ours w/o JT*	16	9.2603	0.2770	0.6802
Ours	16	35.3437	0.9590	0.0167

Table 1. JT* means joint training. We evaluate **image** reconstruction performance w/ or w/o our joint image-video training strategy.

Model / Kernel Size	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**
Image GAN Loss	31.9133	0.9071	0.0436
Video GAN Loss	**32.0262**	**0.9089**	**0.0426**
TemporalConv(3, 1, 1)	30.3332	0.8898	0.0489
TemporalConv(5, 1, 1)	30.8745	0.9004	0.0475
TemporalConv(7, 1, 1)	31.2922	0.9025	0.0458
TemporalConv(5, 3, 3)	31.3516	0.9011	0.0437
TemporalConv(7, 3, 3)	**31.7444**	**0.9074**	**0.0436**

Table 2. Ablation study of the temporal-aware spatial autoencoder with image or video GAN loss, and different kernel sizes.

C. Analysis on textual guidance



	Run ↑	Smoothed	Value	Step	Relative
●	888_caption/tensorboard/version_0	0.1593	0.0738	30,049	20.03 hr
●	baseline/tensorboard/version_0	0.1459	0.1163	51,499	1.18 day

Figure 2. Analysis about the convergence speed on textual guidance.

D. More results on different spatiotemporal modelling strategies.

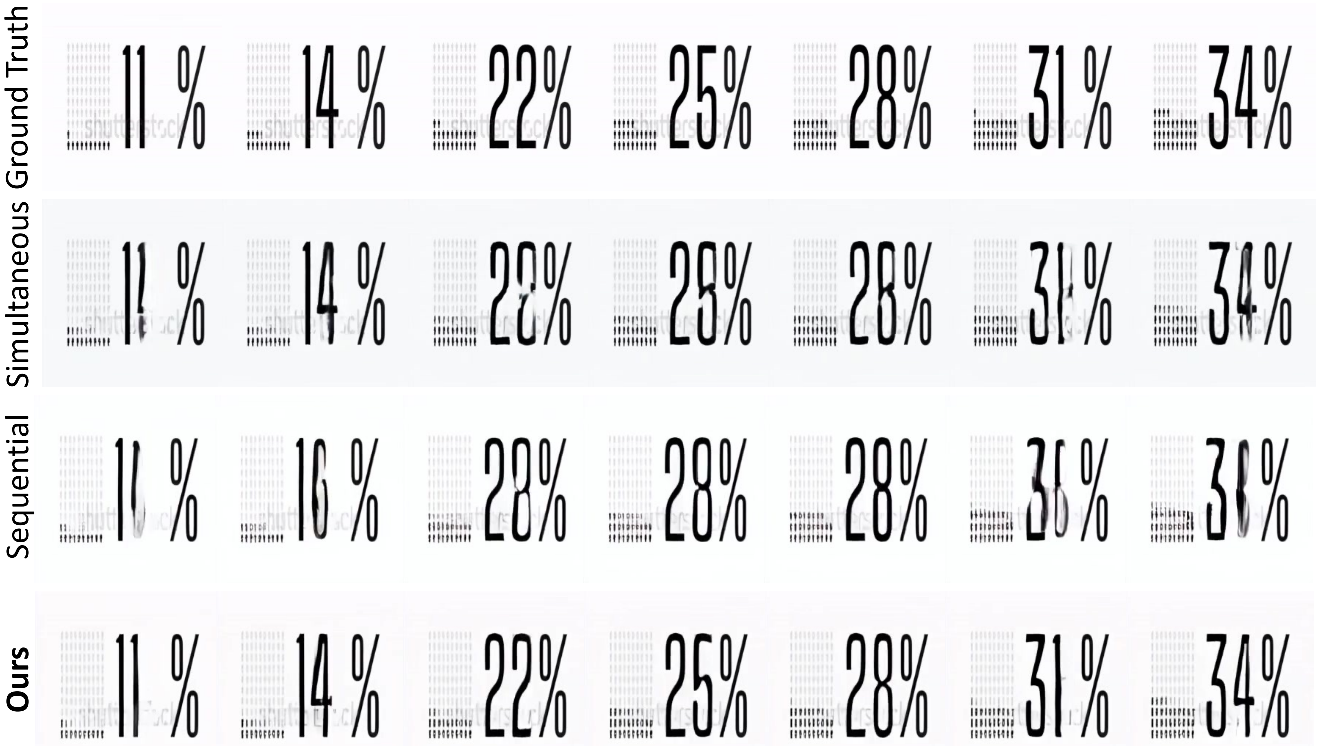


Figure 3. Comparisons among simultaneous spatiotemporal modeling, sequential spatiotemporal modeling and our proposed solution.