1. Experiment Details

In this section, we discuss the additional details of the datasets, video generation models, and other experiment setups. We will release our code to compute FVD with VideoMAE-v2 backbone features and pre-computed features for commonly used video datasets.

Dataset. We conduct our analysis on six datasets, including two widely used video understanding benchmarks Kinetics-400 [2] and Something-Something-v2 [5], three video generation benchmarks FaceForencis [8], Sky Time-lapse [15], and Taichi-HD [9], and the UCF-101 dataset [11] that has been used for both tasks.

Kinetics-400 [2] (K400) contains $267,000$ videos of 10 seconds in 400 action classes. Something-Something-v2 [5] (SSv2) consists of $220,000$ videos of $2-6$ seconds in 174 classes of humans performing basic actions with everyday objects. UCF-101 [11] has $13,320$ videos of, on average, 7 seconds in 101 classes of human actions. Sky Time-lapse [15] (Sky) collects $2,647$ time-lapse videos of the sky in different periods and under various weather conditions. FaceForencis [8] (FFS) contains $1,000$ human talking videos collected from YouTube to facilitate Deepfake detection. We follow the official instructions to process the videos to extract the face region and obtain 704 videos. Taichi-HD [9] (Taichi) is a video dataset of 280 long YouTube videos recording a person performing Taichi, which is pre-processed into $3,335$ short clips. Note that video generation models [4, 17] trained on this dataset often sample every four frames to attain larger motion in each training clip.

Video Generation Models. Our generated videos are from four video generation models, DIGAN [17], TATS [4], StyleGAN-v [10], and PVDM [18]. DIGAN [17] is a GAN-based model that leverages implicit neural representations and computation-efficient discriminators. TATS [4] extends VQGAN [3] for long video generation by designing time-agnostic VAE and hierarchical transformer. DIGAN and TATS-base models are trained on 16 video frames of 128×128 resolution. StyleGAN-v [10] extends the renowned StyleGAN architecture [7] for video generation by employing implicit neural representations. PVDM [18] exploits a latent diffusion architecture and efficient triplane representation. StyleGAN-v and PVDM are evaluated with resolution 256×256 and video length 16 and 128. When computing FVD scores, all four methods generate $2,048$ videos. We follow StyleGAN-v [10] to save the generated videos without severe JPEG compression and sample random clips from the real videos. We can reproduce the reported FVD scores, as shown in Table. 2 in the main paper.

Additional Implementation Deatils. To quantify the FVD temporal sensitivity with video distortion methods, we follow the common practice [12, 16, 17] to sample $2,048$ clips of resolution 128×128 from each video dataset. We apply the distortion in five pre-defined corruption levels following the previous study [6]. To probe the perceptual null space in FVD, we cast the extracted features and weights to float64 to stabilize the optimization process and avoid numerical issues.

In addition, to compute ViT encoder features of VideoMAE [13, 14] and TimeSFormer [1] models, we follow the convention to exploit the pre-logit features. To extract features from the pre-trained VideoMAE encoder-decoder architecture, we take the output of the penultimate layer in the encoder and average across all the patches, which uses essentially the output from the same layer as the fine-tuned VideoMAE model. To reduce memory costs when computing FVD_{128} using VideoMAE-v2 models, we cast all the features to float16, as the FVD score difference between using float16 and float32 is neglectable and often less than 0.03%.

All of our experiments are performed on a single NVIDIA RTX A6000 GPU except for reproducing the StyleGAN-v variants, where we follow the official receipt to train on four NVIDIA RTX A6000 GPUs.

2. Addition Results

Quantifying the temporal sensitivity of FVD. We expand Table 1 in the main paper to include the FVD scores on the six datasets with either spatial (S) or spatiotemporal (ST) distortion using features from the I3D model, three VideoMAE-v2 variants, two TimeSformer models, and VideoMAE-v2 models in Tables 1 and 2. By inspecting the spatial FVDs computed with VideoMAE-v2 features on different datasets, we notice that they vary less than the FVD scores using the I3D features, highlighting their generalization capacity. We also explore the TimeSformer model trained on the SSv2 dataset. Compared with the one trained on the K400 dataset reported in the main paper, it is generally more sensitive to temporal quality change due to the dataset. However, it is still on par with the I3D model as both share the same supervised objective.

Probing the perceptual null space in FVD. We expand Table 7 in the main paper to include FVD and FVD^* on all the models and dataset computed with the I3D model and three VideoMAE-v2 variants in Table 3.

Practical examples. We expand Table 4 in the main paper by showing the FVD changes on all the consecutive 16 frames of the extrapolated generation results using DIGAN in Figure 1. We notice that with longer frames being

Table 1. **Results of analyzing the temporal sensitivity of FVD.** We report FVDs of synthetic videos created from real videos using spatial only or spatiotemporal distortions, where the two sets produce similar frame quality as assessed by FID and only differ in temporal quality. This table includes the results of the I3D model and three VideoMAE-v2 variants.

Dataset	Distortion	Type	FID	FVD_{I3D}	$FVD_{VideoMAE\text{-}v2\text{-}K710}$	$FVD_{VideoMAE\text{-}v2\text{-}SSv2}$	$FVD_{VideoMAE\text{-}v2\text{-}PT}$
UCF-101	Motion Blur	S	133.15	1460.18	121.37	277.10	18.33
		ST	133.69(+0.4%)	1705.27(+16.8%)	147.91(+21.9%)	868.31(+213.4%)	39.21(+113.8%)
	Elastic	S	175.47	979.48	167.21	221.83	7.95
		ST	176.46(+0.6%)	1694.95(+73.0%)	321.96(+92.5%)	1186.91(+435.0%)	58.89(+640.6%)
Sky	Motion Blur	S	79.11	211.08	88.80	127.99	14.22
		ST	79.35(+0.3%)	286.39(+35.7%)	252.01(+183.8%)	733.41(+473.0%)	35.73(+151.2%)
	Elastic	S	72.32	149.23	105.04	142.49	6.97
		ST	72.52(+0.3%)	333.48(+123.5%)	438.19(+317.2%)	1056.40(+641.4%)	60.60(+769.0%)
FFS	Motion Blur	S	80.42	354.49	95.73	199.90	13.61
		ST	79.57(−1.1%)	367.35(+3.6%)	178.96(+87.0%)	717.08(+258.7%)	23.75(+74.4%)
	Elastic	S	161.55	589.07	192.01	164.82	11.14
		ST	161.30(−0.2%)	891.50(+51.3%)	442.62(+130.5%)	969.28(+488.1%)	54.42(+388.4%)
TaiChi	Motion Blur	S	169.76	1016.78	100.83	382.37	25.22
		ST	170.10(+0.2%)	1201.35(+18.2%)	177.51(+76.0%)	1217.34(+218.4%)	47.73(+89.3%)
	Elastic	S	182.99	688.55	100.93	161.51	5.81
		ST	183.21(+0.1%)	1252.72(+81.9%)	372.14(+268.7%)	1467.06(+808.3%)	66.34(+1042.6%)
SSv2	Motion Blur	S	100.65	594.68	89.31	144.95	16.96
		ST	100.62(−0.0%)	678.08(+14.0%)	135.98(+52.3%)	502.09(+246.4%)	29.93(+76.5%)
	Elastic	S	143.16	622.87	216.12	211.98	9.74
		ST	143.91(+0.5%)	980.44(+57.4%)	351.48(+62.6%)	746.91(+252.4%)	48.07(+393.7%)
K400	Motion Blur	S	112.22	996.71	92.11	257.01	17.67
		ST	112.85(+0.6%)	1155.53(+15.9%)	126.96(+37.8%)	785.58(+205.7%)	34.34(+94.3%)
	Elastic	S	146.70	675.53	151.50	241.15	8.61
		ST	146.68(−0.0%)	1189.37(+76.1%)	300.02(+98.0%)	1087.20(+350.8%)	55.01(+539.0%)

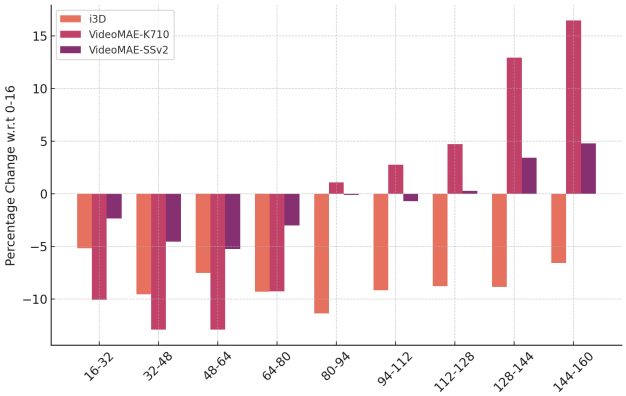

Figure 1. DIGAN [17] trained on the Sky Time-lapse dataset generates periodic artifacts when using extrapolated time steps. We show the percentage change of FVD computed on every 16 frames compared with the first 16 frames.

generated, the motion artifacts become more pronounced. As a consequence, FVD scores computed with VideoMAE features properly capture the reduced temporal quality by providing a larger value. However, FVD scores computed with the I3D backbone are consistently less than from frames 0-16.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 1

[4] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022. 1

[5] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video

Table 2. **Results of analyzing the temporal sensitivity of FVD.** We report FVDs of synthetic videos created from real videos using spatial only or spatiotemporal distortions, where the two sets produce similar frame quality as assessed by FID and only differ in temporal quality. This table includes the results of the I3D model, two TimeSformer variants, and VideoMAE-v1 model.

Dataset	Distortion	Type	FVD_{I3D}	$\text{FVD}_{\text{TimeSfomer-k400}}$	$\text{FVD}_{\text{TimeSfomer-SSv2}}$	$\text{FVD}_{\text{VideoMAE-v1-k400}}$
UCF-101	Motion Blur	Spatial	1460.18	265.77	311.85	26.44
		Spatiotemporal	1705.27(+16.8%)	275.09(+3.5%)	336.51(+7.9%)	46.58(+76.2%)
	Elastic	Spatial	979.48	260.65	261.27	31.82
		Spatiotemporal	1694.95(+73.0%)	313.36(+20.2%)	398.29(+52.4%)	79.85(+150.9%)
Sky	Motion Blur	Spatial	211.08	154.19	133.98	19.39
		Spatiotemporal	286.39(+35.7%)	169.46(+9.9%)	147.69(+10.2%)	62.33(+221.4%)
	Elastic	Spatial	149.23	123.58	137.43	23.47
		Spatiotemporal	333.48(+123.5%)	186.33(+50.8%)	249.93(+81.9%)	99.02(+321.8%)
FFS	Motion Blur	Spatial	354.49	240.65	327.82	21.56
		Spatiotemporal	367.35(+3.6%)	241.97(+0.5%)	311.36(−5.0%)	37.32(+73.1%)
	Elastic	Spatial	589.07	314.96	390.37	32.61
		Spatiotemporal	891.50(+51.3%)	392.19(+24.5%)	472.37(+21.0%)	102.92(+215.6%)
TaiChi	Motion Blur	Spatial	1016.78	342.29	437.08	26.44
		Spatiotemporal	1201.35(+18.2%)	373.35(+9.1%)	499.24(+14.2%)	56.77(+114.7%)
	Elastic	Spatial	688.55	278.37	276.86	20.88
		Spatiotemporal	1252.72(+81.9%)	365.80(+31.4%)	465.57(+68.2%)	105.27(+404.2%)
SSv2	Motion Blur	Spatial	594.68	166.16	167.52	21.23
		Spatiotemporal	678.08(+14.0%)	169.68(+2.1%)	184.42(+10.1%)	32.65(+53.8%)
	Elastic	Spatial	622.87	265.63	186.04	38.13
		Spatiotemporal	980.44(+57.4%)	296.53(+11.6%)	245.35(+31.9%)	84.95(+122.8%)
K400	Motion Blur	Spatial	996.71	203.54	237.63	18.73
		Spatiotemporal	1155.53(+15.9%)	211.09(+3.7%)	254.55(+7.1%)	33.01(+76.2%)
	Elastic	Spatial	675.53	214.95	206.19	25.40
		Spatiotemporal	1189.37(+76.1%)	251.35(+16.9%)	297.65(+44.4%)	65.24(+156.9%)

database for learning and evaluating visual common sense. In *ICCV*, 2017. 1

[6] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 1

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[8] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 1

[9] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 1

[10] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 1

[11] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[12] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *ICLR*, 2021. 1

[13] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 35:10078–10093, 2022. 1

[14] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, pages 14549–14560, 2023. 1

[15] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *CVPR*, 2018. 1

[16] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 1

[17] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *ICLR*, 2022. 1, 2

[18] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, 2023. 1

Table 3. **Results of probing the perceptual null space of FVD.** We report FVDs of normal and frozen generated videos by random sampling (FVD) and sampling to match all the fringe features (FVD*). We color the FVD difference for better visualization: $< 20\%$, $20\% - 40\%$ and $> 60\%$. The drop of FVD on the frozen generated videos indicates the volume of the null space where FVD can be reduced without generating a meaningful motion. I3D has the largest perceptual null space.

Feature Extractor		I3D		VideoMAE-v2-K710		VideoMAE-v2-SSv2		VideoMAE-v2-PT	
Model	Dataset	FVD	FVD*	FVD	FVD*	FVD	FVD*	FVD	FVD*
Normal Generated Videos vs. Real Videos									
DIGAN	UCF-101	562.36	220.89(−60.7%)	358.80	160.13(−55.4%)	378.19	260.77(−31.0%)	2.77	2.67(−3.9%)
DIGAN	Sky	157.13	54.39(−65.4%)	86.58	61.93(−28.5%)	174.79	128.00(−26.8%)	4.72	3.71(−21.5%)
DIGAN	Taichi	132.26	65.72(−50.3%)	58.72	24.45(−58.4%)	313.84	194.17(−38.1%)	4.00	3.66(−8.5%)
TATS	UCF-101	329.92	120.58(−63.5%)	176.98	72.95(−58.8%)	388.79	226.39(−41.8%)	7.92	7.12(−10.1%)
TATS	Sky	125.62	38.42(−69.4%)	100.27	59.83(−40.3%)	213.33	105.69(−50.5%)	18.11	7.87(−56.5%)
TATS	Taichi	124.16	64.17(−48.3%)	37.16	26.08(−29.8%)	274.81	126.53(−54.0%)	5.88	5.34(−9.2%)
StyleGAN-V	Sky	56.63	31.73(−44.0%)	180.97	55.54(−69.3%)	219.85	148.11(−32.6%)	10.04	8.78(−12.5%)
StyleGAN-V	FFS	56.22	25.87(−54.0%)	77.28	61.02(−21.0%)	194.68	135.30(−30.5%)	1.08	1.04(−3.7%)
PVDM	UCF-101	348.81	113.99(−67.3%)	116.01	90.40(−22.1%)	369.14	172.35(−53.3%)	4.51	3.69(−18.2%)
PVDM	Sky	59.95	22.94(−61.7%)	141.48	75.12(−46.9%)	142.50	57.04(−60.0%)	3.63	2.33(−35.7%)
Frozen Generated Videos vs. Real Videos									
DIGAN	UCF-101	1303.13	715.96(−45.1%)	357.61	175.13(−51.0%)	951.59	859.57(−9.7%)	12.61	12.23(−3.1%)
DIGAN	Sky	230.64	115.55(−49.9%)	175.47	142.86(−18.6%)	408.17	362.84(−11.1%)	13.23	12.16(−8.1%)
DIGAN	Taichi	461.79	276.88(−40.0%)	132.96	52.00(−60.9%)	578.61	523.20(−9.6%)	4.40	4.18(−4.9%)
TATS	UCF-101	1157.69	616.25(−46.8%)	247.80	107.41(−56.7%)	908.95	805.88(−11.3%)	14.66	13.68(−6.7%)
TATS	Sky	279.75	126.32(−54.8%)	172.00	140.37(−18.4%)	375.74	353.15(−6.0%)	21.28	15.76(−25.9%)
TATS	Taichi	475.99	312.19(−34.4%)	164.58	64.69(−60.7%)	587.31	530.86(−9.6%)	4.63	4.42(−4.4%)
StyleGAN-V	Sky	206.56	104.27(−49.5%)	224.80	91.71(−59.2%)	503.22	456.24(−9.3%)	23.17	21.60(−6.8%)
StyleGAN-V	FFS	353.79	242.04(−31.6%)	171.38	147.76(−13.8%)	547.24	520.98(−4.8%)	14.08	14.22(+0.9%)
PVDM	UCF-101	1135.61	605.09(−46.7%)	250.52	211.34(−15.6%)	1032.90	898.48(−13.0%)	12.95	12.34(−4.7%)
PVDM	Sky	182.77	94.87(−48.1%)	198.50	140.77(−29.1%)	429.06	395.79(−7.8%)	11.54	11.03(−4.4%)