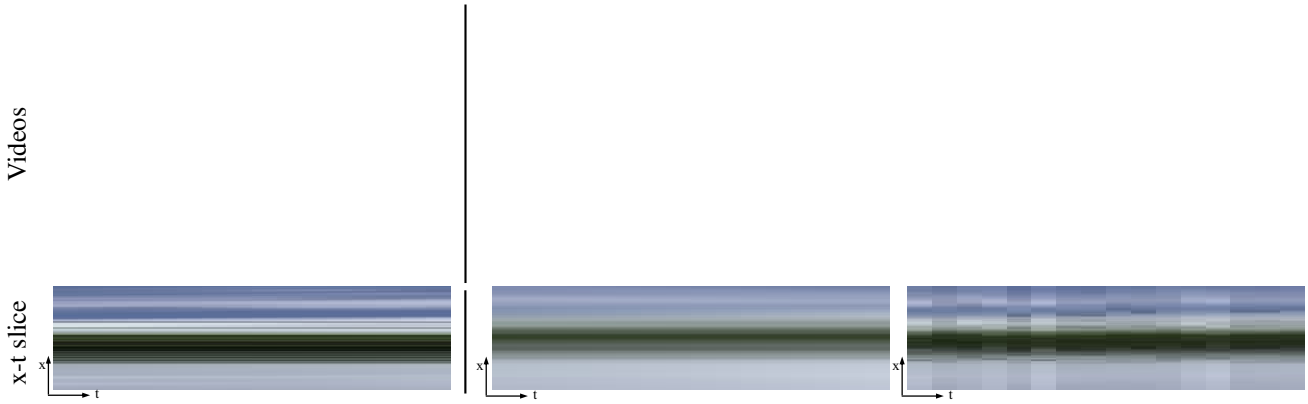
On the Content Bias in Fréchet Video Distance

Songwei Ge[1] Aniruddha Mahapatra[2,3] Gaurav Parmar[2]
Jun-Yan Zhu[2] Jia-Bin Huang[1]
[1]University of Maryland, College Park [2]Carnegie Mellon University [3]Adobe Research
https://content-debiased-fvd.github.io/



(a) Reference Videos

(b) Medium Spatial & No Temporal Corruption
FVD=317.10

(c) Small Spatial & Severe Temporal Corruption
FVD=310.52

Figure 1. **FVD is biased towards per-frame quality than temporal consistency.** FVD [72], a commonly used video generation evaluation metric, should ideally capture *both* spatial and temporal aspects. However, our experiments reveal a strong bias toward individual frame quality. (b) First, we apply mild spatial distortions through local warping, which results in an FVD score of 317.10. (c) Next, we induce slightly less spatial corruptions but *severe* temporal inconsistencies by altering each frame differently. These changes create artifacts that are noticeable to humans and evident in the spatiotemporal x-t slice, as seen in the bottom row, but surprisingly lead to a lower FVD score of 310.52. This discrepancy highlights the metric's bias towards individual frame quality. *We encourage readers to view the videos with Acrobat Reader or visit our website to observe the inconsistencies.*

Abstract

Fréchet Video Distance (FVD), a prominent metric for evaluating video generation models, is known to conflict with human perception occasionally. In this paper, we aim to explore the extent of FVD's bias toward per-frame quality over temporal realism and identify its sources. We first quantify the FVD's sensitivity to the temporal axis by decoupling the frame and motion quality and find that the FVD increases only slightly with large temporal corruption. We then analyze the generated videos and show that via careful sampling from a large set of generated videos that do not contain motions, one can drastically decrease FVD without improving the temporal quality. Both studies suggest FVD's bias towards the quality of individual frames. We further observe that the bias can be attributed to the features extracted from a supervised video classifier trained on the content-biased dataset. We show that FVD with features extracted from the recent large-scale self-supervised video models is less biased toward image quality. Finally, we revisit a few real-world examples to validate our hypothesis.

1. Introduction

Video generation [5, 9, 12, 21, 22, 30, 62] has recently accomplished unprecedented advances driven by the scalable models [29, 65] and growing training data [4, 57]. With rapid progress, it is increasingly crucial to evaluate the model performance accurately. Despite the fruitful literature on assessing video quality [40, 58, 63, 81] and designing image generation evaluation metrics [28, 38, 47, 48, 55], automatically evaluating the quality and diversity of the generated videos, has received less attention [33, 44]. In this paper, we focus on analyzing the bias of Fréchet Video Distance

(FVD) [72], one of the most frequently used metrics for video evaluation.

FVD extends the image generation metric Fréchet Inception Distance (FID) [28] to measure the quality and diversity of generated videos with respect to the training set. Given N features $\mathbf{f}_i$, which are column vectors, extracted from a pre-trained video network, for a set of generated and real videos, we fit a multivariate Gaussian with the mean $\mu = \frac{1}{N}\sum_i \mathbf{f}_i$, and covariance $\mathbf{\Sigma} = \frac{1}{N}\sum_i (\mathbf{f}_i - \mu)(\mathbf{f}_i - \mu)^T$. The performance of the video generator is then measured as the Fréchet distance [19] between the two Gaussian distributions:

$$\text{FVD} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\mathbf{\Sigma}_r + \mathbf{\Sigma}_g - 2\left(\mathbf{\Sigma}_r\mathbf{\Sigma}_g\right)^{\frac{1}{2}}\right), \quad (1)$$

where $(\mu_r, \mathbf{\Sigma}_r)$ and $(\mu_g, \mathbf{\Sigma}_g)$ denote the mean and covariance for real and generated data.

Earlier studies have confirmed that FVD reliably reflects the model performance in various cases, such as training convergence [92], hyperparameter tuning [11, 30], and architecture design [31, 64]. However, several recent studies have reported cases where FVD scores contradict human judgment [11, 20, 64]. A recurrent argument is that FVD tends to value the image quality of individual frames more than the realism of motion. We refer to such bias as the *content bias*, which is inspired by the video generation works in decoupling content and motion [34, 69, 71, 73, 79].

As shown in Figure 1, we motivate our analysis with a simple, controlled setting, where the metric diverges from human perception when weighing spatial and temporal qualities. Specifically, given a set of videos (a), we create two sets of distortion. In (b), we locally warp the frames in each video uniformly, while in (c), we distort the frames differently but with slightly reduced severity. The latter creates additional temporal artifacts. The FVD metric, however, favors video set (c), while most humans would pick video set (b) to be more similar to the reference videos due to the significant temporal inconsistency presented in video set (c).

Building upon this simple example, we present the first systematic study to quantify the content bias and understand its impact using both synthetic and real-world settings. We first distort videos so that the frame quality deteriorates to the same level while the temporal consistency is either intact or, in the other case, significantly decreased. By comparing FVDs on these distorted videos, we can quantify the relative sensitivity of the FVD metric to the temporal consistency. Next, following the previous work on FID analysis [39], we probe the perceptual null space in the FVD metric. Without improving the temporal quality of the generated videos, we can still greatly reduce the FVD scores. Lastly, we revisit a few real-world examples where FVD presents a notable content bias.

Where does the content bias originate from? Previous studies show that the alignment of the FID metric to human perception depends on the choice of the extracted features [1, 10, 39, 46]. In practice, FVD employs an Inflated 3D ConvNet (I3D) model [14], originally trained for action recognition on the Kinetics-400 dataset [14]. The feature space is formed as the output of the logit layer. As a result, these features focus on extracting the semantic information about human actions in the videos.

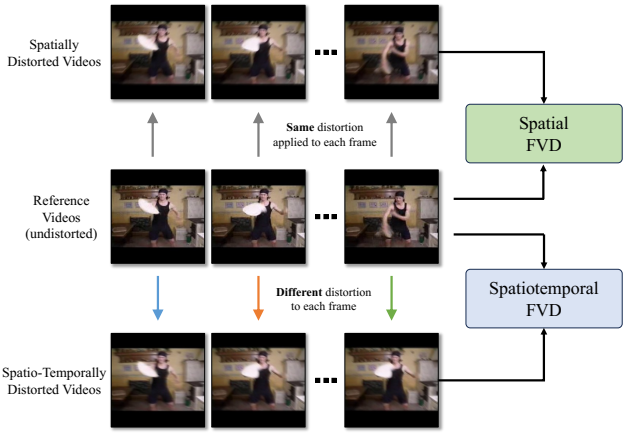


Figure 2. **Analyzing the FVD's sensitivity to temporal consistency.** We distort the same set of videos in spatial only or spatiotemporal manners so that the resulting videos have similar frame quality yet only differ in temporal quality. By comparing the FVD scores of the two distorted video sets, we aim at quantifying the temporal sensitivity of the metric.

Using I3D features thus raises several practical concerns that could undermine the metric's reliability. First, the Kinetics dataset [14] predominantly comprises videos with humans as the protagonists. However, video content diverging from typical Kinetics-400 categories, such as time-lapse landscape videos [86] or first-person riding and biking videos [11], may not produce a meaningful feature representation. Second, the models trained on the Kinetics dataset are shown biased to the appearance of objects and backgrounds instead of motions [16, 32, 42, 43, 59, 75]. For example, to recognize the action "playing saxophone", it is sufficient to detect the presence of a saxophone since this is the only category where a saxophone is presented. Therefore, the features may not capture the musician's motion. Previous works also show that descent classification accuracy can be achieved without modeling the temporal aspect [6, 14].

To verify our hypothesis, we compute FVD scores using features extracted from a self-supervised model [76] trained on diverse dataset and perform the same analysis. Overall, our experiments show that the FVD, computed with I3D features, is strongly biased to the content over the motion, while using features computed with a model trained in a self-supervised manner helps mitigate such bias to a large extent. Our evaluation code and data are available on https://content-debiased-fvd.github.io/.

2. Related Work

Video generation. Various types of generative models have been proposed for video generation such as GANs [54, 60, 69, 71, 73, 74, 79], Autoregressive models [3, 15, 18, 20, 37, 50, 67, 82, 83, 87–90], and implicit neural representations [64, 91]. Following the recent success of text-to-image Diffusion Models [49, 51, 53], several works aim to achieve high-quality results for the text-to-video task. These works leverage diffusion process either in the pixel space [21, 30, 62] or latent space [2, 8, 9, 24, 25, 36, 45, 77, 78, 80, 80, 84, 85, 95] or both [93]. Reliably evaluating the above models remains a challenge. Current works primarily rely on FVD [72] and human perceptual study. While a user study can reflect human preference more accurately, FVD serves as a more scalable evaluation protocol. In this paper, we aim to better understand what aspects the FVD metric values more. Specifically, we analyze its sensitivity to the spatial versus temporal quality.

Evaluation metrics for image and video generation. Many studies have focused on understanding and improving the evaluation metrics for image generation, such as Inception Score [56], FID [28, 39, 46, 48], Perceptual Path Length [35], and precision and recall [38, 55]. Among them, FID is the most commonly adopted one, using the Inception-V3 feature extractor [68] trained on the ImageNet dataset [17]. However, it can sometimes diverge from human judgment, especially on the out-of-domain datasets like human faces [46, 96].

To address the above issue, researchers have introduced several variants [7, 38, 47, 55] and performed analysis to understand FID [1, 10, 39, 48]. For instance, Kynkäänniemi et al. [39] study the role of training data classes in the FID metric and advocate the use of the CLIP model as the feature extractor instead. KID [7] is proposed to improve FID using the squared Maximum Mean Discrepancy (MMD) with a polynomial kernel. KID relaxes the Gaussian assumption in FID and requires fewer samples to compute. Clean-FID [48] shows that the aliasing issue caused by the preprocessing steps could significantly affect the FID scores. Similarly, Skorokhodov et al. [64] studies the "low-level" preprocessing operations in FVD, such as resizing and frame sampling strategies. However, the analysis and improvement of the FVD are much less explored than those of FID.

3. Quantifying the Temporal Sensitivity of FVD

We examine the significance of temporal quality and consistency in FVD calculation. Recent studies suggest that models trained on the Kinetics datasets may not fully leverage the motion information [6, 32, 43, 59], raising a similar question about whether the I3D features in FVD truly capture the motion quality of videos. One way to understand

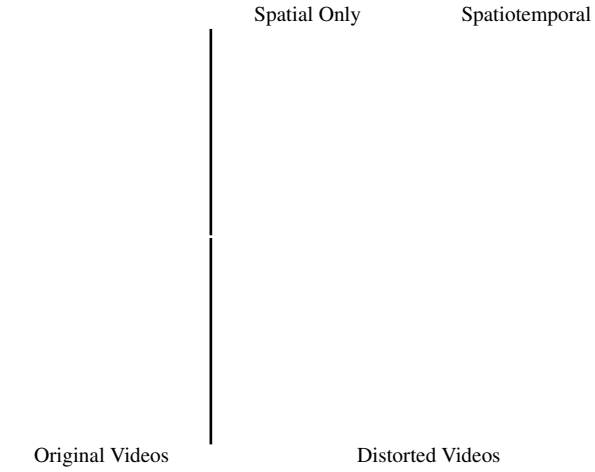


Figure 3. **Visualization of the spatial and spatiotemporal corruptions.** Both corruptions yield similar frame quality, while the spatiotemporal corruption induces additional temporal inconsistency in the video. By comparing the FVD of the spatiotemporal corruption with the spatial corruption, we analyze the temporal sensitivity of the metric. *Best viewed with Acrobat Reader. Please check our website for videos.*

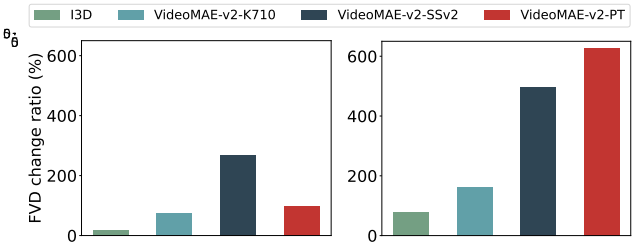


Figure 4. **FVD sensitivity with different video feature extractors.** We show that by substituting the I3D features with ones computed from the VideoMAE-v2 model, the temporal sensitivity can be significantly improved for both motion blur and elastic transformation distortions.

video motion *vs*. content is to undermine one aspect through either spatial or temporal distortion. However, fully decoupling the two aspects is non-trivial [32]. For example, poor frame quality would hinder the creation of a natural motion. As a result, previous approaches suggest creating videos by stitching real frames together, albeit in an incorrect order or from different videos [32, 59, 72].

While this method is useful for analyzing video datasets, it may not be ideal for understanding a video generation metric. This is because generated videos rarely contain frames from irrelevant videos or arrange frames in incorrect order. In contrast, we carefully design distortions that simulate real scenarios to quantify FVD's sensitivity to spatial and temporal video quality.

Video distortion methods. We illustrate our method for

Table 1. **Analyzing FVD temporal sensitivity with video distortions.** We apply spatial only or spatiotemporal distortions to the videos. The two video sets share similar frame quality as assessed by FID. We thus use the FVD ratio to measure the temporal sensitivity of FVD.

Metric	Distortion	UCF-101	Sky Time-lapse	FaceForensics	Taichi-HD	SSv2	Kinectics-400
FID	Spatial	133.15	79.11	80.42	169.76	100.65	112.22
	Spatiotemporal	133.69$_{(+0.4\%)}$	79.35$_{(+0.3\%)}$	79.57$_{(-1.1\%)}$	170.10$_{(+0.2\%)}$	100.62$_{(-0.0\%)}$	112.85$_{(+0.6\%)}$
FVD	Spatial	1460.18	211.08	354.49	1016.78	594.68	996.71
	Spatiotemporal	1705.27$_{(+16.8\%)}$	286.39$_{(+35.7\%)}$	367.35$_{(+3.6\%)}$	1201.35$_{(+18.2\%)}$	678.08$_{(+14.0\%)}$	1155.53$_{(+15.9\%)}$

adding distortions in Figure 2. We apply two relevant distortions to the same set of real videos, aiming to synthesize videos with similar frame quality degradation but large differences in temporal quality. We employ conventional image distortion methods [27, 94] including elastic transformation and motion blur. The elastic transformation locally stretches and contrasts the frames, while the motion blur averages image pixels along a specific direction of motion. We apply the same distortions to each frame to achieve a consistent frame quality drop.

To create spatiotemporal corruptions, as shown in Figure 3, we apply randomly sampled elastic transformation parameters or blur kernels for each frame. This procedure allows us to introduce temporal inconsistency while producing similar frame quality to spatial corruptions.

Experimental setups. After distorting the videos using spatial only and spatiotemporal methods, we compute the FVD score of each video set with respect to the original videos. We apply distortion in five predefined corruption levels [27] and compute the average. To verify that the two distorted sets have similar frame quality, we compute FID [48] on the frames extracted from each set against the original image frames. Finally, we use the relative ratio between changes in FVD and FID to measure temporal sensitivity.

We perform the experiments on several standard video datasets, including Kinetics-400 [14], Something-Something-v2 [23], UCF-101 [66], Sky Time-lapse [86], and FaceForensics [52] datasets. Motivated by the previous finding that the unsupervised models trained on large datasets often produce more reliable features in FID [39, 46], we also compute FVD using a self-supervised video model VideoMAE-v2 [76], which is trained on a mixed set of unlabeled datasets with the Masked Autoencoders (MAE) reconstruction objective [26]. Due to the large gap between the pertaining and downstream tasks, the MAE models are often further fine-tuned on the downstream tasks.

In our experiments, we explore a pretrained model *VideoMAE-v2-PT* and two models fine-tuned on Kinetics-710 dataset [41] (*VideoMAE-v2-K710*) and SSv2 dataset [23] (*VideoMAE-v2-SSv2*). More details about the dataset and experimental setups are included in Supplementary Material Section A.

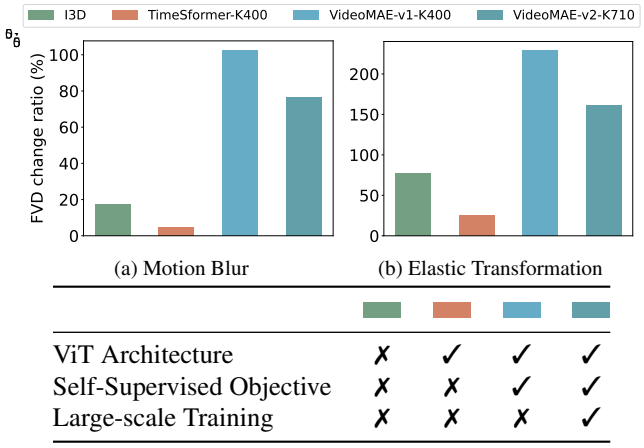


Figure 5. **The origin of FVD sensitivity.** We show the temporal sensitivity in the VideoMAE features is mainly attributed to the self-supervised objective.

FVD temporal sensitivity. We present FID and FVD scores obtained from the videos distorted spatially or spatiotemporally on different datasets in Table 1. We provide more results in Supplementary Material Section B. The minimal FID difference between the spatial and spatiotemporal distortion videos validates our claim that the two distorted video sets share similar frame quality.

Regarding the FVD scores of the spatially distorted videos, several datasets that are not in the distribution of the I3D training data, such as Sky Time-lapse, FaceForensics, and SSv2, generally yield much smaller FVD values compared to in-distribution datasets like Taichi-HD, UCF-101, and Kinectics-400. Given the same level of distortion is applied across different datasets, this suggests that FVD is less sensitive to distortion on the out-of-domain data. We inspect FVD's temporal sensitivity based on its increases induced by temporal inconsistency. Specifically, we compute the relative change of FVD between spatial and spatiotemporal corruptions. We find that FVD sometimes fails to detect the temporal quality decrease. For example, the temporal inconsistency in the FaceForensics dataset only raises FVD by 3%.

To grasp the significance of the FVD increase due to

temporal inconsistency, we compare it with the FVD values computed using VideoMAE-v2 models in Figure 4, where we report the average FVD scores across multiple datasets. When using the VideoMAE models to extract features, we notice a much more pronounced increase of FVD on the videos with temporal inconsistency. For example, when the elastic transformation is adopted to introduce temporal inconsistency, FVD with VideoMAE-v2-PT increases **five** times more than the original FVD with the I3D model.

Compared to VideoMAE variants, the VideoMAE model fine-tuned on the SSv2 dataset consistently exhibits greater temporal sensitivity than the one fine-tuned on the K710 dataset, at least threefold. This difference can be attributed to the SSv2 dataset's emphasis on motion, where different videos share similar visual content, while differences only arise in fine-grained motion cues. Both *VideoMAE-v2-SSv2* and *VideoMAE-v2-PT* exhibit larger sensitivity to the temporal quality. Computing FVD with *VideoMAE-v2-SSv2* features effectively captures mild temporal quality decrease induced by Motion Blur, whereas using *VideoMAE-v2-PT* features proves more sensitive to temporal distortion introduced by Elastic Transformation. Overall, they all exhibit more sensitivity to temporal corruptions compared to FVD computed with the I3D model.

Where does the content bias originate from? Multiple factors could contribute to the increased temporal sensitivity when using features computed from the VideoMAE-v2 model. These factors may encompass the model architecture, training objectives, model capacity, and the dataset. To unravel these intricacies, we further delve into a comparative study with two other models, VideoMAE-v1 [70] and TimeSFormer [6] models.

The VideoMAE-v1 model uses a smaller ViT model size while sharing the same objective as the VideoMAE-v2 model, which helps us demystify the training scales. Due to limited computing resources, we cannot train the VideoMAE-v2 ViT model from scratch with the supervised objective. Instead, we train the smaller ViT with the size of VideoMAE-v1 using the recipe from TimeSFomer. Both models are trained on the Kinetics-400 dataset, sharing the same training dataset as the I3D model. For VideoMAE-v2, we use the *VideoMAE-v2-K710* model, as it shares the most similar fine-tuning dataset, Kinetics-710, with other models. Note that it has the least temporal sensitivity among the three variants, as shown in Figure 4. We use it to make a fair comparison regarding the fine-tuning dataset.

We summarize the distinctions between these models in the table of Figure 5. We perform our temporal sensitivity analysis and report the FVD ratio with different feature extractors in Figure 5. The major improvement in the temporal sensitivity arises when comparing the VideoMAE-v1 and TimeSFormer models. Based on these observations, we conclude that the self-supervised training objective contributes the most to mitigating the content bias.

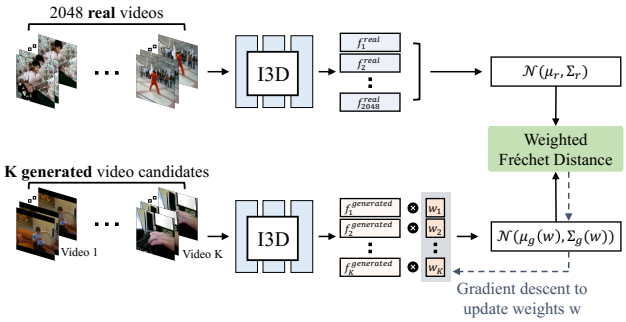


Figure 6. **Probing the perceptual null space in FVD.** We sample a $8\times$ larger set of fake videos and compute a weight for each candidate video by optimizing the weighted Frechet distance. We then use the weights to sample $2,048$ videos to compute the final FVD score.

4. Probing the Perceptual Null Space in FVD

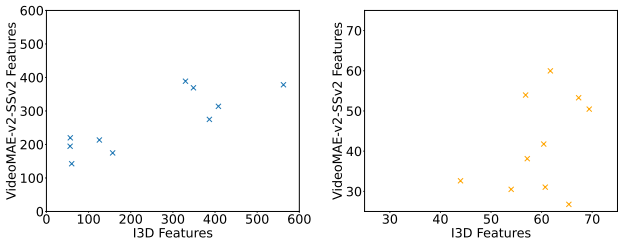
Having observed FID's insensitivity to temporal quality in our synthetic experiments, a natural question arises: how does this bias impact practical evaluation? To address this question, we leverage an analysis method introduced to understand undesired behavior in FID [39]. This tool examines the *perceptual null space*, where the quality of generated images remains similar while the FID score can be effectively adjusted. In our scenario, we generate a large set of candidate videos from the same model and carefully select a subset to lower the FVD score.

As the core assumption of the method, the samples synthesized by the same model should exhibit relatively similar visual quality. However, we observe that the quality may sometimes vary for different generative videos. Therefore, we further extend the analysis to understand the concept of temporal perceptual null space, where we hard-constrain the temporal quality of the generated videos to be the same by using frozen videos.

Resampling method. We adopt the resampling technique proposed by Kynkäänniemi et al. [39]. Given a set of K, where $K > N$, generated videos, we assign a weight $w_i \in \mathbb{R}$ for each video, aiming to minimize the *weighted FVD*. Given the weights, the FVD defined in Equation 1 is reformulated with the weighted mean $\mu_g(\mathbf{w}) = \frac{\sum_i \exp w_i \mathbf{f}_i}{\sum_i \exp w_i}$, and covariance $\mathbf{\Sigma}_g(\mathbf{w}) = \frac{\sum_i \exp w_i (\mathbf{f}_i - \mu)(\mathbf{f}_i - \mu)^T}{\sum_i \exp w_i}$ as:

$$\|\mu_r - \mu_g(\mathbf{w})\|_2^2 + \mathrm{Tr}\left(\mathbf{\Sigma}_r + \mathbf{\Sigma}_g(\mathbf{w}) - 2\left(\mathbf{\Sigma}_r \mathbf{\Sigma}_g(\mathbf{w})\right)^{\frac{1}{2}}\right). \tag{2}$$

This *weighted FVD* serves as the objective for optimizing $\mathbf{w}$. After the optimization process, the resampling is performed



(a) FVD scores (b) FVD decrease by resampling (%)

Figure 7. **FVD decrease induced by resampling across different models and datasets.** We compare FVD computed from the VideoMAE-v2-SSv2 and I3D features. Each dot represents a video generation model trained on a specific dataset. (a) We notice a non-monotonic relationship between the FVD computed with I3D and VideoMAE-v2 features. (b) After resampling, the FVD with VideoMAE features generally decreases less than the FVD with I3D features, since most dots are located in the bottom right area.

with the probability $\frac{\exp w_i}{\sum_i \exp w_i}$. $2,048$ videos are sampled from the candidate set to compute the new FVD score, which we denote as FVD*.

Experimental setups. We experiment with several video generation models, including GANs [64, 91], Transformer [20], and Diffusion models [92]. As these models are evaluated on different benchmarks, we also test with different datasets, including UCF-101 [66], Sky Time-lapse [86], Taichi-HD [61], and FaceForensics [52]. We use the official scripts and checkpoints to sample the videos. For each experiment, we generate a $8\times$ larger candidate set, i.e., $K = 16,384$ videos. To optimize the weights $\mathbf{w}$, we use gradient descent with an initial learning rate of 0.01 and a linear scheduler that decreases the learning rate by 0.1 at every 100 steps. We perform optimization for 300 steps, at which point the weighted FVD scores often converge.

We are especially interested in how much we can reduce the FVD score without improving the temporal quality. To achieve this, we convert each generated video into a *frozen* video by repeating its first frame 16 times. By doing so, we enforce all the videos to contain no motion so that the temporal quality cannot be improved through resampling. Motivated by the temporal sensitivity presented by the VideoMAE-v2 models, we also perform the experiments with the *VideoMAE-v2-SSv2* model as the feature extractor, where the resampling is done by optimizing its specific weights. Further details about the video generation models and experimental setups are included in Supplementary Material Section A.

Observation on the resampling results. We show FVD scores before resampling in Figure 7 (a). We first observe a non-monotonic relationship between the FVD computed with I3D and VideoMAE-v2 features, potentially leading to different model rankings when using these two features. For FVD after resampling, we report the change ratio $\frac{\text{FVD}^*-\text{FVD}}{\text{FVD}}$ in Figure 7 (b). It illustrates a significant drop ($40\% - 70\%$) in FVD computed with I3D features after resampling, while FVD computed with VideoMAE-v2 features experiences a more moderate drop ($25\% - 60\%$).

(a) Candidate videos with the smallest weights.

(b) Candidate videos with the largest weights.

Figure 9. **Candidate videos with the largest and smallest weights.** We visualize the resampling results of DIGAN trained on the Taichi-HD dataset with the 32 largest (most likely to select) and least (least likely to select) weights. We observe clear quality degradation in the samples with the smallest weights. *Best viewed with Acrobat Reader. Please check the website for videos.*

We note clear quality differences after inspecting videos with the highest and lowest weights obtained from optimizing weighted FVD. An example of videos generated by DIGAN trained on the Taichi-HD dataset is shown in Figure 9. We attribute the different observations from the original paper [39] to the unstable performance of the video generator. Since the video generator sometimes generates nonrealistic videos, resampling is beneficial in selecting a group of higher-quality videos, yielding smaller FVD scores.

Temporal perceptual null space. Table 2 shows extensive results, including the FVD of original videos with motions, the FVD of frozen videos, and the weighted FVD of frozen videos after resampling. Despite the absence of motion in the generated videos, one can still reduce FVD by up to half by selectively choosing from the candidate videos when evaluating the Sky Time-lapse dataset. In the worst case, the same or even smaller FVD scores can be achieved compared

Table 2. **Results of probing the temporal perceptual null space of FVD.** We report FVDs of normal and frozen generated videos by random sampling (FVD) and resampling to minimize weighted FVD (FVD*). We color the FVD difference for better visualization: $< 20\%$, $20\% - 40\%$ and $> 40\%$. 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. The gray background indicates the samples where resampling frozen videos can obtain similar or even better FVD than the random generation results with motions.

Model	Dataset	I3D Features			VideoMAE-v2-SSv2 Features		
		FVD	$\text{FVD}_{\text{w/o motion}}$	$\text{FVD}^*_{\text{w/o motion}}$	FVD	$\text{FVD}_{\text{w/o motion}}$	$\text{FVD}^*_{\text{w/o motion}}$
DIGAN [91]	UCF-101 [66]	562.36	1303.13	715.96(−45.1%)	378.19	951.59	859.57(−9.7%)
DIGAN [91]	Sky Time-lapse [86]	157.13	230.64	115.55(−49.9%)	174.79	408.17	362.84(−11.1%)
DIGAN [91]	Taichi-HD [61]	132.26	461.79	276.88(−40.0%)	313.84	578.61	523.20(−9.6%)
TATS [20]	UCF-101 [66]	329.92	1157.69	616.25(−46.8%)	388.79	908.95	805.88(−11.3%)
TATS [20]	Sky Time-lapse [86]	125.62	279.75	126.32(−54.8%)	213.33	375.74	353.15(−6.0%)
TATS [20]	Taichi-HD [61]	124.16	475.99	312.19(−34.4%)	274.81	587.31	530.86(−9.6%)
StyleGAN-V [64]	Sky Time-lapse [86]	56.63	206.56	104.27(−49.5%)	219.85	503.22	456.24(−9.3%)
StyleGAN-V [64]	FaceForensics [52]	56.22	353.79	242.04(−31.6%)	194.68	547.24	520.98(−4.8%)
PVDM [92]	UCF-101 [66]	348.81	1135.61	605.09(−46.7%)	369.14	1032.90	898.48(−13.0%)
PVDM [92]	Sky Time-lapse [86]	59.95	182.77	94.87(−48.1%)	142.50	429.06	395.79(−7.8%)

with randomly selected generated videos *with* motions.

These findings highlight the pronounced content bias inherent in the FVD metric. Conversely, when computing the features for FVD using the VideoMAE-v2 model, which is sensitive to temporal quality, the gap significantly diminishes, and the FVD scores can hardly be decreased through resampling. This emphasizes that the FVD with VideoMAE-v2 has a much smaller temporal perceptual null space. More results are available in Supplementary Material Section B.

5. Case Study: Long Video Generation

Our experiments have revealed that FVD does not sufficiently account for motion in generated videos. We now dive into two case studies from previous works where FVD scores contradict human perception [20, 64, 91]. In both cases, the video generation models are trained on the Sky Time-lapse dataset, which is out-of-domain for the I3D model, and FVD has been shown not to perform well. In addition, both tasks generate longer videos than the standard 16-frame setting, making the motion artifacts more perceptible to humans. Nevertheless, FVD fails to capture these motion artifacts in both experiments.

Case study I [64]. To synthesize long videos, the StyleGAN-v model [64] employs convolutional layers with large reception fields to predict the parameters of the Fourier temporal encoding. We reproduce one of its baselines by substituting such temporal encoding with an LSTM layer. For generating 128-frame videos, the default StyleGAN-V synthesizes realistic motions (Figure 11a), whereas the baseline with continuous LSTM codes and $\sigma^z = 16$ leads to videos with noticeable motion collapses (Figure 11b).

We follow the original study's evaluation protocol and compute the FVD metric by feeding all the 128 frames to the I3D model, termed as FVD_{128}. Note that the I3D model was initially trained on 64 frames, while the global average pooling and convolutional architecture allow it to be applied to any video length. We also compute FVD_{128} with the VideoMAE. Since the VideoMAE uses a ViT with fixed-size positional encoding, we perform interpolation of positional encodings, similar to DINO [13].

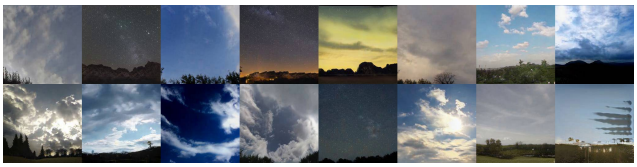
Table 3. StyleGAN-v model [64] with LSTM as the motion codes trained on the Sky Time-lapse dataset generate collapsed motions, whereas FVD computed on the 128 frames favors the results. We show that computing FVD with VideoMAE-v2 features calibrates the conclusion.

Frame #	FVD Feature	StylegGAN-v	w/ LSTM codes
16	I3D	120.11	136.65 (+16.54%)
	VideoMAE-SSv2	223.96	247.25(+23.29%)
	VideoMAE-K710	145.37	154.29(+8.92%)
128	I3D	190.82	172.71(−18.11%)
	VideoMAE-SSv2	332.80	616.74(+283.94%)
	VideoMAE-K710	155.51	191.48(+35.97%)

Contrary to the visual evidence, we observe the same trend as noted by the authors that FVD_{128} computed using the I3D model is lower for the LSTM variant, compared to the original StyleGAN-v, as shown in Table 3. Upon computing FVD_{128} using VideoMAE-v2 features, both using SSv2 and K710, we have them to be in accordance with human preference, i.e., FVD_{128} for the LSTM baseline is much worse compared to the original StyleGAN-v.

Case study II [20, 91]. Though trained on 16-frame video

(a) Default StyleGAN-v.



(b) StyleGAN-v with LSTM motion codes.

Figure 11. **Videos generated by StyelGAN-v and its LSTM variant.** The default StyleGAN-v synthesizes natural motions, while the variant with LSTM motion codes generates repeated patterns. *Best viewed with Acrobat Reader. Please check the website for videos.*

Table 4. DIGAN [91] trained on the Sky Time-lapse dataset with extrapolated time steps generate periodic artifacts, whereas the FVD metric favors the results. We show that computing FVD with VideoMAE-v2 features calibrates the conclusion.

FVD Feature	Frames 0 - 16	Frames 128 - 144
I3D	155.58	141.82(−8.84%)
VideoMAE-SSv2	133.37	150.61(+12.9%)
VideoMAE-K710	250.70	259.29(+3.43%)

clips, DIGAN can be applied to generate longer videos by extrapolating the temporal encodings. However, in the previous study [20], the authors have observed that motion artifacts in the form of repeated changes in the diagonal direction, while FVD computed on the 16-frame chunks, favor the artifacts.

Specifically, to evaluate the long video generation results, they compute FVD at strides of k frames, 16 frames at a time, for the entire video. We compute FVD on frames $0 \rightarrow 16$, $64 \rightarrow 82$, $128 \rightarrow 144$, and so on. The visualization of videos at $0 \rightarrow 16$ and $128 \rightarrow 144$ Figure 13 show more periodic artifacts at $128 \rightarrow 144$ compared to $0 \rightarrow 16$. In contrast, from Table 4, we see that FVD computed using I3D features favors $128 \rightarrow 144$ frames, though FVD computed using VideoMAE features for both SSv2 and K710 follow human judgment.

Recent progress in photorealistic video generation using Diffusion models has enabled the creation of videos with extended durations ($\geq$ 100 frames) [9, 21, 30, 62]. As a result, evaluating long video generation results has become increasingly important. However, according to the two real-world case studies above, FVD with I3D features does not reliably detect motion artifacts in long videos.

(a) Frames 0 - 16.

(b) Frames 128 - 144.

Figure 13. **Videos generated by DIGAN at different extrapolated time steps.** The initial 16 frames generated by DIGAN exhibit natural motions, while the extrapolated frames contain periodic artifacts. *Best viewed with Acrobat Reader. Please check the website for videos.*

6. Discussion

In this paper, we have studied the bias of the FVD on the frame quality. With experiments spanning from synthetic video distortion, to resampling video generation results, to investigating real-world examples, we have concluded that FVD is highly insensitive to the temporal quality and consistency of the generated videos. We have verified the hypothesis that the bias originates from the content-biased video features and show that self-supervised features can mitigate the issues in all the experiments. We hope our work will draw more attention to studying video generation evaluation and designing better evaluation metrics.

Limitations. Several critical aspects of FVD remain underexplored. For example, in addition to the longer time duration, existing methods also generate megapixel resolution videos [9, 21, 30, 62]. However, to compute FVD (I3D or VideoMAE-v2), the video must be resized to a lower (e.g., 224x224) resolution. In addition, many existing methods choose to generate videos not limited to the square aspect ratio, e.g., 16:9, while the FVD metric always requires a square video as the input. Computing FVD using VideoMAE-v2 features is limited by the quadratic cost of attention layers, which could cause issues in evaluating longer video generation.

Acknowledgment. We thank Angjoo Kanazawa, Aleksander Holynski, Devi Parikh, and Yogesh Balaji for their early feedback and discussion. We thank Or Patashnik, Richard Zhang, and Hadi Alzayer for their helpful comments and paper proofreading. We thank Ivan Skorokhodov for his help with reproducing the StyleGAN-v ablation experiments. This work is partly supported by NSF grant No. IIS-239076, the Packard Fellowship, as well as NSF grants No. IIS-1910132 and IIS-2213335.

References

[1] Motasem Alfarra, Juan C Pérez, Anna Frühstück, Philip HS Torr, Peter Wonka, and Bernard Ghanem. On the robustness of quality measures for gans. In *ECCV*, 2022. 2, 3

[2] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 3

[3] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *ICLR*, 2018. 3

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 1

[5] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. 1

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 2, 3, 5

[7] Mikolaj Binkowski, Danica J. Sutherland, Michal Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018. 3

[8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 3

[9] 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 *CVPR*, 2023. 1, 3, 8

[10] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. 2, 3

[11] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *NeurIPS*, 35:31769–31781, 2022. 2

[12] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 1

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2, 4

[15] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *ICCV*, pages 7608–7617, 2019. 3

[16] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 3

[18] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, pages 1174–1183. PMLR, 2018. 3

[19] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 2

[20] 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. 2, 3, 6, 7, 8

[21] 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 *ICCV*, 2023. 1, 3, 8

[22] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning, 2023. 1

[23] 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 database for learning and evaluating visual common sense. In *ICCV*, 2017. 4

[24] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation, 2023. 3

[25] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models, 2023. 3

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 4

[27] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 4

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 1, 2, 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[30] 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, 2, 3, 8

[31] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2

[32] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *CVPR*, 2018. 2, 3

[33] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. 1

[34] Yunseok Jang, Gunhee Kim, and Yale Song. Video prediction with appearance and motion conditions. In *ICML*, pages 2225–2234, 2018. 2

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[36] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *CVPR*, 2023. 3

[37] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, Yong Cheng, Ming-Chang Chiu, Josh Dillon, Irfan Essa, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, David Ross, Grant Schindler, Mikhail Sirotenko, Kihyuk Sohn, Krishna Somandepalli, Huisheng Wang, Jimmy Yan, Ming-Hsuan Yang, Xuan Yang, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024. 3

[38] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019. 1, 3

[39] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *ICLR*, 2022. 2, 3, 4, 5, 6

[40] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019. 1

[41] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer, 2022. 4

[42] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018. 2

[43] Xin Liu, Silvia L Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C Van Gemert. No frame left behind: Full video action recognition. In *CVPR*, 2021. 2, 3

[44] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2023. 1

[45] 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. In *CVPR*, 2023. 3

[46] Stanislav Morozov, Andrey Voynov, and Artem Babenko. On self-supervised image representations for gan evaluation. In *ICLR*, 2020. 2, 3, 4

[47] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *ICML*, pages 7176–7185. PMLR, 2020. 1, 3

[48] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 1, 3, 4

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 3

[50] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014. 3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[52] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 4, 6, 7

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3

[54] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017. 3

[55] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *NeurIPS*, 2018. 1, 3

[56] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 3

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1

[58] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441, 2010. 1

[59] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 535–544, 2021. 2, 3

[60] Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Mostgan-v: Video generation with temporal motion styles. In *CVPR*, 2023. 3

[61] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 6, 7

[62] 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, 3, 8

[63] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018. 1

[64] 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. 2, 3, 6, 7

[65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[66] 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. 4, 6, 7

[67] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 3

[68] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3

[69] 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. 2, 3

[70] 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. 5

[71] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018. 2, 3

[72] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1, 2, 3

[73] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017. 2, 3

[74] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *NeurIPS*, 2016. 3

[75] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 2

[76] 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. 2, 4

[77] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 3

[78] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model, 2023. 3

[79] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *CVPR*, 2020. 2, 3

[80] 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. 3

[81] Zhou Wang, Ligang Lu, and Alan C Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, 2004. 1

[82] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *ICLR*, 2019. 3

[83] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 3

[84] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors, 2023. 3

[85] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation, 2023. 3

[86] 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. 2, 4, 6, 7

[87] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[88] Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation, 2023.

[89] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer. In *CVPR*, pages 10459–10469, 2023.

[90] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation, 2023. 3

[91] 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. 3, 6, 7, 8

[92] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, 2023. 2, 6, 7

[93] 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, 2023. 3

[94] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[95] 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. 3

[96] Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. Hype: A benchmark for human eye perceptual evaluation of generative models. In *NeurIPS*, 2019. 3