Frame by Familiar Frame: Understanding Replication in Video Diffusion Models

Aimon Rahman[1*], Malsha V. Perera[1*], and Vishal M. Patel[1]

Johns Hopkins University, Baltimore MD 21218, USA
{arahma30,jperera4,vpatel36}@jhu.edu

Abstract. Building on the momentum of image generation diffusion models, there is an increasing interest in video-based diffusion models. However, video generation poses greater challenges due to its higher-dimensional nature, the scarcity of training data, and the complex spatiotemporal relationships involved. Image generation models, due to their extensive data requirements, have already strained computational resources to their limits. There have been instances of these models reproducing elements from the training samples, leading to concerns and even legal disputes over sample replication. Video diffusion models, which operate with even more constrained datasets and are tasked with generating both spatial and temporal content, may be more prone to replicating samples from their training sets. Compounding the issue, these models are often evaluated using metrics that inadvertently reward replication. In our paper, we present a systematic investigation into the phenomenon of sample replication in video diffusion models. We scrutinize various recent diffusion models for video synthesis, assessing their tendency to replicate spatial and temporal content in both unconditional and conditional generation scenarios. Our study identifies strategies that are less likely to lead to replication. Furthermore, we propose new evaluation strategies that take replication into account, offering a more accurate measure of a model's ability to generate the original content.

Keywords: Video Generation · Data Replication · Video Diffusion

1 Introduction

Recent advancements in diffusion-based image generative models have paved the way for video generation [14, 32, 38]. However, video synthesis has not reached the same prominence as its image counterpart, primarily due to the immense computational demands and the scarcity of expansive, public video datasets [43]. Most current models limit themselves to producing short, low-resolution videos [15, 25]. The power of the video diffusion framework lies in its ability to utilize simple denoising networks that incorporate a temporal dimension for motion [3, 15]. Such models, when trained on video data, can generate realistic videos full of action and content. Yet, a pressing concern is the potential for training data replication. Even in the image domain, diffusion models are known to replicate content from training datasets, leading to concerns about the originality of the produced content [36, 37].

[*] *Equal Contribution

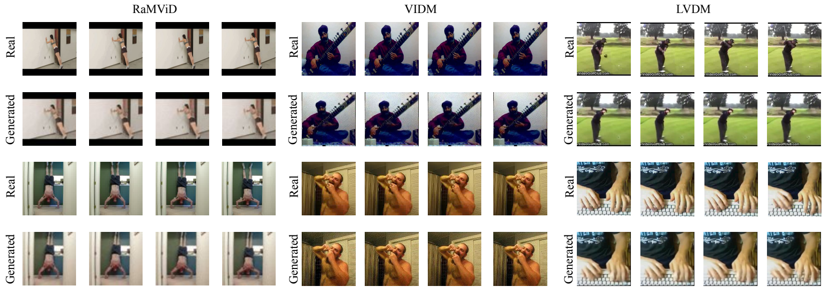

Fig. 1: Diffusion-based video synthesis models can sometimes replicate training data by assembling memorized foreground and background elements. We demonstrate this trend across multiple diffusion models trained on diverse datasets. Such occurrences prompt inquiries regarding data memorization and the ownership of videos produced by diffusion methods. Bottom row: Videos sourced from the RaMViD [16], VIDM [25], and LVDM [12] project websites. Top row: The most similar counterparts from the training dataset.

Exploring Replication in Video vs. Image Generation Models. The issue of training data replication is a well-documented challenge in image generation models, with significant research devoted to understanding its impacts and mechanisms [36, 37]. These models demonstrate a remarkable ability to generate both unique content and, in some instances, partial or complete replications of their training data. In the domain of video generation models, this challenge is amplified due to the additional complexity of generating content that encompasses both static images and their temporal evolution with much smaller training data. Thus, it becomes imperative to examine the extent to which video generation models can innovate in terms of content and motion creation. Moreover, the field of video generation is diverse, covering areas such as video prediction, conditional and unconditional generation, text-conditioned generation, video infilling, etc. This underscores the necessity of investigating how video diffusion models manage the balance between replication and the generation of novel content.

Implication of Data Replication in Video Diffusion Models. The extent of data replication within video generation frameworks holds significant implications, beyond copyright violation, and in particularly in the domain of security and biometrics. A notable concern arises when a video replicates an individual's face from a training dataset, potentially leading to privacy issues. Moreover, a person's unique motion, such as their gait, can be distinct enough to facilitate identification [17, 20, 29]. Replicating such motions, either by mimicking a person's gait or other physical patterns, might have detrimental effects on user authentication processes in behavioral biometrics. Additionally, the recent discovery of motion data for identifying individuals in virtual reality (VR) contexts amplifies these concerns [27, 28]. The implications of video replication thus extend beyond mere copyright infringement, especially when synthetic videos find applications in other downstream areas.

Contributions. Our research centers on the scientific inquiry into video diffusion models, examining their capacity for generating unique content, the extent and frequency of replication, and strategies for mitigating such replication. Our research delves into:

- Defining “replication" in videos. This can be subjective, varying based on content diversity or viewer interpretation. We’ve pinpointed both clear and ambiguous instances of replication, differentiating between content and motion.
- Investigating the frequency of replication in video diffusion models, looking at both content and motion. We aim to determine if these models truly comprehend the actions they generate.
- Analyzing the relationship between the realism of generated videos and content replication. The hypothesis is that hyper-realistic videos might just be reflections of the training dataset.
- Examining video similarity metrics to effectively detect data replication. This will also help set benchmarks for future video diffusion model evaluations, especially since the current metrics, like the FVD, reward similarities to the training dataset.
- Offering recommendations for protocols in training and evaluating future video generation models. Our focus is on establishing guidelines that enhance model performance while ensuring diverse and original content generation, moving beyond current practices that might inadvertently favor replication over innovation.

2 Related Work

Our study intersects with various domains: diffusion models, image/video generation techniques, and the inclination of generative models to mimic samples. Here, we briefly outline the foundational concepts from each area, emphasizing their interrelations and addressing the challenges of video sample replication.

Sample Replication in Generative Models. Recent research has highlighted a trend of training data replication or memorization in popular generative models such as Generative Adversarial Networks (GANs) and diffusion models. As these models become increasingly adept at creating hyper-realistic images, questions arise about the originality of these images versus their being mere duplications from training datasets. The “this person does not exist" phenomenon, famously associated with faces generated by StyleGAN [18], has been scrutinized, revealing the potential to trace back to the original dataset used in the model [47]. It is observed that the tendency of GANs to replicate training data decreases exponentially with the increase in dataset complexity and size [10]. The suggestion has been made that the duplication of training data is a significant factor in this context. By employing non-parametric tests, it is possible to address and mitigate this issue effectively [24]. This phenomenon of replication extends to diffusion models too, where retrieving training data from the model is possible [5]. Although often attributed to training on smaller datasets, this replication behavior is also evident in models trained on larger datasets [36, 37]. The issue becomes particularly concerning in terms of social bias, especially with diffusion models trained on facial datasets, where replication behavior is prominent [30].

Diffusion-based Video Generation. Diffusion Probabilistic Models (DPMs) are a subset of deep generative models known for their ability to incrementally introduce noise into data points [14, 38]. This process is followed by a denoising phase that iteratively cleans the data, resulting in the generation of new samples. DPMs have shown remarkable results in producing high-quality and diverse images, inspiring researchers to explore their potential in video generation, prediction, and interpolation [13,15,16,23,44].

Applying DPMs to video generation is still in its infancy and presents unique challenges [43]. Videos possess higher-dimensional data and intricate spatiotemporal relationships, making the task more complex. Diffusion-based video generation models operate on the principles similar to image diffusion models, with a key distinction in their architecture [15]. These models typically use either a 3D Unet architecture [15, 16], which adds depth to the processing, or a conditional 2D Unet that takes into account the previous frame in a sequence [44]. Most of these models incorporate a temporal layer to capture motion, and variations of this layer are evident in the current research [3, 23, 25, 26]. The process of generating new frames in these models is autoregressive; the creation of each subsequent frame depends on the preceding one. This can be either conditional or unconditional. In a conditional generation, some models use textual descriptions to guide the generation process [26, 43, 48]. Others might use the first frame or a few initial frames as a basis for what is essentially video prediction [16, 44]. When the generation is steered by intermittent frames throughout the video, it is referred to as video infilling. Unconditional generation, on the other hand, relies purely on noise to create videos. Regardless of the approach, these models require an intrinsic understanding of motion and content to perform their tasks effectively. It is also worth noting that these video generation models are often *not publicly available*. They can be resource-intensive, both in terms of training and testing, which might contribute to their limited accessibility.

Video copy detection and localization. Video Copy Detection (VCD) involves identifying pairs of query-reference videos containing copied content without localizing the common content within the videos. In contrast, Video Copy Localization (VCL) requires finding the exact temporal segments within a pair of videos that contain duplicated content. VCD can be performed either at the video-level or frame-level. Video-level approaches utilize standard similarity measures on global representations of the videos to identify copied content [4, 22, 39]. However, these approaches are less effective in partial copy detection tasks as they aggregate with irrelevant content and clutter. Meanwhile, frame-level features with spatio-temporal representations have proven to yield a significant advantage in video retrieval tasks and are advantageous in precisely locating copied segments. Various techniques, such as Fourier-based representations [2], multi-attention networks [46], or transformer-based networks [11, 34], are used for temporal aggregation. Recent works employ a video similarity network that captures fine-grained spatial and temporal structures within pairwise video similarity matrices [21]. In VCL, frame-level feature representations are followed by a temporal alignment module that needs to reveal the time range of one or multiple copied segments between the potential copied video pair. Frame-level features can typically be extracted using image descriptors commonly employed in image copy detection. One popular example is SSCD [31], an image descriptor based on self-supervised learning, which optimizes a descriptor for copy detection through entropic regularization. Temporal Hough voting [8], temporal networks [41], and dynamic programming [7] are examples of some of the simplest VCL methods.

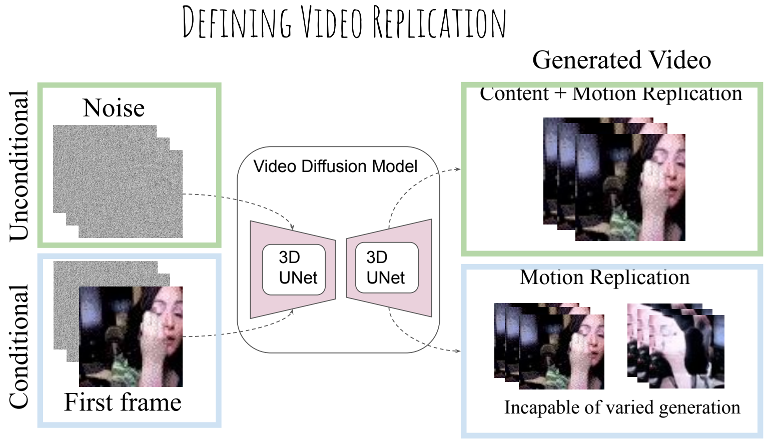

Fig. 2: Definition of replication in video generation domain. Content and motion replication refers to the direct duplication of content and motion from the training dataset, essentially producing a 1:1 copy. On the other hand, motion replication assesses a video generation model's inherent ability to create motion from an initial frame. This initial frame supplies the content context, but the true measure of a video generation network's capability lies in its understanding of the comprehensive content within that first frame. The critical question is whether the network genuinely comprehends and generates subsequent motion, or if it merely replicates sequences it has learned from the training data.

3 Defining Video Replication

Replicating content in the imaging domain generally means that a significant portion of the training image appears in the generated images [36]. What's considered a "significant portion" varies, but it often refers to an easily recognizable region. In essence, it is considered a generated image to have duplicated content if it features an object (either in the foreground or in the background) that mirrors an object from a training image, allowing for slight variations that might arise from data augmentations [36].

However, defining data replication in video generation is a tad more intricate. For unconditional generation, which is the generation from pure random noise without any specific guiding condition, data replication can encompass both the *subject and its motion*. In the domain of conditional generation, where the model is given an initial frame and then predicts the subsequent video, the question arises: *Does the model genuinely understand and generate motion, or is it merely recalling patterns from the training dataset?* This necessitates different definitions of video replication based on the generation context (conditional or unconditional) as illustrated in Figure 2.

For unconditional generation, if the content closely matches that of a training video with minimal differences, it is regarded as replication. Essentially, the replication of visual elements (content) from the training video inevitably leads to the replication of the dynamics or movement patterns (motion) present in the original video. On the other hand, in the conditional context, if a model reproduces the exact movement sequence after being provided with just an initial frame, it is seen as motion replication. In the latter scenario, it is understandable if a model replicates motion when predicting or infilling video. However, to test genuine understanding, alterations can be made to the initial frames—like changing the angle or occluding the frame—to see if the model can still predict plausible subsequent motion. If the model simply replicates the training data to generate motion based on the first frame, this process is referred to as *motion replication* in this paper.

4 Detecting Data Replication in Video Diffusion Models

4.1 Content Replication.

Content replication in the context of text-to-video or unconditional video generation refers to a scenario where the frames generated by the model contain the same or strikingly similar content to what it has seen during training. This means that instead of creating new, original content based on the learned concepts and dynamics, the model reproduces specific examples from its training data, which suggests a lack of true generative capacity or understanding. Due to limited access to publicly available models, our analysis is mostly based on the generated samples showcased on the official websites of the respective papers. We specifically focus on results that are obtained through unconditional generation, ensuring that initial frame contents don't influence the generated samples.

Experimental Setup. Initial findings indicate that a significant portion of the video samples generated unconditionally by the model are direct replicas of the training videos, featuring identical content. To identify these replications, we adapted the Self-Supervised Copy Detector (SSCD) [31], originally designed for image copy detection, for use with video content. Our methodology involves segmenting frames from reference videos, extracting their features, and then concatenating these features. This process is repeated for all training videos. Next, we employ SSCD to extract features from these concatenated frames, denoting these video features as VSSCD. We then calculate the cosine similarity to identify the top matches. Let's denote $R = \{R_1, R_2, ..., R_n\}$ as the set of real videos and $G = \{G_1, G_2, ..., G_m\}$ as the set of generated videos. We calculate the similarity between each pair of VSSCD features of real and generated videos. The similarity between a real video R_i and a generated video G_j is denoted as VSSCD(R_i, G_j). The process can be represented by the following equation:

$$\text{Top-VSSCD} = \max\big\{\text{VSSCD}(R_i, G_j) : 1 \leq i \leq n, 1 \leq j \leq m\big\}, \quad (1)$$

where Top-VSSCD is the highest similarity score among all combinations of real and generated videos. In this work, we use the term 'VSSCD score' to refer to the similarity score between the VSSCD features of two videos. The validity of this VSSCD-based approach is demonstrated in Table 1, where we present the VSSCD scores for exact copies, various augmented copies, and scores corresponding to random videos that do not match the reference.

Then, we evaluate the extent of sample replication in various video generation models and compare these findings with their Fréchet Video Distance (FVD) scores. We examine the output of several models including VIDM [25], VDM [15], RaMViD [16], and LVDM [12]. For this analysis, we only train the VDM model directly on the UCF-101 dataset to generate videos. For the other models - VIDM, RaMViD, and LVDM - we

Table 1: VSSCD Scores for Replication Detection in the UCF-101 [40] and Kinetics-400 [19] Datasets. The table compares the VSSCD scores for exact 1:1 copies, augmented copies, and unrelated videos, demonstrating the reference efficacy of VSSCD in identifying replicated content in videos.

Frame Operation	1:1	Flip	Crop	Occlusion	Translation	Rotation	Random
UCF-101 [40]	1	0.9684	0.9032	0.9998	0.9147	0.8574	0.0788
Kinetics-400 [19]	1	0.9800	0.9026	0.9999	0.9043	0.8031	0.1046

use the video outputs available on their respective websites. All of these models were initially trained on the UCF-101 [40] dataset. Note that, the generation is unconditional, hence content and motion are generated from pure noise.

Table 2: Quantitative comparison of the FVD and VSSCD scores among various video generation networks.

Metrics	VDM [15]	VIDM [25]	LVDM [12]	RaMViD [16]
VSSCD	0.591	0.6347	0.694	0.744
FVD [44]	631	172.77	151.34	152.24

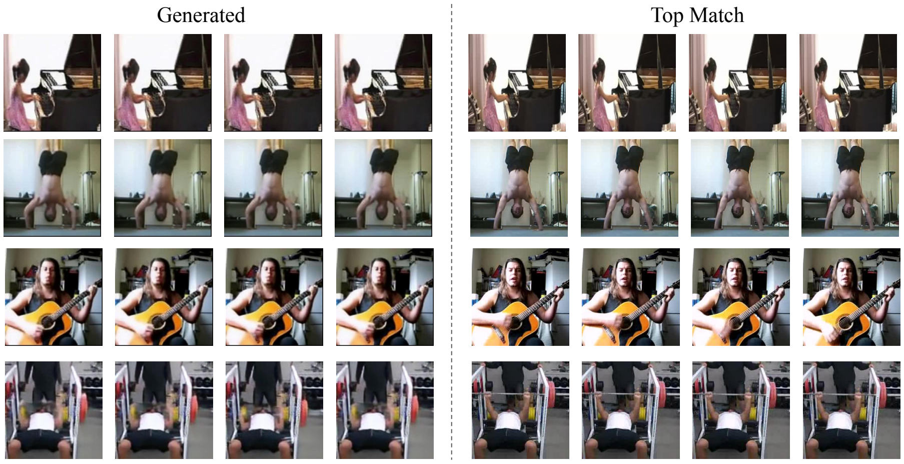

Fig. 3: The highest similarities identified within the UCF-101 dataset compared against outputs generated by an unconditional video generation approach, utilizing a latent video diffusion model (LVDM) [12].

Discussion. We found a tendency among these models to produce videos that replicate training data. This is reflected in Table 2. A shared trait among these models is their training on smaller or limited datasets, such as UCF-101 [40] for VIDM [25], VideoFusion [23], RaMViD [16], etc. This leads us to assume that models trained on limited datasets might be more prone to produce replicated videos due to their limited content understanding. This is evident for both latent and regular video diffusion models, as seen in Figure 1 and Figure 3. This phenomenon is not exclusive to the video domain; similar trends have been observed in image generation models [36, 37]. However, the extent of these trends in image models is not as pronounced as in the video domain. Moreover, what constitutes a “small dataset" varies between the two domains, as video data has to provide both content and motion dynamics.

Observation. Video diffusion models trained from scratch on limited video datasets exhibit a greater tendency to completely replicate the content of the videos within the training dataset.

4.2 Motion Replication

In video prediction tasks, distinguishing between genuine generation and mere replication is challenging, especially when the model is given initial frames to predict subsequent motion. While it is straightforward to spot replication in unconditional generation—by identifying content that mirrors the training set—conditional generation complicates detection. Theoretically, a model with a deep understanding of motion should produce diverse outcomes from the same frame, given the many motion trajectories possible. Producing a singular, training set-specific trajectory might indicate replication, but arguing against it is tough since it might be deemed the ‘best’ result. To probe the model’s grasp of motion, we utilized pre-trained video diffusion models [16, 44] on various datasets.

To evaluate whether a video generation or prediction model possesses a genuine comprehension of the motion dynamics or is merely replicating the learned patterns, we subject the model to a test using the original dataset’s initial frame along with its variants—flipped, cropped, occluded, rotated, and translated. We then assess the model’s performance by comparing the Fréchet Video Distance (FVD) [42] of the output based on the original first frame to that of the augmented versions. We hypothesize that, ideally if the model truly understands motion, the FVD scores should be relatively consistent across all variations.

Discussion. Our results indicate that the model, when provided with an initial frame in its original orientation, is capable of generating subsequent frames effectively. However, it struggles to maintain this performance when presented with minor alterations, such as flipping or cropping, which do not alter the frame’s semantic content. This suggests a tendency towards overfitting on the training dataset. As demonstrated in Table 3, the original orientation consistently results in superior FVD scores, with any form of variation leading to a degradation in performance. This observation is further supported qualitatively, as illustrated in Figure 4.

Observation. Video prediction models frequently memorize the motion dynamics present in the training dataset, resulting in a limited ability to generate novel motion patterns.

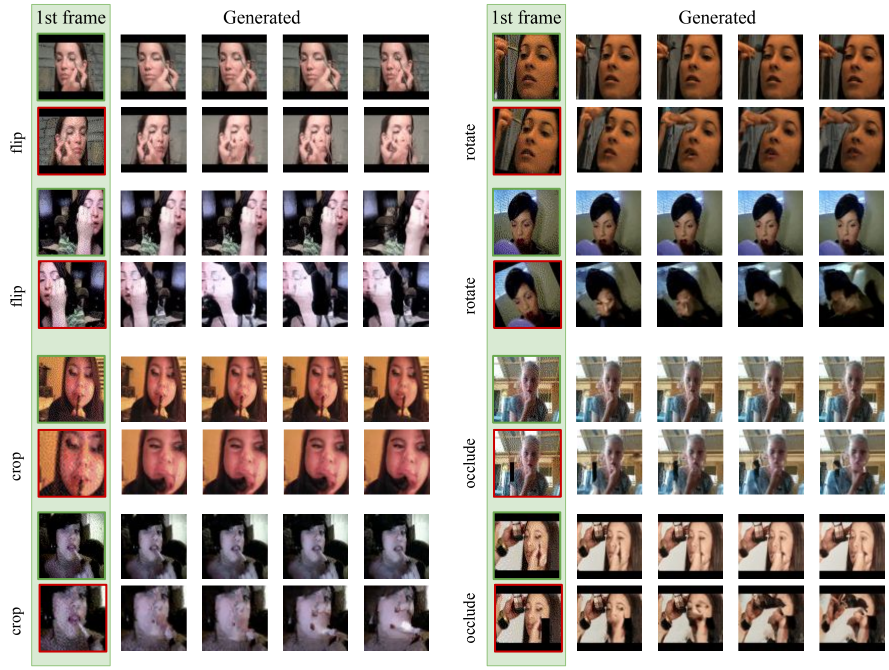

Fig. 4: The initial frame provided as a condition to the video generation model is denoted as the '1^{st} frame'. Frames with a green outline represent their original orientation from the dataset, while those with a red outline signify altered frames. Observations show that the model properly generates motion when presented with frames in their original orientation. However, it struggles to produce consistent motion when given an augmented version of the same image, indicating the model memorized the motion.

Table 3: Quantitative comparison of FVD scores on UCF-101 [40] and Kinetics-400 [19] datasets among various video prediction networks [16].

Frame Orientation	Video Prediction Model [16]	
	UCF-101 [40]	Kinetics-400 [19]
Original	667.64	824.99
Flip	942.02	916.22
Crop	896.46	981.83
Occlusion	867.57	956.31
Translation	873.36	900.33
Rotation	806.02	1022.34

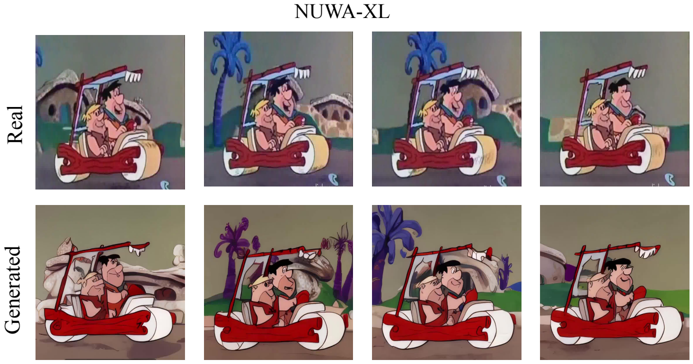

Fig. 5: An instance of replication in a text-to-video (T2V) model [48]. Generated with the text prompt “Fred and Barney driving a car". NUWA-XL has been trained solely on the episodes of “The Flintstones". The replicated segment is from the episode “Disorder in the Court".

5 Replication in Video Diffusion Models

5.1 Replication in Text-to-Video models

Text-to-Video Diffusion Models Utilizing WebVid-10M. The foundation for many text-to-video (T2V) diffusion models [3, 23, 45, 49, 50] has been the WebVid-10M dataset [1]. Unfortunately, this dataset has been retracted due to copyright issues. The construction and operation of these expansive T2V models not only demand substantial computational resources but also face availability restrictions, with most not being accessible to the public. This presents significant challenges for those aiming to replicate these models, as access to both the datasets and the models themselves is restricted. Nonetheless, in Section 6.2, we demonstrate that extending models from the T2I architecture results in a reduced propensity for direct replication, owing to the image generation model's inherent capability for creative output.

T2V Model trained from Scratch. In this section, we evaluate the T2V model trained on a large dataset, with no T2I backbone. For our experiment, we selected the text-to-video model NUWA-XL [48]. While the model itself is not publicly available, videos generated by it are available. Notably, NUWA-XL has been trained on the episodes of “The Flintstones". Given that “The Flintstones" consists of 166 episodes, each approximately 25 minutes long, this constitutes a sizeable dataset for analysis. Our observations revealed instances of replication in NUWA-XL, as illustrated in Figure 5. While these occurrences of replication are less frequent than (47% average top VSSCD) in unconditional models, they underscore that text-to-video models are not entirely immune to replication phenomena.

5.2 Data Requirements for Unique Content: Image vs. Video Diffusion Models

In this section, we draw comparisons between sample replication issues in video diffusion models and their image generation counterparts. We initiate the process by training an image generation diffusion model [14] using individual frames extracted from videos, identifying instances where the model generates unique images. Subsequently, we apply a similar training regimen to a video diffusion model [15] and identify unique video outputs. Our primary focus is to investigate the data requirements for each model type to produce original results. Through our experiments, we find that the image generation model is capable of generating unique samples with as few as 1,000 data points. In contrast, this same amount of data proves to be insufficient for the video model to achieve a similar level of unique output generation. This is illustrated in Figure 6.

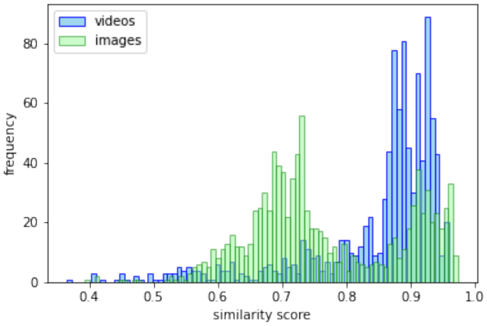

Fig. 6: The histogram illustrates the similarity scores between the generated images or videos and their corresponding training data. Specifically, the similarity score for the generated image samples from the image-based diffusion model is determined by the cosine similarity between the SSCD features of the generated image and the best-matched frame used during model training. Similarly, for samples generated from the video-based diffusion model, the calculation of similarity scores follows the procedure outlined in Section 4.1.

Observation. When training from scratch, the amount of data needed for generating unique videos is significantly greater than that required for the image generation models.

6 Mitigating Video Replication: Recommended Protocols

Video diffusion models mentioned in this paper can differ significantly in their training approaches, architectures, and dataset sizes. Crafting a customized solution for each model type may be unrealistic and exceeds the scope of this study. Instead, we present a set of guidelines aimed at evaluating video diffusion models, particularly emphasizing the assessment of their ability to replicate outcomes. Furthermore, we suggest strategies for training these models on small datasets, which is particularly relevant for minimizing issues of replication in scenarios where resources are limited.

6.1 The Integrated FVD-VSSCD Curve.

The Fréchet Video Distance (FVD) [42] is a commonly employed metric for evaluating video generation models. FVD operates by extracting features from both generated and real videos using a pre-trained feature extractor, typically an I3D model [6] trained on the Kinetics-400 dataset [19]. It then calculates the mean and covariance of these feature distributions for both real and generated videos and computes the distance between them [9]. Ideally, a smaller distance signifies better performance. However, there is an inherent flaw in this metric – if the generated videos are exact replicas of the training data, the distance diminishes, inadvertently rewarding exact replication rather than novel content generation. We can also see this in Table 2, where a high VSSCD score, indicating greater similarity to the training set, corresponds to a better FVD.

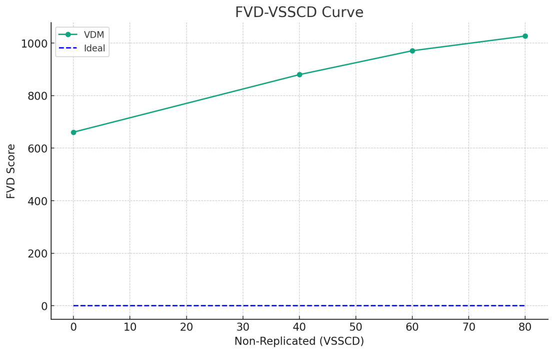

Fig. 7: A representation of the FVD-VSSCD Curve. The y-axis represents the FVD score, which measures the quality of video frames. The x-axis indicates the percentage of non-replicated samples, which reflects the proportion of the generated videos that have been filtered out based on a certain threshold of the VSSCD score (0.6).

To address this limitation, we suggest complementing the FVD results with the VSSCD scores. Our approach involves not just reporting the FVD scores but also recalculating them after excluding the generated videos that are replicates of the training content. Specifically, we remove the generated videos exhibiting various degrees of similarity to the training data and then recalculate the FVD. In an ideal scenario, the resulting graph of the FVD scores versus the percentage of removed replicated samples would display a consistent trend. This would indicate that the videos are not only realistic but also distinct from the training data. We present this analysis in Figure 7, illustrating the relationship between the FVD and replication rates.

6.2 Utilizing Text-to-Image Backbones

In recent methods, several architectures have incorporated a Text-to-Image (T2I) foundation and augmented it with additional temporal layers, enabling them to generate videos [3, 23, 45]. In this approach, the image generation model serves as a foundation for spatial context creation, while added temporal layers focus on learning motion dynamics. Models developed using this method don't rely exclusively on limited video datasets to learn both content and motion, potentially reducing their vulnerability to sample replication. However, it is worth noting that these backbone models, often variants of Stable diffusion [33] for text-to-image tasks, are not immune to replicating sam-

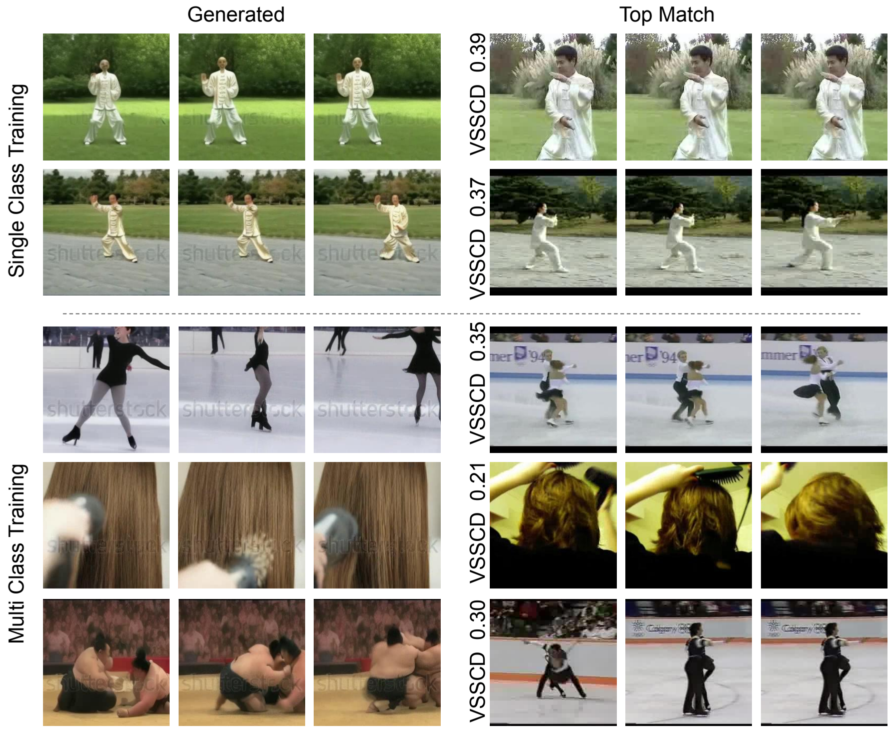

Fig. 8: We fine-tuned the temporal layers of a video diffusion model [45] using two different scenarios. In the first scenario, we trained the model with approximately 100 videos belonging to a single class. In the second scenario, we conducted multi-class training, utilizing around 1,000 different videos across 101 distinct classes. Notably, we observed that in both cases, the similarity between the generated videos and the training data was remarkably low, indicating that the model successfully produced unique video content.

ples from their own image training data [36, 37]. While acknowledging this limitation, our paper does not delve into that specific aspect. Instead, we concentrate on instances of sample replication within the generated videos themselves, where either content or motion is directly mirrored from the video training data in text-to-video models. We utilized Stable Diffusion [33], and expanded it with additional temporal layers [45]. We then trained the model with UCF-101 [40], which resulted in 47% average top VSSCD score, proving to generate less replicated results.

6.3 Fine-tune only Temporal Layers.

As observed earlier, video diffusion models replicate existing content when trained on a small dataset. However, we found that this tendency is significantly reduced when only the temporal layers of the models are fine-tuned with video data. In our experiment, we specifically utilized Modelscope's video generation model [45] and fine-tuned its temporal layer using both the full UCF-101 dataset [40] and its various individual classes.

Notably, even with training on a relatively small sample size of around 100 examples, the model successfully steered clear of reproducing any content from the video training dataset, which is evident in Figure 8. A comparison of average top VSSCD scores can be seen in Figure 9. This success is largely due to the influence of the image generation backbone within the model, which directs the generation of new content. Based on these findings, we advocate for a similar fine-tuning approach, particularly when working with smaller video datasets.

Observation. Leveraging a Text-to-Image (T2I) backbone in video diffusion models demonstrates an enhanced ability to produce unique video content. Fine-tuning only the temporal layers of a pre-trained video diffusion model on a small dataset can effectively address the replication issue in low-resource settings.

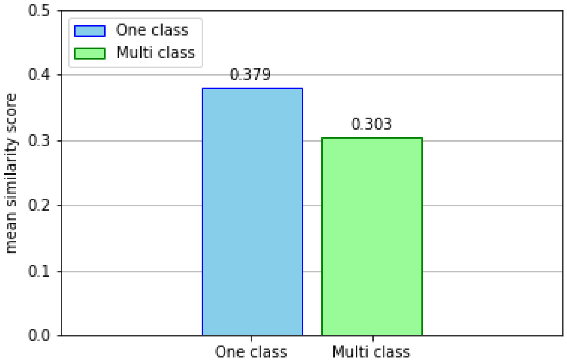

Fig. 9: Average highest VSSCD scores for video generation models trained on a single class versus all classes in the UCF-101 [40] dataset. Both values fall significantly under 0.50, indicating that replication is not evident in the generated contents.

7 Video Replication in SOTA video generation models

Video sample replication is a significant challenge in state-of-the-art models, especially when models and their training datasets are not publicly accessible. In our research, we typically analyze generated videos from project websites and compare them to the closest matches in their training data. This becomes more complicated when the training datasets themselves are not available. In this analysis, we focus on the VideoFusion [23] model, a recent state-of-the-art example where the generated videos are not accessible. To address this, we use screenshots from the model's research paper, representing the generated videos, and match them with the training dataset. Our findings, as illustrated in Figure 10, reveal that even the latest models are susceptible to replicating videos from their training data.

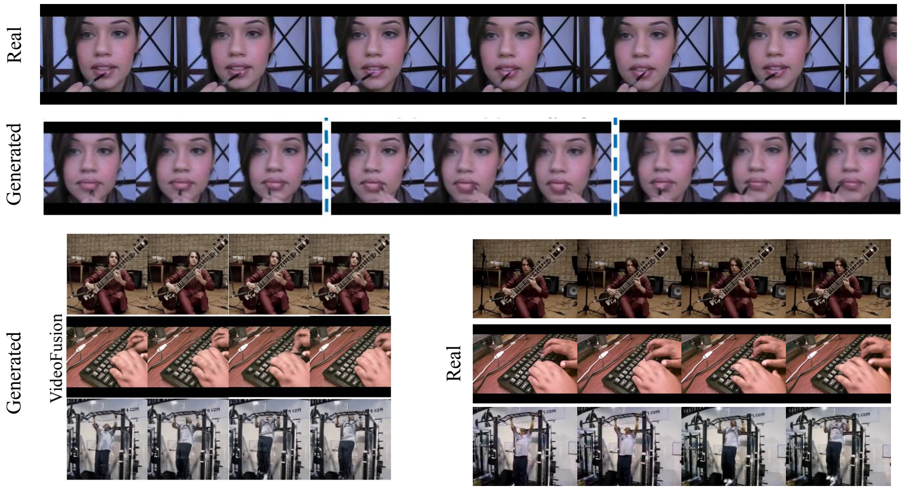

Fig. 10: The highest similarities identified within the UCF-101 dataset compared against outputs generated by an unconditional video generation approach, utilizing VideoFusion [23]. The generated videos are sourced from the research paper.

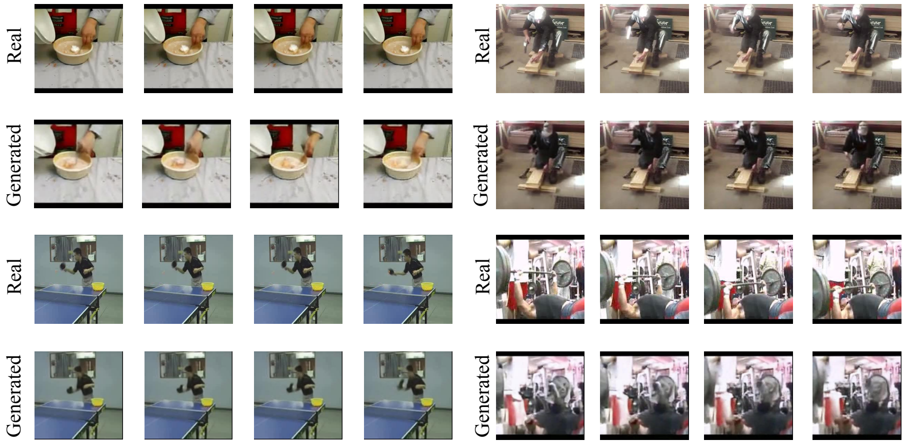

Fig. 11: The highest similarities identified within the UCF-101 dataset compared against outputs generated by an unconditional video generation approach, utilizing LVDM [12]. The generated videos are sourced from the project website.

8 More Qualitative Examples of Video Replication

In this section, we present additional qualitative examples of video replication from various video diffusion models. We examine two distinct types of models: a general video diffusion model operating in the pixel domain [25] and a latent diffusion model [12].

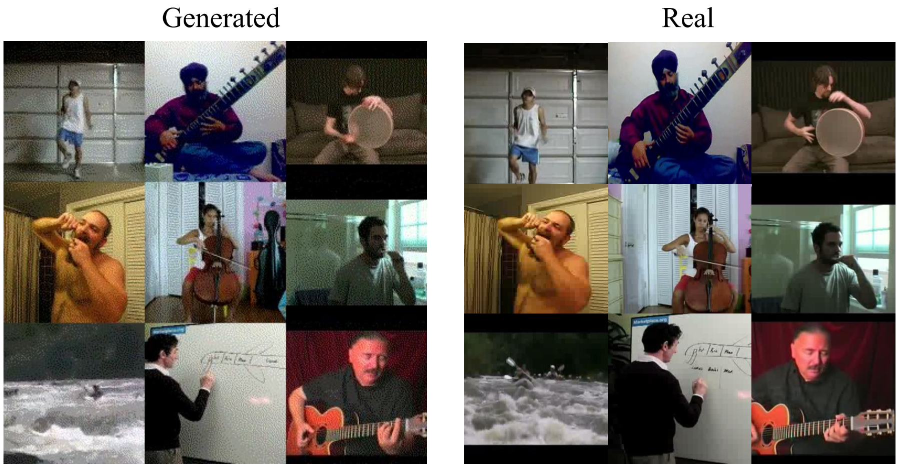

Fig. 12: The highest similarities identified within the UCF-101 dataset compared against outputs generated by an unconditional video generation approach, utilizing VIDM [25]. The generated videos are sourced from the project website.

Figures 12 and 11 provide examples of replication observed in both cases, showcasing the phenomenon across different model architectures.

8.1 Discussion on Replication in Video Diffusion Models

Video diffusion models demonstrate a higher susceptibility to replication compared to image diffusion models, making the originality of generated videos a relatively unexplored area. This raises important questions about the extent to which these models can produce original content. Replication tendencies are evident in both short and long-form videos generated by these models. We propose that if the generated videos lack realism (as seen in models like MakeAVideo [35], LVDM T2V [12], etc.), they are less likely to be replicas. This observation suggests a shift in the focus of current research in this field.

9 Conclusion & Future Work

In our study, we have conducted an in-depth examination of content and motion replication within video generation models. To the best of our knowledge, this is the first comprehensive analysis of replication in the context of video diffusion models. Our focus thus far has centered on the replication of both motion and content aspects. Our future research endeavors will pivot toward exploring the replication of motion across varying content. This involves the application of motion patterns, derived from training data, onto new content scenarios. Such an approach bears significance, particularly in contexts where motion patterns can be as distinctive as biometrics, posing potential risks.

Additionally, we aim to delve more deeper into models trained on extensive datasets. This is motivated by the precedent set by similar large-scale image models, which have demonstrated a tendency for replication. This future line of research promises to enrich our understanding of the nuances and capabilities inherent in video generation technology.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021) 10
2. Baraldi, L., Douze, M., Cucchiara, R., Jegou, H.: Lamv: Learning to align and match videos with kernelized temporal layers. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7804–7813 (2018). https://doi.org/10.1109/CVPR.2018.00814 4
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023) 1, 4, 10, 12
4. Cai, Y., Yang, L., Ping, W., Wang, F., Mei, T., Hua, X.S., Li, S.: Million-scale near-duplicate video retrieval system. In: Proceedings of the 19th ACM International Conference on Multimedia. p. 837–838. MM '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2072298.2072484, https://doi.org/10.1145/2072298.2072484 4
5. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5253–5270 (2023) 3
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 12
7. Chou, C.L., Chen, H.T., Lee, S.Y.: Pattern-based near-duplicate video retrieval and localization on web-scale videos. IEEE Transactions on Multimedia **17**(3), 382–395 (2015). https://doi.org/10.1109/TMM.2015.2391674 4
8. Douze, M., Jégou, H., Schmid, C., Pérez, P.: Compact video description for copy detection with precise temporal alignment. In: Proceedings of the 11th European Conference on Computer Vision: Part I. p. 522–535. ECCV'10, Springer-Verlag, Berlin, Heidelberg (2010) 4
9. Eiter, T., Mannila, H.: Computing discrete fréchet distance (1994) 12
10. Feng, Q., Guo, C., Benitez-Quiroz, F., Martinez, A.M.: When do gans replicate? on the choice of dataset size. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6701–6710 (2021) 3
11. He, X., Pan, Y., Tang, M., Lv, Y., Peng, Y.: Learn from unlabeled videos for near-duplicate video retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1002–1011. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3532010, https://doi.org/10.1145/3477495.3532010 4
12. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022) 2, 6, 7, 15, 16

13. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) 3
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 1, 3, 11
15. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022) 1, 3, 4, 6, 7, 11
16. Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., Dittadi, A.: Diffusion models for video prediction and infilling. arXiv preprint arXiv:2206.07696 (2022) 2, 3, 4, 6, 7, 8, 9
17. Kale, A., Sundaresan, A., Rajagopalan, A., Cuntoor, N.P., Roy-Chowdhury, A.K., Kruger, V., Chellappa, R.: Identification of humans using gait. IEEE Transactions on image processing **13**(9), 1163–1173 (2004) 2
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 3
19. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 7, 9, 12
20. Klempous, R.: Biometric motion identification based on motion capture. 243 (2009) 2
21. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, I.: Visil: Fine-grained spatio-temporal video similarity learning. pp. 6350–6359 (10 2019). https://doi.org/10.1109/ICCV.2019.00645 4
22. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, Y.: Near-duplicate video retrieval with deep metric learning. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 347–356 (2017). https://doi.org/10.1109/ICCVW.2017.49 4
23. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10209–10218 (2023) 3, 4, 8, 10, 12, 14, 15
24. Meehan, C., Chaudhuri, K., Dasgupta, S.: A non-parametric test to detect data-copying in generative models. In: International Conference on Artificial Intelligence and Statistics (2020) 3
25. Mei, K., Patel, V.: Vidm: Video implicit diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 9117–9125 (2023) 1, 2, 4, 6, 7, 8, 15, 16
26. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329 (2023) 4
27. Nair, V., Guo, W., Mattern, J., Wang, R., O'Brien, J.F., Rosenberg, L., Song, D.: Unique identification of 50,000+ virtual reality users from head & hand motion data. arXiv preprint arXiv:2302.08927 (2023) 2
28. Nair, V., Rosenberg, L., O'Brien, J.F., Song, D.: Truth in motion: The unprecedented risks and opportunities of extended reality motion data. arXiv preprint arXiv:2306.06459 (2023) 2
29. Nixon, M.S., Tan, T., Chellappa, R.: Human identification based on gait, vol. 4. Springer Science & Business Media (2010) 2
30. Perera, M.V., Patel, V.M.: Analyzing bias in diffusion-based face generation models. arXiv preprint arXiv:2305.06402 (2023) 3
31. Pizzi, E., Roy, S.D., Ravindra, S.N., Goyal, P., Douze, M.: A self-supervised descriptor for image copy detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14532–14542 (2022) 4, 6

32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685 (2022). https://doi.org/10.1109/CVPR52688.2022.01042 12, 13
34. Shao, J., Wen, X., Zhao, B., Xue, X.: Temporal context aggregation for video retrieval with contrastive learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3268–3278 (January 2021) 4
35. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022) 16
36. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6048–6058 (2023) 1, 2, 3, 5, 8, 13
37. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Understanding and mitigating copying in diffusion models. arXiv preprint arXiv:2305.20086 (2023) 1, 2, 3, 8, 13
38. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 1, 3
39. Song, J., Yang, Y., Huang, Z., Shen, H.T., Hong, R.: Multiple feature hashing for real-time large scale near-duplicate video retrieval. In: Proceedings of the 19th ACM International Conference on Multimedia. p. 423–432. MM ’11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2072298.2072354, https://doi.org/10.1145/2072298.2072354 4
40. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 7, 8, 9, 13, 14
41. Tan, H.K., Ngo, C.W., Hong, R., Chua, T.S.: Scalable detection of partial near-duplicate videos by visual-temporal consistency. In: Proceedings of the 17th ACM International Conference on Multimedia. p. 145–154. MM ’09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1631272.1631295, https://doi.org/10.1145/1631272.1631295 4
42. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019) 8, 11
43. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022) 1, 4
44. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in Neural Information Processing Systems **35**, 23371–23385 (2022) 3, 4, 7, 8
45. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023) 10, 12, 13
46. Wang, K.H., Cheng, C.C., Chen, Y.L., Song, Y., Lai, S.H.: Attention-based deep metric learning for near-duplicate video retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5360–5367 (2021). https://doi.org/10.1109/ICPR48806.2021.9412710 4
47. Webster, R., Rabin, J., Simon, L., Jurie, F.: This person (probably) exists. identity membership attacks against gan generated faces. arXiv preprint arXiv:2107.06018 (2021) 3

48. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. arXiv preprint arXiv:2303.12346 (2023) 4, 10
49. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023) 10
50. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022) 10