Turns Out I'm Not Real: Towards Robust Detection of AI-Generated Videos

Qingyuan Liu, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, Junfeng Yang
Columbia University

{ql2505,ps3391}@columbia.edu   {yunyuntsai,mcz,junfeng}@cs.columbia.edu

Abstract

The impressive achievements of generative models in creating high-quality videos have raised concerns about digital integrity and privacy vulnerabilities. Recent works to combat Deepfakes videos have developed detectors that are highly accurate at identifying GAN-generated samples. However, the robustness of these detectors on diffusion-generated videos generated from video creation tools (e.g., SORA by OpenAI, Runway Gen-2, and Pika, etc.) is still unexplored. In this paper, we propose a novel framework for detecting videos synthesized from multiple state-of-the-art (SOTA) generative models, such as Stable Video Diffusion. We find that the SOTA methods for detecting diffusion-generated images lack robustness in identifying diffusion-generated videos. Our analysis reveals that the effectiveness of these detectors diminishes when applied to out-of-domain videos, primarily because they struggle to track the temporal features and dynamic variations between frames. To address the above-mentioned challenge, we collect a new benchmark video dataset for diffusion-generated videos using SOTA video creation tools. We extract representation within explicit knowledge from the diffusion model for video frames and train our detector with a CNN + LSTM architecture. The evaluation shows that our framework can well capture the temporal features between frames, achieves 93.7% detection accuracy for in-domain videos, and improves the accuracy of out-domain videos by up to 16 points.

1. Introduction

The realm of video creation is undergoing a significant transformation with the advent of video generation tools, such as Stable Video Diffusion [5], SORA by OpenAI [6], RunwayML [3], Pika [2], and Show-1 [23]. These cutting-edge tools are revolutionizing industries from design, marketing, and entertainment to education by creating high-quality video content. The pivotal shift is opening up a myriad of possibilities for creators everywhere, yet poses societal dangers, notably in their widespread use of spreading disinformation, propaganda, scams, and phishing – evidenced by cases like the Taylor Swift deepfakes [4]. The poten-
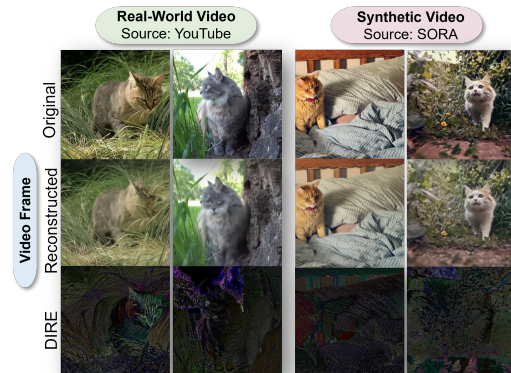

Figure 1. We show the real video frames from YouTube, and fake video from SORA by OpenAI. The explicit knowledge, DIRE, is calculated by the difference between the input original frame and the reconstructed frame from the diffusion model. The reconstructed frame for the SORA-generated video will be visually close to the input original frame, yet the real video from YouTube can not (e.g., the cat face distorted after reconstruction), which inspired us to leverage the DIRE information for training.

tial threats underscore the importance of detecting video generated by these generative models.

Recent advancements have developed detectors that achieve remarkable accuracy in identifying images generated by diffusion-based models [13, 22]. The core method first calculates the Diffusion Reconstruction Error (DIRE) by measuring the difference between an input image and its corresponding reconstructed version from the diffusion model; then, a classifier is trained on the DIRE values to distinguish the AI-generated images from human-created images. Prior work [22] shows that these State-of-the-art detectors have great generalizability, where the detector trained with samples generated from one source of the generative model can also detect samples from other sources. However, we found that these state-of-the-art detectors are not robust enough to detect diffusion-generated video under the same setting due to the following reasons, (1) The detector can easily overfit on the in-domain set when training with DIRE values only, which can not generalize to the out-domain set. (2) The SOTA diffusion-generated image detectors can not

Step 1: Extract explicit knowledge

① Video Frame → Diffusion Model → Reconst. Frame → ② DIRE Frame

Step 2: Train detector

① Video Frame
② DIRE Frame
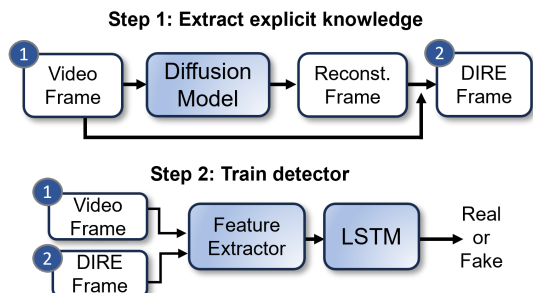→ Feature Extractor → LSTM → Real or Fake

Figure 2. The flow of DIVID. In step 1, given a sequence of video frames, we first generate the reconstructed version of every frame by using the diffusion model. Then, we calculate the DIRE values using the reconstructed frame and their corresponding input frame. In step 2, the CNN+LSTM detector is trained based on sequences of DIRE values and the original RGB frames.

capture the temporal information in multiple video frames. (3) Existing video detection methods for deepfake are not designed for high-quality diffusion-generated video with more variance, which requires more explicit knowledge to detect.

To tackle the challenges, we propose a novel approach for *DIffusion-generated VIdeo Detection*, called **DIVID**. Our method, **DIVID**, carefully investigates the sampling timestep of the diffusion process to generate DIRE values upon multiple frames for real and fake video. We found the tradeoff between the quality of video reconstruction and the proper DIRE values for training the detector. While the sampling step is small, the reconstruction of video frames from the diffusion model can be noisy, and the DIRE values can be large for both real and fake videos. Despite increasing the sampling timestep to a large number, which can improve the video quality after reconstruction from diffusion, it reduces the gap of DIRE values between real and fake. Therefore, a proper sampling timestep is critical for video detection.

We then propose using CNN + LSTM architectures to capture different levels of abstraction features and temporal dependencies in both original RGB frames and DIRE values, as shown in Fig 2. Prior work [22] uses a pre-trained CNN architecture to build the detector, yet it can learn only the features of DIRE values on every single image. To incorporate the temporal features into the model, we change the model architectures and do the two-phase training. We first fine-tune the CNN on the original RGB frame and DIRE values of video, then train the LSTM network based on the feature extractor in the CNN.

Our main contributions are as follows.

- We propose DIffusion-generated VIdeo Detector (DIVID), a new video detection method that can detect video generated from multiple sources of video generation tools.
- We have collected a video dataset, including in-domain trainset / testset, and out-domain testsets. The in-domain train/test sets are from Stable Video Diffusion (SVD) model, and the out-domain is collected from Pika, Runway

Gen-2, and SORA. We will release our benchmark shortly.
- We propose a general training framework by incorporating the temporal information meaningful explicit knowledge in video clips. We observe a simple CNN + LSTM trained with RGB frame + DIRE values can improve the generalizability of detectors on out-domain testsets.

2. Related Works

● **Diffusion-generated Video** Diffusion-based video generation represents a leap forward from static image generation, addressing the complexities of temporal coherence, motion dynamics, and environment consistency. Tools like SORA by OpenAI [6], Stable Video Diffusion [5], MidJourney [1], RunwayML [3], Show-1 [23], Pika [2] and Deep Dream Generator [15] enable users to generate videos with impressive visual and narrative quality. These tools illustrate the breadth of development in AI-driven video content creation, showcasing a range of capabilities from enhancing video quality to generating entirely new content. Each offers unique features tailored to specific creative or technical needs, reflecting the rapid evolution and growing accessibility of video generation technology.

● **Generative Video Detection** The success in generating high-quality images has heightened concerns about security, personal privacy, and digital integrity, emphasizing the need for a robust detector to discern whether samples are from generative models. Recently, Deepfake video, generated by GAN-based models, can perform face manipulation with high realism [11]. Agarwal et al. [11] point out the challenges of detecting deepfake video, where the traditional DNN networks or audio-visual approach based on lipsync inconsistency detection are not robust enough to detect Deepfake. Rössler et al. [18] introduced a face forgery detection technique that begins by tracking and extracting facial information from the sample, followed by training a classifier to detect forgeries. Marra et al. [14] developed multiple CNN-based models for detecting fake images. David et al., [9] proposed to use CNN + LSTM to do the deepfake video detection. However, their approach did not account for cross-model transferability and was found ineffective for generalizing to diffusion-based images [7, 16]. Lorenz et al. [12] modified the widely-used Local Intrinsic Dimensionality (LID) to detect diffusion-based images. Based on the observation that generated images vary less than real images after reconstruction, DIRE [22] proposed to utilize reconstructed error to detect diffusion-generated images and was proved better than traditional RGB-based detectors. Se-DID [13] extended it by ensembling step-wise noise error of corresponding inverse and reverse stages. To facilitate diffusion image detection, benchmarks at image level [22] and fine-grained region level [21] were established for fair comparison. Despite the prior works, the robustness of detectors on diffusion-generated video remains unexplored.

3. Method

3.1. Preliminaries

• **Denoising Diffusion Probabilistic Models (DDPM)** The diffusion model is a novel approach for image generation that was first proposed in [20] inspired by nonequilibrium thermodynamics, strengthening the image quality during the generation process. DDPM operates by gradually transforming random noise into structured images over a series of steps, simulating a reverse diffusion process.

Given an image x_0 sampled from $\mathcal{X}_S$, DDPM relies on learning the reverse of a diffusion process, where a data point x_0 is progressively noised until it becomes a Gaussian noise through a fixed Markov Chain for T steps, and then learning to denoise it back to the original image. Specifically, for every timestep $t \in [1, ..., T]$, they sample data sequence $[x_1, ..., x_T]$ by adding Gaussian noise with variance $\beta_t \in (0, 1)$ during the forward process, which is defined as:

$$q(x_t|x_0) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \,, \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is the noise sampeled from Gaussian distribution, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \Pi_{s=1}^t \alpha_s$. The reverse process then generates a sequence of denoised image $[x_t^g, x_{t-1}^g ..., x_0^g]$ from timestep $t \in [T, ...1]$. For timestep t in the reverse process, the denoised image can be defined as:

$$x_{t-1}^g = \frac{1}{\sqrt{\alpha_t}}\left(x_t^g - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(x_t^g, t)\right) + \sigma_t\epsilon \,, \quad (2)$$

where ϵ_θ is a trainable noise predictor that generates a prediction for the noise at the current timestep and removes the noise. σ_t is the variance of noise. Ideally, the generated sample x_0^g should be moved forward to the distribution of the source domain trained for the diffusion model.

• **Denoising Diffusion Implicit Models (DDIM)** In DDPM, the long stochastic operations can lead to gradient vanishing during the training and huge distortion of the content information. Thus, to train the diffusion more efficiently without content distortion, DDIM [20] modifies the diffusion process by introducing a non-Markovian implicit trajectory. The reverse process in DDIM can be defined as:

$$x_{t-1}^g = \sqrt{\bar{\alpha}_{t-1}}\left(x_t^g - x_{0,t}^g\right) \quad (3)$$
$$+ \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(x_t^g, t) + \sigma_t\epsilon \,,$$

where $x_{0,t}^g$ is the predicted denoised image for x_0 conditioned on x_t^g at the time step t and is defined as:

$$x_{0,t}^g = \frac{x_t^g - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t^g, t)}{\sqrt{\bar{\alpha}_t}}, \quad (4)$$

While $\sigma_t = 0$, the noise term ϵ_θ is ignored, and the sampling process becomes deterministic. When $\sigma_t = \sqrt{1 - \bar{\alpha}_{t-1}/(1 - \bar{\alpha}_t)}\sqrt{1 - \bar{\alpha}_t/\bar{\alpha}_{t-1}}$, the generative process becomes DDPM.

3.2. Diffusion-generated Video Detector (DIVID)

• **Diffusion Reconstruction Error (DIRE)** Previous studies indicate that image detectors designed for GAN or VAE models experience performance drops when identifying images generated by diffusion processes. Moreover, training a binary classifier solely on real and diffusion-generated images does not ensure it will effectively recognize new, unseen diffusion-generated images. In Wang et al., [22], they proposed the DIRE, which can well capture the signal of the diffusion-generated image. Their hypothesis posits that images from diffusion models are sampled from the diffusion process distribution, suggesting that reconstructed diffusion-generated images should closely resemble each other. For any image x_0 from any set $\mathcal{X}$ and a pre-trained diffusion model, the DIRE is calculated by

$$DIRE(x_0) = |x_0 - \mathbf{R}(\mathbf{I}(x_0))| \quad (5)$$

where $\mathbf{I}$ represents the inversion process and $\mathbf{R}$ represents the reconstruction process. The output of DIRE is calculated by the absolute difference between x_0 and its corresponding reconstructed version. DIVID leverages the DIRE to detect the diffusion-generated videos.

• **Detector for Diffusion-generated Videos** To build a robust detector for detecting AI-generated videos, instead of training a binary CNN classifier with video frames, we propose a CNN+LSTM trained with RGB frames + DIRE values. The DIRE values are extracted from an unconditional diffusion model. To well learn the temporal features of RGB frames+DIRE values on every time step, we first train a CNN detector for extracting the frame features on single frames. Then, the RGB frames and DIRE values will be leveraged to train CNN + LSTM model jointly.

Assume we already extract the DIRE of the video frame set from a video sample Q, denoted as $\{q_1, q_2, ..., q_T\}$ and T is the total time step of frame. The ϕ is the feature extractor of CNN model. The feature of DIRE set from ϕ is $\{\phi(q_1), \phi(q_2), ..., \phi(q_T)\}$. For a feature $\phi(q_t)$ of DIRE frame q_t on time step $t \in [1, 2, ..., T]$, the working of LSTM in DIVID can be described as follows. The LSTM generated hidden representation a^t at every time step in the LSTM cells for output prediction (Eq. 8), which inquires the input feature frame $\phi(q_t)$ and the inherent information of hidden representation a^{t-1} from the previous step. The cell state c^t in Eq. 7 memorizes the past cell state and helps retain the information from the past for a^t. f^t, u^t, and o^t are the forgetting, updating, and output gate layers. We initially generate a candidate cell state $c^{\sim t}$ using Eq. 6, where W_a^c and W_x^c are the weight parameters.

$$c^{\sim t} = \tanh(W_a^c a^{t-1} + W_x^c \phi(q_t)) \quad (6)$$
$$c^t = f^t \cdot c^{t-1} + u^t \cdot c^{\sim t} \quad (7)$$
$$a^t = o^t \cdot \tanh(c^t) \quad (8)$$

	Detector Architecture	Evaluation Metrics	
		Acc.	AP
RGB	CNN	90.16	97.02
RGB	CNN+LSTM	90.16	97.39
DIRE [22]	CNN	92.74	97.46
DIVID/ DIRE only	CNN+LSTM	**93.68**	97.66
DIVID/ DIRE + RGB	CNN+LSTM	91.33	**98.20**

Table 1. Detection performance on the in-domain testset. DIVID outperforms our baseline architectures regarding accuracy (Acc.) and average precision (AP). RGB represents the original pixel frame values from raw video.

	Model	Out-domain			Total Avg.
		Gen-2	Pika	SORA	
RGB	CNN	65.42	78.04	60.05	67.84
RGB	CNN+LSTM	**67.76**	84.11	60.80	70.89
DIRE [22]	CNN	50.93	60.75	54.77	55.48
DIVID / DIRE only	CNN+LSTM	60.75	80.37	60.8	67.3
DIVID / DIRE + RGB	CNN+LSTM	66.82	**86.92**	**61.01**	**71.58**

Table 2. Detection performance on out-domain testsets. We compare DIVID with three baselines and show the detection accuracies on SORA [6], Pika [2], and Gen-2 [3].

By training the CNN+LSTM with DIRE+RGB frame features, DIVID improves the detection accuracy of video from both in-domain and out-domain.

4. Experiment

Experimental Setting • **Dataset** We contruct a dataset for evaluating our method shown in Table 3 by using public video generation tools, including Stable Video Diffusion (SVD) [10], Pika [2], Gen-2 [3], SORA [6]. Our source data are from ImageNet Video Visual Relation Detection (VidVRD) [19], which contains 1k source videos. For fake video generation with SVD, the process starts from randomly cropping each source video into a 25-frame video clip. Then, the first image frame will be sent to SVD to generate a corresponding 25-frame fake video clip. The video resolution is 1025×576. • **Model** We use ADM [8], an unconditional 256×256 diffusion model trained on ImageNet-1K [17], as our reconstruction model to generate the DIRE representation for every video frame. The CNN classifier is a ResNet50 model with pre-trained weight on ImageNet-1K. The LSTM follows a one-layer architecture with hidden size 2048, which can handle multiple types of representation (e.g., DIRE, RGB values of the original frame) extracted from the CNN feature extractor. • **Baseline and Implementation Details** DIVID is trained based on DIRE and original RGB features extracted from video frames with a CNN + LSTM model. We set the training batchsize as 128. In the training of LSTM, we use a 32-consecutive sequence of 4 frames in each batch. • **Evaluation Metrics** We evaluate the detection performance based on the prediction of every frame and calculate accuracy and average precision.

Experimental Results We show the detection performance for in-domain video set in Table 1. The three

	Video Source	Denoising Condition	Generated Model	# of Clips (real/fake)
In-domain	VidVRD [19]	Image2Video	SVD-XT [10]	1k/1k
Out-domain	VidVRD (test)	Image2Video	Pika [2]	107/107
	VidVRD (test)	Image2Video	Gen-2 [3]	107/107
	YouTube/SORA [6]	x	x	207/191

Table 3. Composition of our video dataset. It includes real video clips from VidVRD [19] and fake video clips generated from SVD, Pika, and Gen-2. The fake video generation is conditioned on the video frame from VidVRD. We also collected real videos from YouTube and SORA [6] website as our 3rd out-domain test set based on the same them (e.g., A cat on the bed).
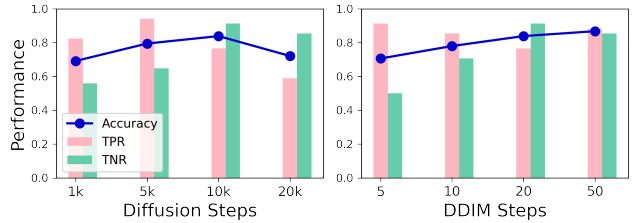

Figure 3. Analysis on diffusion steps and ddim step for DIVID. The left-side subfigure shows the performance on different diffusion steps from 1k to 10k, we freeze the ddim step as 20 for all. The right-side subfigure shows the performance on different ddim steps from 5 to 50, and the diffusion step is fixed as 10k for all.

baselines, including original trained under CNN and CNN+LSTM, and DIRE [22] trained with CNN. Compared with the three baselines, DIVID achieves 98.20% average precision (AP) and has better detection accuracy, and outperforms them by 0.94% to 3.52%. In Table 2, we show the performance of DIVID on three out-domain testsets, including SORA, Pika, and Gen-2. For every baseline, we show the checkpoint results that have the best out-domain detection accuracy. We observe that DIVID has competitive in-domain results as baselines and also achieves better generalizability on out-domain testset. DIVID improves the out-domain average accuracy by 0.69% to 16.1%. In Figure 3. we compare different diffusion steps and DDIM steps for generating the DIRE values in our DIVID.

5. Conclusion

We propose a general framework, DIVID for diffusion-generated video detection. We have collected a new video benchmark dataset, including fake videos generated from SVD, SORA, Pika, and Gen-2. Unlike prior SOTA detectors that only use the DIRE values to train the CNN detector, DIVID leverages both RGB frames and DIRE values with a simple CNN+LSTM, which can capture the temporal information and extract explicit knowledge from multiple video frames. The evaluation shows that DIVID achieves better detection performance on in- and out- domain testsets by up to 3.52 and 16.1 points. Our work highlights the importance of increasing the generalizability of current SOTA detectors.

References

[1] Midjourney. (2023). midjourney (v5.1) [text-to-image model]. https://www.midjourney.com/. 2

[2] Pika labs. (2024) [text-to-video model]. https://www.pika.art/. 1, 2, 4

[3] Runway ml. gen2. (2024) [text-to-video model]. https://research.runwayml.com/. 1, 2, 4

[4] Look what you made me do: Why deepfake taylor swift matters, 2024. 1

[5] 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. 1, 2

[6] 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, 2, 4

[7] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[9] David Güera and Edward J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. 2

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 4

[11] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *CoRR*, abs/1812.08685, 2018. 2

[12] Peter Lorenz, Ricard L. Durall, and Janis Keuper. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 448–459, 2023. 2

[13] RuiPeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023. 1, 2

[14] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389, 2018. 2

[15] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. 2

[16] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes, 2024. 2

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 4

[18] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images, 2019. 2

[19] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM International Conference on Multimedia*, Mountain View, CA USA, 2017. 4

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[21] Sai Wang, Ye Zhu, Ruoyu Wang, Amaya Dharmasiri, Olga Russakovsky, and Yu Wu. Deter: Detecting edited regions for deterring generative manipulations, 2023. 2

[22] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023. 1, 2, 3, 4

[23] 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. *arXiv preprint arXiv:2309.15818*, 2023. 1, 2