

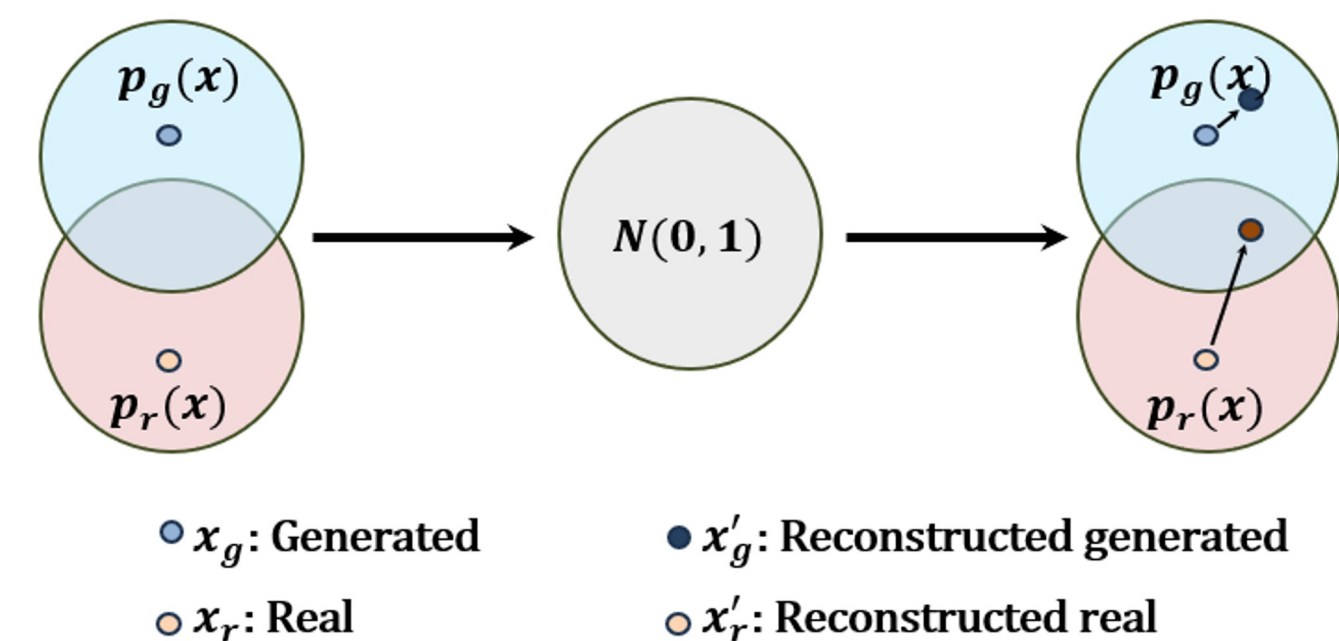
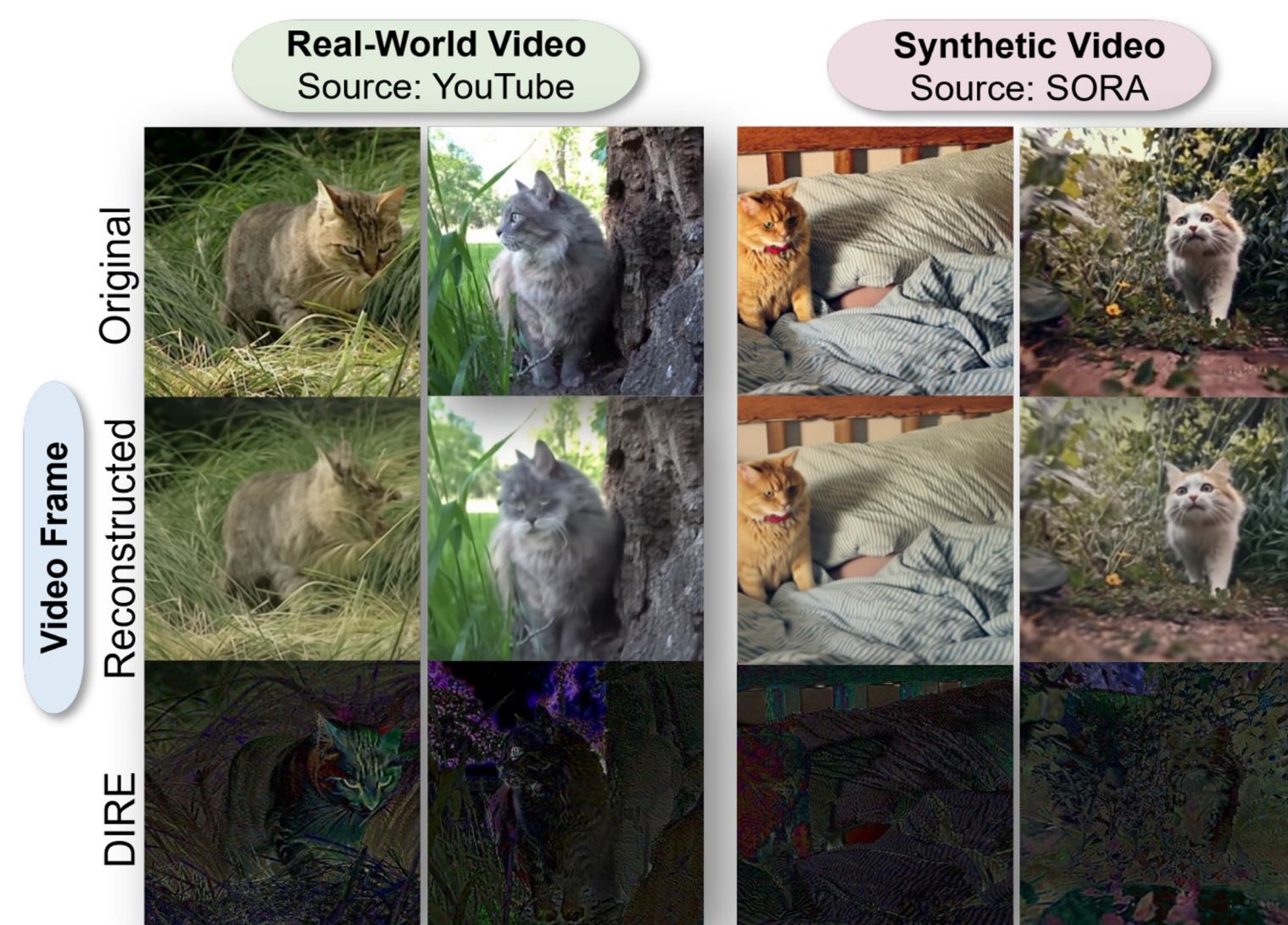
## Motivation & Background

- The impressive achievements of generative models in creating high-quality videos have raised concerns about digital integrity and privacy vulnerabilities
- Recent works have widely studied Deepfakes videos detection, which can identify GAN-generated samples
  - However, Deepfake detectors only focus on detecting the artifact in face features
  - The robustness of them on **diffusion-generated videos** is unexplored



- Diffusion-based video generation represents a leap forward from static image generation, addressing the complexities of temporal coherence, motion dynamics, and environment consistency.
- Diffusion Reconstruction Error (DIRE) can help to distinguish the human-generated videos and diffusion-generated videos.
  - Calculated by the difference between the input original frame and the reconstructed frame from the diffusion model

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

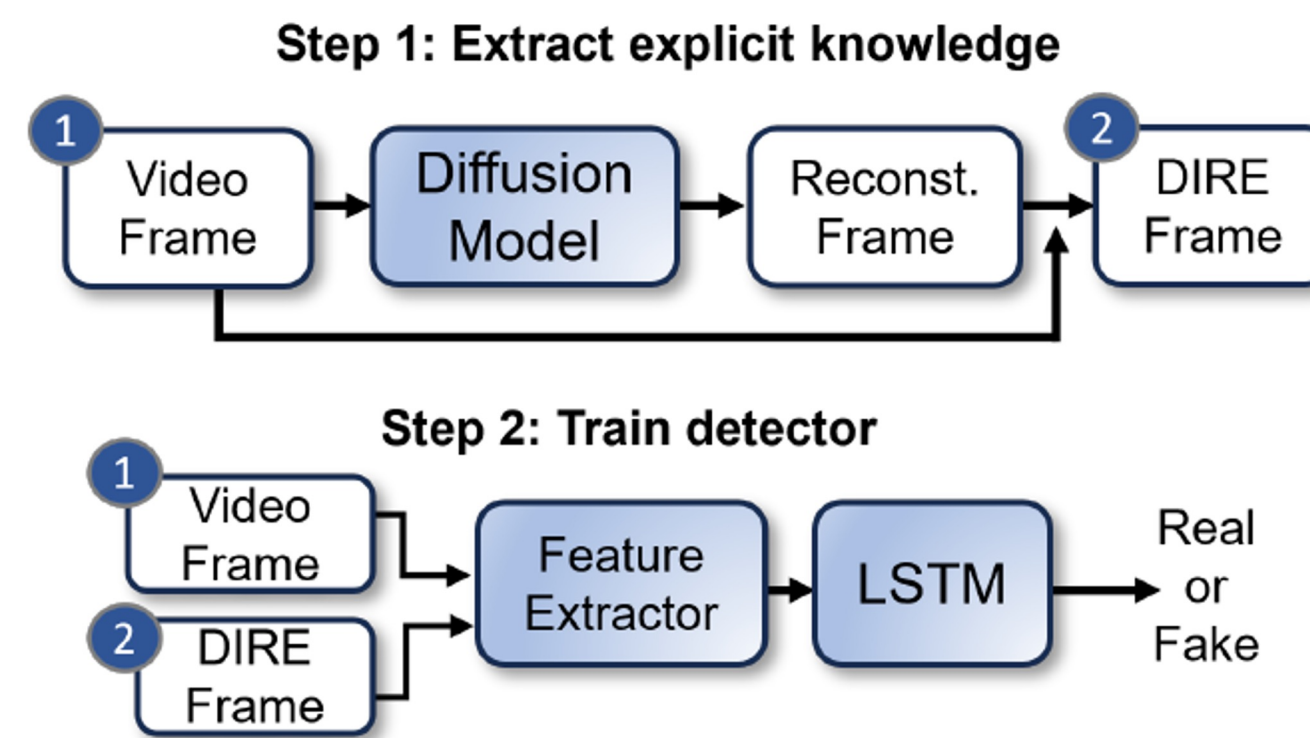


- Images from diffusion models are sampled from the diffusion process distribution, which means reconstructed diffusion-generated images should closely resemble each other.

Figure 1. We show the real video frames from YouTube, and fake video from SORA by OpenAI.

## Methodology

- We propose a novel approach for **Diffusion-generated Video Detection**, called **DIVID**. Our method, **DIVID**, carefully investigates the sampling time step of the diffusion process to generate DIRE values upon multiple frames for real and fake video
- Key insights**
  - We observe the information of DIRE could well capture the difference between real and fake video.
  - The sampling/diffusion step are important to amplify the value of DIRE for larger diffusion/sampling step can increase the value of DIRE.



### The flow of DIVID

- Given a sequence of video frames, we first generate the reconstructed version of every frame
- Calculate the DIRE using the reconstructed frame and their corresponding input frame.
- CNN+LSTM detector is trained based on sequences of DIRE values and the original RGB frames.

## Dataset Collection

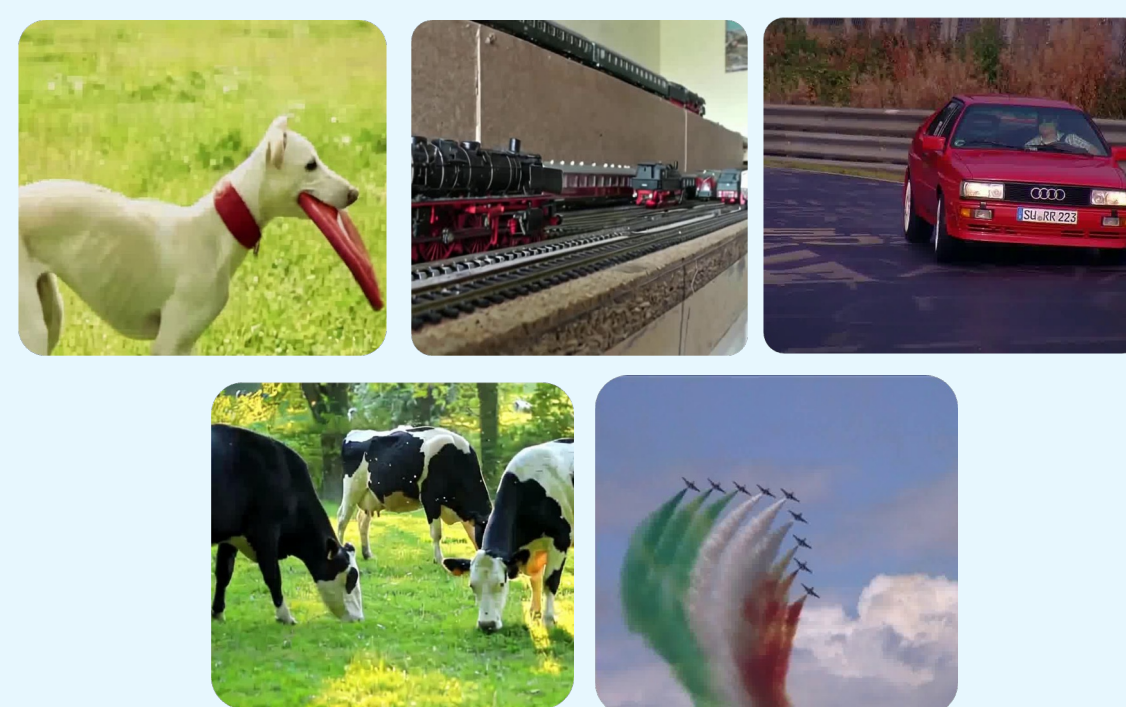
- We collected an AI-generated video dataset, including:
  - In-domain datasets: generated from Stable Video Diffusion (SVD) model.
  - Out-domain datasets: generated from Pika, Runway Gen-2, and SORA.

	Video Source	Denoising Condition	Generated Model	# of Clips (real/fake)
In-domain	VidVRD [18]	Image2Video	SVD-XT [9]	2k/2k
Out-domain	VidVRD (test)	Image2Video	Pika [2]	107/107
	VidVRD (test)	Image2Video	Gen-2 [3]	107/107
	YouTube/SORA [6]	x	x	207/191

### Details of Video dataset.

- Real video clips are collected from VidVRD [4]
- Fake video clips generated from SVD, Pika, and Gen-2
- We also collected real videos from YouTube and SORA [5] website as our 3<sup>rd</sup> out-domain test set based on the same theme (e.g., A cat on the bed).
- Visualization of video dataset samples**
  - Unlike previous deepfake and motion detection datasets, our dataset covers a wider range of topics.

### SVD (In-Domain Sets)



### Out-Domain Sets



## Experiment & Ablation Study

### Experimental Setting

#### Model

- We use ADM [8], an unconditional 256×256 diffusion model trained on ImageNet-1K [17], as our reconstruction model.
- The CNN classifier is a ResNet50 model. The LSTM follows a one-layer architecture with hidden size 2048

#### 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 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%

Detector Architecture		Evaluation Metrics		
		Acc.	AUC	AP
RGB	CNN	90.16	96.78	97.02
RGB	CNN+LSTM	90.16	97.15	97.39
DIRE [21]	CNN	92.74	97.35	97.46
DIVID/ DIRE only	CNN+LSTM	93.68	97.31	97.66
DIVID/ DIRE + RGB	CNN+LSTM	91.33	97.95	98.20

Table 1. Detection performance on the in-domain testset. DIVID outperforms 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 [21]	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

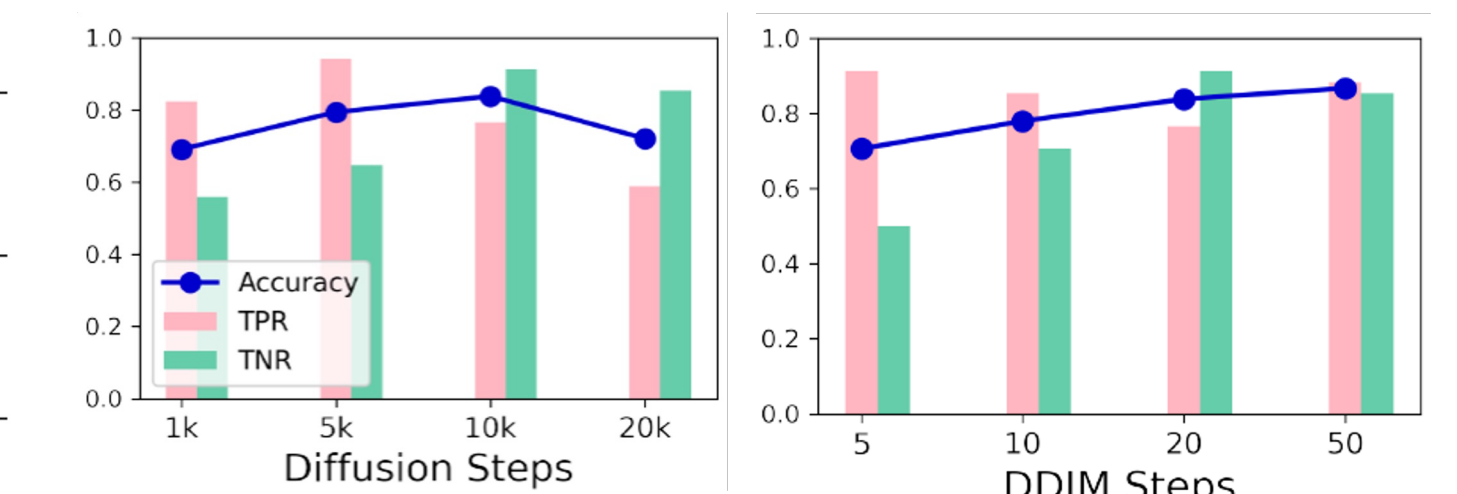


Figure 2. Analysis on diffusion steps and ddim step for DIVID

Table 2. Detection performance on out-domain testsets.

## Conclusion

- We propose a general framework, DIVID for diffusion-generate video detection, which can capture the temporal information and extract explicit knowledge from multiple video frames.
- We collect a novel video benchmark dataset to detect AI generated videos, including real videos, and fake videos generated from SVD, SORA, Pika, and Gen-2.
- Our work highlights the importance of increasing the generalizability of current SOTA detectors.

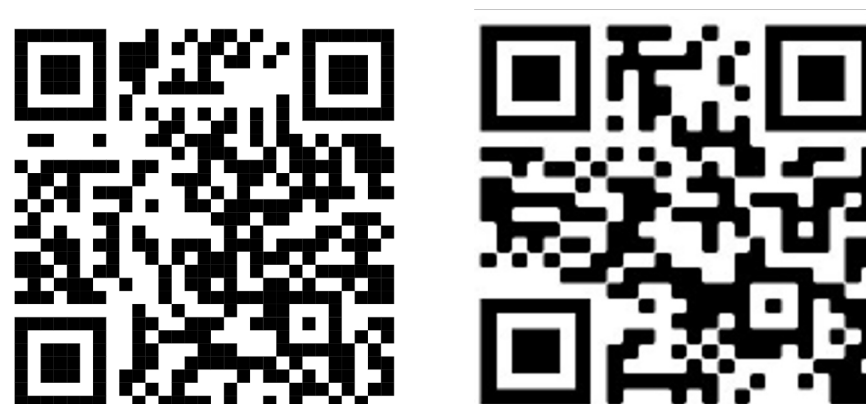
## Future Study

- Explore and improve the video detector architecture
- Investigate efficient way to generate synthetic videos
- Improve generalizability on out-domain video detection

### Reference:

- Pavel Korshunov, et al. "Deepfakes: a new threat to face recognition? assessment and detection"
- Andreas Rossler, et al. "Faceforensics++: Learning to detect manipulated facial images"
- Zhendong Wang, et al. "Dire for diffusion-generated image detection"
- Xindi Shang, et al. "Video visual relation detection"
- Tim Brooks, et al. "Video generation models as world simulators"

## Contact



Paper

Personal Website