DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark

Haoxing Chen[1†], **Yan Hong**[1†], **Zizheng Huang**[1,2], **Zhuoer Xu**[1], **Zhangxuan Gu**[1*], **Yaohui Li**[2], **Jun Lan**[1], **Huijia Zhu**[1], **Jianfu Zhang**[3*], **Weiqiang Wang**[1], **Huaxiong Li**[2]
[1]Ant Group
[2]Nanjing University
[3]Shanghai Jiao Tong University
hx.chen@hotmail.com

Abstract

Recently, video generation techniques have advanced rapidly. Given the popularity of video content on social media platforms, these models intensify concerns about the spread of fake information. Therefore, there is a growing demand for detectors capable of distinguishing between fake AI-generated videos and mitigating the potential harm caused by fake information. However, the lack of large-scale datasets from the most advanced video generators poses a barrier to the development of such detectors. To address this gap, we introduce the first AI-generated video detection dataset, GenVideo. It features the following characteristics: (1) a large volume of videos, including over one million AI-generated and real videos collected; (2) a rich diversity of generated content and methodologies, covering a broad spectrum of video categories and generation techniques. We conducted extensive studies of the dataset and proposed two evaluation methods tailored for real-world-like scenarios to assess the detectors' performance: the cross-generator video classification task assesses the generalizability of trained detectors on generators; the degraded video classification task evaluates the robustness of detectors to handle videos that have degraded in quality during dissemination. Moreover, we introduced a plug-and-play module, named Detail Mamba (DeMamba), designed to enhance the detectors by identifying AI-generated videos through the analysis of inconsistencies in temporal and spatial dimensions. Our extensive experiments demonstrate DeMamba's superior generalizability and robustness on GenVideo compared to existing detectors. We believe that the GenVideo dataset and the DeMamba module will significantly advance the field of AI-generated video detection. Our code and dataset will be aviliable at https://github.com/chenhaoxing/DeMamba.

1 Introduction

Advancements in generative models [Zhang et al., 2023a, Chen et al., 2023a, Li et al., 2023] have been impressive, enabling the creation of highly realistic images with less effort and expertise. As these models become capable of generating sufficiently realistic images, more researchers are exploring how to improve video creation [Blattmann et al., 2023a, Liu et al., 2023a, Wang et al., 2023a, GoogleAI]. Currently, certain generative algorithms, such as Sora [Brooks et al., 2024] and Gen2 [Research, 2023], are capable of producing high-quality videos through the use of straightforward inputs, including text and images. While these generative algorithms can reduce manual labor and enhance creativity, they also introduce risks Barrett et al. [2023]. For example, they could be utilized to misinform the public

*Corresponding author.† Equal contribution.

Preprint. Under review.

in critical domains such as politics or economics. A notable incident involved an AI-generated video of Taylor Swift that spread widely on Twitter, harming her reputation. This situation highlights the pressing need for technology that can detect these fake videos and avoid potential harm.

To assist in developing robust and highly generalizable detectors, we have created the first million-scale dataset of AI-generated videos, named GenVideo. GenVideo leverages state-of-the-art models to generate massive amounts of video, providing comprehensive training and validation for detectors of AI-generated videos. Unlike deepfake video datasets [Gu et al., 2021a, Xu et al., 2023a, Gu et al., 2022a] which focus on human face videos, GenVideo encompasses a broad spectrum of scene contents and motion variations, closely simulating the real-world authentication challenges posed by video generation models in various practical settings. GenVideo includes 1,078,838 generated videos and 1,223,511 real videos. The fake videos consist of those generated in-house and those collected from the internet, while the real videos come from the Youku-mPLUG [Xu et al., 2023b], Kinetics-400 [Kay et al., 2017a], and MSR-VTT [Xu et al., 2016a] datasets. Due to the scale of the data, we can prevent detectors from merely learning the content differences between real and fake videos, instead focusing on subtle signs that determine video authenticity. We propose two tasks that align with real-world detection challenges: (1) cross-generator video classification, where a trained detector is tasked with identifying videos from unseen generators; and (2) degraded video classification, where the detector assesses videos that have been degraded, such as those with low resolution, compression artifacts, or Gaussian blur. GenVideo can significantly advance the development of detectors aimed at identifying AI-generated videos in society.

In this paper, we evaluated state-of-the-art detection models [Qian et al., 2020, Tan et al., 2024, Gu et al., 2021a, Ni et al., 2022, Radford et al., 2021] on GenVideo. However, the generalization capabilities of these models are compromised due to the limitations of existing image detection methods, which cannot model temporal inconsistencies, and video detection methods, which struggle to efficiently model local spatial inconsistencies. In the show "Detail" [2], Kobe Bryant offered insights into basketball nuances, including Jason Tatum's toe positioning when catching the ball, advice that significantly improved Tatum's ability to penetrate defenses. As shown in Figure 1, generated videos often exhibit both spatial and temporal artifacts, and modeling only one aspect (either spatial or temporal) may not be sufficient to cover all types of artifacts. Building a detector with satisfactory generalization performance requires modeling the spatial-temporal local details. In this paper, we introduce a plug-and-play module called Detail Mamba (DeMamba), which leverages a structured state space model to capture spatial-temporal inconsistencies across different regions, thereby discerning the authenticity of videos. Extensive experiments on GenVideo demonstrate that DeMamba can be used as a plug-and-play addition to existing feature extractors, significantly enhancing the generalizability and robustness of models.

Our contributions are summarized as follows:

- We introduce the first million-scale dataset for AI-generated video detection, GenVideo, which includes fake videos from various scenes, contents, and models.
- We design two tasks to evaluate the performance of detectors: cross-generator video classification and degraded video classification.
- We propose a plug-and-play detector, DeMamba, capable of modeling spatial-temporal inconsistencies. Extensive experimental results validate the generalizability and robustness of our DeMamba in identifying AI-generated videos.

2 Related works

2.1 Video generation methods

Video generation methods Henschel et al. [2024], Zhou et al. [2024] have become powerful tools for producing high-quality video content from textual or image prompts. Currently, video generation primarily encompasses two major tasks: Text-to-Video (T2V) and Image-to-Video (I2V). T2V involves inputting a text prompt to the model to generate videos based on textual instructions, while I2V aims to generate videos based on an input image, describing videos' content or specific frames.

[2] https://www.youtube.com/watch?v=NZmEY3n97P4

Figure 1: Spatial and temporal artifacts in generated videos. We illustrate the spatial and temporal artifacts present in the generated videos. Artifacts: (a) errors in local appearance, (b) frequency inconsistency: average spectrum of video frames for real videos and fake videos generated, (c) temporal inconsistency.

Based on the types of these video generation methods, they can be separated into three categories: diffusion-based methods with Unets, diffusion-based methods with Transformers, and other methods.

Diffusion-Unet. In the realm of video synthesis Ma et al. [2024a], Zhang et al. [2024], Bar-Tal et al. [2024], Wei et al. [2023], Ho et al. [2022], Girdhar et al. [2023], Feng et al. [2023], Xu et al. [2023c], Hu et al. [2023], Ni et al. [2023], Girdhar et al. [2023] recent advancements in diffusion-unet based methods have demonstrated significant progress. Text2Video-ZeroKhachatryan et al. [2023] introduces a cost-effective approach, enriching latent codes with motion dynamics for temporal consistency. Animatediff Guo et al. [2023a] introduces a plug-and-play motion module for animating personalized text-to-image models without model-specific tuning, while Pia Zhang et al. [2023b] designs a personalized image animator for motion controllability aligned with conditioned images. Similarly, Lavie Wang et al. [2023b] leverages temporal self-attentions and rotary positional encoding, emphasizing joint image-video fine-tuning for high-quality video outputs. I2VGen-XL Wang et al. [2023c] makes preliminary attempts in video generation with visual guidance from images. VideoCrafter Chen et al. [2023b, 2024] and DynamiCrafter Xing et al. [2023] enhance video quality through synthesized images and leveraging motion priors from text-to-video models to animate open-domain images. ModelScopeT2V Wang et al. [2023d] incorporates spatial-temporal blocks to ensure consistent frame generation and smooth movement transitions. SVD Blattmann et al. [2023b] evaluates the critical multi-stage training required for successful video latent diffusion models. VideoComposer Wang et al. [2023e] creatively concatenates image embeddings with style embeddings to enhance the visual continuity of generated videos. SEINE Chen et al. [2023c] proposes a random-mask video diffusion model to push the boundaries of text-driven video synthesis.

Diffusion-Transformer. In the evolving field of video generation, diffusion-transformer based methods [Brooks et al., 2024, Ope, 2024] have garnered considerable attention due to their flexibility and efficiency in handling sequential data. Latte [Ma et al., 2024b] enhances the transformer methodology by extracting spatio-temporal tokens from input videos and modeling video distribution in the latent space using a series of Transformer blocks. Cogvideo [Hong et al., 2022] leverages a transformer-based model optimized with a multi-frame-rate hierarchical training strategy, which enhances learning efficiency and video quality. Sora [Ope, 2024, Brooks et al., 2024] adopts a DiT-based generative architecture [Li et al., 2022], for video generation, showcasing the versatility of transformer-based models in adapting to the specific demands of video synthesis.

Others. In addition to diffusion-based models, Generative Adversarial Network(GAN) Shen et al. [2023], Wang et al. [2023f] and autoregressive models [GoogleAI, Kondratyuk et al., 2023] are also applied for video generation. Notable contributions in this domain include[Yoo et al., 2023] and [Lei et al., 2023], who have explored the fundamentals of transformer applications in video generation. FlashVideo [Lei et al., 2023] focuses on accelerating transformer models for video generation. VideoPoet [Kondratyuk et al., 2023] utilizes a decoder-only transformer architecture to process multimodal inputs and generate creative videos. Magvit [Yu et al., 2023] employs a masked generative video transformer that quantizes videos into spatial-temporal visual tokens using a 3D tokenizer. A few works [Shen et al., 2023, Ghosh et al., 2024, Wang et al., 2023f] introduce temporal layers into GAN for video generation.

Table 1: Statistics of real and generated videos in the GenVideo dataset.

Video Source	Type	Task	Time	Resolution	FPS	Length	Training Set	Testing Set	Total Count
Kinetics-400 [Kay et al., 2017b]	Real	-	17.05	224-340	-	5-10s	260,232	-	1,213,511
Youku-mPLUG [Xu et al., 2023b]		-	23.07	-	-	10-120s	953,279	-	
MSR-VTT [Xu et al., 2016b]	Real	-	16.05	-	-	10-30s	-	10,000	10,000
ZeroScope [zer, 2024]	Fake	T2V	23.07	1024×576	8	3s	132,465	-	1,048,575
I2VGen-XL [Wang et al., 2023c]		I2V	23.12	1280×720	8	2s	61,391	-	
SVD [Blattmann et al., 2023b]		I2V	23.12	1024×576	8	4s	149,026	-	
VideoCrafter [Chen et al., 2024]		T2V	24.01	1024×576	8	2s	37,970	-	
Pika [pik, 2022]		T2V&I2V	24.02	1088×640	24	3s	96,058	-	
DynamiCrafter [Xing et al., 2023]		I2V	24.03	1024×576	8	3s	44,681	-	
SD [Zhang et al., 2023b]		T2V&I2V	23-24	512-1024	8	2-6s	199,838	-	
SEINE [Chen et al., 2023c]		I2V	24.04	1024×576	8	2-4s	2,408	-	
Latte [Ma et al., 2024b]		T2V	24.03	512×512	8	2s	149,979	-	
OpenSora [Ope, 2024]		T2V	24.03	512×512	8	2s	174,759	-	
ModelScope [Wang et al., 2023d]	Fake	T2V	23.03	256×256	8	4s	-	700	8,588
MorphStudio [Mor, 2023]		T2V	23.08	1280×720	8	2s	-	700	
MoonValley [moonvalley.ai, 2022]		T2V	24.01	1024×576	16	3s	-	626	
HotShot [Hot, 2023]		T2V	23.10	672×384	8	1s	-	700	
Show_1 [Zhang et al., 2023c]		T2V	23.10	576×320	8	4s	-	700	
Gen2 [Esser et al., 2023]		I2V&T2V	23.09	896×512	24	4s	-	1,380	
Crafter [Chen et al., 2023b]		T2V	23.04	256×256	8	4s	-	1,400	
Lavie [Wang et al., 2023a]		T2V	23.09	1280×2048	8	2s	-	1,400	
Sora [Brooks et al., 2024]		T2V	24.02	-	-	-60s	-	56	
WildScrape		T2V&I2V	24	512-1024	8-16	2-6s	-	926	
Total Count	-	-	-	-	-	-	2,262,086	19,588	2,271,674

2.2 AI-Generated content detection

AI-generated visual content may raise concerns about the spread of misinformation. As a result, considerable efforts have been made to design forgery detection models and establish benchmarks in this field. In recent years, a significant amount of research has focused on detecting generated images [Guo et al., 2023b, Lorenz et al., 2023, Wu et al., 2023, Wang et al., 2023g] with help of AI-Generated Image datasets [Wang et al., 2022, Zhu et al., 2023a, Wang et al., 2023g], particularly those from unseen generative models. To date, studies in [Gu et al., 2021a, Xu et al., 2023a, Gu et al., 2022a] have addressed the detection of Deepfake videos, but there is a lack of research specifically dedicated to detecting generated videos on a wider range beyond human faces. We hope that this paper will make pioneering and insightful contributions to this research area.

3 GenVideo

3.1 Overviews of GenVideo

In response to the critical need for evaluating the generalizability of datasets and detectors (*i.e.*, the capacity of training detectors to accurately recognize unseen videos from the open world) and the robustness of these detectors (*i.e.*, their ability to maintain high performance against various corruptions to fake videos), we have developed the GenVideo dataset. This dataset is characterized by two main features:

- Large scale: The GenVideo dataset is organized hierarchically, encompassing cross generators such as diffusion-based generators and transformer-based generators, and cross architectures within the same type of generator like different motion modules combined with the same T2I base model Guo et al. [2023a], Zhang et al. [2023b]. This structure facilitates covering a broader range of generated content and producing fake videos on a larger scale. The training (*resp.*, testing) set in GenVideo contains a total of $2,262,086$ (*resp.*, $19,588$) video clips, comprising $1,213,511$ (*resp.*, $10,000$) real videos and $1,048,575$ (*resp.*, $8,588$) fake videos.
- Diverse content: GenVideo includes a wide array of high-quality fake videos sourced from open-source websites, along with videos produced using both user-trained and officially provided pre-trained video generation models, including T2V and I2V models. The generated video content encompasses a diverse range of scenes, including landscapes, people, buildings, objects, and more. The duration of the videos is primarily between 2 to 6 seconds, and the aspect ratios of the video resolutions vary widely. This diverse collection ensures a comprehensive set of fake videos, significantly enriching the understanding of AI-generated video detection across numerous real-world contexts, and enhancing the generalizability and robustness of detectors.

Evaluation objectives. To avoid trivial detection caused by the same distribution from the same generator, as observed in previous AI-generated image detection datasets [Wang et al., 2020, Zhu et al., 2023b], we conduct two tasks to verify the performance of detection models: cross-generator generalization and degraded video classification. Cross-generator generalization refers to the model being trained on data generated by some generators and validated on unseen data generated by other generators, which is meant to test the model's generalization ability. Degraded video classification, on the other hand, is used to validate the model's robustness by testing its ability to recognize videos of different types of degradation.

3.2 Organization of GenVideo

The GenVideo dataset primarily consists of real videos and fake videos shown in Table 1. The real videos are mainly sourced from existing datasets related to video action dataset [Kay et al., 2017a] and video description dataset [Xu et al., 2023b, 2016b]. The fake videos are obtained through external web scraping, internal generation pipelines based on open-source projects, and a number of existing video evaluation datasets [Liu et al., 2023b].

Considering the emergence of video generation models, which primarily focus on diffusion-based methods [Xing et al., 2023, Zhang et al., 2023b, Wang et al., 2023c, Blattmann et al., 2023b, Chen et al., 2024, Xing et al., 2023, Zhang et al., 2023b, Chen et al., 2023c] and methods based on auto-regressive models Ma et al. [2024b], Ope [2024], the training set of the GenVideo dataset predominantly comprises videos generated by these two popular types of algorithms shown in Table 1. Additionally, following [Liu et al., 2023b], we generate $96,058$ videos using the service provided by Pika website pik [2022]. To balance the quantity ratio between real videos and fake videos, we sampled 260,232 and 953,279 video clips from the existing video datasets Kinetics-400 [Kay et al., 2017b] and Youku-mPLUG [Xu et al., 2023b], respectively, to form the white sample of the training set.

For the testing set, the real videos are sourced from the MSR-VTT dataset [Xu et al., 2016b], which is a large video description dataset. The fake videos are mainly sourced from two parts: the first part comes from the Evalcrafter benchmark [Liu et al., 2023b], which is used to assess the temporal smoothness, quality, and other metrics of different generation models. The second part of the data comes from external web scraping, covering generated videos from existing popular video generation methods [Zhang et al., 2024, Yang et al., 2024, Bar-Tal et al., 2024, Ma et al., 2024a, Wang et al., 2023e, Ren et al., 2024, Wang et al., 2023h, Qing et al., 2023, Ho et al., 2022, Ge et al., 2023, Guo et al., 2023c, Tian et al., 2024, Kondratyuk et al., 2023, Yoo et al., 2023, Ni et al., 2023, Zeng et al., 2023, Wei et al., 2023, Feng et al., 2023, Xu et al., 2023c, Hu et al., 2023, Yu et al., 2023], which includes diffusion-based methods [Blattmann et al., 2023a, Xing et al., 2023, Wang et al., 2023d, Zeng et al., 2023, Wei et al., 2023, Feng et al., 2023, Xu et al., 2023c, Hu et al., 2023], auto-regressive-based models [Kondratyuk et al., 2023, Ma et al., 2024b, Brooks et al., 2024], and other models [Yoo et al., 2023, Yu et al., 2023]. This data encompasses most of the currently available video generation methods and advanced derivative methods of mainstream video generation techniques. Those scraped data are denoted as WildScape in Table 1.

3.3 Video collection details of GenVideo

We synthesize fake videos and gather real videos to construct the GenVideo dataset utilizing the hierarchical structure and the corresponding generators. It's crucial to underline that the primary objective of an AI-generated video detection dataset is to achieve *robust and generalizable detection capabilities*, rather than solely focusing on *video quality for assessment purposes*. A diverse and large-scale collection ensures that the dataset encompasses a wide range of video categories, facilitating detailed evaluations of AI-generated video detection algorithms and their effectiveness across various contexts.

Fake video collection: The guiding principle for collecting fake videos is to ensure maximal diversity in content and generators. We prioritize generating additional fake videos using the most recent generators due to their superior quality. To collect diverse fake videos from different resources as training samples, we have established a video generation pipeline for text-to-video generation and image-to-video generation. This pipeline facilitates the production of videos using popular generative mechanisms, including diffusion-based models [Xing et al., 2023, Zhang et al., 2023b, Guo et al.,

2023a, Wang et al., 2023c, Blattmann et al., 2023b, Chen et al., 2024, Xing et al., 2023, Zhang et al., 2023b, Chen et al., 2023c], transformer-based methods [Ma et al., 2024b, Ope, 2024] and service-based method Pika website [pik, 2022]. We selected common categories such as humans, animals, and plants as foreground keywords, and typical scenes like "in the park" or "on the lawn" as background keywords. Leveraging a large language model [Le Scao et al., 2023], these foreground and background keywords are expanded into comprehensive textual prompts to guide text-to-video generation. For image-to-video generation, we employed various text-to-image models, including different versions of Stable Diffusion (SD [Rombach et al., 2022], SDXL [Podell et al., 2023]). Using the enriched textual prompts generated by the large language model, we created corresponding images, which were subsequently used as input for image-to-video models to generate the final videos.

To assemble a representative testing set, we investigated current video generators based on different model architectures [Blattmann et al., 2023a, Brooks et al., 2024, Ghosh et al., 2024] and scraped example videos from their projects, as illustrated by WildScape in Table 1. This includes prominent video generation models such as VideoPoet [Kondratyuk et al., 2023], Emu [Girdhar et al., 2023], and Sora [Brooks et al., 2024]. Additionally, we collected videos generated by various condition-guided models [Wei et al., 2023, Feng et al., 2023, Xu et al., 2023c] that focus on social contexts and characters. We also included some non-mainstream generation algorithms, such as those based on latent flow diffusion models [Ni et al., 2023], masked generative video transformer [Yu et al., 2023], or autoregressive models [Gupta et al., 2023]. This approach ensures coverage of both popular algorithms and those generating high-quality content, particularly around character-centric videos. We integrated existing video quality evaluation dataset Liu et al. [2023a] that include typical generation methods and demonstrate relatively high generation quality.

Real video collection: Considering that fake videos from generators are limited to specific domains determined by training datasets such as Kinetics-400 [Kay et al., 2017b] and Youku-mPLUG [Xu et al., 2023d], we sample parts of videos from those datasets as the real part of the GenVideo dataset. Specifically, we randomly sample $953,279$ videos from Youku-mPLUG [Xu et al., 2023d] and randomly slice 10-second segments from each video to form real samples.

4 DeMamba

4.1 Preliminaries

Structured State Space Sequence models (S4) [Gu et al., 2022b, 2021b, Smith et al., 2023] are grounded in continuous systems, facilitating the mapping of a one-dimension function or sequence, denoted as $x(t) \in \mathbb{R}^L$ to $y(t) \in \mathbb{R}^L$, via an intermediary hidden state $h(t) \in \mathbb{R}^N$. In a formal context, S4 leverage the subsequent ordinary differential equation to represent the input data:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ embodies the system's evolutionary matrix, with $\mathbf{B} \in \mathbb{R}^{N\times L}$ and $\mathbf{C} \in \mathbb{R}^{L\times N}$ serving as the projection matrices. To navigate the transition from continuous to discrete modeling in contemporary S4, the Mamba framework utilizes a timescale parameter $\triangle$, facilitating the conversion of $\mathbf{A}$ and $\mathbf{B}$ into their discrete equivalents $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ through the Zero-Order Hold methodology [Gu et al., 2022b], expressed as:

$$\bar{\mathbf{A}} = \exp(\triangle\mathbf{A}), \quad \bar{\mathbf{B}} = \triangle\mathbf{A}^{-1}(\exp(\triangle\mathbf{A} - \mathbf{I})) \cdot \triangle\mathbf{B}, \quad h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \quad y_t = \mathbf{C}h_t. \tag{2}$$

Contrary to traditional models that primarily rely on linear time-invariant S4, Mamba [Gu and Dao, 2023] distinguishes itself by implementing a Selection mechanism computed with Scan for S4 (S6). Within the S6 framework, parameters $\mathbf{B} \in \mathbb{R}^{B\times L\times N}$, $\mathbf{C} \in \mathbb{R}^{B\times L\times N}$, and $\triangle \in \mathbb{R}^{B\times L\times D}$ are inherently derived from the input $\mathbf{x} \in \mathbb{R}^{B\times L\times D}$, formulating an intrinsic structure for contextual perceptiveness and adaptive modulation of weights.

4.2 AI-generated video detection with DeMamba module

Overview. As illustrated in Figure 2, our proposed method comprises a feature encoder, a DeMamba block, and an MLP classification head. Specifically, we employ state-of-the-art vision encoders (*e.g.*, CLIP [Radford et al., 2021] and XCLIP [Ni et al., 2022]) to encode the input video frames

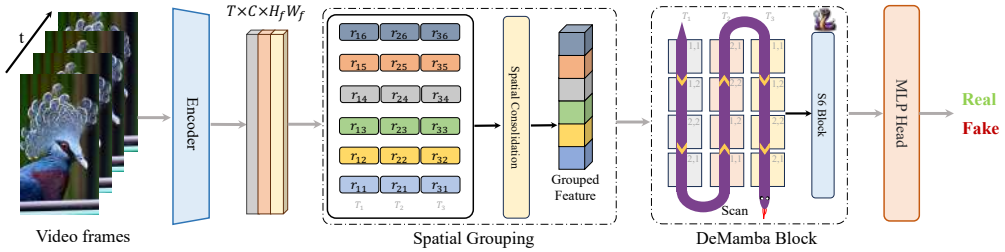

Figure 2: The overall framework of our Detail Mamba (DeMamba).

$\boldsymbol{X}^v \in \mathbb{R}^{3\times T\times H\times W}$ into a sequence of features, denoted by $\boldsymbol{F} \in \mathbb{R}^{T\times C\times H_f\times W_f}$, where C symbolizes the channel dimensionality, and H_f, W_f represent the spatial dimensions, *i.e.*, height and width of the feature maps, respectively. Following this, the extracted features are spatially grouped, and the DeMamba module is applied to model the intra-group feature consistency. Finally, we aggregate the features from different groups to determine whether the input video is generated by AI.

DeMamba block. We first apply spatial consolidation: given the feature $\boldsymbol{F}$, we split it into s^2 zones along both the height and width dimension where each zone of $\boldsymbol{F}$ is denoted as $\boldsymbol{F}_{jk} \in \mathbb{R}^{T\times C\times(H_f/s)\times(W_f/s)}$, where $j,k = \{1,...,s\}$. In Figure 2, we adapt the 1D Mamba layer for handling spatial-temporal input by expanding its capability to a 3D scan. In the previous Mamba approaches [Gu and Dao, 2023, Zhu et al., 2024, Liu et al., 2024], a sweep-scan mechanism was utilized, which might not effectively capture the inherent contextual relationships between adjacent tokens. To address this limitation, we propose a continuous scan strategy for each segmented region, aimed at maintaining spatial continuity throughout the entire scanning phase. Suppose a zone consists of four spatial positions: (1,1), (1,2), (2,1), and (2,2), corresponding to the top-left, top-right, bottom-left, and bottom-right corners, respectively. The sweep-scan order is (1,1) -> (1,2) -> (2,1) -> (2,2), whereas in the continuous scan, the order is (1,1) -> (1,2) -> (2,2) -> (2,1). This method organizes spatial tokens based on their proximity and subsequently aligns them sequentially across successive frames. It facilitates the coherent integration of spatial and temporal dynamics, enhancing the capability of the model to capture complex spatial-temporal relationships. After modeling the spatial-temporal inconsistency of each partitioned region using DeMamba, we can obtain the feature $\boldsymbol{F}'_{jk} \in \mathbb{R}^{T\times C\times(F_hF_w/s^2))}$, where $j,k = \{1,...,s\}$.

Classification head. To leverage more comprehensive features for classification, we aggregate both global and local features. Specifically, we temporally and spatially average the input features $\boldsymbol{F}$ before the DeMamba block to obtain the global feature $\boldsymbol{F}^{\text{global}} \in \mathbb{R}^C$, and average pool the temporal and spatial features $\boldsymbol{F}'_{jk}$ after the DeMamba processing into pooled features $\boldsymbol{F}^{\text{pool}}_{jk} \in \mathbb{R}^C$. Then we concatenate the local features with the global features and apply a simple MLP for classification:

$$y_{\text{pred}} = \text{Sigmoid}(\text{MLP}([\boldsymbol{F}^{\text{global}}; \boldsymbol{F}^{\text{pool}}_{11}; ....; \boldsymbol{F}^{\text{pool}}_{ss}])). \tag{3}$$

Finally, we use binary cross-entropy loss to train our model to classify real/fake videos.

5 Experiments

5.1 Implementation details

Datasets. To comprehensively analyze the performance of various detectors, we divided the dataset into two distinct parts: the basic training set D_{train} and the out-of-domain test set $D_{\text{v-ood}}$. D_{train} and $D_{\text{v-ood}}$ contain fake videos created by different generative methods and different real videos. D_{train} and $D_{\text{v-ood}}$ contain fake videos produced by different generative methods and real videos from different sources. D_{train} includes 1,213,511 real videos and 1,048,575 generated videos produced using 10 baseline generative methods. $D_{\text{v-ood}}$ contains 10,000 real videos and 8,588 generated videos created with 10 generative methods. For detailed information about the data, please refer to **Section 3 and Section 3.3**.

Evaluation metrics. Consistent with the methodologies employed in prior studies, our evaluation framework primarily focuses on reporting accuracy (ACC) and average precision (AP) to assess the

Table 2: Training parameter settings.

Task	Model	Batchsize	LR	Samples per epoch	Epochs	Optimizer	Scheduler	Loss
many-to-many generalization	F3Net	1024	1e-5	1024×2000	30	AdamW	[20, 25], lr * 0.1	BCE
	NPR	1024	1e-5	1024×2000	30		[20, 25], lr * 0.1	
	STIL	128	1e-5	128×4000	30		[20, 25], lr * 0.1	
	VideoMAE-B	128	1e-5	128×4000	30		[20, 25], lr * 0.1	
	CLIP-B-LP	1024	1e-6	1024×2000	30		[20, 25], lr * 0.1	
	CLIP-B-FT	1024	1e-6	1024×2000	10		-	
	XCLIP-B-LP	128	1e-6	128×4000	30		[20, 25], lr * 0.1	
	XCLIP-B-FT	128	1e-6	128×4000	10		-	
	DeMamba-CLIP-LP	128	1e-6	1024×2000	30		[20, 25], lr * 0.1	
	DeMamba-CLIP-FT	128	1e-6	1024×2000	10		-	
	DeMamba-XCLIP-LP	128	1e-6	128×4000	30		[20, 25], lr * 0.1	
	DeMamba-XCLIP-FT	128	1e-6	128×4000	10		-	
one-to-many generalization	STIL	128	1e-5	128×400	10	AdamW	-	BCE
	NPR	1024	1e-5	1024×2000	10		-	
	XCLIP-B-FT	1024	1e-6	128×400	10		-	
	DeMamba-XCLIP-FT	128	1e-6	128×400	10		-	

Table 3: Comparisons to the SOTAs in mean accuracy (ACC) and average precision (AP) on the many-to-many generalization task.

Model	Detection level	Metric	Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Real	Avg.
F3Net	Image	ACC	83.93	99.71	98.62	77.57	57.00	36.57	99.52	99.71	89.43	76.78	99.14	83.45
		AP	68.27	99.89	99.67	89.35	85.24	63.17	99.58	99.89	93.80	88.41	-	88.73
NPR	Image	ACC	91.07	99.57	99.49	24.29	89.64	57.71	97.12	99.86	94.29	87.80	97.46	85.30
		AP	67.17	99.14	99.20	22.76	93.91	61.76	96.33	99.72	94.15	90.40	-	82.45
STIL	Video	ACC	78.57	98.14	98.04	76.00	61.79	53.29	99.36	97.36	94.57	65.01	98.72	83.71
		AP	57.21	99.08	99.32	86.19	82.24	70.43	99.25	98.96	97.18	81.32	-	87.12
VideoMAE-B	Video	ACC	67.86	96.00	98.41	96.14	77.14	80.43	97.44	96.93	96.29	68.36	99.71	88.61
		AP	66.49	98.85	99.77	99.27	96.55	95.31	99.49	99.69	99.27	90.74	-	94.54
CLIP-B-PT	Image	ACC	85.71	82.43	90.36	71.00	79.29	75.43	89.62	86.29	82.14	75.16	57.22	79.67
		AP	6.78	43.56	70.88	29.97	52.97	35.36	55.52	66.03	44.23	42.99	-	44.83
DeMamba -CLIP-PT	Video	ACC	58.93	96.43	93.12	68.00	69.36	69.00	89.14	91.86	96.14	56.59	98.06	$80.60_{+0.93}$
		AP	25.87	95.14	96.23	73.43	83.31	75.49	90.17	95.06	95.05	69.95	-	$79.97_{+35.14}$
CLIP-B-FT	Image	ACC	94.64	99.86	91.38	77.29	88.14	86.00	99.68	99.79	84.29	84.67	97.38	91.19
		AP	80.67	99.67	95.24	82.20	93.48	88.62	99.55	99.79	86.93	89.08	-	91.52
DeMamba CLIP-B-FT	Video	ACC	95.71	100.00	98.70	69.14	92.43	93.29	100.00	100.00	83.57	82.94	99.44	$92.29_{+1.10}$
		AP	85.50	100.00	99.59	76.15	96.78	96.99	99.97	100.00	89.80	89.72	-	$93.45_{+1.93}$
XCLIP-B-PT	Video	ACC	81.34	82.15	83.35	80.98	81.82	81.55	82.14	82.98	81.93	81.10	81.37	81.88
		AP	16.39	72.16	87.77	39.86	65.57	54.26	75.23	84.80	61.60	55.28	-	61.29
DeMamba -XCLIP-PT	Video	ACC	66.07	95.86	94.64	77.86	75.36	80.29	90.89	92.50	96.00	66.41	95.12	$84.64_{+2.76}$
		AP	18.26	93.50	94.72	69.94	78.08	71.50	83.95	92.23	93.54	68.10	-	$76.38_{+15.09}$
XCLIP-B-FT	Video	ACC	82.14	99.57	93.62	61.29	79.36	69.71	97.92	99.79	77.14	83.59	98.14	85.66
		AP	64.42	99.73	96.78	70.98	90.35	77.28	97.34	99.84	82.01	88.97	-	86.77
DeMamba -XCLIP-FT	Video	ACC	98.21	100.00	99.86	65.43	94.86	98.86	100.00	100.00	92.86	89.09	99.42	$\mathbf{94.42}_{+8.76}$
		AP	93.32	100.00	99.97	85.55	98.97	99.60	99.98	100.00	97.77	95.75	-	$\mathbf{97.10}_{+10.43}$

effectiveness of the detectors. The accuracy calculation is based on a threshold value of 0.5. For image-based detection techniques, we consolidate frame-level predictions to derive the corresponding video-level predictions, ensuring a coherent analysis across different media formats. It is noteworthy that when evaluating the performance on a dataset generated by a specific synthesis method, we calculate the ACC of that synthesis method based on the dataset itself. Additionally, in the process of computing the AP, we take into account real videos to achieve a more comprehensive assessment.

Baselines. 1) CLIP (ICML'21) [Radford et al., 2021] is an innovative model that connects images with textual descriptions, enabling it to perform exceptionally well across various vision and language tasks. 2) F3Net (ECCV'20) [Qian et al., 2020] addresses the detection of sophisticated face manipulations using a combination of frequency-aware components and loc al statistics. Its dual-stream approach effectively identifies forgery patterns. 3) VideoMAE (NeurIPS'22) [Tong et al., 2022] adapts the Masked Autoencoder [He et al., 2022] concept for video, enhancing feature extraction and improving performance on downstream tasks through self-supervised training. 4) XCLIP (ECCV'22) [Ni et al., 2022] adapts existing image-language models for video recognition, adding a new cross-frame attention mechanism to improve the exchange of temporal information and tailor prompts specifically for videos. 5) STIL (MM'21) [Gu et al., 2021a] proposes a novel spatial-temporal inconsistency

Table 4: Comparisons to the SOTAs in mean accuracy (ACC) and average precision (AP) on the one-to-many generalization task.

Training subset	Model	Detection level	Metric	Testing subset											Avg.
				Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Real	
Pika	NPR	Image	ACC	55.36	77.57	71.88	4.86	7.21	4.29	86.26	60.29	71.43	31.53	99.52	51.83
			AP	45.74	91.55	92.71	21.80	44.32	22.74	95.04	90.03	84.91	60.88	-	64.97
	STIL	Image	ACC	75.00	79.43	94.49	57.86	53.14	64.14	97.12	85.29	69.43	62.42	92.43	75.52
			AP	22.35	71.62	93.19	40.61	53.24	47.73	94.94	85.82	58.99	61.91	-	63.04
	XCLIP-B-FT	Video	ACC	67.86	91.29	96.23	12.00	22.36	9.14	99.84	83.43	75.57	51.84	99.64	64.47
			AP	71.08	97.53	99.44	44.68	72.69	38.37	99.96	97.32	88.00	74.00	-	78.31
	DeMamba -XCLIP-FT	Video	ACC	92.86	97.29	98.48	38.29	53.50	41.43	99.84	94.07	77.29	64.15	98.65	**77.80**$_{+13.33}$
			AP	77.75	98.42	99.16	52.97	76.72	56.24	99.80	97.91	82.83	74.81	-	**81.66**$_{+3.35}$
SEINE	NPR	Image	ACC	46.43	78.57	63.70	21.86	7.00	3.29	92.97	89.29	33.86	24.84	99.70	51.05
			AP	36.30	92.63	85.02	52.68	25.69	11.05	97.80	97.78	64.64	47.48	-	61.11
	STIL	Video	ACC	71.43	80.43	88.48	67.71	54.57	55.71	93.93	89.57	72.00	50.11	92.27	74.20
			AP	23.89	71.01	88.18	52.17	54.49	41.23	84.73	87.38	58.72	46.51	-	60.83
	XCLIP-B-FT	Video	ACC	85.71	95.43	76.23	65.86	35.93	37.00	99.68	99.00	75.57	49.78	99.80	74.54
			AP	85.89	97.97	94.40	92.81	81.68	77.68	98.48	98.91	92.27	67.91	-	88.80
	DeMamba -XCLIP-FT	Video	ACC	94.64	98.43	92.17	82.43	52.29	54.00	99.52	99.14	79.29	57.88	98.99	**82.61**$_{+8.07}$
			AP	83.74	99.01	97.66	90.82	84.11	73.30	99.72	99.73	89.72	76.45	-	**89.43**$_{+0.63}$
OpenSora	NPR	Image	ACC	55.36	76.29	55.51	58.57	76.50	22.43	74.92	83.07	29.86	60.37	95.95	62.62
			AP	25.65	75.24	65.02	55.12	82.42	20.75	72.65	86.84	28.13	64.50	-	57.63
	STIL	Video	ACC	32.14	45.43	56.45	35.14	45.07	34.57	57.83	63.14	19.86	43.95	98.13	48.33
			AP	6.94	55.62	75.99	43.55	68.06	44.01	63.84	80.59	29.39	57.58	-	52.56
	XCLIP-B-FT	Video	ACC	67.86	75.86	67.46	70.86	73.14	43.57	79.87	86.29	33.43	63.17	98.10	69.06
			AP	48.28	81.39	81.84	77.38	86.08	51.87	83.41	93.18	39.27	72.74	-	71.54
	DeMamba -XCLIP-FT	Video	ACC	55.36	87.43	81.30	73.14	85.21	73.14	89.62	90.07	44.86	58.10	97.30	**75.95**$_{+6.89}$
			AP	25.89	86.63	87.27	74.38	91.12	76.01	86.41	93.83	48.74	67.92	-	**73.82**$_{+2.28}$

Table 5: Robustness evaluation of different detectors on many-to-many generalization task.

Model	Detection level	Metric	Original	Compression		Transformation		Watemarks		Gaussian noise	Color transform
				CRF = 28	JEPG	Flip	Crop	Text	Image		
F3Net	Image	ACC	83.45	77.55$_{-5.90}$	82.04$_{-1.41}$	81.45$_{-2.00}$	69.71$_{-13.74}$	76.90$_{-7.55}$	78.67$_{-4.78}$	82.94$_{-0.51}$	82.93$_{-0.52}$
		AP	88.73	87.39$_{-1.34}$	**81.02**$_{-7.71}$	87.70$_{-1.73}$	67.05$_{-21.68}$	87.01$_{-1.72}$	87.93$_{-0.70}$	87.17$_{-1.56}$	87.17$_{-1.56}$
NPR	Image	ACC	85.30	80.83$_{-4.47}$	73.92$_{-11.38}$	83.82$_{-1.48}$	71.69$_{-13.61}$	83.26$_{-2.04}$	82.43$_{-2.87}$	84.94$_{-0.36}$	85.03$_{-0.47}$
		AP	82.45	76.12$_{-6.33}$	53.72$_{-28.73}$	79.05$_{-3.40}$	36.36$_{-46.09}$	67.16$_{-15.28}$	77.02$_{-5.43}$	57.55$_{-24.90}$	76.81$_{-5.64}$
STIL	Video	ACC	83.71	78.14$_{-5.57}$	73.79$_{-9.92}$	81.97$_{-1.74}$	77.58$_{-6.13}$	77.64$_{-6.07}$	74.16$_{-9.55}$	83.44$_{-0.27}$	83.47$_{-0.24}$
		AP	87.12	85.04$_{-2.08}$	59.38$_{-27.74}$	85.21$_{-1.91}$	76.38$_{-10.74}$	82.59$_{-4.53}$	83.09$_{-4.03}$	74.54$_{-12.58}$	86.65$_{-0.47}$
CLIP-B-FT	Image	ACC	91.19	82.92$_{-8.27}$	76.78$_{-12.41}$	89.23$_{-12.41}$	66.91$_{-24.28}$	79.65$_{-11.54}$	78.58$_{-12.61}$	88.50$_{-2.69}$	88.81$_{-2.38}$
		AP	91.52	90.42$_{-1.10}$	85.33$_{-12.41}$	91.05$_{-0.14}$	76.35$_{-15.17}$	88.94$_{-2.58}$	91.05$_{-0.47}$	84.55$_{-6.97}$	91.05$_{-0.50}$
XCLIP-B-FT	Video	ACC	85.66	84.17$_{-1.49}$	53.96$_{-31.70}$	84.48$_{-1.18}$	56.24$_{-30.42}$	85.06$_{-1.60}$	82.84$_{-2.82}$	82.94$_{-2.72}$	85.41$_{-0.25}$
		AP	86.77	86.53$_{-0.24}$	54.11$_{-32.66}$	86.76$_{-0.01}$	46.10$_{-40.67}$	86.71$_{-0.06}$	86.74$_{-0.03}$	68.05$_{-14.89}$	86.01$_{-0.76}$
DeMamba -CLIP-FT	Video	ACC	92.29	85.63$_{-6.66}$	83.49$_{-8.80}$	90.59$_{-1.70}$	62.25$_{-30.04}$	83.34$_{-8.95}$	87.14$_{-5.15}$	90.63$_{-1.66}$	91.50$_{-0.79}$
		AP	93.45	90.93$_{-2.52}$	74.03$_{-19.42}$	93.44$_{-0.01}$	61.12$_{-32.33}$	90.08$_{-3.37}$	93.35$_{-0.10}$	82.99$_{-10.46}$	93.39$_{-0.06}$
DeMamba -XCLIP-FT	Video	ACC	**94.42**	**90.52**$_{-3.90}$	**93.26**$_{-1.16}$	**94.26**$_{-0.16}$	**72.98**$_{-22.44}$	**89.59**$_{-4.83}$	**91.06**$_{-3.36}$	**94.36**$_{-0.06}$	**93.82**$_{-0.60}$
		AP	**97.10**	**96.00**$_{-1.10}$	76.29$_{-20.81}$	**96.76**$_{-0.34}$	**67.18**$_{-29.92}$	**94.30**$_{-2.80}$	**97.07**$_{-0.03}$	**85.05**$_{-12.05}$	**97.05**$_{-0.05}$

learning method, which captures the spatial-temporal inconsistencies in forged videos by introducing a spatial inconsistency module, a temporal inconsistency module, and an information supplementation module. 6) NPR (CVPR'24) [Tan et al., 2024] is a concept introduced to understand and describe the complex, local interdependencies between image pixels that arise from upsampling processes, particularly evident in synthetic images created by generative models like GANs and diffusion models, which are used to identify and analyze the structural artifacts resulting from these operations.

Training settings. As shown in Table 2, we present the different model training parameter settings for many-to-many and one-to-many generalization tasks. All of our experiments were conducted on a system equipped with 8 Tesla A-100 80G GPUs and an Intel(R) Xeon(R) Platinum 8369B CPU @ 2.90GHz.

5.2 Task1: cross generator generalization

Due to the rapid iteration of generation methods, we propose a cross-dataset generalization task to test the generalization performance of detectors. Specifically, it consists of two types of generalization tasks: 1) the many-to-many generalization task, and 2) the one-to-many generalization task.

Many-to-many generalization task. This task involves training on 10 baseline categories and then testing on each subset and the average detection performance on $D_{\text{v-ood}}$. As shown in Table 3, video models achieve better recognition accuracy compared to image models because video models can model temporal sequences. Moreover, our DeMamba model can be effectively integrated into existing

Table 6: Ablation study of DeMamba.

Model	Metric	Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Real	Avg.
w/o Demamba Block	ACC	82.14	99.57	93.62	61.29	79.36	69.71	97.92	99.79	77.14	83.59	98.14	85.66$_{-8.76}$
	AP	64.42	99.73	96.78	70.98	90.35	77.28	97.34	99.84	82.01	88.97	-	86.77$_{-10.43}$
w/o global feat	ACC	97.33	99.85	99.45	67.66	93.55	97.88	100.00	100.00	91.34	86.54	98.55	93.83$_{-0.59}$
	AP	87.32	99.35	99.22	82.44	97.99	99.42	99.42	99.85	97.32	94.98	-	95.73$_{-1.37}$
DeMamba -XCLIP-FT	ACC	98.21	100.00	99.86	65.43	94.86	98.86	100.00	100.00	92.86	89.09	99.42	**94.42**
	AP	93.32	100.00	99.97	85.55	98.97	99.60	99.98	100.00	97.77	95.75	-	**97.10**

Table 7: Influence of different zone sizes in DeMamba.

Model	Zone size	Metric	Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Real	Avg.
DeMamba -CLIP-FT	1×1	ACC	100.00	99.86	84.78	95.57	81.43	82.29	100.00	99.79	78.29	81.10	97.64	90.98
		AP	93.80	99.64	91.78	96.03	90.04	86.13	99.92	99.81	82.12	86.96	-	92.62
	2×2	ACC	95.71	100.00	98.70	69.14	92.43	93.29	100.00	100.00	83.57	82.94	99.44	**92.29**
		AP	85.50	100.00	99.59	76.15	96.78	96.99	99.97	100.00	89.80	89.72	-	**93.45**
	7×7	ACC	98.21	100.00	98.41	65.00	94.00	95.43	100.00	100.00	80.86	84.34	98.19	91.62
		AP	87.28	99.99	99.06	73.51	97.47	97.00	99.76	99.97	85.75	89.95	-	92.97
	14×14	ACC	100.00	99.86	93.41	88.57	81.71	70.14	100.00	99.93	76.43	80.02	98.70	89.89
		AP	94.94	99.42	97.03	93.56	94.02	84.56	99.12	99.22	77.60	89.46	-	92.89
DeMamba -XCLIP-FT	1×1	ACC	91.07	100.00	99.35	73.71	89.50	96.29	100.00	100.00	94.71	81.21	99.37	93.20
		AP	86.65	99.99	99.83	90.64	97.85	98.81	99.91	99.99	98.43	92.18	-	96.43
	2×2	ACC	98.21	100.00	99.86	65.43	94.86	98.86	100.00	100.00	92.86	89.09	99.42	**94.42**
		AP	93.32	100.00	99.97	85.55	98.97	99.60	99.98	100.00	97.77	95.75	-	**97.10**
	7×7	ACC	96.43	100.00	99.57	56.00	88.71	96.00	100.00	100.00	94.43	79.48	99.64	91.84
		AP	96.02	100.00	99.97	82.85	99.69	99.52	99.98	100.00	86.51	97.45	-	96.20
	14×14	ACC	76.79	100.00	99.06	22.86	84.71	95.14	99.52	100.00	87.71	72.25	99.87	85.26
		AP	80.43	100.00	99.76	65.69	98.49	99.31	99.79	99.91	96.87	88.31	-	92.86

models, achieving significant improvements. For instance, by incorporating the DeMamba module into the XCLIP, the DeMamba-XCLIP-FT attains an average accuracy of 94.42% and an average AP of 97.10%, which represents a 10.22% increase in accuracy and 12.02% increase in AP compared to the original XCLIP. Note that PT (Partially Tuning) indicates that the backbone is frozen, with only the other parts being fine-tuned, while FT (Full Training) tuning the entire model.

One-to-many generalization task. Following AI-generated image detection setting [Tan et al., 2024, Corvi et al., 2023, Wang et al., 2023g], we also perform a one-to-many generalization task. Unlike the many-to-many generalization task, the one-to-many generalization task involves training on one baseline category and then testing on each subset and the average detection performance on $D_{\text{v-ood}}$. As shown in Table 4, our DeMamba-XCLIP-FT achieves better generalization performance in three one-to-many generalization tasks due to the learning of spatial-temporal inconsistency in DeMamba.

5.3 Task2: degraded video classification

In practical detection scenarios, the robustness of the detector to perturbations is also of paramount importance. In this regard, we investigated the impact of perturbations on the detector on 8 different types: H.264 compression, JPEG compression, FLIP, Crop, text watermark, image watermark, Gaussian noise, and color transform. More specific details about the perturbations can be found in **Appendix B.2**. Table 5 shows the performance of the models trained in the many-to-many task under the influence of these perturbations. We can see that in the case of degraded data, DeMamba-XCLIP-FT still achieves the best performance, indicating that our model has good robustness in the face of degraded data.

5.4 Ablation study

Ablation testing. We conducted ablation experiments to validate the effectiveness of DeMamba. As shown in Table 6, DeMamba effectively enhances the generalization performance of the model. Additionally, when using fused features, the model achieves its best performance.

Influence of different zone sizes. We investigated the impact of zone size in dividing zones in DeMamba on modeling temporal inconsistency. As shown in Table 7, the best performance is observed when the zone size is 2. Smaller zones enable the model to concentrate more on local

Table 8: Influence of scanning order in DeMamba.

Model	Scan Rrder	Metric	Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Real	Avg.
DeMamba -XCLIP-FT	sweep	ACC	95.21	100.00	92.55	61.77	92.69	96.33	99.87	99.32	81.53	73.11	98.79	90.11$_{-3.31}$
		AP	93.32	99.95	97.32	85.55	95.88	97.54	99.66	97.54	87.13	80.82	-	93.47$_{-3.63}$
	continuous	ACC	98.21	100.00	99.86	65.43	94.86	98.86	100.00	100.00	92.86	89.09	99.42	**94.42**
		AP	93.32	100.00	99.97	85.55	98.97	99.60	99.98	100.00	97.77	95.75	-	**97.10**

details, leading to superior modeling performance. However, excessively small zones may result in the loss of spatial contextual information. Therefore, selecting an appropriate zone size is crucial.

Influence of scanning orders. As shown in Table 8, the continuous scan proposed in this paper effectively enhances performance compared to the traditional scanning method.

6 Broader impacts

Our research focuses on utilizing machine learning to detect generated videos. We have introduced the first million-scale AI-generated video detection dataset and developed the DeMamba model. These efforts are crucial for protecting digital content and preventing the spread of misinformation. However, there is a potential for these tools to be misused, leading to competition between video generation and detection technologies. We aim to advocate for the ethical use of technology and promote creative research into tools that verify media authenticity. We believe this will help protect the public from the harm of misinformation, enhance the clarity and authenticity of information dissemination, and ensure the protection of personal privacy.

7 Conclusion and limitation

This paper introduces GenVideo, a dataset specifically designed for detecting fake videos generated by generative models. GenVideo is characterized by its large-scale nature, as well as the rich diversity of generated content and methods. We propose two tasks that mimic real-world scenarios, namely the cross-generator video classification task and the degraded video classification task, to evaluate the detection performance of existing detectors on GenVideo. Additionally, we introduce a plug-and-play effective detection model called Detail Mamba (DeMamba), which distinguishes AI-generated videos by analyzing inconsistencies in the spatial-temporal dimensions. This model has demonstrated its strong generalization and robustness across multiple tasks. We hope that this research will inspire the creation and improvement of other detection technologies, providing new avenues for the development of authentic and reliable AI-generated content applications.

The main limitation of this article lies in the suboptimal training efficiency of the proposed DeMamba, a common issue with the Mamba model. Consequently, we encourage the community to design more lightweight and generalized detection models to facilitate the regulation of AI-generated content.

References

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, pages 3813–3824, 2023a.

Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Jun Lan, Xing Zheng, Yaohui Li, Changhua Meng, Huijia Zhu, and Weiqiang Wang. Diffute: Universal text editing diffusion model. In *NeurIPS*, 2023a.

Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. In *NeurIPS*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

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. *arXiv preprint arXiv:2310.11440*, 2023a.

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*, 2023a.

GoogleAI. Veo. `https://deepmind.google/technologies/veo/`. Accessed: 2024-05.

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. URL `https://openai.com/index/sora/`.

Runway Research. Text driven video generation, 2023. URL `https://research.runwayml.com/gen2`.

Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, 2023.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *ACM Multimedia*, pages 3473–3481, 2021a.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. TALL: thumbnail layout for deepfake video detection. In *ICCV*, pages 22601–22611, 2023a.

Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *ECCV*, volume 13672, pages 596–613, 2022a.

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, Chenliang Li, Qi Qian, Maofei Que, Ji Zhang, Xiao Zeng, and Fei Huang. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362*, 2023b.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017a.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016a.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, volume 12357, pages 86–103, 2020.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, 2024.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, volume 13664, pages 1–18, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024.

Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*, 2024a.

David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. *arXiv preprint arXiv:2312.04433*, 2023.

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.

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. *arXiv preprint arXiv:2311.10709*, 2023.

Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, et al. Dreamoving: A human video generation framework based on diffusion models. *arXiv preprint arXiv:2312.05107*, 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023c.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, pages 18444–18455, 2023.

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 *ICCV*, pages 15954–15964, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023a.

Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. *arXiv preprint arXiv:2312.13964*, 2023b.

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*, 2023b.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. I2vgen-xl, 2023c. URL `https://modelscope.cn/models/damo/Image-to-Video/summary`.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023b.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

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. *arXiv preprint arXiv:2310.12190*, 2023.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023d.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023b.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 36, 2023e.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*, 2023c.

Open-sora: Democratizing efficient video production for all, 2024. URL `https://github.com/hpcaitech/Open-Sora`.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024b.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *ACM Multimedia*, pages 3530–3539, 2022.

Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Mostgan-v: Video generation with temporal motion styles. In *CVPR*, pages 5652–5661, 2023.

Yuhan Wang, Liming Jiang, and Chen Change Loy. Styleinv: A temporal style modulated inversion network for unconditional video generation. In *ICCV*, pages 22851–22861, 2023f.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Jaehoon Yoo, Semin Kim, Doyup Lee, Chiheon Kim, and Seunghoon Hong. Towards end-to-end generative modeling of long videos with memory-efficient bidirectional transformers. In *CVPR*, pages 22888–22897, 2023.

Bin Lei, Caiwen Ding, et al. Flashvideo: A framework for swift inference in text-to-video generation. *arXiv preprint arXiv:2401.00869*, 2023.

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

Partha Ghosh, Soubhik Sanyal, Cordelia Schmid, and Bernhard Schölkopf. Raven: Rethinking adversarial video generation with efficient tri-plane networks. *arXiv preprint arXiv:2401.06035*, 2024.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *ICCV*, pages 3155–3165, 2023b.

Peter Lorenz, Ricard L Durall, and Janis Keuper. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *ICCV*, pages 448–459, 2023.

Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800*, 2023.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *ICCV*, pages 22388–22398, 2023g.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *arXiv preprint arXiv:2306.08571*, 2023a.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017b.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016b.

Zeroscope-v2-xl, 2024. URL `https://huggingface.co/cerspense/zeroscope_v2_XL`.

Pika art, 2022. URL `https://pika.art/`.

Morph studio, 2023. URL `https://www.morphstudio.com/`.

moonvalley.ai. moonvalley.ai, 2022. URL `https://moonvalley.ai/`.

Hotshot, 2023. URL `https://huggingface.co/hotshotco/Hotshot-XL`.

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*, 2023c.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, pages 7346–7356, 2023.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pages 8692–8701, 2020.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *NeurIPS*, 2023b.

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. *arXiv preprint arXiv:2310.11440*, 2023b.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. *arXiv preprint arXiv:2402.03162*, 2024.

Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. *arXiv preprint arXiv:2312.15770*, 2023h.

Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*, 2023.

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*, pages 22930–22941, 2023.

Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023c.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.

Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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. *arXiv preprint arXiv:2312.06662*, 2023.

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, et al. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362*, 2023d.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2022b.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In *NeurIPS*, pages 572–585, 2021b.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *ICLR*, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 15979–15988, 2022.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, pages 1–5, 2023.

Appendix

A Model details

We provide detailed information about the methods used in this paper, as shown in Table 9. Our model only requires a small addition of parameters on the XCLIP-B model to achieve significant performance improvements.

Table 9: Model details and its performance on many-to-many generalization task.

Model	Detection level	Input size	Param (M)	FLOPs(G)	ACC	AP
CLIP-B	Image	$8 \times 224 \times 224$	151.46	149.34	91.19	91.52
NPR	Image	$8 \times 224 \times 224$	**1.44**	**14.08**	85.30	82.45
F3Net	Image	$8 \times 299 \times 299$	48.31	145.04	83.45	88.73
VideoMAE-B	Video	$16 \times 224 \times 224$	86.54	147.52	88.61	94.54
STIL	Video	$8 \times 224 \times 224$	22.69	38.06	83.71	87.12
XCLIP-B	Video	$8 \times 224 \times 224$	121.26	141.28	85.66	86.77
DeMamba-XCLIP	Video	$8 \times 224 \times 224$	125.37	147.61	**94.42**	**97.10**

B Experiment details

B.1 Implementation details

Data pre-processing. For each video, we uniformly sample frames for alignment. For videos longer than 3 seconds, we sample 2 frames every second. For videos shorter than 3 seconds, our sampling frequency is $\frac{8}{\text{length}}$ seconds. The pseudo-code for Pytorch-like is as follows:

```
    video_length = get_video_length(video_path)
    os.makedirs(os.path.dirname(image_path), exist_ok=True)
    if video_length >= 3 :
        inter_val = 2
        os.system(f"cd {image_path} | ffmpeg -loglevel quiet -i {video_path}
        -r {inter_val} {image_path}%d.jpg")
    else:
        inter_val = math.ceil(8 / video_length)
        os.system(f"cd {image_path} | ffmpeg -loglevel quiet -i {video_path}
        -r {inter_val} {image_path}%d.jpg")
```

Dataset augmentation. During training and testing, we randomly select 8 or 16 consecutive frames from the video after frame sampling, and resize each frame to 224×224. To enhance generalizability of models, we introduced random data augmentation during training, including HorizontalFlip, ImageCompression, GaussNoise, GaussianBlur, and Grayscale. The pseudo-code for Pytorch-like is as follows:

```
    aug_list  = [albumentations.Resize(224, 224)]
    if random.random() < 0.5:
        aug_list.append(albumentations.HorizontalFlip(p=1.0))
    if random.random() < 0.5:
        quality_score = random.randint(50, 100)
        aug_list.append(albumentations.ImageCompression(
        quality_lower=quality_score, quality_upper=quality_score))
    if random.random() < 0.3:
        aug_list.append(albumentations.GaussNoise(p=1.0))
    if random.random() < 0.3:
        aug_list.append(albumentations.GaussianBlur(blur_limit=(3, 5), p=1.0))
    if random.random() < 0.001:
        aug_list.append(albumentations.ToGray(p=1.0))
    aug_list.append(albumentations.Normalize(mean=(0.485, 0.456, 0.406),
    std=(0.229, 0.224, 0.225), max_pixel_value=255.0, p=1.0))
    trans = albumentations.Compose(aug_list)
```

B.2 Details in degraded video classification task

We applied the following transformations to the videos in $D_{\text{v-ood}}$ and utilized models trained on many-to-many tasks for testing. Here, we provide the specific implementation details for each task of degraded video classification:

(1) H.264 compression: H.264, also known as Advanced Video Coding (AVC), is a widely used video compression standard. In this paper, we set the crf to 28 to compress the video.

(2) JPEG compression: JPEG compression is a widely used image compression standard designed for efficient compression of digital images. The JPEG algorithm is based on the characteristics of the human visual system, taking advantage of the insensitivity of human eyes to the loss of image details, thus achieving lossy compression of data. In this paper, we set the quality to 35 for the degradation experiment.

(3) FLIP: We randomly select either Horizontal Flip or Vertical Flip with equal probability for the degradation experiment.

(4) Crop: We randomly crop the video from the original video with a scale of 71% to 93%.

(5) Text watermark: We randomly add textual watermarks on the random position of videos.

(6) Image watermark: We randomly added visual watermarks on the random position of videos.

(7) Gaussian noise: We add Gaussian blur to the video with a setting of $\sigma = 7$.

(8) Color transform: We randomly select one color transformation from brightness, contrast, saturation, hue and set the parameter to 0.5.

C Visualization of dataset

As shown in Figure 3-22, we provide visualizations of samples from the dataset. From these figures, it can be seen that our dataset possesses diverse content.

Figure 3: ZeroScope [Wang et al., 2023d] generated samples visualization.

Figure 4: I2VGen-XL [Wang et al., 2023c] generated samples visualization.

Figure 5: SVD [Blattmann et al., 2023b] generated samples visualization example.

Figure 6: VideoCrafter [Chen et al., 2024] generated samples visualization.

Figure 7: Pika [pik, 2022] generated samples visualization.

Figure 8: DynamicCrafter [Xing et al., 2023] generated samples visualization.

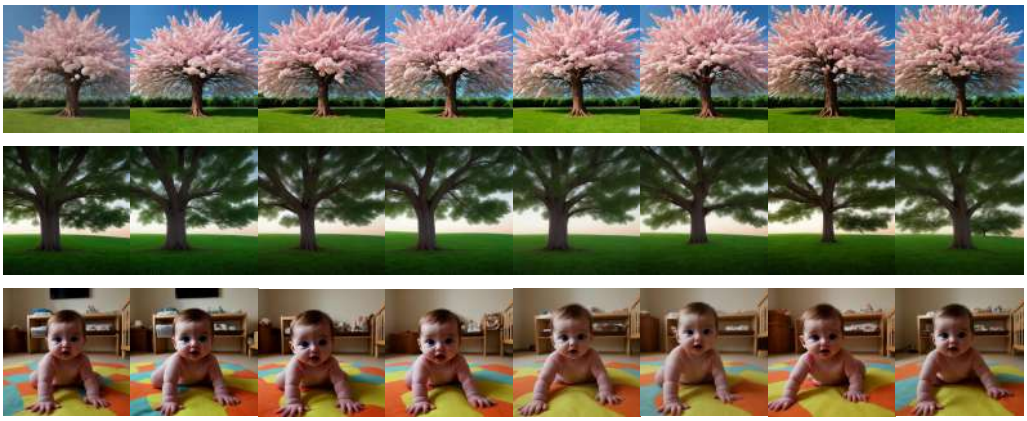

Figure 9: SD Zhang et al. [2023b] generated samples visualization.

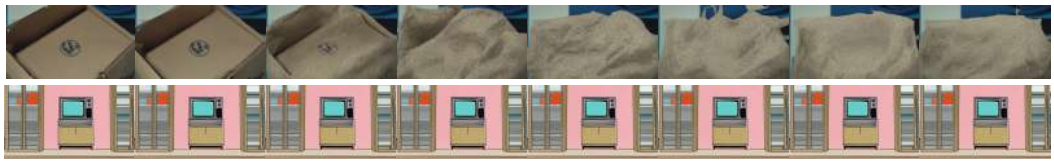

Figure 10: SEINE [Chen et al., 2023c] generated samples visualization.

Figure 11: Latte [Ma et al., 2024b] generated samples visualization.

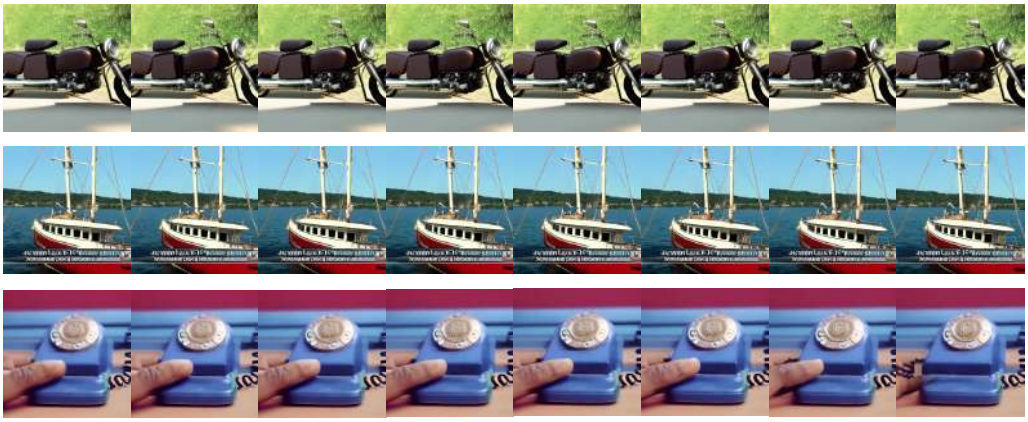

Figure 12: OpenSora [Ope, 2024] generated samples visualization.

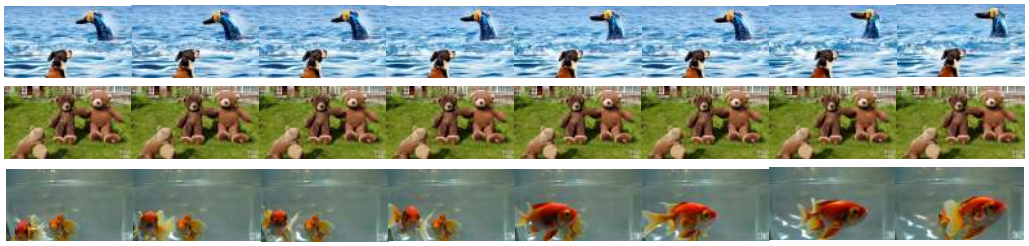

Figure 13: ModelScope [Wang et al., 2023d] generated samples visualization.

Figure 14: MorphStudio [Mor, 2023] generated samples visualization.

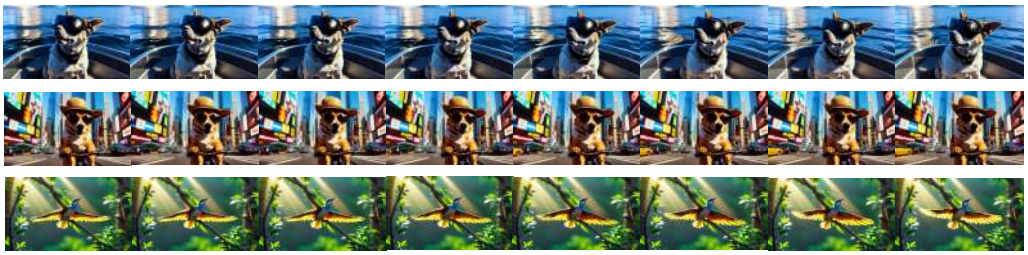

Figure 15: MoonValley [moonvalley.ai, 2022] generated samples visualization.

Figure 16: HotShot [Hot, 2023] generated samples visualization.

Figure 17: Show_1 [Zhang et al., 2023c] generated samples visualization.

Figure 18: Gen2 [Esser et al., 2023] generated samples visualization.

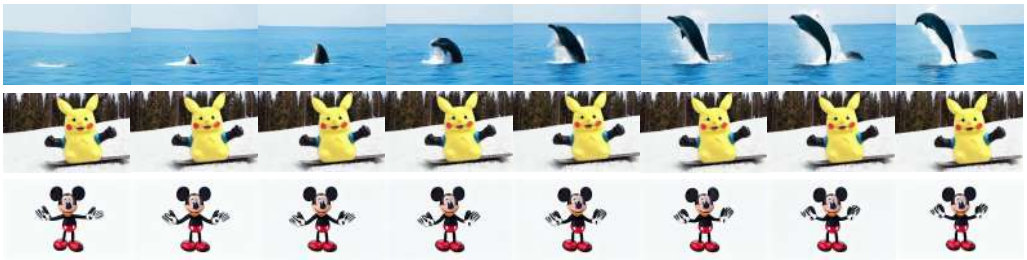

Figure 19: Lavie [Wang et al., 2023a] generated samples visualization.

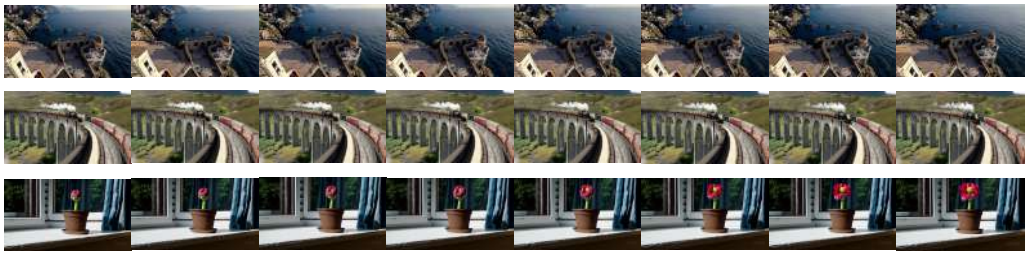

Figure 20: Sora [Brooks et al., 2024] generated samples visualization.

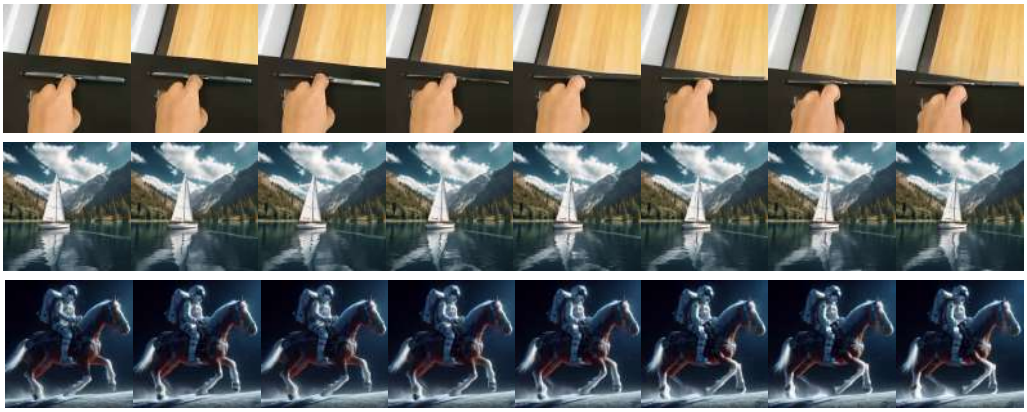

Figure 21: WildScrape Wei et al. [2023], Feng et al. [2023], Xu et al. [2023c] sample visualization.

Figure 22: Crafter [Chen et al., 2023b] generated samples visualization.