中图法分类号:TP392　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)04-1059-13

论文引用格式:Zheng T P, Chen Y X, Wen X Z, Li Y C and Wang Z Y. 2025. Diffusion model-generated video dataset and detection benchmarks. Journal of Image and Graphics, 30(4):1059-1071(郑天鹏, 陈雁翔, 温心哲, 李严成, 王志远. 2025. 扩散模型生成视频数据集及其检测基准研究. 中国图象图形学报, 30(4):1059-1071)[DOI:10.11834/jig.240259]

扩散模型生成视频数据集及其检测基准研究

郑天鹏,陈雁翔*,温心哲,李严成,王志远

合肥工业大学计算机与信息学院,合肥　230061

摘　要:**目的**　扩散模型在视频生成领域取得了显著的成功,目前用于视频生成的扩散模型简单易用,也更容易让此类视频被随意滥用。目前,视频取证相关的数据集更多聚焦在人脸伪造领域上,缺少通用场景的描述,使生成视频检测的研究具有局限性。随着视频扩散模型的发展,视频扩散模型可以生成通用场景视频,但目前生成视频数据集类型单一,数据量少,且部分数据集不包含真实视频,不适用于生成视频检测任务。为了解决这些问题,提出了包含文本到视频(text to video, T2V)和图像到视频(image to video, I2V)两种方法的多类型、大规模的生成视频数据集与检测基准。**方法**　使用现有的文本到视频和图像到视频等扩散视频生成方法,生成类型多样、数量规模大的生成视频数据,结合从网络获取的真实视频数据得到最终数据集。T2V视频生成中,使用15种类别的提示文本生成场景丰富的T2V视频,I2V使用下载的高质量图像数据集生成高质量的I2V视频。为了评估数据集生成视频的质量,使用目前先进的生成视频评估方法对视频的生成质量进行评估,以及使用视频检测方法进行生成视频的检测工作。**结果**　创建了包含T2V和I2V两类生成视频的通用场景生成视频数据集,扩散模型生成视频数据集(diffusion generated video dataset,DGVD)并结合当前先进的生成视频评估方法EvalCrafter和AIGCBench提出了包含T2V和I2V的生成视频质量估计方法。生成视频检测基准使用了4种图像级检测方法CNNdet(CNN detection)、DIRE(diffusion reconstruction error)、WDFC(wavelet domain forgery clues)和DIF(deep image fingerprint)以及6种视频级检测方法I3D(inflated 3D)、X3D(expand 3D)、C2D(convnets 2D)、Slow、SlowFast和MViT(multiscale vision Transformer),其中图像级检测方法无法对未知数据进行有效检测,泛化性较差,而视频级检测方法能够对同一骨干网络实现方法生成的视频有较好的表现,具有一定泛化能力,但仍然无法在其他网络中实现较好的指标。**结论**　本文创建了生成类别丰富、场景多样的大规模视频数据集,该数据集和基准完善了生成视频检测任务在此类场景下数据集和基准不足的问题,有助于促进生成视频检测领域的发展。论文相关数据集与代码下载地址:https://cstr.cn/31253.11.sciencedb.22031和https://github.com/ZenT4n/DVGD。

关键词:视频生成;扩散模型;生成视频检测;提示文本生成;视频质量评估

Diffusion model-generated video dataset and detection benchmarks

Zheng Tianpeng, Chen Yanxiang*, Wen Xinzhe, Li Yancheng, Wang Zhiyuan

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei　230061, China

Abstract: **Objective** Diffusion models have shown remarkable success in video generation. An example is Open AI sora, where we can simply use a text or image to generate a video. However, this convenient video generation also raises concerns regarding the potential abuse of generated videos for deceptive purposes. Existing detection techniques primarily tar-

收稿日期:2024-05-16;**修回日期**:2024-09-08;**预印本日期**:2024-09-15

*通信作者:陈雁翔　chenyx@hfut.edu.cn

基金项目:国家自然科学基金项目(61972127)

Supported by: National Natural Science Foundation of China(61972127)

get face videos. However, datasets specialized for the detection of forged scene videos generated by diffusion models are lacking, and many existing datasets suffer from limitations such as single conditional modality and insufficient data volume. To address these challenges, we propose a multiconditional generated video dataset and corresponding detection benchmark. Traditional detection methods for generated videos often rely on a single conditional modality, such as text or image, which restricts their ability to effectively detect a wide range of generated videos. For instance, algorithms trained solely on videos generated by text to video (T2V) model may fail to identify videos generated by image to video (I2V) model. By introducing multiconditional generated videos, we aim to provide a comprehensive and robust dataset that encompasses T2V and I2V generated videos. The proposed dataset development process involves collecting diverse multiconditional generated videos and real videos downloaded from the Internet. Each generative method produces substantial videos that can be utilized to train the detection model. These diverse conditions and large number of generated videos ensure that the dataset encompasses the broad features of diffusion videos, thereby enhancing the effectiveness of generated video detection models trained on this dataset. **Method** Our generated video dataset can provide training data for detection, allowing the detector to recognize whether the video is AI-generated or not. It uses existing advanced diffusion models (T2V and I2V) to generate numerous videos. Prompt text is one of the keys to generate high-quality videos. We use ChatGPT to generate these prompt texts. To obtain general prompt texts, we set 15 entity words, such as dog and cat, and then combine them with a template as the input for ChatGPT. In this way, we gain 231 prompt texts for generating T2V videos. Different from the T2V model, the I2V model uses the image as the input condition to make a moving image. Using the existing advanced T2V and I2V methods for video generation, combined with real videos obtained from the web to build the final dataset. The generation quality of the video is evaluated using the existing advanced methods of generated video evaluation. We combine the metrics of the EvalCrafter and AIGCBench to evaluate generated videos. For the detection module, we use advanced video detection methods (four image-level video detectors and six video-level video detectors) to evaluate the performance of existing detectors on our dataset with different experiments. **Result** We introduce a generalized scene generated video dataset (diffusion-generated video dataset, DGVD) and corresponding detection benchmarks constructed using multiple generated video detection methods. A generated video quality estimation method containing T2V and I2V is proposed by combining the current state-of-the-art generated video evaluation methods EvalCrafter and AIGCBench. Generated video detection experiments are conducted on four image-level detectors (CNN detection, CNNdet; diffusion reconstruction error, DIRE; wavelet domain forgery clues, WDFC; and deep image fingerprint, DIF) and six video-level detectors (inflated 3D, I3D; expand 3D, X3D; convnets 2D, C2D; Slow; SlowFast; and multiscale vision Transformer, MViT). We set two experiments, within-classes detection and cross-classes detection. For within-classes detection, we train and evaluate the performance of the detectors using the T2V test dataset. For cross-classes detection, we train the detectors on the T2V test dataset but evaluated their performance on the I2V test dataset. Experimental results demonstrate that the image-level detection methods are unable to effectively detect unknown data and exhibit poor generalization. Conversely, the video-level detection methods perform well on the videos generated by methods implemented in the same backbone network. However, they still cannot achieve good generalizability in other classes. These results indicate that existing video detectors are unable to identify the majority of videos generated by the diffusion model. **Conclusion** We work, introduce a novel dataset, DGVD, that encompasses a diverse array of categories and generative scenarios to address the need for advancements in generated video detection. By providing a comprehensive dataset and corresponding benchmarks, we offer a challenging environment for training and evaluating detection models. This dataset and its corresponding benchmarks highlight the current gaps in generated video detection and serve as basis for further progress in the field. We hope to reach significant strides toward enhancing the robustness and effectiveness of generated video detection systems, ultimately driving innovation and advancement in the field. The dataset and code in this paper can be downloded at https://cstr.cn/31253.11.sciencedb.22031 and https://github.com/ZenT4n/DVGD.

Key words: video generation; diffusion model; generated video detection; prompt text generation; video quality evaluation

0 引 言

视频生成是当前人工智能发展的一大领域，特别是与人脸视频生成相关的工作。为了应对人脸生成视频可能带有的欺骗性与误导性，研究者创建了相当多的数据集用于人脸视频检测任务。FF++(FaceForensics++)(Rössler等,2019)是一个由1 000条原始视频序列组成的取证数据集，这些视频序列经过4种自动人脸处理方法处理。DFDC(deepfake detection challenge)(Dolhansky等,2020)挑战赛则为人脸生成视频提供了8种人脸生成方法和数量众多的人脸生成视频。Celeb-DF(Celeb-deepfake forensics)(Li等,2020)则提供了不同种族、年龄和性别的人脸视频数据集。李伟等人(2024)使用图像修复技术篡改人脸，提供了基于修复技术的人脸篡改数据集。这些数据集大多是生成人脸进行换脸，或是重新生成面部的某些区域，但直接生成场景的视频数据集仍然缺乏。为此本文采用生成整个视频的方式，针对通用场景生成视频，从生成对象到实现方法弥补了原有数据集的不足。

近年来，随着图像扩散模型和视频扩散模型的问世，使用生成式AI(artificial intelligence)生成的视觉内容不断增加，人们可以使用简单的文本生成高质量图像或视频，甚至可以达到以假乱真的程度。视频扩散模型的快速发展使得生成高质量视频变得越来越容易，通用场景视频的生成质量越来越高。VDM(video diffusion model)(Ho等,2022)首次将扩散图像模型中的UNet网络扩展为3D UNet，并通过图像视频协同训练的方式实现了文本到视频的生成。Blattmann(2023)提出了Video LDM模型，它仿效LDM(latent diffusion model)将图像映射到潜在空间的方法，将视频映射到潜在空间中，从而加速了视频生成的过程。Peebles和Xie(2023)使用Transformer取代了扩散模型中的UNet网络，提出了DiT(diffusion Transformers)模型，进一步提高了生成速度和在一些分辨率下的生成效果。视频生成技术将会向着高分辨率和高质量、可解释的生成控制的方向发展(刘安安 等,2024)。这些方法的出现让扩散视频模型可以生成内容丰富、高质量的视频，使得原有的人脸生成视频数据集已经无法满足现有的生成视频检测工作，亟需通用场景的生成视频数据集来解决这方面问题。

总的来说，目前生成视频数据集存在以下问题。首先，生成视频数据集和基准多集中在人脸领域，而通用场景生成视频数据集和基准相对较少。其次，目前通用场景的生成视频数据集和基准只包含文本到视频(text to video，T2V)或图像到视频(image to video，I2V)方法生成的视频，如EvalCrafter(Liu等,2024)和AIGCBench(artificial intelligence generated content benchmark)(Fan等,2023)分别针对T2V视频和I2V视频的生成。而当前模型能够接受除文本模态以外的输入，例如I2VGen-XL(Zhang等,2023)、Mm-Diffusion(Rombach等,2022)、Codi(Tang等,2023)等。这些模型可以利用文本、图像或音频等多种模态的信息来生成视频。OpenAI的sora(Brooks等,2024)在T2V、I2V等方法生成的视频的时间一致性和生成质量上也取得了显著的提升。可见单一类别的数据集只涉及了目前视频生成中的一部分，缺乏多样性。最后，通过表1可以发现，目前大部分通用场景生成视频数据集并不包含真实视频，仅有生成视频，无法直接用于生成视频检测任务。

为了解决这些问题，本文创建了扩散模型生成视频数据集(diffusion generated video dataset，DGVD)和检测基准。该数据集和基准包含3个部分：数据集、数据质量评估模块和生成视频检测模块。

表1 与现有生成视频数据集对比

Table 1 Comparison of existing generated video datasets

数据集	视频类型	T2V	I2V	生成方法	生成视频	真实视频
EvalCrafter	场景	√	×	13	9 100	0
AIGCBench	场景	×	√	5	3 928	0
GVF	场景	√	×	4	3 856	964
DGVD(本文)	场景	√	√	4	108 615	27 058

注："√"和"×"分别表示包含和不包含。

DGVD旨在收集目前主流的扩散模型生成的视频，并将其用于生成视频的检测任务。为了提高数据的类型多样性和质量，数据集中的生成视频来源于4种不同的生成技术，包括两种文本到视频生成(T2V)方法和两种图像到视频生成(I2V)方法。在T2V生成视频中，使用了15种类别的提示文本，生成场景丰富的视频。在I2V视频中，使用高质量的自然风景图像，生成了高质量的自然场景视频，因此I2V生成视频在质量上要高于T2V生成视频。数据质量评估模块参考了EvalCrafter(Liu等，2024)和AIGCBench(Fan等，2023)对生成视频质量的评价标准，分别从中提取部分指标，从视频质量、时间一致性、动态评分和条件—视频对齐4个方面对数据集的生成视频进行了评估。生成视频检测模块使用了4种图像级检测方法和6种视频级检测方法，从空间和时间两个维度对生成视频进行检测，并验证了现有检测器的泛化能力。通过构建多样化的数据集和全面综合的评估指标，推动生成视频检测领域的进步，弥补当前领域在数据集和基准上的缺失。论文相关数据集与代码下载地址：https://cstr.cn/31253.11.sciencedb.22031和https://github.com/ZenT4n/DVGD。

1 现有数据集与视频生成

1.1 生成视频数据集

目前公开的视频检测数据集大多与人脸相关，而通用场景相关的数据集主要集中在视频生成的评估基准上，因此通用场景相关的视频检测数据集仍然稀缺。表1列举了目前通用场景领域的视频数据集。EvalCrafter(Liu等，2024)和AIGCBench(Fan等，2023)分别提出了T2V和I2V的生成视频流程，但其直接提供的视频数量较少，且数据集并不包含真实视频，因此无法直接用于生成视频检测工作。GVF(generated video forensics)(Ma等，2024)是基于现有的生成视频评估数据基准FETV(fine-grained evaluation of open-domain text-to-video)(Liu等，2023b)生成的数据集。它主要用于生成视频检测，但同样存在以下缺点：数量不足，无法支持大规模训练；方法类型单一，所有视频都为T2V类型，没有考虑其他类型的生成方法。而本文所创建的数据集同时考虑了T2V生成方法与I2V生成方法，并且生成了大量的生成视频，弥补了当前数据集数据量与类型多样性的不足。

1.2 视频生成方法

数据集的生成过程如图1所示，包含T2V视频生成模块、I2V视频生成模块以及真实视频处理模块。T2V视频生成模块由提示文本的生成和视频生成两个部分组成。相较而言，I2V视频生成模块较为简单，仅包含输入图像数据和图像到视频生成。

文本生成视频(T2V)模型只需提供简单的文本提示，就能够生成相应的视频。

与T2V模型不同，I2V模型使用图像作为输入，通过让图像中物体运动或镜头移动实现视频效果。SVD(stable video diffusion)(Blattmann等，2023)就可将图像生成为有运动逻辑的视频。然而，仅使用图像作为输入并不能保证模型生成理想的视频。因此，为了提高视频生成的可控性，一些I2V模型会引入文本提示来进一步控制视频内容。例如，Animate-diff(Guo等，2024)、Videocrafter(Chen等，2023)和I2VGen-XL(Zhang等，2023)等扩散视频生成模型可以同时实现T2V和I2V功能，生成更具控制效果的视频。

目前，T2V和I2V的扩散视频生成方法多种多样，生成质量良莠不齐。在实验了多种视频生成方法后，为了选择生成数量多、质量高且具有可调整性的视频，综合考虑了现有生成模型生成视频的质量、视频推理时间以及可调整性(能否修改视频分辨率、时长等参数)等因素。最终选择了ModelScope(Wang等，2023a)、ZeroScope、SVD和I2VGen-XL这4个方法作为数据集的视频生成方法。

ModelScope模型在LDM的基础上加入了时空卷积与时空注意力模块，相较于非diffusion视频生成模型和基于像素的diffusion视频生成模型，它的生成速度更快、文本对齐程度更高。ZeroScope则是在Modelscope的基础上进行了微调，实现了无水印、高质量的视频生成。SVD是一个用于高分辨率视频生成的图像到视频生成扩散模型。而I2VGen-XL则利用专门设计的UNet在隐空间中进行时空建模，然后通过解码器重建出最终视频。

1.3 文本到视频生成

1.3.1 用于视频生成的提示文本

目前T2V生成视频方法所使用的提示文本有两种：1)基于现有的文图数据集或文本视频数据集中

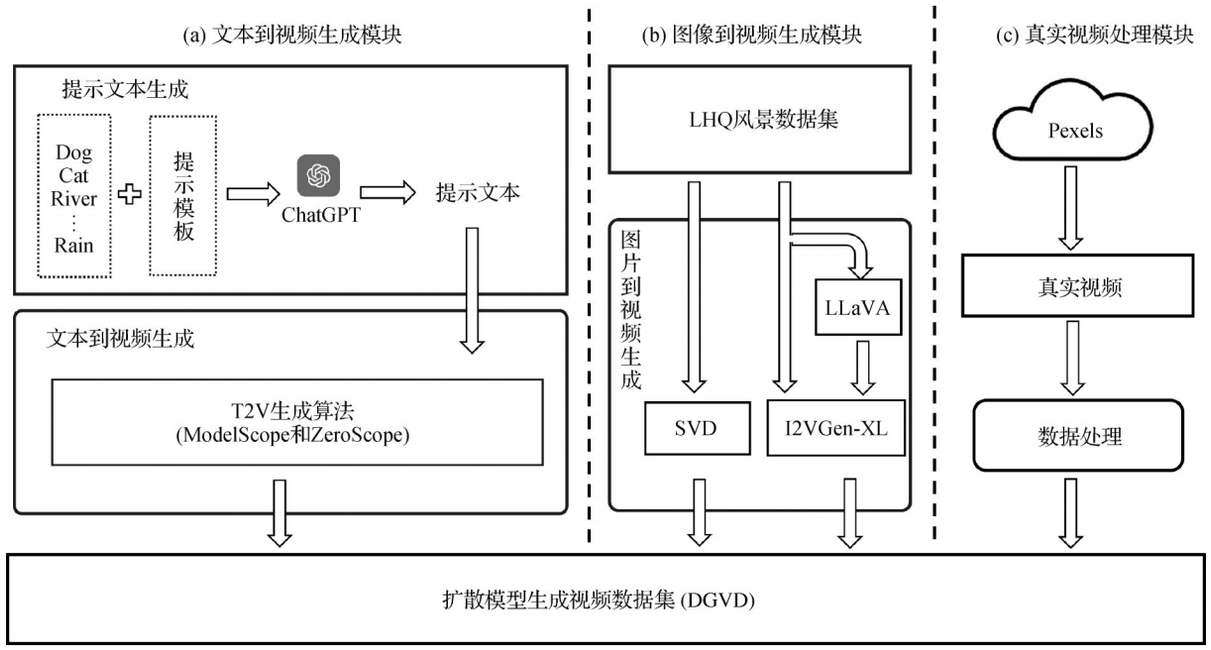


图 1　不同方法的视频生成方案以及真实视频处理流程

Fig. 1　Video generation solutions for diverse methodologies and real video processing pipeline

((a)T2V generation;(b)I2V generation;(c)real video processing)

的文本进行视频生成（Liu 等，2023b）；2）使用不同的物体、场景和事件等组合不同的文本，然后由大语言模型润色和人工选择的半自动生成（Liu 等，2024）。在构造生成视频检测数据集时，使用现有文本—图像数据集或者文本—视频数据集的文本作为描述文本。这不符合人类在使用 T2V 模型生成视频时的习惯。人类在使用生成视频时通常直接使用更为简单和清晰的文本作为生成模型的提示输入，并非使用图像或视频的文本描述，所以数据集的提示文本生成沿用了第 2 种方法。

大语言模型如 GPT-4（OpenAI 等，2024），在自然语言处理领域展现出了惊人的效果。经过有目的的提示工程，GPT-4 甚至可以生成用于专业领域的文本。如图 1(a)所示，数据集的 T2V 的提示文本，来源于 OpenAI ChatGPT 生成。为了增加生成文本的多样性和描述场景的多样性，设置了 15 个实体类别，包括{dog, cat, river, …}等 15 个类别。使用人工书写的提示文本作为指导和提示模板："Can you generate n sentences about [class], require the sentence has the word [class]."，让 ChatGPT 为每一个类别生成 10～20 句提示文本。这样的模板不仅提供了生成文本的数量，还要求 ChatGPT 在生成文本中包含所需的实体类别，从而保证最终生成的视频不会受到其他提示词的影响。

1.3.2　视频生成

基于上述生成的提示文本，使用预训练的 T2V 模型可以生成与文本描述相关的视频。在经过测试后，筛选掉无法正确生成目标类别的提示文本，最终保留了共 231 句提示文本。需要注意的是，ModelScope 预训练模型对同一文本生成的视频存在随机性，即使每次输入的文本相同，但生成的视频是不相同的。基于设计的生成管道，每一个实体类别都生成了 3 000 条视频。与 ModelScope 使用相同的流程，ZeroScope 为每一个实体类别都生成了 2 000 条视频。

1.4　图像到视频生成

目前大部分视频生成模型对于有规律、运动简单如瀑布、流水和烟花等事物的视频生成效果更好。其作为视频背景或插入到真实的视频中更不易察觉。为了保证生成视频的质量，使用 LHQ（landscapes high-quality）（Skorokhodov 等，2021）数据集作为 I2V 模型的输入。LHQ 是自然风景数据集，包含了 90 000 幅高质量、高分辨率的自然风景图像，其图像从 Unsplash（60 k）和 Flickr（30 k）获取。

SVD 可以使用不同分辨率的图像输入，同样输出不同分辨率的图像，但经过测试，SVD 生成高分辨率的视频效果最好，降低分辨率会使得生成视频的稳定性下降，出现不正常的光晕或画面变形。

I2VGen-XL可以使用不同分辨率的图像输入，同样输出不同分辨率的图像，但与SVD不同的是，I2VGen-XL每幅图像都需要一条文本提示。为了提高I2VGen-XL的生成质量，每一幅图像都提供了符合该图像内容的文本描述。如图1(b)所示，每幅图像输入I2VGen-XL的同时，使用LLaVA(Liu等，2023a)大模型对图像进行描述，将描述文本与对应图像同时作为I2VGen-XL模型的输入。与SVD相同，调整分辨率和时长会使得视频出现不正常的光晕或画面变形。

SVD和I2VGen-XL会基于输入的图像，推理图像内容的运动，并以此生成视频。输入的图像和生成的视频会构成真假图像视频对。

1.5 真实视频收集与处理

在生成视频检测任务中，除了需要生成的视频，还需要相应的真实视频。如图1(c)所示，真实视频在提供高质量且完全免费素材的网站Pexels上下载，总共获取了2 537条分辨率为1 080 P的视频，每个视频长度在3 s到30 min不等。视频的检索使用树林、街道和雨天等与生成数据集中视频描述接近的标签，并加上白天和夜晚两个时间场景的标签，最终两两组合作为搜索条件。下载的视频并不完全适用于生成视频检测，需要进行数据处理工作。真实视频的处理包括：视频删除和视频分割。在下载的真实视频中，存在部分视频重复和视频内容与目标标签不符的情况，通过人工检索将这部分视频删除。另外，由于原始真实视频长度跨度较大，存在30 min以上的长视频，视频检测模型并不需要时间过长的视频，大部分生成模型的视频时长仅在1 min以内。因此，为了提高数据利用率，对每个视频进行了分割，分割大小为4 s。通过处理，最终获得27 058条时长在4 s内、分辨率为1 920×1 080像素的高清视频用于生成视频检测任务。

2 数据统计

2.1 视频数据统计

数据集总共使用231条提示文本，全部都由ChatGPT生成。通过这些提示文本使用T2V模型生成了75 000条视频。抽取LHQ数据集中的33 615幅图像在I2V模型上生成了对应数量的视频。所有视频生成工作均在一张NVIDIA A800 GPU上进行。在ModelScope上生成了45 000条视频，ZeroScope上生成了30 000条视频；在SVD和I2VGen-XL分别生成了19 998条和13 617条视频。最终，生成的视频数量达到108 615条，并经过人工筛选与处理，从真实视频数据中得到27 058条视频，为后续研究提供了丰富的数据内容。

表2统计了每个生成方法生成的视频数量、视频分辨率、帧率和时长等信息，以及每种方法需要的条件。

表2 视频数据统计信息

Table 2 Video statistics

方法	条件	数量	分辨率/像素	帧速率/(帧/s)	时长/s
ModelScope	文本	45 000	256×256	8	3
ZeroScope	文本	30 000	768×256	8	3
SVD	图像	19 998	1 024×1 024	8	4
I2VGen	图像+文本	13 617	1 280×704	8	4
真实视频	/	27 058	1 920×1 080	24	2～4

2.2 数据属性

数据集保留了生成视频时使用的提示文本和图像。数据集$\boldsymbol{D}$中包含视频$\boldsymbol{v}$、图像$\boldsymbol{i}$和文本$\boldsymbol{t}$，不同的子数据集有不同的组合。视频统一格式为MP4，图像统一为JPG，文本与视频名称和图像名称存储在文本文件中。

对于T2V数据集，每个数据项包含视频和文本的文本视频对$\{\boldsymbol{v},\ \boldsymbol{t}\}\in\boldsymbol{D}_{\mathrm{T2V}}$，表示视频$\boldsymbol{v}$与其对应的提示文本$\boldsymbol{t}$。

对于I2V数据集，生成视频使用条件为图像或者文本—图像对，使用SVD生成视频的子数据集的数据项表示为$\{\boldsymbol{v},\ \boldsymbol{i}\}\in\boldsymbol{D}_{\mathrm{I2V}}$，表示使用的真实图像$\boldsymbol{i}$和其生成的视频$\boldsymbol{v}$，最终组成真假图像视频对。而I2VGen-XL使用图像和文本作为生成条件，所以数

据项表示为$\{\boldsymbol{v},\ \boldsymbol{i},\ \boldsymbol{t}\}\in\boldsymbol{D}_{\mathrm{I2V}}$，表示生成的视频$\boldsymbol{v}$，使用的真实图像$\boldsymbol{i}$和提示文本$\boldsymbol{t}$，其同样能构成真假图像视频对。真实视频的数据项表示为$\{\boldsymbol{v}\}\in\boldsymbol{D}_{\mathrm{R}}$。图 2 展示了数据集的样例与每种生成方法使用的输入。

本文对不同生成方法提供了更为详细的输入和输出样例。T2V 方法提供了输入的提示文本以及输出的部分视频帧。I2V 方法则提供了输入的图像以及输出的部分视频帧。图 3—6 分别为 ModelScope 文本输入及其对应的生成样例、ZeroScope 输入文本及其对应的生成样例、Stable video diffusion 的输入图像及其对应的生成样例、I2VGen-XL 的输入图像及其对应的生成样例。

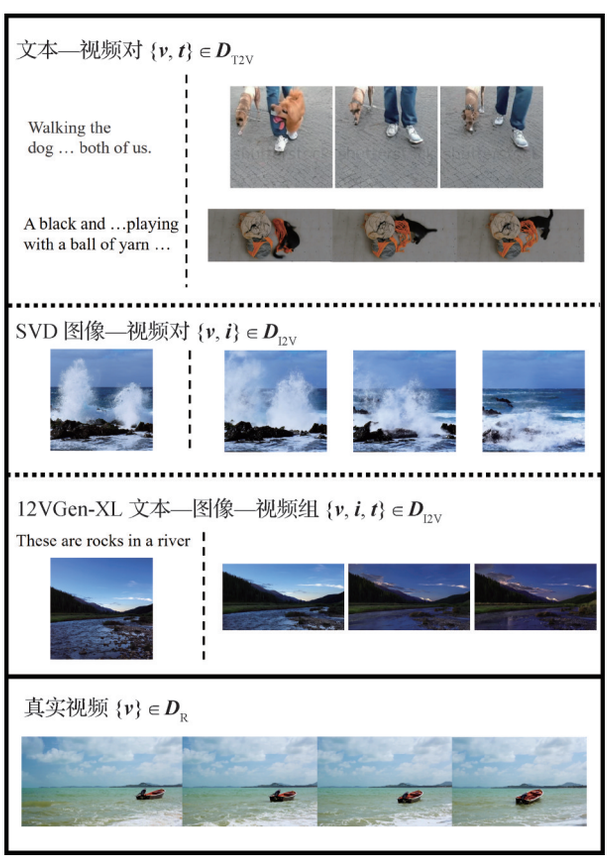


图 2　几种生成方法对应的输入与输出视频样例

Fig. 2　Sample inputs and their corresponding output videos for different generation methods

数据集$\boldsymbol{D}$包含真实视频和生成视频两个子数据集，分别用 real 和 fake 表示。生成视频子数据集根据类别和生成方法划分为 T2V-ModelScope，T2V-ZeroScope，I2V-SVD 和 I2V-I2VGen-XL 4 个子数据集。T2V 和 I2V 表示视频类别，后缀表示生成方法。

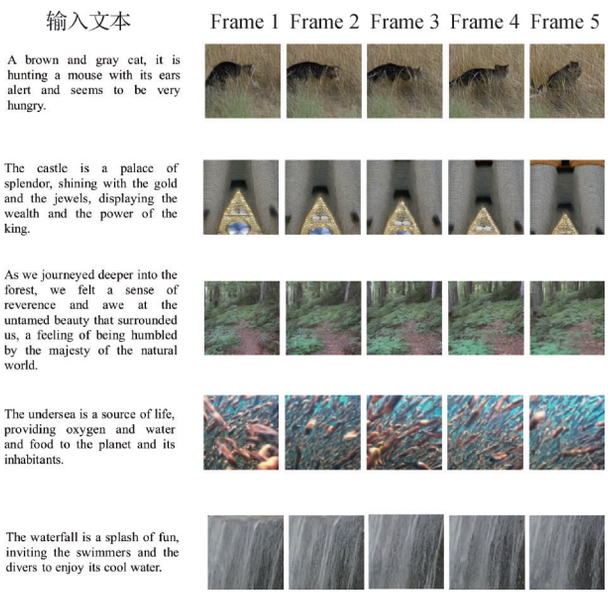


图 3　ModelScope 输入文本及其对应的生成样例

Fig. 3　Text inputs of ModelScope and generated samples

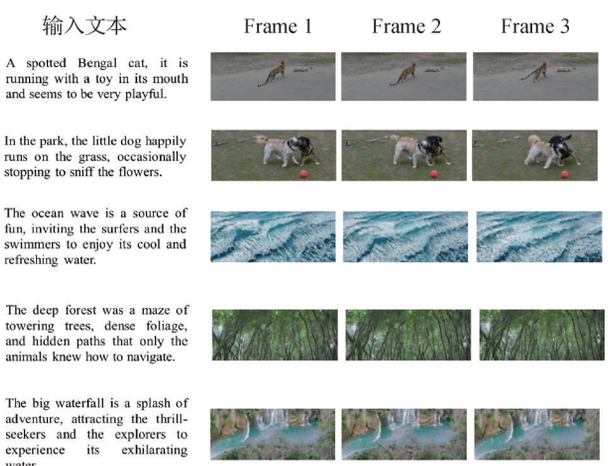


图 4　ZeroScope 输入文本及其对应的生成样例

Fig. 4　Text inputs of ZeroScope and generated samples

3　数据质量验证

随着 AIGC（artificial intelligence generated content）的发展，生成评估领域的工作也在不断推进，为了检验数据集的生成质量，从众多的生成视频评估方法中选择了 EvalCrafter（Liu 等，2024）和 AIGCBench（Fan 等，2023）。EvalCrafter 是 T2V 生成视频基准，从不同方面对 T2V 模型进行了评价，包括生成视频的视觉质量、文本—视频对齐、运动质量和时间一致性等方面，每个方面都集合了多种视频评估方法，通过加权平均的方式获得每一个方面的评

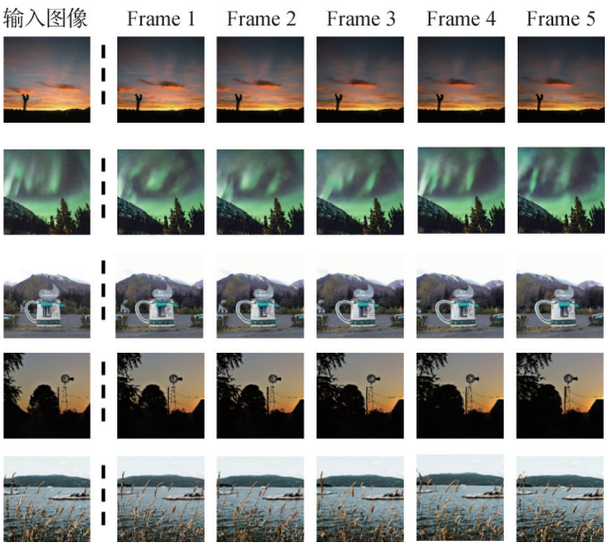


图5　Stable video diffusion的输入图像及其对应的生成样例

Fig. 5　Image inputs of Stable video diffusion and generated samples

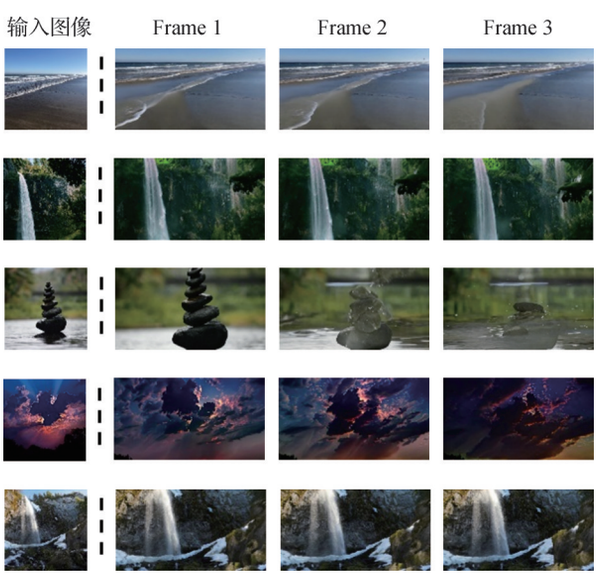


图6　I2VGen-XL的输入图像及其对应的生成样例

Fig. 6　Image inputs of I2VGen-XL and generated samples

分。而AIGCBench针对I2V视频在控制—视频对齐、运动效果、时间一致性和视频质量等方面进行了评估。AIGCBench使用现有的文本—图像和文本—视频数据集进行视频生成，所以部分评估指标包含了视频参考，为了适应本数据集，删去了需要参考视频的评估指标。

EvalCrafter针对其提示文本的分类设置了部分关于人脸和人体姿势相关的评估指标，而本数据集的提示文本设计与EvalCrafter不同。因此，为了适应数据集，去除了EvalCrafter中关于人脸和人体姿势的相关评估指标，删除了人脸和人体姿势相关权重，并提高了其他指标的权重。

在综合了EvalCrafter和AIGCBench的评估指标后，对T2V和I2V生成视频形成了新的指标。它包含4个方面：视频质量、时间一致性、动态效果和条件—视频对齐。关于T2V和I2V所含有的评估指标可见表3。评估指标包含：1）视频质量：Inception Score（IS），Video Quality Assessment（VQA）（VQAa，VQAt）；2）时间一致性：temp_score，warping_error；3）动态评分：flow_score；4）条件控制对齐，根据T2V和I2V分为：（1）文本—视频对齐：Text-Video CLIP，sd_score，blip_bleu；（2）图像—视频对齐：MSE（mean squared error），SSIM（structure similarity index measure），Image-Video CLIP。

表3　T2V和I2V视频评估指标

Table 3　T2V and I2V video metrics

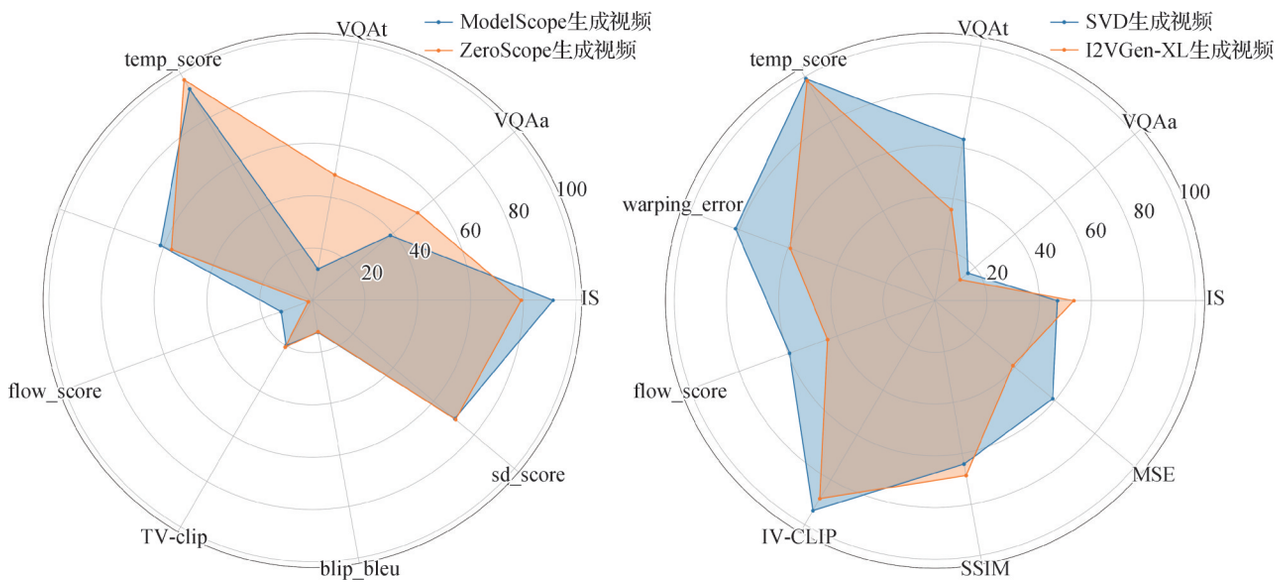
方面	指标	T2V	I2V
视频质量	IS ↑	√	√
	VQA ↑ (VQAt, VQAa)	√	√
时间一致性	temp_score ↑	√	√
	warping_error ↓	√	√
动态评分	flow_score ↑	√	√
图像—视频对齐	MSE ↓	×	√
	SSIM ↑	×	√
	IV-CLIP ↑	×	√
文本—视频对齐	TV-CLIP ↑	√	×
	blip_bleu ↑	√	×
	sd_score ↑	√	×

注："√"表示包含，"×"表示不包含，"↑"表示值越高越好，"↓"表示值越低越好。

各评估指标具体说明如下：Inception Score（IS）：计算图像清晰度和多样性；Video Quality Assessment（VQA）（VQAa，VQAt）：评估视频的美学指标（VQAa）和技术性指标（VQAt）；temp_score：两帧之间的一致性；warping_error：每次取两帧，一帧进行部分像素的变化，计算两帧最小距离；flow_score：视频两帧之间的平均稠密流；sd_score：生成视频与相同文本下，SD（stable diffusion）模型的嵌入相似程度；blip_bleu：提示文本与视频描述文本的一致性。MSE：输入图像与第1帧之间的均方误差；SSIM：输入图像与第1帧之

间的结构相似性；Text-Video CLIP 和 Image-Video CLIP：条件（本文，图像）和视频的CLIP相似性评分。

本文使用新的评估指标分别为T2V和I2V生成视频进行了质量评估验证。验证结果如图7所示，在T2V生成视频中，ZeroScope生成的视频在综合评分上高于ModelScope，而I2V生成视频中，SVD生成的视频在多项评分上都高于I2VGen-XL生成的视频。



(a) T2V视频质量　　(b) I2V视频质量

图7　生成视频质量评估结果

Fig. 7　Generated video quantity evaluation results ((a) T2V video quantity; (b) I2V video quantity)

4　检测基准

4.1　生成视频检测实验

DGVD包含了两类扩散视频生成方法，为了全面评估生成视频检测效果，设计了类内检测任务和跨类检测任务。类内检测任务主要评估训练数据与测试数据为相同类别方法的检测情况，如ModelScope和ZeroScope同为T2V方法；跨类检测任务主要评估训练数据与测试数据为不同类别方法的检测情况。使用T2V方法生成的视频所训练的模型在I2V视频上的表现，评估模型的泛化能力。

在使用图像级检测器时，为了充分利用视频信息，使用关键帧提取方法，提取9帧图像，合并作为输入。评估指标使用了二分类问题常用的AP（average precision）、ACC（accuracy）、AUC（area under the curve）和F1值。

DGVD对4类生成方法均划分了训练集、验证集和测试集。每一个检测器都使用ModelScope生成的视频和真实视频划分的训练集对进行训练，随后在ModelScope和ZeroScope划分的测试集进行类内检测任务，在SVD和I2VGen-XL划分的测试集中进行跨类检测任务。

4.2　检测方法

基准实现了10种检测器来进行真伪检测任务，选择这些方法主要有以下标准。1）考虑经典的应用在视频或图像上的方法，如I3D（inflated 3D）（Carreira和Zisserman，2017）等，或是当前广泛使用的一些标准方法；2）将检测器分为图像级检测器和视频级检测器，在视频级上，又分为基于CNN（convolutional neural network）的检测器和基于Transformer的检测器；3）选择的方法是易于实现的。如表4所示，使用的方法有CNNDet（CNN detection）（Wang等，2020），WDFC（wavelet domain forgery clues）（Deng等，2023），DIF（deep image fingerprint）（Sinitsa和Fried，2024），DIRE（diffusion reconstruction error）（Wang等，2023b），I3D（Carreira和Zisserman，2017），C2D（convnets 2D）（Wang等，2018），Slow和SlowFast（Feichtenhofer等，2019），X3D（expend 3D）（Feichtenhofer，2020）和Mvit（multiscale vision Transformer）

(Fan等,2021)等。

CNNDet用于区分图像是真实图像还是由CNN生成的图像。它使用ProGAN生成的图像数据集进行训练,并验证了在CNN生成的图像上训练的取证模型对其他CNN合成方法也能表现出惊人的泛化程度。

表4 检测器

Table 4 Detector

方法	检测级别	预训练数据集
CNNDet	图像级	ProGAN
DIRE	图像级	LSUN
WDFC	图像级	DDPM
DIF	图像级	SD14
I3D	视频级	Kinetics-400
X3D	视频级	Kinetics-400
C2D	视频级	Kinetics-400
Slow	视频级	Kinetics-400
SlowFast	视频级	Kinetics-400
MViT	视频级	Kinetics-400

DIRE针对扩散生成的图像进行训练,使用新的图像表示方式,扩散重建误差(DIRE),它通过预先训练的扩散模型测量输入图像与其重建图像之间的误差,检测图像是否为扩散模型生成。

WDFC针对扩散生成的图像,应用离散傅里叶变换和Haar离散小波变换(Haar discrete wavelet transform,HDWT)获取生成图像的高频伪影,利用该伪影作为生成图像的检测线索。

DIF利用CNN会产生图像指纹的现象来训练一个额外的CNN,以模仿目标图像生成器的指纹,深度图像指纹(DIF)。利用DIF作为生成图像的检测线索。

I3D提出了膨胀3D卷积,将非常深的卷积网络的滤波器和池化内核扩展到3D,从而实现对视频数据的特征提取。

X3D是一个高效视频网络,不同于一般的3D模型仅扩展时间维度,它沿着空间、时间、宽度和深度等多个网络轴逐步扩展一个微小的二维架构。

SlowFast使用快慢网络架构,实现两个分支分别对时间与空间维度进行处理分析,slow分支使用较少的帧数以及较大的通道数来学习空间语义信息,而fast分支使用较大的帧数以及较少的通道数来学习运动信息。而Slow则是仅仅使用SlowFast中的slow分支。

Mvit基于Transformer实现,结合多尺度特征层次结构,使得早期层次可以高空间分辨率下运行,对简单的低级视觉信息进行建模。而更深的层可以有效地关注空间粗糙但复杂的高级特征,以对视觉语义进行建模。

由此生成视频检测方法可以分为基于空间的检测方法和基于时空的检测方法。空间检测方法考虑视频的空间信息,使用图像级检测器对生成视频进行检测。空间检测能够验证视频在图像级别上的生成能力是否欺骗图像级的检测器,判断当前视频生成模型是否过于重视视频的生成而忽略了图像生成的能力。时空检测方法考虑目前经典的视频方法,大部分视频生成模型在时间一致性上表现不佳,难以生成动作流畅的视频。基于时空的检测方法,使用视频级检测器在提取视频空间特征的同时,关注时域信息,提取视频整体特征进行检测任务。

4.3 结果分析和评价

4.3.1 类内检测结果

使用训练好的检测器,总共在4个测试数据集上对模型进行了评估。根据表5,对于类内检测,所有的模型都在ModelScope测试集中得到了良好的表现,特别是图像级检测器的性能。但是图像级检测器对ZeroScope生成视频的检测效果有所下降,特别是CNNDet和DIRE两个方法,准确度均下降到15%左右,而使用伪影或指纹进行检测的方法WDFC和DIF效果比CNNDet和DIRE要好。

而视频级检测器在ModelScope和ZeroScope上都有着很好的表现,仅在ZeroScope视频检测结果比ModelScope视频略差,I3D、X3D、C2D和SlowFast等方法在ZeroScope视频上表现更好。

4.3.2 跨类检测结果

目前大多数生成视频检测器使用T2V视频进行训练。该实验在T2V训练集上训练的模型,在I2V测试集上进行测试,旨在评估使用T2V视频训练的检测模型在I2V视频上的泛化性能。从表5结果可见,两类检测器对SVD和I2VGen-XL生成视频的检测能力明显下降,这表明在T2V视频数据上训练的检测器,对于I2V视频数据的泛化能力不足。对于

SVD 视频，CNNDet、DIRE、WDFC、DIF、I3D 和 X3D 这几个检测器的 AP 表现较高，它们对于真实视频的检测能力仍有保留，但无法很好地检测生成视频。C2D、Slow、SlowFast 和 Mvit 表现出了较高的 ACC，但是 AP 较低。而几乎所有的检测器在 I2VGen-XL 生成视频的检测能力上表现都不好。值得注意的是，WDFC 和 DIF 两个使用伪影或指纹的方法，在 I2VGen-XL 视频检测中表现出了较好的结果。这些结果突出了泛化性能的重要性，以及在不同生成模型上的检测器的挑战，为进一步提高模型的泛化能力提供了有益的参考和启示。

4.3.3 总体评价

图像级检测器对于已知数据能够达到很好的效果，在 ModelScope 测试数据中表现良好，但是对于同一骨干网络下的 ZeroScope 测试数据表现不佳，出现高 AP 的现象，说明此类检测器只学习到了已知数据的特征，缺乏泛化性。而视频级检测器对于同一骨干网络下的数据都有较好的表现，如表 5 中，视频级检测器在 ModelScope 和 ZeroScope 都有着很好的表现，而 I2V 视频数据上表现不足，效果基本与图像级检测器持平。说明视频级检测器有着更好的特征提取能力，但仍然无法满足目前的视频检测工作。

表 5　视频检测结果

Table 5　Video detection results

/%

方法	ModelScope				ZeroScope			
	ACC	AP	F1	AUC	ACC	AP	F1	AUC
CNNDet	**100.00**	99.98	99.97	**100.00**	15.27	97.18	26.49	88.09
DIRE	99.94	**100.00**	99.92	**100.00**	15.33	94.78	26.52	80.75
WDFC	99.29	99.65	99.43	99.97	76.55	88.27	69.76	88.63
DIF	97.56	98.35	97.99	97.95	87.51	97.54	85.91	87.60
I3D	99.01	99.41	98.68	99.21	95.44	99.18	87.01	97.31
X3D	99.15	99.50	98.86	99.32	**98.63**	**99.75**	95.71	**99.19**
C2D	98.52	99.11	98.03	98.81	97.74	98.13	97.56	97.94
Slow	97.41	98.38	96.61	97.89	81.17	84.37	82.67	82.82
SlowFast	98.14	98.83	97.55	98.48	97.76	98.07	**97.57**	97.94
Mvit	99.99	99.99	**99.99**	99.99	81.94	96.74	62.86	89.36
方法	SVD				I2VGen-XL			
	ACC	AP	F1	AUC	ACC	AP	F1	AUC
CNNDet	21.30	75.41	35.09	45.20	28.41	75.20	44.24	59.43
DIRE	21.13	**82.57**	35.11	56.36	28.44	84.96	44.27	72.81
WDFC	57.80	78.90	2.22	58.31	55.92	77.96	33.46	62.02
DIF	**73.65**	64.05	55.53	**69.09**	**76.86**	**85.29**	**76.89**	**81.08**
I3D	21.33	78.71	35.11	50.00	28.62	71.63	44.35	50.12
X3D	21.35	78.72	35.13	50.04	28.63	71.64	44.35	50.13
C2D	57.51	42.51	73.01	50.02	28.76	71.87	44.26	50.31
Slow	57.35	42.47	72.88	49.89	28.36	71.68	44.04	49.93
SlowFast	57.46	42.50	72.95	50.00	29.31	72.05	44.42	50.65
Mvit	57.55	42.55	**73.04**	50.05	28.65	71.64	44.35	50.14

注：加粗字体表示各列最优结果。

5 结 论

本文提出了一个功能模块和生成视频类型丰富的生成视频检测框架，包含扩散模型生成视频数据集、视频质量评估和生成视频检测3个模块。数据集为通用场景的生成视频检测任务提供了训练数据，解决了生成视频数据集的稀缺问题。并且结合EvalCrafter和AIGCBench为数据集进行了质量评估。在检测基准中，可以使用多种检测器进行生成视频的检测工作，并提供了统一的评估指标。相比其他数据集，本文创建的数据集生成视频数量更多，达到10 k以上，生成方法更为丰富和多样，同时包含T2V方法和I2V方法，为生成视频检测工作提供了更全面的训练数据。实验结果还表明，目前视频检测器即使在同一类型或同一方法上的检测结果表现良好，但在不同方法生成的视频上的泛化性较差。

由于当前生成模型并不能为所有场景都提供高质量的视频生成，目前T2V数据集主要局限于生成质量较高的场景，动物上只有猫狗等，大部分场景都是自然环境。I2V模型则只使用了LHQ数据集，其包含的场景多为自然风景，将来可以扩充其他场景，进一步提高场景的复杂性。随着视频生成模型的不断发展，后续应能够扩充包含更多场景、事件的高质量视频，提高数据集的多样性。

参考文献(References)

Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, Levi Y, English Z, Voleti V, Letts A, Jampani V and Rombach R. 2023. Stable video diffusion: scaling latent video diffusion models to large datasets [EB/OL]. [2024-05-16]. https://arxiv.org/pdf/2311.15127.pdf

Brooks T, Peebles B, Holmes C, DePue W, Guo Y, Jing L, Schnurr D, Taylor J, Luhman T, Luhman E, Ng C, Wang R and Ramesh A. 2024. Video generation models as world simulators [EB/OL]. [2024-05-16]. https://openai.com/research/video-generation-models-as-world-simulators

Carreira J and Zisserman A. 2017. Quo Vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Chen H X, Xia M H, He Y Q, Zhang Y, Cun X D, Yang S S, Xing J B, Liu Y F, Chen Q F, Wang X T, Chao W and Shan Y. 2023. VideoCrafter1: open diffusion models for high-quality video generation [EB/OL]. [2024-05-16] https://arxiv.org/pdf/2310.19512.pdf

Deng Y F, Deng X, Duan Y P and Xu M. 2023. Diffusion-generated fake face detection by exploring wavelet domain forgery clues//Proceedings of 2023 International Conference on Wireless Communications and Signal Processing (WCSP). Hangzhou, China: IEEE: 1-6 [DOI: 10.1109/WCSP58612.2023.10404721]

Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M L and Ferrer C C. 2020. The DeepFake detection challenge (DFDC) dataset [EB/OL]. [2024-05-16]. https://arxiv.org/pdf/2006.07397.pdf

Fan F D, Luo C J, Gao W L and Zhan J F. 2023. AIGCBench: comprehensive evaluation of image-to-video content generated by AI. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(4): #100152 [DOI: 10.1016/j.tbench.2024.100152]

Fan H Q, Xiong B, Mangalam K, Li Y H, Yan Z C, Malik J and Feichtenhofer C. 2021. Multiscale vision Transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6804-6815 [DOI: 10.1109/ICCV48922.2021.00675]

Feichtenhofer C. 2020. X3D: Expanding architectures for efficient video recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 200-210 [DOI: 10.1109/CVPR42600.2020.00028]

Feichtenhofer C, Fan H Q, Malik J and He K M. 2019. SlowFast networks for video recognition//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 6201-6210 [DOI: 10.1109/ICCV.2019.00630]

Guo Y W, Yang C Y, Rao A Y, Liang Z Y, Wang Y H, Qiao Y, Agrawala M, Lin D H and Dai B. 2024. AnimateDiff: animate your personalized text-to-image diffusion models without specific tuning//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR

Ho J, Salimans T, Gritsenko, A, Chan W, Norouzi M and Fleet D J. 2022. Video diffusion models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #628

Li W, Huang T Q, Huang L Q, Zheng A K and Xu C. 2024. Large-scale datasets for facial tampering detection with inpainting techniques. Journal of Image and Graphics, 29(7): 1834-1848(李伟, 黄添强, 黄丽清, 郑翱鲲, 徐超. 2024. 面向人脸修复篡改检测的大规模数据集. 中国图象图形学报, 29(7): 1834-1848) [DOI: 10.11834/jig.230422]

Li Y Z, Yang X, Sun P, Qi H G and Lyu S W. 2020. Celeb-DF: a large-scale challenging dataset for DeepFake forensics//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]

Liu A A, Su Y T, Wang L J, Li B, Qian Z X, Zhang W M, Zhou L N,

Zhang X P, Zhang Y D, Huang J W and Yu N H. 2024. Review on the progress of the AIGC visual content generation and traceability. Journal of Image and Graphics, 29(6): 1535-1554(刘安安, 苏育挺, 王岚君, 李斌, 钱振兴, 张卫明, 周琳娜, 张新鹏, 张勇东, 黄继武, 俞能海. 2024. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报, 29(6): 1535-1554)[DOI: 10.11834/jig.240003]

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023a. Visual instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #1516

Liu Y F, Cun X D, Liu X B, Wang X T, Zhang Y, Chen H X, Liu Y, Zeng T Y, Chan R and Shan Y. 2024. EvalCrafter: benchmarking and evaluating large video generation models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 22139-22149 [DOI: 10.1109/CVPR52733.2024.02090]

Liu Y X, Li L, Ren S H, Gao R D, Li S C, Chen S S, Sun X and Hou L. 2023b. FETV: a benchmark for fine-grained evaluation of open-domain text-to-video generation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: 2723

Ma L, Zhang J J, Deng H P, Zhang N Y, Liao Y and Yu H Y. 2024. DeCoF: generated video detection via frame consistency [EB/OL]. [2024-05-16]. https://arxiv.org/pdf/2402.02085v2.pdf

OpenAI, Achiam J, Adler S and Agarwal S. 2024. GPT-4 Technical Report [EB/OL]. [2024-05-16]. https://arxiv.org/pdf/2303.08774.pdf

Peebles W and Xie S N. 2023. Scalable diffusion models with Transformers//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4172-4182. [DOI: 10.1109/ICCV51070.2023.00387]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685. [DOI: 10.1109/CVPR52688.2022.01042]

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Nießner M. 2019. FaceForensics++: learning to detect manipulated facial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1-11 [DOI: 10.1109/ICCV.2019.00009]

Sinitsa S and Fried O. 2024. Deep image fingerprint: towards low budget synthetic image detection and model lineage analysis//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 4055-4064. [DOI: 10.1109/WACV57701.2024.00402]

Skorokhodov I, Sotnikov G and Elhoseiny M. 2021. Aligning latent and image spaces to connect the unconnectable//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 14124-14133. [DOI: 10.1109/ICCV48922.2021.01388]

Tang Z N, Yang Z Y, Zhu C G, Zeng M and Bansal M. 2023. Any-to-any generation via composable diffusion//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: ACM: #707

Wang J N, Yuan H J, Chen D Y, Zhang Y Y, Wang X and Zhang S W. 2023a. Modelscope text-to-video technical report [EB/OL]. [2024-05-16]. https://arxiv.org/pdf/2308.06571.pdf

Wang S Y, Wang O, Zhang R, Owens A and Efros A A. 2020. CNN-generated images are surprisingly easy to spot… for now//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8692-8701. [DOI: 10.1109/CVPR42600.2020.00872]

Wang X L, Girshick R, Gupta A and He K M. 2018. Non-local neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7794-7803. [DOI: 10.1109/CVPR.2018.00813]

Wang Z D, Bao J M, Zhou W G, Wang W L, Hu H Z, Chen H and Li H Q. 2023b. DIRE for diffusion-generated image detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 22388-22398. [DOI: 10.1109/ICCV51070.2023.02051]

Zhang S W, Wang J Y, Zhang Y Y, Zhao K, Yuan H J, Qin Z W, Wang X, Zhao D L and Zhou J R. 2023. I2VGen-XL: high-quality image-to-video synthesis via cascaded diffusion models [EB/OL]. [2024-05-16]. https://arxiv.org/pdf/2311.04145.pdf

作者简介

郑天鹏，男，硕士研究生，主要研究方向为多媒体内容安全。E-mail: zhengtp@mail.hfut.edu.cn

陈雁翔，通信作者，女，教授，主要研究方向为多媒体信息处理和多媒体内容安全。E-mail: chenyx@hfut.edu.cn

温心哲，男，硕士研究生，主要研究方向为多媒体内容安全。E-mail: wenxinzhe@mail.hfut.edu.cn

李严成，男，硕士研究生，主要研究方向为多媒体内容安全。E-mail: lyc123@mail.hfut.edu.cn

王志远，男，硕士研究生，主要研究方向为信息安全。E-mail: zhiyuanwang@mail.hfut.edu.cn