The Role of Video Generation in Enhancing Data-Limited Action Understanding

Wei Li[1,3], **Dezhao Luo**[4], **Dongbao Yang**[1] ✉, **Zhenhang Li**[1,3]
Weiping Wang[1], **Yu Zhou** [2] ✉
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2] VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[3] School of Cyber Security, University of Chinese Academy of Sciences
[4]Queen Mary University of London
{liwei1,wangweiping,lizhenhang, yangdongbao}@iie.ac.cn, yzhou@nankai.edu.cn,
dezhao.luo@qmul.ac.uk

Abstract

Video action understanding tasks in real-world scenarios always suffer data limitations. In this paper, we address the data-limited action understanding problem by bridging data scarcity. We propose a novel method that employs a text-to-video diffusion transformer to generate annotated data for model training. This paradigm enables the generation of realistic annotated data on an infinite scale without human intervention. We proposed the information enhancement strategy and the uncertainty-based label smoothing tailored to generate sample training. Through quantitative and qualitative analysis, we observed that real samples generally contain a richer level of information than generated samples. Based on this observation, the information enhancement strategy is proposed to enhance the informative content of the generated samples from two aspects: the environments and the characters. Furthermore, we observed that some low-quality generated samples might negatively affect model training. To address this, we devised the uncertainty-based label smoothing strategy to increase the smoothing of these samples, thus reducing their impact. We demonstrate the effectiveness of the proposed method on four datasets across five tasks and achieve state-of-the-art performance for zero-shot action recognition.

1 Introduction

Over the past decade, deep learning [LeCun *et al.*, 2015; Miech *et al.*, 2019] has brought remarkable progress benefiting from large-scale annotated data [Carreira *et al.*, 2018; Sultani *et al.*, 2018], especially in the computer vision community such as CLIP with 400M image-text pairs [Radford *et al.*, 2021]. However, obtaining high-quality datasets is often complex, time-consuming, and costly [Carreira *et al.*, 2018; Grauman *et al.*, 2022], requiring a large amount of manual annotation. In particular in the field of video understanding, human annotations can be inaccurate (50% samples are not aligned in the HowTo100M dataset [Miech *et al.*, 2019]). A lack of a well-annotated dataset with sufficient samples will limit the learning of models and result in poor generalization ability, facing challenges in real-world applications such as abnormal action detection [Sultani *et al.*, 2018; Lv *et al.*, 2023] and long-tail action recognition [Perrett *et al.*, 2023; Grauman *et al.*, 2022].

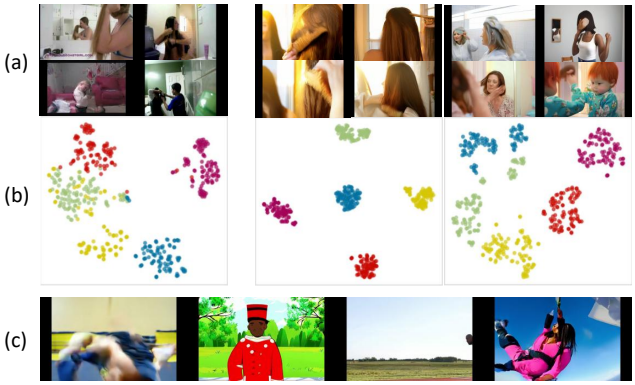

Figure 1: The samples (a) and t-SNE visualizations (b) of the synthetic dataset and the real dataset. From left to right, they are: the HMDB-51 dataset, the synthetic HMDB-51 dataset with the basic strategy and the synthetic HMDB-51 dataset with our proposed information enhancement strategy. (c) Unsatisfactory synthetic videos.

In order to solve the data-limited problem in video understanding, previous methods often focus on designing data augmentation strategies [Yun *et al.*, 2020; Kim *et al.*, 2020; Li *et al.*, 2022b] and transfer knowledge [Kim *et al.*, 2025; Li *et al.*, 2024; Rasheed *et al.*, 2023; Luo *et al.*, 2023] from other modalities, such as images, where data is easier to collect. However, we argue that these solutions are suboptimal as a) data augmentation does not create novel semantics for general knowledge learning [Trabucco *et al.*, 2024], and b) knowledge learned from other modalities does not provide necessary information for video understanding, specifically, images do not include temporal relations between video frames.

With the development of diffusion models [Yang *et al.*,

✉ Corresponding author.

2024; Liu *et al.*, 2024; Esser *et al.*, 2023], existing methods propose to synthesis datasets to improve model learning [He *et al.*, 2023; Tian *et al.*, 2024; Luo *et al.*, 2024; Feng *et al.*, 2023]. There is a problem in the utilization of generated samples for training video understanding models, manifested by the deficient information within the samples, which leads to suboptimal training efficiency. As illustrated in Figure 1 (a) and (b), our finding is that due to the complex and rich information of real videos, real samples have higher intra-class distances and lower inter-class distances than synthetic samples; while the videos generated by the diffusion transformer tend to include information with simple and similar content, which limits the learnable information contained in the synthetic samples . This also makes the classification difficulty of synthetic samples lower than that of real samples.

To solve the problem of deficient information within generated samples, we propose an information enhancement strategy to enhance the effective information of synthetic samples, which refers to context information conducive to action understanding. This strategy injects various character and environment information that conforms to specific human actions into the generated video and alleviates the domain gap between synthetic and real samples. Due to the limited capabilities of the text2video model, some synthetic videos perform unsatisfactorily and contain less effective semantic information, as shown in Figure 1 (c). To alleviate the impact of low-quality synthetic samples on model training, we propose uncertainty-based label smoothing. We calculate the uncertainty of the generated samples with CLIP to measure their quality and adjust the smoothness of label smoothing with uncertainty, which can prevent the model from overfitting low-quality samples and alleviating its impact.

We conducted extensive experiments on four datasets (Kinetics-600, UCF-101, HMDB-51, UCF-Crime) across five tasks (few-shot, zero-shot, base-to-novel, long-tail, abnormal action detection) and demonstrated the effectiveness of our method in data-limited action understanding. We achieved state-of-the-art performance for zero-shot action recognition tasks on Kinetics-600, UCF-101, and HMDB-51.

In summary, our contributions are as follows:

(1) To the best of our knowledge, we are the first to investigate the effectiveness of data generated by video diffusion transformers for enhanced data-limited video understanding that includes action recognition and action detection tasks.

(2) To solve the problem of deficient information within generated samples, we propose two strategies: the information enhancement strategy and the uncertainty-based label smoothing. These two strategies significantly improve the training efficiency of the generated samples.

(3) We conduct extensive experiments on four datasets across five data-limited action understanding tasks to demonstrate the effectiveness of our proposed method, and our method achieves SOTA performance for zero-shot action recognition.

2 Related Work

2.1 Training with Synthetic Samples

The methods of training with synthetic samples can be broadly categorized into two main streams: generative-based and graphics engine-based methods. The graphics engine-based methods use real objects' 3D models to arrange, reconstruct, or move them according to certain rules and then render them to generate training samples. It has found usage in a myriad of fields, including object detection [Gaidon *et al.*, 2016; Cabon *et al.*, 2020], optical flow estimation [Dosovitskiy *et al.*, 2015; Kim *et al.*, 2022], auto driving [Dosovitskiy *et al.*, 2017], etc.

With the explosion development of AIGC, image generation models have become of sufficient quality to generate training samples. Earlier works on image understanding have concentrated on the generation of samples or data augmentation with GANs [Baranchuk *et al.*, 2021; Li *et al.*, 2022a]. But with the emergence of the diffusion model, many methods focus on generating samples using diffusion models given its realistic, high-quality and controllable generative ability. [He *et al.*, 2023] found that diffusion model synthetic samples can significantly improve the classification results of zero-shot, few-shot and transfer learning, and designed strategies to filter the noise of synthetic samples and enhance their diversity. [Feng *et al.*, 2023] proposes a novel test-time prompt tuning method which leverages diffusion models to generate augmented data. [Zhou *et al.*, 2023] proposes diffusion inversion, which invert images to the diffusion model's latent space, and generates novel samples by conditioning the generative model. [Feng *et al.*, 2024] fine-tunes a diffusion model to enhance the capability of the object detection model. [Wang *et al.*, 2024] studies the reasons why synthetic samples sometimes harms contrastive learning performance from the perspective of data inflation and data augmentation, and proposes adaptive inflation to improve the contrastive learning. [Trabucco *et al.*, 2024] uses a diffusion model to edit and change the semantics of the image to address the lack of diversity in previous image augmentation methods.

Within the realm of video understanding, most works still rely on graphics engines or GANs [Guo *et al.*, 2022; Kim *et al.*, 2022]. Unlike these previous works, this paper focuses on leveraging text2video diffusion transformers to improve data-limited action understanding that includes action recognition and action detection tasks. To the best of our knowledge, no prior research has delved into this domain.

2.2 Text-to-Video Diffusion Models

In recent years, text-to-video generation has undergone significant advancements, particularly driven by the rapid development of diffusion models. Previous video generation models based on diffusion typically followed a UNet-based architecture. Video Diffusion Models (VDMs) [Ho *et al.*, 2022] extend image synthesis diffusion models to video generation by training jointly on both image and video data. Gen-1 [Esser *et al.*, 2023] presents a structure and content-aware model that modifies videos guided by example images or texts. With the impressive capabilities demonstrated by Sora [Liu *et al.*, 2024], which uses Transformer as the

backbone of diffusion models, i.e. Diffusion Transformers (DiT), has gradually become mainstream. A series of DiT-based works have emerged, including CogVideoX [Yang *et al.*, 2024] and Hunyuan Video [Kong *et al.*, 2024]. These works are capable of generating high-resolution videos with coherent actions, which makes it possible to apply the generated long-tail video data for downstream task training. In this paper, we utilize CogVideoX [Yang *et al.*, 2024] to generate all training samples because it performs well in generating human-related videos and includes rich motion information.

3 Method

3.1 Overview

We address the data-limited action understanding problem by bridging the data scarcity. To this end, we adopt the text2video model to generate the required data according to the target task. Our approach consists of two steps: 1. Generate a video dataset with a text2video model based on given labels; 2. Train the action understanding model with the generated video dataset, as shown in Figure 2 (a). This paradigm is concise yet general and can produce infinite-scale annotated data without human intervention.

3.2 Video Sample Generator

We employ the text2video model to solve the problem of data-limited action understanding by generating synthetic samples. Given a data-limited dataset $D = \{v_1, v_2, \ldots, v_{n-1}, v_n\}$ with a class name list $C = \{c_1, c_2, \ldots, c_{k-1}, c_k\}$, we use GPT-4o to generate text descriptions of humans performing specific actions, then generate the synthetic video dataset D' based on them:

$$D' = DiT(gpt(prompt_{act}, c)) \tag{1}$$

where DiT represents the text2video diffusion transformer.

Although this paradigm is straightforward and simple, it has the problem of insufficient information in the generated samples, which leads to suboptimal training efficiency. When action descriptions are generated solely on the basis of action names, text-to-video models tend to produce videos with simplistic and repetitive content with these descriptions, leading to lower information richness in the generated videos compared to real videos.

3.3 Information Enhancement Strategy

To enrich the action-related information that serves as a crucial factor in action understanding datasets to train action understanding models effectively, we propose the information Enhancement Strategy towards generated samples. Action-relevant information, such as the environment in which the action frequently occurs, the objects that often appear when the action is performed, the characteristics of the people who always perform the action, benefits action understanding model training, as it encompasses rich semantics and contextual details that are relevant to human actions. A well-constructed generated dataset should encompass sufficient action-related information, as this determines the upper limit of knowledge that the model can acquire from it.

Given an action category name c, we construct the contextual information most relevant to this action. We mainly focus on information from two perspectives: the environment where the action occurs and the appearance characteristics of the person performing the action. We generate contextual information through GPT-4o, where we generate 4 different environment descriptions $Env = gpt(prompt_{env}, c)$ and 16 different character descriptions $Char = gpt(prompt_{char}, c)$ for each action. Finally, we use this contextual information to enhance the content of the generated video. The process of generating action description information is shown in Figure 2 (b).

$$D' = DiT(gpt(Char, Env, prompt_{act}, c)) \tag{2}$$

3.4 Uncertainty-Based Label Smoothing

Due to the limited capabilities of the text2video model, generated videos sometimes have suboptimal generation effects and contain less effective semantic information for training. Such samples may be detrimental to the training of the model, as shown in Figure 1.

To address this problem, we propose uncertainty-based label smoothing. Given a synthetic video sample v', we first use CLIP [Radford *et al.*, 2021] to calculate its similarity S with all category names C, and then use the entropy of the similarity to measure its uncertainty.

$$\begin{aligned} S &= \{CLIP(v', c) | c \in C\}, \\ H(S) &= -\sum_{i=1}^{k} S_i log\, S_i \end{aligned} \tag{3}$$

After that, we dynamically adjust the smoothing of label smoothing according to the uncertainty of the sample:

$$q_i' = \begin{cases} 1-\varepsilon, & if\ i = y \\ \varepsilon/(K-1), & otherwise \end{cases}, \varepsilon = wH(S) \tag{4}$$

whle w represents the weighting of uncertainty-based label smoothing, $y = [q_1, q_2, ..., q_{k-1}, q_k]$ represents the original one-hot label. As low-quality generated samples often exhibit greater uncertainty, leading to an increased degree of label smoothing in such cases, as shown in Figure 2 (c). In this way, we can prevent the model from overfitting to low-quality samples and mitigate their impact on model training.

4 Experiment

4.1 Implementation Details

The text2video model we use is CogVideoX-2B [Yang *et al.*, 2024]. For each dataset, we generate 128 videos for each category with 50 inference steps. We need to emphasize that the training data of CogVideoX-2B does not contain the annotated UCF-101, HMDB-51 or Kinetics-600 datasets we used in our experiments, so there is no data leakage. For information enhancement strategy, we use GPT-4o to generate multiple different environment descriptions at once, with the following prompt like *"Given a phrase describing an action, generate four different scene appearance descriptions that*

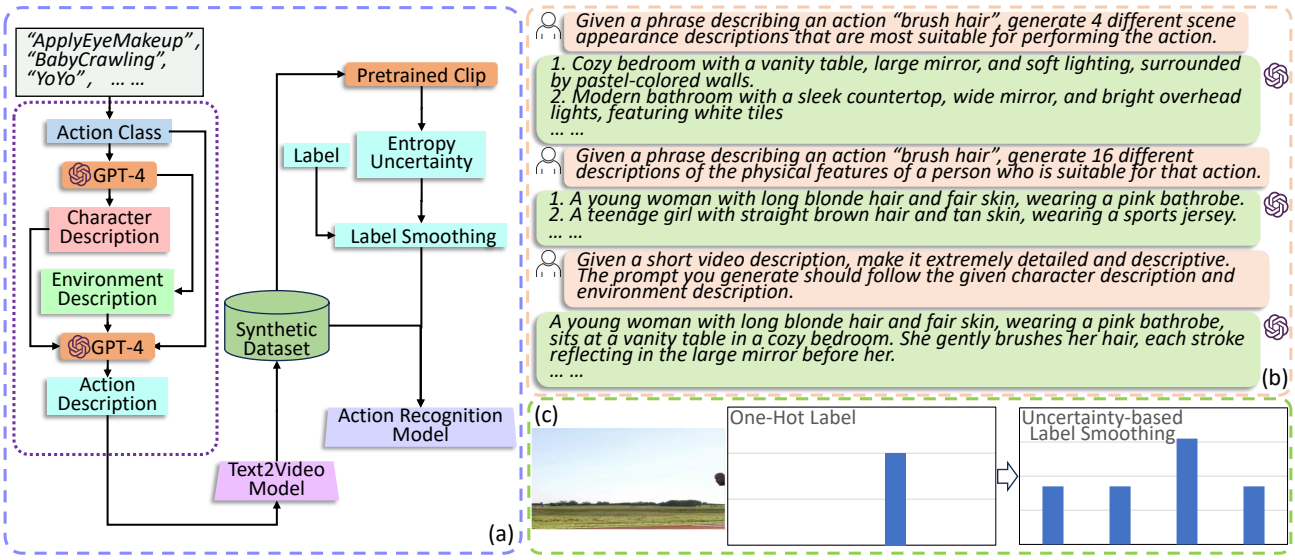

Figure 2: (a): The overall structure of our proposed method. We designed two strategies for generating sample training, information enhancement strategy (left) and uncertainty-based label smoothing strategy (right). (b): The process of generating action description information through proposed information enhancement strategy. (c): Uncertainty-based label smoothing uses a higher smoothness for low-quality generated samples with higher uncertainty.

are most suitable for performing the action.". Similarly, the prompt for generating character descriptions is as follows: "*Given a phrase describing an action, generate 16 different descriptions of the physical features of a person who is suitable for that action.*". In subsequent training, we adopt TC-CLIP [Kim *et al.*, 2025] in most experiments except for abnormal action detection. For abnormal action detection, we use the X-CLIP-B/32 [Ni *et al.*, 2022] following [Lv *et al.*, 2023]. For tasks where real samples are not available such as zero-shot, we train the model with synthetic samples only. For tasks where real samples are available such as few-shot, long-tail, etc., we pre-train the model with synthetic samples and then fine-tune with the real samples. We set the w to 0.3 in uncertainty-based label smoothing. After calculating the uncertainty of all samples in a dataset, normalization is performed on them to constrain the scale.

4.2 Main Results

Zero-shot Action Recognition

In zero-shot action recognition, only the category names of the target dataset are available. The model needs to perform action recognition without training on target dataset. We conduct zero-shot action recognition experiments on the UCF-101, HMDB-51, and Kinetics-600 datasets. We report the Top-1 accuracy for UCF-101 and HMDB-51, Top-1 and Top-5 accuracy for Kinetics-600.

Table 1 shows the zero-shot action recognition results of the proposed method. The proposed method has significant improvements across all three datasets, especially the HMDB-51, where the Top-1 accuracy of zero-shot action recognition is improved by 5% after combining the generated sample training. It should be noted that when our method is based solely on the VIT-B backbone, our performance on HMDB-51 even surpasses those of methods based on the VIT-L backbone ([Zhu *et al.*, 2023; Akbari *et al.*, 2023]). The rich contextual information related to actions is introduced by the proposed method, leading to the improvement of zero-shot action recognition.

The proposed method is generic and can be adapted to various models. We train using the generated samples on three zero-shot action recognition models, including ViFi-CLIP, BIKE, and TC-CLIP. The results are shown in Table 1 bottom. The proposed method improves all three models, especially the Top-1 accuracy of BIKE on Kinetics-600 is improved by almost 10%.

Few-shot Action Recognition

The purpose of the few-shot action recognition task is to enable the model to accurately classify action categories using only a few labeled samples per category, thereby alleviating the reliance on data. We conducted the few-shot action recognition experiments with the UCF-101 and HMDB-51 datasets in Table 2. We first pre-train the model with generated samples and then fine-tune it on each dataset with only K samples per category, where K is in 2, 4, 8 and 16.

The results demonstrate that the proposed method achieves improvements in all K-Shot settings, which can be attributed to the effective initialization of the model with the proposed method. Table 2 shows that the improvement gradually decreases as the K increases from 2 to 16. This phenomenon could be attributed to the fact that the information provided by the generated samples partially overlaps with that provided by the real samples. The proposed information enhancement strategy enhances the action-related information in the generated samples, allowing the proposed method to remain effective even at high values of K=16 in HMDB-51 and K=8 in UCF-101.

Method	HMDB-51	UCF-101	K600 (Top-1)	K600 (Top-5)	All (Top-1)
Vanilla CLIP [Radford *et al.*, 2021]	40.8 ± 0.3	63.2 ± 0.2	59.8 ± 0.3	83.5 ± 0.2	54.6
ActionCLIP [Wang *et al.*, 2021]	49.1 ± 0.4	68.0 ± 0.9	56.1 ± 0.9	83.2 ± 0.2	57.7
X-CLIP [Ni *et al.*, 2022]	44.6 ± 5.2	72.0 ± 2.3	65.2 ± 0.4	86.1 ± 0.8	60.6
Vita-CLIP [Wasim *et al.*, 2023]	48.6 ± 0.6	75.0 ± 0.6	67.4 ± 0.5	-	63.7
Open-VCLIP [Weng *et al.*, 2023]	53.9 ± 1.2	83.4 ± 1.2	73.0 ± 0.8	93.2 ± 0.1	70.1
OTI [Zhu *et al.*, 2023]	54.2 ± 1.3	83.3 ± 0.3	66.9 ± 1.0	-	68.1
OST [Chen *et al.*, 2024]	55.9 ± 1.2	79.7 ± 1.1	75.1 ± 0.6	94.6 ± 0.2	70.2
FROSTER [Huang *et al.*, 2024]	54.8 ± 1.3	84.8 ± 1.1	74.8 ± 0.9	-	71.5
IMP-MoE-L (VIT-L) [Akbari *et al.*, 2023]	59.1	91.5	76.8	-	75.8
OTI (VIT-L) [Zhu *et al.*, 2023]	59.3 ± 1.7	88.1 ± 1.0	70.6 ± 0.5	-	72.7
ViFi-CLIP [Rasheed *et al.*, 2023]	51.3 ± 0.6	76.8 ± 0.7	71.2 ± 1.0	92.2 ± 0.3	66.4
ViFi-CLIP [Rasheed *et al.*, 2023]+Ours	60.8 ± 0.4	81.9 ± 1.2	76.0 ± 1.0	94.3 ± 0.2	72.9
BIKE [Wu *et al.*, 2023]	53.3 ± 1.1	79.6 ± 0.3	68.5 ± 1.3	90.9 ± 0.4	67.1
BIKE [Wu *et al.*, 2023]+Ours	56.5 ± 0.7	87.7 ± 1.0	78.0 ± 1.1	94.7 ± 0.3	74.1
TC-CLIP [Kim *et al.*, 2025]	56.0 ± 0.3	85.4 ± 0.8	78.1 ± 1.0	95.7 ± 0.3	73.2
TC-CLIP [Kim *et al.*, 2025]+Ours	61.0 ± 1.0	86.5 ± 0.7	78.6 ± 0.8	96.1 ± 0.2	75.3

Table 1: Comparison with state-of-the-arts on zero-shot action recognition.

Method	HMDB-51				UCF-101			
	K=2	K=4	K=8	K=16	K=2	K=4	K=8	K=16
Vanilla CLIP [Radford *et al.*, 2021]	41.9	41.9	41.9	41.9	63.6	63.6	63.6	63.6
ActionCLIP [Wang *et al.*, 2021]	47.5	57.9	57.3	59.1	70.6	71.5	73.0	91.4
X-CLIP [Ni *et al.*, 2022]	53.0	57.3	62.8	64.0	76.4	83.4	88.3	91.4
ViFi-CLIP [Rasheed *et al.*, 2023]	57.2	62.7	64.5	66.8	80.7	85.1	90.0	92.7
OST [Chen *et al.*, 2024]	59.1	62.9	64.9	68.2	82.5	87.5	91.7	93.9
TC-CLIP [Kim *et al.*, 2025]	58.6	63.3	65.5	68.8	86.8	90.1	92.0	94.3
TC-CLIP [Kim *et al.*, 2025]+Ours	63.0	65.6	68.9	71.4	88.1	91.3	92.9	94.0

Table 2: Comparison with state-of-the-arts on few-shot action recognition. All the models are directly fine-tuned from CLIP.

Base-to-Novel Generalization

The task of base-to-novel generalization is proposed to evaluate the generalization of a model to unseen classes when only samples from half of the classes are available for training. In each dataset, 16 samples per category are selected from half of the categories to construct the base split for training, while the remaining half of the categories serve as the novel split for evaluation.

Table 3 shows the Top-1 accuracy of the base classes, the novel classes, and their harmonic mean (HM) in UCF-101 and HMDB-51. The proposed method shows noticeable gains for base and novel categories.

Long-Tail Action Recognition

The concept of "long-tail problem" refers to the phenomenon where the imbalance in the distribution of samples across different classes leads to a significantly larger number of samples in the minority head classes compared to the tail classes, as models tend to perform well on the head classes but exhibit poorer performance on the tail classes. Following [Perrett *et al.*, 2023], we construct long-tail action recognition datasets based on UCF-101 and HMDB-51.

The results of long-tail action recognition are presented in Table 4. The proposed method can improve the performance of the tail and the few categories while maintaining the performance of the head categories.

Abnormal Action Detection

Video anomaly action detection aims to identify abnormal events or actions within videos. The primary objective is to pinpoint the specific time window when an anomalous activity occurs, such as crimes, traffic accidents, or other illegal activities, etc. This task is challenging because weak labels are provided only at the video level, whereas the model needs to make frame-level predictions for abnormal actions. We employ the MIL (Multiple Instance Learning) baseline method following [Lv *et al.*, 2023] for abnormal action detection. Initially, we first pre-train the model on generated samples and subsequently fine-tune it on real samples. We report the AUC and AUC_A [Lv *et al.*, 2021] of the frame-level ROC results on the UCF-Crime dataset.

The results for abnormal action detection are presented in Table 5. The proposed method is effective for abnormal action detection, leading to a 2.23% improvement in AUC and a 2.54% improvement in AUC_A compared to MIL baseline. The collection of real-world abnormal action videos is often constrained by real conditions, such as the rarity of the occurrence of anomalous actions. The proposed methods en-

Method	HMDB-51			UCF-101		
	Base	Novel	HM	Base	Novel	HM
Vanilla CLIP [Radford *et al.*, 2021]	53.3	46.8	49.8	78.5	63.6	70.3
ActionCLIP [Wang *et al.*, 2021]	69.1	37.3	48.5	90.1	58.1	70.7
X-CLIP [Ni *et al.*, 2022]	69.4	45.5	55.0	89.9	58.9	71.2
ViFi-CLIP [Rasheed *et al.*, 2023]	73.8	53.3	61.9	92.9	67.7	78.3
Open-VCLIP [Weng *et al.*, 2023]	70.3	50.4	58.9	94.8	77.5	85.3
FROSTER [Huang *et al.*, 2024]	74.1	58.0	65.1	95.3	80.0	87.0
TC-CLIP [Kim *et al.*, 2025]	73.3	59.1	65.5	95.4	81.6	88.0
TC-CLIP [Kim *et al.*, 2025]+Ours	77.7	64.0	70.2	95.5	85.2	90.1

Table 3: Comparison with state-of-the-arts on base-to-novel generalization. All the models are directly fine-tuned from CLIP.

Method	HMDB-51				UCF-101			
	Few	Tail	Head	Acc	Few	Tail	Head	Acc
TC-CLIP [Kim *et al.*, 2025]	49.41	79.63	87.08	60.39	76.13	66.48	80.03	73.46
TC-CLIP [Kim *et al.*, 2025]+Ours	52.55	80.37	87.50	62.40	79.30	68.36	86.25	78.77

Table 4: Results on long-tail action recognition.

Method	AUC	AUC_A
SVM Baseline	50.00	50.00
Sohrab et al. [Sohrab *et al.*, 2018]	58.50	-
BODS [Wang and Cherian, 2019]	68.26	-
GODS [Wang and Cherian, 2019]	70.46	-
Zhang et al. [Zhang *et al.*, 2019]	78.66	-
Motion-Aware [Zhu and Newsam, 2019]	79.10	62.18
Wu et al. [Wu *et al.*, 2020]	82.44	-
MIL	81.80	59.90
MIL+Ours	84.03	62.44

Table 5: Results on abnormal action detection.

Method	HMDB-51
TC-CLIP [Kim *et al.*, 2025]	56.0
+Basic	57.6
+Cha	60.0
+Env	59.7
+IE	61.0

Table 6: Ablation study about information enhancement strategy. "IE" refers to proposed information enhancement strategy with character information and environment information.

Method	UCF-101	HMDB-51
TC-CLIP [Kim *et al.*, 2025]	85.4	56.0
+LS	85.8	60.7
+UW	86.0	60.4
+UF	86.5	60.3
+UL	86.5	61.0

Table 7: Ablation study about uncertainty-based label smoothing.

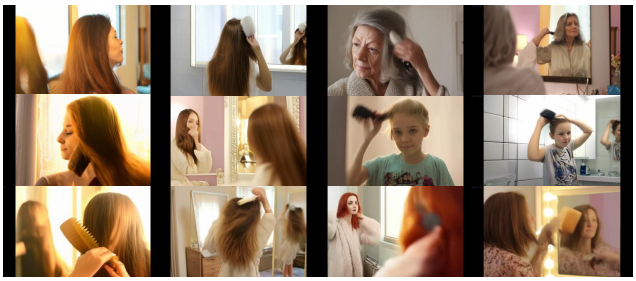

Figure 3: Visualization of generated samples. The strategies adopted from left to right are: "Basic", "Env", "Cha" and "IE".

able the low-cost construction of anomalous action datasets, which effectively enhances the model's performance.

4.3 Ablation Studies

We discuss the effectiveness of the proposed strategies in Table 6 and Table 7. We report the Top-1 accuracy under zero-shot setting for for the ablation study.

The proposed information enhancement strategy ("IE") consists of two components: environmental information enhancement and character information enhancement. We discuss them in Table 6. The "Basic" strategy refers to generating action descriptions solely based on category names and then using them to create videos; The "Env" strategy incorporates environmental information where the action occurs; the "Cha" strategy incorporates character appearance information about the person performing the action.

The results show that incorporating either environmental information or character information in the synthetic samples improves zero-shot accuracy. The information enhancement strategy, which combines environmental information and character information, can lead to a 3.4% improvement compared to the basic strategy. This indicates that enhanc-

ing the action-related information of the generated samples is beneficial for training action understanding models. Figure 3 illustrates the visualization of the "brush hair" action samples generated using the four aforementioned strategies. When employing the basic strategy, the video generation model tends to generate similar videos in which the characters consistently appear as Caucasian women with long brown hair. When the "Env" strategy is adopted, various contextual details that assist in understanding the "brush hair" action are included in generated videos, such as full-length mirrors and bathrooms with tiled backgrounds. This information enhances the knowledge encompassed by the generated dataset. We set the information enhancement strategy as the default setting for the remaining experiments.

In Table 7, we discuss three strategies to mitigate the impact of low-quality generated samples on training. "LS" refers to the label smoothing without additional design toward low-quality generated samples. "UW" denotes the uncertainty-based weighting strategy, where we weight the loss based on the uncertainty of each sample. "UF" denotes the uncertainty-based filtering strategy that removes high-uncertainty samples from the synthetic dataset. "UL" denotes the uncertainty-based label smoothing strategy, where we adjust the smoothness of the label smoothing based on the uncertainty of each sample.

Based on the experiments, uncertainty-based label smoothing shows the best performance among three strategies, leading to a 1.1% improvement on the UCF-101 dataset. This validates the effectiveness of the proposed uncertainty-based label smoothing strategy, which can prevent the model from overfitting low-quality samples and alleviating its impact.

5 Analysis

Method	DB	MMD_{lin}	MMD_{brf}	MD_{poly}
Basic	2.120	0.115	0.098	0.130
IE	4.310	0.112	0.092	0.122
HMDB-51	5.040	0.000	0.000	0.000
UCF-101	2.840	0.036	0.031	0.042

Table 8: Davies Bouldin scores (DB) and Maximum Mean Discrepancy (MMD) of real and synthetic datasets. "IE" refers to generated HMDB-51 dataset with information enhancement strategy, "Basic" refers to generated HMDB-51 dataset with basic strategy. We list the MMD values between various datasets and the HMDB-51 dataset.

We conducted qualitative analysis on synthetic datasets and real datasets, as shown in Figure 1. We extract CLIP features from the HMDB-51 dataset, the synthetic HMDB-51 dataset with the basic strategy, and the synthetic HMDB-51 dataset with the information enhancement strategy. Subsequently, we make the t-SNE visualization of these datasets based on the extracted features.

In Figure 1, it is evident that the boundaries between different categories of synthetic samples are quite distinct. In contrast, the boundaries between samples of different categories in the real dataset are more blurred and lack clarity. This could be attributed to the higher complexity of real samples compared to synthetic samples. This complexity stems from various aspects such as changes in lighting, camera angles, and background details, among others, leading real samples to contain more information. When generating samples based on the basic strategy, the text2video model tends to produce videos with similar content, resulting in low levels of effective information contained in the generated datasets. This limits the upper bound of knowledge that the synthetic data can contain.

The proposed information enhancement strategy effectively increases the information content in the generated samples. As illustrated in Figure 1, the proposed method improves the information contained in the generated samples, bringing their distribution patterns closer to real samples.

We perform quantitative analyzes in Table 8. Davies-Bouldin (DB) is a metric used to assess the effectiveness of clustering algorithms, where a smaller value indicates clearer separation between clusters and tighter intra-cluster cohesion. We employ the Davies-Bouldin score to measure the complexity of the datasets. The first column of Table 8 presents the Davies-Bouldin scores for the generated datasets and real datasets. Real datasets (UCF-101,HMDB-51) exhibit higher DB scores compared to generated samples due to their complexity. The Maximum Mean Discrepancy(MMD) [Gretton *et al.*, 2006] is a commonly used metric for measuring the domain gap between different datasets, where a higher MMD indicates a larger domain gap. In columns 2, 3, and 4 of Table 8, we present the MMD between various datasets and the HMDB-51 dataset. MMD between real datasets, such as UCF-101 and HMDB-51, exhibit lower values, while the MMD between generated datasets and real datasets are higher, indicating a larger domain gap. The proposed strategy reduces the domain gap between generated samples and real samples, which is manifested as lower MMD values.

6 Conclusion

In this paper, we tackle the challenge of data-limited action understanding by introducing a generic method that leverages synthetic video data generated through the text2video diffusion transformer. This approach enables the cost-effective creation of large-scale annotated video datasets, significantly mitigating the data scarcity in action understanding tasks. To enhance the utility of generated samples, we propose the information enhancement strategy and uncertainty-based label smoothing. We validate the proposed method on four datasets across five tasks and achieve SOTA performance for zero-shot action recognition.

Acknowledgements

Supported by the National Natural Science Foundation of China (Grant NO 62376266 and 62406318), Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China.

References

[Akbari *et al.*, 2023] Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig

Adam. Alternating gradient descent and mixture-of-experts for integrated multimodal perception. *NeurIPS*, 36:79142–79154, 2023.

[Baranchuk *et al.*, 2021] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[Cabon *et al.*, 2020] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

[Carreira *et al.*, 2018] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[Chen *et al.*, 2024] Tongjia Chen, Hongshan Yu, Zhengeng Yang, Zechuan Li, Wei Sun, and Chen Chen. Ost: Refining text knowledge with optimal spatio-temporal descriptor for general video recognition. In *CVPR*, pages 18888–18898, 2024.

[Dosovitskiy *et al.*, 2015] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015.

[Dosovitskiy *et al.*, 2017] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[Esser *et al.*, 2023] 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.

[Feng *et al.*, 2023] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, pages 2704–2714, 2023.

[Feng *et al.*, 2024] Chengjian Feng, Yujie Zhong, Zequn Jie, Weidi Xie, and Lin Ma. Instagen: Enhancing object detection by training on synthetic dataset. In *CVPR*, pages 14121–14130, 2024.

[Gaidon *et al.*, 2016] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, pages 4340–4349, 2016.

[Grauman *et al.*, 2022] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022.

[Gretton *et al.*, 2006] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *NeurIPS*, 19, 2006.

[Guo *et al.*, 2022] Xi Guo, Wei Wu, Dongliang Wang, Jing Su, Haisheng Su, Weihao Gan, Jian Huang, and Qin Yang. Learning video representations of human motion from synthetic data. In *CVPR*, pages 20197–20207, 2022.

[He *et al.*, 2023] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023.

[Ho *et al.*, 2022] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 35:8633–8646, 2022.

[Huang *et al.*, 2024] Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. In *ICLR*, 2024.

[Kim *et al.*, 2020] Taeoh Kim, Hyeongmin Lee, MyeongAh Cho, Ho Seong Lee, Dong Heon Cho, and Sangyoun Lee. Learning temporally invariant and localizable features via data augmentation for video recognition. In *ECCV*, pages 386–403. Springer, 2020.

[Kim *et al.*, 2022] Yo-whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? *NeurIPS*, 35:35710–35723, 2022.

[Kim *et al.*, 2025] Minji Kim, Dongyoon Han, Taekyung Kim, and Bohyung Han. Leveraging temporal contextualization for video action recognition. In *ECCV*, pages 74–91. Springer, 2025.

[Kong *et al.*, 2024] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[Li *et al.*, 2022a] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *CVPR*, pages 21330–21340, 2022.

[Li *et al.*, 2022b] Wei Li, Dezhao Luo, Bo Fang, Xiaoni Li, Yu Zhou, and Weiping Wang. Video motion perception for self-supervised representation learning. In *International conference on artificial neural networks*, pages 508–520. Springer, 2022.

[Li *et al.*, 2024] Wei Li, Dezhao Luo, Dongbao Yang, and Weiping Wang. Large language model for action anticipation. In *International Conference on Artificial Neural Networks*, pages 207–222. Springer, 2024.

[Liu *et al.*, 2024] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[Luo *et al.*, 2023] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23045–23055, 2023.

[Luo *et al.*, 2024] Dezhao Luo, Shaogang Gong, Jiabo Huang, Hailin Jin, and Yang Liu. Generative video diffusion for unseen cross-domain video moment retrieval. *CoRR*, 2024.

[Lv *et al.*, 2021] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *TIP*, 30:4505–4515, 2021.

[Lv *et al.*, 2023] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *CVPR*, pages 8022–8031, 2023.

[Miech *et al.*, 2019] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.

[Ni *et al.*, 2022] 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*, pages 1–18. Springer, 2022.

[Perrett *et al.*, 2023] Toby Perrett, Saptarshi Sinha, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Use your head: Improving long-tail video recognition. In *CVPR*, pages 2415–2425, 2023.

[Radford *et al.*, 2021] 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. PMLR, 2021.

[Rasheed *et al.*, 2023] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, pages 6545–6554, 2023.

[Sohrab *et al.*, 2018] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *ICPR*, pages 722–727. IEEE, 2018.

[Sultani *et al.*, 2018] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018.

[Tian *et al.*, 2024] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *NeurIPS*, 36, 2024.

[Trabucco *et al.*, 2024] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *ICLR*, 2024.

[Wang and Cherian, 2019] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *ICCV*, pages 8201–8211, 2019.

[Wang *et al.*, 2021] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[Wang *et al.*, 2024] Yifei Wang, Jizhe Zhang, and Yisen Wang. Do generated data always help contrastive learning? In *ICLR*, 2024.

[Wasim *et al.*, 2023] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *CVPR*, pages 23034–23044, 2023.

[Weng *et al.*, 2023] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *ICML*, pages 36978–36989. PMLR, 2023.

[Wu *et al.*, 2020] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, pages 322–339. Springer, 2020.

[Wu *et al.*, 2023] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *CVPR*, pages 6620–6630, June 2023.

[Yang *et al.*, 2024] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[Yun *et al.*, 2020] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*, 2020.

[Zhang *et al.*, 2019] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *ICIP*, pages 4030–4034. IEEE, 2019.

[Zhou *et al.*, 2023] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023.

[Zhu and Newsam, 2019] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.

[Zhu *et al.*, 2023] Yan Zhu, Junbao Zhuo, Bin Ma, Jiajia Geng, Xiaoming Wei, Xiaolin Wei, and Shuhui Wang. Orthogonal temporal interpolation for zero-shot video recognition. In *ACMMM*, pages 7491–7501, 2023.