Journal of Intelligent Computing and Networking
https://www.ffspub.com/index.php/jicn/index
ISSN 3079-9228 (print)
ISSN 3107-2798 (online)
E-mail: jicn.office@ffspub.com

Article

Semantic-Aware Contrasting with Explicitly Disentangled Representations for AI-Generated Video Detection

Jiaquan Yu[1], Jingyi Xie[1], Zeyu Zhang[1], Yunhang Qian[2], Xiaobin Hu[2], Jiawei Liu[1,†]
[1]*MoE Key Lab of BIPC, University of Science and Technology of China, Hefei 230027, China*
[2]*School of Computing, National University of Singapore, 119077, Singapore*
E-mail: jwliu6@ustc.edu.cn

Received: September 18, 2026 / Revised: November 3, 2026 / Accepted: November 6, 2026 / Published online: November 16, 2026

Abstract: The rapid proliferation of AI-generated videos (AIGVs) has raised critical concerns about content authenticity and trustworthy media consumption, underscoring the need for robust AIGV detection. Despite recent progress, existing benchmarks suffer from a fundamental limitation: real and fake videos are often collected or generated from different semantic distributions. As a result, detectors may rely on spurious correlations involving objects, scenes, or actions rather than intrinsic artifacts introduced by generative models. This limitation calls for a benchmark that explicitly aligns real and generated videos at the semantic level, enabling detection methods to be evaluated and optimized under controlled content conditions. To this end, we introduce SAVGE, a large-scale semantically aligned AIGV detection dataset comprising 2,369,822 videos, including 424,115 real videos and 1,945,707 generated videos. SAVGE organizes samples into real–fake semantic groups by using captions paired with real videos as prompts for text-to-video generation, thereby providing both group-level correspondence and textual semantic anchors. Based on SAVGE, we further propose Explicit Disentangling and Semantic-aware Contrasting (EDSC), a model-agnostic training framework for learning artifact-focused representations with reduced semantic interference. EDSC first extracts and aligns visual semantic features with textual representations, and then explicitly suppresses semantic information in the artifact branch while contrasting real and fake samples within each semantic group. By integrating caption-guided semantic supervision with intra-group contrastive learning in a two-phase optimization strategy, EDSC encourages detectors to capture generation-related cues under matched semantic conditions. Extensive experiments on SAVGE and GenVideo demonstrate that EDSC improves most evaluated backbone–metric combinations, enhances robustness under challenging evaluation settings, and strengthens cross-dataset generalization. These results highlight the importance of semantic alignment in AIGV detection and validate explicit semantic–artifact disentanglement as an effective strategy for robust detector learning.

Keywords: AI-Generated Video Detection; Representation Disentanglement Learning; Contrastive Learning; Video Analysis
https://doi.org/10.64509/jicn.24.159

1 Introduction

Recent advances in generative modeling have made it increasingly practical to synthesize high-fidelity videos with realistic appearance, motion, and temporal coherence [1–8]. These AI-generated videos (AIGVs) are expanding the creative possibilities of entertainment, media production, education, and virtual environments. At the same time, their growing realism and accessibility raise serious concerns about misinformation, impersonation, and the erosion of trust in digital content. As synthetic videos become increasingly difficult for humans to distinguish from camera-captured ones, robust AIGV detection is becoming increasingly important for content authentication.

A central challenge in AIGV detection is distinguishing generation artifacts from semantic content. Ideally, a detector should determine whether a video is real or generated without relying on the objects, scenes, or actions it depicts. This requirement makes dataset construction particularly important: when real and fake samples differ substantially in semantics, models may learn spurious content biases rather than intrinsic traces left by generative models [9]. Therefore, a strong benchmark should not only contain diverse synthetic videos but also control the semantic gap between real and fake samples.

† Corresponding author: Jiawei Liu
* Academic Editor: Xueyang Fu
© 2026 The authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Chen *et al.* [10] took an important step toward this goal by introducing GenVideo [10], the first large-scale benchmark for AIGV detection. GenVideo contains diverse, high-quality fake videos synthesized by multiple generation techniques and defines practical evaluation tasks such as cross-generator video classification and degraded video classification. These designs make GenVideo a valuable benchmark for studying detector robustness under realistic conditions. Nevertheless, GenVideo [10] still lacks explicit semantic correspondence between real and fake samples. As illustrated in Figure 1, visual manipulation datasets such as FaceForensics++ [11] and CASIA [12] generate fake content by directly manipulating real content, naturally preserving strong semantic alignment across the two domains. In contrast, GenVideo does not systematically provide the caption-based real–fake groups used in SAVGE. This semantic mismatch obscures the intended detection signal: models may exploit content-distribution differences rather than learn generation-related artifacts. It also limits the use of learning strategies, such as explicit semantic disentanglement, that require semantically aligned supervision.

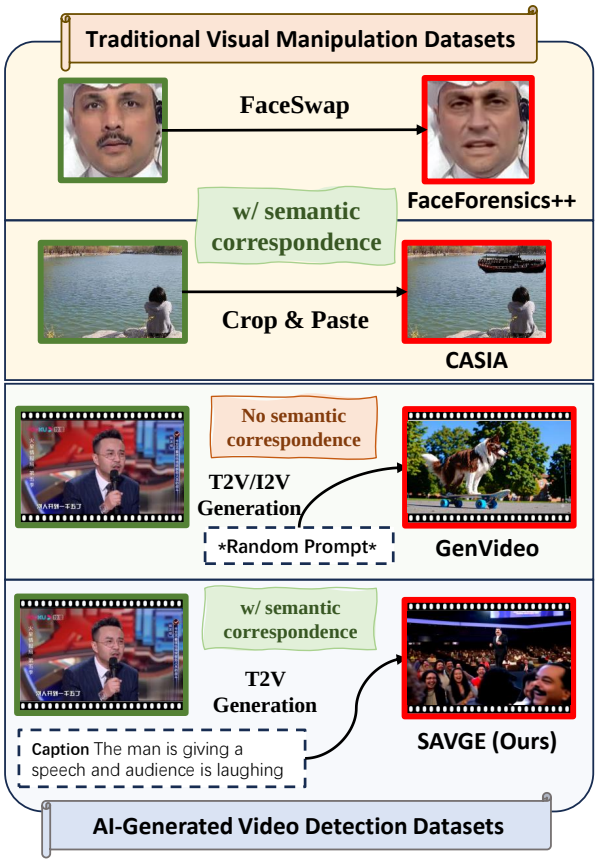


Figure 1: Comparison of representative visual manipulation and AIGV detection datasets. FaceForensics++ [11] and CASIA [12] preserve semantic correspondence because fake content is produced by manipulating real content. FaceForensics++ contains videos, whereas CASIA contains images and is included only as an image-tampering example. By contrast, existing AIGV benchmarks such as GenVideo [10] generate fake videos from independent prompts, leading to weak semantic alignment. SAVGE addresses this gap by synthesizing fake videos from captions paired with real videos.

To bridge this gap, we introduce SAVGE, a large-scale AIGV detection dataset built around semantically aligned real–fake video groups. SAVGE treats semantic correspondence as a core dataset property: each training group contains one real video and multiple fake videos that share the same semantic description. Specifically, real videos are collected from existing video-captioning datasets, and their captions are used as prompts for state-of-the-art text-to-video generation models to synthesize corresponding fake videos. This design converts the original video captions from auxiliary annotations into explicit semantic anchors that connect real and generated samples at the group level. To make this correspondence reliable, we further apply a retrieval-based ranking and filtering strategy that compares generated candidates with their paired real videos, removes severely misaligned samples, and retains semantically consistent fake videos. After this filtering process, SAVGE contains 424,115 real videos and 1,945,707 fake videos, making it both large in scale and structurally aligned for controlled AIGV detection research. The resulting organization enables comparisons between real and fake videos under matched semantic conditions, reduces the risk that detectors exploit content-distribution shortcuts, and supports fine-grained analyses of generation artifacts across different semantic contexts. Moreover, the paired captions and group identifiers make SAVGE suitable for learning strategies that require semantic supervision, such as semantic–artifact disentanglement and group-level contrastive learning.

Building on SAVGE, we further propose a model-agnostic Explicit Disentangling and Semantic-aware Contrasting (EDSC) training framework for robust AIGV detection. EDSC decomposes video representations into semantic and artifact components and encourages the artifact component to be discriminative while remaining insensitive to semantic variation. It operates through a two-phase training strategy. In the Semantic Extraction and Alignment (SEA) phase, EDSC uses captions as semantic anchors and aligns the visual semantic component with textual representations by maximizing cross-modal mutual information. In the Semantic Suppression and Intra-Group Contrasting (SSGC) phase, EDSC suppresses semantic information in the artifact component by minimizing intra-modal mutual information, while a group-level contrastive objective separates real and fake videos within each semantic group. Through this alternating process, EDSC learns semantically invariant representations that better capture generation artifacts. Extensive experiments on SAVGE and GenVideo validate the effectiveness and generalizability of EDSC, and further demonstrate the value of semantic alignment for robust AIGV detection.

Our main contributions are summarized as follows:

- We introduce SAVGE, a large-scale AIGV detection dataset that organizes real and fake videos into semantically aligned groups, with textual captions serving as semantic anchors.
- We design a scalable caption-guided generation and filtering pipeline that uses real-video captions as prompts, ranks candidate generated videos, and filters them to construct reliable semantically aligned real–fake groups.

- We propose EDSC, a model-agnostic framework that learns semantically invariant yet discriminative artifact representations through explicit disentanglement learning and semantic-aware contrasting learning.
- We conduct extensive experiments on the SAVGE and GenVideo datasets, demonstrating the effectiveness and generalizability of EDSC and highlighting the importance of semantic alignment for robust AIGV detection.

2 Related Work

2.1 Video Generation Methods

Early video generation studies were largely driven by Generative Adversarial Networks (GANs), which extend image-level adversarial synthesis to the spatiotemporal domain by jointly modeling appearance and motion. Representative methods include TGAN [13], which introduces a temporal generator to capture latent dynamics, MoCoGAN [14], which disentangles motion and content representations to improve temporal coherence, and StoryGAN [15], which incorporates structured story-level conditioning to encourage semantic consistency across frames. These GAN-based models established important foundations for video synthesis and can produce visually sharp clips, but they are often limited by training instability, restricted temporal range, and weak semantic controllability.

Recent progress has therefore shifted toward diffusion-based and transformer-enhanced generation pipelines that support prompt-conditioned, scalable video synthesis [1–6, 16–18]. Compared with earlier adversarial approaches, diffusion models provide a more stable generative framework and can better exploit textual or visual conditions to control video semantics. For example, Imagen Video [3] demonstrates high-definition text-to-video generation with cascaded diffusion models, VideoBooth [5] incorporates image prompts for efficient personalized generation, ModelScopeT2V [6] introduces spatiotemporal modules to improve frame consistency and motion smoothness, and VideoFusion [16] decomposes the diffusion process to better capture video structure. More recent systems, such as Latte [17] and OpenSora [18], further integrate attention-based architectures to strengthen long-range temporal dependency modeling and improve scalability.

Overall, progress in video generation has lowered the barrier to producing high-quality synthetic videos and increased the difficulty of distinguishing them from real-world content. Accordingly, this paper focuses on AI-generated video detection and constructs SAVGE using multiple text-based video generation methods to synthesize fake videos whose semantics are aligned with real video captions.

2.2 AI-Generated Visual Content Detection

Detection of AI-generated visual content has evolved from early GAN-artifact detectors based on low-level statistics and CNN features [19, 20] to methods using spatiotemporal cues and hybrid pixel-frequency representations [21–23]. Although effective, these methods often depend on generator-specific artifacts and may degrade under changes in synthesis models, compression pipelines, or data distributions. This issue is further amplified by diffusion-based generation, whose high fidelity, semantic controllability, and artifact patterns differ from earlier GAN-generated content [24, 25]. Related image-forensics research has likewise emphasized learning generic forgery cues across source qualities through multi-prior fusion [26].

Reliable benchmarks are therefore essential for assessing whether detectors learn generalizable generation cues rather than dataset-specific biases. Representative deepfake benchmarks, including FaceForensics++ [11], Celeb-DF [27], and DFDC [28], provide manipulated facial videos across diverse compression levels, synthesis qualities, and real-world conditions. Because they focus mainly on face manipulation, fake samples are typically produced by modifying real source videos and thus retain strong correspondence in identity, scene context, and motion. While useful for studying manipulation artifacts, this setting does not fully reflect AI-generated video (AIGV) detection, where an entire video can be synthesized from a text prompt and real–fake semantic distributions may be mismatched.

To support AIGV detection, Chen *et al.* introduced GenVideo [10], the first large-scale benchmark in this area, which provides diverse generated videos and practical evaluation protocols. However, its generated videos are not explicitly paired with semantically corresponding real videos, creating a semantic distribution gap that may encourage detectors to exploit content biases rather than intrinsic generation artifacts. To address this limitation, we introduce SAVGE, a large-scale AIGV detection dataset organized around semantically aligned real–fake video groups with textual annotations. Using captions as semantic anchors, SAVGE enables controlled evaluation and supports learning strategies that disentangle semantic information from artifact-sensitive representations.

Recent benchmarks include GenVidBench [29], which studies cross-source and cross-generator detection and includes real-video-derived test pairs; GenWorld [30], which examines generated real-world simulation videos; and Ivy-Fake [31], which supports explainable image and video detection. D3 [32] detects generated videos using second-order temporal features without detector training. SAVGE combines caption-based semantic groups with retrieval filtering and uses this structure directly in EDSC training.

2.3 Feature Disentanglement Learning for AI-Generated Visual Content Detection

Feature disentanglement learning aims to separate visual representations into interpretable factors, making it appealing for forgery detection, where models should focus on manipulation artifacts rather than semantic content. Otherwise, detectors may rely on objects, scenes, or motions as shortcuts and generalize poorly to unseen generators or real-world data. Prior work [9, 33–35] has shown the value of this idea for AI-generated visual content detection. UCF [9] improves cross-domain face forgery detection by disentangling shared semantics from forgery cues, while FDML [35] combines disentanglement with multi-view learning to mine intrinsic forgery traces. However, extending such strategies to AIGV detection remains difficult since existing datasets often lack semantic correspondence between real and fake videos, making generation artifacts hard to separate from content differences.

SAVGE addresses this issue by organizing real and fake videos into semantically aligned groups and providing captions as semantic anchors. This structure enables our EDSC framework to disentangle semantic and artifact representations and apply semantic-aware contrasting within each group, encouraging semantically invariant yet artifact-sensitive features.

3 SAVGE: Semantically Aligned AI-Generated Video Dataset

SAVGE is a large-scale dataset for AI-generated video detection, comprising 2,369,822 videos in total, including 424,115 real videos sourced from multiple real-world datasets and 1,945,707 AI-generated fake videos. Unlike GenVideo [10], which lacks explicit semantic alignment between real and fake videos, SAVGE organizes samples into semantic groups. Each group contains one real video and its caption, together with four to six fake videos for training or one fake video for testing. The caption serves as the semantic anchor. This structure provides direct supervision for learning with captions and for contrasting real and generated videos within the same semantic group. SAVGE is intended to support research on robust AI-generated video detection and related tasks.

SAVGE and GenVideo include several common generation methods, but differ in their source videos, sample organization, and filtering. We therefore evaluate the same SAVGE-trained detectors on both test sets, as described in Section 5. At the same time, we recognize the rapid evolution of generative techniques. Therefore, we plan to update SAVGE on an annual basis by incorporating newly emerged generation methods, keeping the benchmark comprehensive, current, and representative of real-world challenges.

3.1 Video Collection and Generation

To obtain high-quality textual annotations for each semantic group, SAVGE uses several video-captioning datasets as sources of real videos, including Youku-mPLUG [36], ActivityNet [37], and 20BN-s-s V2 [38]. These datasets provide a rich variety of real-world video content across multiple domains and scenarios. We carefully select a subset from each dataset to ensure diverse semantics and video styles, forming the base set of positive samples in SAVGE.

Each real video serves as the basis of a unique semantic group and is treated as the sole positive sample in that group. To construct semantically aligned negative samples, we use the caption of each real video as the prompt for text-to-video generation. Specifically, SAVGE employs a diverse set of state-of-the-art text-to-video models to synthesize fake videos, including ZeroScope [39], VideoCrafter [40], Wan2.1 [41], SD [42], Latte [17], OpenSora [18], ModelScope [6], MorphStudio [43], Show_1 [44], Gen2 [45], Crafter [40], and Lavie [46]. As shown in Table 1, the development and test splits each use six generation sources. Formally, we denote the i-th real video as $V^{i,0}$ and its generated fake videos as $\{V^{i,j}\}_{j=1}^{6}$, where j indexes the six generation sources used for the corresponding split.

ZeroScope uses 576w generation at 576×320 followed by XL upscaling to 1024×576; Table 1 lists the final output resolution.

3.2 Ranking and Filtering for Fake Videos

Because current text-to-video models do not always preserve prompt semantics faithfully, some generated videos may deviate from their corresponding real videos in subject, scene, or action. To reduce such mismatches, we introduce a video–video retrieval ranking mechanism to identify and discard low-quality generated samples.

Table 1: Statistics of real and generated videos in our proposed SAVGE dataset.

Video Source	Type	Resolution	FPS	Length	Train Set	Test Set	Total Count
ActivityNet [37]	Real	-	-	~120s	13,289	8,090	424,115
Youku-mPLUG [36]		-	-	10-120s	176,510	5,379	
20BN-s-s V2 [38]		-	-	~15s	214,316	6,531	
ZeroScope [39]	Fake	1024×576	8	3s	190,605	-	1,925,707
VideoCrafter [40]		1024×576	8	2s	291,082	-	
Wan2.1 [41]		1280 × 720	16	5s	337,192	-	
SD [42]		512-1024	8	2-6s	298,661	-	
Latte [17]		512×512	8	2s	405,655	-	
OpenSora [18]		512×512	8	2s	402,512	-	
ModelScope [6]	Fake	256×256	8	4s	-	2,451	20,000
MorphStudio [43]		1280×720	8	2s	-	2,983	
Show_1 [44]		576×320	8	4s	-	3,913	
Gen2 [45]		896×512	24	4s	-	3,246	
Crafter [40]		256×256	8	4s	-	3,089	
Lavie [46]		1280×2048	8	2s	-	4,318	
Total Count	-	-	-	-	2,329,822	40,000	2,369,822

3.3 Ranking and Filtering for Fake Videos

Because current text-to-video models do not always preserve prompt semantics faithfully, some generated videos may deviate from their corresponding real videos in subject, scene, or action. To reduce such mismatches, we introduce a video–video retrieval ranking mechanism to identify and discard low-quality generated samples.

Specifically, for the i-th semantic group with generated videos $\{V^{i,j}\}_{j=1}^{6}$, we adopt VideoMAE V2 [47] to compute a similarity score $s_j \in [-1,1]$ between the real video $V^{i,0}$ and each fake sample $V^{i,j}$. We first remove groups whose mean similarity over all six candidates is below 0.50, then filter candidates within the remaining groups. For the training set, we retain fake videos whose retrieval score is greater than 0.6 or whose rank is among the top four within the group, balancing semantic consistency with sample diversity. The top-four rule maintains the group size for training, while the score threshold admits additional well-matched candidates. For the test set, we apply a stricter criterion and keep only the top-ranked fake video in each group. After filtering, SAVGE contains 2,369,822 videos in total, including 424,115 real and 1,945,707 fake ones.

To analyze the effect of the retrieval-based filtering strategy, we visualize the "Fake Video–Caption" and "Fake Video–Real Video" similarity-score distributions before and after the filtering process. As shown in Figure 2, filtering shifts the similarity-score distributions toward higher values, indicating that the filtering process removes generated samples with weak semantic correspondence to their paired real videos. This quantitative trend is consistent with the qualitative examples in Figure 3, where the retained fake videos generally preserve the primary subjects, scenes, and actions described by the captions of their corresponding real videos. Meanwhile, these examples still exhibit visible appearance discrepancies and generation artifacts, suggesting that filtering improves semantic consistency without removing the subtle cues required for detection. Overall, the proposed filtering strategy enhances semantic alignment while preserving the diversity and discriminative characteristics of fake samples needed for robust AIGV detection.

3.4 Dataset Formulation

Formally, SAVGE is organized as a series of semantic groups, with the i-th group denoted as:

$$G_i = \left\{ V^{i,0}, \{V^{i,j}\}_{j=1}^{L_i}, T_i \right\} \tag{1}$$

where $V^{i,0}$ is the real video, $\{V^{i,j}\}_{j=1}^{L_i}$ is the set of fake videos generated from the same textual prompt T_i, and L_i is the number of fake videos in the i-th group. For the training set, $4 \le L_i \le 6$, while for the test set, L_i is fixed to 1.

Table 1 shows that the sources differ in resolution, frame rate, and duration. Semantic grouping matches content rather than these technical properties. Source and caption biases, VideoMAE-based selection, technical video differences, and the limited generator collection may therefore still affect detection.

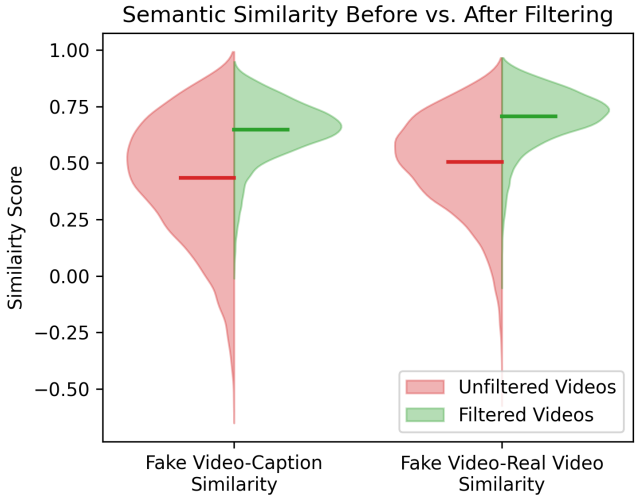


Figure 2: "Fake Video–Caption" and "Fake Video–Real Video" similarity-score distributions before and after the filtering process. The distributions show that retained fake videos are more semantically aligned with their corresponding real videos or captions than videos in the unfiltered set, indicating that the filtering process reduces real–fake semantic mismatch.

4 EDSC: Explicit Disentangling and Semantic-aware Contrasting for AIGV Detection

Building on the aligned structure of SAVGE, we introduce Explicit Disentangling and Semantic-aware Contrasting (EDSC), a training framework that moves AIGV detection beyond conventional binary supervision. Rather than treating semantics as an uncontrolled source of bias, EDSC uses the captions and semantic group annotations in SAVGE to disentangle semantic content from generation-related artifacts in the learned representation. This design is motivated by the observation that generalizable detectors should rely on forgery traces instead of category-, scene-, or action-level shortcuts, a principle also supported by recent generalizable deepfake detection research [9]. At a high level, EDSC uses textual annotations as semantic anchors to localize semantic factors in video features, suppresses these factors from the artifact branch, and further contrasts real and generated videos under matched semantic conditions. In this way, EDSC turns the semantic alignment of SAVGE into a form of supervision, enabling detectors to learn artifact-specific representations that capture more intrinsic differences between real and AI-generated videos.

4.1 Overview

EDSC is a model-agnostic training framework designed to exploit the semantic structure of SAVGE for robust AIGV detection. Given a video and its paired caption, EDSC uses a video encoder to produce two complementary visual representations: a visual semantic feature $\boldsymbol{f}_{SV}$, which is encouraged to capture the caption-related content of the video, and a visual artifact feature $\boldsymbol{f}_{AV}$, which is optimized to retain generation-specific cues for real–fake classification. In parallel, a text



Figure 3: Example video samples from the SAVGE dataset. Each fake video is generated using the caption of a real video as the prompt, and the resulting real–fake set is organized as a semantic group.

encoder maps the caption into a semantic reference feature $\boldsymbol{f}_S$, providing an explicit semantic anchor for aligning and isolating content information. During training, $\boldsymbol{f}_{SV}$ and $\boldsymbol{f}_S$ serve as auxiliary supervisory signals that guide disentanglement, while only $\boldsymbol{f}_{AV}$ is used as the final detection representation. This design encourages the detector to make decisions based on generation artifacts rather than semantic shortcuts such as objects, scenes, or actions.

The overall pipeline of EDSC is illustrated in Figure 4. EDSC is optimized through two alternating phases: Semantic Extraction and Alignment (SEA) and Semantic Suppression and Intra-Group Contrasting (SSGC). In SEA, EDSC maximizes cross-modal mutual information between each projected semantic feature and the basic feature from the other modality. It also minimizes the distance between $\boldsymbol{f}_{SV}$ and $\boldsymbol{f}_S$. This objective encourages $\boldsymbol{f}_{SV}$ to encode caption-consistent content, such as objects, scenes, and actions, and therefore establishes an explicit semantic anchor inside the visual representation. By first localizing semantic information in a dedicated branch, EDSC provides a reliable target for removing content-related factors from the artifact branch in the following phase. In the SSGC phase, EDSC minimizes the intra-modal mutual information between $\boldsymbol{f}_{SV}$ and the visual artifact feature $\boldsymbol{f}_{AV}$, encouraging $\boldsymbol{f}_{AV}$ to discard semantic cues and preserve generation-related artifacts. Meanwhile, an intra-group contrastive objective is applied in the $\boldsymbol{f}_{AV}$ space, where real and fake videos sharing the same semantic description are contrasted to strengthen category separation under matched content conditions. The SEA and SSGC phases are alternated until convergence, yielding a semantically invariant artifact representation for robust real–fake classification. At inference time, EDSC requires only the video encoder and the artifact feature $\boldsymbol{f}_{AV}$, with no text input or semantic branch.

4.2 Encoders and Feature Extraction

EDSC employs a video encoder and a text encoder to extract basic features from the two modalities. Notably, EDSC imposes no restrictions on the choice of encoders. For example, one may adopt a unified vision-language model such as XCLIP [48], or alternatively combine two single-modal models such as VideoMAE [47] and BERT [49]. This flexibility makes EDSC adaptable to diverse encoder architectures and practical deployment requirements.

Specifically, we denote the basic features extracted from the video encoder as $\boldsymbol{f}_V \in \mathbb{R}^{d_V}$ and the features extracted from the text encoder as $\boldsymbol{f}_T \in \mathbb{R}^{d_T}$. We then project these features into a latent space using linear projection heads P_{SV}, P_{AV}, and P_S to obtain the visual semantic feature $\boldsymbol{f}_{SV}$, visual artifact feature $\boldsymbol{f}_{AV}$, and semantic reference feature $\boldsymbol{f}_S$, respectively, as follows:

$$\boldsymbol{f}_{SV} = P_{SV}(\boldsymbol{f}_V), \tag{2}$$

$$\boldsymbol{f}_{AV} = P_{AV}(\boldsymbol{f}_V), \tag{3}$$

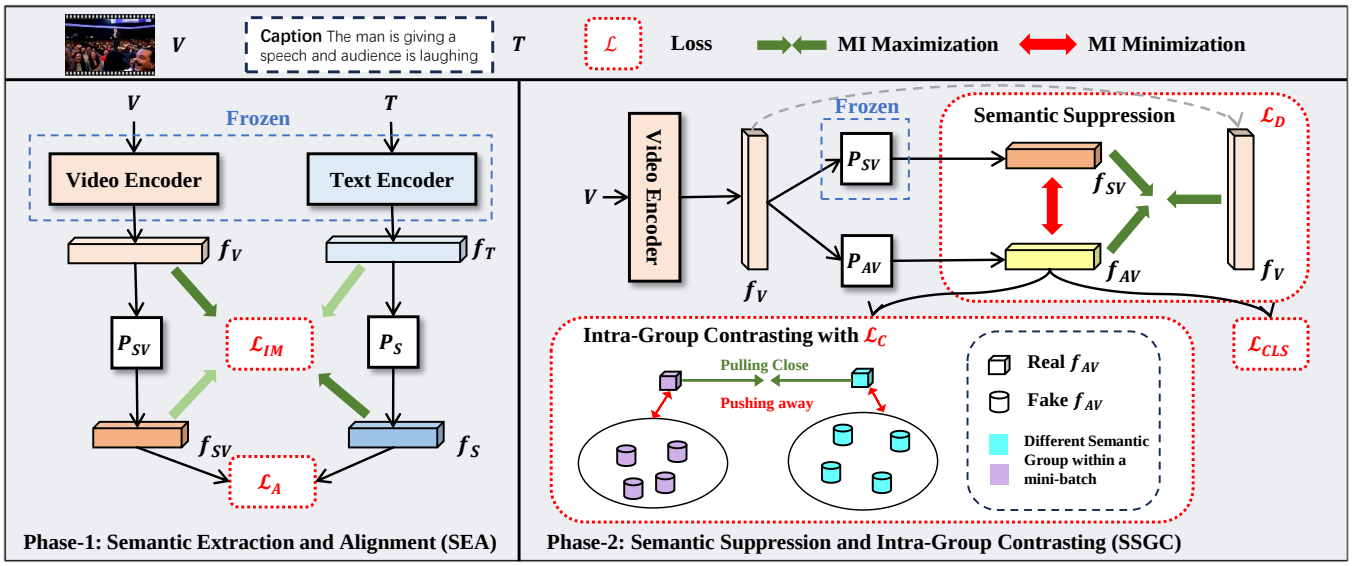


Figure 4: The overall framework of our proposed EDSC approach. It consists of two alternating phases: (1) Semantic Extraction and Alignment (SEA) and (2) Semantic Suppression and Intra-Group Contrasting (SSGC). The frozen boxes indicate fixed parameters in the corresponding phase, not fixed output features.

$$\boldsymbol{f}_S = P_S(\boldsymbol{f}_T) \tag{4}$$

where $\boldsymbol{f}_{SV}$, $\boldsymbol{f}_{AV}$, $\boldsymbol{f}_S \in \mathbb{R}^d$ and d is the dimension of the latent space. The three projected features are then used in the subsequent SEA and SSGC phases.

4.3 Semantic Extraction and Alignment Phase

The SEA phase aligns the visual semantic feature $\boldsymbol{f}_{SV}$ with the information described in the text annotation, preparing it to guide semantic suppression in the SSGC phase. To achieve this goal, EDSC combines cross-modal mutual-information maximization with a distance constraint between the projected semantic features to encourage semantic alignment, using Deep InfoMax [50] with Jensen-Shannon divergence (JSD-DIM) [51], which provides stable mutual information estimation. The JSD-DIM estimator $I_\theta^{(J)}(X,Y)$ for two generic random variables $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$ is formulated as follows:

$$\begin{aligned} I_\theta^{(J)}(X,Y) = \mathbb{E}_{p(x,y)}\left[-\log\left(1+e^{-T_\theta(x,y)}\right)\right] \\ -\mathbb{E}_{p(x)p(y)}\left[\log\left(1+e^{T_\theta(x,y)}\right)\right] \end{aligned} \tag{5}$$

where $T_\theta : \mathscr{X} \times \mathscr{Y} \mapsto \mathbb{R}$ is a learnable statistics network parameterized by θ. Maximizing $I_\theta^{(J)}(X,Y)$ estimates and increases the mutual information between X and Y.

During this phase, we freeze the video and text encoder parameters and optimize the semantic projection heads P_{SV} and P_S. The phase objective $\mathscr{L}_{SEA}$ is formulated as follows:

$$\mathscr{L}_{IM} = I_{\theta_{VT}}^{(J)}(\boldsymbol{f}_{SV}, \boldsymbol{f}_T) + I_{\theta_{TV}}^{(J)}(\boldsymbol{f}_S, \boldsymbol{f}_V), \tag{6}$$

$$\mathscr{L}_A = \|\boldsymbol{f}_{SV} - \boldsymbol{f}_S\|_2, \tag{7}$$

$$\mathscr{L}_{SEA} = -\mathscr{L}_{IM} + \mathscr{L}_A \tag{8}$$

$\mathscr{L}_{SEA}$ comprises two terms. The first term, $\mathscr{L}_{IM}$, maximizes the cross-modal mutual information between the projected semantic features $\boldsymbol{f}_{SV}/\boldsymbol{f}_S$ and the basic features $\boldsymbol{f}_T/\boldsymbol{f}_V$ from the other modality. Since the encoders are frozen in this phase, $\mathscr{L}_{IM}$ encourages $\boldsymbol{f}_{SV}$ and $\boldsymbol{f}_S$ to capture the semantic information shared between the two modalities. The second term, $\mathscr{L}_A$, minimizes the distance between $\boldsymbol{f}_{SV}$ and $\boldsymbol{f}_S$, further enforcing semantic alignment. Overall, $\mathscr{L}_{SEA}$ encourages $\boldsymbol{f}_{SV}$ to learn video semantic components that are aligned with the information described in the text annotation.

4.4 Semantic Suppression and Intra-Group Contrasting Phase

The SSGC phase aims to transform the artifact branch into a discriminative yet semantically invariant representation. It contains two complementary objectives. First, using the aligned semantic feature $\boldsymbol{f}_{SV}$ learned in the SEA phase as guidance, SSGC suppresses the semantic information contained in the visual artifact feature $\boldsymbol{f}_{AV}$ while preserving the non-semantic information derived from the basic video feature $\boldsymbol{f}_V$. This encourages $\boldsymbol{f}_{AV}$ to focus on generation-related artifacts rather than caption-level content. Second, SSGC applies intra-group contrastive learning in the $\boldsymbol{f}_{AV}$ space, where real and fake videos with the same semantic description are explicitly separated. By combining semantic suppression with contrastive separation under aligned content, SSGC produces an artifact-focused feature $\boldsymbol{f}_{AV}$ that is more suitable for robust real–fake classification.

To achieve feature disentanglement, we adopt JSD-DIM for intra-modal mutual information estimation as follows:

$$\mathscr{L}_D = -I_{\theta_{AS}}^{(J)}(\boldsymbol{f}_{AV}, \boldsymbol{f}_{SV}) + I_{\theta_W}^{(J)}([\boldsymbol{f}_{SV}||\boldsymbol{f}_{AV}], \boldsymbol{f}_V) \tag{9}$$

By minimizing $I_{\theta_{AS}}^{(J)}(\boldsymbol{f}_{AV}, \boldsymbol{f}_{SV})$, we encourage $\boldsymbol{f}_{AV}$ to capture information that is distinct from $\boldsymbol{f}_{SV}$. However, this term alone may lead to trivial solutions. To mitigate this issue, the second term $I_{\theta_W}^{(J)}([\boldsymbol{f}_{SV}||\boldsymbol{f}_{AV}], \boldsymbol{f}_V)$ encourages $\boldsymbol{f}_{AV}$ to retain information that is complementary to $\boldsymbol{f}_{SV}$ and sufficient to recover the information contained in the basic feature $\boldsymbol{f}_V$. These two terms jointly produce an explicitly disentangled

$\boldsymbol{f}_{AV}$ with reduced semantic information. We maximize $\mathscr{L}_D$ (equivalently, minimize $-\mathscr{L}_D$ in $\mathscr{L}_{SSGC}$), thereby minimizing $I^{(J)}_{\theta_{AS}}$ while maximizing $I^{(J)}_{\theta_W}$.

Furthermore, we apply semantic group-level feature contrasting to enforce a clearer separation between the two categories. To this end, we construct a mini-batch of $N = (1+L)M$ samples, including M real–fake video groups. The i-th group contains one real video and L fake videos sharing the same textual annotation, whose visual artifact features are denoted as $\boldsymbol{f}^{i,0}_{AV}$ and $\{\boldsymbol{f}^{i,k}_{AV}\}^{L}_{k=1}$, respectively. We use the InfoNCE loss [52] to formulate the learning objective:

$$\mathscr{L}_C = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{sim(\boldsymbol{f}^{i,0}_{AV}, \boldsymbol{f}^{i,*}_{AV})}{\sum_{k\in[L]\cup\{*\}} sim(\boldsymbol{f}^{i,0}_{AV}, \boldsymbol{f}^{i,k}_{AV})} \tag{10}$$

where $[L] = \{1,\ldots,L\}$, $\boldsymbol{f}^{i,*}_{AV}$ is sampled from all other real features $\{\boldsymbol{f}^{j,0}_{AV}\}_{j\neq i}$ within the mini-batch as a positive sample, and $sim(f,g) = \exp(f\cdot g/\tau)$ is the similarity function. Intuitively, $\mathscr{L}_C$ pulls the real visual artifact feature $\boldsymbol{f}^{i,0}_{AV}$ close to the positive sample $\boldsymbol{f}^{i,*}_{AV}$ while pushing it away from all fake features $\{\boldsymbol{f}^{i,k}_{AV}\}^{L}_{k=1}$ within the same semantic group.

Finally, $\boldsymbol{f}_{AV}$ is used for classification. We apply a linear transformation followed by a sigmoid function to map each $\boldsymbol{f}^{i,k}_{AV}$ to a probability score $\hat{y}^{i,k}$. The classification loss is formulated as a weighted binary cross-entropy loss: Real videos are assigned label 1 and fake videos label 0; the real-video term is weighted by L to balance the L fake samples in each group.

$$\mathscr{L}_{CLS} = -\frac{1}{M}\sum_{i=1}^{M}\left[L\log\hat{y}^{i,0} + \sum_{k=1}^{L}\log(1-\hat{y}^{i,k})\right] \tag{11}$$

The overall loss function for the SSGC phase is formulated as:

$$\mathscr{L}_{SSGC} = -\mathscr{L}_D + \mathscr{L}_C + \mathscr{L}_{CLS} \tag{12}$$

4.5 Training and Inference Scheme

Algorithm 1 summarizes the training and inference pipeline of EDSC. In particular, it describes the alternating optimization scheme adopted during training, where the model first performs the SEA phase to align the visual semantic representation with the textual semantic anchor, and then executes the SSGC phase to suppress semantic information in the artifact branch while enforcing intra-group discrimination between real and fake samples. This alternating design enables the semantic branch to provide guidance from the learned semantic projection for disentanglement while progressively refining the semantically invariant artifact representation used for final prediction. During inference, only the visual encoder E_V, the projection head P_{AV}, and the classifier C are retained; neither captions nor the semantic branch are needed. In SSGC, E_V, P_{AV}, and C are updated; P_{SV} is frozen, and the text branch remains unchanged. The output of P_{SV} may still change as E_V is updated.

Algorithm 1 Training and Inference of EDSC

Require: Video encoder E_V; text encoder E_T; projection heads P_{SV}, P_{AV}, and P_S; classifier C; SAVGE training set with semantic groups; number of training iterations T.
Ensure: Trained video encoder E_V, projection head P_{AV}, and classifier C.
Training: Initialize the parameters of E_V, E_T, P_{SV}, P_{AV}, P_S, and C.
for $t = 1$ to T **do**
  Sample a mini-batch of M semantic groups, each containing one real video, L fake videos, and the shared text annotation.
  Extract basic video features $\boldsymbol{f}_V \leftarrow E_V(\boldsymbol{x})$ and text features $\boldsymbol{f}_T \leftarrow E_T(\boldsymbol{s})$.
  Obtain $\boldsymbol{f}_{SV} \leftarrow P_{SV}(\boldsymbol{f}_V)$, $\boldsymbol{f}_{AV} \leftarrow P_{AV}(\boldsymbol{f}_V)$, and semantic reference feature $\boldsymbol{f}_S \leftarrow P_S(\boldsymbol{f}_T)$.
  if t is in the SEA phase **then**
    Compute $\mathscr{L}_{SEA} = -\mathscr{L}_{IM} + \mathscr{L}_A$.
    Update P_{SV} and P_S using $\mathscr{L}_{SEA}$; freeze E_V and E_T.
  else
    Compute disentanglement loss $\mathscr{L}_D$ between $\boldsymbol{f}_{SV}$, $\boldsymbol{f}_{AV}$, and $\boldsymbol{f}_V$.
    Compute intra-group contrastive loss $\mathscr{L}_C$ on $\boldsymbol{f}_{AV}$ within each semantic group.
    Compute classification loss $\mathscr{L}_{CLS}$ from $\hat{y} = C(\boldsymbol{f}_{AV})$.
    Form $\mathscr{L}_{SSGC} \leftarrow -\mathscr{L}_D + \mathscr{L}_C + \mathscr{L}_{CLS}$.
    Update E_V, P_{AV}, and C using $\mathscr{L}_{SSGC}$; freeze P_{SV}.
  end if
end for
Inference: For a test video, compute $\boldsymbol{f}_V \leftarrow E_V(\boldsymbol{x})$, $\boldsymbol{f}_{AV} \leftarrow P_{AV}(\boldsymbol{f}_V)$, and output the prediction $\hat{y} = C(\boldsymbol{f}_{AV})$.

5 Experiments

5.1 Datasets

We conduct a comprehensive evaluation of EDSC on the SAVGE dataset, which is divided into a development set and a test set for training and benchmarking, respectively. Detailed statistics of the SAVGE development and test sets are provided in Table 1. For model development, we further split the development set into training and validation subsets with a ratio of 9:1. In addition, we perform cross-dataset evaluation on GenVideo [10] to assess the generalization capability of EDSC and compare the relative difficulty of the two benchmarks.

5.2 Comparative Methods

Since EDSC is model-agnostic, it can be applied to a wide range of AIGV detection methods. We select several state-of-the-art methods for comparison, including three frame-level methods, F^3Net [53], NPR [54], and CLIP-B-FT [55], and three video-level methods, TALL [56], XCLIP-B-FT [48], and DeMamba-XCLIP-B-FT [10]. For methods using CLIP or XCLIP, we adopt their respective text encoders to extract the semantic reference feature $\boldsymbol{f}_S$; for the other methods, we use a pre-trained BERT-base model [49].

5.3 Implementation Details

For methods without EDSC, we adopt a unified training protocol with a batch size of 128, a learning rate of 1×10^{-6},

and the AdamW optimizer for 10 epochs. Each mini-batch is constructed by sampling 64 real videos and 64 fake videos to maintain class balance during training.

For methods with EDSC, we construct each mini-batch from 25 semantic groups, with each group containing 1 real video and 4 fake videos, resulting in a batch size of 125. We use the same learning rate of 1×10^{-6} and the AdamW optimizer, and train for 10 epochs. In the first 2,000 iterations of each epoch, we optimize only the SEA phase and then switch to the SSGC phase for the remaining iterations. The output dimensions of the projection heads P_{SV}, P_{AV}, and P_S are set to $d = 512$. The structure of the statistics network T_θ follows the original Deep InfoMax design [50]. The temperature τ in $\mathscr{L}_C$ is set to 0.1 for all experiments.

Each backbone uses the same frame sampling and input size with and without EDSC.

For evaluation metrics, we report average precision (AP), recall, and the F1 score to assess detector effectiveness. Recall and the F1 score are computed using a decision threshold of 0.5, whereas AP is threshold-independent. For frame-based detection methods, we average frame-level predictions to obtain the corresponding video-level predictions. All results are reported as the average of five independent runs.

5.4 In-dataset Evaluation on the SAVGE Dataset

We first evaluate a diverse set of baseline methods on the proposed SAVGE dataset, with the results reported in Table 2. Among all methods without EDSC, DeMamba-XCLIP-B-FT [10] achieves the strongest overall performance, with an F1 score of 0.8476 and an AP of 0.8719, outperforming both frame-level and video-level counterparts. This advantage is likely attributable to its dedicated Mamba-based temporal modeling module, which is well suited to capturing complex spatiotemporal dependencies in video content. The strong baseline performance of DeMamba-XCLIP-B-FT also indicates the importance of effective temporal modeling for AIGV detection on semantically aligned video benchmarks.

After incorporating EDSC, all methods improve on most evaluation metrics, demonstrating that the proposed framework can improve most metrics for detectors built on different backbone architectures. In particular, DeMamba-XCLIP-B-FT [10]+EDSC achieves the best overall performance among all compared methods, reaching an F1 score of 0.8651 and an AP of 0.8760. These gains suggest that EDSC complements strong baseline encoders by further suppressing semantic interference and promoting artifact-sensitive representations, thereby improving discriminability even when the underlying detector is already highly competitive.

XCLIP-B-FT [48] shows the largest F1 gain from EDSC, with improvements of approximately 0.06/0.11/0.04 in recall/F1/AP, respectively. One possible explanation is that XCLIP, as a unified vision-language model, is naturally compatible with the semantic supervision introduced by EDSC and can therefore exploit the disentanglement objective more effectively. Taken together, these results show that EDSC is a general and effective training framework for improving the accuracy and robustness of AIGV detectors on SAVGE.

5.5 Cross-dataset Evaluation on the GenVideo Dataset

We further evaluate the generalization ability of EDSC on the GenVideo dataset [10] by directly testing models trained on the SAVGE development set on the GenVideo test set. This setting provides a more stringent assessment of cross-dataset transferability, since the training and evaluation data differ in collection protocol, semantic organization, and sample construction. As reported in Table 2, methods equipped with EDSC improve most backbone–metric combinations relative to their corresponding baseline versions, demonstrating that the representations learned by EDSC generalize beyond the SAVGE benchmark. Among all compared methods,

Table 2: In-dataset and cross-dataset evaluation on the SAVGE and GenVideo [10] datasets. All methods below are trained solely on the SAVGE development set.

Method	Detection Level	EDSC	Tested on SAVGE			Tested on GenVideo		
			Recall	F1	AP	Recall	F1	AP
BERT+F^{3}Net [53]	Frame	-	0.7371	0.7634	0.8123	0.8032	0.7961	0.8655
		✓	0.7962	0.7825	0.8569	0.8683	0.8912	0.9306
BERT+NPR [54]	Frame	-	0.7832	0.7213	0.7944	0.8235	0.8162	0.7937
		✓	0.8345	0.7921	0.8212	0.8912	0.8657	0.8345
CLIP-B-FT [55]	Frame	-	0.8341	0.8235	0.8654	0.9291	0.8065	0.9344
		✓	0.8587	0.8452	0.8341	0.9344	0.9127	0.9215
BERT+TALL [56]	Video	-	0.7663	0.7219	0.7891	0.8396	0.8215	0.8568
		✓	0.7978	0.7452	0.8345	0.8412	0.8097	0.8701
XCLIP-B-FT [48]	Video	-	0.7954	0.7059	0.8096	0.8471	0.8042	0.8760
		✓	0.8513	0.8122	0.8520	0.9295	0.8650	0.9487
DeMamba-XCLIP-B-FT [10]	Video	-	0.8568	0.8476	0.8719	0.9457	**0.9143**	0.9638
		✓	**0.8712**	**0.8651**	**0.8760**	**0.9503**	0.9029	**0.9794**

DeMamba-XCLIP-B-FT [10]+EDSC achieves the highest cross-dataset recall and AP, at 0.9503 and 0.9794, respectively, although its F1 decreases. These results further suggest that the gains brought by EDSC are not limited to in-dataset fitting, but extend to scenarios in which detectors are applied to previously unseen data distributions.

Most reported metrics are higher on GenVideo than on SAVGE. The datasets differ in their real-video sources, generation pipelines, and filtering, so these results do not identify a single cause for that difference. SAVGE specifically supports evaluation under caption-based semantic grouping; the cross-dataset results test whether the learned representations also transfer to another collection.

EDSC improves 32 of the 36 means in Table 2. CLIP-B-FT AP decreases by 3.13 and 1.29 percentage points on SAVGE and GenVideo, respectively, suggesting that semantic suppression may also remove useful ranking information. On GenVideo, BERT+TALL and DeMamba F1 decrease by 1.18 and 1.14 points despite higher AP: better ranking need not improve F1 at the fixed 0.5 threshold. These five-run means show gains across most settings, but do not establish the significance of small differences.

5.6 Robustness Evaluation on the SAVGE Dataset

Following the evaluation protocol in [10], we assess robustness on the SAVGE dataset by applying several common data transformations to the original videos, including JPEG compression, Gaussian noise, frame flipping, cropping, and color transformation. Under this setting, we evaluate BERT+TALL [56], XCLIP-B-FT [48], and DeMamba-XCLIP-B-FT [10], together with their EDSC-enhanced counterparts. The corresponding results are reported in Table 3. We use frame-wise JPEG quality 35 and additive zero-mean Gaussian noise with $\sigma = 0.05$ on $[0, 1]$ pixels. Crop width and height are each 160–208 pixels on 224×224 frames; the same window is used across a video and resized back to 224×224. All model variants largely preserve their performance under frame flipping and color transformation. This behavior is expected, since these transformations neither substantially alter semantic content nor introduce pronounced artifacts that interfere with detection. In contrast, cropping leads to severe performance degradation for DeMamba-XCLIP-B-FT and its EDSC-enhanced counterpart, with AP drops of 0.27 and 0.26, respectively. A plausible explanation is that cropping may remove visually informative regions that contain crucial generation artifacts, preventing the detectors from forming reliable predictions from incomplete spatial evidence. JPEG compression and Gaussian noise also reduce detection performance, although their effects differ. In particular, DeMamba-XCLIP-B-FT [10] suffers a substantial AP drop of 0.21 after JPEG compression, highlighting video compression as a major challenge for robust AIGV detection. Although DeMamba-XCLIP-B-FT [10]+EDSC also degrades under JPEG compression, the AP drop is reduced to 0.14, indicating that EDSC improves resilience to compression artifacts to some extent. By comparison, the impact of Gaussian noise is more moderate, with AP drops of 0.10 and 0.05

Table 3: Robustness evaluation on the SAVGE dataset by applying various data transformations.

Method	Metric	Original	JPEG	Gaussian	Flip	Crop	Color Trans.
BERT+TALL [56]	**Recall**	0.7663	0.7359 $_{-0.03}$	0.7548 $_{-0.01}$	0.7515 $_{-0.01}$	0.6029 $_{-0.16}$	0.7350 $_{-0.03}$
	F1	0.7219	0.5483 $_{-0.17}$	0.6735 $_{-0.05}$	0.7003 $_{-0.02}$	0.4625 $_{-0.26}$	0.6968 $_{-0.03}$
	AP	0.7891	0.6673 $_{-0.12}$	0.6856 $_{-0.10}$	0.7596 $_{-0.03}$	0.4715 $_{-0.32}$	0.7616 $_{-0.03}$
BERT+TALL [56] + EDSC	**Recall**	0.7978	0.7661 $_{-0.03}$	0.7850 $_{-0.01}$	0.7947 $_{-0.00}$	0.6292 $_{-0.17}$	0.7939 $_{-0.00}$
	F1	0.7452	0.6415 $_{-0.10}$	0.6892 $_{-0.06}$	0.7327 $_{-0.01}$	0.5895 $_{-0.16}$	0.7439 $_{-0.00}$
	AP	0.8345	0.7762 $_{-0.06}$	0.7765 $_{-0.06}$	0.8081 $_{-0.03}$	0.5794 $_{-0.26}$	0.8343 $_{-0.00}$
XCLIP-B-FT [48]	**Recall**	0.7954	0.7646 $_{-0.03}$	0.7813 $_{-0.01}$	0.7805 $_{-0.01}$	0.6946 $_{-0.10}$	0.7732 $_{-0.02}$
	F1	0.7059	0.5758 $_{-0.13}$	0.6736 $_{-0.03}$	0.6771 $_{-0.03}$	0.4551 $_{-0.25}$	0.6730 $_{-0.03}$
	AP	0.8096	0.6286 $_{-0.18}$	0.7250 $_{-0.08}$	0.7882 $_{-0.02}$	0.4987 $_{-0.31}$	0.8002 $_{-0.01}$
XCLIP-B-FT [48] + EDSC	**Recall**	0.8513	0.8097 $_{-0.04}$	0.8204 $_{-0.03}$	0.8339 $_{-0.02}$	0.6963 $_{-0.16}$	0.8474 $_{-0.00}$
	F1	0.8122	0.7194 $_{-0.09}$	0.7646 $_{-0.05}$	0.7926 $_{-0.02}$	0.7069 $_{-0.11}$	0.8051 $_{-0.01}$
	AP	0.8520	0.7539 $_{-0.10}$	0.8016 $_{-0.05}$	0.8338 $_{-0.02}$	0.5750 $_{-0.28}$	0.8415 $_{-0.01}$
DeMamba-XCLIP-B-FT [10]	**Recall**	0.8568	0.8371 $_{-0.02}$	0.8419 $_{-0.01}$	0.8542 $_{-0.00}$	0.6794 $_{-0.18}$	0.8514 $_{-0.01}$
	F1	0.8476	0.7043 $_{-0.14}$	0.7841 $_{-0.06}$	0.8399 $_{-0.01}$	0.6350 $_{-0.21}$	0.8471 $_{-0.00}$
	AP	0.8719	0.6590 $_{-0.21}$	0.7764 $_{-0.10}$	0.8711 $_{-0.00}$	0.6039 $_{-0.27}$	0.8698 $_{-0.00}$
DeMamba-XCLIP-B-FT [10] + EDSC	**Recall**	0.8712	0.8643 $_{-0.01}$	0.8702 $_{-0.00}$	0.8710 $_{-0.00}$	0.7095 $_{-0.16}$	0.8686 $_{-0.00}$
	F1	0.8651	0.7597 $_{-0.11}$	0.8169 $_{-0.05}$	0.8633 $_{-0.00}$	0.6398 $_{-0.23}$	0.8624 $_{-0.00}$
	AP	0.8760	0.7410 $_{-0.14}$	0.8281 $_{-0.05}$	0.8703 $_{-0.01}$	0.6184 $_{-0.26}$	0.8755 $_{-0.00}$

for DeMamba-XCLIP-B-FT [10] and DeMamba-XCLIP-B-FT [10]+EDSC, respectively. Overall, these results suggest that, while EDSC remains vulnerable to severe degradations such as cropping and JPEG compression, it learns more robust representations under milder perturbations such as Gaussian noise, highlighting its potential for deployment in challenging real-world scenarios.

5.7 Ablation Study

We conduct ablation studies to examine the contribution of different components in EDSC. We adopt BERT+TALL [56], XCLIP-B-FT [48], and DeMamba-XCLIP-B-FT [10] as backbone models and evaluate the following variants:

5.7.1 SEA Phase

We first evaluate the effectiveness of the two objectives used in the SEA phase by separately removing the mutual information term $\mathscr{L}_{IM}$ and the feature alignment term $\mathscr{L}_A$. As shown in Table 4, removing $\mathscr{L}_{IM}$ causes a substantial performance drop across all three backbones. For example, on XCLIP-B-FT, recall/F1/AP decrease from 0.8513/0.8122/0.8520 with the full EDSC framework to 0.6741/0.6239/0.7116 without $\mathscr{L}_{IM}$. Similar degradations are observed on BERT+TALL and DeMamba-XCLIP-B-FT. This result indicates that cross-modal mutual information maximization is essential for extracting reliable semantic factors from the video representation because it explicitly encourages the visual semantic branch to preserve information shared with the textual caption. Without this term, the semantic branch cannot provide a stable anchor for subsequent semantic suppression in the artifact branch.

Removing $\mathscr{L}_A$ leads to a milder but still consistent degradation compared with the full EDSC framework. In particular, XCLIP-B-FT achieves 0.8453/0.7972/0.8461 in recall/F1/AP without $\mathscr{L}_A$, which is lower than the full EDSC result of 0.8513/0.8122/0.8520. This suggests that, although mutual information alignment captures shared cross-modal semantics, the direct distance constraint between $\boldsymbol{f}_{SV}$ and $\boldsymbol{f}_S$ further improves the compactness and consistency of the semantic space. Overall, these results demonstrate that $\mathscr{L}_{IM}$ plays the dominant role in semantic extraction, while $\mathscr{L}_A$ provides complementary alignment that stabilizes the SEA phase and improves the final artifact representation.

5.7.2 Feature Disentanglement in SSGC Phase

We next examine the effect of feature disentanglement by removing the disentanglement loss $\mathscr{L}_D$ from the SSGC phase. In this variant, $\boldsymbol{f}_{AV}$ is no longer explicitly constrained to be independent of the semantic feature $\boldsymbol{f}_{SV}$, and the detector is optimized mainly by the contrastive loss $\mathscr{L}_C$ and classification loss $\mathscr{L}_{CLS}$. As shown in Table 4, removing $\mathscr{L}_D$ consistently degrades performance across the three backbones. For XCLIP-B-FT, recall/F1/AP drop from 0.8513/0.8122/0.8520 with the full EDSC framework to 0.8320/0.7761/0.8128. A similar decline is observed for DeMamba-XCLIP-B-FT, whose recall/F1/AP decrease from 0.8712/0.8651/0.8760 to 0.8315/0.8142/0.8496. These results indicate that explicit semantic suppression is crucial for preventing the artifact branch from retaining content-related shortcuts. By reducing the dependence between $\boldsymbol{f}_{AV}$ and $\boldsymbol{f}_{SV}$, $\mathscr{L}_D$ guides the detector toward generation-related cues and plays an important role in learning robust artifact representations.

5.7.3 Intra-group Contrasting in SSGC Phase

We further study the contribution of intra-group contrastive learning by removing $\mathscr{L}_C$ and optimizing only the disentanglement loss $\mathscr{L}_D$ and classification loss $\mathscr{L}_{CLS}$ on $\boldsymbol{f}_{AV}$. As shown in Table 4, removing $\mathscr{L}_C$ consistently weakens the full EDSC framework. For XCLIP-B-FT, recall/F1/AP decrease from 0.8513/0.8122/0.8520 to 0.8427/0.7856/0.8267. For DeMamba-XCLIP-B-FT, the corresponding scores drop from 0.8712/0.8651/0.8760 to 0.8615/0.8342/0.8548. These results show that classification supervision alone is insufficient to fully exploit the semantic grouping structure of SAVGE. By contrasting real and fake videos within the same semantic group, $\mathscr{L}_C$ explicitly enlarges the inter-category

Table 4: Ablation study on the effectiveness of various EDSC designs. The methods are trained and tested on the SAVGE dataset.

Method	BERT+TALL			XCLIP-B-FT			DeMamba-XCLIP-B-FT		
	Recall	F1	AP	Recall	F1	AP	Recall	F1	AP
w/o EDSC	0.7663	0.7219	0.7891	0.7954	0.7059	0.8096	0.8568	0.8476	0.8719
EDSC	0.7978	0.7452	0.8345	0.8513	0.8122	0.8520	0.8712	0.8651	0.8760
Ablation study on SEA Phase									
SEA w/o $\mathscr{L}_{IM}$	0.7058	0.6154	0.6927	0.6741	0.6239	0.7116	0.7336	0.7160	0.7397
SEA w/o $\mathscr{L}_A$	0.7854	0.7148	0.8128	0.8453	0.7972	0.8461	0.8644	0.8279	0.8497
Ablation study on SSGC Phase									
SSGC w/o $\mathscr{L}_D$	0.7601	0.7034	0.7798	0.8320	0.7761	0.8128	0.8315	0.8142	0.8496
SSGC w/o $\mathscr{L}_C$	0.7765	0.7107	0.7956	0.8427	0.7856	0.8267	0.8615	0.8342	0.8548
Ablation study on training strategy									
EDSC w/ Joint Train	0.4767	0.4133	0.3759	0.5395	0.4678	0.3916	0.5985	0.5037	0.4668

margin under matched content conditions, helping $\boldsymbol{f}_{AV}$ capture more discriminative generation artifacts. Nevertheless, the degradation caused by removing $\mathscr{L}_C$ is generally smaller than that caused by removing $\mathscr{L}_D$, suggesting that intra-group contrasting mainly complements feature disentanglement by further refining category separation in the artifact feature space.

5.7.4 Two-phase Training

Finally, we investigate whether the proposed alternating optimization strategy is necessary by jointly optimizing $\mathscr{L}_{SEA}$ and $\mathscr{L}_{SSGC}$ in a single stage. As shown in Table 4, joint training leads to severe performance degradation for all three backbones. For XCLIP-B-FT, recall/F1/AP drop to 0.5395/0.4678/0.3916, which is not only far below the full EDSC result of 0.8513/0.8122/0.8520 but also substantially worse than the baseline without EDSC. The same trend appears for DeMamba-XCLIP-B-FT, where joint training obtains only 0.5985/0.5037/0.4668, compared with 0.8712/0.8651/0.8760 for the full EDSC framework. These results indicate that semantic extraction and semantic suppression should not be optimized simultaneously from the beginning. Without a dedicated SEA phase to first establish a reliable semantic anchor, the semantic and artifact branches may interfere with each other, making the suppression objective unstable and preventing $\boldsymbol{f}_{AV}$ from learning meaningful generation-related cues. Therefore, the alternating two-phase strategy helps stabilize disentanglement and improve the performance of EDSC.

Overall, the ablation results obtained with BERT+TALL [56], XCLIP-B-FT [48], and DeMamba-XCLIP-B-FT [10] show a consistent trend: all core components of EDSC contribute positively across different backbone architectures. Although the magnitude of improvement varies across models, feature disentanglement, intra-group contrasting, and the two-phase training strategy each provide consistent performance benefits. These observations further indicate that the effectiveness of EDSC is attributed to the generality of the training mechanism rather than to a property of any particular backbone.

5.8 Qualitative Results

Figure 5 presents t-SNE visualizations of feature representations learned by different baseline detectors and their EDSC-enhanced counterparts on the SAVGE test set. A clear trend can be observed across all three backbones: after incorporating EDSC, the feature distributions of real and fake videos become more compact within each category and more separable across categories. This suggests that EDSC guides the encoders toward representations that better capture generation-related discrepancies rather than being dominated by semantic content variation.

More specifically, the original feature spaces produced by BERT+TALL [56], XCLIP-B-FT [48], and DeMamba-XCLIP-B-FT [10] exhibit noticeable overlap between real and fake samples, indicating that the learned representations

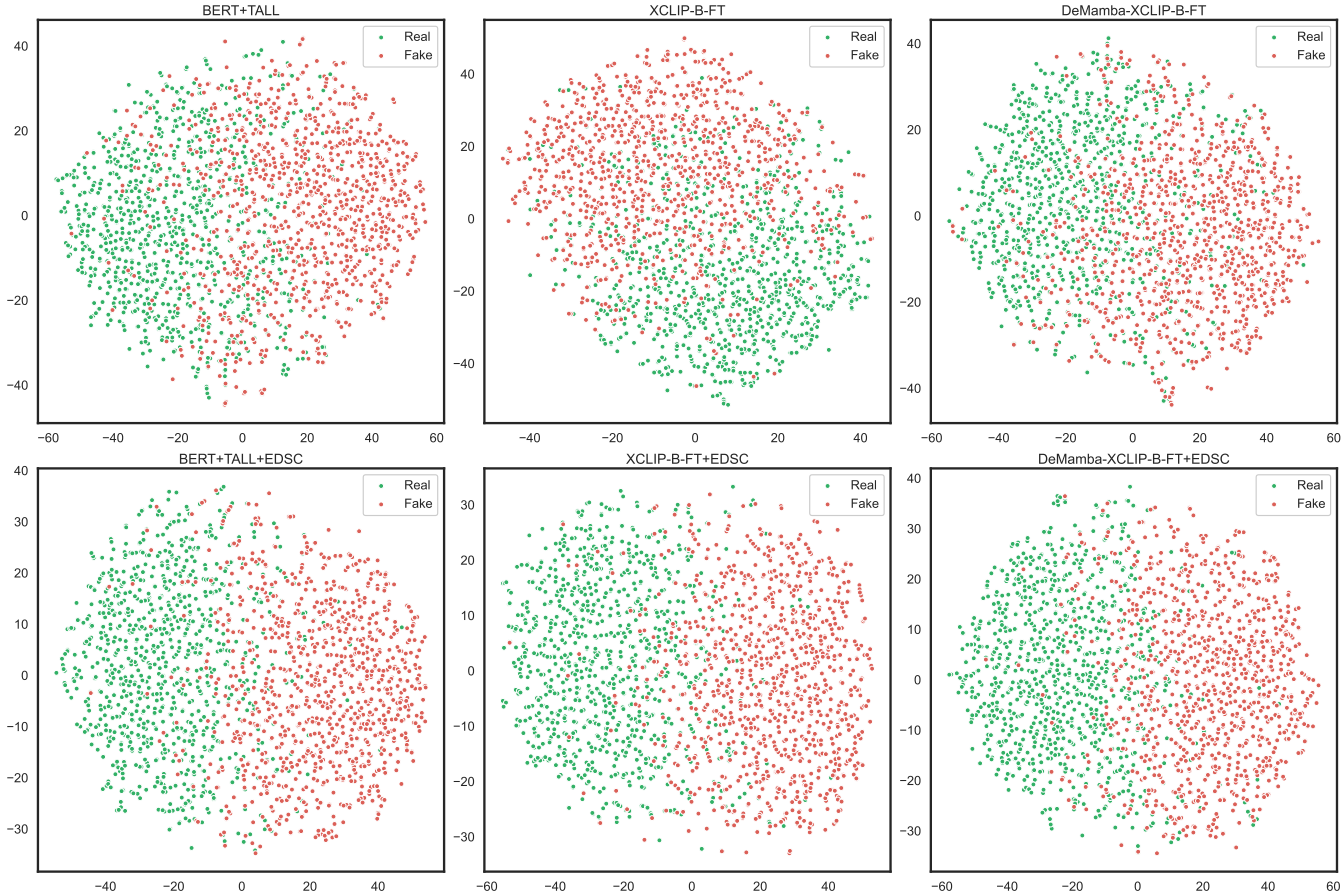


Figure 5: t-SNE visualization of the learned feature representations produced by BERT+TALL, XCLIP-B-FT, and DeMamba-XCLIP-B-FT, together with their EDSC-enhanced counterparts, on the SAVGE test set.

still contain ambiguous components that hinder reliable discrimination. In contrast, the EDSC-enhanced versions form more distinct clusters with clearer category boundaries. This observation is consistent with the design of EDSC, which explicitly suppresses semantic information in the artifact branch and strengthens intra-group discrimination through semantic-aware contrasting.

The visualization also shows that the benefit of EDSC is not limited to a specific backbone architecture. Although the degree of separation varies across models, the same qualitative improvement is consistently observed after applying EDSC, demonstrating its compatibility with different detector designs. These results provide intuitive evidence that the proposed framework learns more discriminative representations, helping explain the quantitative gains reported in Table 2.

6 Discussion on Broader Impacts

AI-generated videos offer substantial value for creative production, education, and entertainment, but they also introduce risks related to misinformation, impersonation, and the erosion of trust in digital media. Robust detection methods are therefore important for supporting content authenticity and the responsible deployment of generative video technologies. In this work, SAVGE and EDSC are designed to advance AIGV detection by enabling the study of artifact-aware and semantically grounded representations. In particular, the semantic alignment provided by SAVGE may facilitate the development of detectors that generalize better across generators and application scenarios.

Related concerns are especially important in medical AI, where recent work has surveyed agentic clinical systems, established unified benchmarks for multimodal medical multi-agent systems, and developed evidence-aware distillation for medical vision–language models [57–59]. Although these applications differ from media forensics, they similarly underscore the need to ground model decisions in relevant multimodal evidence and to evaluate robustness across domains.

However, we acknowledge the dual-use nature of this research. Although the dataset and methodology are intended to strengthen detection, they may also inform efforts to evade existing detectors. For this reason, we advocate responsible data release and transparent usage policies.

In the long term, AIGV detection should be viewed as one component of a broader governance framework that also includes platform policy, public awareness, and human oversight. Care is especially needed in deployment, since false positives may affect legitimate content and creators. We hope this work supports future research on trustworthy and responsible synthetic media ecosystems.

7 Conclusion

In this paper, we introduce SAVGE, a large-scale dataset for AI-generated video (AIGV) detection. By constructing semantically aligned real–fake video pairs and providing textual descriptions as semantic anchors, SAVGE offers rich supervisory signals for learning robust and semantically grounded AIGV detectors. Building on this dataset, we further propose Explicit Disentangling and Semantic-aware Contrasting (EDSC), a model-agnostic framework that learns artifact-focused and more generalizable representations. Specifically, EDSC leverages the textual annotations in SAVGE to explicitly disentangle semantically invariant artifact features while using the provided grouping information to enforce a clearer separation between real and fake categories. Extensive experiments on both SAVGE and GenVideo demonstrate the effectiveness of EDSC and show overall improvements in robustness and cross-dataset generalization. We hope that SAVGE and EDSC will serve as useful resources for advancing future research in AIGV detection.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62476260 and the Fundamental Research Funds for the Central Universities under Grant WK2100000057.

Author Contributions

Conceptualization, J.Y., J.X. and J.L.; methodology, J.Y. and J.X.; validation, Z.Z. and Y.Q.; formal analysis, J.Y., J.X., Z.Z. and Y.Q.; investigation, J.Y., J.X., Z.Z. and Y.Q.; resources, X.H. and J.L.; data curation, J.Y., J.X., Z.Z. and Y.Q.; writing—original draft preparation, J.Y. and J.X.; writing—review and editing, X.H. and J.L.; visualization, Z.Z. and Y.Q.; supervision, X.H. and J.L.; funding acquisition, J.L.

Conflict of Interest

All the authors declare that they have no conflict of interest.

Data Available

SAVGE uses real videos and textual annotations from ActivityNet Captions (https://cs.stanford.edu/people/ranjaykrishna/densevid/), Youku-mPLUG (https://github.com/X-PLUG/Youku-mPLUG), and Something-Something V2 (https://www.qualcomm.com/developer/software/something-something-v-2-dataset). GenVideo, used for cross-dataset evaluation, is provided through the authors' repository (https://github.com/chenhaoxing/DeMamba). Inquiries about access to SAVGE should be directed to the corresponding author, Jiawei Liu (jwliu6@ustc.edu.cn).

References

[1] Wang, S., Liu, J., Wang, W., Jin, Y., Du, J., Han, Z.: MMGT: Motion Mask Guided Two-Stage Network for Co-Speech Gesture Video Generation. IEEE Transactions on Circuits and Systems for Video Technology **36**(2), 1446–1459 (2026) https://doi.org/10.1109/TCSVT.2025.3604109

[2] Wang, C., Hu, P., Zhao, H., Guo, Y., Gu, J., Dong, X., Han, J., Xu, H., Liang, X.: Uniadapter: All-in-one control for flexible video generation. IEEE Transactions

on Circuits and Systems for Video Technology **35**(6), 6059–6073 (2025) https://doi.org/10.1109/TCSVT.2025.3532495

[3] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., *et al.*: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) https://doi.org/10.48550/arXiv.2210.02303

[4] Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18444–18455 (2023)

[5] Jiang, Y., Wu, T., Yang, S., Si, C., Lin, D., Qiao, Y., Loy, C.C., Liu, Z.: Videobooth: Diffusion-based video generation with image prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6689–6700 (2024). https://doi.org/10.1109/CVPR52733.2024.00639

[6] Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: ModelScope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023) https://doi.org/10.48550/arXiv.2308.06571

[7] Yang, Y., Yue, G., Zhou, W., Mao, X., Wang, R., Zhao, B.: Expressive Human Volumetric Video Generation With Rich Text. IEEE Transactions on Circuits and Systems for Video Technology **36**(4), 5424–5436 (2026) https://doi.org/10.1109/TCSVT.2025.3628996

[8] Wang, C., Yan, S., Chen, Y., Wang, X., Wang, Y., Dong, M., Yang, X., Li, D., Zhu, R., Clifton, D.A., *et al.*: Denoising Reuse: Exploiting Inter-Frame Motion Consistency for Efficient Video Generation. IEEE Transactions on Circuits and Systems for Video Technology **35**(9), 8436–8451 (2025) https://doi.org/10.1109/TCSVT.2025.3548728

[9] Yan, Z., Zhang, Y., Fan, Y., Wu, B.: UCF: Uncovering Common Features for Generalizable Deepfake Detection. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 22355–22366 (2023). https://doi.org/10.1109/ICCV51070.2023.02048

[10] Chen, H., Hong, Y., Huang, Z., Xu, Z., Gu, Z., Li, Y., Lan, J., Zhu, H., Zhang, J., Wang, W., Li, H.: DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark. arXiv preprint arXiv:2405.19707 (2024) https://doi.org/10.48550/arXiv.2405.19707

[11] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to Detect Manipulated Facial Images. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV, pp. 1–11 (2019). https://doi.org/10.1109/ICCV.2019.00009

[12] Dong, J., Wang, W., Tan, T.: CASIA Image Tampering Detection Evaluation Database. In 2013 IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP, pp. 422–426 (2013). https://doi.org/10.1109/CHINASIP.2013.6625374

[13] Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2849–2858 (2017). https://doi.org/10.1109/ICCV.2017.308

[14] Tulyakov, S., Liu, M.-Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1526–1535 (2018). https://doi.org/10.1109/CVPR.2018.00165

[15] Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., Gao, J.: StoryGAN: A Sequential Conditional GAN for Story Visualization. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6322–6331 (2019). https://doi.org/10.1109/CVPR.2019.00649

[16] Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Notice of Removal: VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10209–10218 (2023). https://doi.org/10.1109/CVPR52729.2023.00984

[17] Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.-F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024) https://doi.org/10.48550/arXiv.2401.03048

[18] Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-Sora: Democratizing Efficient Video Production for All. arXiv preprint arXiv:2412.20404 (2024) https://doi.org/10.48550/arXiv.2412.20404

[19] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8692–8701 (2020). https://doi.org/10.1109/CVPR42600.2020.00872

[20] Durall, R., Keuper, M., Keuper, J.: Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7887–7896 (2020). https://doi.org/10.1109/CVPR42600.2020.00791

[21] Li, Y., Chang, M.-C., Lyu, S.: In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7 (2018). https://doi.org/10.1109/WIFS.2018.8630787

[22] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In Proceedings of International Conference on Machine Learning, pp. 813–824 (2021)

[23] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging Frequency Analysis for Deep Fake Image Recognition. In Proceedings of the 37th International Conference on Machine Learning, pp. 3247–3258 (2020)

[24] Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. In Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 77771–77782 (2023)

[25] Ojha, U., Li, Y., Lee, Y.J.: Towards Universal Fake Image Detectors that Generalize Across Generative Models. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24480–24489 (2023). https://doi.org/10.1109/CVPR52729.2023.02345

[26] Zhang, J., Jian, L., Ma, Z., Ma, J., Huang, X.: Forgery Localization Via Extracting Generic Features and Multiple Prior Fusion. Journal of Intelligent Computing and Networking **2**(2), 1–13 (2026) https://doi.org/10.64509/jicn.22.94

[27] Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3204–3213 (2020). https://doi.org/10.1109/CVPR42600.2020.00327

[28] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The DeepFake Detection Challenge (DFDC) Dataset. In arXiv Preprint arXiv:2006.07397 (2020). https://doi.org/10.48550/arXiv.2006.07397

[29] Ni, Z., Yan, Q., Huang, M., Yuan, T., Tang, Y., Hu, H., Chen, X., Wang, Y.: GenVidBench: A 6-Million Benchmark for AI-Generated Video Detection. Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence, 15582–15590 (2026) https://doi.org/10.1609/aaai.v40i18.38587

[30] Chen, W., Zheng, W., Zheng, Y., Chen, L., Zhou, J., Lu, J., Duan, Y.: GenWorld: Towards Detecting AI-generated Real-world Simulation Videos. arXiv preprint arXiv:2506.10975 (2025) https://doi.org/10.48550/arXiv.2506.10975

[31] Jiang, C., Dong, W., Zhang, Z., Yu, F., Peng, W., Yuan, X., Bi, Y., Zhao, M., Zhou, Z., Si, C., *et al.*: Ivy-Fake: A Unified Explainable Framework and Benchmark for Image and Video AIGC Detection. In Proceedings of the 2026 International Conference on Multimedia Retrieval, pp. 2438–2447 (2026). https://doi.org/10.1145/3805622.3810615

[32] Zheng, C., Suo, R., Lin, C., Zhao, Z., Yang, L., Liu, S., Yang, M., Wang, C., Shen, C.: D3: Training-Free AI-Generated Video Detection Using Second-Order Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12852–12862 (2025). https://doi.org/10.1109/ICCV51701.2025.01194

[33] Chen, H., Sun, J., Liu, Y., Shi, H., Hu, D.: Self-Supervised Disentangled Representation Learning via Compositional Invariance. IEEE Transactions on Circuits and Systems for Video Technology **36**(3), 3201–3214 (2026) https://doi.org/10.1109/TCSVT.2025.3625076

[34] Dang, Z., Luo, M., Jia, C., Dai, G., Wang, J., Chang, X., Wang, J.: Disentangled Representation Learning With Transmitted Information Bottleneck. IEEE Transactions on Circuits and Systems for Video Technology **34**(12), 13297–13310 (2024) https://doi.org/10.1109/TCSVT.2024.3443122

[35] Yu, M., Li, H., Yang, J., Li, X., Li, S., Zhang, J.: FDML: Feature disentangling and multi-view learning for face forgery detection. Neurocomputing **572**, 127192 (2024) https://doi.org/10.1016/j.neucom.2023.127192

[36] Xu, H., Ye, Q., Wu, X., Yan, M., Miao, Y., Ye, J., Xu, G., Hu, A., Shi, Y., Xu, G., Li, C., Qian, Q., Que, M., Zhang, J., Zeng, X., Huang, F.: Youku-mPLUG: A 10 Million Large-scale Chinese Video-Language Dataset for Pre-training and Benchmarks. arXiv preprint arXiv:2306.04362 (2023) https://doi.org/10.48550/arXiv.2306.04362

[37] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In Proceedings of the IEEE International Conference on Computer Vision, pp. 706–715 (2017). https://doi.org/10.1109/ICCV.2017.83

[38] Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., *et al.*: The "something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision, pp. 5842–5850 (2017). https://doi.org/10.1109/ICCV.2017.622

[39] cerspense: ZeroScope-v2-576w. Hugging Face (2023). https://huggingface.co/cerspense/zeroscope_v2_576w

[40] Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., *et al.*: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023) https://doi.org/10.48550/arXiv.2310.19512

[41] Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., *et al.*: Wan: Open and Advanced Large-Scale Video Generative Models. arXiv preprint arXiv:2503.20314 (2025) https://doi.org/10.48550/arXiv.2503.20314

[42] Zhang, Y., Xing, Z., Zeng, Y., Fang, Y., Chen, K.: PIA: Your Personalized Image Animator via Plug-and-Play Modules in Text-to-Image Models. arXiv preprint arXiv:2312.13964 (2023) https://doi.org/10.48550/arXiv.2312.13964

[43] Morph studio. https://www.morphstudio.com/ Accessed 2026-09-18

[44] Zhang, D.J., Wu, J.Z., Liu, J.-W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision **133**, 1879–1893 (2025) https://doi.org/10.1007/s11263-024-02271-9

[45] Runway Research: Gen-2: Generate Novel Videos with Text, Images or Video Clips. https://research.runwayml.com/gen2 Accessed 2026-09-18

[46] Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., *et al.*: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023) https://doi.org/10.48550/arXiv.2309.15103

[47] Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling video masked autoencoders with dual masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14549–14560 (2023). https://doi.org/10.1109/CVPR52729.2023.01398

[48] Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In Proceedings of European Conference on Computer Vision, pp. 1–18 (2022). https://doi.org/10.1007/978-3-031-19772-7_1

[49] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423

[50] Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In Proceedings of 7th International Conference on Learning Representations, ICLR (2019)

[51] Nowozin, S., Cseke, B., Tomioka, R.: f-GAN: training generative neural samplers using variational divergence minimization. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 271–279 (2016)

[52] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) https://doi.org/10.48550/arXiv.1807.03748

[53] Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Proceedings of European Conference on Computer Vision, pp. 86–103 (2020). https://doi.org/10.1007/978-3-030-58610-2_6

[54] Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28130–28139 (2024). https://doi.org/10.1109/CVPR52733.2024.02657

[55] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In Proceedings of International Conference on Machine Learning, pp. 8748–8763 (2021)

[56] Xu, Y., Liang, J., Jia, G., Yang, Z., Zhang, Y., He, R.: Tall: Thumbnail layout for deepfake video detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22658–22668 (2023). https://doi.org/10.1109/ICCV51070.2023.02071

[57] Hu, X., Qian, Y., Yu, J., Liu, J., Ji, X., Tang, P., Xu, C., Liu, J., *et al.*: The Landscape of Medical Agents: A Survey. TechRxiv (2026) https://doi.org/10.36227/techrxiv.176581395.56964766/v4

[58] Qian, Y., Hu, X., Yu, J., Xin, S., Chen, X., Zhang, J., Jiang, P.-T., Liu, J., Li, H.B.: MedMASLab: A Unified Orchestration Framework for Benchmarking Multimodal Medical Multi-Agent Systems. arXiv preprint arXiv:2603.09909 (2026) https://doi.org/10.48550/arXiv.2603.09909

[59] Qian, Y., Yu, J., Liu, J., Wang, M., Li, H.B., Hu, X.: Med-OPD: Improving Medical Vision-Language Models via Evidence-Aware On-Policy Distillation. arXiv preprint arXiv:2607.16303 (2026) https://doi.org/10.48550/arXiv.2607.16303