

Detecting Hyper-Realistic Videos Generated by Diffusion Models via Text-Guided Semantic Enhancement

Shum Hing Ling Zhenyu Zhou Ajay Kumar *
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University, Hong Kong

{hing-ling.shum, zhenyu.zhou}@connect.polyu.hk, ajay.kumar@polyu.edu.hk

Abstract

This paper addresses two key challenges in detecting diffusion-generated videos: generalization to unseen but sophisticated video synthesizers and threats from the multimodal manipulations where the deepfakes combine the text prompts and visual synthesis to create realistic forgeries. We propose a novel multimodal framework built on a transformer-based architecture, which was originally designed for image forgery detection. Our framework extends this architecture by integrating two complementary components: (i) spatio-temporal feature extraction and (ii) a text-guided enrichment module which uses a frozen Vision Language Models (VLMs) text encoder when prompts are available and a small set of learnable default embeddings when no prompt is provided. Trained on our self-curated dataset comprising KlingAI and StableDiffusion Samples, we present cross-dataset performance from unseen but hyper-realistic fake video generators comprising Sora, Luma, Pika, and Runway. Our model can achieve high accuracy and outperformance results, which demonstrate cross-model generalization for detecting hyper-realistic AI-generated videos.

1. Introduction

Recent advances in diffusion models have revolutionized video generation, enabling text-to-video and image-to-video synthesis with unprecedented realism [33, 7]. Commercial platforms like OpenAI’s Sora [35], Luma Dream Machine [31], and Stability AI’s Stable Video Diffusion (SVD-XT) [43] have democratized high-quality video creation, but simultaneously pose significant threats to information integrity and media authenticity. Current video forgery detection methods face two critical limitations. First, [17] mentioned [3, 30, 48, 49] still struggle with cross-model generalization—detectors trained on one generator often fail when confronted with videos from unseen, so-



Figure 1. Overview of our approach to consolidate the spatial feature and the dual-path text-guided feature modules.

phisticated synthesizers [57]. This limitation stems from their reliance on generator-specific artifacts rather than fundamental forgery patterns. Second, existing multimodal approaches neither leverage the full generation context [34] nor address scenarios where no text prompt [27, 42, 49] is provided. Modern video synthesis is inherently multimodal [22, 41], guided by text prompts that provide semantic context for generation, yet purely visual detection methods overlook this crucial information [32, 45].

These limitations are particularly problematic in real-world applications where: (1) new generators emerge faster than detection methods can adapt, and (2) generated content lacks accessible prompt information during detection. Current multimodal detection approaches [42, 27, 49] either require prompts at inference time or cannot handle prompt-absent scenarios effectively.

1.1. Our Work

Inspired by contrastive language-image learning [40], these works [34, 18, 20, 54, 56] employ a frozen pre-trained model paradigm with an attached classifier for forgery detection by aligning the vision features with the language features in the multimodal space. Followed by [42, 27, 49],

*Corresponding Author

they utilize the language to guide the model during training. Besides the spatial information of image, the frequency domain [19, 12] information also can enhance the performance of fake detection. Therefore, to address these gaps, we introduce a novel multimodal framework (see overview in Fig. 1) that unifies spatio-temporal feature extraction with text-guided feature enrichment. Our approach builds upon a transformer-based architecture originally designed for image forgery detection [27], extending it with two key innovations: (a) Spatio-Temporal Feature Extraction: We adapt the forgery-aware visual encoder to process video sequences, capturing frame-level artifacts through transformer-based temporal modeling. (b) Text-Guided Feature Enrichment: We introduce a dual-path semantic module that processes text prompts via a frozen CLIP [40] text encoder when available, while falling back to learnable default embeddings when text prompts are not available. These embeddings are initialized randomly and optimized end-to-end to capture semantic patterns distinguishing real from generated content. After training, each set of learnable embeddings becomes a semantic fingerprint. Therefore, these embeddings are able to capture meaningful semantic patterns that aid in detecting fake videos, even though they originate from random noise.

The framework introduced in this paper can achieve strong cross-dataset generalization by learning semantic representations that are not influenced by artifacts specific to individual generators. We train and test on text-to-video (KlingAI [22]) and image-to-video (SVD-XT [43]) dataset samples as within-generator performance, then evaluate on four unseen (or unknown) generators: Sora [35], Luma [31], Pika [38], and Runway [41] as cross-generator performance. Our comparative experimental results in this paper achieve outperforming results over state-of-the-art (SOTA) methods, across unseen generators, and validate the merit of our approach in detecting hyper-realistic fake videos.

2. Related Work

Visual Forgery Detection: To efficiently detect the fake images or fake videos, many recent works are proposed to detect the forgery based on the image-based [23, 48, 55] and frequency-based [13, 19, 10, 15, 37, 44, 50, 12]. Moreover, the video-level forgery detection task also has been explored in the literature [9, 6, 21, 25, 53]. [39] introduced the frequency analysis into the detection framework, by utilizing decomposed high-frequency components. [22, 43, 35, 38, 31, 41] can now generate fake videos more easily. For detecting the diffusion-based video, the most recent work [42] proposed the MM-Det model by using the Large Multimodal Models (LMMs) with In-and-Across Frame Attention (IAFA) mechanism.

Cross-model Generalization: To enhance generalization, [47] adopt various data augmentations and large-scale GAN

images to improve the generalization to unseen testing data. The self-supervised learning methods [16, 36] can be used to improve the model generalization by learning more robust features. Recent efforts [51] have improved generalization to unseen forgery techniques. Li et al. [26] introduce KID, a multi-task learning approach that injects “real-data” prior into ViT-based backbones. Yermakov et al. [4] propose Human Action CLIPS, leveraging the CLIP’s ViT-L/14 visual encoder [40] for cross-dataset robustness. Liu et al. [29] develop LAVID, employing LVLMs through an agentic framework to adapt reasoning for novel artifacts. DIVID [28] uses a CNN + LSTM architecture to capture the temporal features and dynamic variations between frames of out-of-domain videos. These methods underscore the value of semantic signals but remain limited when prompt information is unavailable at test time and do not provide a comparison with the state-of-the-art FatFormer [27].

Multimodal Detection Frameworks: Inspired by the contrastive language-image pre-training [40], many works [18, 20, 54, 56, 5] have used the pre-trained paradigm by freezing the pre-trained weights and adopting an attached classifier for forgery detection. The UniFD [34] explores the potential of (vision-language models) VLMs, i.e., CLIP [40], for synthetic image detection. Furthermore, the FatFormer [27] presents a novel forgery-aware adaptive transformer approach based on the CLIP [40]. Prior work has begun to combine vision and language for video forgery: (1) Some approaches [42, 27, 49] append prompt embeddings to visual features but require prompt access at inference. (2) Others [34, 18, 20, 54, 56] train separate text and image pipelines without a unified consolidation strategy. However, there is no effort to learn a fallback embedding for the no-prompt setting or jointly optimize semantic and visual streams in a single transformer-based architecture.

3. Our Framework

The block diagram of our framework is shown Fig. 2, and the key task is to achieve binary classification on unknown input videos v . Each video input undergoes are firstly preprocessed and represented as tensors of shape $[B, T, C, H, W]$, where B is the batch size, T is the number of uniformly sampled frames, C represents number of channels (RGB), while $H \times W$ represents the spatial dimensions of each of the video frames. Since our backbone visual encoder operates on individual frames, we reshape the 5D video tensor to $[B \times T, C, H, W]$ for efficient batch processing. After feature extraction, we reorganize the outputs to $[B, T, D]$, where D represents the feature vector dimension for subsequent temporal modeling.

3.1. Visual Feature Extraction and Representation

3.1.1 Visual Feature Extraction

Each frame I_t is processed by a CLIP ViT-L/14 [40] visual encoder $\phi(\cdot)$ with weights inherited from a transformer-

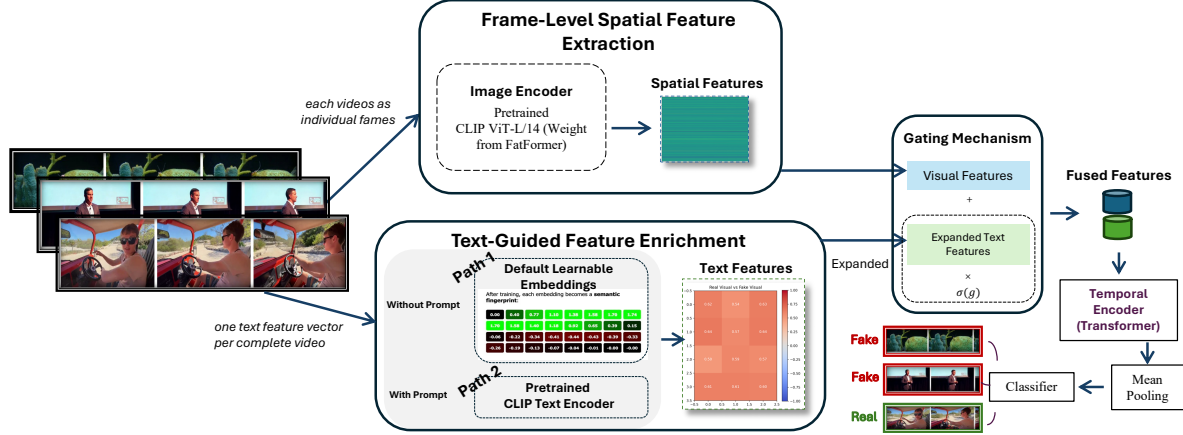


Figure 2. Our framework to accurately detect hyper-realistic fake videos.

based image forensics model [27]:

$$\mathbf{f}_t = \phi(I_t), \mathbf{f}_t \in \mathbb{R}^{1024} \quad (1)$$

, where \mathbf{f}_t represents the 1024-dimensional feature vector for the frame t , and $\phi(\cdot)$ denotes the visual encoder function. We preserve the full 1024-dimensional space to uncover subtle diffusion artifacts. To mitigate domain shifts across generator sources d , we incorporate learned layer normalization:

$$\hat{\mathbf{f}}_t = \text{LayerNorm}_d(\mathbf{f}_t), \hat{\mathbf{f}}_t \in \mathbb{R}^{1024} \quad (2)$$

, where $\text{LayerNorm}_d(\cdot)$ represents dataset-specific normalization parameters, and $\hat{\mathbf{f}}_t$ denotes the normalized frame features. This step generates features, which are used to build a composite sequence $\{\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_T\}$.

3.1.2 Dual-Path Text-Guided Feature Enrichment

Our key innovation lies in the dual-path semantic enrichment module that maintains text or semantic guidance even in the absence of text prompts or descriptions. Our approach can handle two real-world scenarios.

Path 1: Prompt-Available Scenario: When text prompts p are available with unknown videos, we extract semantic embeddings \mathbf{s} using a frozen CLIP text encoder [40]:

$$\mathbf{s} = \psi(p), \mathbf{s} \in \mathbb{R}^{768} \quad (3)$$

where $\psi(\cdot)$ represents the frozen CLIP text encoder [40], and \mathbf{s} denotes the extracted text embedding of dimension 768.

Path 2: Prompt-Free Scenario: When text prompts or video text descriptions are unavailable, we generate a small set of learnable default embeddings matrix [42]:

$$\mathbf{S}_{\text{default}} \in \mathbb{R}^{k \times 768} \quad (4)$$

, where $\mathbf{S}_{\text{default}}$ represents the learnable embedding matrix with k embeddings initialized randomly and optimized during training, initialized from $N(0, 1)$ and trained end-to-end

to generate representative semantics. Then, we index into the learnable default matrix $\mathbf{S}_{\text{default}}$ in Eq. (4) to pick one prototype $\mathbf{s} = \frac{\mathbf{S}_{\text{default}}(j)}{\|\mathbf{S}_{\text{default}}(j)\|_2} \in \mathbb{R}^{768}$ row by $j = \text{idx}(d) \in \{0, 1, \dots, k-1\}$, where $\text{idx}(d)$ is a simple lookup from video domain d . Regardless with or without prompt generators, generated videos would maintain coherent global semantic information, although the semantic of each frame has differences. Therefore, a simple lookup $\text{idx}(d)$ from video domain d can get the global video semantic information to align multimodal and enhance the performance which can be proved in Tab. 6.

In both pathways, we project the 768-dimensional text embeddings up to 1024 dimensions to align with our visual features. This upward projection preserves the rich information captured in the visual pathway and is performed using a learnable linear transformation:

$$\mathbf{e} = \mathbf{s}\mathbf{W}_{\text{proj}}, \mathbf{e} \in \mathbb{R}^{1024} \quad (5)$$

, where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{1024 \times 768}$ represents the learnable projection matrix, and \mathbf{e} denotes the resulting 1024-dimensional text feature. The resulting projected text features are then L2-normalized for consistent scaling.

$$\hat{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}, \hat{\mathbf{e}} \in \mathbb{R}^{1024} \quad (6)$$

, where $\|\cdot\|_2$ denotes the L2 norm, and $\hat{\mathbf{e}}$ represents the normalized text features. Generated videos with or without text prompts would maintain the same global semantic embedding $\hat{\mathbf{e}}$. Therefore, these normalized text features are then scaled along the temporal dimension to match the number of frames, resulting in a tensor of shape $[T, 1024]$. Concretely, we form:

$$\mathbf{E} = [\hat{\mathbf{e}}, \hat{\mathbf{e}}, \dots, \hat{\mathbf{e}}] \in \mathbb{R}^{T \times 1024} \quad (7)$$

3.1.3 Gated Consolidation of Text and Visual Features

In both pathways, we merge the normalized visual features $\mathbf{F} \in \mathbb{R}^{T \times 1024}$ with the aligned text features $\mathbf{E} \in \mathbb{R}^{T \times 1024}$

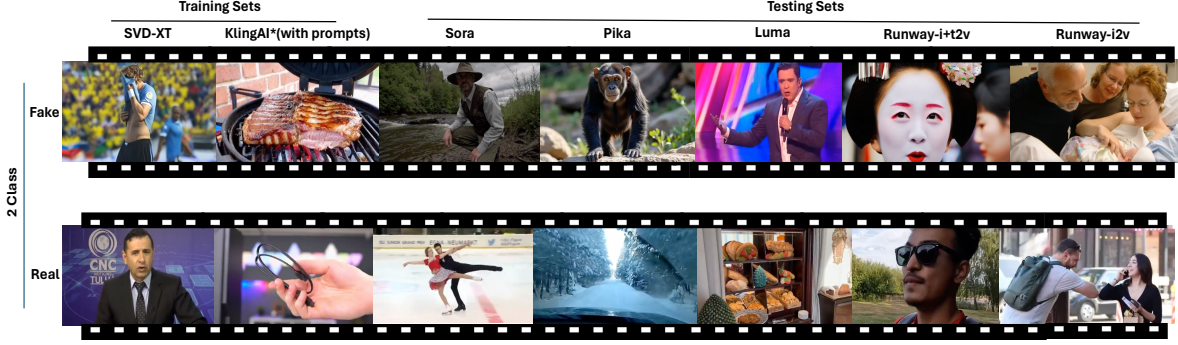


Figure 3. Overview and samples of images from our hyper-realistic fake/real video dataset.

via a learnable gating mechanism:

$$\mathbf{Z} = \mathbf{F} + \sigma(g)\mathbf{E} \quad (8)$$

Here, g is a learnable gating parameter and $\sigma(g) = \frac{1}{1+e^{(-g)}}$ is the sigmoid function that constrains in the $[0, 1]$ range, dynamically balancing visual and semantic cues. This mechanism enables the network to dynamically determine the importance of semantic (text) features based on their relevance to fakeness detection. The output or enriched feature \mathbf{Z} maintain the same dimensions $[T, 1024]$ and can capture both visual imprints and semantic context.

3.2. Temporal Modeling and Classification

3.2.1 Transformer Encoder Architecture

After text-visual consolidation, the enriched feature sequence $\mathbf{Z} \in \mathbb{R}^{T \times 1024}$ is processed by our temporal encoder to uncover frame-to-frame relationships. We employ a Transformer-based encoder consisting of two layers with eight attention heads per layer. Each Transformer layer implements self-attention through the conventional approach [46]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (9)$$

, where \mathbf{Q} , \mathbf{K} , and \mathbf{V} are query, key, and value matrices derived through learned projections of the input features, and $d_k = 1024$ is the dimension of the key vectors. The temporal encoder aggregates frame-level information into a coherent video representation:

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{Z}), \mathbf{H} \in \mathbb{R}^{T \times 1024} \quad (10)$$

After using the Transformer Encoder, the extracted embedding \mathbf{Z} of T frames of a video will be processed by self-attention, which assigns different weights to each frame. Finally, to aggregate these frame features and generate a single fixed-length representation for the entire video, we incorporate mean pooling across the temporal dimension:

$$\mathbf{h} = \frac{1}{T} \sum_{t=1}^T \mathbf{H}_t, \mathbf{h} \in \mathbb{R}^{1024} \quad (11)$$

This vector \mathbf{h} is expected to encapsulate the consolidated spatial and semantic imprints across all video frames and serve as the key input for the classification.

3.2.2 Network Training and Classification

The consolidated video representation $\mathbf{h} \in \mathbb{R}^{1024}$ pooled from the unknown video is used for a two-class classification using a lightweight MLP. This classifier firstly reduces the dimensionality of the data while extracting the most discriminative features for fake detection:

$$\hat{\mathbf{y}} = \text{Dropout}(\text{ReLU}(\mathbf{W}_1\mathbf{h} + \mathbf{b}_1))\mathbf{W}_2 + \mathbf{b}_2 \quad (12)$$

, where $\mathbf{W}_1 \in \mathbb{R}^{512 \times 1024}$, $\mathbf{b}_1 \in \mathbb{R}^{512}$ and $\mathbf{W}_2 \in \mathbb{R}^{2 \times 512}$, $\mathbf{b}_2 \in \mathbb{R}^2$ are the trainable parameters. The output logits $\hat{\mathbf{y}} \in \mathbb{R}^2$ represent the model's prediction for real (class 0) and fake (class 1) probabilities. We train the entire network in an end-to-end manner using cross-entropy loss between $\hat{\mathbf{y}}$ and the ground-truth label. During the inference, we use softmax to obtain class probabilities and accordingly generate predicted labels.

4. Experiments

4.1. Dataset Development

To evaluate of our framework's performance across generators, we use the KlingAI [22] and SVD-XT [43] as within-dataset. For unseen generators, we collected the latest diffusion-based models, Pika [38], Luma [31], and Runway [41], and the well-known Sora [35]. Therefore, we develop a balanced dataset of 1,088 video clips (544 fake, 544 real) [1] covering diverse content categories to evaluate cross-generator generalization capabilities from modern detectors. And the hyper-realistic video dataset is origined in Tab. 1.

Real Video Acquisition: Our real videos, as shown in Fig. 4, are sourced from three categories: (1) facial content from the CelebV-Text dataset [52], (2) scenery from the deployed YouTube-8M [2] on MM-Det [42], and (3) activities from the Panda-70M dataset [8]. This diversity ensures balanced pairing with synthetic samples across content types.

Table 1. Dataset statistics for hyper-realistic video detection.

Dataset Type	Subset	Real Videos	Fake Videos	Generation Approach	Model Version (Release Date)	Prompts Available	Prompts/Image Sources	Dataset Usage
Training	KlingAI	105	105	Text-to-Video	Kling 1.0 (2024-6-6)	✓	Panda-70M [8] (Train)	Train/Val/Test
	Stable Video Diffusion (SVD-XT)	325	325	Image-to-Videos	stabilityai 1.0 (2023-7-23)	✗	Open Images-V7 [14] (Train)	Train/Val/Test
Unseen Generator	Pika	33	33	Text-to-Video	Pika 2.2 (2025-2-27)	✗	Panda-70M (Test)	Test Only
	Sora	33	33	Text-to-Video	Sora Turbo (2024-12-9)	✗	Panda-70M (Test)	Test Only
	Luma	33	33	Text-to-Video	Ray 2 (2025-1-16)	✗	Panda-70M (Test)	Test Only
	Runway-i2v	8	8	Image-to-Video	Gen 4 (2025-3-31)	✗	Open Images-V7 (Test)	Test Only
	Runway-i&t2v	7	7	Image+Text-to-Video	Gen 4 (2025-3-31)	✗	Self-crafted prompts + Open Images-V7 (Test)	Test Only
Total		544	544					1088



Figure 4. Samples from our captured real video datasets from Panda-70M [8], CelebV-Text [52], and MM-Det [42].



Figure 5. Illustration for our process to generate the training dataset KlingAI [22] fake videos from text prompts.

Training sets (Within-Dataset): We create the training sets for training, validation, and testing from two distinct generation approaches:

- KlingAI (Text-to-Video) [22]: 105 fake videos generated using text prompts from the Panda-70M dataset, paired with 105 semantically similar real videos, as shown in Fig. 5.
- SVD-XT (Image-to-Video) [43]: 325 fake videos generated from Google Open Images V7 dataset [14] static images, paired with 325 corresponding real videos, as shown in Fig. 6.

Testing Sets (Unseen Generators): To evaluate cross-model generalization without any fine-tuning, we create test sets from four sophisticated generators not seen during training:

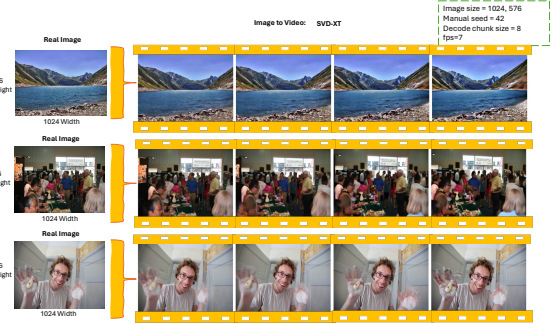


Figure 6. Illustration for our process to generate the training dataset SVD-XT [43] fake videos from images.

- Pika [38], Sora [35], Luma [31]: Each contains 33 text-to-video generated samples paired with 33 real videos, as shown in Fig. 7.
- Runway [41]: 8 image-to-video samples (Runway-i2v) and 7 hybrid image+text-to-video samples (Runway-i+t2v), each paired with real counterparts, as shown in Fig. 8.

Each fake video across all testing sets was paired with a semantically similar real video, resulting in a balanced test collection of 228 videos (114 real, 114 fake).

Content Diversity and Distribution: Our datasets encompass six main content categories with the following approximate distribution: human activities (40%), scenery (19%), food (12%), sports activities (10%), human face (8%), vehicles (6%), and animals (5%). This diversity helps ensure our model generalizes across different visual contexts.

Train-Validation-Test Split: For our training subsets (KlingAI [22] and SVD-XT [43]), we employed a 60:10:30 split for train, validation, and test partitions, respectively, while maintaining class balance within each partition. This resulted in approximately 258 videos for training, 43 for validation, and 129 for testing from these training subsets alone.

Class Balance and Sampling: We maintained a perfect 1:1 ratio between real and fake videos in both training and testing sets to ensure unbiased evaluation. Each generator

Table 2. Data preprocessing.

Parameter	Specification
Duration	4 seconds
Resolution	360p (height fixed, width autoadjusted)
Codec	H.264 video, AAC audio
Encoding Preset	Fast
Aspect Ratio	16:5 standard
Brightness/ Contrast	Standardized across all videos
Frame Extraction	8 frames uniformly sampled per video




Text to Video: Sora <i>"A man is sitting in a car and talking to the camera."</i>	Text to Video: Luma <i>"A newborn baby is wrapped in a blue blanket and appears to be crying."</i>	Text to Video: Pika <i>"There are sausages cooking on a grill, and a person is using tongs to turn them over."</i>
		

Figure 7. Example from unseen text-to-video generators (left to right): Sora [35], Luma [31], Pika [38].

subset (KlingAI [22], SVD-XT [43], Pika [38], Sora [35], Luma [31], and Runway [41]) has an equal number of real and fake samples.

Data Preprocessing and Normalization: We standardized all our videos in both classes using FFmpeg with the specifications in Tab. 2. All videos are standardized to 256×256 resolution, center-cropped to 224×224 , and normalized using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). This standardization ensures consistent processing across diverse video sources and eliminates potential artifacts from video compression or resolution differences as confounding factors.

Watermark Removal: All videos in our dataset were verified to be watermark-free to ensure the detector learns from content patterns rather than watermark artifacts. For Sora [35] videos, which typically include OpenAI’s watermark in the bottom-right corner, we apply consistent cropping (120 pixels from the bottom-right) to all videos in Sora’s [35] fake videos.

4.2. Implementation Details

Model Architecture Specifications: Our framework employs a 1024-dim CLIP ViT-L/14 backbone [40], a Transformer encoder, a projected frozen text encoder, 16 embeddings, adaptive fusion, and MLP classification, in Tab. 3. In our implementation, all feature processing and fusion occur within a unified 1024-dimensional latent space.

Training Configuration: During the training, we employed 16 learnable default embeddings to better capture the diversity of generation characteristics across different models, as in Tab. 4. We utilized the adaptive strategy to adjust text-visual gating value and 100% of the available text prompts for the KlingAI [22] dataset to maximize the

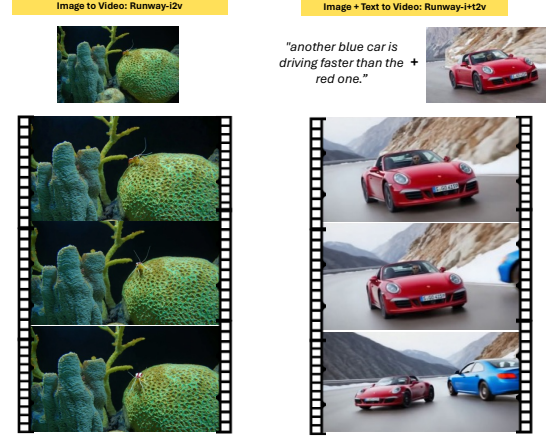


Figure 8. Example from unseen image/image&text to video generator: Runway [41].

Table 3. Our framework architecture specifications.

Component	Specification
Visual Backbone	CLIP ViT-L/14 (inherited from FatFormer [27])
Feature Dimension	1024
Temporal Encoder	Transformer (2 layers, 8 attention heads)
Text Encoder	Frozen CLIP Text Encoder
Text Feature Dimension	768 (projected to 1024)
Default Embeddings	16 learnable embeddings (1024-dim each)
Text-Visual Fusion	Adaptive gating mechanism
Dataset Normalization	Layer normalization per dataset source
Classification Head	2-layer MLP (1024→512→2) with ReLU activation

Table 4. Training hyperparameters.

Parameter	Value
Optimizer	AdamW
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-4}
Batch Size	8
Epochs	5
Number of Default Embeddings	16
Text-Visual Gating Value	Learnable adaptive gating
Prompt Availability	100% of available prompts
Temporal Model	Transformer
Temporal Layers	2
Attention Heads	8

Table 5. Training infrastructure.

Resource	NVIDIA A100-SXM4 GPU (40GB VRAM)
Platform	Google Colab Pro+
Training Time	22-23 mins for 5 epochs
Inference Time	0.05 seconds per video (8 frames)
Memory Usage	6-8GB GPU memory (batch size 8)
Parallel Processing	Single GPU training

semantic information available to the model.

Hardware: Training is conducted on A100-SXM4 GPU (40GB VRAM) with approximately 23 minutes training time for 5 epochs in Tab. 5. The efficient training is enabled by pre-trained backbone initialization.

Evaluation Metrics: We employ cross-generator evaluation, training exclusively on KlingAI and SVD-XT, then testing on four unseen generators. Performance is measured using Accuracy (ACC) and Area Under ROC Curve (AUC), like in [42]. Optimal classification thresholds are determined per dataset using validation sets (thresholds 0.1-0.9 in 0.05 increments).

Table 6. Performance comparison of video forgery detection. Note: Best results are in **bold**. “-” indicates results not reported in the respective references.

Dataset Type	Subset	2-class supervision	
		FatFormer	Ours
Training	Kling-t2v	95.24/98.07	95.24/99.29
	SVD-XT-i2v	96.92/ 99.75	97.95/99.70
Unseen Generator	Pika-t2v	87.88/97.80	95.45/98.35
	Sora-t2v	81.82/94.12	86.36/94.58
	Luma-t2v	63.64/80.81	75.76/83.29
	Runway-i2v	68.75/87.5	75.00/89.06
	Runway-i&t2v	85.71/95.92	92.86/100.00
Mean		82.85/93.42	88.37/94.90

Table 7. Performance analysis using video generation approach (ACC%/AUC%). Note: Best results are in **bold**.

Generation Approach	Generators	FatFormer [27]	Ours	Improvement
Text-to-Video	KlingAI	82.15/92.70	88.20/93.88	+6.05/+1.18
	Pika Sora Luma			
Image-to-Video	SVD-XT Runway-i2v	82.84/93.63	86.48/94.38	+3.64/+0.75
Hybrid (Image&Text)	Runway-i&t2v	85.71/95.92	92.86/100.00	+7.15/+4.08
Mean	All Unseen	77.56/91.23	85.09/93.06	+7.53/+1.83

4.3. Main Experimental Results

Our semantic-assisted framework significantly outperforms three recent baselines, i.e., FatFormer [27], using only videos from unseen generators without prompts. Results (ACC%/AUC%, Tab. 6) show large margins: Pika [38]: 95.45%/98.35%; RUNWAY-i&t2v [41]: 92.86%/100%; Sora [35]: 86.36%/94.58%; Luma [31]: 75.76%/83.29%.

Comparison with State-of-the-Art Methods: Our framework demonstrates consistent improvement over previous state-of-the-art methods across most datasets in Tab. 6. On KlingAI [22] and SVD-XT [43] training datasets, our approach achieves comparable or superior performance (KlingAI: 95.45%/99.29%, SVD-XT: 97.95%/99.70%), demonstrating effective learning of training distributions. The most significant improvements are observed on unseen generators, particularly on Pika [38] with a +7.57% accuracy improvement over FatFormer [27], demonstrating superior cross-model generalization. Across all generators (Trainig + Unseen Generator), our framework achieves 88.37%/94.90% Tab. 6, substantially outperforming prior methods.

Cross-Generator Generalization Analysis: We analyze performance across different generation approaches to understand our framework’s robustness in Tab. 7. Our framework achieves 88.20%/93.88%, across text-to-video generators (KlingAI [22], Pika [38], Sora [35], Luma [31]), outperforming FatFormer (82.15%/92.70%) by +6.06% accuracy. This demonstrates the effectiveness of our text-guided approach for detecting sophisticated text-to-video content. On image-to-video content (SVD-XT [43], Runway-i2v [41]), our approach maintains strong performance (86.48%/94.38%), showing robust detection of mo-

Table 8. Number of learnable default embeddings.

#. default embeddings	Mean Test (per-subset)
1	86.10/94.05
2	86.10/94.05
4	86.61/93.94
8	86.21/ 94.83
16	87.61/92.17

tion synthesis artifacts regardless of generation paradigm. Most notably, on Runway’s [41] hybrid image&text-to-video content, our framework achieves 92.86%/100.00%, matching state-of-the-art performance while demonstrating robustness to complex multimodal generation techniques. Across all unseen generators, our framework achieves 85.09%/93.06% Tab. 7, substantially outperforming the FatFormer [27]. As shown in Tab. 6, all the models obtain lower performance on the Luma and Runway datasets. The reason maybe from the domain gap between the past and current diffusion-based models. These models are trained on earlier KlingAI and SVD-XT models struggle to generalize to these newer generators. On the other hand, this kind observation proves the robustness of the Luma and Runway generators.

4.4. Ablation Study

We conduct five ablation experiments to verify the effectiveness of key components in our framework. Unless specified otherwise, we report the Mean Test (per-subset) ACC%/AUC% in a set of 4 learnable default embedding, learned gating parameter, and 100% prompt available inputs, while using Transformer as the temporal encoder across test datasets (Pika [38], Sora [35], Luma [31], and Runway-i&t2v [41]) in this ablation study session.

Larger Number of Learnable Default Embeddings: It tests how the number of default embeddings (e.g., 1, 2, 4, 8) affects performance. In the implementation, we use 4 default embeddings. We observe that performance increases substantially with 16 embeddings, improving accuracy by +2.80% compared to our baseline 4-embedding configuration, Problem: Inconsistency with Tab. 4 in this ablation study. The result in Tab. 8 presents the effect of varying the #. number of these embeddings (1, 2, 4, 8, 16) and demonstrates that a larger embedding space better captures the diversity of generation characteristics across different models. While the minimal differences exist between 1, 2, and 4 embeddings, suggesting that even a small number of learnable embeddings provides reasonable performance.

High Text-Influence Gate Works Reasonably Well for Hybrid Generator: A fixed gate value of 0.75 achieves the best performance with +1.54% ACC and +0.10% AUC over our learned approach. Tab. 9 compares our learnable gate parameter against fixed values in 4 quartiles (0.25, 0.5, 0.75, 1). The result indicates that text features should contribute significantly (approximately 75%) to the final representation for optimal detection performance. While our learned gate approach doesn’t outperform the best fixed value in

Table 9. Gating mechanism analysis.

	gate value	Mean Test (per-subset)
fixed	0.25	85.36/93.92
	0.5	86.36/93.88
	0.75	88.15/94.04
	1	87.65/94.12
learned	adjusts during training	86.61/93.94

Table 10. Comparison between Fixed Gating and Learned Gating.

	Mean Test (per-subset) fixed 0.75 gating parameter	Mean Test (per-subset) learned gating parameter
Runway-i&t2v	85.71/100	92.86/100
Runway-i2v	75.00/90.63	75.00/89.06

this controlled experiment, it eliminates the need for manual tuning, potentially offering better adaptability to unseen generators.

Comparison between Fixed Gating and Learned Gating: To further evaluate our Gate Mechanism, we conducted an extra ablation experiment. The results in Tab. 10 demonstrate that while a fixed high text-influence gate (0.75) works reasonably well for hybrid generation, the learned gate significantly improves performance on hybrid content (+7.15% accuracy) while maintaining similar performance on pure image-to-video content. It suggests our implemented adaptive approach can better balance the influence of text and visual features based on the content type.

The learned gate appears to recognize when text information is more relevant (for hybrid generations) and adjusts accordingly, while not overemphasizing text features when they are less informative (for pure image-to-video content). However, we believe that this observation about Runway could be presented as a direction for future work, specifically investigating how detection methods can be further enhanced for hybrid generation techniques where the boundary between real and synthetic content becomes even more blurred.

Prompt Free Inconsistency Detection: It evaluate performance when different percentages of training samples have prompts available (0%, 25%, 50%, 75%, 100%) to show how our model leverages available text guidance in Tab. 11. The best results are achieved at 75% prompt availability, outperforming both the 0% baseline (+2.46% ACC, +0.34% AUC) and the 100% configuration (+0.94% ACC, +0.06% AUC). Tab. 11 investigates how varying % percentages (0% 100%) of training samples with text prompts affect performance. This suggests that some diversity in feature representation, i.e., fixing prompted and default embeddings, benefits the model by exposing it to both scenarios during training. This finding validates our dual-path approach to text feature processing and demonstrates its robustness even with partial prompt information.

4.5. Summary

We built our framework using a Transformer temporal encoder trained with 100% prompt-available samples, 16 learnable default embeddings, and an adaptive gating mech-

Table 11. Prompt availability impact.

train samples with prompts (%)	Mean Test (per-subset)
0	85.09/93.66
25	84.85/93.76
50	85.09/93.66
75	87.55/94.00
100	86.61/93.94

anism. Ablation findings then demonstrate (1) Temporal modeling: LSTM (+0.79% ACC/+1.00% AUC) can outperform Transformer, suggesting sequential processing is valuable. (2) Embedding size: Moving from 4 to 16 defaults boosts ACC by +1.00%, showing larger semantic capacity aids generalization. (3) Gating: A fixed 0.75 gate matches or exceeds the learned gate (+1.54% ACC), though the learned version avoids manual tuning. (4) Prompt mix: Training with 75% prompts yields the best mean ACC/AUC (+2.46%/+0.34% over 0%), validating the dual-path design. These results confirm both our final architectural choices and the specific effects of each component on cross-model generalization.

5. Conclusions and Further Work

This paper introduced a novel multimodal framework for detecting hyper-realistic diffusion-generated fake videos that addresses two critical challenges: cross-model generalization and multimodal manipulation detection. By extending the transformer-based forgery detection mechanism with spatio-temporal modeling and integrating a dual-path text-guided enrichment module, our framework achieves robust performance across unseen generators. Our key innovation lies in learnable default embeddings that provide semantic context when prompts are unavailable. This enables practical deployment of our framework in real-world scenarios where generation prompts are not available. A comprehensive evaluation on hyperrealist unseen generators (Pika [38], Sora [35], Luma [31], Runway [41]) demonstrates superior cross-model generalization, achieving 85.09% mean accuracy with 93.06% AUC in Tab. 7.

Performance on the most challenging generator (Luma [31]) reaches 75.76% accuracy, indicating room for improvement on sophisticated content. Computational cost remains significant due to the ViT-L/14 backbone, limiting real-time applications. Enhancing training data diversity, developing adaptive text-visual fusion mechanisms, and exploring video-specific pretraining could further improve robustness and efficiency. The semantic enhancement approach presented in this paper can also be used to detect a range of synthesized (fake) biometric images, e.g. fingerprints [11] or iris [24], and opens promising directions for multimodal deepfake detection. The further extension will also evaluate the performance of this forgery detection model under adversarial attacks. Any potential of misuse of our framework can be prevented by deployed under strict access control like by the law enforce departments.

References

- [1] Web link to download our collected dataset. https://web.comp.polyu.edu.hk/csajaykr/Hyper_Realistic_Videos_Dataset.zip, 2025. 4
- [2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 4
- [3] A. Aghasanli, D. Kangin, and P. Angelov. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 467–474, 2023. 1
- [4] M. Bohacek and H. Farid. Human action clips: Detecting ai-generated human motion. *arXiv preprint arXiv:2412.00526*, 2024. 2
- [5] N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee, et al. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025. 2
- [6] C. Chang, Z. Liu, X. Lyu, and X. Qi. What matters in detecting ai-generated videos like sora? *arXiv preprint arXiv:2406.19568*, 2024. 2
- [7] H. Chen, Y. Hong, Z. Huang, Z. Xu, Z. Gu, Y. Li, J. Lan, H. Zhu, J. Zhang, W. Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024. 1
- [8] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-w. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 4, 5
- [9] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1133–1143, 2024. 2
- [10] C. T. Doloriel and N.-M. Cheung. Frequency masking for universal deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13466–13470. IEEE, 2024. 2
- [11] C. Dong and A. Kumar. Synthesis of multi-view 3d fingerprints to advance contactless fingerprint identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13134–13151, 2023. 8
- [12] C. Dong, A. Kumar, and E. Liu. Think twice before detecting gan-generated fake images from their spectral domain imprints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7865–7874, 2022. 2
- [13] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 2
- [14] Google LLC. Open images v7 dataset. <https://storage.googleapis.com/openimages/web/index.html>, 2020. Accessed: 2025-06-22. 5
- [15] A. A. Hasanaath, H. Luqman, R. Katib, and S. Anwar. Fsbic: Deepfake detection with frequency enhanced self-blended images. *Image and Vision Computing*, page 105418, 2025. 2
- [16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2
- [17] Y. Hong, J. Feng, H. Chen, J. Lan, H. Zhu, W. Wang, and J. Zhang. Wildfake: A large-scale and hierarchical dataset for ai-generated images detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3500–3508, 2025. 1
- [18] T. Huang, B. Dong, Y. Yang, X. Huang, R. W. Lau, W. Ouyang, and W. Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22157–22167, 2023. 1, 2
- [19] Y. Jeong, D. Kim, S. Min, S. Joe, Y. Gwon, and J. Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 48–57, 2022. 2
- [20] S. Jiao, Y. Wei, Y. Wang, Y. Zhao, and H. Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36:35631–35653, 2023. 1, 2
- [21] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, 57(6):159, 2024. 2
- [22] Kling AI. Kling ai. <https://www.klingai.com/global/>, n.d. Accessed: 2025-06-22. 1, 2, 4, 5, 6, 7
- [23] P. Korshunov, A. Jain, and S. Marcel. Custom attribution loss for improving generalization and interpretability of deepfake detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8972–8976. IEEE, 2022. 2
- [24] A. Kumar. *Iris and Periocular Recognition Using Deep Learning*. Elsevier, 2024. 8
- [25] R. Kundu, H. Xiong, V. Mohanty, A. Balachandran, and A. K. Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28050–28060, 2025. 2
- [26] T. Li, Y. Guo, Z. Liu, H. Peng, and Y. Wang. Deepfake detection via knowledge injection. *arXiv preprint arXiv:2503.02503*, 2025. 2
- [27] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024. 1, 2, 3, 6, 7
- [28] Q. Liu, P. Shi, Y.-Y. Tsai, C. Mao, and J. Yang. Turns out i’m not real: Towards robust detection of ai-generated videos. *arXiv preprint arXiv:2406.09601*, 2024. 2
- [29] Q. Liu, Y.-Y. Tsai, R. Zha, V. Li, P. Shi, C. Mao, and J. Yang. Lavid: An agentic lvlm framework for diffusion-generated video detection. *arXiv preprint arXiv:2502.14994*, 2025. 2

- [30] P. Lorenz, R. L. Durall, and J. Keuper. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 448–459, 2023. 1
- [31] Luma AI. Luma dream machine: New freedoms of imagination. <https://lumalabs.ai/dream-machine>, n.d. Accessed: 2025-06-22. 1, 2, 4, 5, 6, 7, 8
- [32] H.-H. Nguyen-Le, V.-T. Tran, D.-T. Nguyen, and N.-A. Le-Khac. Passive deepfake detection across multi-modalities: A comprehensive survey. *arXiv preprint arXiv:2411.17911*, 2024. 1
- [33] Z. Ni, Q. Yan, M. Huang, T. Yuan, Y. Tang, H. Hu, X. Chen, and Y. Wang. Genvidbench: A challenging benchmark for detecting ai-generated video. *arXiv preprint arXiv:2501.11340*, 2025. 1
- [34] U. Ojha, Y. Li, and Y. J. Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 1, 2
- [35] OpenAI. Sora. <https://openai.com/sora/>, n.d. Accessed: 2025-06-22. 1, 2, 4, 5, 6, 7, 8
- [36] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2
- [37] S. Peng, T. Zhang, L. Gao, X. Zhu, H. Zhang, K. Pang, and Z. Lei. Wmamba: Wavelet-based mamba for face forgery detection. *arXiv preprint arXiv:2501.09617*, 2025. 2
- [38] Pika. Pika. <https://pika.art/>, n.d. Accessed: 2025-06-22. 2, 4, 5, 6, 7, 8
- [39] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020. 2
- [40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 3, 6
- [41] Runway. Tools for human imagination. <https://runwayml.com/>, n.d. Accessed: 2025-06-22. 1, 2, 4, 5, 6, 7, 8
- [42] X. Song, X. Guo, J. Zhang, Q. Li, L. Bai, X. Liu, G. Zhai, and X. Liu. On learning multi-modal forgery representation for diffusion generated video detection. *arXiv preprint arXiv:2410.23623*, 2024. 1, 2, 3, 4, 5, 6
- [43] Stability AI. stable-video-diffusion-img2vid-xt. <https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt>, Jan. Published on Hugging Face. Accessed: 2025-06-22. 1, 2, 4, 5, 6, 7
- [44] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024. 2
- [45] D. S. Vahdati, T. D. Nguyen, A. Azizpour, and M. C. Stamm. Beyond deepfake images: Detecting ai-generated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4397–4408, 2024. 1
- [46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [47] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 2
- [48] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 1, 2
- [49] H. Wu, J. Zhou, and S. Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800*, 2023. 1, 2
- [50] S. Xiao, Y. Guo, H. Peng, Z. Liu, L. Yang, and Y. Wang. Generalizable ai-generated image detection based on fractal self-similarity in the spectrum. *arXiv preprint arXiv:2503.08484*, 2025. 2
- [51] A. Yermakov, J. Cech, and J. Matas. Unlocking the hidden potential of clip in generalizable deepfake detection. *arXiv preprint arXiv:2503.19683*, 2025. 2
- [52] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, and W. Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023. 4, 5
- [53] M. Zeng, X. Liao, C. Chen, N. Lin, Z. Wang, C. Chen, and A. Yang. Chameleon: On the scene diversity and domain variety of ai-generated videos detection. *arXiv preprint arXiv:2503.06624*, 2025. 2
- [54] G. Zhang, L. Wang, G. Kang, L. Chen, and Y. Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19148–19158, 2023. 1, 2
- [55] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021. 2
- [56] H. Zhu, Y. Wei, X. Liang, C. Zhang, and Y. Zhao. Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22257–22267, 2023. 1, 2
- [57] Y. Zou, P. Li, Z. Li, H. Huang, X. Cui, X. Liu, C. Zhang, and R. He. Survey on ai-generated media detection: From non-mlm to mlm. *arXiv preprint arXiv:2502.05240*, 2025. 1