SVD-Det: A Lightweight Framework for Video Forgery Detection Using Semantic and Visual Defect Cues

Tsung-Shan Yang[1], Tianyu Zhang[2], Feng Qian[2], Bing Yan[2], C.-C. Jay Kuo[1]
[1] University of Southern California, Los Angeles, CA 90089
[2] ByteDance Ltd., San Jose, CA 95110
{tsungsha, jckuo}@usc.edu
{tianyu.z, feng.qian, yanbing}@bytedance.com

Abstract

With the rapid proliferation of AI-generated content (AIGC) on multimedia platforms, efficient and reliable video forgery detection has become increasingly important. Existing approaches often rely on either visual artifacts or semantic inconsistencies, but suffer from high computational costs, limiting their deployment at scale. In this work, we propose ***SVD-Det****, a lightweight and efficient pipeline that leverages both* ***Semantic and Visual Defect cues*** *to detect forged videos. SVD-Det fuses spatiotemporal representations from raw RGB frames and compression-induced distortions using a 3D-Swin Transformer, and augments semantic understanding via CLIP-based embeddings. To integrate these heterogeneous modalities, we introduce* ***Domain-Query Attention (DoQA)****, a novel attention mechanism that hierarchically aggregates spatial and temporal features. Experiments across seven video generation domains demonstrate that SVD-Det not only achieves state-of-the-art detection performance but also reduces model size and inference time by over 97% and 98%, respectively, compared to LMM-based baselines. Our results highlight the practicality and robustness of SVD-Det for scalable AIGC detection in real-world scenarios.*

1. Introduction

With the rapid advancement of artificial intelligence (AI) generation models, AI-generated content (AIGC) has become increasingly prevalent on streaming multimedia platforms. As illustrated in Figure 1, individuals can now easily create and share high-quality videos via social video platforms. However, this surge in generated content poses a significant threat to the reliability of information delivered through web applications. While most existing methods focus on detecting forged images, their high computational cost often limits their applicability to video data.

Rather than directly training an AIGC detector, we begin by posing a fundamental question: "How do humans perceive that a video is artificial?" Humans can often detect AI-generated videos due to unnatural artifacts and implausible content. Beyond visual appearance, the way a video is encoded and transmitted also plays a crucial role in perception. Prior studies [26, 36] suggest that compression and transmission artifacts, such as blurring and compression, can obscure these revealing signs, making detection more difficult.

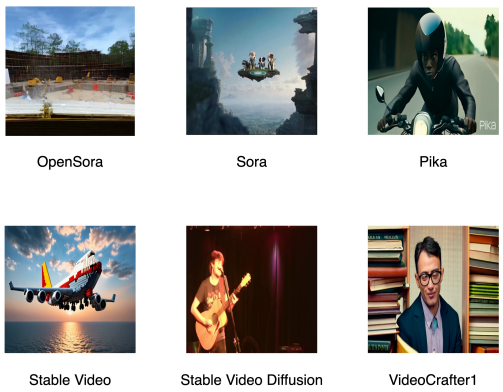

Figure 1. The example frames from AI-generated videos, including OpenSora, Sora, Pika, Stable Video, Stable Video Diffusion, and VideoCrafter1.

Currently, Visual Language Models (VLMs) deliver impressive performance across a wide range of tasks. However, these models cannot process entire videos at once; instead, they analyze samples of frames. For example, Figure 2 illustrates a failure case encountered in an online chatbot dialogue. This example highlights the challenges faced in real-world applications. Additionally, statistics indicate that over 30 million clips are uploaded to social media platforms daily. Examining all these clips using VLM agents would be prohibitively expensive, which hampers the com-

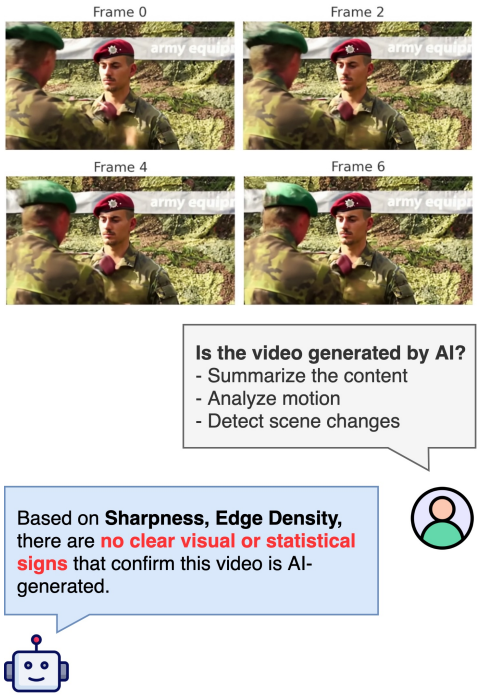

Figure 2. A failure case in online chatbot applications arises when the language model cannot identify inconsistencies among the frames of an AI-generated video.

mercial viability of the models due to high inference costs.

To address these challenges, we propose a lightweight and efficient video forgery detection model named **SVD-Det**, which integrates both **semantic** and **visual defect** cues. The effectiveness of each modality is visualized in figure 3. Specifically, SVD-Det extracts spatial and temporal information using 3D Swin Transformer [14] from both the raw RGB video and its compressed version. It also captures high-level semantic content via the CLIP visual encoder [20]. To unify these heterogeneous cues, we introduce a novel attention module, **Domain-Query Attention (DoQA)**, which enables effective multi-modal fusion and ultimately achieves better performance in detecting AI-generated content. Our main contributions are summarized as follows:

- We propose **SVD-Det**, a lightweight and efficient pipeline for video forgery detection, achieving state-of-the-art performance while reducing the number of parameters by 98% and the inference time by 97% compared to existing methods.
- We exploit compression artifacts as discriminative features, providing a computationally efficient alternative to reconstruction-based detection.
- We introduce **Domain-Query Attention (DoQA)**, a novel attention mechanism for fusing visual and semantic features effectively.

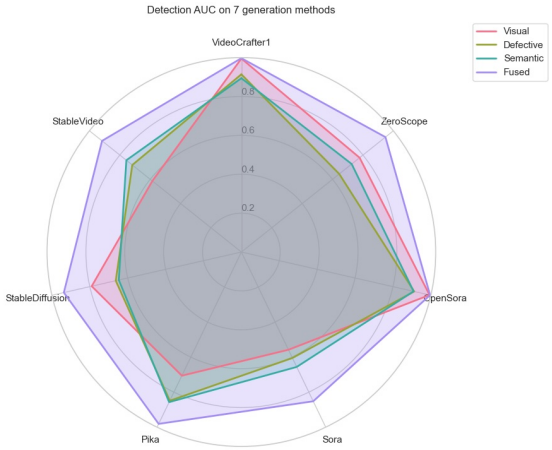

Figure 3. Performance of visual, defective, and semantic features. The 'fused' refers to the SVD-Det that incorporates features from all three modalities.

- We validate the proposed method on large-scale, user-uploaded content from a popular multimedia platform *without fine-tuning on the test set*, demonstrating strong robustness and generalizability in real-world applications.

2. Related Works

2.1. Forgery Image Detection

Recent research on forgery image detection [11, 28] has made significant advancements with the development of generative adversarial networks (GANs) [6] and diffusion-based techniques [9]. These approaches take advantage of the artificial patterns created by generative defects in images. Specifically, Guo et al. [7] proposed a hierarchical decision-making process to identify the type of generation used. However, their non-parametric method struggles to keep pace with the rapid evolution of generative techniques. Tan et al. [25] highlighted the importance of upsampling in generative methods and expanded the detection capabilities beyond CNN-based methods [16].

The advancement of generative models has limited the generalizability of forgery detection due to the removal of obvious unnatural patterns. To enhance detection capabilities, researchers are leveraging the semantic information from foundation models and Visual Language Models (VLMs). For instance, Ojha et al. [18] addressed the challenges posed by high-quality generated images by utilizing the CLIP model [20] to identify unnatural scenes.

2.2. Forgery Video Detection

Forgery video detection has evolved alongside various video manipulation techniques. Early research was influ-

enced by methods for face swapping. Zhao et al. [35] localized facial representations and identified artificial patterns within video frames. Haliassos et al. [8] focused on facial landmarks and demonstrated the inconsistencies in lip movement present in generated videos. Zheng et al. [36] incorporated temporal information into forgery detection by refactoring frame features using temporal transformer encoders. Additionally, Wang et al. [31] proposed an optimization strategy that alternately updates the parameters of spatial and temporal filters.

As video generation technology continues to advance, diffusion-based models [10, 37] are producing smooth and impressive videos. To identify high-quality generated content, video forgery detection must consider not only visual inconsistencies but also counterfeit semantic features. Wu et al. [32] harnessed the capabilities of large VLMs to develop a detection pipeline that incorporates reasoning and machine perception through prompting. Zhang et al. [34] introduced a question-and-answer pipeline using VLMs to detect unnatural patterns in videos. Additionally, Song et al. [24] enhanced performance in artificial intelligence-generated content (AIGC) detection by aligning features from a CNN backbone with tokens from LLM responses.

While VLMs achieve remarkable results, their intensive computational costs pose a significant challenge for practical applications. Additionally, video embedding inference still relies on a frame-by-frame approach, resulting in frequent input/output requirements during processing. To reduce computational overhead, we propose an end-to-end optimization pipeline that integrates a Video Swin Transformer [14] with a CLIP visual encoder [20]. This approach captures the underlying semantic representations of the data instead of relying on fine-tuned VLM tokens.

2.3. Defects of AI-generated Videos

Current AIGC detection methods focus on identifying clues related to defect patterns. Corvi et al. [4] analyzed the residual patterns reconstructed from a denoising model and showed that diffusion-based models produce forensic traces distinct from other generation models. Also, they note that 'well-trained' detectors cannot be generalized to 'unseen' patterns that were not present during training. Vahdati et al. [26] further visualized the frequency responses of denoised images and demonstrated that the model can be significantly affected by compression artifacts. Nguyen et al. [17] investigated the frequency responses of the manipulated images.

The diffusion process consists of an iterative denoising procedure. In this context, Wang et al. [30] proposed the DIRE loss, which calculates the reconstruction error arising from the denoising process. Diffusion-generated images can be nearly perfectly reconstructed using a pretrained model, whereas real images lack this reconstruction fidelity, revealing a key forensic distinction. Building on the concept of DIRE loss, Luo et al. [15] explored reconstruction loss in the latent space, $LaRE^2$, demonstrating trends similar to those observed with DIRE loss.

Despite the strong performance of VLMs, their high computational demands and frame-wise inference requirements hinder real-time deployment. To address this, we have developed a lightweight end-to-end pipeline for detecting AIGC and analyzing reconstruction errors stemming from compression artifacts.

3. Method

Inspired by the demands of streaming multimedia applications, uploaded videos are typically stored in various binary formats determined by codec settings and compression standards. While compression algorithms are designed to align with human perceptual quality, they often produce disproportionately large bitstreams in the presence of subtle inconsistencies, especially introduced in AI-generated content. To address this, our model leverages two input sources: the raw RGB video and its compressed counterpart, which captures distortion artifacts.

Aside from visual defects, humans often rely on semantic understanding to judge the authenticity of video content. Unnatural themes or implausible scenes are quickly flagged by human perception. To incorporate semantic representations into our process, we utilize the visual foundation model, CLIP [20], to extract the video's topic.

We introduce a novel model, **SVD-Det**, which utilizes both **Semantic** and **Visual Defect** cues to detect AI-generated videos. After extracting features from the respective modalities, our pipeline integrates them using a Domain-Query Attention mechanism. Spatial and temporal features are aggregated independently to enhance representational efficiency. The final prediction is derived from visual and semantic scores, each computed through separate feedforward networks (FFNs).

3.1. Model Overview

The proposed method, SVD-Det, is illustrated in Figure 4. SVD-Det does not depend on extensive training with compressed videos; instead, it utilizes the artifacts created by compression. The blocking artifacts that result from compression can also influence how humans perceive the video. While existing methods mainly target forgery detection at the frame level, maintaining temporal consistency across video clips is crucial for human perception. Rather than simply combining decisions from individual frames, SVD-Det utilizes the 3D-Swin block from the Video Swin Transformer [14] to capture temporal information through consecutive patches. The two sequences are concatenated and used as visual features.

Furthermore, we incorporate the CLIP visual encoder [20] to extract high-level semantic representations,

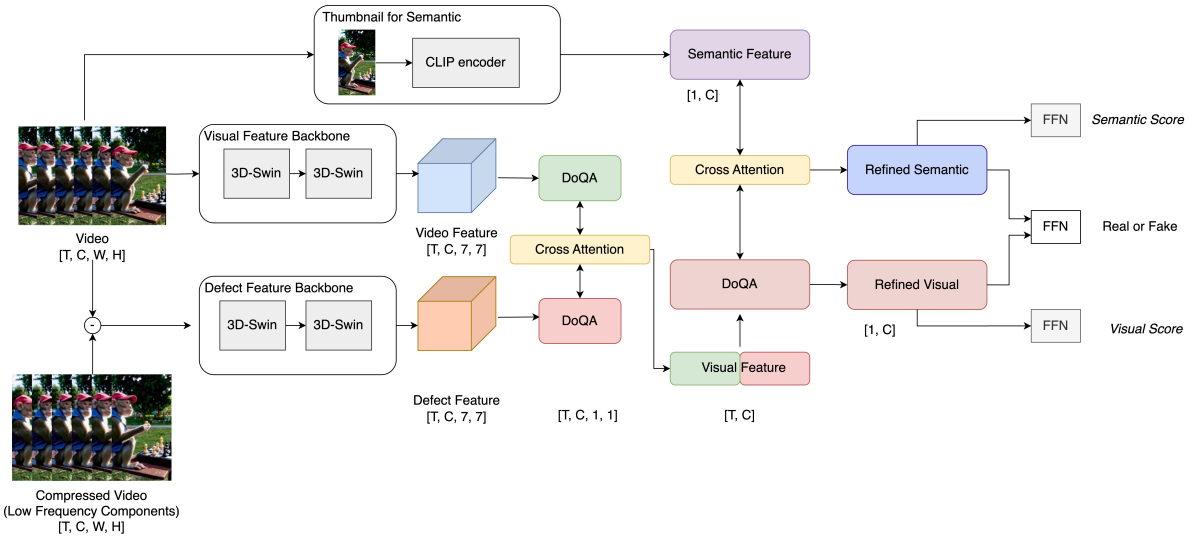

Figure 4. The proposed SVD-Det consists of an overall pipeline with three main components: (1) visual features, (2) defective features, and (3) semantic features. The visual and defective features are processed using a Video Swin Transformer [14], while the semantic features rely on a pretrained CLIP model [20]. Spatial attention is applied through a newly proposed Domain-Query Attention block. The refined visual and semantic information is then sent to separate feedforward networks (FFNs) to compute visual and semantic scores, which are used for the final decision.

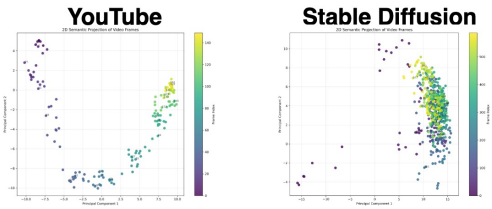

Figure 5. The semantic visualization of frames. The left image shows the video from YouTube, while the right image displays the frames from Stable Diffusion.

allowing us to identify forged clips based on unrealistic patterns. Figure 5 demonstrates that the semantic information is closely related across the generated clips. The compression features are elaborated in section 3.2, and the DoQA block structure is illustrated in section 3.3.

3.2. Compression Artifact

Conventional metrics for assessing compression quality, such as Video Multimethod Assessment Fusion (VMAF), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR), evaluate transient effects and noise levels statistically. While these metrics generally align well with real-world video distributions, they do not perform as effectively for AI-generated videos. Specifically, AI-generated videos often contain unnatural patterns that lead these metrics to incorrectly compensate for artifacts by suggesting higher bitrates, resulting in a deviation from accurate compression assessment. The frequency analysis is presented in Figure 6.

The bitrate partition in AI-generated videos often results in the preservation of unnatural patterns. As a result, while compressed AIGC maintains high-frequency patterns, it loses quality in low-frequency areas. When compressing videos at the same Constant Rate Factor (CRF), the distortions observed in real videos differ from those in counterfeit ones.

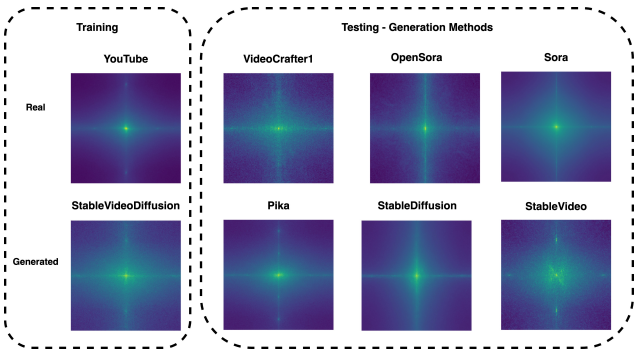

Figure 6. The frequency analysis of compression artifacts involves comparing the original and compressed frames using a two-dimensional Fourier transform. The training data comprises real videos sourced from YouTube, as well as videos generated by Stable Video Diffusion. The generated testing data includes VideoCraft1, OpenSora, Sora, Pika, Stable Diffusion, and Stable Video, organized from left to right and top to bottom.

As shown in Figure 6, the frequency-domain residue distributions reveal this contrast: real videos exhibit distortions

uniformly across all frequencies, whereas AI-generated videos show dominant low-frequency distortions. The vertical and horizontal axes intersecting at the origin in the frequency domain correspond to grid-like artifacts in the spatial domain. Consequently, SVD-Det processes two input streams—raw RGB frames and compression-induced distortion—to better capture these differences.

3.3. Domain-Query Attention (DoQA)

We propose a Domain-Query Attention (DoQA) block, consisting of two cascaded attention layers that gather information from neighboring tokens. The structure of this module is illustrated in Figure 7. To establish an efficient processing pipeline, we decompose operations across the three spatiotemporal dimensions into separate spatial and temporal stages. First, the DoQA block is applied to the spatial dimensions (height and width) in order to extract a compact representation for each frame. Then, a second DoQA block aggregates these frame-level representations across the temporal dimension.

The input to the first self-attention layer is formed by concatenating a learnable token with a sequence of neighboring tokens. The resulting feature tensor $F \in \mathbb{R}^{L \times C}$ and the pooled domain token $\hat{Q} \in \mathbb{R}^{C}$ is computed as follows:

$$\hat{Q},\ F = \texttt{self-attn}(\hat{X}) \tag{1}$$

$$\hat{X} = \texttt{concate}([DOM], X), \tag{2}$$

where $X \in \mathbb{R}^{L \times C}$ denotes the sequence of input tokens of length L and channel dimension C, and $[DOM] \in \mathbb{R}^{C}$ is a learnable token used to indicate the source of the features. The function `self-attn` denotes the self-attention mechanism [27], and `concate` denotes the concatenation operation along the sequence dimension.

To further concentrate the representation of the domain, the pooled domain token is passed through a multi-layered perceptron and then forwarded to a second cross-attention layer along with the sequence of input tokens:

$$Q = \texttt{MLP}(\hat{Q}) \tag{3}$$

$$V = \texttt{cross-attn}(Q, F), \tag{4}$$

where $Q \in \mathbb{R}^{C}$ denotes the refined pooled domain token, and $F \in \mathbb{R}^{L \times C}$ represents the sequence of feature tokens. The function `MLP` is a two-layer feedforward network, and `cross-attn` denotes the cross-attention mechanism [27]. The final representation $V \in \mathbb{R}^{C}$ is obtained by using Q as query and F as both the key and value in the cross-attention operation.

A learnable token is introduced to summarize contextual information from neighboring tokens. The representations can be further encoded into a compact vector, enabling the model to represent video content effectively while avoiding the curse of dimensionality. This design stabilizes training and improves generalization. The details of ablation studies can be found in section 4.7.

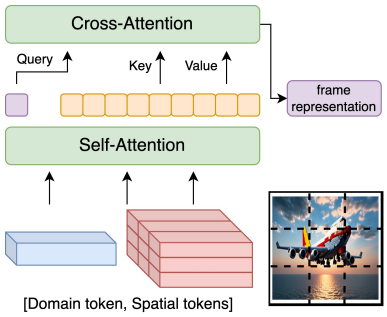

Figure 7. The proposed attention mechanism consists of several steps. For example, in the spatial merging block, the input for the first self-attention layer is formed by concatenating domain tokens with spatial tokens. In this setup, the domain token acts as the query, while the spatial tokens serve as both the keys and values in the second cross-attention layer. To simplify the explanation, the multi-layer perceptrons (MLPs) that connect the two attention mechanisms have been omitted.

3.4. Loss Function

To address the imbalanced label distribution in the detection dataset, we adopt the weighted focal loss [12] instead of the standard cross-entropy loss for binary classification. The classification loss is defined as:

$$L_{cls} = \alpha(1-p)^{\gamma}\log(p), \tag{5}$$

where α and γ are hyperparameters that control the weighting and focusing factor, respectively, and p denotes the predicted probability for the ground-truth class. The weighted focal loss down-weights easy examples and emphasizes learning from hard or minority-class examples, thereby improving the model's sensitivity to true positives in imbalanced settings.

Since the features consist of two modalities, namely visual and semantic, we introduce two auxiliary classification losses during training. The visual and semantic scores are obtained from two corresponding feedforward networks (FFNs), and both are supervised using the same ground-truth labels as the primary classification task. The auxiliary losses, L_{vis} and L_{sem}, follow the same hyperparameter configuration as the main classification loss L_{cls}. The overall objective function is defined as:

$$L_{total} = L_{cls} + w_{aux}(L_{vis} + L_{sem}), \tag{6}$$

where L_{total} represents the final objective function, and w_{aux} is a weighting factor that balances the primary classification loss with the auxiliary modality-specific losses.

4. Experiments

4.1. Experimental Settings

We utilize the tiny 3D-Swin transformer [14] as the backbone for processing RGB videos and the residues from the compressed video. Our semantic backbone is based on the CLIP model [20]. The backbone parameters are initialized with a warm start using pretrained weights, while the parameters in the attention blocks are initialized with a cold start. We employ the AdamW optimizer, setting the learning rate to 10^{-6} for the backbone and 10^{-5} for the attention blocks, with a weight decay of 10^{-3}. The weights for the auxiliary semantic and visual loss are set to 0.2.

4.2. Dataset

In this experiment, we follow the setup outlined in the DVF dataset [24] to ensure fair comparisons. The training set includes 1,000 real videos sourced from YouTube and 1,973 fake videos generated by Stable Video Diffusion [2]. Ten percent of the videos are set aside as a validation set, while the remainder is used for training. For evaluation, the real videos are obtained from Internvid10M [29], and the fake videos are generated using VideoCraft1 [3], ZeroScope, Opensora [37], Sora, Pika, Stable Diffusion [21], and Stable Video.

To minimize the effects of thresholding, we employ the Area Under the Curve (AUC) for performance evaluation. In our comparison of state-of-the-art methodologies, we assess techniques that utilize various backbone architectures. For frame-level forgery detection, we select CNNDet [28], F3Net [19], and HiFi-Net [7]. In contrast, ViViT [1], TALL [33], and TS2-Net [13] focus on integrating spatial and temporal information from consecutive frames at the video level. Using semantic information from CLIP, Clip-Raising [5], Uni-FD [18], and DE-FAKE [22] are designed to detect unrealistic videos based on their content. Additionally, DIRE [30] employs a reconstruction loss derived from diffusion-based generation, as described in the DDIM formulas [23]. Finally, MM-Det [24] leverages a large language model to extract features and combine textual and visual representations for effective detection.

4.3. Video Forgery Detection

Table 1 demonstrates that the proposed **SVD-Det** model achieves superior average AUC performance compared to state-of-the-art methods across seven AI-generated video domains. On average, SVD-Det surpasses MM-Det by 2.7% in AUC. Specifically, it outperforms MM-Det in five domains: VideoCraft1 (+4.2%), ZeroScope (+0.8%), Open-Sora (+10.8%), Pika (+2.2%), and Stable Video (+1.8%). For the remaining two domains, Sora and Stable Diffusion, SVD-Det delivers competitive results, with only marginal drops of 1.0% and 2.0%, respectively. These results highlight the robustness and generalizability of SVD-Det across diverse generation frameworks.

In comparison to DIRE [30], which focuses solely on defective features, SVD-Det incorporates additional information beyond just distribution, resulting in an AUC performance that is over 30% higher. Similarly, SVD-Det outperforms various semantic detection methods, such as Clip-Raising [5], Uni-FD [18], and DE-FAKE [22], by 20%. Furthermore, SVD-Det showcases its capability to integrate features from different modalities, achieving a 10% higher AUC compared to visual feature-based models like HiFi-Net [7], F3-Net [19], and ViVit [1].

4.4. Open-World Study

As shown in Table 2, we compare the proposed AIGC detector (**SVD-Det**) with the previous state-of-the-art method (MM-Det [24]) under an open-world detection setting. The dataset used in this study is not open-source. The data is collected from publicly available videos on a multimedia platform. Detailed dataset statistics are provided in the Appendix. It is important to note that the proposed model is not fine-tuned on this dataset, ensuring that the evaluation truly reflects its ability to generalize to unseen data.

From Table 2, it can be observed that SVD-Det consistently outperforms MM-Det across both the well-constructed (DVF) and open-world datasets. While the performance gap on DVF is modest (+2.7% AUC), the improvement on the open-world dataset is substantial (+22.6% AUC). This large gain demonstrates the robustness and generalizability of the proposed method in handling diverse, in-the-wild AI-generated content where training and testing domains differ significantly. These results highlight the effectiveness of incorporating multi-modal cues and the proposed Domain-Query Attention in boosting performance without any domain-specific fine-tuning.

4.5. Efficiency

While MM-Det represents a significant breakthrough in the field, it is important to consider the associated computational costs for real-world deployment. The process of frame-wise reconstruction requires frequent read and write operations on the disk, which can lead to high deployment expenses. Additionally, fine-tuning and inference of Visual Language Models (VLMs) necessitate substantial GPU resources. Currently, when a video is input into the VLM application, it is sampled frame by frame, and each frame is processed in a frame-wise manner. The temporal information is lost in the dialogues.

While VLMs are quite powerful, their answers are not always completely accurate. Figure 2 illustrates a failure case in video forgery detection using an online chatbot application. The VLM shows the capability to apply metrics from signal processing, including sharpness and brightness.

Method	VideoCrafter1	ZeroScope	OpenSora	Sora	Pika	Stable Diffusion	Stable Video	Avg
DIRE [30]	55.9	61.8	53.8	60.5	65.8	62.7	69.9	62.1
Raising [5]	63.8	60.7	64.1	68.8	70.7	78.2	62.8	67.0
TALL [33]	76.0	65.9	62.1	64.3	72.3	65.8	79.8	69.5
DE-FAKE [22]	74.7	68.2	55.8	64.1	85.6	85.4	70.6	71.2
TS2-Net [13]	61.8	70.6	75.5	78.0	78.2	62.1	78.6	72.1
Uni-FD [18]	75.0	71.2	76.6	73.1	76.2	80.2	66.7	74.1
CNNDet [28]	87.4	88.2	78.0	63.8	77.3	73.5	78.9	78.2
ViViT [1]	79.1	78.4	77.7	69.4	83.1	82.1	83.6	79.1
F3Net [19]	89.7	80.5	69.3	70.8	88.9	84.4	85.1	81.3
HiFi-Net [7]	90.2	89.7	80.1	70.1	87.8	89.2	83.1	84.3
MM-Det [24]	93.5	94.0	88.8	**86.2**	95.9	**95.7**	89.9	92.0
SVD-Det(Ours)	**99.7**	**94.8**	**99.6**	85.2	**98.1**	93.7	**91.7**	**94.7**

Table 1. Video forgery detection performance on the DVF dataset measured by AUC (%). The data is taken from [24]. The highest performance is marked in **bold**, while the second highest is underlined.

Data Source	Well-Constructed DVF	Open-World Social Media
MM-Det [24]	92.0	48.6
SVD-Det	**94.7**	**71.2**

Table 2. Comparison between the proposed method and the state-of-the-art AIGC detector under an open-world setting. Performance is measured by AUC (%), with open-world testing data collected from the public multimedia platform.

While its planning and generalization capabilities are impressive, the decision-making process can be computationally intensive. These metrics do enhance the reliability of the reasoning process; however, in the specific area of detection, SVD-Det provides a more economical solution and outperforms the VLM.

Table 3 presents a comparison of the training and inference costs between MM-Det and SVD-Det. During the inference stage, SVD-Det achieves a remarkable 97% reduction in model size and an impressive 98% decrease in inference time. In addition to the inference stage, the training process costs are also essential for deployment considerations. SVD-Det employs an end-to-end training scheme, whereas MM-Det relies on a Visual Language Model (VLM) to generate visual question-answering pairs, which are then used to fine-tune a 7 billion parameter Language Model.

4.6. Visualization

The attention map illustrated in Figure 8 highlights the artifact in the video through the use of the proposed DoQA module. This module enables the model to concentrate on unnatural patterns and artifacts, even as the time sequence extends. In contrast, when the input length increases, the heatmaps generated by the model without DoQA begin to diverge.

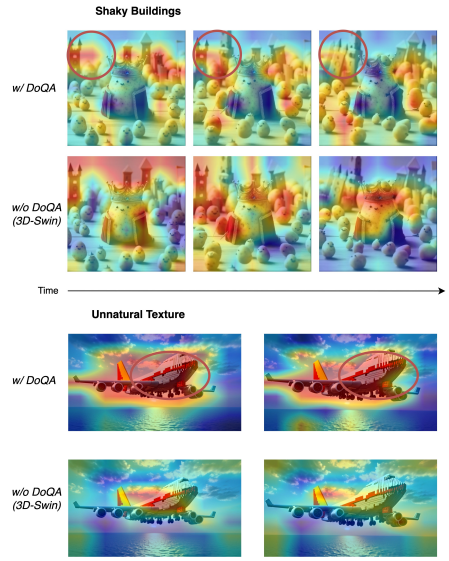

Figure 8. Visualization of attention heatmaps w/ and w/o DoQA. The first clip is sampled from Soar, while the second clip is taken from Stable Video. The upper row displays the attention heatmaps produced using DoQA, and the lower row shows the heatmaps generated without DoQA. These images represent the last layer of attention heatmaps.

4.7. Ablation Studies

In Table 4, we analyze features from different modalities. The performance of the original RGB feature indicates that the generation methods produce high-quality content, with no significant temporal inconsistencies in the videos. A similar trend is observed with the compression distortion features, which can reveal certain patterns. However, distortion should not be considered the sole indicator of quality. While semantic features can identify high-quality but unrealistic videos, they struggle to detect inconsistent spatial

	Avg AUC(%)	Model Size	Inference Time (sec/clip)	Training Recipe
MM-Det	92.0	7B (40x)	150 (50x)	(1) Prepare VQA pairs from larger VLM (2) Finetune 7B-VLM with VQA pairs (3) Train attention blocks
SVD-Det(Ours)	**94.7**	**180M (1x)**	**3 (1x)**	**End-to-end**

Table 3. The efficiency comparison between SVD-Det and MM-Det [24] is presented, with inference time measured on a single Tesla V100 GPU.

Features			Video-Crafter1	Zero-Scope	Open-Sora	Sora	Pika	Stable Diffusion	Stable Video	Avg
Visual	Defective	Semantic								
✓			99.5	77.8	99.3	55.8	70.6	79.0	58.6	77.2
	✓		91.4	64.4	91.2	60.5	84.9	66.3	71.8	75.8
		✓	89.4	72.6	90.9	65.6	85.7	64.7	75.7	68.5
✓	✓		93.5	67.5	93.3	64.2	87.8	61.1	76.6	77.7
✓		✓	93.7	64.8	93.7	60.0	79.5	76.8	78.8	79.6
✓	✓	✓	**98.0**	**70.7**	**99.2**	**70.6**	**80.6**	**78.4**	**79.7**	**81.0**

Table 4. The ablation studies demonstrate the functionalities of the features. The merging is performed through cross-attention. The ✓ mark indicates that the feature is applied during the model training stage.

and temporal patterns.

By combining visual and semantic cues, we observe improved performance across all generated videos. Ultimately, the final model, which incorporates three modalities, semantic, visual, and defective features, achieves the best results.

Attention	Open-Sora	Sora	Pika	Stable Diffusion	Stable Video	Avg
Cross-Attention	99.2	70.6	80.6	78.4	79.7	81.0
Self-Atteion	91.6	62.6	82.3	59.7	78.8	73.8
DoQA	**99.6**	**85.2**	**98.1**	**93.7**	**91.7**	**94.7**

Table 5. The ablation studies demonstrate the functionalities of the proposed Domain-Query Attention mechanism. All experiments are conducted using features from three modalities: visual, defective, and semantic information. To save space, we omit the results in VideoCrafter1 and ZeroScope($\sim$0.95), but the detailed experiment results can be found in the supplementary materials.

To validate the proposed Domain-Query Attention (DoQA), we present the performance results in Table 5. The self-attention mechanism uses the concatenation of tokens from both modalities, which increases the length of the input sequence and requires more computational resources. This complexity can also hinder the model's convergence. In contrast, cross-attention restricts the generalizability of features across different modalities. Therefore, the proposed DoQA effectively captures domain information from the query modality and integrates it with information from other modalities. Additionally, the cascaded structure of DoQA facilitates faster convergence.

5. Conclusion

In this paper, we presented **SVD-Det**, a novel and efficient framework for video forgery detection that jointly leverages semantic information and visual compression artifacts. Motivated by the way humans perceive AI-generated videos—through both unnatural appearance and implausible semantic content—we designed a dual-branch architecture that captures spatiotemporal patterns via a 3D Swin Transformer and semantic cues via CLIP embeddings. To effectively fuse multi-modal signals, we proposed the **Domain-Query Attention (DoQA)** module, which hierarchically aggregates spatial and temporal features using a lightweight attention mechanism.

Our extensive experiments on seven challenging AIGC domains demonstrate that SVD-Det achieves state-of-the-art performance while being significantly more computationally efficient than prior LMM-based methods. It reduces inference time by 98% and model size by 97%, making it practical for deployment in large-scale content moderation systems. We believe that SVD-Det opens new directions for scalable and interpretable video forgery detection in the era of generative media.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 6, 7

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi,

Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6

[3] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 6

[4] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[5] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024. 6, 7

[6] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[7] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 2, 6, 7

[8] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 3

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[11] Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon 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

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[13] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, pages 319–335. Springer, 2022. 6, 7

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2, 3, 4, 6

[15] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024. 3

[16] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3091–3095. IEEE, 2022. 2

[17] Tai D Nguyen, Shengbang Fang, and Matthew C Stamm. Videofact: detecting video forgeries using attention, scene context, and forensic traces. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8563–8573, 2024. 3

[18] Utkarsh Ojha, Yuheng Li, and Yong Jae 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. 2, 6, 7

[19] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020. 6, 7

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 3, 4, 6

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[22] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pages 3418–3432, 2023. 6, 7

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[24] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. *arXiv preprint arXiv:2410.23623*, 2024. 3, 6, 7, 8

[25] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2

[26] Danial Samadi Vahdati, Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Beyond deepfake images: Detecting ai-generated videos. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 4397–4408, 2024. 1, 3

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[28] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei 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, 6, 7

[29] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 6

[30] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 3, 6, 7

[31] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4129–4138, 2023. 3

[32] Guangyang Wu, Weijie Wu, Xiaohong Liu, Kele Xu, Tianjiao Wan, and Wenyi Wang. Cheap-fake detection with llm using prompt engineering. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 105–109. IEEE, 2023. 3

[33] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023. 6, 7

[34] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pages 399–415. Springer, 2024. 3

[35] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 3

[36] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021. 1, 3

[37] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 3, 6