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], 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}@bytedance.com

Abstract

This supplementary file provides a detailed explanation of the compression distortion in the real and generated videos of SVD-Det.

1. Semantic Information

Generated videos heavily depend on the provided prompts and initial frames. As a result, the semantic information within the time sequence contains redundant data along the temporal direction, which can hinder convergence during the training phase. To visualize the semantic behavior of the generated clips, we show the frame-wise representations obtained from the trained CLIP encoder in a scatter plot.

The values for the x- and y-axes are derived from the PCA deposition and project the vectors onto the largest and second-largest eigenvectors. As illustrated in Figure 1, the representations derived from a real-world video exhibit greater diversity in the CLIP features. In contrast, the representations generated by Stable Diffusion and Stable Video closely align with both the input prompt and the initial frame, as demonstrated in Figures 3 and 4. Conversely, the frames produced by Pika, shown in Figure 2, are diverse but become more grouped as time progresses.

2. Ablation Study

Attention	Video-Crafter1	Zero-Scope	Open-Sora	Sora	Pika	Stable Diffusion	Stable Video	Avg
Cross-Attention	98.0	70.7	99.2	70.6	80.6	78.4	69.7	81.0
Self-Atteion	89.3	62.4	91.6	62.6	82.3	59.7	68.8	73.8
DoQA	99.7	94.8	99.6	85.2	98.1	93.7	91.7	**94.7**

Table 1. 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.

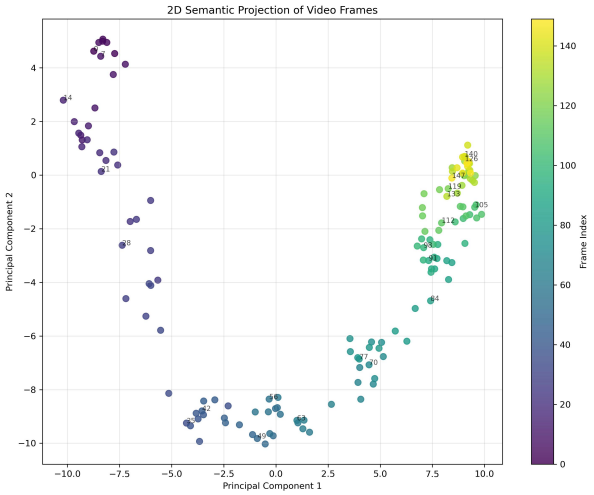

Figure 1. The semantic representations from the real dataset.

Table 1 presents the detailed results of experiments conducted on various attention mechanisms, which support the conclusions drawn in Section 4.7. The simplified metrics are displayed in Table 5. Notably, the use of the Domain-Query Attention (DoQA) mechanism leads to performance improvements in the domains of VideoCrafter1 and Zero-Scope.

3. Compression Defects

The figures below show that the user-generated videos depict high-frequency defects, i.e., randomly scattered dots on an image. On the contrary, the generative frames contain more patch-like defects after compression distortions. From left to right, there are the original frame, the compressed frame, and the difference between the two.

The residue consists of block-wise patterns in diffusion-based models. While user-generated videos exhibit homogeneous patterns, the residues are distributed in a dot-

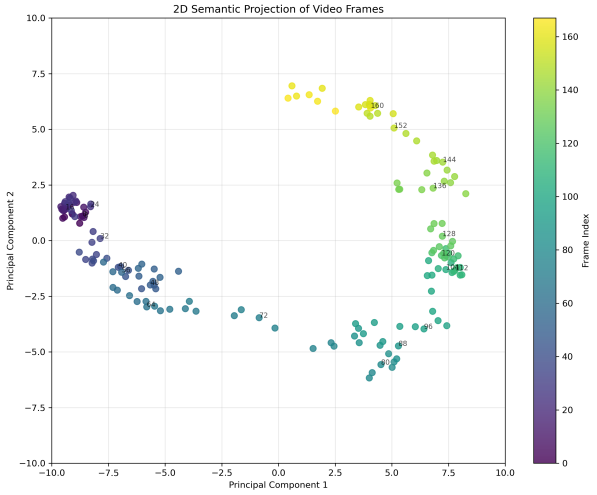

Figure 2. The semantic representations from the Pika dataset.

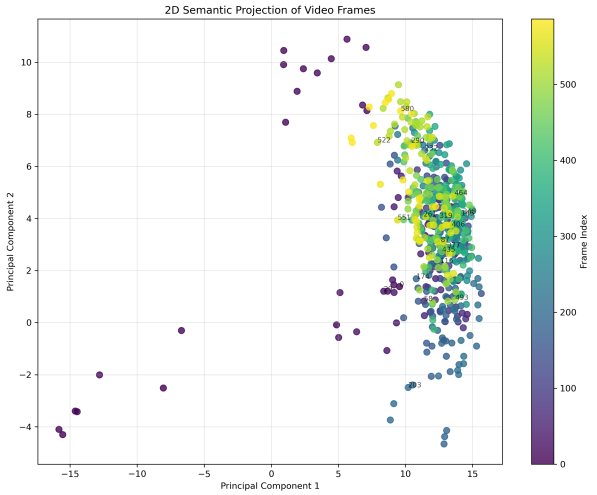

Figure 3. The semantic representations from the Stable Diffusion dataset.

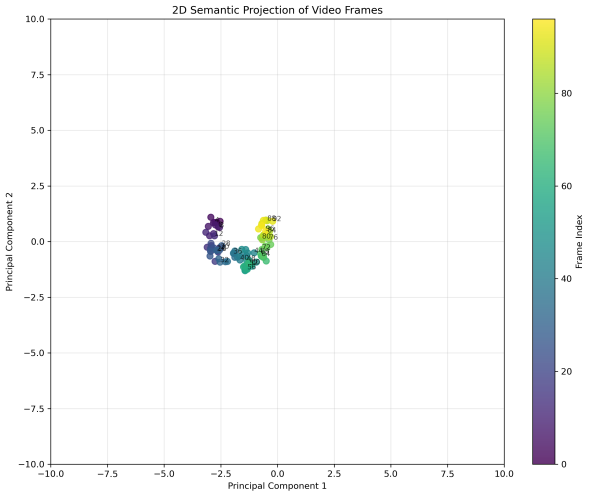

Figure 4. The semantic representations from the Stable Video dataset.

wise pattern. In contrast, diffusion-generated videos display patch-wise residue patterns, confirming the presence of low-frequency distortion mentioned in the paper.

The visualization patterns are illustrated in Figure 5, 6, 7, 8, 9, 10, 11, and 12.

Figure 5. Compressed frames from the user-generated dataset.

Figure 6. Compressed frames from the StableVideoDiffusion dataset.

Figure 7. Compressed frames from the VideoCrafter1 dataset.

Figure 8. Compressed frames from the Sora dataset.

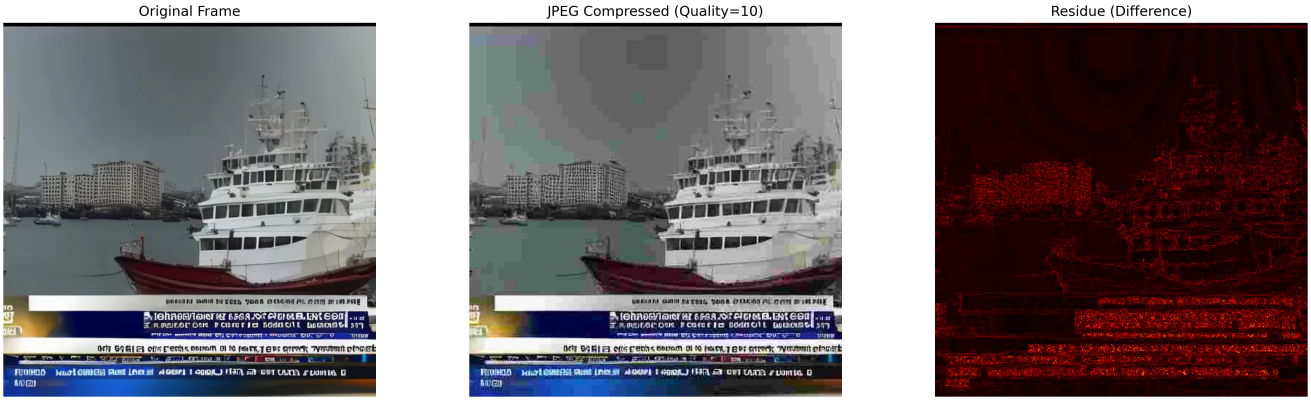

Figure 9. Compressed frames from the OpenSora dataset.

Figure 10. Compressed frames from the Pika dataset.

Figure 11. Compressed frames from the StableVideo dataset.

Figure 12. Compressed frames from the StableDiffusion dataset.