SWIFT: Sliding Window Reconstruction for Few-Shot Training-Free Generated Video Attribution

Supplementary Material

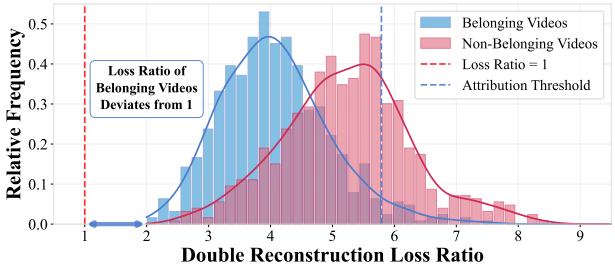


Figure 1. Performance of AEDR in video modality (belonging model: HunyuanVideo [3], non-belonging model: EasyAnimate [7]). The loss ratio for attributed videos deviates from 1, and the increased attribution threshold (Threshold=5.79) results in non-belonging videos being incorrectly classified as belonging.

1. Performance Degradation of AEDR

Although AEDR [6] performs excellently in image modality, its accuracy significantly drops when transferred to video modality. The cause of this phenomenon lies in the introduction of complex temporal characteristics in video modality, which disrupts the core assumption of AEDR: the loss ratio for belonging samples in two consecutive reconstructions should be close to 1, while the ratio for non-belonging samples is significantly greater than 1. Specifically, video VAEs need to simultaneously account for both spatial and temporal dimensions, and their latent space exhibits a high degree of temporal coupling—motion information and dependencies between adjacent frames are compressed and shared within the latent space, which considerably increases the difficulty of reconstruction. Therefore, this complex temporal coupling and cross-frame dependencies lead to a systematic increase in the loss of the first reconstruction, causing the loss ratio for belonging videos to deviate from 1 as well (see Fig. 1). The failure of AEDR stems from its focus solely on spatial consistency during reconstruction, while videos also entail temporal consistency.

2. More Details of the S-Video

Our dataset (S-Video) comprises a total of 4,000 videos, with 500 real videos and 3,500 generated videos. The real videos were randomly sampled from the OpenVidHD [4], while the generated videos were produced using 700 randomly selected prompts from OpenVidHD and five different generative models. Fig. 2 presents samples of videos generated using the same prompts across the five models. Fig. 3 displays samples of real videos, and Tab. 1 provides a list of the prompts used in the generation process.

Table 1. Examples of text prompts used to create the S-Video.

Prompt Examples
The video features a young man with a black cap and a black t-shirt, standing in a skate park. He is smiling and looking directly at the camera. The skate park is filled with ramps and rails, and the ground is covered in a light blue paint. The background is slightly blurred, but it appears to be a cloudy day. The man's casual attire and the skate park setting suggest that this is a casual.
In the video, a person is seen preparing a salad in a white bowl placed on a wooden table. The salad consists of various ingredients such as lettuce, tomatoes, beans, and avocado. The person is using a spoon to mix the ingredients together. The table also has a pen and a bowl of chips. The overall style of the video is casual and homey, capturing a simple yet enjoyable cooking moment.
The video features a woman with blonde hair and blue eyes, wearing a black top and gold earrings. She is making a peace sign with her right hand and has her mouth open as if she is speaking or singing. The background shows a staircase with a white railing and a white wall. The style of the video is casual and informal, with a focus on the woman's expression and gesture.
In the video, a man and a woman are seated at a dining table in a kitchen, enjoying a meal together. The man is wearing a yellow and black plaid hoodie, while the woman is dressed in a pink hoodie. They are both eating from white bowls filled with a green and yellow dish, possibly a soup or stew. The kitchen is equipped with modern appliances, including a microwave and an oven.

3. Evaluation Metrics

In this paper, we utilize accuracy (Acc) as the metric to evaluate the attribution performance of the algorithm. The goal of SWIFT is to determine whether a given video belongs to the target model, which involves binary classification (belonging vs. non-belonging). The evaluation takes into account four key outcomes: True Positives (TP), where belonging videos are correctly identified; False Positives (FP), where non-belonging videos are misclassified as belonging; False Negatives (FN), where belonging videos are mistakenly identified as non-belonging; and True Negatives (TN), where non-belonging videos are correctly identified.

"The video captures a man riding a motorcycle on a city street. He is wearing a black helmet and a black jacket, and he has a backpack on the back of the motorcycle. The motorcycle is black and red, and it has a red stripe on the side. The man is riding on the right side of the street, and he is in motion. The street is lined with buildings, and there are trees in the background. The sky is overcast, and the lighting is soft. The video is shot in a realistic style, and it captures the man and his motorcycle in great detail."

(a) The following five video samples are generated using this prompt.

(b) A sample video generated by HunyuanVideo [3].

(c) A sample video generated by Wan2.1 [5].

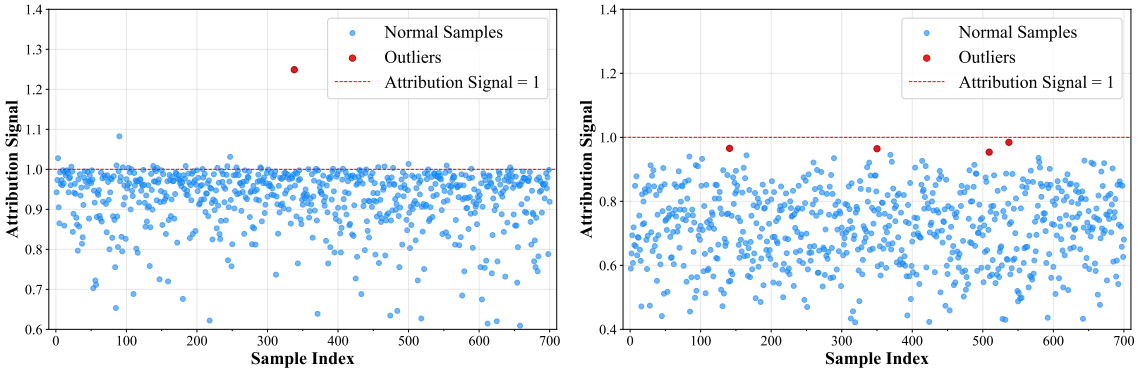
(d) A sample video generated by LTX-Video [1].

(e) A sample video generated by EasyAnimate [7].

(f) A sample video generated by Wan2.2 [5].

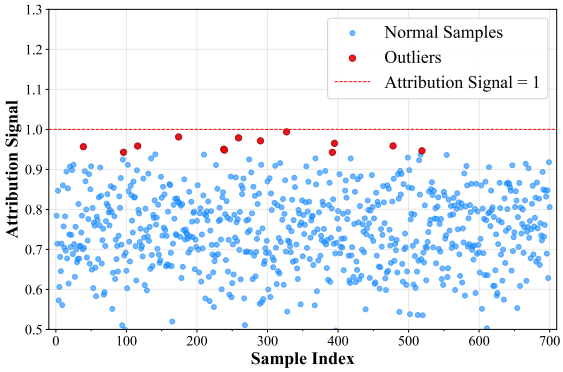
Figure 2. Video samples generated by the five models using the same prompt.

Figure 3. A real video sample randomly drawn from the OpenVidHD-1M [4].

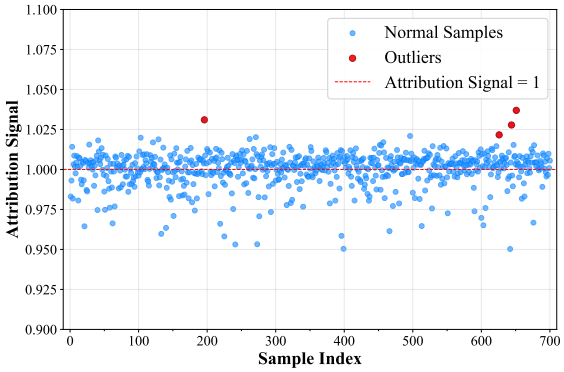


(a) Attribution signal distribution of HYV [3].

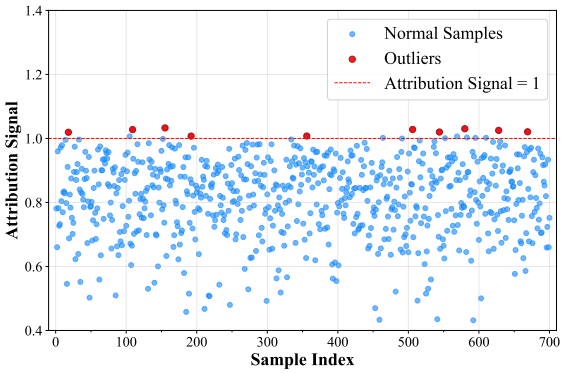
(b) Attribution signal distribution of Wan2.1 [5].



(c) Attribution signal distribution of LTX [1].



(d) Attribution signal distribution of EA [7].



(e) Attribution signal distribution of Wan2.2 [5].

Figure 4. The distribution of attribution signals and outliers for the five models mentioned in this paper.

4. Distribution of Attribution Signal

Due to the variations in model architecture and training data, the distribution of the attribution signal t in the five video generation models discussed in this paper exhibits significant differences. As a result, we determined specific threshold values for each model. By testing on the belonging videos for each model, the distribution of the attribution signal t for each model is shown in Fig. 4. It is observed that the attribution signals for the HYV [3], Wan2.1 [5], EA [7], and Wan2.2 [5] are relatively dispersed, with the majority of the attribution signals in the belonging videos for these four models being less than 1. This is the fundamental reason why the HYV [3], EA [7], and Wan2.2 [5] are capable of achieving zero-shot attribution.

In contrast, the attribution signal distribution for LTX [1] is highly concentrated, roughly symmetrically around 1, which provides an intuitive explanation for its inferior attribution performance compared to the aforementioned models. Furthermore, we observed that each model's belonging videos contain a small number of outliers (represented as red dots in Fig. 4), which can be attributed to the inherent diversity in the generated videos. To address this issue, we employed Kernel Density Estimation [2], a non-parametric method that makes no assumptions about the underlying distribution and is inherently robust to outliers.

Table 2. Corrupted levels under different windows (MSE Loss).

Window	W_0	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
HYV($\times 10^{-4}$)	5.413	6.592	6.574	**6.674**	5.442	-	-	-	-
LTX($\times 10^{-3}$)	1.072	1.081	1.081	**1.087**	1.082	1.084	1.081	1.077	1.060

5. Verification of Theoretical Analysis

Our theoretical analysis (see Section 3.2) aims to "maximize the reconstruction discrepancy," i.e., maximize corruption (can be quantified by the MSE Loss). An intuitive and expedited strategy is to select W_{N-1} as the corrupted window; for the other 4 models in the paper (excluding LTX [1]), W_3 aligns with the analysis ($N = 4$). However, LTX deviates from this intuition due to its unique decoder with a denoising step. The results presented in Tab. 2 show that selecting W_3 causes the greatest corruption in LTX, thus adhering to the corruption-maximizing theory.

References

[1] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2, 3, 4

[2] JooSeuk Kim and Clayton D Scott. Robust Kernel Density Estimation. *The Journal of Machine Learning Research*, 13 (1):2529–2565, 2012. 3

[3] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2, 3

[4] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A Large-Scale High-Quality Dataset For Text-to-Video Generation. In *International Conference on Learning Representations*, 2025. 1, 2

[5] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3

[6] Chao Wang, Zijin Yang, Yaofei Wang, Weiming Zhang, and Kejiang Chen. AEDR: Training-Free AI-Generated Image Attribution via Autoencoder Double-Reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 1

[7] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. EasyAnimate: A High-Performance Long Video Generation Method based on Transformer Architecture. *arXiv preprint arXiv:2405.18991*, 2024. 1, 2, 3