SAGA: Source Attribution of Generative AI Videos

Rohit Kundu[1,2], Vishal Mohanty[2], Hao Xiong[3], Shan Jia[2], Athula Balachandran[2], Amit K. Roy-Chowdhury[1]
[1]University of California, Riverside, [2]YouTube (Google), [3]Google DeepMind
{rohit.kundu@email, amitrc@ece}.ucr.edu; {rohitkun, vishalmohanty, haoxg, shanjia, athula}@google.com

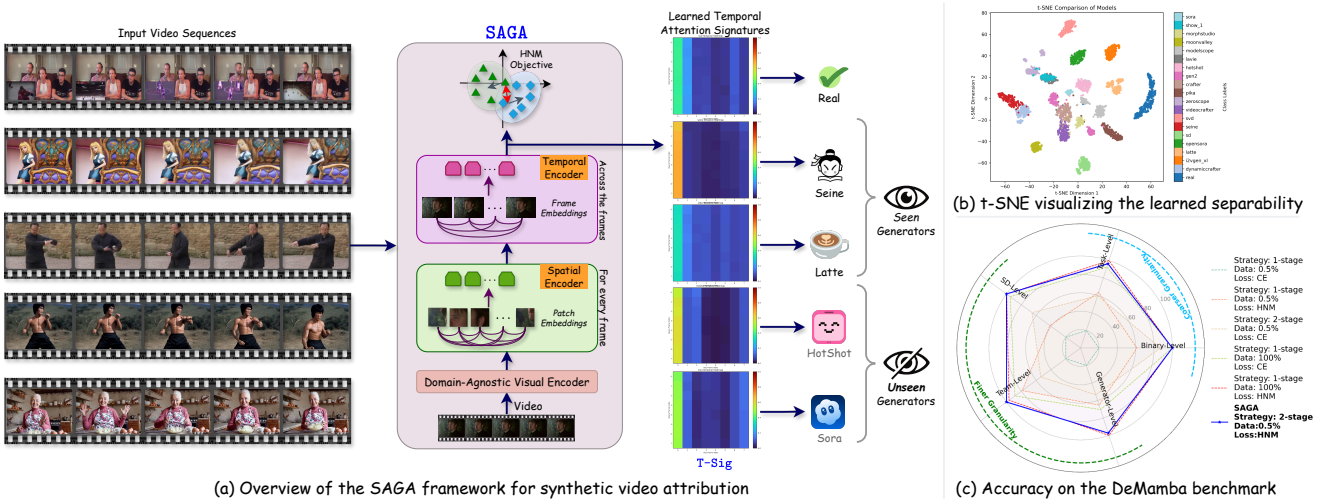


Figure 1. **SAGA: Data-Efficient & Interpretable AI Video Source Attribution.** (a) Temporal Attention Signatures (T-Sigs): SAGA pioneers AI video source attribution. Our novel T-Sigs provide interpretability, showing unique fingerprints for Real, Seen, and even Unseen generators. (b) Feature Separability: t-SNE visualization of learned features demonstrates clear generator clusters. (c) Multi-Granular Performance & Data Efficiency: SAGA excels across 5 attribution levels. Radar chart shows our 2-stage training method using the Hard Negative Mining (HNM) objective, using only 0.5% labeled data, matches fully supervised performance and surpasses baselines.

Abstract

The proliferation of generative AI has led to hyper-realistic synthetic videos, escalating misuse risks and outstripping binary real/fake detectors. We introduce SAGA (Source Attribution of Generative AI videos), the first comprehensive framework to address the urgent need for AI-generated video source attribution at a large scale. Unlike traditional detection, SAGA identifies the specific generative model used. It uniquely provides multi-granular attribution across five levels: authenticity, generation task (e.g., T2V/I2V), model version, development team, and the precise generator, offering far richer forensic insights. Our novel video transformer architecture, leveraging features from a robust vision foundation model, effectively captures spatio-temporal artifacts. Critically, we introduce a data-efficient pretrain-and-attribute strategy, enabling SAGA to achieve state-of-the-art attribution using only 0.5% of source-labeled data per class, matching fully supervised performance. Furthermore, we propose Temporal Attention Signatures (T-Sigs), a novel interpretability method that visualizes learned temporal differences, offering the first explanation for why different video generators are distinguishable. Extensive experiments on public datasets, including cross-domain scenarios, demonstrate that SAGA sets a new benchmark for synthetic video provenance, providing crucial, interpretable insights for forensic and regulatory applications. The project page is https://rohit-kundu.github.io/SAGA.

1. Introduction

The rapid advancement of AI-driven video synthesis, spanning text-to-video (T2V) [43, 49, 59] and image-to-video (I2V) [5, 9, 53] systems, has democratized content creation but also heightened concerns over misuse and misinformation [26, 31], exemplified by incidents like AI-generated wildfire videos causing public alarm [36]. Current defenses largely focus on binary real/fake detection [7, 10, 22, 54]. However, as generative models multiply and evolve at an unprecedented pace [24, 25], merely detecting a video as synthetic is insufficient.

The critical need has **shifted from *whether* it's fake to *what is its source***? Identifying the specific generative model or family, called source attribution [46, 52, 56], is paramount for effective digital forensics [2], intellectual property enforcement [4, 41], and developing robust adversarial countermeasures [32, 60].

Attributing synthetic videos to their source is a far more complex challenge than traditional DeepFake detection or even image source attribution [14, 52, 56]. While image-based methods offer a starting point, they fundamentally fail to address video-specific complexities. We identify three key barriers: (1) **Temporal Dynamics:** Videos possess unique temporal fingerprints and inconsistencies resulting from the generation process, entirely missed by static image analysis. (2) **Increased Model Diversity:** The video generation pipeline involves more diverse architectures and stages (e.g., frame synthesis, motion models), creating a vastly larger and more complex attribution space. (3) **Video Compression:** Unlike image compression, video codecs introduce complex spatio-temporal artifacts that can obscure or destroy subtle generator-specific traces. These challenges necessitate a novel approach designed specifically for the video domain.

To address this significant gap, we introduce SAGA (*Source Attribution of Generative AI videos*), the first large-scale, comprehensive framework dedicated to multi-granular source attribution of AI-generated videos. Moving decisively beyond binary detection, SAGA pinpoints the origin of a synthetic video across five crucial levels (denoted "−L") of granularity *using only 0.5% of the data*: (1) BIN−L (real/synthetic); (2) TASK−L (real vs. T2V vs. I2V); (3) SD−L (differentiating between Stable Diffusion versions e.g., [33, 37]); (4) TEAM−L (attributing to development teams, aiding misuse tracking); and (5) GEN−L (precise model ID). This multi-granular approach is crucial in practice: for example, when two generators are highly similar, an in-the-wild video may yield low-confidence predictions at the GEN−L, but higher confidence at coarser levels such as SD version or team, still providing valuable forensic insight. Furthermore, unlike prior works, SAGA provides *Temporal Attention Signatures* or T−Sig as shown in Fig. 1(a), which ***offer crucial interpretability into why generative models are distinguishable***.

By averaging frame-to-frame attention scores across multiple videos from a common source, we derive unique visual 'fingerprints' (T−Sig) for each generator. To the best of our knowledge, this is the first work to visually explain video attribution performance: T−Sigs highlight the subtle but stable temporal artifacts, such as characteristic motion dynamics or frame-to-frame inconsistencies, that SAGA learns to leverage for fine-grained source identification.

To achieve this, SAGA employs a novel multi-headed attention video transformer to effectively capture temporal inconsistencies. To enhance in-the-wild robustness, we initialize our model with rich visual features from a foundational vision encoder [1], mitigating domain gap issues [6, 8, 28]. Addressing the common scenario of abundant binary labels but scarce multi-class source labels, we propose a pretrain-and-adapt strategy. We first build a strong visual representation by pretraining a binary (real vs. fake) classifier. Subsequently, this base model is efficiently adapted to the multi-class source attribution challenge, utilizing a contrastive objective with hard-negative mining (HNM). Remarkably, this adaptation, allows SAGA to match the performance of a model trained with 100% of the source-labeled data, *even when using only 0.5% of the labeled examples* for the adaptation phase (Fig. 1(c)). This highlights exceptional data efficiency for the complex task of synthetic video source attribution.

Table 1. **Characteristic Comparison:** Unlike prior methods, SAGA performs video source attribution with only 0.5% labeled data, is evaluated on a large corpus of generators from open-source datasets and provides interpretable analyses.

Aspect	Existing Methods	SAGA
Binary Classification	✓	✓
Source Attribution	Generator-level	Multi-tiered
Number of Generators Evaluated	4 ↓	20 ↑
Data-Efficient Training	✗	✓
Intra-data evaluation	✓	✓
Cross-data evaluation	✗	✓
Quantitative Evaluation	✓	✓
Qualitative Analysis	✗	✓

In summary, our main contributions are as follows:

- We pioneer large-scale, AI-generated **video source attribution** by introducing SAGA. This framework moves beyond traditional binary (real/fake) detection, addressing the more complex challenge of identifying the specific origin of synthetic videos. SAGA demonstrates exceptional data efficiency, achieving robust attribution performance using only 0.5% of the labeled source data per class, on par with fully supervised methods.
- We establish the first comprehensive, **multi-granular framework** for video source attribution, encompassing five distinct levels: BIN−L (authenticity), TASK−L (T2V/I2V), SD−L (base model version), TEAM−L (development team origin), and GEN−L (specific model). This hierarchy provides richer, more practical forensic insights than possible with single-level analysis.
- We introduce **Temporal Attention Signatures** (T−Sigs), a novel interpretability method specifically designed for AI-generated video source attribution. Derived from SAGA's learned temporal attention patterns, T−Sigs provide the first visual means in the synthetic video literature to understand why different generators are distinguishable, by exposing their unique, inherent temporal artifacts.
- We conduct extensive evaluations across **19 distinct video generators** from two public datasets (DeMamba [7] and DVF [42]). Our results validate SAGA's effectiveness and robustness in multi-granular source attribution under both in-domain and cross-domain scenarios, setting a new benchmark for this emerging field.

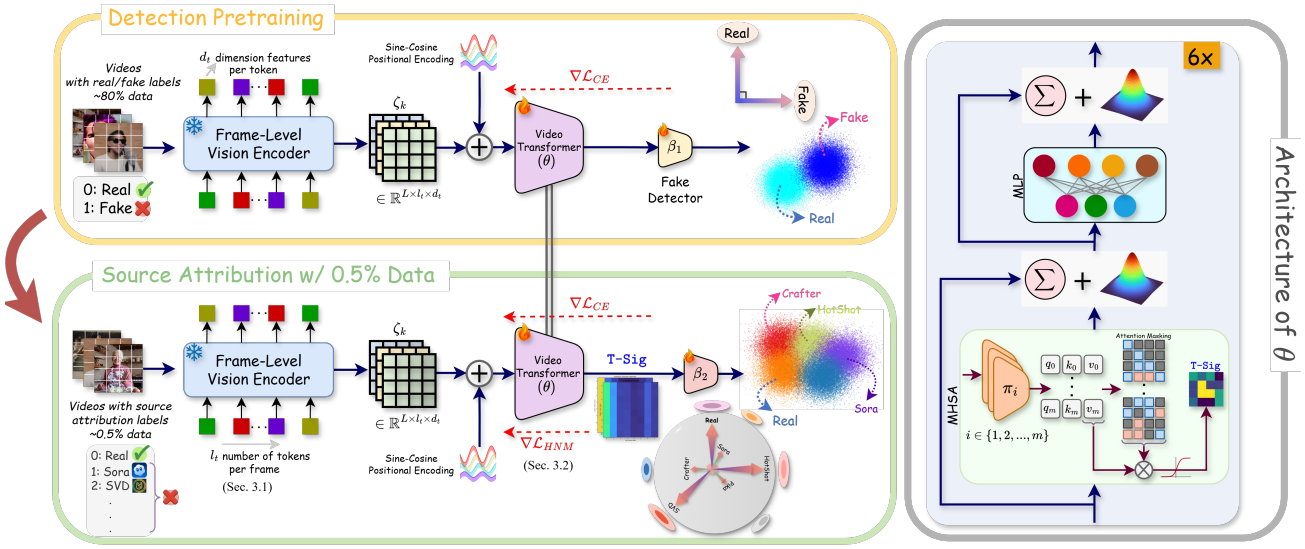


Figure 2. Overall framework of SAGA with a two-stage training approach. In Stage-1, each video x_k with real/fake labels is processed through a frozen foundational vision encoder to extract image-level features z_m, which are stacked in temporal order to form the video representation ζ_k. Positional encoding is added, and the sequence is passed through our video transformer architecture θ (Sec. 3.1) to obtain ϕ_k. The classifier β_1 maps ϕ_k to real or fake classes using a cross-entropy loss ($\mathcal{L}_{CE}$). In Stage-2, the pretrained video transformer is adapted for attribution into n_c classes (n_c defined by the attribution task; see supplementary Sec. S-1) using only 0.5% of source labeled data. Stage-2 incorporates an additional hard negative mining objective ($\mathcal{L}_{\text{HNM}}$, Sec. 3.2) along with $\mathcal{L}_{CE}$ for the attribution task.

2. Related Work

Synthetic Content Detection: Early efforts in synthetic content detection primarily targeted images, a focus driven by the evolution of generative models, from early GAN models [20, 21, 61] to the more recent and powerful diffusion models [35, 37, 38]. The authors in [11, 12] investigated the detection of diffusion-generated images from GAN detectors through spatial and frequency domain analyses, and revealed the distinctive forensic traces left by generation models. Several studies explored diverse features to distinguish realistic diffusion-generated images, including reconstruction errors[51], CLIP-based representations [30, 45], and up-sampling artifacts [44]. However, as demonstrated in [7, 46], image-centric approaches fall short when applied to videos, where the capture of different spatial traces or temporal artifacts is essential for effective detection. Synthetic video detection remains comparatively underexplored but has gained more focus in the recent two years. DeMamba [7] addresses this by employing a structured state space model that continuously scans spatial and temporal zones to capture subtle generative artifacts, enabling robust real/fake classification across diverse video generators and outperforming image-based detectors on video-specific inconsistencies. A large AI-generated video dataset with 19 T2V/I2V generators is also proposed. UNITE [22] tackles face-manipulation and synthetic video detection using a foundation model with a transformer and attention-diversity loss. While these methods advance video authenticity detection, they are primarily confined to binary classification and do not address the more challenging task of source attribution.

Source Model Attribution: Research in source attribution has been largely focused on synthetic images. [15] proposed an open-world discovery and attribution pipeline that iteratively combines out-of-distribution detection, clustering, and supervised refinement, enabling the discovery and attribution of images from both known and unknown GANs in a scalable manner. POSE [57] further advanced this by simulating open-set samples using lightweight augmentation models to better model the imperceptible traces left by unknown generative models. Wang et al. [52] tackled the origin attribution problem from a model-agnostic and alteration-free perspective, proposing a reverse-engineering approach that leverages reconstruction loss: if an image can be more accurately inverted by a given model, it is likely to have been generated by that model. Collectively, these works highlight the shift from simple real/fake detection to fine-grained, open-set, and model-agnostic attribution in images. However, most of these methods are tailored for static images and cannot address the unique spatiotemporal challenges of source attribution in synthetic videos, where temporal consistency and motion artifacts play a critical role.

To the best of our knowledge, the only prior work to attempt source attribution in videos is by Vahdati et al. [46], whose study is limited to only 4 generators with closed-source videos and focuses on only generator-level attribution. In contrast, SAGA provides a far more comprehensive benchmark (as shown in Table 1) on 19 video generators, spanning multiple levels of attribution granularity along with interpretable T-Sig analyses for actionable provenance analysis.

3. Proposed Method

Given a video x_k, the goal is to predict its source label y_k from a set of n_c possible classes. At the binary level (BIN-L), we have $n_c = 2$ and for source attribution, $n_c > 2$. Given a dataset $(x_k, y_k)_{k=1}^{N} \in \mathcal{X}$, the SAGA model learns to map x_k to y_k under the source attribution setting, supporting both binary and fine-grained multi-class source attribution tasks.

Instead of training a n_c-class model from scratch, we introduce a two-stage training protocol that builds the source attribution model on top of a pre-trained binary classifier trained with extensive real/fake data. In stage-1, we pretrain a video transformer model (Sec. 3.1) for binary real vs. fake classification, as they are abundantly available, using only cross-entropy (CE) loss. In stage-2, we perform source attribution the model through a contrastive objective (Sec. 3.2), using only 0.5% of source labeled examples to efficiently adapt to fine-grained attribution. The pretraining is done once, and it acts as the common starting point for all levels of attribution.

3.1. Video Transformer

AI-generated videos inherently exhibit domain gap [6, 8, 28], which is critical to address since the aim of SAGA is to be used in-the-wild. To enhance robustness, we extract potentially domain-agnostic features by leveraging a powerful visual encoder pretrained on web-scale image-text data. Given a video instance $x_k \in \mathcal{X}$, we process each frame g_m (resized to a fixed resolution) using the frozen pretrained encoder. This produces a tokenized embedding $z_m \in \mathbb{R}^{l_t \times d_t}$ for each frame, where $m \in 1,2,...,L$ with L denoting the number of frames per video. The dimension of these embeddings is determined by the chosen encoder, where l_t is the number of tokens per frame and d_t is the token feature dimension. The embeddings for all frames in x_k are concatenated in temporal order, resulting in a video-level representation $\zeta_k \in \mathbb{R}^{L \times l_t \times d_t}$, which serves as input to our trainable video transformer. The resulting set of encoded videos is thus represented as $\mathcal{Z} = \zeta_k \,|\, x_k \in \mathcal{X}$. The SAGA's video-transformer model employs a multi-head self-attention (MHSA) transformer architecture [48] (θ) tailored for video attribution to obtain $\phi_k = \theta(\zeta_k)$. It processes sequences of frame embeddings, effectively capturing temporal dependencies for robust video-level predictions.

Our novel video transformer architecture processes the frame-level token embeddings $\zeta_k \in \mathbb{R}^{L \times l_t \times d_t}$ in a hierarchical manner: first, by applying spatial self-attention within each frame's tokens, and second, by applying temporal self-attention across the frame-level representations.

Spatial Encoder: To capture relationships between spatial patches within individual frames, the input tokens for each frame are initially processed independently. We employ a single standard transformer encoder block (detailed below). This block refines the l_t token embeddings for each of the L frames. The output tokens for each frame are then average pooled across the token dimension, resulting in a single feature vector $\in \mathbb{R}^{d_t}$ for each frame.

Temporal Encoder: The sequence of L frame-level feature vectors is then passed to the Temporal Encoder. Sinusoidal positional encodings are added to these vectors to inject temporal order information. The Temporal Encoder consists of $D = \text{depth} + 1$ stacked standard transformer encoder blocks. Each of these encoder blocks contains:

- A Multi-Head Self-Attention (MHSA) layer with $N_h = 12$ parallel attention heads, using scaled dot-product attention to model inter-frame dependencies.
- Layer Normalization, residual connections, and dropout, to ensure training stability and prevent overfitting.
- A two-layer feed-forward network (MLP) with GELU activation [19] for non-linear transformations.

This stacked architecture allows the model to build progressively complex representations of temporal dynamics and inconsistencies. The Temporal Attention Signatures (T-Sigs) are extracted from the attention scores of the MHSA layer in the penultimate block encoder block of this Temporal Encoder. During inference, the attention scores over several videos are extracted and normalized to produce T-Sigs. These attention scores highlight which frames the model attends to when processing the sequence, revealing patterns characteristic of the video's source.

3.2. Contrastive Objective

With a pre-trained binary classifier as the foundation, Stage-2 adapts the model for multi-class source attribution. To address the limited availability of fine-grained labeled data, we incorporate a contrastive loss with hard negative mining (HNM), enabling effective attribution even with a small number of samples per generator, since CE-loss alone proved to be suboptimal in this scenario (Table 7, Fig. 5). Given an anchor embedding $\mathbf{a}$, a positive embedding $\mathbf{p}$ (same class), and a negative embedding $\mathbf{n}$ (different class), the triplet loss encourages the following margin constraint:

$$\|\mathbf{a}-\mathbf{p}\|_2^2 + \alpha < \|\mathbf{a}-\mathbf{n}\|_2^2, \tag{1}$$

where $\alpha > 0$ is a margin hyperparameter. The loss is defined as:

$$\mathcal{L}_{\text{triplet}} = \max\big(0, \|\mathbf{a}-\mathbf{p}\|_2^2 - \|\mathbf{a}-\mathbf{n}\|_2^2 + \alpha\big). \tag{2}$$

Semi-hard negatives, which are most commonly used in the literature [18, 39, 55], are those that are further from the anchor than the positive, but within the margin as,

$$\|\mathbf{a}-\mathbf{p}\|_2^2 < \|\mathbf{a}-\mathbf{n}\|_2^2 < \|\mathbf{a}-\mathbf{p}\|_2^2 + \alpha. \tag{3}$$

Thus, for each anchor-positive pair, the negative $\mathbf{n}$ is selected such that for a batch $\mathcal{B}$:

$$\mathcal{L}_{\text{semi-HNM}} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} \max\Bigg(0, \|\mathbf{a}_i-\mathbf{p}_i\|_2^2 - \min_{\substack{j \\ y_j \neq y_i \\ \|\mathbf{a}_i-\mathbf{p}_i\|_2^2 < \|\mathbf{a}_i-\mathbf{n}_j\|_2^2 \\ < \|\mathbf{a}_i-\mathbf{p}_i\|_2^2 + \alpha}} \|\mathbf{a}_i-\mathbf{n}_j\|_2^2 + \alpha\Bigg) \tag{4}$$

Hard negatives are those that are closer to the anchor than the positive that is, $\|\mathbf{a}-\mathbf{n}\|_2^2 < \|\mathbf{a}-\mathbf{p}\|_2^2$. Thus, the HNM loss is:

$$\mathcal{L}_{\text{HNM}} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\max\left(0, \|\mathbf{a}_i-\mathbf{p}_i\|_2^2 - \min_{\substack{j \\ y_j \neq y_i}} \|\mathbf{a}_i-\mathbf{n}_j\|_2^2 + \alpha\right). \quad (5)$$

This focuses the model on the most challenging negatives within the batch. In our source attribution task, some generators produced embeddings with overlapping t-SNE clusters when trained with CE-loss alone (Fig. 5). This is because CE-loss maximizes class separation in logit space but does not enforce geometric separation in the embedding space. Semi-hard negative mining selects negatives that satisfy Eq. 3 , i.e., they are farther than the positive but still within the margin and therefore yield non-zero loss; by contrast, easy negatives satisfy $\|\mathbf{a}-\mathbf{n}\|_2^2 > \|\mathbf{a}-\mathbf{p}\|_2^2+\alpha$ and contribute zero loss, while hard negatives satisfy $\|\mathbf{a}-\mathbf{n}\|_2^2 \leq \|\mathbf{a}-\mathbf{p}\|_2^2$. In heavily overlapping clusters many negatives are hard rather than semi-hard, and semi-hard mining omits them (Fig. 3), which limits the gradient signal needed to separate overlapping modes.

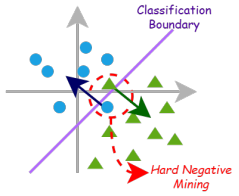


Figure 3. HNM enables better separation boundaries between classes while semi-HNM will exclude these samples from the loss.

HNM, on the other hand, always selects the most difficult negative within the batch:

$$\mathbf{n}_{\text{hard}} = \arg\min_{\substack{j \\ y_j \neq y_i}} \|\mathbf{a}_i-\mathbf{n}_j\|_2^2, \quad (6)$$

and the gradient is given by:

$$\nabla_\theta \mathcal{L}_{\text{HNM}} \propto 2(\mathbf{a}-\mathbf{n}_{\text{hard}}) - 2(\mathbf{a}-\mathbf{p}). \quad (7)$$

This mechanism directly pushes the anchor away from the nearest negative, forcing separation even when clusters overlap. Let $\mathcal{S}_c$ denote the embedding manifold for class c. For overlapping classes c and c', hard mining minimizes,

$$\min_\theta \mathbb{E}_{(\mathbf{a},\mathbf{p})\sim\mathcal{S}_c}\left[\max_{\mathbf{n}\sim\mathcal{S}_{c'}}\left(\|\mathbf{a}-\mathbf{p}\|_2^2 - \|\mathbf{a}-\mathbf{n}\|_2^2+\alpha\right)\right], \quad (8)$$

which is equivalent to maximizing the minimum inter-class margin α. This is particularly important when the intra-class variance σ^2_{intra} is comparable to or exceeds the difference between the margin and inter-class variance σ^2_{inter} as, $\sigma^2_{\text{intra}} \geq \alpha - \sigma^2_{\text{inter}}$. Thus, along with the CE-loss ($\mathcal{L}_{CE}$), the final loss function to train SAGA becomes $\lambda\cdot\mathcal{L}_{CE} + (1-\lambda)\cdot\mathcal{L}_{HNM}$.

In our experiments, CE-loss with semi-HNM was insufficient to separate overlapping generators, resulting in a mean accuracy of 70.31% on the GEN-L task (see t-SNE visualizations in Fig. 5 and quantitative results in the supplementary). In contrast, incorporating HNM with CE-loss markedly improved performance, achieving a mean accuracy of 94.99% by effectively enforcing separation between samples from different generator classes ($\mathcal{S}c$ vs. $\mathcal{S}c'$).

Table 2. TASK-L attribution performance (Accuracy) under different settings. SAGA performs almost perfectly, compared to the **100% data** setting.

Setting	Real	T2V	I2V	Overall
Strategy: 1-stage Data: 0.5%	99.20%	97.41%	66.20%	82.41%
Strategy: 1-stage Data: 100%	99.97%	99.93%	99.97%	**99.96%**
Strategy: 2-stage Data: 0.5% ***(Our Setting)***	99.79%	99.32%	91.12%	**98.20%**

Table 3. SD-L attribution performance under different settings. SAGA performs marginally better than the **100% data** setting.

Setting	Real	SD 1.4	SD 1.5	SD 2.1	SDXL	Overall
Strategy: 1-stage Data: 0.5%	99.84%	0.00%	0.00%	99.75%	99.28%	59.77%
Strategy: 1-stage Data: 100%	99.99%	99.90%	99.99%	99.80%	92.09%	**98.35%**
Strategy: 2-stage Data: 0.5% ***(Our Setting)***	99.95%	97.02%	98.15%	99.14%	98.20%	**98.49%**

Table 4. Evaluation of SAGA on BIN-L task under in-domain and various cross-domain settings. The results demonstrate the strong generalization and robustness of SAGA for authenticity verification, even on unseen generators.

Trained on	Tested on	Accuracy	Precision	Recall
All DeMamba generators (80% data in training)	All DeMamba generators (20% unseen data in evaluation)	99.94%	100.00%	99.89%
DeMamba *train* set generators (10 generators)	DeMamba *val* set generators (9 generators)	99.86%	100.00%	99.72%
T2V Generators (12 generators)	I2V Generators (4 generators)	99.98%	99.98%	99.98%
SD 2.1 generators (6 generators)	All remaining generators with known SD backbones (5 generators)	99.94%	99.90%	99.96%
Generators from: Alibaba Group, Stability AI, Tencent AI Lab, and Pika AI (8 Generators)	All remaining generators (11 generators)	99.16%	99.98%	98.41%

4. Experiments

Datasets: The training is performed on the DeMamba [7] dataset, with 19 different AI video generators and 1M real videos. More details can be found in the supplementary material, including how we define different attribution levels. Additionally, we use the DVF dataset [42] covering 8 video generators for cross-data evaluations. Implementation details are provided in the supplementary.

Real/Fake Detection Results: We first comprehensively evaluate the performance of SAGA in binary classification in Table 4, conducting both in-domain and cross-generator evaluations. For cross-generator analysis, we train on the *train* split of DeMamba and evaluate on the *val* split, as well as train on generators from a specific team or SD version backbone or generation task and test on the remaining generators. The results demonstrate that, for the BIN-L task, SAGA achieves robust authenticity verification

Table 5. SOTA and SAGA comparison on DeMamba [7] for in-domain and DVF [42] for cross-domain evaluation. **Best** and second-best performances are highlighted.

Dataset	Method	Precision	Recall	Accuracy
DeMamba [7]	TALL [54]	87.91%	88.52%	88.42%
	F3Net [34]	88.73%	81.88%	86.04%
	NPR [44]	82.45%	84.08%	83.45%
	STIL [16]	87.12%	82.22%	85.35%
	MINTIME-CLIP-B [7]	91.55%	87.62%	89.98%
	FTCN-CLIP-B [7]	92.21%	86.18%	89.67%
	CLIP-B-PT [7]	44.83%	81.74%	41.82%
	DeMamba-CLIP-PT [7]	79.97%	78.86%	79.98%
	XCLIP-B-PT [7]	61.29%	81.93%	65.83%
	DeMamba-XCLIP-PT [7]	76.38%	83.59%	79.31%
	XCLIP-B-FT [7]	86.77%	84.41%	86.07%
	SAGA (BIN-L)	**100.00%**	**99.89%**	**99.94%**
DVF [42]	CNNDet [50]	-	-	78.20%
	DIRE [51]	-	-	62.10%
	Raising [13]	-	-	67.00%
	UNI-FD [30]	-	-	74.10%
	F3Net [34]	-	-	81.30%
	ViViT [3]	-	-	79.10%
	TALL [54]	-	-	69.50%
	TS2-Net [27]	-	-	72.10%
	DE-FAKE [40]	-	-	72.10%
	HifiNet [17]	-	-	84.30%
	DVF [42]	-	-	92.00%
	SAGA (BIN-L)	**99.35%**	**96.14%**	**95.39%**

Table 6. TEAM-L performances (Accuracy) under different settings. SAGA performs better than the **100% data** setting on average. The proposed 2-stage training significantly improves performance on certain teams as highlighted.

Team	Strategy: 1-stage Data: 0.5%	Strategy: 1-stage Data: 100%	Strategy: 2-stage Data: 0.5% *(Our Setting)*
Real	99.54%	99.95%	98.86%
Alibaba Group	97.57%	99.80%	97.63%
Hotshot Co.	94.96%	95.68%	98.56%
HPC AI Tech	92.99%	99.95%	96.85%
MoonValley	98.52%	100.00%	100.00%
MorphStudio	88.27%	77.78%	83.95%
OpenAI	13.33%	80.00%	66.67%
Personal: Sterling	60.96%	81.96%	95.80%
Pika	88.81%	99.76%	95.84%
Runway ML	74.62%	90.53%	92.42%
Shanghai AI Lab-1	95.98%	99.91%	97.93%
Shanghai AI Lab-2	0.00%	99.99%	96.26%
Show Lab	98.10%	100.00%	98.10%
Stability AI	91.20%	99.98%	98.65%
Tencent AI Lab	84.24%	98.76%	93.19%
Overall	80.55%	**94.94%**	**97.77%**

across diverse and previously unseen data sources, exhibiting strong generalization and minimal sensitivity to domain shifts. To rigorously benchmark SAGA against existing state-of-the-art (SOTA) binary detectors, we conducted evaluations on both the DeMamba [7] and DVF [42] datasets, as presented in Table 5. Notably, all competing SOTA methods on the DVF dataset [42] are trained and tested within the DVF dataset, following standard in-domain protocols. In contrast, SAGA is trained exclusively on DeMamba and evaluated directly on DVF, constituting a challenging cross-dataset generalization scenario. Despite this, SAGA significantly outperforms the SOTA baselines, highlighting its superior robustness and generalizability for the BIN-L task across diverse data distributions.

Fine-grained Attribution Results: We further evaluate SAGA under multi-granular source attribution tasks across three training regimes: (1) 1-stage training with only 0.5% labeled data, (2) 1-stage training with 100% data, and (3) our proposed 2-stage training framework using 0.5% labeled data for all attribution levels (attribution description in the supplementary).

On TASK-L task (Table 2), SAGA achieves strong performance in distinguishing real, T2V, and I2V videos. This exposes the distinctive patterns left by T2V and I2V generation methods. With full data, the model nearly saturates accuracy across all classes (99.96% overall). In the low-data setting, performance drops substantially for I2V (66.20%), indicating limited data hinders generalization to this class. Our two-stage training framework substantially mitigates this drop, boosting I2V accuracy to 91.12% and overall accuracy to 98.20%.

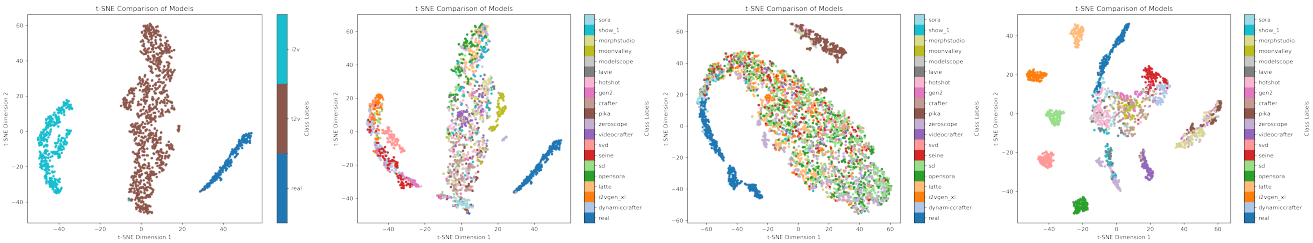
Evaluation on the SD-L task in Table 3 demonstrates that full-data training yields high accuracy across all SD versions (98.35% overall). However, in the low-data regime, the model struggles to distinguish SD 1.4 and SD 1.5 (both 0%), while maintaining high accuracy for real, SD 2.1, and SDXL. The two-stage approach closes this gap, achieving over 97% for all SD versions. On the TEAM-L task (Table 6), the model achieves high accuracy for most teams with full data (94.94% overall). In the low-data setting, performance varies widely across teams, with some (e.g., Shanghai AI Lab-2, OpenAI) at or near 0%. The two-stage training strategy dramatically improves robustness, yielding 97.77% overall and consistently high accuracy across almost all teams.

Table 7 presents results for the most challenging setting: GEN-L attribution. Using only cross-entropy loss in the low-data regime, the model performs poorly (24.55% overall), but adding a hard negative contrastive loss boosts accuracy to 65.80%. Our two-stage framework with HNM achieves 94.99% overall, a substantial improvement over single-stage approaches. With full data, the model achieves up to 97.41% accuracy, highlighting the benefit of both data scale and contrastive learning for fine-grained attribution.

t-SNE Analysis: To further interpret the representations learned by SAGA, we conduct a t-SNE [47] and T-Sig analyses of the feature embeddings produced by the model under different attribution settings. Specifically, we visualize the embedding outputs of the last (6^{th}) encoder of θ for videos from the validation set when the model is trained for all attribution levels. First, Fig. 4 (a) and (b) with TASK-L attribution results show that embeddings for real, T2V, and I2V samples form clearly separable clusters, demonstrating effective discrimination among these broad categories. However, when the same embeddings are colored by generator, substantial overlap is observed among most generators, with only a few, such as MorphStudio [43] and SVD [5], forming distinct clusters. This suggests that while SAGA is highly effective at coarse-grained attribution, it does not inherently separate individual generators at this level. Fig. 4 (c) shows the t-SNE plot for the SAGA model trained for BIN-L classification. Here, all generators except Pika [23] collapse

Table 7. GEN-L classification results (Accuracy) with different settings of the SAGA framework. SAGA is able to achieve results close to the **100%** setting, by only using 0.5% of source labeled data (100% data setting has ∼1.6M training data). In many cases (as highlighted) the performance is close to 0.00% for certain difficult generators, but the $\mathcal{L}_{HNM}$ objective has been able to mitigate these missed detections even while using a small fraction of the data. especially while using the proposed 2-stage training.

Generators	**Strategy: 1-stage Data: 0.5%**		**Strategy: 2-stage Data: 0.5%**		**Strategy: 1-stage Data: 100%**	
	$\mathcal{L}_{CE}$ Only	$\mathcal{L}_{CE}+\mathcal{L}_{HNM}$	$\mathcal{L}_{CE}$ Only	$\mathcal{L}_{CE}+\mathcal{L}_{HNM}$ ***(Our Setting)***	$\mathcal{L}_{CE}$ Only	$\mathcal{L}_{CE}+\mathcal{L}_{HNM}$
Real	0.00%	99.04%	98.12%	99.95%	99.91%	99.21%
DynamiCrafter	0.00%	0.00%	0.20%	56.64%	99.53%	75.58%
I2VGen-XL	0.24%	0.00%	9.19%	96.87%	95.13%	96.84%
Latte	0.00%	0.00%	36.89%	98.33%	99.30%	97.84%
OpenSora	98.60%	0.00%	1.23%	91.00%	99.82%	96.70%
SD	99.82%	97.10%	3.36%	91.28%	98.86%	97.76%
SEINE	0.00%	95.47%	1.48%	89.25%	70.79%	95.95%
SVD	0.00%	0.00%	0.06%	91.45%	99.63%	96.68%
VideoCrafter	0.00%	91.25%	27.23%	92.27%	62.71%	95.23%
ZeroScope	0.00%	0.00%	68.11%	91.50%	99.91%	95.95%
Pika	99.58%	93.54%	15.01%	89.99%	98.44%	93.43%
Crafter	0.00%	0.00%	78.79%	79.80%	85.52%	94.95%
Gen2	0.38%	0.00%	69.32%	84.85%	11.36%	78.79%
HotShot	0.00%	92.09%	94.96%	96.40%	42.45%	95.68%
Lavie	0.00%	55.94%	79.37%	81.82%	3.85%	92.66%
ModelScope	0.00%	0.00%	87.59%	97.81%	8.76%	97.08%
MoonValley	0.00%	97.78%	97.78%	99.26%	97.78%	100.00%
MorphStudio	0.00%	14.81%	90.12%	81.48%	0.00%	82.10%
Show_1	0.00%	0.00%	92.38%	98.10%	45.71%	100.00%
Sora	0.00%	0.00%	93.33%	73.33%	66.67%	60.00%
Overall	24.55%	65.80%	55.13%	**94.99%**	70.51%	**97.41%**

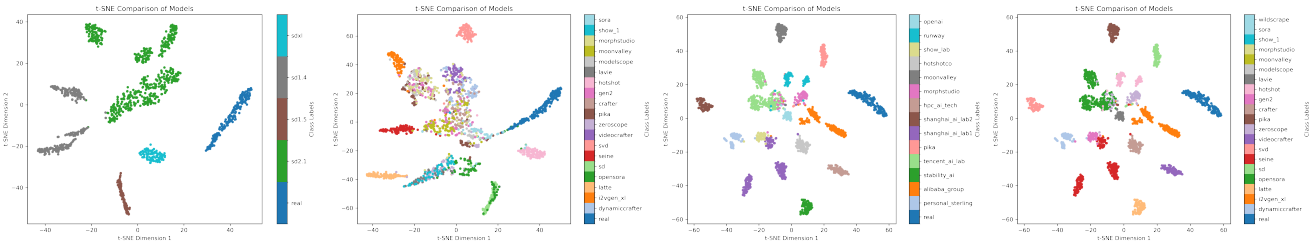


(a) SAGA (TASK-L) on different tasks

(b) SAGA (TASK-L) on different generators

(c) SAGA (BIN-L) on different generators

(d) SAGA (10-class) on all generators



(e) SAGA (SD-L) on different SD versions

(f) SAGA (SD-L) on different generators

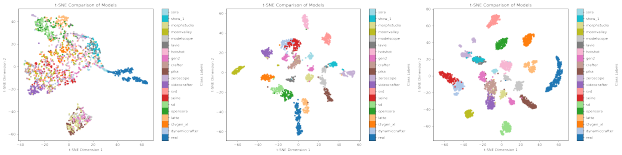
(g) SAGA (TEAM-L) on different teams

(h) SAGA (TEAM-L) on different generators

Figure 4. t-SNE visualization of SAGA's learned representations trained on the TASK-L, BIN-L, SD-L and TEAM-L attribution tasks, respectively. Even when supervised at coarser levels, SAGA distinctly clusters individual generators, revealing strong fine-grained discriminative ability.

into a single "fake" cluster, indicating that the model learns to aggregate all synthetic sources together for the binary task, with minimal separation among generators.

Fig. 4 (d) visualizes embeddings from a SAGA model trained on 10 specific generators from the DeMamba *train* set. The seen generators form distinct clusters, and notably, several unseen generators such as Hotshot [29], Show_1 [58], and MorphStudio [43] also appear as separable clusters. This indicates that the model, even trained on a subset of generators, can recognize distributional differences and cluster unseen sources, highlighting its potential for generalization in open-set scenarios. For the SD-L and TEAM-L models, t-SNE projections (Fig. 4 (e) - (h)) reveal that the learned representations not only cluster according to the supervised SD version or team labels, but also



(a) CE-loss only (b) Semi-HNM (c) HNM

Figure 5. t-SNE visualization of SAGA on the GEN-L attribution task with different loss functions.

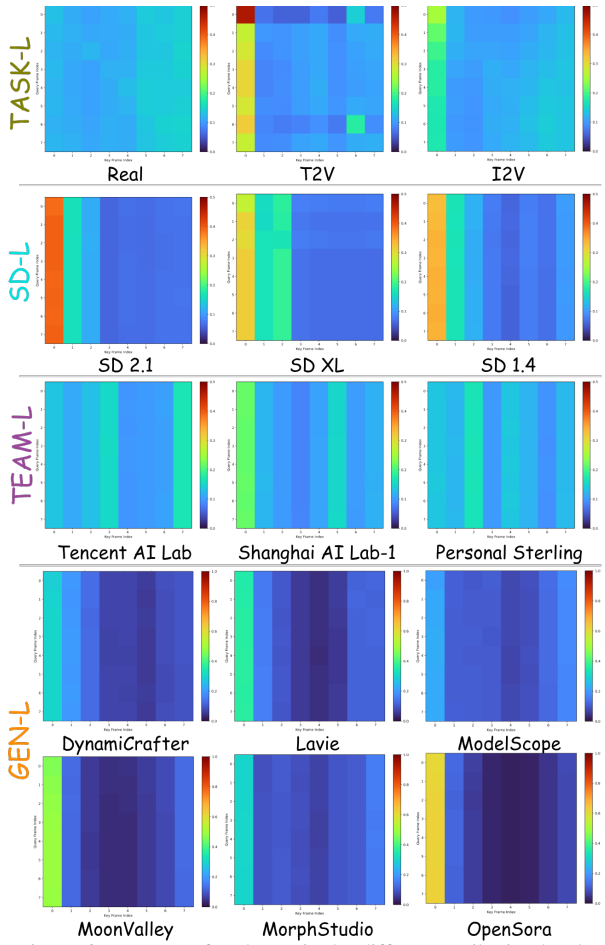


Figure 6. T-Sigs for classes in the different attribution levels.

often separate individual generators within each group. This indicates that SAGA captures fine-grained differences between generators, even when supervision is provided only at a coarser level. This level of separation indicates that the model is sensitive to subtle distributional differences introduced by specific generator architectures or research teams, enabling it to infer whether an unknown generator shares an SD backbone or team affiliation, or represents a completely novel source.

The t-SNE analysis for the GEN-L attribution task in Fig. 5 highlights the impact of different loss functions on SAGA's ability to learn discriminative embeddings. It demonstrates that HNM is highly effective in enforcing discriminative representations for the most fine-grained generator attribution.

T-Sigs Analysis: T-Sigs reveal how SAGA uses temporal cues to distinguish AI-video sources. These signatures, visualized in Fig. 6, are derived by averaging frame-to-frame attention across a large number of videos per class.

Stable and unique T-Sigs emerge for each class in Fig. 6. Despite content variations, videos from the same source yield consistent signatures, indicating shared temporal artifacts. Crucially, these signatures are visually distinct between classes validating SAGA's ability to capture and differentiate based on class-specific temporal inconsistencies.

As shown in Fig. 1(a), even completely unseen generators produce unique and discernible T-Sigs, distinct from training classes and each other. This suggests SAGA learns fundamental temporal characteristics of synthetic generation, beyond just memorizing training patterns. The ability to produce novel signatures for unknown sources indicates strong potential for open-set recognition, allowing SAGA to flag content from new generators, which is vital for real-world deployment. In essence, T-Sigs demonstrate that SAGA keys in on subtle yet consistent generator-specific temporal fingerprints, enabling accurate and interpretable source attribution.

Ablation Results: We evaluate the performance of the SAGA framework on the GEN-L attribution task using different loss functions, including CE-loss and contrastive objectives with semi-hard and hard negative mining strategies. Fig. 5 (quantitative results in the supplementary) show the superiority of the HNM loss for GEN-L source attribution tasks. Table 7 further presents quantitative comparisons under various training regimes: single-stage training with 0.5% labeled samples, single-stage training with 80% of the dataset, and our proposed two-stage training framework utilizing 0.5% labeled data. Across all settings, the contrastive objective with HNM consistently surpasses the CE loss baseline, demonstrating its effectiveness for fine-grained generator-level attribution. We also evaluate the effect of varying the number of samples in second-stage training on source attribution performance, finding that higher sample counts lead to improved results. More details are provided in the supplementary.

5. Conclusion

We introduced SAGA, the first comprehensive framework designed for the critical task of multi-granular source attribution of AI-generated videos, moving beyond inadequate binary detection. By combining a novel video transformer with features from a vision foundation model and a data-efficient two-stage contrastive training strategy, SAGA achieves state-of-the-art performance across five attribution levels, from binary to fine-grained generator ID, even with only 0.5% labeled data and in cross-dataset setups. Our introduction of Temporal Attention Signatures (T-Sigs) provides novel interpretability, visually explaining why generators are distinguishable. SAGA establishes a robust benchmark for AI video provenance, offering crucial capabilities for digital forensics and the responsible governance of generative AI.

Acknowledgements: This work was supported in part by funding from YouTube (Google LLC).

References

[1] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[2] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Tania Sari Bonaventura, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, et al. Deepfake media forensics: State of the art and challenges ahead. *arXiv preprint arXiv:2408.00388*, 2024. 2

[3] 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

[4] Rosa Maria Ballardini, Kan He, and Teemu Roos. Ai-generated content: authorship and inventorship in the age of artificial intelligence. In *Online Distribution of Content in the EU*, pages 117–135. Edward Elgar Publishing, 2019. 2

[5] 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. 1, 6

[6] Baoying Chen and Shunquan Tan. Featuretransfer: Unsupervised domain adaptation for cross-domain deepfake detection. *Security and Communication Networks*, 2021(1):9942754, 2021. 2, 4

[7] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024. 1, 2, 3, 5, 6

[8] Rui Chen, Lei Sun, Jing Tang, Geng Li, and Xiangxiang Chu. Finger: Content aware fine-grained evaluation with reasoning for ai-generated videos. *arXiv preprint arXiv:2504.10358*, 2025. 2, 4

[9] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[10] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *arXiv preprint arXiv:2408.17052*, 2024. 1

[11] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 973–982, 2023. 3

[12] 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

[13] 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

[14] Shengbang Fang, Tai D Nguyen, and Matthew C Stamm. Open set synthetic image source attribution. *arXiv preprint arXiv:2308.11557*, 2023. 2

[15] Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14094–14103, 2021. 3

[16] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021. 6

[17] 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. 6

[18] Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2821–2829, 2017. 4

[19] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[22] Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 3

[23] Pika Labs. Pika art. `https://pika.art/`, 2022. 6

[24] Hannah Lee, Changyeon Lee, Kevin Farhat, Lin Qiu, Steve Geluso, Aerin Kim, and Oren Etzioni. The tug-of-war between deepfake generation and detection. *arXiv preprint arXiv:2407.06174*, 2024. 1

[25] Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. Fakebench: Probing explainable fake image detection via large multimodal models. *arXiv preprint arXiv:2404.13306*, 2024. 1

[26] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025. 1

[27] 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

[28] Qingxuan Lv, Yuezun Li, Junyu Dong, Sheng Chen, Hui Yu, Huiyu Zhou, and Shu Zhang. Domainforensics: Exposing face forgery across domains via bi-directional adaptation. *IEEE Transactions on Information Forensics and Security*, 2024. 2, 4

[29] John Mullan, Duncan Crawbuck, and Aakash Sastry. Hotshot-xl. https://github.com/hotshotco/hotshot-xl, 2023. 7

[30] 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. 3, 6

[31] OpenAI. Sora by openai. https://openai.com/sora/, 2024. 1

[32] Ben Pinhasov, Raz Lapid, Rony Ohayon, Moshe Sipper, and Yehudit Aperstein. Xai-based detection of adversarial attacks on deepfake detectors. *arXiv preprint arXiv:2403.02955*, 2024. 2

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[34] 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

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[36] Reuters. Ai-generated video purports to show apocalyptic scenes of los angeles wildfires, 2025. 1

[37] 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. 2, 3

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022. 3

[39] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4

[40] 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

[41] Jan Smits and Tijn Borghuis. Generative ai and intellectual property rights. In *Law and artificial intelligence: Regulating AI and applying AI in legal practice*, pages 323–344. Springer, 2022. 2

[42] 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. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 5, 6

[43] Morph Studio. Morph studio. https://www.morphstudio.com/, 2024. 1, 6, 7

[44] 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. 3, 6

[45] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7184–7192, 2025. 3

[46] 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 Conference on Computer Vision and Pattern Recognition*, pages 4397–4408, 2024. 2, 3

[47] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[48] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, Ł Kaiser, and I Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[49] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1

[50] 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. 6

[51] 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

[52] Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Where did i come from? origin attribution of ai-generated images. *Advances in neural information processing systems*, 36: 74478–74500, 2023. 2, 3

[53] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 1

[54] 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. 1, 6

[55] Hong Xuan, Abby Stylianou, and Robert Pless. Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2474–2482, 2020. 4

[56] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4662–4670, 2022. 2

[57] Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15856–15865, 2023. 3

[58] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024. 7

[59] 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, 2024. 1

[60] Shuai Zhou, Chi Liu, Dayong Ye, Tianqing Zhu, Wanlei Zhou, and Philip S Yu. Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Computing Surveys*, 55(8):1–39, 2022. 2

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3