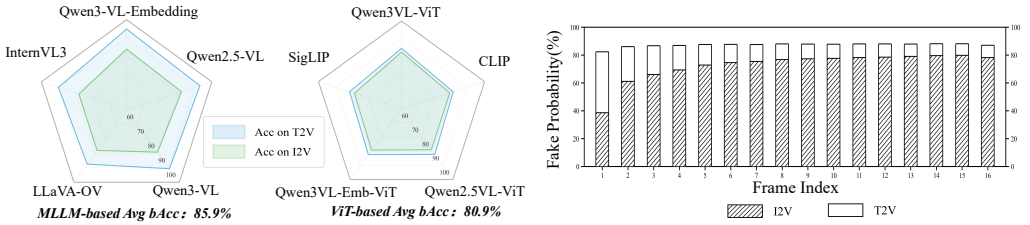
# ATTENTION MEETS ARTIFACTS: BI-LEVEL TEMPORAL RE-ALIGNMENT FOR AIGV DETECTION

**Anonymous authors**
Paper under double-blind review




(a) Cross-paradigm performance comparison between different backbones.

(b) Frame-wise fake probability.

Figure 1: **Key observations across video-generation paradigms.** (a) MLLM-based detectors outperform ViT-based ones, yet show pronounced cross-paradigm imbalance, especially on I2V. (b) Frame-level fake probabilities stay uniform for T2V but increase toward later frames for I2V.

## ABSTRACT

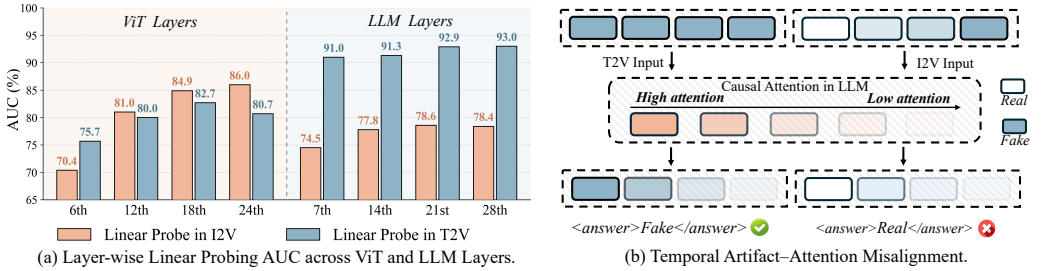
Recent advances in video generation have raised growing security concerns, making AI-generated video (AIGV) detection increasingly essential. Existing methods predominantly rely on Multi-modal Large Language Models (MLLMs), but overlook the discrepancies among samples produced by different generation paradigms. However, we empirically find that MLLM-based detectors exhibit pronounced performance imbalance across different generation paradigms, with a notable weakness on image-to-video (I2V) samples. We attribute this to **Temporal Artifact–Attention Misalignment**. In particular, unlike entirely synthesized T2V samples, I2V samples are conditioned on a real initial image, and thus the forgery artifacts become progressively more pronounced toward later frames. The LLM decoder, however, attends predominantly to earlier frames owing to its causal attention mechanism; as a result, the most discriminative forgery cues in later frames are left largely unattended. To accommodate diverse generation paradigms, we propose a joint optimization strategy termed **Bi-level Temporal Re-Alignment (BiTRA)** that operates at both the extrinsic and intrinsic levels. Specifically, at the extrinsic level, we introduce *"Selective Hierarchical Artifact Injection"*, an adaptive query mechanism that establishes an additional cross-attention pathway for global forgery information access, leveraging a temporal-artifact prior for discriminative information selection. At the intrinsic level, *"Temporal Attention Smoother"* is guided by the Kullback-Leibler margin loss to mitigate the excessive attention bias inherent in the LLM decoder. Extensive experiments demonstrate the effectiveness of our designs on all generation paradigms and confirm our superiority over existing state-of-the-art methods.

## 1 INTRODUCTION

Video generation (Singer et al., 2022; Ho et al., 2022; Blattmann et al., 2023; Xing et al., 2024) has rapidly evolved, enabling synthetic videos with higher visual fidelity. While these developments greatly expand the capabilities of AI-generated video (AIGV), they also pose growing challenges to digital media forensics, making reliable AIGV detection increasingly essential. Previous AIGV detection methods focus on exploring spatial and temporal cues, which can be divided into discriminative representation learning (Battocchio et al., 2025; Kundu et al., 2025; Chen et al., 2026; Li et al., 2026e; Feng et al., 2026) and explicit forensic prior modeling (Zheng et al., 2025; Ben Hayun et al., 2026; Corvi et al., 2026; Internò et al., 2026; Zhang et al., 2026). Recent advances in large-scale multimodal pretraining (Alayrac et al., 2022; Lin et al., 2024) have established Multimodal Large Language Models (MLLMs) as a promising framework for AIGV detection (Tan et al., 2026). Leveraging the strong multimodal perception and reasoning capabilities, recent approaches enhance



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




(a) Layer-wise Linear Probing AUC across ViT and LLM Layers.

(b) Temporal Artifact–Attention Misalignment.

Figure 2: **Temporal Artifact-Attention Misalignment.** (a) Layer-wise linear probing on the frozen Qwen3-VL-Embedding-2B: I2V forgery evidence is strengthened within the ViT but sharply weakened in the LLM, whereas T2V evidence is further enhanced. (b) Under native causal attention, the prediction token is biased toward early frames, misaligned with the later-emerging I2V artifacts.

forgery discrimination through fine-grained artifact supervision (Li et al., 2026d), artifact-aware reward design (Park et al., 2026), and evidence-driven agentic reasoning (Li et al., 2026b). Despite these advances, they still largely treat videos from different generation paradigms uniformly, overlooking the intrinsic differences between their generation processes. Our empirical analysis further reveals that, while MLLM-based detectors outperform ViT-based detectors overall, they exhibit a pronounced cross-paradigm imbalance, with notably weaker performance on image-to-video (I2V).

To uncover the origin of this imbalance, we analyze it from two complementary perspectives: (i) how forgery artifacts differ across generation paradigms, and (ii) how MLLMs process these paradigm-dependent artifacts. First, we apply representative SOTA AIGI detectors, including Effort (Yan et al., 2024) and DDA (Chen et al., 2025), to each frame of generated videos, revealing similar temporal patterns across detectors and among generators within the same generation paradigm (Fig. 6 in the appendix). Concretely, as shown in Fig. 1(b), frame-wise fake probabilities remain relatively uniform over frames in T2V videos, whereas in I2V videos they start markedly lower and progressively increase over subsequent frames. This difference stems from their distinct conditioning mechanisms: T2V synthesizes all frames solely from a text prompt, whereas I2V is additionally anchored to a real initial frame, whose realistic visual cues dominate the early frames but gradually fade as generation proceeds, making later frames increasingly deviate from real-image statistics and thus more detectable. Second, applying layer-wise linear probing to the internal representations of the MLLM, we observe that detection performance remains relatively flat across the layers of the visual encoder, with little disparity between the two paradigms. However, as the representations propagate through the LLM decoder, the performance on I2V videos degrades substantially, whereas that on T2V videos improves markedly, thereby further exacerbating the cross-paradigm imbalance, as shown in Fig. 2(a). Our further analysis reveals that the LLM decoder exhibits a pronounced attention bias toward early-frame visual tokens, assigning substantially less attention to later frames. As discussed by Tian et al. (2025), this bias primarily originates from the causal self-attention mechanism and is widely observed in open-source MLLMs (Liu et al., 2023; Li et al., 2026c) (Fig. 7 in the appendix). Such a bias is largely harmless for T2V videos, whose artifacts are uniformly distributed, but detrimental for I2V videos, whose most discriminative artifacts emerge precisely in the under-attended later frames. We therefore attribute the cross-paradigm performance disparity to **Temporal Artifact-Attention Misalignment**, where the LLM's early-frame attention bias fails to align with the divergent temporal artifact distributions of different generation paradigms, as illustrated in Fig. 2(b).

However, naive remedies, *i.e.*, modifying the causal attention mask (Tian et al., 2025) or discarding early frames (Kang et al., 2025), fail to resolve this issue, as they disrupt the pretrained causal information flow or discard the leading I-frames crucial for video representation learning (Tab. 16 and Tab. 18 in the appendix). To address this issue, we propose **Bi-level Temporal Re-Alignment (BiTRA)**, a joint optimization strategy that operates at both the extrinsic and intrinsic levels. At the extrinsic level, **Selective Hierarchical Artifact Injection (SHAI)** establishes an additional cross-attention pathway for global forgery information access and leverages the temporal artifact prior to guide discriminative information selection. At the intrinsic level, **Temporal Attention Smoother (TAS)** employs a Kullback-Leibler margin loss to mitigate excessive temporal attention bias inherent in the LLM decoder. In summary, our key contributions are three-fold:

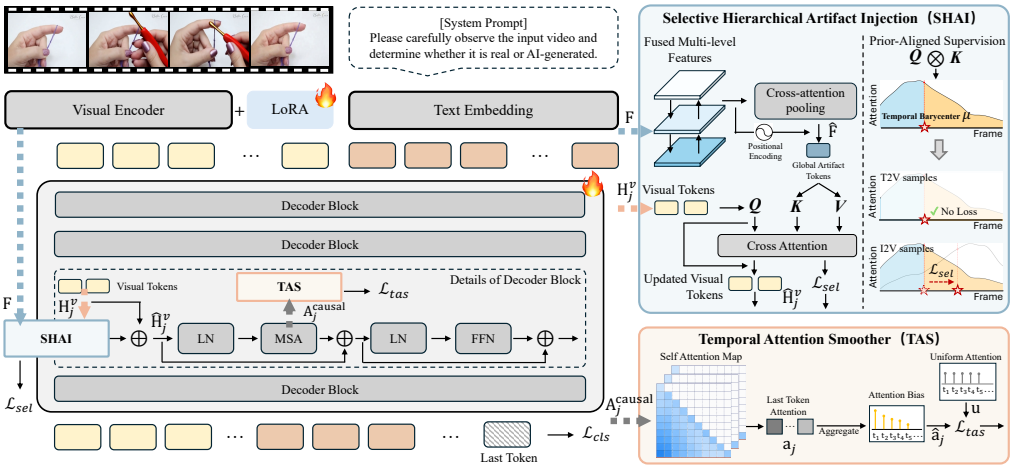



Figure 3: **The overall framework of BiTRA.** At the extrinsic level, Selective Hierarchical Artifact Injection builds a cross-attention pathway for global artifact access under a temporal-artifact prior. At the intrinsic level, Temporal Attention Smoother mitigates the temporal attention bias.

- We reveal a pronounced cross-paradigm disparity in MLLM-based AIGV detectors and attribute it to Temporal Artifact-Attention Misalignment.
- We propose Bi-level Temporal Re-Alignment (BiTRA), equipped with Selective Hierarchical Artifact Injection (SHAI) and Temporal Attention Smoother (TAS).
- Extensive experiments demonstrate the state-of-the-art performance of BiTRA on three popular AIGV datasets.

## 2 Related Work

### 2.1 AI-Generated Video Detection

Forgery detection has evolved from Deepfake detection (Li et al., 2018; 2020; Haliassos et al., 2021) to AIGI detection (Wang et al., 2020; Ojha et al., 2023; Wang et al., 2023; Tan et al., 2024), and more recently to AIGV detection. Conventional AIGV detectors can be divided into two categories: spatio-temporal representation learning (Bai et al., 2024; Ma et al., 2024; He et al., 2024; Liu et al., 2024; Chang et al., 2024; Ji et al., 2024; Kundu et al., 2025; Chen et al., 2026) and prior-based forensic modeling (Zheng et al., 2025; Ben Hayun et al., 2026; Internò et al., 2026; Zhang et al., 2026; Corvi et al., 2026; Fang et al., 2026).

### 2.2 MLLM-based AI-Generated Video Detection

Leveraging strong visual understanding and cross-modal reasoning, MLLMs have emerged as a promising framework for AIGV detection. Existing methods mainly follow two directions: discriminative representation learning (Song et al., 2024) and post-training optimization through supervised fine-tuning and reinforcement learning (Wen et al., 2025a;b). Specifically, VidGuard-R1 (Park et al., 2026) employs GRPO with artifact-aware rewards; Skyra (Li et al., 2026d) combines fine-grained supervision with SFT and RL; VideoVeritas (Tan et al., 2026) adopts perceptual RL to enhance local artifact perception; and Hermes (Li et al., 2026b) introduces evidence-driven planning and aggregation for interpretable detection. Despite these advances, current methods remain largely paradigm-agnostic, disregarding the intrinsic discrepancies in their underlying generation processes.

## 3 Method

The overall pipeline of our proposed **Bi-level Temporal Re-Alignment (BiTRA)** is illustrated in Fig. 3, designed to address **Temporal Artifact-Attention Misalignment** through complementary extrinsic and intrinsic re-alignment. At the extrinsic level, **Selective Hierarchical Artifact Injection (SHAI)** establishes an additional cross-attention pathway for global forgery information access, leveraging a temporal-artifact prior for discriminative information selection. At the intrinsic level, **Temporal Attention Smoother (TAS)** is guided by the Kullback-Leibler margin loss to mitigate the excessive temporal attention bias inherent in the LLM decoder. By integrating extrinsic- and

intrinsic-level adjustments within a unified architecture, BiTRA learns to concentrate on the most discriminative representations.

### 3.1 PRELIMINARIES

Given a text prompt $c$ and a real initial image $I_0$, a synthetic video with $T$ frames $\hat{V} = \{x_t\}_{t=0}^{T-1}$ can be formulated as

$$\hat{V}^{\mathrm{T2V}} \sim G_\theta(z, c),\ z \sim \mathcal{N}(0, \mathbf{I});\ \ \hat{V}^{\mathrm{I2V}} \sim G_\theta(z, c, I_0),\ \text{s.t. } \hat{x}^{(0)} \approx I_0. \tag{1}$$

Here, $\hat{V}^{\mathrm{T2V}}$ and $\hat{V}^{\mathrm{I2V}}$ are T2V and I2V samples, respectively. $G_\theta(\cdot)$ denotes the generative model and $z$ is a random noise vector sampled from a Gaussian distribution.
Concretely, for T2V, all frames are synthesized from pure noise under the same global text condition. As a result, forgery artifacts are typically distributed relatively uniformly throughout the video. In contrast, for I2V, the frames are additionally anchored by a real initial image $I_0$, which serves as a strong realistic conditioning signal in the early frames and gradually fades (Fig. 1(b)).
However, causal self-attention in the LLM decoder is strongly biased toward early-frame visual tokens, attending much less to later frames, while the most discriminative forgery artifacts in I2V videos emerge toward later frames, leaving these crucial cues largely unattended (Fig. 2(b)).

### 3.2 SELECTIVE HIERARCHICAL ARTIFACT INJECTION (SHAI)

To address this issue, a naive remedy is to replace causal attention over visual tokens with bidirectional attention. While this alleviates the disparity between T2V and I2V, it breaks the native causal information flow and consequently degrades overall performance(Tab. 18 in the appendix). Based on these observations, we further propose an adaptive extrinsic injection for global forgery information access and leverage the temporal artifact prior to guide discriminative information selection.

**Hierarchical Forgery Abstraction.** To capture visual forgery cues at both low and semantic levels for complementary representations (Raghu et al., 2021), we select $\mathcal{S}_1 = \{\ell_i\}_{i=1}^{L_1}$ from all Vision Transformer (ViT) layers, and extract features $\mathbf{F}_i \in \mathbb{R}^{B\times T\times N\times d}$ from the $i$-th level, where $L_1$ is the number of ViT extraction layers, $T$ denotes the number of input video frames, $N$ and $d$ denote the number of visual tokens per frame and their dimension, respectively. Then, based on the multi-level features, a bottom-up fusion can be formulated as

$$\mathbf{F}_{i+1} = \mathbf{F}_{i+1} + \Phi\left(\mathbf{F}_i\right), \quad \Phi\left(\mathbf{x}\right) = \mathbf{W}_2\,\mathrm{GELU}(\mathbf{W}_1\,\mathrm{LN}(\mathbf{x}))\,, \tag{2}$$

where $\Phi\left(\cdot\right)$ denotes a lightweight bottleneck adapter for cross-level feature transformation with two learnable projectors $\mathbf{W}_1$ and $\mathbf{W}_2$, $\mathrm{GELU}(\cdot)$ is the activation function, and $\mathrm{LN}(\cdot)$ is the layer-normalization operation. Similarly, a top-down fusion can be defined in the same manner. Such a bidirectional fusion not only enriches shallow features with high-level semantics, but also feeds fine-grained artifact cues back into deeper representations, which is crucial for AIGV detection (Lin et al., 2017; Zhao et al., 2021; Chen et al., 2022).
In addition, for the fused visual features $f_i^t \in \mathbb{R}^{B\times N\times d}$ of the $t$-th frame at the $i$-th visual level, we compress the number of tokens and incorporate position embeddings:

$$\widehat{f}_i^t = \mathbf{W}\left[\mathrm{Pool}\left(\mathrm{LN}\left(f_i^t\right)\right) + \mathbf{E}_{pos}^t + \mathbf{E}_{pos}^i\right], \tag{3}$$

where $\mathrm{Pool}(\cdot)$ denotes the pooling operation, $\mathbf{E}_{pos}^t$ and $\mathbf{E}_{pos}^i$ are the temporal and depth position embeddings, respectively. Subsequently, we gather $\widehat{f}_i^t$ over all frames and levels into global artifact tokens $\widehat{\mathbf{F}} \in \mathbb{R}^{B\times(TL_1)\times d'}$.

**Global Artifact Injection.** With the fused global artifact tokens, we introduce an additional cross-attention pathway in the LLM decoder for global forgery information access. Specifically, we select $\mathcal{S}_2 = \{\ell_j\}_{j=1}^{L_2}$ from LLM layers and extract visual hidden states $\mathbf{H}_j^v$ from the $j$-th level, serving as query to grasp forgery cues from global artifact tokens $\widehat{\mathbf{F}}$:

$$\mathbf{A}_j^v = \mathrm{MSA}(\mathbf{H}_j^v, \widehat{\mathbf{F}}, \widehat{\mathbf{F}}), \quad \widehat{\mathbf{H}}_j^v = \mathbf{H}_j^v + \mathbf{A}_j^v, \tag{4}$$

where $\mathrm{MSA}(\cdot)$ denotes the multi-head self attention operation, $\widehat{\mathbf{F}}$ serves as both the key and value, and $L_2$ is the number of LLM injection layers. The residual update is applied only to visual tokens, with the structure and masking scheme of the native causal self-attention unchanged. Different from existing position-wise injection mechanisms, such as DeepStack (Bai et al., 2025) and Deep

Table 1: Balanced accuracy (%) comparison of BiTRA and competing detectors on ViF-Bench (Li et al., 2026d) under T2V and I2V evaluation. The best and second-best results are in bold and underlined.

| Method | T2V Evaluation | | | | | | | | | | | | I2V Evaluation | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Wan2.1 -1.3B | CogV -X1.5 | Wan2.2 -5B | Hunyuan Video | Wan2.2 -14B | SkyReels -V2 | LTX-Video -13B | Hailuo -V2 | Pika -V2 | PixVerse -V4-5 | Kling -V1 | Sora -V2 | Wan2.2 -5B | Hunyuan Video | VACE -1.3B | Wan2.2 -14B | SkyReels -V2 | LTX-Video -13B | Gen4 -Turbo | |
| NSG-VD (NIPS'25) | 50.0 | 49.4 | 50.3 | 50.3 | 50.0 | 49.7 | 49.1 | 50.4 | 48.3 | 48.7 | 49.3 | 49.7 | 49.7 | 49.7 | 50.0 | 49.4 | 49.1 | 50.0 | 50.5 | 49.7 |
| ReStraV (NIPS'25) | 43.3 | 54.0 | 37.3 | 53.8 | 41.0 | 56.8 | 37.2 | 57.0 | 58.6 | 40.2 | 46.6 | 47.8 | 38.0 | 55.8 | 49.5 | 47.8 | 55.8 | 45.8 | 55.2 | 48.5 |
| STALL (CVPR'26) | 58.3 | 62.0 | 55.3 | 58.8 | 64.5 | 63.5 | 69.6 | 63.7 | 60.2 | 70.8 | 66.5 | 61.0 | 47.3 | 52.8 | 51.5 | 48.5 | 50.0 | 51.0 | 53.3 | 58.3 |
| D3 (ICCV'25) | 48.5 | 50.9 | 49.7 | 49.2 | 43.9 | 34.7 | 33.7 | 42.8 | 55.0 | 39.3 | 33.0 | 47.0 | 54.6 | 46.7 | 51.2 | 52.5 | 45.5 | 47.2 | 49.8 | 46.1 |
| DeMamba (arXiv'24) | 65.5 | 65.8 | 65.8 | 65.9 | 65.8 | 65.8 | 64.1 | 64.2 | 66.9 | 66.8 | 67.0 | 63.7 | 60.0 | 60.3 | 61.8 | 61.8 | 64.0 | 64.1 | 62.5 | 64.3 |
| CoCoDetect (CVPR'26) | 82.5 | 82.2 | 82.8 | 83.3 | 83.1 | 82.8 | 80.9 | 84.2 | 86.0 | 85.3 | 86.8 | 81.8 | 81.9 | 82.5 | 59.2 | 82.2 | 81.8 | 82.5 | 86.5 | 82.0 |
| AIGVDetect (ICLR'26) | 85.9 | 85.6 | 85.0 | 85.2 | 85.6 | 85.9 | 84.4 | 87.9 | 87.3 | 87.3 | 88.6 | 62.5 | 75.5 | 77.0 | 54.9 | 72.7 | 76.9 | 82.5 | 87.4 | 80.9 |
| BusterX++ (arXiv'25) | 54.9 | 59.4 | 52.4 | 59.2 | 62.4 | 65.8 | 56.3 | 61.7 | 76.8 | 75.7 | 52.8 | 52.3 | 50.3 | 49.7 | 50.9 | 49.7 | 50.0 | 50.0 | 50.9 | 56.9 |
| VideoVeritas (ICML'26) | 70.1 | 73.5 | 44.8 | 79.1 | 69.3 | 52.5 | 51.8 | 76.5 | 73.6 | 83.3 | 85.3 | 59.1 | 54.9 | 75.5 | 52.5 | 60.4 | 74.2 | 70.9 | 54.1 | 66.4 |
| Skyra (CVPR'26) | 97.0 | 96.4 | 92.1 | 94.8 | 96.4 | 95.8 | 95.9 | 96.0 | 96.4 | 96.1 | 94.7 | 91.0 | 87.6 | 93.6 | 79.1 | 84.6 | 79.0 | 83.7 | 79.5 | 91.0 |
| **BiTRA (Ours)** | **99.1** | **99.1** | **99.1** | **99.1** | **99.1** | **99.1** | **99.1** | **98.9** | **99.0** | **99.0** | **98.9** | **98.3** | **99.1** | **99.1** | **86.2** | **99.1** | **99.1** | **99.1** | **99.1** | **98.3** |

Residual Injection (Lin et al., 2026), which only enlarge the representation capacity of the LLM decoder (Tab. 12b in the appendix), our cross-attention-based injection allows each visual token to adaptively retrieve relevant cues based on its own query to mitigate the attention bias.

**Prior-Aligned Supervision.** To further enhance compatibility across generation paradigms, we supervise the selection of information for injection using prior knowledge of paradigm-dependent artifact distributions. As shown in Fig. 1(b), T2V videos exhibit relatively uniform frame-wise fake probabilities over time, whereas I2V videos contain weaker forgery signals in early frames that progressively intensify toward later ones. Based on this temporal prior, we explicitly supervise the artifact-injection process for I2V samples, enabling the cross-attention module to adaptively select more discriminative forgery cues for injection.

Concretely, we reshape the cross-attention map at the $j$-th selected LLM layer along the temporal and visual-level dimensions to obtain $\mathbf{A}_j^v \in \mathbb{R}^{B\times T\times N'\times L_1}$, where $N'$ denotes the number of visual token queries in the LLM decoder. We then aggregate the attention weights over visual queries and visual levels to obtain the frame-wise injection distribution:

$$p_j^v = \frac{1}{N'L_1}\sum_{n=1}^{N'}\sum_{i=1}^{L_1}\mathbf{A}_j^v. \tag{5}$$

Here, $p_j^v \in \mathbb{R}^{B\times T}$ quantifies the amount of artifact information retrieved from each frame at the $j$-th injection layer. After normalizing $p_j^v$ over the temporal dimension, we assign each frame $t$ a normalized temporal coordinate $\tau_t$ and compute the temporal barycenter $\mu_j$ of the distribution:

$$\tau_t = \frac{t}{T-1}, \qquad \mu_j = \sum_{t=0}^{T-1} p_j^v \tau_t. \tag{6}$$

The larger $\mu_j$ is, the more injection weight is assigned to later frames, where I2V forgery cues are generally more pronounced. Accordingly, for I2V samples, we define the temporal selection loss of the cross-attention pathway as

$$\mathcal{L}_{\text{sel}} = \frac{1}{L_2}\sum_{j=1}^{L_2}\left[m-\mu_j\right]_+^2. \tag{7}$$

Here, $[\cdot]_+ = \max\{0,\cdot\}$ and $m \in [0,1]$ denotes the temporal margin. When $\mu_j < m$, the loss encourages the injection distribution to shift toward later temporal positions. Once $\mu_j \geq m$, no additional penalty is imposed. This prior supervision guides cross-attention selection for I2V samples toward middle-to-late frames, thereby enhancing the utilization of late-frame artifact information.

### 3.3 Temporal Attention Smoother (TAS)

SHAI addresses **Temporal Artifact-Attention Misalignment** from the artifact side by establishing an additional cross-attention pathway for global forgery information access. We further introduce **Temporal Attention Smoother (TAS)** to mitigate excessive temporal attention bias in the LLM decoder at the intrinsic level.

At the $j$-th LLM injection layer, we constrain the causal-attention weights from the last token used for prediction. Let $\mathbf{A}_j^{\text{causal}} \in \mathbb{R}^{B\times M\times M}$ denote the native self-attention map of causal attention, where $M$ is the overall sequence length. We extract the attention of the last hidden state $\mathbf{a}_j \in$

Table 2: Average performance (%) of BiTRA and competing detectors on GenVideo (Chen et al., 2026), averaged over the ten video generators covered by its official test split, following the official evaluation protocols.

| Method | bAcc | Recall | F1 |
|---|---|---|---|
| NSG-VD (NIPS'25) | 91.5 | 88.0 | 90.9 |
| ReStraV (NIPS'25) | 70.4 | 75.5 | 70.8 |
| STALL (CVPR'26) | 77.8 | 79.3 | 77.9 |
| D3 (ICCV'25) | 78.5 | 71.4 | 76.8 |
| DeMamba (arXiv'24) | 90.5 | 94.3 | 90.8 |
| CoCoDetect (CVPR'26) | 86.5 | <u>96.9</u> | 87.8 |
| AIGVDetect (ICLR'26) | <u>95.4</u> | 91.2 | 90.6 |
| BusterX++ (arXiv'25) | 90.7 | 96.3 | <u>91.2</u> |
| VideoVeritas (ICML'26) | 74.8 | 90.7 | 78.0 |
| Skyra (CVPR'26) | 91.0 | 87.8 | 90.3 |
| **BiTRA (Ours)** | **97.7** | **98.2** | **97.7** |

Table 3: Balanced accuracy (%) of BiTRA and competing detectors on the GenBuster (ID) (Wen et al., 2025a) and GenBuster++ (OOD) (Wen et al., 2025b) test sets.

| Method | ID | OOD | Avg. |
|---|---|---|---|
| NSG-VD (NIPS'25) | 61.4 | 53.3 | 57.3 |
| ReStraV (NIPS'25) | 58.6 | 54.5 | 56.5 |
| STALL (CVPR'26) | 69.5 | 72.3 | 70.9 |
| D3 (ICCV'25) | 60.3 | 50.5 | 55.4 |
| DeMamba (arXiv'24) | 91.2 | <u>86.1</u> | <u>88.7</u> |
| CoCoDetect (CVPR'26) | <u>94.4</u> | 50.0 | 72.2 |
| AIGVDetect (ICLR'26) | 84.3 | 69.6 | 76.9 |
| BusterX++ (arXiv'25) | 79.2 | 81.7 | 80.4 |
| Skyra (CVPR'26) | 43.4 | 58.9 | 51.1 |
| **BiTRA (Ours)** | **94.9** | **89.8** | **92.3** |
| VideoVeritas* (ICML'26) | <u>95.7</u> | <u>91.4</u> | <u>93.6</u> |
| **BiTRA* (Ours)** | **98.0** | **94.3** | **96.1** |

$\mathbb{R}^{B\times(TN')}$ over all visual tokens, and aggregate it within each frame to obtain a normalized per-frame distribution $\widehat{\mathbf{a}}_j \in \mathbb{R}^{B\times T}$.

To avoid over-constraining the native attention pattern, TAS adopts a margin-based smoothing loss for all samples:

$$\mathcal{L}_{\text{tas}} = \frac{1}{L_2}\sum_{j=1}^{L_2}\left[\frac{D_{\text{KL}}\left(\widehat{\mathbf{a}}_j \| \mathbf{u}\right)}{\log T} - \delta\right]_+, \tag{8}$$

where $\mathbf{u} = (1/T, \ldots, 1/T)$ denotes the uniform reference distribution and $\delta$ denotes the smoothing margin. The normalized KL divergence measures the deviation from uniform temporal allocation, with $\log T$ as its theoretical upper bound. The constraint is activated only when this deviation exceeds the margin, allowing moderate non-uniformity while suppressing excessive concentration.

### 3.4 Overall Loss

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{sel}}\mathcal{L}_{\text{sel}} + \lambda_{\text{tas}}\mathcal{L}_{\text{tas}}, \tag{9}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss applied to the last-layer hidden state of the final token in the LLM, $\lambda_{\text{sel}}$ and $\lambda_{\text{tas}}$ are hyper-parameters to balance the regularization objectives. These two regularizers operate in complementary attention spaces: $\mathcal{L}_{\text{sel}}$ guides temporal information selection in the additional cross-attention pathway of SHAI, whereas $\mathcal{L}_{\text{tas}}$ controls temporal concentration in the native causal attention.

## 4 Experiments

### 4.1 Datasets and Evaluation Metrics

Strictly following previous work (Chen et al., 2026; Wen et al., 2025b; Li et al., 2026d; Zhang et al., 2026; Internò et al., 2026; Zheng et al., 2025), we evaluate our method on three popular AIGV detection benchmarks under their official splits and protocols. For **ViF-Bench** (Li et al., 2026d), we conduct evaluations by training the models on the training split with 16 frames per video and testing them on both T2V and I2V test splits. For **GenVideo** (Chen et al., 2026), we conduct a complementary evaluation across diverse video generators, using its official splits and 8 frames per video. For **GenBuster** (Wen et al., 2025a) and **GenBuster++** (Wen et al., 2025b), we train the model on GenBuster and evaluate it on the in-distribution (ID) GenBuster test set as well as the out-of-distribution (OOD) GenBuster++ test set. For fair comparison with other work (Li et al., 2026d; Tan et al., 2026; Wen et al., 2025b; Li et al., 2026e; Feng et al., 2026; Chen et al., 2026; Zheng et al., 2025; Ben Hayun et al., 2026; Internò et al., 2026; Zhang et al., 2026), both **Balanced Accuracy (bAcc)**, **Recall**, and **F1 score** are reported under threshold 0.5.

Table 4: Ablation studies on Selective Hierarchical Artifact Injection and Temporal Attention Smoother, reported in bAcc (%).

| SHAI | TAS | T2V | I2V | Overall |
|---|---|---|---|---|
| × | × | 92.3 (+0.0) | 84.0 (+0.0) | 89.2 (+0.0) |
| ✓ | × | 97.0 (+4.7) | 92.9 (+8.9) | 95.5 (+6.3) |
| × | ✓ | 94.1 (+1.8) | 88.4 (+4.4) | 92.0 (+2.8) |
| ✓ | ✓ | **99.0** (+6.7) | **97.2** (+13.2) | **98.3** (+9.1) |

Table 5: Ablation studies on the individual components of Selective Hierarchical Artifact Injection, in bAcc (%).

| GAI | HFA | $\mathcal{L}_{\text{sel}}$ | T2V | I2V | Overall |
|---|---|---|---|---|---|
| × | × | × | 94.1 (+0.0) | 88.4 (+0.0) | 92.0 (+0.0) |
| ✓ | × | × | 95.3 (+1.2) | 91.9 (+3.5) | 94.1 (+2.1) |
| ✓ | ✓ | × | 98.5 (+4.4) | 94.5 (+6.1) | 97.0 (+5.0) |
| ✓ | ✓ | ✓ | **99.0** (+4.9) | **97.2** (+8.8) | **98.3** (+6.3) |

### 4.2 Implementation Details

Following the pre-processing pipeline of Skyra (Li et al., 2026d), we utilize Qwen3-VL-Embedding-2B (Li et al., 2026a) as our default backbone based on PyTorch and Transformer. We adapt the visual encoder using LoRA ($r = 16$, $\alpha = 32$) and fully fine-tune the LLM as our baseline, with the fixed learning rates of $1e-4$ and $2e-5$, respectively. SHAI extracts features from visual layers $\mathcal{S}_1 = \{6, 12, 18, 24\}$, and both $\mathcal{L}_{\text{sel}}$ and $\mathcal{L}_{\text{tas}}$ are performed at LLM layers $\mathcal{S}_2 = \{7, 14, 21, 28\}$. We set $m = 0.5$, $\delta = 0.4$, and $\lambda_{\text{sel}} = \lambda_{\text{tas}} = 0.1$. We use AdamW (Loshchilov & Hutter, 2017) with batch size of 64, for 10 epochs on ViF-Bench and 5 epochs on other datasets. For compared methods, we report results from their original papers when available, and otherwise reproduce them using official codes.

### 4.3 Experimental Results

**Comparison with competitive detectors.** We compare BiTRA with state-of-the-art AIGV detectors on three benchmarks. **On ViF-Bench**, BiTRA surpasses Skyra (Li et al., 2026d) by +7.3% in average bAcc, achieving the best performance on both T2V and I2V evaluations (Tab. 1). **On GenVideo**, compared to AIGVDetect (Li et al., 2026e) which utilizes ViT from Qwen2.5-VL for classification, BiTRA achieves the best average bAcc of 97.7% across ten generators, with a gain of +2.3% (Tab. 2). **On GenBuster and GenBuster++**, BiTRA trained solely on the GenBuster training set achieves bAcc performance of 94.9% and 89.8% on the ID and OOD test sets, respectively, yielding the best average of 92.3%. Notably, it outperforms the second-best DeMamba (Chen et al., 2026) by +3.7% on GenBuster++, demonstrating superior OOD generalization without auxiliary data. For fair comparison with VideoVeritas (Tan et al., 2026), BiTRA* incorporates the same additional training data and improves the ID and OOD results to 98.0% and 94.3%, respectively, achieving the best average bAcc of 96.1% and surpassing VideoVeritas* by +2.5% (Tab. 3).

### 4.4 Ablation Studies

**The Impact of Main Components in BiTRA.** In Tab. 4, we first ablate Selective Hierarchical Artifact Injection (SHAI). We can see that the additional cross-attention injection with forgery-artifact prior can bring +6.3% and +8.9% improvement on overall and I2V bACC, respectively. Such gains also validate the necessity of SHAI at the extrinsic level. Moreover, Temporal Attention Smoother (TAS) can also improve 2.8% average bAcc, further mitigating the excessive temporal attention bias in the LLM decoder at the intrinsic level. Finally, the performance gap between T2V and I2V shrinks from 8.3% to 1.8%, demonstrating the complementarity of the two components in alleviating Temporal Artifact-Attention Misalignment.

**The Impact of components in Selective Hierarchical Artifact Injection.** In Tab. 5, we ablate the effectiveness of components in SHAI. First, Global Artifact Injection (GAI) improves the overall bACC from 92.0% to 94.1% by providing adaptive access to global artifact information. Subsequently, Hierarchical Forgery Abstraction (HFA) further gives a gain of +2.9% by integrating complementary forensic cues across multiple ViT depths. Lastly, Prior-Aligned Supervision with $\mathcal{L}_{\text{sel}}$ achieves 98.3% overall bAcc, with I2V improving from 94.5% to 97.2%, validating the benefit of aligning the temporal allocation of injected artifact information with the temporal evolution of I2V forgery cues.

**Compatibility with other multi-modal foundation models.** In Tab. 6, we transfer our Bi-level Temporal ReAlignment (BiTRA) to other multi-modal backbones of both NTP-based (Chen et al., 2023; Li et al., 2024) and embedding-based (Li et al., 2026a) paradigms. These results show that

Table 6: Compatibility of BiTRA with NTP-based and embedding-based multimodal foundation models on ViF-Bench. All results are reported in bACC (%).

| Backbone | Type | Baseline | | | + BiTRA (Ours) | | |
|---|---|---|---|---|---|---|---|
| | | T2V | I2V | Overall | T2V | I2V | Overall |
| LLaVA-0.5B | NTP | 84.9 | 79.1 | 82.8 | **89.8** (+4.9) | **82.7** (+3.6) | **87.2** (+4.4) |
| InternVL-2B | NTP | 86.4 | 79.9 | 84.0 | **91.7** (+5.3) | **85.5** (+5.6) | **89.4** (+5.4) |
| Qwen2.5-VL-Emb-7B | Emb. | 94.5 | 85.6 | 91.2 | **98.0** (+3.5) | **92.4** (+6.8) | **95.9** (+4.7) |
| Qwen3-VL-Emb-2B | Emb. | 92.3 | 84.0 | 89.2 | **99.0** (+6.7) | **97.2** (+13.2) | **98.3** (+9.1) |
| Qwen3-VL-Emb-8B | Emb. | 92.9 | 83.1 | 88.7 | **97.7** (+4.8) | **94.3** (+11.2) | **96.5** (+7.8) |

Table 7: Comparison with general attention-sink mitigation methods on ViF-Bench; $(\cdot)$ denotes the gain over baseline.

| Method | bAcc | Recall | F1 |
|---|---|---|---|
| SoFA | 83.7 (-5.5) | <u>96.4</u> | 85.4 |
| VAR | <u>90.0</u> (+0.8) | 91.5 | <u>89.7</u> |
| HwRMSNorm | 82.0 (-7.2) | 78.7 | 79.8 |
| **BiTRA (Ours)** | **98.3** (+9.1) | **98.6** | **98.2** |

Table 8: Cross-paradigm bACC (%) on AIGVD-Bench (Ma et al., 2026). † and ‡ denote training with exclusively T2V and I2V fake samples.

| Method | T2V | I2V | V2V | Avg. |
|---|---|---|---|---|
| Baseline† | 95.9 (+0.0) | 78.5 (+0.0) | 73.8 (+0.0) | 82.7 (+0.0) |
| **BiTRA**† | **96.9** (+1.0) | **81.1** (+2.6) | **78.3** (+4.5) | **85.4** (+2.7) |
| Baseline‡ | 85.1 (+0.0) | 95.1 (+0.0) | 89.0 (+0.0) | 89.7 (+0.0) |
| **BiTRA**‡ | **87.4** (+2.3) | **95.5** (+0.4) | **94.0** (+5.0) | **92.3** (+2.6) |

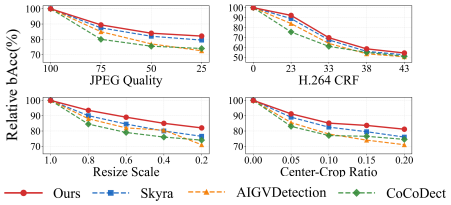
our method is compatible with various backbones, *e.g.*, +5.4% average bAcc with InternVL-2B and +7.8% average bAcc with Qwen3-VL-Emb-8B compared to the baseline.

**Comparison with general attention-sink mitigation methods.** In Tab. 7, we compare BiTRA with representative attention-sink mitigation methods, *i.e.*, soft-bidirectional masking of SoFA (Tian et al., 2025), visual-sink attention recycling of VAR (Kang et al., 2025), and head-wise RMSNorm (Li et al., 2026c). These methods target general vision-language understanding, where evidence spreads over the whole input, so attention is flattened uniformly, blind to *where* the evidence lies. Among them, compared to our baseline, VAR yields only a marginal gain ($+0.8\%$ bAcc) in that sink-token recycling partially relieves early-frame dominance, yet it is spread uniformly rather than focusing on artifact-rich later frames. All three methods underperform BiTRA by $8.3\%$–$16.3\%$ bAcc, confirming that paradigm-conditioned forensic adaptation, rather than generic attention correction, constitutes the critical factor for temporal re-alignment.

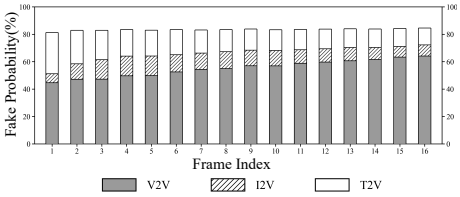
**Detailed ablations of proposed designs.** We further ablate the detailed configurations of BiTRA in the appendix. **For Selective Hierarchical Artifact Injection**, bidirectional fusion performs better than unidirectional ones, and the cross-attention injection is better than injection strategies such as addition and concatenation. Supervision loss $\mathcal{L}_{\text{sel}}$ with temporal barycenter is better than other variants such as ranking loss of APR (Rios et al., 2021). **For Training Strategies**, jointly adapting the visual encoder and LLM performs best, with LoRA for the visual encoder and full fine-tuning for the LLM. **For Hyper-parameters**, both the number of ViT extraction layers $L_1$ and that of LLM injection layers $L_2$ are set to 4, which performs best. In both cases, the layers are selected at uniform intervals. **For Alternative Remedies**, modifying the causal attention mask or discarding early frames yields no consistent gain: the former disrupts the pretrained causal information flow, and the latter hurts real-video specificity by discarding leading I-frames (Tab. 16 & 18 in the appendix).

### 4.5 Further Analysis

**Analysis of Video-to-Video (V2V) samples.** We further extend the frame-wise artifact analysis to AIGVDBench (Ma et al., 2026), which covers text-to-video (T2V), image-to-video (I2V), and video-to-video (V2V) generation. Using the same set of SOTA AIGI detectors (Yan et al., 2024; Chen et al., 2025), we find that **T2V exhibits relatively uniform fake confidence over time, whereas both I2V and V2V show progressively increasing trends, with V2V maintaining lower overall probabilities than I2V**, as shown in Fig. 4(b). This suggests that the two visually condi-

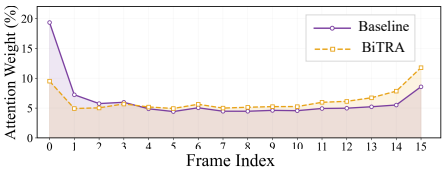



(a) Robustness under common video perturbations on ViF-Bench.

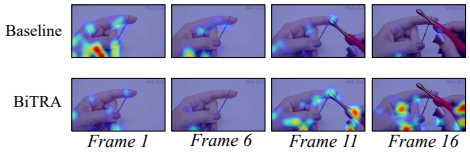



(b) Frame-wise fake probability across generation paradigms.

Figure 4: (a) Robustness under common video perturbations on ViF-Bench, in Relative bAcc (%). (b) Frame-wise fake probability across generation paradigms on AIGVDBench: T2V stays uniform, while I2V and V2V increase over time, with V2V lower than I2V.




(a) Temporal Casual Attention Weight.




(b) Casual Attention Heatmaps on I2V.

Figure 5: Mitigation of temporal attention bias on I2V videos. (a) Frame-wise attention weights over 16 frames: BiTRA suppresses the initial-frame dominance and yields a smoother distribution. (b) Attention heatmaps: BiTRA shifts attention toward later, artifact-rich frames.

tioned paradigms, I2V and V2V, exhibit similar temporal artifact dynamics. V2V, however, shows weaker forgery signals because it inherits not only appearance but also motion and temporal structure from the source video, allowing it to preserve content and dynamics more faithfully despite the gradual accumulation of transformation errors and temporal drift.

In Tab. 8, consistent with these observations, **models trained on I2V generalize better to V2V than those trained on T2V, indicating that V2V shares more transferable forensic cues with I2V**. Moreover, **BiTRA consistently improves the performance of T2V, I2V, and V2V under both training settings, demonstrating its compatibility and generalization across paradigms**.

**Analysis of Robustness.** Following AIGVDetect (Li et al., 2026e), We further evaluate the robustness of our method under common video perturbations, including compression, downscaling, and center cropping. BiTRA consistently outperforms compared methods across all perturbation levels, showing stronger robustness for AIGV detection (Fig. 4(a)).

**Visualization of the mitigated temporal causal attention bias.** We first visualize the frame-wise causal attention distributions of all samples in Fig. 5(a). The baseline concentrates heavily on the first frame (nearly 20%) and rapidly decays thereafter, where forgery evidence is relatively weak. In contrast, BiTRA suppresses this initial spike (down to about 9.5%) and progressively increases its attention toward the later frames, which contain stronger artifacts. Moreover, the attention heatmaps on I2V samples are illustrated in Fig. 5(b). We can see that the baseline scatters its focus over early, artifact-poor regions, whereas BiTRA increasingly attends to the forged tool-hand regions in later frames (*e.g.*, Frames 11 and 16). These results confirm that BiTRA mitigates the temporal attention bias and better re-aligns attention with the temporal evolution of forgery artifacts.

## 5 Conclusion

We reveal distinct temporal forgery dynamics across T2V and I2V and identify Temporal Artifact–Attention Misalignment between later-emerging I2V artifacts and early-biased causal attention as a key cause of degraded I2V performance. To address this issue, we propose Bi-level Temporal Re-Alignment (BiTRA), integrating Selective Hierarchical Artifact Injection (SHAI) for selective global artifact access and Temporal Attention Smoother (TAS) for mitigating temporal attention bias while preserving the native causal structure. Experiments on three benchmarks demonstrate state-of-the-art performance and improved cross-paradigm balance. Our analysis further extends to V2V, revealing that it shares I2V-like temporal artifact dynamics and likewise benefits from BiTRA.

### AI USE STATEMENT

In this work, we used generative AI tools only for minor language polishing and grammar checking of the manuscript. We did not use generative AI tools for research ideation, experimental design, data analysis, or the generation of research results. All AI-assisted text has been reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

### ETHICS STATEMENT

This work studies the detection of AI-generated videos, which contributes positively to digital media forensics and the mitigation of misinformation. All datasets used in this work are publicly available benchmarks, and no human subjects or private personal data are involved. We foresee no direct negative societal impact from this research.

### REPRODUCIBILITY STATEMENT

We describe the datasets and evaluation metrics in the Datasets and Evaluation Metrics subsection, and provide the full training configurations, including backbones, optimizers, learning rates, and hyper-parameter settings, in the Implementation Details subsection. Additional ablation configurations are reported in the appendix.

## REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. Ai-generated video detection via spatial-temporal anomaly learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 460–470. Springer, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Joy Battocchio, Stefano Dell'Anna, Andrea Montibeller, and Giulia Boato. Advance fake video detection via vision transformers. In *Proceedings of the 2025 ACM Workshop on Information Hiding and Multimedia Security*, pp. 1–11, 2025.

Omer Ben Hayun, Roy Betser, Meir Yossef Levi, Levi Kassel, and Guy Gilboa. Training-free detection of generated videos via spatial-temporal likelihoods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16299–16310, 2026.

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.

Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. What matters in detecting ai-generated videos like sora? *arXiv e-prints*, pp. arXiv–2406, 2024.

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. *Science China Information Sciences*, 69(6):162103, 2026.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Riccardo Corvi, Davide Cozzolino, Ekta Prashnani, Shalini De Mello, Koki Nagano, and Luisa Verdoliva. Seeing what matters: Generalizable ai-generated video detection with forensic-oriented augmentation. *Advances in neural information processing systems*, 38:26418–26446, 2026.

Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. *Advances in neural information processing systems*, 32, 2019.

Ziyuan Fang, Tianyi Wei, Guanjie Wang, Weiming Zhang, Nenghai Yu, and Wenbo Zhou. Real data lies: Unveiling and closing the quality shortcut in generalizable AI-generated video detection. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=sYzf4xhthY`.

Huidong Feng, Wentao Chen, Jie Chen, Xinqi Cai, Ruolong Ma, Yinglin Zheng, Yuxin Lin, and Ming Zeng. Cocovideo: The high-quality commercial-model-based contrastive benchmark for ai-generated video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11346–11356, 2026.

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*, pp. 5039–5049, 2021.

Peisong He, Leyao Zhu, Jiaxing Li, Shiqi Wang, and Haoliang Li. Exposing ai-generated videos: A benchmark dataset and a local-and-global temporal defect based detection method. *arXiv preprint arXiv:2405.04133*, 2024.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. Ai-generated video detection via perceptual straightening. *Advances in neural information processing systems*, 38:20672–20705, 2026.

Lichuan Ji, Yingqi Lin, Zhenhua Huang, Yan Han, Xiaogang Xu, Jiafei Wu, Chong Wang, and Zhe Liu. Distinguish any fake videos: Unleashing the power of large-scale data and motion features. *arXiv preprint arXiv:2405.15343*, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28050–28060, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001–5010, 2020.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026a.

Shuaibo Li, HAO Pengfei, Hongtao Wu, Jianfeng Dong, Ping Li, Xiaohong Liu, and Lei Zhu. Hermes: An evidence-driven agentic framework for trustworthy and explainable ai-generated video detection. In *Forty-third International Conference on Machine Learning*, 2026b.

Siquan Li, Kaiqi Jiang, Jiacheng Sun, and Tianyang Hu. The structural origin of attention sink: Variance discrepancy, super neurons, and dimension disparity. In *Forty-third International Conference on Machine Learning*, 2026c.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: Ai-generated video detection via grounded artifact reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4482–4493, 2026d.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, 2018. doi: 10.1109/WIFS.2018.8630787.

Zhengcen Li, Chenyang Jiang, Hang Zhao, Shiyang Zhou, Yunyang Mo, Feng Gao, Fan Yang, Qiben Shan, Shaocong Wu, and Jingyong Su. Preserving forgery artifacts: AI-generated video detection at native scale. In *The Fourteenth International Conference on Learning Representations*, 2026e. URL https://openreview.net/forum?id=XD43lfRCg6.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, et al. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models. *arXiv preprint arXiv:2606.15880*, 2026.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

Qingyuan Liu, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang. Turns out i'm not real: Towards robust detection of ai-generated videos. *arXiv preprint arXiv:2406.09601*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Long Ma, Jiajia Zhang, Hongping Deng, Ningyu Zhang, Yong Liao, and Haiyang Yu. Decof: Generated video detection via frame consistency. *arXiv preprint arXiv*, 2402, 2024.

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your one-stop solution for ai-generated video detection. *arXiv preprint arXiv:2601.11035*, 2026.

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*, pp. 24480–24489, 2023.

Kyoungjun Park, Yifan Yang, Juheon Yi, Shicheng Zheng, Muhammad Muaz, Yifei Shen, Dongqi Han, Caihua Shan, and Lili Qiu. Vidguard-r1: AI-generated video detection and explanation via reasoning MLLMs and RL. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gXjOsBcXIR.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Annette Rios, Chantal Amrhein, Noëmi Aepli, and Rico Sennrich. On biasing transformer attention towards monotonicity. *arXiv preprint arXiv:2104.03945*, 2021.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

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. *Advances in Neural Information Processing Systems*, 37:122054–122077, 2024.

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*, pp. 28130–28139, 2024.

Hao Tan, Senyuan Shi, Zichang Tan, Zijian Yu, Huijia Zhu, Weiqiang Wang, Jun Wan, Zhen Lei, et al. Videoveritas: Ai-generated video detection via perception pretext reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10599–10609, 2025.

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*, pp. 8695–8704, 2020.

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*, pp. 22445–22455, 2023.

Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Busterx: Mllm-powered ai-generated video forgery detection and explanation. *arXiv preprint arXiv:2505.12620*, 2025a.

Haiquan Wen, Tianxiao Li, Zhenglin Huang, Yiwei He, and Guangliang Cheng. Busterx++: Towards unified cross-modal ai-generated content detection and explanation with mllm. *arXiv preprint arXiv:2507.14632*, 2025b.

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*, pp. 399–417. Springer, 2024.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for ai-generated video detection. *Advances in neural information processing systems*, 38:174683–174730, 2026.

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*, pp. 2185–2194, 2021.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free ai-generated video detection using second-order features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12852–12862, 2025.

| Method | Metric | T2V Evaluation | | | | | | | | | | | | I2V Evaluation | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Wan2.1 -1.3B | CogV -X1.5 | Wan2.2 -5B | Hunyuan Video | Wan2.2 -14B | SkyReels -V2 | LTX -13B | Hailuo -V2 | Pika -V2 | PixVerse -V4.5 | Kling -V1 | Sora -V2 | Wan2.2 -5B | Hunyuan Video | VACE -1.3B | Wan2.2 -14B | SkyReels -V2 | LTX -13B | Gen4 -Turbo | |
| NSG-VD | bAcc | 50.0 | 49.4 | 50.3 | 50.3 | 50.0 | 49.7 | 49.1 | 50.4 | 48.3 | 48.7 | 49.3 | 49.7 | 49.7 | 49.7 | 50.0 | 49.4 | 49.1 | 50.0 | 50.5 | 49.7 |
| | Recall | 99.4 | 98.2 | 100.0 | 100.0 | 99.4 | 98.8 | 97.5 | 100.0 | 96.0 | 96.7 | 97.9 | 98.7 | 98.8 | 98.8 | 99.4 | 98.2 | 97.6 | 99.4 | 100.0 | 98.7 |
| | F1 | 66.5 | 66.0 | 66.8 | 66.8 | 66.5 | 66.3 | 65.7 | 66.8 | 65.0 | 65.3 | 65.9 | 66.2 | 66.3 | 66.3 | 66.5 | 66.0 | 65.7 | 66.5 | 66.9 | 66.2 |
| ReStraV | bAcc | 43.3 | 54.0 | 37.3 | 53.8 | 41.0 | 56.8 | 37.2 | 57.0 | 58.6 | 40.2 | 46.6 | 47.8 | 38.0 | 55.8 | 49.5 | 47.8 | 55.8 | 45.8 | 55.2 | 48.5 |
| | Recall | 57.0 | 72.0 | 59.0 | 75.4 | 51.0 | 97.0 | 39.7 | 80.2 | 92.5 | 51.9 | 61.9 | 65.4 | 55.0 | 77.0 | 47.5 | 85.5 | 80.4 | 75.5 | 76.0 | 68.4 |
| | F1 | 50.1 | 61.0 | 48.5 | 62.0 | 46.4 | 69.2 | 38.7 | 65.1 | 69.1 | 46.5 | 53.7 | 55.6 | 47.0 | 63.5 | 48.5 | 62.1 | 64.5 | 58.2 | 62.9 | 56.4 |
| STALL | bAcc | 58.3 | 62.0 | 55.3 | 58.8 | 64.5 | 63.5 | 69.6 | 63.7 | 60.2 | 70.8 | 66.5 | 61.0 | 47.3 | 52.8 | 51.5 | 48.5 | 50.0 | 51.0 | 53.3 | 58.3 |
| | Recall | 51.0 | 58.5 | 45.0 | 51.8 | 63.5 | 61.5 | 73.9 | 61.1 | 54.3 | 76.5 | 67.6 | 57.7 | 29.0 | 40.0 | 37.5 | 31.5 | 34.7 | 36.5 | 38.3 | 51.0 |
| | F1 | 55.0 | 60.6 | 50.1 | 55.7 | 64.1 | 62.8 | 70.8 | 62.7 | 57.7 | 72.4 | 66.9 | 59.7 | 35.5 | 45.9 | 43.6 | 38.0 | 41.0 | 42.7 | 45.0 | 54.2 |
| D3 | bAcc | 48.5 | 50.9 | 49.7 | 49.2 | 43.9 | 34.7 | 33.7 | 42.8 | 55.0 | 39.3 | 33.0 | 47.0 | 54.6 | 46.7 | 51.2 | 52.5 | 45.5 | 47.2 | 49.8 | 46.1 |
| | Recall | 50.9 | 55.8 | 53.4 | 52.5 | 41.7 | 23.3 | 21.5 | 39.0 | 64.7 | 32.0 | 17.9 | 48.0 | 63.2 | 47.3 | 56.4 | 58.9 | 45.1 | 48.5 | 55.0 | 46.1 |
| | F1 | 49.7 | 53.2 | 51.5 | 50.8 | 42.6 | 26.3 | 24.5 | 38.3 | 57.9 | 33.6 | 19.8 | 46.3 | 58.2 | 47.1 | 53.6 | 55.3 | 45.2 | 47.9 | 48.2 | 44.7 |
| DeMamba | bAcc | 65.5 | 65.8 | 65.8 | 65.9 | 65.8 | 65.8 | 64.1 | 64.2 | 66.9 | 66.8 | 67.0 | 63.7 | 60.0 | 60.3 | 61.8 | 61.8 | 64.0 | 64.1 | 62.5 | 64.3 |
| | Recall | 99.4 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 96.9 | 97.8 | 100.0 | 100.0 | 100.0 | 96.0 | 88.5 | 89.1 | 92.1 | 92.1 | 96.3 | 96.3 | 92.0 | 96.7 |
| | F1 | 74.2 | 74.5 | 74.5 | 74.6 | 74.5 | 74.5 | 72.9 | 73.2 | 75.1 | 75.1 | 75.2 | 72.5 | 68.9 | 69.2 | 70.7 | 70.7 | 72.8 | 72.9 | 71.0 | 73.0 |
| CoCoDetect | bAcc | 82.5 | 82.2 | 82.8 | 83.3 | 83.1 | 82.8 | 80.9 | 84.2 | 86.0 | 85.3 | 86.8 | 81.8 | 81.9 | 82.5 | 59.2 | 82.2 | 81.8 | 82.5 | 86.5 | 82.0 |
| | Recall | 98.8 | 98.2 | 99.4 | 100.0 | 100.0 | 99.4 | 95.6 | 97.8 | 100.0 | 100.0 | 99.3 | 95.3 | 97.6 | 98.8 | 52.2 | 98.2 | 96.9 | 98.8 | 95.5 | 95.9 |
| | F1 | 85.0 | 84.7 | 85.3 | 85.7 | 85.6 | 85.3 | 83.4 | 86.1 | 87.7 | 87.2 | 88.3 | 83.9 | 84.4 | 85.0 | 56.1 | 84.7 | 84.2 | 85.0 | 87.6 | 83.9 |
| AIGVDetect | bAcc | 85.9 | 85.6 | 85.0 | 85.2 | 85.6 | 85.9 | 84.4 | 87.9 | 87.3 | 87.3 | 88.6 | 62.5 | 75.5 | 77.0 | 54.9 | 72.7 | 76.9 | 82.5 | 87.4 | 80.9 |
| | Recall | 100.0 | 99.4 | 98.2 | 98.2 | 99.4 | 100.0 | 96.9 | 100.0 | 100.0 | 99.3 | 99.3 | 52.7 | 79.1 | 82.2 | 38.0 | 73.6 | 81.5 | 93.3 | 96.4 | 88.8 |
| | F1 | 87.6 | 87.3 | 86.7 | 86.9 | 87.3 | 87.6 | 86.1 | 89.2 | 88.8 | 88.7 | 89.7 | 58.4 | 76.3 | 78.1 | 45.8 | 73.0 | 77.9 | 84.2 | 88.4 | 81.5 |
| BusterX++ | bAcc | 54.9 | 59.4 | 52.4 | 59.2 | 62.4 | 65.8 | 56.3 | 61.7 | 76.8 | 75.7 | 52.8 | 52.3 | 50.3 | 49.7 | 50.9 | 49.7 | 50.0 | 50.0 | 50.9 | 56.9 |
| | Recall | 10.3 | 19.4 | 5.5 | 18.9 | 25.5 | 32.1 | 13.1 | 24.1 | 54.3 | 52.0 | 5.7 | 5.3 | 1.2 | 0.0 | 2.4 | 0.0 | 0.6 | 0.6 | 2.7 | 14.4 |
| | F1 | 18.6 | 32.3 | 10.3 | 31.6 | 40.4 | 48.4 | 23.1 | 38.6 | 70.1 | 68.1 | 10.7 | 10.1 | 2.4 | 0.0 | 4.7 | 0.0 | 1.2 | 1.2 | 5.2 | 21.9 |
| VideoVeritas | bAcc | 70.1 | 73.5 | 44.8 | 79.1 | 69.3 | 52.5 | 51.8 | 76.5 | 73.6 | 83.3 | 85.3 | 59.1 | 54.9 | 75.5 | 52.5 | 60.4 | 74.2 | 70.9 | 54.1 | 66.4 |
| | Recall | 58.6 | 65.4 | 8.0 | 76.3 | 57.1 | 23.3 | 22.1 | 70.6 | 64.3 | 84.0 | 89.3 | 37.8 | 28.2 | 69.3 | 23.5 | 39.3 | 66.9 | 60.1 | 23.4 | 50.9 |
| | F1 | 66.2 | 71.1 | 12.6 | 78.5 | 65.0 | 32.9 | 31.4 | 75.0 | 70.9 | 83.4 | 85.9 | 48.1 | 38.5 | 73.9 | 33.0 | 49.8 | 72.2 | 67.4 | 33.8 | 57.4 |
| Skyra | bAcc | 97.0 | 96.4 | 92.1 | 94.8 | 96.4 | 95.8 | 95.9 | 96.0 | 96.4 | 96.1 | 94.7 | 91.0 | 87.6 | 93.6 | 79.1 | 84.6 | 79.0 | 83.7 | 79.5 | 91.0 |
| | Recall | 100.0 | 98.8 | 90.3 | 95.7 | 98.8 | 97.6 | 98.1 | 98.5 | 99.3 | 98.7 | 96.5 | 88.7 | 81.2 | 93.3 | 64.2 | 75.2 | 64.0 | 73.6 | 66.1 | 88.4 |
| | F1 | 97.1 | 96.5 | 92.0 | 94.9 | 96.5 | 95.8 | 96.0 | 96.1 | 96.5 | 96.2 | 94.8 | 90.8 | 86.7 | 93.6 | 75.4 | 82.9 | 75.3 | 81.9 | 76.3 | 90.3 |
| BiTRA (Ours) | bAcc | 99.1 | 99.1 | 99.1 | 99.1 | 99.1 | 99.1 | 99.1 | 98.9 | 99.0 | 99.0 | 98.9 | 98.3 | 99.1 | 99.1 | 86.2 | 99.1 | 99.1 | 99.1 | 99.1 | **98.3** |
| | Recall | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 98.7 | 100.0 | 100.0 | 74.2 | 100.0 | 100.0 | 100.0 | 100.0 | **98.6** |
| | F1 | 99.1 | 99.1 | 99.1 | 99.1 | 99.1 | 99.1 | 99.1 | 98.9 | 99.0 | 99.0 | 98.9 | 98.3 | 99.1 | 99.1 | 84.3 | 99.1 | 99.1 | 99.1 | 99.1 | **98.2** |

Table 9: Detailed bAcc, Recall, and F1 results (%) on ViF-Bench across twelve T2V and seven I2V generators. Each method occupies three rows, and our BiTRA results are highlighted in gray.

# A Full Experimental Results

We provide complete experimental results using **Balanced Accuracy (bAcc)**, **Recall**, and **F1 score**. Unless otherwise specified, all evaluation protocols and implementation settings follow those described in the main paper.

## A.1 Complete Benchmark Results

We provide complete bAcc, Recall, and F1 results on all three benchmarks. Tab. 9 covers twelve T2V and seven I2V generators on ViF-Bench, Tab. 10 shows results across ten GenVideo generators, and Tab. 11 presents results on the in-distribution (ID) GenBuster test set and the out-of-distribution (OOD) GenBuster++ test set under different training-data settings. Across these benchmarks, BiTRA consistently achieves the strongest overall performance. Its gains are particularly evident on I2V generators and the GenBuster++ OOD test set, further demonstrating balanced cross-paradigm detection and strong generalization.

## A.2 Additional Ablation Results

The main paper summarizes the principal findings of the ablation studies. Here, we provide complete results for the detailed architectural choices, layer configurations, and training strategies. Unless otherwise specified, all experiments are conducted on ViF-Bench using Qwen3-VL-Embedding-2B, with bAcc, Recall, and F1 score as metrics. **For Selective Hierarchical Artifact Injection (SHAI)**, as shown in Tab.12, bidirectional propagation outperforms its unidirectional variants in Hierarchical Forgery Abstraction (HFA), and cross-attention performs better than Deep-VRM and sequence concatenation for global artifact injection. Notably, although position-wise injection, the paradigm shared by DeepStack (Bai et al., 2025) and Deep-VRM (Lin et al., 2026), improves over the baseline, it underperforms even sequence concatenation and lags far behind cross-attention, suggesting that simply enlarging the injection capacity without allocating features according to the temporal evolution of forgery artifacts cannot resolve the misalignment. Tab. 13(a) further shows that the temporal-barycenter objective in $\mathcal{L}_{\text{sel}}$ outperforms PAM, Q-PAD (Cuturi et al., 2019), and the APR ranking objective (Rios et al., 2021). **For Hyper-parameters**, Tab. 13(b) and Tab. 14 show that Hierarchical Forgery Abstraction (HFA), Global Artifact Injection (GAI) and Temporal Attention Smoother (TAS) perform best using four uniformly distributed layers. Accordingly, we set the number of ViT extraction layers $L_1$ and LLM injection layers $L_2$ to 4. **For Training Strategies**, Tab. 15 demonstrates that jointly adapting the visual encoder and LLM outperforms adapting either component individually. Among the evaluated strategies, applying LoRA (Hu et al., 2022) to the visual

| Method | Metric | Video Generator | | | | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sora | Morph Studio | Gen2 | HotShot | LaVie | Show-1 | Moon Valley | Crafter | Model Scope | Wild Scrape | |
| NSG-VD | bAcc | 81.7 | 98.3 | 96.7 | 91.7 | 90.8 | 88.3 | 95.8 | 94.2 | 88.4 | 88.8 | 91.5 |
| | Recall | 68.3 | 98.3 | 100.0 | 92.5 | 87.5 | 80.0 | 98.3 | 94.2 | 78.6 | 82.5 | 88.0 |
| | F1 | 78.9 | 98.3 | 96.8 | 91.7 | 90.5 | 87.3 | 95.9 | 94.2 | 87.1 | 88.0 | 90.9 |
| ReStraV | bAcc | 59.3 | 91.9 | 58.0 | 80.6 | 56.2 | 61.3 | 67.1 | 88.5 | 78.9 | 62.2 | 70.4 |
| | Recall | 68.0 | 100.0 | 63.6 | 100.0 | 49.7 | 65.1 | 61.4 | 99.8 | 99.1 | 48.4 | 75.5 |
| | F1 | 62.6 | 92.5 | 60.2 | 83.8 | 53.2 | 62.7 | 65.1 | 89.6 | 82.4 | 56.2 | 70.8 |
| STALL | bAcc | 78.6 | 75.4 | 85.9 | 71.8 | 80.4 | 77.5 | 83.7 | 80.2 | 73.7 | 71.1 | 77.8 |
| | Recall | 78.6 | 76.4 | 96.3 | 68.4 | 84.0 | 79.4 | 88.5 | 83.6 | 71.3 | 66.3 | 79.3 |
| | F1 | 78.6 | 75.7 | 87.2 | 70.8 | 81.1 | 77.9 | 84.5 | 80.9 | 73.1 | 69.6 | 77.9 |
| D3 | bAcc | 79.5 | 78.6 | 77.7 | 77.6 | 79.0 | 79.9 | 78.4 | 79.0 | 76.6 | 78.5 | 78.5 |
| | Recall | 69.6 | 71.9 | 70.1 | 70.0 | 72.1 | 73.7 | 72.2 | 72.1 | 70.3 | 71.5 | 71.4 |
| | F1 | 77.2 | 77.0 | 75.9 | 75.7 | 77.4 | 78.5 | 76.9 | 77.5 | 75.1 | 76.8 | 76.8 |
| DeMamba | bAcc | 88.8 | 90.5 | 89.5 | 94.6 | 90.4 | 91.2 | 89.8 | 90.7 | 89.1 | 90.0 | 90.5 |
| | Recall | 91.3 | 94.7 | 93.1 | 98.2 | 94.3 | 95.3 | 93.8 | 94.7 | 93.7 | 94.0 | 94.3 |
| | F1 | 89.1 | 90.9 | 89.9 | 94.8 | 90.8 | 91.6 | 90.2 | 91.1 | 89.6 | 90.4 | 90.8 |
| CoCoDetect | bAcc | 91.1 | 86.6 | 87.2 | 85.7 | 85.0 | 86.7 | 86.5 | 86.7 | 87.0 | 82.5 | 86.5 |
| | Recall | 90.4 | 99.4 | 99.3 | 97.3 | 98.3 | 97.6 | 99.7 | 99.0 | 98.4 | 89.3 | 96.9 |
| | F1 | 91.0 | 88.1 | 88.6 | 87.2 | 86.8 | 88.0 | 88.1 | 88.1 | 88.3 | 83.6 | 87.8 |
| AIGVDetect | bAcc | 90.9 | 98.4 | 99.6 | 94.3 | 99.2 | 96.0 | 99.3 | 99.3 | 91.3 | 85.6 | 95.4 |
| | Recall | 82.1 | 97.1 | 99.5 | 89.0 | 98.8 | 92.3 | 99.1 | 99.1 | 83.0 | 71.6 | 91.2 |
| | F1 | 65.3 | 95.8 | 98.4 | 91.5 | 98.0 | 93.3 | 96.5 | 98.2 | 88.0 | 81.5 | 90.6 |
| BusterX++ | bAcc | 83.7 | 92.8 | 93.3 | 88.9 | 91.4 | 92.3 | 92.1 | 92.6 | 92.4 | 87.7 | 90.7 |
| | Recall | 92.3 | 98.4 | 99.6 | 92.6 | 96.9 | 97.2 | 99.5 | 99.0 | 97.6 | 90.2 | 96.3 |
| | F1 | 85.0 | 93.2 | 93.7 | 89.3 | 91.8 | 92.7 | 92.6 | 93.1 | 92.8 | 88.0 | 91.2 |
| VideoVeritas | bAcc | 56.3 | 78.1 | 79.0 | 73.9 | 75.6 | 76.1 | 80.0 | 77.0 | 78.1 | 73.6 | 74.8 |
| | Recall | 60.7 | 95.9 | 99.6 | 87.4 | 92.9 | 92.0 | 99.8 | 95.8 | 96.0 | 87.0 | 90.7 |
| | F1 | 58.1 | 81.4 | 82.6 | 77.0 | 79.2 | 79.4 | 83.3 | 80.7 | 81.5 | 76.7 | 78.0 |
| Skyra | bAcc | 79.9 | 94.4 | 96.1 | 88.5 | 83.5 | 95.3 | 95.9 | 94.3 | 95.5 | 86.6 | 91.0 |
| | Recall | 66.3 | 95.9 | 97.9 | 82.7 | 72.0 | 96.2 | 98.1 | 94.3 | 95.6 | 78.6 | 87.8 |
| | F1 | 76.8 | 94.5 | 96.2 | 87.8 | 81.4 | 95.3 | 96.0 | 94.3 | 95.5 | 85.4 | 90.3 |
| BiTRA (Ours) | bAcc | 95.5 | 98.6 | 98.5 | 98.1 | 97.8 | 97.4 | 98.6 | 98.5 | 98.6 | 95.4 | **97.7** |
| | Recall | 92.9 | 100.0 | 100.0 | 99.1 | 98.6 | 97.7 | 100.0 | 100.0 | 100.0 | 94.1 | **98.2** |
| | F1 | 95.4 | 98.6 | 98.5 | 98.2 | 97.8 | 97.4 | 98.7 | 98.5 | 98.6 | 95.4 | **97.7** |

Table 10: Complete bAcc, Recall, and F1 results (%) on GenVideo across ten video generators. Each method occupies three rows, and our BiTRA results are highlighted in gray.

| Method | GenBuster Test (ID) | | | GenBuster++ (OOD) | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|
| | bAcc | Recall | F1 | bAcc | Recall | F1 | bAcc | Recall | F1 |
| NSG-VD | 61.4 | 71.3 | 64.8 | 53.3 | 76.0 | 61.9 | 57.3 | 73.7 | 63.4 |
| ReStraV | 58.6 | 22.7 | 35.4 | 54.5 | 24.0 | 34.5 | 56.5 | 23.4 | 35.0 |
| STALL | 69.5 | 73.1 | 70.5 | 72.3 | 68.6 | 71.2 | 70.9 | 70.9 | 70.9 |
| D3 | 60.3 | 64.7 | 62.0 | 50.5 | 46.1 | 48.2 | 55.4 | 55.4 | 55.1 |
| DeMamba | 91.2 | 92.9 | 91.4 | 86.1 | 77.6 | 84.8 | 88.7 | 85.3 | 88.1 |
| CoCoDetect | 94.4 | 91.2 | 94.2 | 50.0 | **100.0** | 66.7 | 72.2 | 95.6 | 80.4 |
| AIGVDetect | 84.3 | 84.3 | 91.5 | 69.6 | 69.8 | 69.6 | 76.9 | 77.1 | 80.6 |
| BusterX++ | 79.2 | **97.9** | 83.1 | 81.7 | 95.7 | 84.3 | 80.4 | **96.8** | 83.7 |
| Skyra | 43.4 | 33.0 | 36.8 | 58.9 | 31.6 | 43.4 | 51.1 | 32.3 | 40.1 |
| BiTRA (Ours) | **94.9** | 93.4 | **94.8** | **89.8** | 81.8 | **88.9** | **92.3** | 87.6 | **91.8** |
| VideoVeritas* | 95.7 | 99.2 | 95.8 | 91.4 | 94.9 | 91.7 | 93.6 | 97.1 | 93.8 |
| BiTRA* (Ours) | **98.0** | **99.4** | **98.0** | **94.3** | **98.1** | **94.5** | **96.1** | **98.8** | **96.2** |

Table 11: Complete performance comparison (%) on the GenBuster test set (ID) and the GenBuster++ test set (OOD). * denotes methods that use additional training data following VideoVeritas (Tan et al., 2026), while all other methods are trained only on the GenBuster training set as default. The best result within each training-data setting is highlighted in bold.

encoder and fully fine-tuning the LLM achieves the best performance and is adopted in our main experiments as default.

### A.3 Prior-Aligned Supervision Variants

We compare the temporal-barycenter objective adopted in our proposed Prior-Aligned Supervision with three alternative objectives. For clarity, we omit the batch dimension and denote the normalized frame-wise injection distribution at the $j$-th LLM injection layer by $\mathbf{p}_j^v = \{p_{j,t}^v\}_{t=0}^{T-1}$. All objectives are applied only to I2V samples. Let $K = T/2$ and divide the frames into the early and late sets:

$$\mathcal{E} = \{0, \ldots, K-1\}, \qquad \mathcal{L} = \{K, \ldots, T-1\}. \tag{10}$$

| Propagation | bAcc | Recall | F1 |
|---|---|---|---|
| Top-down | 97.4 | 97.4 | 97.2 |
| Bottom-up | 97.8 | 97.5 | 97.6 |
| Bidirectional | **98.3** | **98.6** | **98.2** |

**(a)** Propagation strategy in HFA.

| Injection | bAcc | Recall | F1 |
|---|---|---|---|
| Deep-VRM | 91 | 88.4 | 90.3 |
| Concat. | 93.8 | 96.1 | 93.6 |
| Cross-Attn. | **98.3** | **98.6** | **98.2** |

**(b)** Global artifact injection strategy.

Table 12: Detailed ablation studies of SHAI on ViF-Bench. (a) Bottom-up and Top-down denote shallow-to-deep and deep-to-shallow feature propagation in Hierarchical Forgery Abstraction (HFA), respectively. (b) Deep-VRM and Concat. denote deep residual injection and sequence concatenation in Global Artifact Injection (GAI), respectively. All metrics are reported in percentage (%).

**Progressive Attention-Mass (PAM).** PAM compares the total injection mass assigned to the early and late halves of the video:

$$P_j^{\text{early}} = \sum_{t\in\mathcal{E}} p_{j,t}^v, \qquad P_j^{\text{late}} = \sum_{t\in\mathcal{L}} p_{j,t}^v. \tag{11}$$

It penalizes the model when the total late-frame mass is smaller than the early-frame mass:

$$\mathcal{L}_{\text{PAM}} = \frac{1}{L_2}\sum_{j=1}^{L_2}\left[P_j^{\text{early}} - P_j^{\text{late}}\right]_+^2, \tag{12}$$

where $[x]_+ = \max(0, x)$. This objective provides a coarse constraint on the relative attention mass of the two temporal halves without explicitly modeling the distribution within each half.

**Quantile-Wise Progressive Attention Dominance (Q-PAD).** Q-PAD compares the distributions of early- and late-frame injection weights through quantile-wise matching. Specifically, we sort the two sets of weights as

$$e_{j,(1)} \le \cdots \le e_{j,(K)}, \qquad l_{j,(1)} \le \cdots \le l_{j,(K)}, \tag{13}$$

where $\{e_{j,(k)}\}_{k=1}^K$ and $\{l_{j,(k)}\}_{k=1}^K$ are the sorted injection weights from $\mathcal{E}$ and $\mathcal{L}$, respectively. We then require each late-frame quantile to dominate its corresponding early-frame quantile:

$$\mathcal{L}_{\text{Q-PAD}} = \frac{1}{L_2 K}\sum_{j=1}^{L_2}\sum_{k=1}^{K}\left[e_{j,(k)} - l_{j,(k)}\right]_+^2. \tag{14}$$

Compared with PAM, Q-PAD considers the distribution of injection weights rather than only their total mass.

**All-to-All Progressive Ranking (APR).** APR imposes a stronger pairwise ranking constraint by requiring every late-frame weight to be no smaller than every early-frame weight:

$$\mathcal{L}_{\text{APR}} = \frac{1}{L_2 K^2}\sum_{j=1}^{L_2}\sum_{l\in\mathcal{L}}\sum_{e\in\mathcal{E}}\left[p_{j,e}^v - p_{j,l}^v\right]_+^2. \tag{15}$$

This all-to-all constraint explicitly enforces late-frame dominance but can be overly restrictive because it requires every early–late frame pair to satisfy the prescribed ordering.

Here, $[x]_+ = \max(0, x)$. In contrast to these variants, our temporal-barycenter objective constrains the first temporal moment of the complete injection distribution using the normalized frame coordinates. It therefore preserves continuous temporal-position information without relying on a hard early–late partition or enforcing all pairwise orderings. As shown in Tab. 13(a), the temporal-barycenter objective achieves the best bAcc, Recall, and F1 performance among all compared variants.

## B Temporal Input and Attention Variants

We conduct additional experiments to investigate how temporal input organization and attention masking affect the model's utilization of forgery cues distributed across video frames. Specifically, we examine frame dropping, forward-reverse embedding fusion, and bidirectional attention as alternative strategies to further validate the effectiveness of BiTRA.

| Objective | bAcc | Recall | F1 |
|---|---|---|---|
| PAM | 97.2 | 95.7 | 96.7 |
| Q-PAD | 97.3 | 96.5 | 96.9 |
| APR | 97.0 | 97.3 | 96.8 |
| Temporal barycenter | **98.3** | **98.6** | **98.2** |

**(a)** Prior-aligned supervision objective.

| Layers | bAcc | Position | bAcc |
|---|---|---|---|
| 1 | 96.8 | Early | 97.3 |
| 2 | 97.8 | Late | 97.8 |
| 4 | **98.3** | Uniform | **98.3** |
| 8 | 97.8 | – | – |

**(b)** Number and placement of TAS layers.

Table 13: Detailed comparisons of (a) different objectives for prior-aligned injection supervision and (b) the number and placement of LLM layers selected for TAS on ViF-Bench.

| Configuration | bAcc | Recall | F1 |
|---|---|---|---|
| *Number of ViT levels* | | | |
| 1 | 95.6 | 97.1 | 95.3 |
| 2 | 97.0 | 97.3 | 96.8 |
| 4 | **98.3** | **98.6** | **98.2** |
| 8 | 97.5 | 97.5 | 97.2 |
| *ViT-level placement* | | | |
| Early | 97.4 | 97.3 | 97.1 |
| Late | 96.8 | 95.6 | 96.4 |
| Uniform | **98.3** | **98.6** | **98.2** |

**(a)** ViT-level configuration for HFA.

| Configuration | bAcc | Recall | F1 |
|---|---|---|---|
| *Number of LLM levels* | | | |
| 1 | 94.0 | 98.2 | 94.1 |
| 2 | 96.9 | 95.8 | 96.5 |
| 4 | **98.3** | **98.6** | **98.2** |
| 8 | 97.5 | 97.5 | 97.2 |
| *Injection-layer placement* | | | |
| Early | 97.8 | 96.3 | 97.4 |
| Late | 97.6 | 97.1 | 97.3 |
| Uniform | **98.3** | **98.6** | **98.2** |

**(b)** LLM-level configuration for GAI.

Table 14: Effects of layer number and placement on ViF-Bench: (a) ViT levels used for hierarchical artifact abstraction, and (b) LLM layers used for global artifact injection. All metrics are reported in percentage (%).

### B.1 Frame Dropping

**Experimental Setup.** We investigate whether the cross-paradigm performance disparity can be attributed to the temporal positions of the retained frames, and whether discarding early frames can mitigate the weakness of MLLM-based detectors on I2V videos. Specifically, we train and evaluate Qwen3-VL-Embedding-2B under five matched input settings: using the complete video, dropping the first $k$ frames, and dropping the last $k$ frames, where $k \in \{1, 2\}$. The same frame-dropping rule is applied during both training and inference. We report the overall bAcc together with the T2V and I2V Recalls.

**Analysis.** As shown in Tab. 16, dropping trailing frames causes a moderate decrease in overall bAcc and reduces I2V Recall from 81.4% to approximately 78%, as removing these frames reduces discards the most discriminative forgery cues. In contrast, dropping leading frames substantially improves I2V Recall, reaching 92.3% when the first two frames are removed. This result confirms that the artifact-poor early frames strongly interfere with I2V detection.

Despite the improvement in I2V Recall, dropping leading frames reduces the overall bAcc to below 79%, indicating a substantial increase in false positives on real videos. One possible explanation is that the first frame of a video typically corresponds to, or is located near, an intra-coded frame (I-frame), which provides relatively complete and stable appearance information without relying on inter-frame prediction. Such information can serve as an important cue for recognizing authentic videos. Removing the leading frames therefore weakens these stable real-content cues and leaves a larger proportion of inter-predicted frames, whose compression and motion-compensation irregularities may be confused with forgery artifacts. Consequently, the detector becomes more biased toward fake predictions, reducing real-video specificity and overall bAcc.

**Conclusion.** The degraded I2V performance primarily results from overemphasizing artifact-poor early frames, while later frames contain stronger forgery cues. However, simply discarding early frames improves I2V Recall at the cost of real-video specificity and overall bAcc, making frame dropping an ineffective solution. This motivates preserving the complete video while re-aligning attention with the temporal evolution of forgery artifacts.

| ViT | LLM | bAcc | Recall | F1 |
|---|---|---|---|---|
| Adapted | Frozen | 88.3 | 90.6 | 87.7 |
| Frozen | Adapted | 94.6 | 93.7 | 94.3 |
| Adapted | Adapted | **98.3** | **98.6** | **98.2** |
| LoRA | Full | **98.3** | **98.6** | **98.2** |
| Full | LoRA | 94.1 | 94.5 | 93.9 |
| LoRA | LoRA | 95.3 | 97.8 | 95.2 |
| Full | Full | 97.2 | 98.3 | 97.1 |

Table 15: Comparison of component adaptation and fine-tuning strategies on ViF-Bench. The upper block determines which components are adapted, while the lower block compares their adaptation strategies.

| Input Setting | bAcc | Recall | |
|---|---|---|---|
| | | T2V | I2V |
| Full video | **89.2** | 98.2 | 81.4 |
| Drop last ($k = 1$) | 87.8 | 98.7 | 78.5 |
| Drop last ($k = 2$) | 85.6 | 97.7 | 78.2 |
| Drop first ($k = 1$) | 77.8 | 98.6 | 86.7 |
| Drop first ($k = 2$) | 78.1 | **99.4** | **92.3** |

Table 16: Experimental results of frame dropping analysis on ViF-Bench. bAcc is computed over the complete test set, while T2V and I2V results are reported using Recall. Each detector is trained and evaluated with the same input setting. All results are reported in percentage (%), with the best result in each column highlighted in bold.

### B.2 Forward–Reverse Video-Embedding Fusion

**Experimental Setup.** We investigate whether jointly incorporating chronological and reversed temporal information can improve video-level forgery representations. Let $\mathcal{V} = (x_1, \ldots, x_T)$ denote a video containing $T$ uniformly sampled frames, where $x_t \in \mathbb{R}^{H\times W\times 3}$ is the RGB frame at temporal position $t$, and $H$ and $W$ denote its spatial resolution. From $\mathcal{V}$, we construct the chronological and reversed frame sequences as

$$\mathcal{V}_{\text{fwd}} = (x_1, \ldots, x_T), \qquad \mathcal{V}_{\text{rev}} = (x_T, \ldots, x_1). \tag{16}$$

The two sequences are independently processed by the same Qwen3-VL-Embedding-2B model. For each forward pass, we use the final-layer representation of the last token as the video-level embedding:

$$\mathbf{z}_{\text{fwd}} = E_{\text{Qwen}}(\mathcal{V}_{\text{fwd}}), \qquad \mathbf{z}_{\text{rev}} = E_{\text{Qwen}}(\mathcal{V}_{\text{rev}}). \tag{17}$$

Here, $E_{\text{Qwen}}(\cdot)$ denotes Qwen3-VL-Embedding-2B followed by last-token feature extraction, while $\mathbf{z}_{\text{fwd}}, \mathbf{z}_{\text{rev}} \in \mathbb{R}^d$ denote the resulting chronological and reversed video embeddings, respectively, with $d$ being the embedding dimension.

We then fuse the chronological and reversed embeddings using a residual fusion module. Specifically, the two embeddings are first concatenated and transformed by a feed-forward network (FFN), after which the transformed feature is added to both original embeddings:

$$\begin{aligned} \mathbf{z}_{\text{cat}} &= [\mathbf{z}_{\text{fwd}}; \mathbf{z}_{\text{rev}}], \\ \mathbf{z}_{\text{fuse}} &= \mathbf{z}_{\text{fwd}} + \mathbf{z}_{\text{rev}} + \text{FFN}(\mathbf{z}_{\text{cat}}), \\ \hat{y} &= \sigma(\mathbf{W}_c \mathbf{z}_{\text{fuse}} + b_c). \end{aligned} \tag{18}$$

Here, $[\cdot;\cdot]$ denotes feature concatenation, such that $\mathbf{z}_{\text{cat}} \in \mathbb{R}^{2d}$. The fusion network $\text{FFN} : \mathbb{R}^{2d} \to \mathbb{R}^d$ projects the concatenated representation back to the original embedding dimension. Consequently, $\mathbf{z}_{\text{fuse}} \in \mathbb{R}^d$ contains both the original directional representations and their learned interactions. Finally, $\mathbf{W}_c \in \mathbb{R}^{1\times d}$ and $b_c \in \mathbb{R}$ are the learnable weight and bias of the binary classifier, respectively; $\sigma(\cdot)$ denotes the sigmoid function; and $\hat{y} \in (0, 1)$ is the predicted probability that the input video is fake. This strategy operates entirely at the video-embedding level. It requires two

| Method | bAcc | | | Recall | | |
|---|---|---|---|---|---|---|
| | Overall | T2V | I2V | Overall | T2V | I2V |
| Baseline | 89.2 | 92.3 | 84.0 | 92.0 | 98.2 | 81.4 |
| Fwd.–Rev. Fusion | 94.3 | 95.8 | 91.5 | 95.6 | 98.9 | 89.9 |
| **BiTRA (Ours)** | **98.3** | **99.0** | **97.2** | **98.6** | **99.9** | **96.3** |

Table 17: Comparison of forward–reverse video-embedding fusion and BiTRA on ViF-Bench. For forward–reverse fusion, the chronological and reversed frame sequences are independently encoded by Qwen3-VL-Embedding-2B, and their final-layer last-token embeddings are fused for prediction. All results are reported in percentage (%), with the best result in each column highlighted in bold.

| Visual Attention | bAcc | | | Recall | |
|---|---|---|---|---|---|
| | Overall | T2V | I2V | T2V | I2V |
| Causal | **89.2** | **92.3** | 84.0 | **98.2** | 81.4 |
| Bidirectional | 86.8 | 86.2 | **87.9** | 96.0 | **99.4** |

Table 18: Comparison between causal and bidirectional visual attention on ViF-Bench.

complete Qwen3-VL-Embedding-2B forward passes but leaves the native causal attention mechanism within each encoding pass unchanged.

**Analysis.** As shown in Tab. 17, forward–reverse embedding fusion improves the overall bAcc from 89.2% to 94.3%. The improvement is more pronounced on I2V: its bAcc increases by 7.5 points and its Recall increases by 8.5 points, whereas the corresponding T2V gains are considerably smaller. This result further demonstrates that temporal ordering has a substantial influence on I2V detection. By combining the embeddings obtained from both temporal orders, the classifier can utilize the natural chronological representation while incorporating stronger forgery cues from later I2V frames through the reversed representation.

However, this approach requires two complete forward passes through Qwen3-VL-Embedding-2B for every video, resulting in substantially higher computation and inference latency. Its overall bAcc of 94.3% also remains below the 98.3% achieved by BiTRA. In contrast, BiTRA re-aligns artifact utilization and temporal attention within a single MLLM processing pipeline, avoiding duplicated video-embedding computation while achieving better performance.

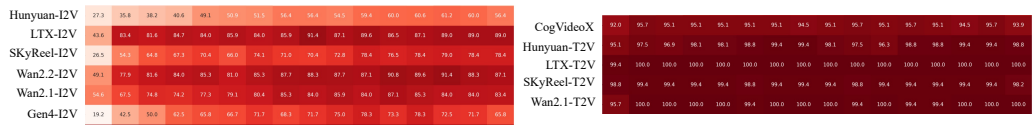
**Conclusion.** Forward–reverse video-embedding fusion confirms the importance of temporal ordering for I2V detection. However, it requires two complete Qwen3-VL-Embedding forward passes, resulting in approximately $20.3$ TFLOPs per sample, nearly $1.88\times$ the $10.8$ TFLOPs required by BiTRA. Despite this substantially higher computational cost, its detection performance remains inferior to BiTRA. These results indicate that explicitly re-aligning temporal artifacts and attention within a unified single-pass architecture provides a more accurate and computationally efficient solution.

### B.3 Causal vs. Bidirectional Visual Attention

**Experimental Setup.** We investigate whether directly removing the causal constraint among visual tokens can alleviate temporal attention bias. The baseline retains the native causal self-attention of the LLM decoder. In the bidirectional variant, we modify only the visual-to-visual portion of the attention mask, allowing every visual token to attend to visual tokens from both preceding and subsequent frames. All remaining attention-mask rules and model components are kept unchanged.

**Analysis.** As shown in Tab. 18, bidirectional visual attention improves I2V bAcc from 84.0% to 87.9% and I2V Recall from 81.4% to 99.4%. Meanwhile, the T2V–I2V bAcc gap decreases from 8.3 to 1.7 points. Allowing early-frame visual tokens to access information from later frames therefore substantially alleviates the cross-paradigm imbalance. This result further confirms that the native causal information flow restricts the utilization of the stronger forgery cues appearing in later I2V frames.

However, this improvement comes at the cost of T2V and overall performance. T2V bAcc decreases from 92.3% to 86.2%, while the overall bAcc drops from 89.2% to 86.8%. Directly replacing causal




(a) Fake Probability on I2V.

(b) Fake Probability on T2V.

Figure 6: Detailed per-generator frame-wise fake probability across real, I2V, and T2V videos on ViF-Bench. Each cell reports the mean predicted fake probability (%) at one of 16 uniformly sampled temporal positions for the corresponding video generator. Darker colors indicate higher fake probabilities. Probabilities are averaged over Effort (Yan et al., 2024) and DDA (Chen et al., 2025).

visual attention with bidirectional attention disrupts the information flow learned during MLLM pretraining and removes useful temporal directionality. Consequently, although bidirectional attention reduces the excessive preference for early frames, it also degrades the representations formed under the native causal structure.

**Conclusion.** Bidirectional visual attention alleviates the T2V–I2V imbalance but sacrifices overall detection performance by disrupting the pretrained causal information flow. This trade-off motivates BiTRA to preserve the native causal structure while smoothing excessive temporal attention bias and improving the utilization of forgery artifacts distributed throughout the video.

# C Detailed Diagnostics

The main paper summarizes the key observations leading to Temporal Artifact–Attention Misalignment. In this section, we provide more detailed analyses from two complementary perspectives: the temporal evolution of frame-wise forgery artifacts across generation paradigms and the layer-wise allocation of causal attention over video frames.

## C.1 Frame-Wise Fake Probability

**Experimental Setup.** We adopt representative SOTA AIGI detectors, *i.e.*, Effort (Yan et al., 2024) and DDA (Chen et al., 2025), with their officially released checkpoints, as frame-level forgery probes. Each detector independently processes the 16 uniformly sampled frames of every video. We first verify that the temporal trends are consistent across the two detectors, and then average their predicted fake probabilities at the same temporal position for each video generator, producing the heatmap shown in Fig. 6.

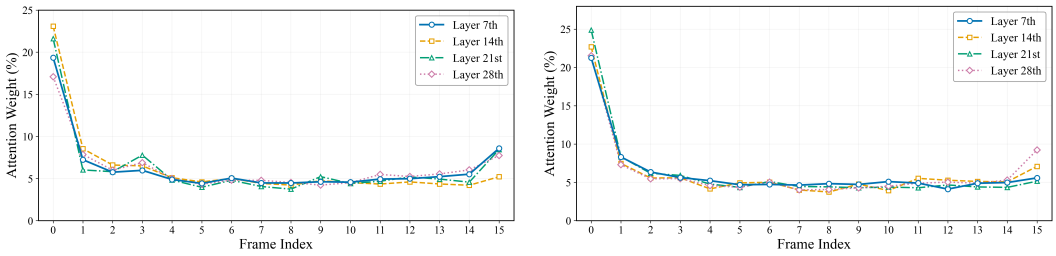
**Analysis.** As shown in Fig. 6, T2V videos generally exhibit high and temporally stable fake probabilities, indicating that their forgery artifacts are distributed relatively uniformly throughout the sequence. In contrast, I2V videos show a distinct progressive pattern: their fake probabilities are markedly lower in the initial frames and gradually increase toward later temporal positions. Although the absolute probabilities vary across generators, this temporal trend is consistently observed among different I2V models.

This difference is consistent with the conditioning mechanisms of the two generation paradigms. T2V models synthesize every frame solely from textual conditions, causing forgery artifacts to appear throughout the sequence. In contrast, I2V generation is additionally anchored by an input image, whose realistic visual content strongly influences the initial frames. As generation proceeds, this constraint gradually weakens, while generation errors and temporal inconsistencies accumulate, making the forgery artifacts increasingly distinguishable.

**Conclusion.** Real and T2V videos exhibit relatively stable frame-wise fake probabilities, whereas I2V videos show progressively increasing fake probabilities from early to later frames. This paradigm-dependent temporal artifact distribution provides the basis for diagnosing Temporal Artifact–Attention Misalignment.

## C.2 Analysis of Temporal Attention Bias

**Experimental Setup.** We analyze frame-wise attention distributions across the LLM decoder layers to determine whether the models exhibit a systematic preference for particular temporal positions. We compare two open-source pretrained MLLMs, InternVL3 and Qwen3-VL-Embedding. At each LLM layer, we use the final token as the query and extract its attention weights over all visual tokens. The weights corresponding to the visual-token interval of each frame are summed and




**(a)** InternVL3 **(b)** Qwen3-VL-Embedding

Figure 7: Layer-wise temporal attention distributions of two open-source pretrained MLLMs: (a) InternVL3 and (b) Qwen3-VL-Embedding. Each row corresponds to an LLM decoder layer, while each column represents one of the 16 sampled frames. Darker colors indicate a larger proportion of final-token attention allocated to the corresponding frame.

normalized across the 16 frames, yielding a layer-wise frame-attention distribution. The resulting distributions are further averaged over attention heads and evaluation samples.

**Analysis.** As shown in Fig. 7, both InternVL3 and Qwen3-VL-Embedding exhibit a consistent temporal attention bias across the LLM decoder layers. In both architectures, the final token allocates substantially more attention to preceding frames. This shared pattern indicates that the bias is not specific to a particular MLLM architecture, but is inherent in the native causal attention mechanism of the LLM decoder.

Such a temporal preference is particularly unfavorable for I2V detection. As shown in Sec. C.1, I2V forgery artifacts are relatively weak in early frames and become increasingly pronounced toward later frames. Excessive attention to preceding frames therefore limits the utilization of the more discriminative forgery cues emerging at later temporal positions.

**Conclusion.** Both InternVL3 and Qwen3-VL-Embedding consistently favor preceding frames across the LLM decoder layers, confirming that temporal attention bias is shared across different MLLM architectures. This result further supports our diagnosis of Temporal Artifact–Attention Misalignment in I2V detection.

## D Additional Attention Heatmap Visualizations

To qualitatively examine BiTRA's spatial and temporal attention, we visualize one example from each of ViF-Bench, GenVideo, and GenBuster++. The first two are I2V videos, while the third is a T2V video. For each example, we compare attention heatmaps of the baseline and BiTRA using identical frames and inference settings, where warmer colors indicate higher prediction-token attention to spatial visual tokens.

As shown in Fig. 8, for the two I2V examples, the baseline mainly attends to preceding frames, where forgery artifacts are relatively weak. In contrast, BiTRA progressively shifts its attention toward later frames and focuses more strongly on spatial regions containing discriminative forgery cues. This behavior is consistent with Prior-Aligned Supervision, which applies $\mathcal{L}_{\text{sel}}$ specifically to I2V samples and guides artifact selection and injection toward the more informative middle-to-late frames. For the T2V example, whose forgery artifacts are more uniformly distributed over time, BiTRA produces a correspondingly more balanced attention distribution across the video frames. This pattern reflects the effect of Temporal Attention Smoother (TAS), which mitigates excessive attention concentration on preceding frames without forcing the model to favor a particular temporal position. These results demonstrate that BiTRA neither simply shifts attention toward later frames nor imposes a fixed temporal distribution. Instead, its two components jointly adapt attention allocation to the paradigm-dependent temporal evolution of forgery artifacts.

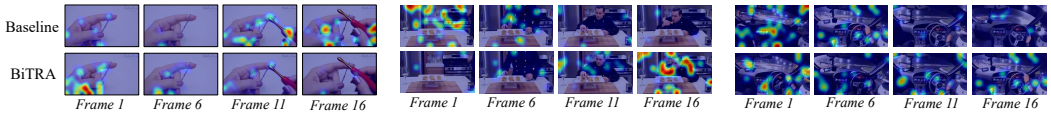



Figure 8: Attention heatmap comparisons between the baseline and BiTRA on three video examples selected from ViF-Bench , GenVideo , and GenBuster++ .