GENVIDCLIP: TEMPORAL-AWARE CLIP WITH MULTIMODAL ALIGNMENT FOR AI-GENERATED VIDEO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancement of AI-driven video generation has substantially increased the risk of highly realistic visual forgeries. Existing video forgery detectors often model video content in a homogeneous manner, making sparse and subtle forgery traces vulnerable to being overwhelmed by large amounts of normal spatial and temporal information, which leads to temporal feature dilution. To address this issue, we propose GenVidCLIP, a temporal-aware three-branch forgery detection framework built upon a Vision-Language Model. The core design explicitly decomposes video forgery representation into three complementary components: a Spatial Semantic Stream (SSS) for capturing global semantic context, a Temporal Kinematic Stream (TKS) for identifying sparse high-response temporal artifacts, and a text semantic branch for providing learnable authenticity-aware semantic priors. To effectively integrate these heterogeneous representations, we further introduce a Contrastive Residual Gating Network (CRGN), which models the divergence between spatial and temporal features and selectively injects localized temporal evidence into the global spatial representation before semantic alignment with text prototypes. Experiments on GenVidBench-6M show that GenVidCLIP achieves 85.05% ACC, 90.29% F1, 87.74% TNR on HD real videos, and 99.59% AUROC. On the cross-dataset GenVideo many-to-many protocol, it obtains 97.46% AP and 93.79% F1. Additional evaluations on Celeb-DF and AEGIS-Hard demonstrate transfer beyond the primary benchmark while revealing remaining challenges under diverse manipulation and generation distributions.

1 INTRODUCTION

Recently, generative Artificial Intelligence (AI) technologies like Sora Brooks et al. (2024) and Gen-2 Runway (2023) have achieved breakthrough progress in spatial coherence, temporal continuity, and resolution. These advancements enable the synthesis of AI-Generated Content (AIGC) with extreme visual fidelity, blurring the boundary between reality and forgery. However, the widespread accessibility and abuse of this technology precipitate severe risks of misinformation dissemination. Faced with increasingly realistic videos, developing robust detection models with strong cross-generator generalization has become an urgent proposition in multimedia security.

Nevertheless, existing detection methods exhibit significant limitations against such highly realistic content. Current paradigms primarily fall into two categories. The first focuses on spatial artifact mining Li et al. (2020a), which largely fails against AIGC videos characterized by highly smoothed, realistic spatial textures. The second directly employs 3D CNNs Carreira and Zisserman (2017) or long-sequence spatiotemporal attention Bertasius et al. (2021) to conduct homogeneous modeling over entire video clips. Since advanced AIGC videos often contain only sparse temporal artifacts, global sequence modeling can drown these discriminative signals in redundant background content, resulting in severe temporal feature dilution Zhao et al. (2021) across long video clips.

To address these bottlenecks, we propose GenVidCLIP, a temporal-aware three-branch AI-generated video detector built on Vision-Language Models (VLMs) Radford et al. (2021) with explicit spatial-temporal decomposition at the visual representation level.

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.

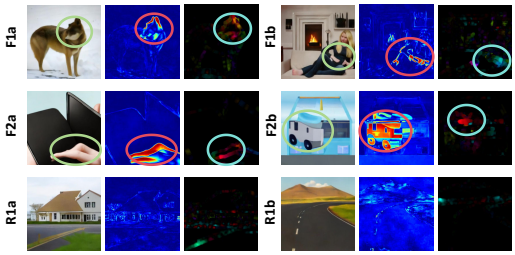

Figure 1: **Motivation for explicit decoupling.** Rows 1–2: AI-generated videos whose kinematic representations expose temporal anomalies (circled); row 3: pristine real videos. Each group shows raw RGB, L1 difference, and optical flow.

The first core innovation lies in explicitly decoupling global spatial semantics from microscopic spatiotemporal artifacts at the visual level. As demonstrated in Figure 1, while raw RGB frames possess deceptive spatial realism, their corresponding kinematic representations uniquely expose temporal anomalies (highlighted by colored circles) that remain entirely invisible to standard spatial branches, contrasting sharply with the stable patterns of real videos. To exploit this, we design a dual-stream visual architecture consisting of a Spatial Semantic Stream (SSS) and a Temporal Kinematic Stream (TKS). Instead of treating the video as a homogeneous sequence, the TKS employs a hybrid kinematic sampling strategy—combining optical flow and L1 pixel differences—to explicitly lock onto high-frequency dynamic inconsistencies. By isolating these critical moments, GenVidCLIP fundamentally prevents sparse forgery traces from being diluted by the massive volume of normal background frames.

Furthermore, simply concatenating spatial and temporal features often leads to semantic conflicts. To address this, we introduce the Contrastive Residual Gating Network (CRGN), which dynamically evaluates the divergence between the SSS and TKS, adaptively injecting localized temporal anomalies into the global spatial context. We also leverage the zero-shot capabilities of VLMs through an implicit semantic alignment mechanism that constrains fused visual representations within a shared latent space guided by continuous text prompts, thereby calibrating the model against visually realistic yet synthetic content during authenticity-aware discrimination.

In summary, our main contributions are as follows:

1. We propose GenVidCLIP, a temporal-aware three-branch architecture that explicitly decouples spatial semantics from temporal artifacts. By capturing highfrequency temporal irregularities, the temporal branch substantially improves the models sensitivity to forgery cues.
2. We introduce the Contrastive Residual Gating Network, which dynamically injects localized temporal anomalies into global spatial features through learned gating, while an implicit semantic alignment strategy effectively overcomes the bottleneck of crossmodal alignment.
3. Extensive experiments on GenVidBench-6M and GenVideo demonstrate competitive detection and cross-generator transfer. GenVidCLIP achieves 85.05% ACC, 90.29% F1, 87.74% HD-real TNR, and 99.59% AUROC on GenVidBench-6M, while Celeb-DF and AEGIS-Hard further characterize cross-task robustness across distinct manipulation distributions.

2 Related Work

2.1 Spatiotemporal Video Forgery Detection

The trajectory of AI-generated video detection has evolved from static image-level analysis to global spatiotemporal modeling and multimodal semantic alignment. Early research primarily targeted microscopic flaws in the spatial domain, analyzing frequency anomalies or local edge inconsistencies Li et al. (2020a); Frank et al. (2020). While these approaches exhibited high sensitivity to early face-swapping technologies Li and Lyu (2019), their effectiveness is severely constrained today, as the textures of modern AI-generated content (AIGC) have become increasingly seamless. To overcome the limitations of single-frame detection, researchers introduced 3D CNNs like I3D Carreira and Zisserman (2017) and SlowFast Feichtenhofer et al. (2019), followed by Video Transformers such as TimeSformer Bertasius et al. (2021) and VideoMAE Tong et al. (2022), and more recently State Space Models (SSMs) like DeMamba Chen et al. (2024) for processing long sequences with linear complexity. Although these architectures excel at extracting global motion, they encounter bottlenecks when facing highly realistic AIGC videos. Because forgery traces can be sparse and subtle, homogeneous aggregation over long clips may attenuate localized discriminative signals. We treat this as a motivating hypothesis rather than strict causal evidence, and complement the qualitative observation with dataset-level motion statistics in Appendix E. More recently, several works have

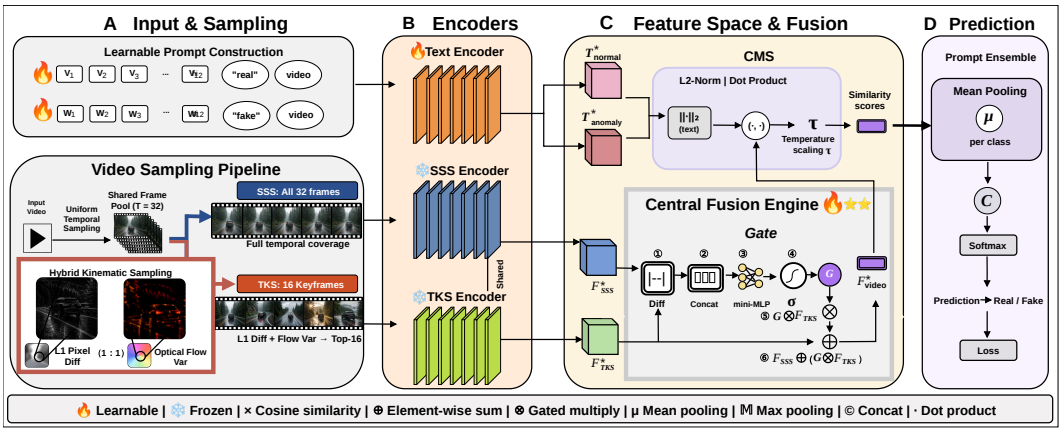

Figure 2: Framework overview of our proposed GenVidCLIP method. In the Central Fusion Engine, F^*_{SSS} denotes the peak anomaly signal extracted from the SSS encoder output via max pooling, and F^*_{TKS} denotes the corresponding semantic context features from the TKS encoder. G is the dynamic gating mask produced by the CRGN module. F_{video} is the final fused visual representation fed into the cross-modal similarity module for final prediction.

explored frequency-domain and reconstruction-based cues for AIGC detection Tan et al. (2024), as well as multi-scale temporal attention mechanisms. Despite this progress, none of these methods explicitly decouple kinematic artifacts from global spatial semantics prior to fusion, motivating the explicit kinematic-spatial decomposition adopted in TKS.

2.2 Multimodal Semantic Alignment

Furthermore, to enhance zero-shot generalization, introducing cross-modal semantic priors from Vision-Language Models (VLMs) like CLIP Radford et al. (2021) has emerged as a key strategy. Representative works typically employ hard prompts and cross-attention mechanisms for feature fusion Ojha et al. (2023) to provide high-order semantic constraints Zhou et al. (2022). Nevertheless, static prompts struggle to cover highly variable AIGC distributions, and dense explicit attention can inadvertently disrupt the native manifold of the pre-trained VLM Wortsman et al. (2022). To circumvent these drawbacks, GenVidCLIP employs multi-stage continuous prompt learning to dynamically adapt to unknown generative distributions. By performing minimalist implicit alignment within a shared latent space, our design maximizes the utility of textual priors while strictly preserving the original representation space of the pre-trained VLM.

3 Method

3.1 Task Formulation and Architecture Overview

Given an input video sequence $\mathbf{V}$ and a set of text prompts $\mathbf{P}$ describing forgery concepts, the multimodal AI-generated video detection task aims to learn a discriminative network. By mining local anomalous traces in the spatiotemporal dimensions of $\mathbf{V}$ and aligning them with the semantic priors from $\mathbf{P}$, the network outputs a predictive probability to accurately classify $\mathbf{V}$ as either real or AI-generated. To overcome the prevalent temporal feature dilution bottleneckwhere weak, highly sparse forgery traces are easily submerged by overwhelming normal backgroundsGenVidCLIP is designed as an end-to-end cascaded architecture with two collaborative stages (Figure 2).

In the first stage (*Multimodal Feature Extraction*), the SSS uniformly samples frames to capture macroscopic context F_S, the TKS extracts kinematic keyframes to produce local features F_T, and the text branch generates semantic priors T via continuous prompt learning. In the second stage (*CRGN Fusion and Alignment*), CRGN derives a dynamic gate from the difference between F_S and F_T and injects selected temporal evidence into the spatial baseline to obtain F_{video}. Cosine-similarity alignment then anchors F_{video} to the text prototypes T in a shared latent space.

3.2 THREE-BRANCH MULTIMODAL FEATURE EXTRACTION

The architecture consists of a Spatial Semantic Stream (SSS), a Temporal Kinematic Stream (TKS), and a text semantic branch. The SSS adopts a uniform sampling strategy. Given an input video sequence $V \in \mathbb{R}^{L\times C\times H\times W}$ (where L is the frame count), a set of N image frames $V_S = \{I_{s_1}, I_{s_2}, \ldots, I_{s_N}\}$ is obtained through uniform interval sampling. These frames are processed by a pre-trained ViT to extract frame-level representations, then aggregated via the `[CLS]` token to produce the spatial feature $F_S \in \mathbb{R}^D$, serving as a stable semantic baseline. By providing full temporal coverage, the SSS captures the macroscopic spatial context that anchors detection, while leaving localized anomaly extraction to the complementary TKS. This design prevents the SSS from wasting capacity on sparse artifacts that would be diluted during global aggregation.

To address high-frequency artifacts, the TKS employs a hybrid kinematic sampling strategy evaluating inter-frame kinematic responses. For adjacent frames I_t and I_{t+1}, the TKS first downsamples them into grayscale tensors $Y_t, Y_{t+1} \in \mathbb{R}^{H'\times W'}$. Two complementary metrics are computed in parallel. The pixel-level mean frame difference captures local intensity mutations:

$$D_t = \frac{1}{H'W'} \sum_{x=1}^{H'} \sum_{y=1}^{W'} |Y_{t+1}(x,y) - Y_t(x,y)| \tag{1}$$

Second, to capture spatial geometric distortions from temporal incoherence, we compute the inter-frame motion magnitude via dense optical flow (u_t, v_t):

$$M_t(x,y) = \sqrt{u_t(x,y)^2 + v_t(x,y)^2} \tag{2}$$

We then compute the spatial variance of this magnitude matrix to quantify abnormal local motion dispersion across each frame and its spatial regions:

$$O_t = \frac{1}{H'W'} \sum_{x=1}^{H'} \sum_{y=1}^{W'} \left(M_t(x,y) - \mu_t\right)^2 \tag{3}$$

where μ_t represents the spatial expectation of the optical-flow magnitude at frame t. To remove scale differences between the two kinematic metrics, we apply min-max normalization $\mathcal{N}(\cdot)$ to each response sequence X before combining the two cues:

$$\mathcal{N}(X_t) = \frac{X_t - \min(X)}{\max(X) - \min(X)} \tag{4}$$

By mapping D_t and O_t into the same dimensional space, the comprehensive kinematic anomaly score S_t for the t-th frame can be defined as their linear fusion:

$$S_t = \mathcal{N}(D_t) + \mathcal{N}(O_t) \tag{5}$$

Based on the score set of the entire sequence, the TKS employs a Top-K operator to extract the indices of the K keyframes with the strongest responses. A temporal re-ordering is then performed to ensure the temporal coherence of the sampling, thereby constructing the local anomaly set $V_T = \{I_{t_1}, I_{t_2}, \ldots, I_{t_K}\}$. After the local keyframe set V_T is processed by ViT for feature extraction, it is fed into the temporal encoder of the TKS. Fundamentally different from the `[CLS]` aggregation method of the SSS, to prevent high-frequency local artifacts from being diluted by normal background features during long-term temporal modeling, the TKS forcibly applies 1D temporal max pooling after the self-attention calculation. This mechanism acts as an explicit anomaly filter, forcing the network to focus on and retain the temporal anchors with the strongest forgery detection responses in the sequence, ultimately outputting a highly discriminative temporal kinematic feature $F_T \in \mathbb{R}^D$. Finally, to provide semantic prior guidance for authenticity discrimination, the text branch directly adopts the object-agnostic prompt learning strategy from AnomalyCLIP Zhou et al. (2024), applying it without modification to the video forgery detection setting. Given that static hard prompts struggle to cover the complex and variable AI-generated video forgery distributions, this branch constructs a set of discrete language templates describing realistic attributes and temporal inconsistencies. Within the pre-trained CLIP text encoder, class-specific learnable continuous context vectors are introduced in the word embedding space. These vectors are concatenated with the discrete templates to build composite prompts, which are then processed by the text encoder to generate the global text semantic prototypes $T_c \in \mathbb{R}^D$ (where $c \in \{0, 1\}$). These prototypes act as abstract semantic anchors rather than definitive labels, equipping it with the ability to dynamically fit unknown generator distributions.

3.3 Contrastive Residual Gating Network and Semantic Alignment

After obtaining F_S from the SSS and F_T from the TKS, we fuse them via the CRGN module, which avoids disrupting the pre-trained VLM's manifold. The key insight is that pristine videos exhibit high spatiotemporal consistency, whereas AI-generated videos show significant feature divergence. The CRGN quantifies this divergence through element-wise absolute differences:

$$D_{ST} = |F_S - F_T| \tag{6}$$

where D_{ST} denotes the spatiotemporal difference matrix, F_S represents the spatial semantic feature, F_T is the local temporal kinematic feature, and $|\cdot|$ indicates element-wise absolute value. Subsequently, F_S, F_T, and D_{ST} are concatenated along the channel dimension and fed into a Multi-Layer Perceptron (MLP) to generate a dynamic gating mask $G \in (0,1)^D$:

$$G = \sigma\left(\text{MLP}([F_S \parallel F_T \parallel D_{ST}])\right) \tag{7}$$

where σ is the Sigmoid function and $\parallel$ denotes channel-wise concatenation. Finally, the mask G acts as an adaptive feature selector for the TKS features. The selected high-frequency anomalies are then superimposed as local forgery residuals onto the macroscopic semantic baseline of the SSS, yielding a highly refined comprehensive visual representation F_{video}:

$$F_{video} = F_S + G \odot F_T \tag{8}$$

where F_{video} is the final refined visual representation, and $\odot$ denotes the element-wise multiplication operation. While preserving the global semantics of the video, this mechanism precisely refines high-frequency artifacts, mitigating conflicts when fusing heterogeneous visual branches. Compared to concatenation or cross-attention alternatives, this residual gating design offers two key advantages: (1) the mask G is conditioned on the explicit divergence D_{ST}, providing a principled criterion for anomaly injection; and (2) the additive residual structure preserves the pre-trained spatial representations in F_S, preventing catastrophic feature drift during fine-tuning.

Cross-Modal Fusion via Implicit Feature Alignment. Rather than employing explicit cross-attention, we perform implicit metric alignment. L_2-normalized F_{video} and text prototypes T_c are compared via cosine similarity with a learnable temperature τ:

$$p_c = \exp(\tau) \cdot \left\langle \frac{F_{video}}{\|F_{video}\|_2}, \frac{T_c}{\|T_c\|_2} \right\rangle \tag{9}$$

where τ is a learnable temperature. This alignment forces visual artifacts to cluster toward their text anchors, achieving cross-modal integration without additional parameters.

Training Objective. The logits $\{p_c\}_{c=0}^{1}$ are supervised with label-smoothed cross-entropy:

$$\mathcal{L} = -\sum_{c=0}^{1} \tilde{y}_c \log \frac{\exp(p_c)}{\sum_j \exp(p_j)}, \qquad \tilde{y}_c = (1-\epsilon)y_c + \frac{\epsilon}{2}. \tag{10}$$

Here y_c is the one-hot label and $\tilde{y}_c$ is the smoothed target. We use CosineAnnealingLR and differential learning rates as detailed in Appendix A. At inference time, the predicted class is $\hat{y} = \arg\max_c p_c$.

4 Experiments

4.1 Comparison with State-of-the-Art Methods

Training and Evaluation Protocol. We train GenVidCLIP on a subset of only 3,200 videos from the GenVidBench-6M Pair 1 split—0.1% of the full training data (∼3.2M videos)—while CNN/Transformer baselines in Table 1 were trained on the complete split Ni et al. (2026). The full GenVidCLIP model contains approximately 470.32M parameters, of which approximately 42.38M are trainable; the remaining parameters are frozen pre-trained components. Thus, only about 9.0% of the model is optimized on this small training subset. Pair 2 is reserved for testing; for GenVideo we adopt DeMamba's Many-to-Many protocol Chen et al. (2024).

We benchmark our proposed framework against a comprehensive suite of state-of-the-art (SOTA) video forgery detection methods. Table 1 includes representative 3D CNNs Feichtenhofer et al. (2019),

Table 1: Comparison on the GenVidBench-6M cross-source and cross-generator task. Per-generator columns report fake-video recall (%); HD reports TNR on authentic HD videos. AUROC is threshold-independent. The GenVidCLIP row reports our final training setting.

Method	Type	MuseV	SVD	CogV	Mora	HD (Real)	ACC	F1	AUROC
I3D (CVPR'17) Carreira and Zisserman (2017)	CNN	32.72	12.04	76.44	72.30	95.04	60.21	67.38	84.06
TRN (ECCV'18) Zhou et al. (2018)	CNN	43.52	27.16	58.30	97.26	97.56	66.67	72.49	92.13
SlowFast (ICCV'19) Feichtenhofer et al. (2019)	CNN	87.14	29.80	93.07	55.23	96.61	70.06	81.13	88.65
TSM (ICCV'19) Lin et al. (2019)	CNN	95.94	73.16	36.44	91.72	96.30	73.88	82.49	95.31
TPN (CVPR'20) Yang et al. (2020)	CNN	56.34	24.30	71.87	73.37	99.07	66.23	73.39	95.67
TIN (AAAI'20) Shao et al. (2020)	CNN	68.38	41.44	57.88	79.67	96.41	67.91	73.74	72.96
TimeSformer (ICML'21) Bertasius et al. (2021)	Trans.	69.94	29.13	67.82	98.47	72.47	71.53	77.31	76.71
VideoSwin (CVPR'22) Liu et al. (2022)	Trans.	90.24	27.72	91.64	88.14	99.10	80.39	86.36	96.05
MViTv2-S (CVPR'22) Li et al. (2022)	Trans.	77.08	44.89	99.91	76.77	86.98	80.45	85.66	90.29
UniFormerV2 (ICCV'23) Li et al. (2023)	Trans.	27.97	12.84	94.79	78.21	99.46	65.31	73.03	96.73
TALL (ICCV'23) Xu et al. (2023)	Trans.	59.50	50.30	67.70	86.00	64.60	65.80	75.60	71.00
MINTIME (TIFS'24) Coccomini et al. (2024)	Trans.	29.80	10.80	80.90	92.00	93.70	61.80	69.40	82.30
Qwen2.5-VL (arXiv'25) Bai et al. (2025)	MLLM	25.86	27.06	68.51	43.26	71.15	74.30	55.49	96.25
GPT-4.1mini (2025) OpenAI (2025)	MLLM	26.07	33.78	94.07	57.19	87.64	59.62	66.72	95.83
VidGuard-R1 (ICLR'26) Park et al. (2026)	MLLM	36.52	16.02	99.35	76.94	99.94	66.09	72.77	91.23
GenVidCLIP	**Hybrid**	78.34	72.09	93.01	94.07	87.74	**85.05**	**90.29**	**99.59**

Video Transformers Liu et al. (2022), and recent Multimodal Large Language Models (MLLMs) including Qwen2.5-VL Bai et al. (2025), GPT-4.1mini OpenAI (2025), and VidGuard-R1 Park et al. (2026). DeMamba-based variants are retained in the GenVideo comparison in Table 2, where we follow its Many-to-Many evaluation protocol Chen et al. (2024).

Performance on GenVidBench-6M. Table 1 summarizes cross-source and cross-generator performance on GenVidBench-6M. GenVidCLIP achieves 85.05% ACC, 90.29% F1, 87.74% HD-real TNR, and 99.59% AUROC. Among the methods reported in Table 1, GenVidCLIP attains the highest overall ACC, F1, and AUROC, while maintaining strong recall across multiple generator subsets.

We also evaluate Qwen2.5-VL Bai et al. (2025), GPT-4.1mini OpenAI (2025), and VidGuard-R1 Park et al. (2026) under their native zero-shot or prompted pipelines without task-specific fine-tuning. Their lower F1 scores (55–73%) indicate that generic vision-language alignment alone does not reliably capture the subtle temporal artifacts in this benchmark. This comparison spans different evaluation paradigms and is therefore used only as a reference. VidGuard-R1, for example, attains near-perfect real-video specificity but only 16.02% recall on the dynamic SVD subset.

4.2 Cross-Dataset and Generalization Robustness

We evaluate zero-shot generalization using the strict GenVideo Many-to-Many protocol Chen et al. (2024), with the comparison against representative baselines reported in Table 2.

GenVidCLIP reaches 97.46% AP and 93.79% F1. Table 5 further reports zero-shot accuracy across nine unseen generators: GenVidCLIP obtains the highest average accuracy (0.9420) and remains strong on HotShot (0.9638), where FTCN and MINTIME drop to 0.1771 and 0.2643, respectively.

Figure 3 shows that GenVidCLIP avoids several generator-specific collapses observed in other baselines and maintains an average recall of 94.20%, providing complementary evidence of cross-generator robustness without implying universal generalization.

External robustness. We additionally evaluate GenVidCLIP without dataset-specific fine-tuning on Celeb-DF Li et al. (2020b) and AEGIS-Hard Li et al. (2025). GenVidCLIP reaches 82.4% AUROC on Celeb-DF, indicating transfer to localized face-swapping manipulations. On AEGIS-Hard, it obtains 0.68 $\mathrm{Acc}_{\mathrm{all}}$, 0.68 Macro-F1, and 0.747 AUROC. We treat these results as external robustness evidence rather than matched benchmarks for the primary task.

Protocol note. Celeb-DF evaluates localized face manipulation, whereas AEGIS-Hard is a broader generated-video stress test. The lower AEGIS-Hard scores expose a remaining domain gap. As

Table 2: Comparison on the GenVideo many-to-many task. Best results in **bold**.

Method	Level	Recall (%)	AP (%)	F1 (%)
F3Net (ECCV'20) Qian et al. (2020)	Image	81.88	88.73	80.24
NPR (CVPR'24) Tan et al. (2024)	Image	84.08	82.45	71.64
CLIP-B-PT (ICML'21) Radford et al. (2021)	Image	81.74	44.83	21.57
CLIP-B-FT (ICML'21) Radford et al. (2021)	Image	90.57	91.52	76.31
STIL (MM'21) Gu et al. (2021)	Video	82.22	87.12	78.23
VideoMAE-B (NeurIPS'22) Tong et al. (2022)	Video	87.50	94.54	89.55
MINTIME-CLIP (TIFS'24) Coccomini et al. (2024)	Video	87.62	91.55	82.31
FTCN-CLIP (ICCV'21) Zheng et al. (2021)	Video	86.18	92.21	87.55
TALL (ICCV'23) Xu et al. (2023)	Video	88.52	87.91	74.19
DeMamba-CLIP-PT Chen et al. (2024)	Video	78.86	79.97	73.49
DeMamba-CLIP-FT Chen et al. (2024)	Video	91.58	93.45	89.19
XCLIP-B-PT Ni et al. (2022)	Video	81.93	61.29	38.39
DeMamba-XCLIP-PT Chen et al. (2024)	Video	83.59	76.38	64.77
XCLIP-B-FT Ni et al. (2022)	Video	84.41	86.77	76.62
DeMamba-XCLIP-FT Chen et al. (2024)	Video	93.92	97.10	90.20
Qwen2.5-VL (arXiv'25) Bai et al. (2025)	MLLM	54.23	77.43	70.39
GPT-4.1mini (2025) OpenAI (2025)	MLLM	65.44	78.02	72.54
VidGuard-R1 (ICLR'26) Park et al. (2026)	MLLM	87.66	94.38	90.00
Ours	**Hybrid**	**94.20**	**97.46**	**93.79**

detailed in Appendix C, AUROC is not uniformly reported for all AEGIS baselines, so unavailable values remain unreported rather than inferred from other metrics.

Table 3: Cross-task AUROC (%) on Celeb-DF, a localized face-swapping benchmark used to evaluate transfer beyond whole-video AI-generated content detection. GenVidCLIP is evaluated without dataset-specific fine-tuning; higher AUROC indicates better separation between authentic and manipulated videos.

	FWA	Xcep.-23	Xcep.-40	DSP-FWA	**Ours**
AUROC (%)	56.9	65.3	65.5	64.6	**82.4**

Table 4: AEGIS-Hard external stress test reported as Acc_{all}/Macro-F1 under zero-shot (ZS), reasoning (Reason), and LoRA settings. GenVidCLIP is evaluated without dataset-specific fine-tuning in the ZS setting; "–" denotes an unavailable or unreported result. Values from different columns correspond to protocol-specific evaluation settings and are not directly interchangeable.

Method	ZS	Reason	LoRA
Qwen2.5-VL 3B	.52/.48	.47/.46	.56/.47
Qwen2.5-VL 7B	.59/.52	.57/.48	.61/.55
Video-LLaVA-HF 7B	.50/.33	.46/.45	–
GenVidCLIP	**.68/.68**	–	–

Table 5: Per-generator accuracy on GenVideo under the Many-to-Many zero-shot protocol. "Avg" is the mean across nine unseen generators; bold marks the best result in each column.

Model	Sora	Morph	Gen2	HotShot	Lavie	Show-1	Moon	Crafter	Wild	Avg
F3Net Qian et al. (2020)	0.8393	0.9971	0.9862	0.7577	0.5700	0.3657	0.9952	0.9971	0.7678	0.8188
NPR Tan et al. (2024)	0.9107	0.9957	0.9949	0.2429	0.8964	0.5771	0.9712	0.9986	0.8780	0.8408
STIL Gu et al. (2021)	0.7857	0.9814	0.9804	0.7600	0.6179	0.5329	0.9936	0.9736	0.6501	0.8222
VideoMAE Tong et al. (2022)	0.6786	0.9600	0.9841	0.9614	0.7714	0.8043	0.9744	0.9693	0.6836	0.8750
MINTIME Coccomini et al. (2024)	0.8929	1.0000	0.9899	0.2643	0.9679	0.9814	0.9984	1.0000	0.8238	0.8762
FTCN Zheng et al. (2021)	0.8750	1.0000	0.9891	0.1771	0.9771	0.9186	1.0000	1.0000	0.8283	0.8618
TALL Xu et al. (2023)	0.9107	0.9828	0.9783	0.8300	0.7657	0.7957	0.9952	0.9893	0.6631	0.8852
CLIP-B-PT Radford et al. (2021)	0.8571	0.8243	0.9036	0.7100	0.7929	0.7543	0.8962	0.8629	0.7516	0.8174
DeMamba-XCLIP-FT Chen et al. (2024)	**0.9821**	**1.0000**	**0.9986**	0.6543	0.9486	**0.9886**	**1.0000**	**1.0000**	0.8909	0.9392
Qwen2.5-VL Bai et al. (2025)	0.5834	0.5568	0.5400	0.3389	0.4213	0.3800	0.8115	0.6287	0.7143	0.5409
GPT-4.1mini OpenAI (2025)	0.4302	0.6695	0.5623	0.5411	0.6300	0.5587	0.9143	0.6712	0.6892	0.6496
VidGuard-R1 Park et al. (2026)	0.6629	0.9586	0.9623	0.8271	**0.9429**	0.7200	0.9792	0.9807	0.7863	0.8766
GenVidCLIP	0.8367	0.9638	0.9906	**0.9638**	0.8976	0.9103	0.9890	0.9774	**0.9136**	**0.9420**

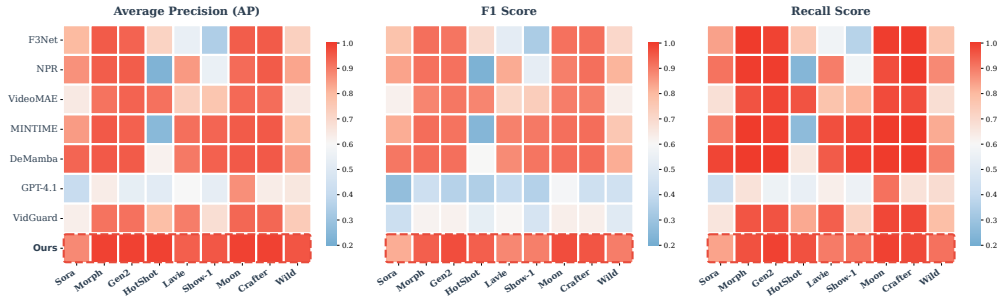

Figure 3: **Zero-shot generalization performance heatmaps on the GenVideo dataset.** We compare the Average Precision (AP), F1 Score, and Recall across 9 unseen video generators. As highlighted by the dashed red box, GenVidCLIP maintains a stable performance block across the nine generative domains under the same zero-shot evaluation protocol.

4.3 Ablation Studies

We conduct systematic ablation studies on GenVidBench to validate the contribution of each module (Table 6). Relying solely on SSS yields 0.9151 AP, whereas introducing explicit kinematic modeling consistently improves performance: optical-flow TKS reaches 0.9472 AP, L1-based TKS reaches 0.9491 AP, and combining both cues gives the strongest result with 0.9692 AP, 0.9273 F1, and 0.8810 ACC. Table 6(b) further shows that CRGN improves F1 and ACC, supporting selective residual injection rather than direct feature merging. For temporal density, $K{=}16$ provides the best balance, whereas denser sampling slightly reduces F1 by reintroducing less informative frames into the selected set used for temporal anomaly modeling.

One may ask whether the TKS gains arise from kinematic selection or simply from processing more frames. Three observations argue against the latter explanation. First, the SSS baseline already processes 32 frames spanning the full video, so temporal coverage itself is not the bottleneck. Second, both single-cue TKS variants improve upon SSS, and their combination is stronger than either alone. Third, the keyframe-density sweep is non-monotonic and peaks at $K{=}16$, which is inconsistent with a simple "more frames is always better" effect. Dataset-level motion statistics in Appendix E provide complementary correlational evidence for source-dependent temporal responses.

Table 6: Ablation on TKS, CRGN, and keyframes.

Configuration	AP	F1	ACC
(a) Feature Extraction			
SSS (baseline)	0.9151	0.8797	0.8005
+ TKS (Opt. Flow)	0.9472	0.8967	0.8242
+ TKS (L1 Pixel)	0.9491	0.8919	0.8215
+ L1 & Opt. Flow	**0.9692**	**0.9273**	**0.8810**
(b) Residual Gating			
w/o CRGN gating	**0.9693**	0.9189	0.8693
w/ CRGN gating	0.9692	**0.9273**	**0.8810**

Configuration	AP	F1	ACC
(c) Keyframe Density			
4 Keyframes	0.9665	0.9113	0.8575
8 Keyframes	0.9611	0.9040	0.8430
12 Keyframes	0.9673	0.9071	0.8485
16 Keyframes	**0.9692**	**0.9273**	**0.8810**
18 Keyframes	0.9682	0.9101	0.8787
20 Keyframes	0.9668	0.9119	0.8746

4.4 Qualitative Analysis

Figure 4 visualizes TKS and CRGN responses. Although individual RGB frames remain visually realistic, L1 differences and optical-flow fields reveal abrupt local changes and motion discontinuities. The CRGN responses concentrate on informative regions while suppressing much of the static background, and the selected Top-K keyframes coincide with high-response moments. These observations complement the quantitative ablations without being treated as causal evidence.

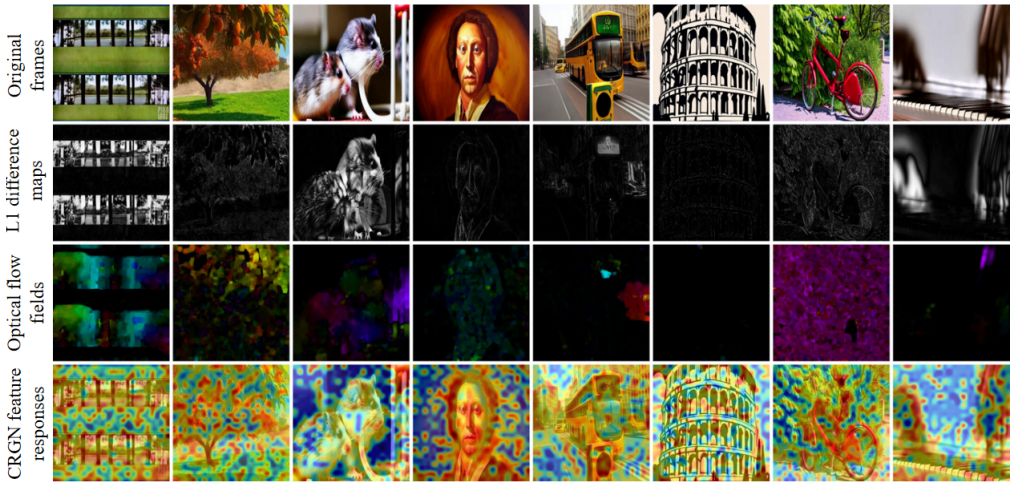

Figure 4: **Qualitative visualization of GenVidCLIP.** Top to bottom: Original frames, L1 differences, optical flow, and CRGN responses. Selected keyframes highlighted in red.

Robustness to editing artifacts. Hard cuts, abrupt scene changes, and camera transitions can also produce large L1 or optical-flow responses and may therefore be selected by TKS. To examine whether these sparse responses displace the continuous spatial context, we compare the fused representation with the SSS representation: the mean adjacent-frame cosine similarity within SSS is 0.9728, the fused-vs.-SSS cosine similarity is 0.9376, and 93% of videos exceed 0.90 similarity (Appendix F). These results indicate that the global SSS context remains dominant for most videos, although edit-induced false positives remain possible. We therefore view the consistency analysis as evidence that sparse TKS cues complement, rather than replace, the full-coverage SSS context; it does not imply immunity to scene transitions or editing artifacts.

Observed failure modes. Qualitative inspection identifies three recurring cases (Appendix G): extreme handheld camera shake can mimic abnormal motion, heavy compression can introduce temporal flicker, and near-static generated videos may provide too little kinematic evidence for TKS. These cases motivate scene-cut-aware filtering and complementary non-motion cues rather than treating kinematic responses as universally reliable. In particular, future temporal scorers could separate global camera motion from localized residual motion, while low-response near-static clips may benefit from complementary spatial or frequency-domain evidence.

Practical implications. Taken together, these analyses suggest that TKS should be treated as a complementary source of localized temporal evidence rather than a standalone decision cue. The

continuous SSS branch preserves global semantic context when motion evidence is weak or confounded, while CRGN regulates how strongly TKS features are injected into the final representation. Because the current implementation requires full-video Top-K selection and dense optical flow, deployment-oriented variants should prioritize lower-cost temporal scoring while preserving the same spatial–temporal complementarity.

5 Conclusion

We presented GenVidCLIP, a three-branch framework that separates global spatial semantics, sparse temporal kinematic cues, and text-based authenticity priors before residual fusion. TKS isolates high-response moments through L1/optical-flow sampling, while CRGN selectively injects temporal evidence into the spatial representation before visual-text alignment. Trained on 3,200 videos, GenVidCLIP achieves 85.05% ACC, 90.29% F1, 87.74% HD-real TNR, and 99.59% AUROC on GenVidBench-6M, together with strong GenVideo cross-generator transfer. Celeb-DF and AEGIS-Hard further demonstrate cross-task robustness while exposing a remaining domain gap. The principal limitations are the cost of dense optical flow, full-video access for Top-K selection, and sensitivity to camera edits, compression artifacts, and near-static generated content. Future work will focus on more efficient, edit-aware temporal scoring and complementary cues for low-motion cases.

AI USE STATEMENT

Generative AI tools were used for language polishing, Chinese–English translation, organization and presentation of experimental results, literature search, verification and formatting of references, and assistance with the formatting of tables, figure captions, and LaTeX content.

Generative AI was not used for methodology development, experimental design, or the generation of experimental data or results. All AI-assisted content, citations, and formatting revisions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

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 (CVPR)*, pp. 5001–5010, 2020.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. "Leveraging Frequency Analysis for Deep Fake Image Recognition." In *International Conference on Machine Learning (ICML)*, pp. 3247–3258, PMLR, 2020.

Yuezun Li and Siwei Lyu. "Exposing DeepFake Videos By Detecting Face Warping Artifacts." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 46–52, 2019.

Joao Carreira and Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset." In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6299–6308, 2017.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. "Is Space-Time Attention All You Need for Video Understanding?" In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR 139, pp. 813–824, 2021.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. "DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark." *arXiv preprint arXiv:2405.19707*, 2024.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. "Learning Self-Consistency for Deepfake Detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15023–15033, 2021.

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 (CVPR)*, pp. 24480–24489, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. "Learning to Prompt for Vision-Language Models." *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. "Robust Fine-Tuning of Zero-Shot Models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, 2022.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. "Video Generation Models as World Simulators." *OpenAI Technical Report*, 2024. [Online]. Available: `https://openai.com/index/video-generation-models-as-world-simulators/`

Runway. "Gen-2: Generate novel videos with text, images, or video clips." *Runway Research*, 2023. [Online]. Available: `https://runway.com/research/gen-2`

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision." In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, PMLR, 2021.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. "SlowFast Networks for Video Recognition." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6202–6211, 2019.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training." In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 10078–10093, 2022.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. "AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection." In *International Conference on Learning Representations (ICLR)*, 2024.

Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. "Temporal Pyramid Network for Action Recognition." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 591–600, 2020.

Hao Shao, Shengju Qian, and Yu Liu. "Temporal Interlacing Network." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11966–11973, 2020.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. "Temporal Relational Reasoning in Videos." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 803–818, 2018.

Ji Lin, Chuang Gan, and Song Han. "TSM: Temporal Shift Module for Efficient Video Understanding." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7083–7093, 2019.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. "UniFormerV2: Unlocking the Potential of Image ViTs for Video Understanding." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1632–1643, 2023.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. "Video Swin Transformer." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3202–3211, 2022.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4804–4814, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. "Qwen2.5-VL Technical Report." *arXiv preprint arXiv:2502.13923*, 2025.

OpenAI. "Introducing GPT-4.1 in the API." *OpenAI*, 2025. [Online]. Available: `https://openai.com/index/gpt-4-1/`

Kyoungjun Park, Yifan Yang, Juheon Yi, Muhammad Muaz, Shicheng Zheng, Yifei Shen, Dongqi Han, Caihua Shan, and Lili Qiu. "VidGuard-R1: AI-Generated Video Detection and Explanation via Reasoning MLLMs and RL." In *International Conference on Learning Representations (ICLR)*, 2026.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. "Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 86–103, 2020.

Chuangchuang Tan, Huan Liu, 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 (CVPR)*, pp. 28130–28139, 2024.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. "Spatiotemporal Inconsistency Learning for DeepFake Video Detection." In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, pp. 3473–3481, 2021.

Davide Alessandro Coccomini, Giorgos Kordopatis-Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. "MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection." *IEEE Transactions on Information Forensics and Security*, 19:6084–6096, 2024.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. "Exploring Temporal Coherence for More General Video Face Forgery Detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15044–15054, 2021.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. "TALL: Thumbnail Layout for Deepfake Video Detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22658–22668, 2023.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. "Expanding Language-Image Pretrained Models for General Video Recognition." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 1–18, 2022.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. "GenVidBench: A 6-Million Benchmark for AI-Generated Video Detection." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 18, pp. 15582–15590, 2026.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216, 2020.

Jieyu Li, Xin Zhang, and Joey Tianyi Zhou. "AEGIS: Authenticity Evaluation Benchmark for AI-Generated Video Sequences." In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, pp. 13346–13353, 2025. doi: 10.1145/3746027.3758295.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. "TALL: Thumbnail Layout for Deepfake Video Detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22658–22668, 2023.

Davide Alessandro Coccomini, Giorgos Kordopatis-Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. "MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection." *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 6084–6096, 2024. doi: 10.1109/TIFS.2024.3409054.

Meng Du, Hongchang Chen, Ran Li, Junjie Zhang, Qi Ouyang, Shibo Zhang, and Shuxin Liu. "G2VD: Generalizable AI-Generated Video Detection via Counterfactual Intervention and Causal Disentanglement." *arXiv preprint arXiv:2607.04607*, 2026.

A Implementation Details

Architecture. GenVidCLIP builds upon a ViT-L-14-336 backbone pre-trained by OpenAI CLIP (loaded via open_clip). The vision encoder consists of 24 Transformer layers with width 1024, patch size 14, and 16 attention heads, producing 768-dimensional embeddings. The text Transformer has 12 layers, width 768, and 12 heads. Both the SSS Encoder and the TKS Encoder are lightweight 2-layer Transformers with 8 heads and dropout 0.1. The CRGN gate MLP maps from 3×768 to 768 dimensions via two linear layers with LayerNorm, GELU activation, and a final Sigmoid gate. Input frames are resized and center-cropped to 336×336 pixels and normalized with CLIP's standard mean and standard deviation.

Training. We train for 15 epochs on a subset of 3,200 videos using a single NVIDIA GeForce RTX 4090 (24 GB). We use AdamW with weight decay 0.05, a head learning rate of 10^{-4}, and a backbone learning rate of 10^{-6} ($0.01\times$ the head rate). The scheduler is CosineAnnealingLR with $\eta_{\min} = 10^{-6}$. Physical batch size is 2 with 16 gradient accumulation steps, yielding an effective batch size of 32. We employ mixed-precision training, gradient checkpointing, and label smoothing of 0.1. The random seed is fixed to 111.

Data Sampling. The SSS uniformly samples $N{=}32$ frames spanning the entire video. The TKS extracts $K{=}16$ keyframes via the hybrid kinematic strategy: each frame is downsampled to 128×128 grayscale, and both the L1 pixel difference and the Farneback dense optical flow variance (pyr_scale=0.5, levels=3, winsize=15, iterations=3, poly_n=5, poly_sigma=1.2) are computed. After min-max normalization and linear fusion, the Top-K frames are selected and temporally re-ordered.

Prompt Learning. We adopt an object-agnostic prompt design with 9 normal templates (e.g., "real {}", "natural {}") and 14 anomaly templates (e.g., "fake {}", "AI-generated {}", "{} with temporal flickering"). The learnable context length is 12, with deep compound prompts of length 4 injected at depth 9 in the text Transformer.

Dataset Protocols. For GenVidBench-6M, we follow the original benchmark split: Pair 1 for training and Pair 2 for testing (cross-source and cross-generator). For GenVideo, we adopt the Many-to-Many evaluation protocol from DeMamba Chen et al. (2024).

B Dataset Statistics

Table 7 summarizes the dataset statistics used in our experiments. GenVidBench-6M Ni et al. (2026) is a large-scale benchmark containing exactly 6,488,432 video clips from diverse generative engines (CogVideo, Mora, MuseV, SVD) as well as real high-definition (HD) videos collected from public sources. The full training split (Pair 1) and test split (Pair 2) each contain 3,244,216 videos. Due to computational constraints and for efficient training, our training phase utilizes a subset of 3,200 videos uniformly sampled from Pair 1 to stress-test data efficiency, as detailed in Section 4. Meanwhile, the complete Pair 2 split is used to rigorously evaluate cross-source and cross-generator generalization. GenVideo Chen et al. (2024) provides a complementary benchmark with 9 additional unseen generators for zero-shot evaluation under the Many-to-Many protocol, utilizing its official test split of 10,000 real and 8,588 fake videos.

Table 7: Dataset statistics of the full GenVidBench-6M and GenVideo benchmarks. Note: Our model was trained on a 3,200-video subset of Pair 1.

Dataset	Split	#Real	#Fake	Generators
GenVidBench-6M	Pair 1 (Train)	1,622,108	1,622,108	Subset of generators
	Pair 2 (Test)	1,622,108	1,622,108	CogVideo, Mora, MuseV, SVD
GenVideo	M2M (Test)	10,000	8,588	Sora, Morph, Gen2, HotShot, Lavie, Show-1, Moon, Crafter, Wild

C Additional External Robustness Results

Celeb-DF focuses on face-swapping and localized facial manipulation rather than whole-video generation, while AEGIS-Hard provides a difficult external stress test for AI-generated video detection. The corresponding results are reported in Tables 3 and 4 in the main paper. We treat both evaluations as external robustness analyses rather than matched primary benchmarks. For AEGIS-Hard, AUROC is available for GenVidCLIP but not uniformly reported for all comparison baselines, so missing values are left unreported rather than inferred.

D Parameter and Data Efficiency

GenVidCLIP is not designed to minimize the total size of the pre-trained backbone; instead, it follows a parameter-efficient task-adaptation regime. The full model contains approximately 470.32M parameters, but only 42.38M (9.0%) are updated during training, while the remaining 427.94M parameters (91.0%) stay frozen. More importantly, this adaptation is performed using only 3,200 training videos, compared with approximately 3.2M videos used by the supervised CNN/Transformer baselines. Table 8 therefore focuses on the task-specific optimization footprint and training-data scale rather than raw model size alone.

Table 8: Comparison of task-specific optimization scale and training-data efficiency. "Updated" denotes the fraction of model parameters optimized during task-specific training. End-to-end baselines update the full network, whereas GenVidCLIP updates only 9.0% of its parameters.

Method	Backbone	Trainable	Updated	Training Videos
I3D Carreira and Zisserman (2017)	3D-R50	∼25M	100%	∼3.2M
SlowFast Feichtenhofer et al. (2019)	R50+R50	∼34M	100%	∼3.2M
TimeSformer Bertasius et al. (2021)	ViT-B/16	∼121M	100%	∼3.2M
VideoSwin Liu et al. (2022)	Swin-B	∼88M	100%	∼3.2M
VideoMAE-B Tong et al. (2022)	ViT-B/16	∼87M	100%	∼3.2M
DeMamba Chen et al. (2024)	ViT-B/16	∼4.1M	∼3.3%	∼3.2M
GenVidCLIP (Ours)	**ViT-L/14-336**	**42.38M**	**9.0%**	**3,200**

Several observations emerge from Table 8. First, GenVidCLIP updates only 42.38M parameters, corresponding to 9.0% of the full model, while 91.0% of the pre-trained parameters remain frozen. This yields a substantially smaller optimization footprint than fully fine-tuned CNN/Transformer baselines, which update their entire networks. Second, GenVidCLIP uses only 3,200 training videos—approximately three orders of magnitude fewer than the ∼3.2M videos used by these supervised baselines—while retaining competitive detection performance. DeMamba uses fewer trainable parameters (4.1M) and therefore represents a particularly lightweight adaptation regime; our advantage instead lies in combining limited task-specific optimization with strong data efficiency. General-purpose MLLMs are excluded from the trainable-parameter comparison because they are evaluated zero-shot in our experiments and therefore follow a fundamentally different adaptation protocol.

For completeness, GenVidCLIP processes $N+K=48$ frames at 336×336 resolution per video, which increases the per-video inference cost relative to conventional 224^2 video backbones. This cost primarily comes from the ViT-L visual encoder; Farneback optical flow is computed on only 128×128 grayscale frames as CPU-side preprocessing and can be parallelized independently of GPU inference. Thus, Table 8 should be interpreted as a comparison of task-specific optimization and data efficiency rather than a claim of minimum inference cost.

E Dataset-Level Temporal Statistics

To test whether the temporal-artifact motivation extends beyond a few qualitative examples, we uniformly sample 32 frames from 1,000 videos and aggregate video-level p95 values of L1 frame difference, dense optical-flow magnitude, and optical-flow variance by source. These statistics are intended as correlational support for the motivation, not as causal proof.

Table 9: Dataset-level temporal-response statistics on GenVidBench-6M.

Source	n	Frame diff	Flow mag	Flow var
CogVideo	198	0.055	2.408	63.98
Mora	206	0.056	2.264	44.71
SVD	199	0.042	2.055	13.13
MuseV	193	0.010	0.423	1.75
HD Real	204	0.008	0.324	1.03

CogVideo and Mora show stronger temporal responses, whereas SVD and especially MuseV exhibit weaker motion cues closer to real HD videos. This pattern is consistent with the observed difficulty differences, while we deliberately avoid interpreting the correlation as strict causality.

F TKS AND CONTINUOUS SSS CONTEXT

Real-world edits such as hard cuts, abrupt scene changes, camera transitions, and shot changes can generate large frame-difference or optical-flow responses and may therefore be selected by TKS. To examine whether sparse TKS evidence substantially displaces the continuous SSS context, we compare the final fused video representation against the SSS global representation. SSS uses 32 uniformly sampled frames and TKS uses 16 high-response keyframes.

Table 10: Fused-vs-SSS representation consistency analysis.

Metric	Result	Interpretation
Mean adjacent-frame cosine similarity among SSS frame features	0.9728	Strong local continuity in the SSS sequence.
Cosine similarity between fused video feature and SSS global feature	0.9376	Fused representation remains close to the continuous global context.
Videos with fused-vs-SSS similarity $>$ 0.90	93%	Alignment holds for most videos.

These results show that the fused representation remains strongly aligned with the SSS global context, but they do not imply that TKS is immune to editing artifacts. We therefore explicitly retain real-world cuts and scene transitions as a limitation and leave scene-cut-aware filtering to future work.

G FAILURE CASE ANALYSIS

We identify three dominant failure modes through qualitative inspection of misclassified samples (Figure 5):

(a) Real videos with extreme camera motion. Handheld or action-camera footage with rapid panning or shaking produces elevated optical-flow variance and L1 pixel differences, which can mimic the kinematic signature of generative artifacts and cause false positives on HD real videos.

(b) Real videos with heavy compression. Heavily compressed real videos (e.g., low-bitrate social media re-uploads) introduce blockiness and temporal flickering that closely resemble AI-generated artifacts, confusing the TKS.

(c) Near-static AI-generated content. When a generated video depicts a mostly static scene with minimal motion, the kinematic sampling strategy finds few discriminative keyframes, and the TKS feature F_T becomes uninformative. In such cases, the model relies primarily on the SSS, which may lack sufficient spatial cues for detection.

These observations suggest future improvements including scene-cut/transition-aware filtering, motion-aware scoring that distinguishes camera motion from local generative inconsistencies, and complementary frequency-domain cues for near-static content.

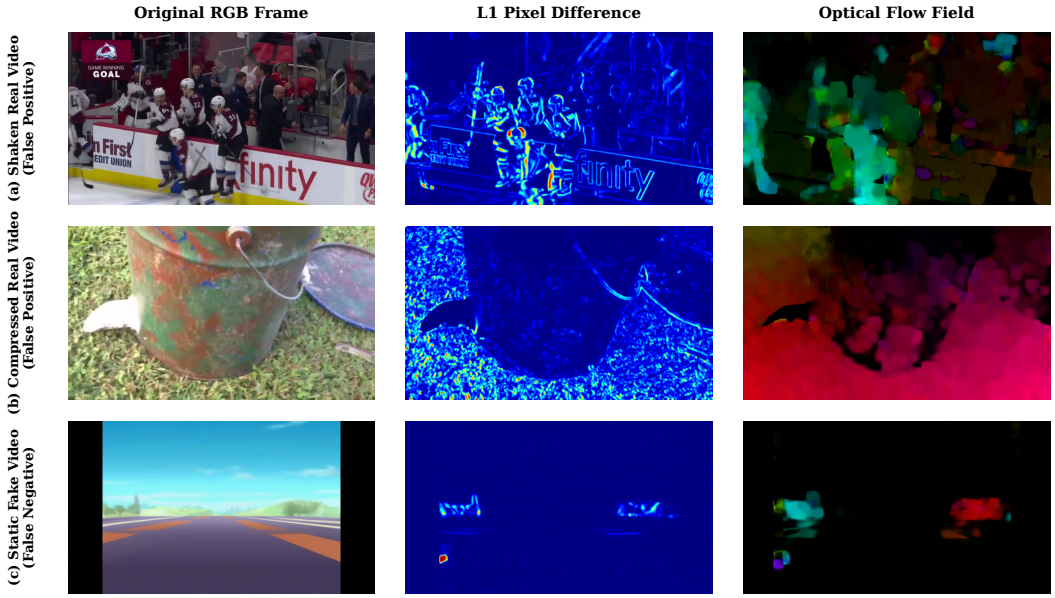

Figure 5: **Representative failure cases.** (a) A real video with extreme handheld camera shake that triggers high kinematic scores, leading to a false positive. (b) A real video with heavy compression artifacts resembling AI-generated flickering. (c) An AI-generated video with minimal motion (near-static scene), where the TKS fails to find discriminative keyframes.

H Broader Impact

GenVidCLIP is designed to detect AI-generated videos and combat visual misinformation. We highlight both positive and negative societal implications.

Positive impact. Robust deepfake detection is essential for preserving trust in visual media. Our method can support content moderation platforms, journalistic fact-checking, and judicial evidence verification by automatically flagging potentially synthetic videos.

Potential negative impact. False positivesreal videos incorrectly flagged as AI-generatedcould unjustly discredit authentic content or harm individuals whose genuine recordings are questioned. We therefore recommend using automated detections as advisory signals for human review rather than as standalone judgments. Publishing detection methodologies may also enable adversarial adaptation by generator developers; however, transparent research can support stronger defensive evaluation and countermeasures.