HuRF: Human-Centric Representation Learning for Generalizable Digital Human Video Forensics

Anonymous authors
Paper under double-blind review

Abstract

Native digital humans, whose appearance, motion, and identity are synthesized or reconstructed end-to-end, pose growing challenges to media authenticity. Existing AI-generated video and face-centric deepfake detectors often rely on generator-specific visual artifacts, temporal irregularities, or editing traces, which do not transfer reliably across heterogeneous digital human generation mechanisms. We therefore seek transferable forensic evidence from the human subject itself, which provides a shared representational basis across different generation mechanisms. Based on this insight, we propose **HuRF**, a **hu**man-centric **r**epresentation learning framework for generalizable digital human video **f**orensics. HuRF leverages intermediate representations from a pose-oriented human foundation model, retaining dense RGB appearance alongside fine-grained awareness of human structure. It further distills region-level semantic relations from a human parsing teacher to suppress background interference, and adaptively calibrates temporal evidence across generation mechanisms, reducing reliance on unreliable motion-naturalness cues. In this way, the detector focuses on transferable evidence from the human subject rather than generator- or background-specific shortcuts. To systematically evaluate this problem, we construct **HuDigivid**, containing over 35K videos from 19 known generators, together with nearly 1K collected in-the-wild digital human videos with unknown generators. Extensive experiments reveal distinct failure modes of existing detector families and show that HuRF achieves the strongest overall performance under cross-generator, cross-paradigm, and in-the-wild evaluations while remaining robust to common video degradations.

1 Introduction

AI-generated video (AIGV) technology (Ho et al., 2022; Peebles & Xie, 2023) has advanced rapidly in visual fidelity, motion coherence, and controllability, with growing use in content creation, live streaming, and virtual services. Among human-centric generated videos, native digital humans pose a distinct forensic challenge because their appearance, motion, and identity are synthesized or reconstructed end-to-end rather than locally edited from real recordings. Reliably verifying such content has therefore become an important forensic problem, being unique in AIGV detection and distinct from facial deepfake detection.

A key challenge of native digital human forensics lies in the mismatch between existing forensic evidence assumptions and this human-centric authenticity objective. As illustrated in Fig. 1, general AIGV detectors are developed for open-domain content authenticity and typically exploit generation-related cues from arbitrary visual content (Ma et al., 2025; Ojha et al., 2023; Corvi et al., 2026). However, digital human forensics focuses on the authenticity of the human subject itself, making non-human content or scene-level cues potentially misleading. Meanwhile, video-level detectors may be misled by realistic motion patterns inherited from real videos in explicit-guided digital human generation, where temporal naturalness becomes less discriminative for authenticity verification. In contrast, traditional deepfake detectors target edited real videos and rely on manipulation traces introduced during identity modification (Shiohara & Yamasaki, 2022), which may not exist in end-to-end synthesized digital humans. Therefore, existing forensic paradigms based on either general generation cues or editing traces may not provide transferable evidence for native dig-

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.

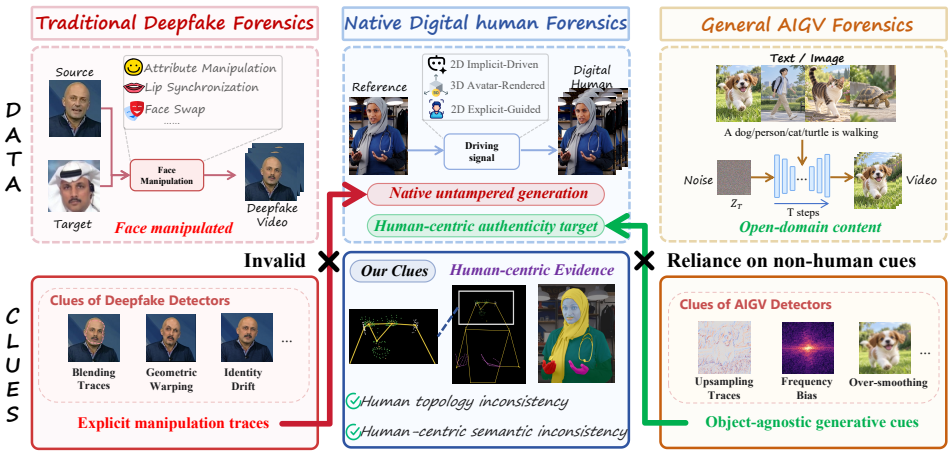


Figure 1: Comparison of forensic characteristics across traditional deepfake, general AIGV, and native digital human videos. Traditional deepfake detection relies on manipulation traces from identity editing, while general AIGV detection exploits generation-related cues from open-domain content. In contrast, native digital human forensics requires human-centric evidence beyond generator-specific cues and non-human content.

ital human verification. Fig. 2 shows differing mean zero-shot performance among representative pretrained detectors. Appendix Table B1 reports their per-generator AUCs, while Table 2 further demonstrates the transfer limitations of different detector families under cross-generator and cross-paradigm evaluations. These observations call for forensic representations centered on the human subject rather than generator-specific or non-human cues.

We address this problem from a human-centric perspective. Although various kinds of digital human systems differ substantially in how they generate video, what they generate remains the same—the human subject. We therefore use the human subject as a shared representational basis for organizing forensic evidence across generation mechanisms and propose **HuRF (Human-Centric Representation Learning for Digital Human Video Forensics)**. HuRF leverages intermediate representations from a pose-oriented human foundation model (Khirodkar et al., 2026) to retain dense appearance while internalizing fine-grained human structure, and uses **Human Semantic Relation Distillation (HSRD)** to organize evidence through human-part relations while suppressing background interference. Besides, **Temporal Adaptive Calibration (TAC)** is introduced to adaptively calibrate temporal evidence across generation mechanisms, reducing the dependence on motion-naturalness cues for pose-driven digital humans whose motion patterns are inherited from real videos. Since human-centric representation learning and temporal adaptation capture complementary but distinct forensic evidence, we adopt a two-stage training strategy to progressively optimize them, with the semantic teacher discarded during inference.

Systematically evaluating this problem also requires data that span heterogeneous digital human generation mechanisms. Conventional deepfake datasets (Dolhansky et al., 2020; Khalid et al., 2021; Cai et al., 2024) mainly target editing-based manipulations of real human videos, whereas general AIGV datasets (Ma et al., 2026; Ni et al., 2026; Wen et al., 2025; Li et al., 2026b) cover open-domain generated content. Recent digital human datasets such as **DigiFakeAV** (Liu et al., 2026) and **AvatarShield** (Xu et al., 2026) move closer to our setting, but remain largely confined to 2D generation and provide limited coverage of fundamentally different synthesis and rendering paradigms. Such limited coverage may bias evaluation toward specific generation mechanisms rather than the broader digital human generation space. We therefore construct **HuDi-**

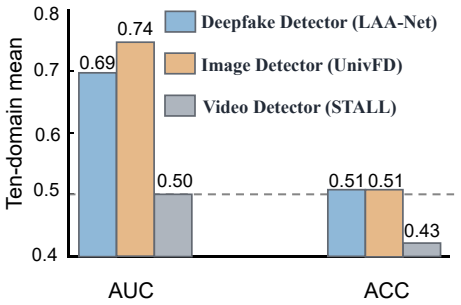


Figure 2: Ten-domain mean zero-shot AUC and ACC@0.5 of pretrained detectors. Per-generator AUC is provided in Appendix Table B1.

givid, containing over 35K videos from 19 generators spanning **2D Implicit-Driven**, **2D Explicit-Guided**, and **3D Avatar-Rendered Generation**, together with two in-the-wild test sets.

Extensive experiments reveal substantial differences in the transfer behavior of existing detector families, whereas HuRF achieves consistently strong cross-generator, cross-paradigm, in-the-wild, and cross-dataset performance while remaining robust to common post-processing degradations. Our contributions are summarized as follows:

- We propose **HuRF**, a human-centric representation learning framework that organizes forensic evidence around human appearance and structure, region-level semantic relations, and adaptively calibrated temporal evidence across heterogeneous generation mechanisms.
- We construct **HuDigivid**, a large-scale digital human video dataset with over 35K videos from 19 controlled generators across three heterogeneous generation paradigms and two collected in-the-wild test sets.
- We **perform systematic analysis** on existing detector families and show their sensitivity to generator-specific visual traces, paradigm-dependent temporal cues, and editing-oriented supervision, while demonstrating HuRF's strong generalization across controlled, in-the-wild, and cross-dataset evaluations.

2 Related Work

2.1 Datasets for AI-Generated and Digital Human Videos

Recent AIGV datasets have grown substantially in scale and diversity. GenBuster-200K (Wen et al., 2025) incorporates recent high-fidelity generators and real-world data. GenWorld (Chen et al., 2025a) extends evaluation to driving, navigation, and manipulation scenarios, studying the physical plausibility of world-model videos through multi-view consistency. AIGVDBench (Ma et al., 2026) covers 31 open- and closed-source models and over 440K videos across T2V, I2V, and V2V tasks. These datasets substantially broaden the coverage of general AIGV detection, but remain open-domain, where humans constitute only one category among many rather than the organizing subject.

Human-video authenticity has also been extensively studied through deepfake datasets. FaceForensics++ (Rossler et al., 2019) standardized evaluation across classic facial manipulations; Celeb-DF (Li et al., 2020) reduced obvious synthesis artifacts; and DFDC (Dolhansky et al., 2020) expanded scale and identity diversity. DeepfakeBench (Yan et al., 2023) unified nine datasets and 15 detectors under common processing and evaluation protocols. This line of work primarily studies manipulation of existing real content, which differs from synthesizing or reconstructing a complete human subject.

Recent datasets have begun to target human-centric synthetic videos. DigiFakeAV (Liu et al., 2026) contains 60K audio-visual samples from five digital human generators, together with cloned speech. AvatarShield (Xu et al., 2026) constructs FakeHumanVid with roughly 15K videos from nine pose-, audio-, or text-driven methods. These efforts establish digital human forensics as a distinct setting from both conventional deepfakes and open-domain AIGV, but remain largely restricted to 2D generation. Compared with these datasets, HuDigivid expands the generation space to three paradigms including 2D Implicit-Driven Generation, 2D Explicit-Guided Generation, and 3D Avatar-Rendered Generation, while further incorporating collected in-the-wild digital humans for cross-generator, cross-paradigm, and real-world evaluation.

2.2 Forensics for AI-Generated Videos and Deepfakes

General AIGV forensics initially focused on generator-specific artifacts, but recent studies have shifted toward more generalizable spatiotemporal evidence. DeCoF (Ma et al., 2025) explores frame-level consistency, while DeMamba (Chen et al., 2026) models fine-grained spatial-temporal patterns. More recent works further investigate representation dynamics (Internò et al., 2026) and physical motion regularities (Zhang et al., 2026) to improve robustness. However, these methods mainly capture generic video properties and do not explicitly model human-specific structures.

Deepfake forensics mainly studies manipulated real videos, where editing operations leave characteristic traces. Early approaches learn local blending artifacts (Shiohara & Yamasaki, 2022), while later methods improve generalization by exploiting spatial-temporal patterns (Xu et al., 2023) or adapting large-scale visual representations (Cui et al., 2025). Nevertheless, these methods are designed for editing-based manipulations and may not transfer to native digital humans, whose appearance and motion are generated end-to-end.

3 HUDIGIVID

3.1 DATASET CONSTRUCTION

HuDigivid targets generalization-oriented digital human forensics by covering diverse generation mechanisms and data sources. For authenticity verification, we consider videos of digitally generated humans as fake content and naturally recorded ones as real content.

Real data collection. We collect a real-video pool of 15,371 human videos from HDTF (Zhang et al., 2021), DFDC (Dolhansky et al., 2020), public TikTok datasets (Jafarian & Park, 2021), and videos from YouTube, Bigo, Douyin, TikTok, and JD, covering news presentation, social media, and live commerce. Besides serving as authentic samples, selected videos also provide reference identities and conditioning inputs for controlled generation, with detailed source-to-generator correspondence provided in Appendix Section A.1.

Controlled fake data construction. Using 19 methods, we construct 19,275 controlled digital human videos: 9 from 2D Implicit-Driven Generation, 5 from 2D Explicit-Guided Generation, and 5 from 3D Avatar-Rendered Generation (Fig. 3(b)), with corresponding pipelines illustrated in Fig. 3(a). For 2D Explicit-Guided Generation, DWPose (Yang et al., 2023) sequences extracted from real videos are combined with reference images to provide explicit motion and identity conditions. 2D Implicit-Driven generators infer motion from higher-level conditions such as audio or driving videos together with a reference identity, while 3D Avatar-Rendered methods reconstruct an identity-specific 3D representation from images or videos and subsequently render it into video.

In-the-wild fake data collection To complement controlled generation, we independently collect 953 in-the-wild digital human videos, including 481 JDAvatar videos collected from JD Live and 472 BaiduAvatar videos collected from Baidu Live. These platform-collected videos may involve multiple unknown generation methods and processing pipelines, making them more heterogeneous than the controlled subsets; we therefore reserve them as independent real-world test sets.

Overall, HuDigivid contains 35,599 videos: 15,371 real human videos, 19,275 controlled digital human videos from 19 generators, and 953 independently collected in-the-wild digital human videos. Detailed generation configurations, reference sources, and access arrangements are provided in Appendix Table A1 and Section A.1, while detailed dataset compositions are provided in Appendix Tables A2 and A3.

3.2 EVALUATION PROTOCOL DESIGN

Digital human generators exhibit diverse synthesis mechanisms, resulting in different forensic challenges. A reliable detector should generalize beyond specific generator identities, transfer across generation paradigms, and remain effective under real-world deployment scenarios. Therefore, we design two complementary evaluation protocols to assess cross-generator and cross-paradigm generalization, respectively. Both protocols are further evaluated on in-the-wild digital datasets.

Cross-Generator Protocol. The *Cross-Generator Protocol* evaluates generalization to unseen generators with diverse training coverage. The training set contains synthetic videos from all three generation paradigms and real videos introduced in Sec. 3.1. The test set uses held-out generators across the same paradigms, with real samples collected from platforms unseen during training.

Cross-Paradigm Protocol. The *Cross-Paradigm Protocol* evaluates generalization beyond the observed training paradigm under limited training coverage. It restricts synthetic training videos to 2D Implicit-Driven Generation, with real samples adjusted to match the evaluation setting. The test set contains held-out generators across all three paradigms, evaluating both unseen-generator and unseen-paradigm transfer.

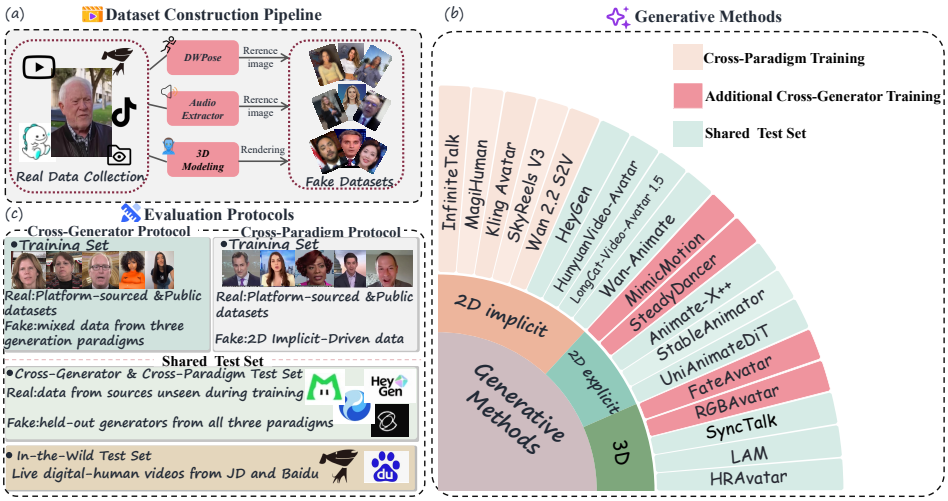


Figure 3: Overview of HuDigivid: (a) data construction pipeline, (b) 19 generators across three paradigms with protocol-specific fake coverage, and (c) evaluation protocols.

4 METHOD

4.1 OVERVIEW

To learn transferable forensic evidence across heterogeneous digital human generators, we propose HuRF, as illustrated in Fig. 4. Although different generators employ distinct synthesis mechanisms, they all need to reproduce the appearance, structure, and motion of the human subject. HuRF therefore organizes forensic learning around complementary **human-centric spatial and temporal evidence**. Specifically, we leverage intermediate representations from a pose-oriented human foundation model to preserve dense appearance while introducing **fine-grained structural awareness**. **Human Semantic Relation Distillation (HSRD)** further organizes spatial features through human-part relations and suppresses background interference, while **Temporal Adaptive Calibration (TAC)** adaptively calibrates the reliance on temporal evidence across different generation mechanisms by dynamically adjusting the contribution of the temporal branch. As human-centric representation learning and temporal evidence calibration focus on complementary but different objectives, we adopt a two-stage training strategy to progressively optimize these components.

4.2 THE DETAILS OF HURF

Pose-oriented representations provide strong structural priors but lack explicit semantic relationships among human regions. We therefore distill human-part relational knowledge from a parsing model, rather than directly fusing heterogeneous representations or enforcing point-wise alignment, thereby avoiding additional inference cost while preserving discriminative forensic features.

Human Semantic Relation Distillation (HSRD). HSRD transfers fine-grained human semantic knowledge from a frozen parsing model into the forensic representation through relational distillation. Specifically, we use Sapiens2-Seg (Khirodkar et al., 2026) as the semantic teacher, which predicts 29 semantic classes, including one background class and 28 human-related foreground classes covering different human parts. For the n-th frame and spatial position i, the logits of the foreground classes form the human semantic response $g_{n,i} \in \mathbb{R}^{28}$. Instead of directly matching segmentation features, HSRD transfers the semantic relationships among different spatial locations. For two positions i and j, their semantic relation in the teacher space is defined as:

$$R^T_{n,i,j} = \exp\left[\tau\left(\frac{g^T_{n,i} g_{n,j}}{\|g_{n,i}\|_2 \|g_{n,j}\|_2} - 1\right)\right], \tag{1}$$

where τ is the temperature coefficient.

Since background regions may occupy a considerable portion of digital human videos, directly supervising all spatial relations may introduce irrelevant correlations. We therefore introduce **Human-**

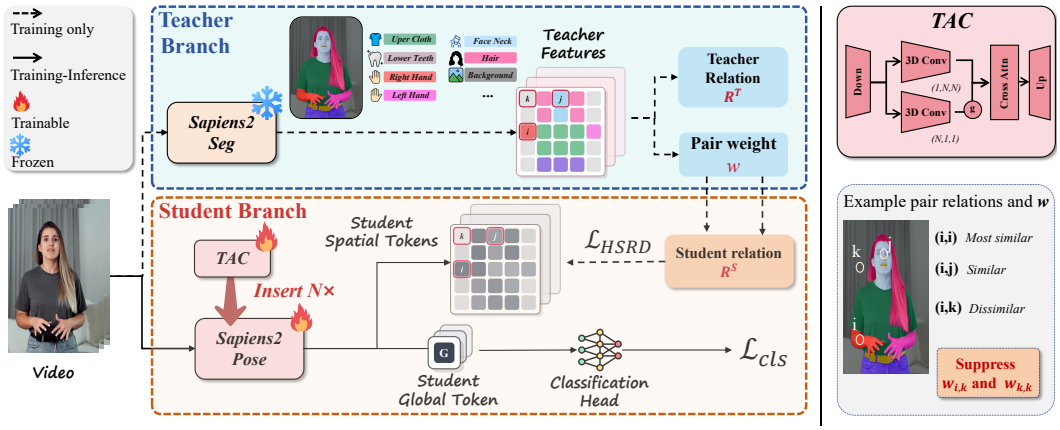


Figure 4: Overview of HuRF for human-centric digital human video forensics.

Aware Relation Weighting (HARW) to emphasize relations involving human-related regions. Let $q_{n,i,c}$ denote the predicted probability of semantic class c, where $c = 0$ represents the background class. The human relevance score is defined as $p_{n,i} = 1 - q_{n,i,0}$, and the pair-wise weight is:

$$w_{n,i,j} = p_{n,i} p_{n,j}. \tag{2}$$

The student representation is extracted from Sapiens2-Pose, which takes video frames as input and provides human-aware visual representations. Specifically, we use the output from block 32 of the 40-block Sapiens2-Pose Transformer, as this intermediate layer balances visual information and human structural awareness while avoiding the stronger task-specific bias of the final layer. The spatial token at position i of frame n is denoted as $z_{n,i} \in \mathbb{R}^D$. The corresponding relation in the student feature space is computed as:

$$R^S_{n,i,j} = \frac{z^T_{n,i} z_{n,j}}{\|z_{n,i}\|_2 \|z_{n,j}\|_2}. \tag{3}$$

Finally, HSRD minimizes the HARW-weighted discrepancy between teacher semantic relations and student forensic relations:

$$\mathcal{L}^{(n)}_{\text{HSRD}} = \frac{\sum_i \sum_j w_{n,i,j} \left(R^S_{n,i,j} - R^T_{n,i,j}\right)^2}{\sum_i \sum_j w_{n,i,j}}. \tag{4}$$

HSRD transfers fine-grained human-part semantic relations from the parsing teacher into the forensic representation, allowing the student encoder to internalize human semantic knowledge and establish human-aware relational structures for forensic analysis.

Temporal Adaptive Calibration (TAC). Temporal cues complement the within-frame relations learned by HSRD, but their forensic reliability varies across generation mechanisms. For example, 2D Explicit-Guided Generation inherits motion trajectories from real videos, **making realistic motion itself less discriminative**. We therefore introduce TAC to adaptively regulate temporal updates according to both the current representation and its temporal residual. TAC extends the spatiotemporal adapter of Yan et al. (2025b) and is inserted after each of the first 32 Pose Transformer blocks. From block output X, we obtain bottleneck features $U = \text{GELU}(W_{\text{down}} X)$, with patch tokens reshaped into a spatiotemporal map; U_ℓ is its frame-ℓ slice. For $a \in \{s, t\}$, $C_{a,k}$ denotes a depthwise 3D convolution on U: the spatial branch $a = s$ uses a $(1, k, k)$ kernel, and the temporal branch $a = t$ uses a $(k, 1, 1)$ kernel, where $k \in \{3, 5, 7\}$. Their multi-scale residuals are

$$\Delta_a(U) = \frac{1}{3} \sum_{k \in \{3,5,7\}} C_{a,k}(U), \qquad a \in \{s, t\}. \tag{5}$$

Here $[\Delta_a(U)]_\ell$ is the frame-ℓ slice of the residual. We spatially pool U_ℓ and the temporal residual as $p_\ell = \text{Pool}_{HW}(U_\ell)$ and $r_\ell = \text{Pool}_{HW}([\Delta_t(U)]_\ell)$. The frame-wise temporal gate is

$$g_\ell = 2\sigma(\text{MLP}([p_\ell, r_\ell])), \tag{6}$$

where $[\cdot, \cdot]$ denotes concatenation and $g_\ell \in (0, 2)$ is a scalar broadcast across spatial positions and channels. The spatial and temporal branch outputs at frame ℓ are

$$V_{s,\ell} = U_\ell + [\Delta_s(U)]_\ell, \qquad V_{t,\ell} = U_\ell + g_\ell [\Delta_t(U)]_\ell. \tag{7}$$

Thus, only the temporal residual is adaptively calibrated: $g_\ell < 1$ attenuates temporal evidence, whereas $g_\ell > 1$ amplifies it. The spatial and temporal branch features then interact through bidirectional cross-attention, with the gate initialized near one to preserve the original temporal update at initialization. After block 32, the final features are aggregated for clip-level authenticity prediction.

4.3 Training Strategy

Adapting the pose-oriented encoder shifts its representation toward authenticity discrimination, while TAC learns temporal interactions on top of this representation. Joint optimization may cause temporal gradients to interfere with frame-level representation adaptation. We therefore adopt progressive two-stage training.

Stage 1: Human-Centric Forensic Adaptation. We freeze the original Sapiens2-Pose parameters and train only the inserted LoRA weights and video classifier, comprising approximately 0.52M trainable parameters in total, using authenticity supervision and HARW-weighted HSRD.

Stage 2: Spatiotemporal Adaptation. Starting from the Stage-1 checkpoint, we freeze the adapted encoder and LoRA, insert TAC into the encoder, and update only TAC and the classifier. HSRD remains active to preserve established human-region semantics.

Both stages optimize:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{HSRD}} \mathcal{L}_{\text{HSRD}}. \tag{8}$$

where $\mathcal{L}_{\text{cls}}$ is the clip-level binary classification loss and $\mathcal{L}_{\text{HSRD}}$ is the relation distillation loss. At inference, the teacher is discarded, and HuRF predicts authenticity directly from the original video.

5 Experiments

5.1 Experimental Setup

Baselines and metrics. We compare HuRF with 11 representative baselines from four families: image-level AIGC detectors (e.g., Effort (Yan et al., 2024) and DGS-Net (Yan et al., 2025a)), video-level AIGC detectors, general video classification models (e.g., TimeSformer (Bertasius et al., 2021), VideoMAE (Tong et al., 2022), and UniFormerV2 (Li et al., 2023)), and deepfake detectors. All baselines use official pretrained weights and are fine-tuned on HuDigivid following their original training protocols. Deepfake methods retain self-synthesized supervision when required. We report video-level ROC-AUC and class-balanced ACC@0.5.

Implementation details. We use Sapiens2-Pose-1B and Sapiens2-Seg-1B as the forensic student and semantic teacher, respectively. HuRF takes 8-frame video clips with a resolution of 224×224 as input. Image-level AIGC and frame-based deepfake detectors independently process 8 sampled frames and average frame-level predictions, while video-level baselines follow their original temporal settings. During training, Stage 1 applies rank-8 LoRA modules to Sapiens2-Pose Transformer blocks 27–32 for human-centric forensic adaptation. In Stage 2, we freeze the adapted encoder and LoRA parameters, and optimize TAC modules inserted after the first 32 Transformer blocks together with the classification head. We set the HSRD temperature to $\tau = 20$ and the distillation loss weight to $\lambda_{\text{HSRD}} = 1$. We train HuRF for one epoch using Adam with a learning rate of 2×10^{-4}, weight decay of 5×10^{-4}, and an effective batch size of 32.

5.2 Generalization under the Cross-Generator Protocol

We evaluate all methods under the *Cross-Generator Protocol*, where training covers diverse generation mechanisms and testing uses held-out generators to assess generalization beyond specific generator identities. As reported in Table 1 and Appendix Table C1, existing methods achieve competitive performance under this setting, indicating that diverse training exposure enables them to capture transferable generation-related cues. Nevertheless, HuRF achieves the best performance with a mean AUC of 0.9164 and ACC@0.5 of 0.8348, demonstrating more generalizable forensic evidence across heterogeneous digital human generators.

Table 1. Held-out-generator generalization under the *Cross-Generator Protocol* in video-level ROC-AUC.

Method	2D Implicit				2D Explicit			3D			Cross-domain
	HeyGen	HunyuanAvatar	LongCat 1.5	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
Image-level AIGC detectors											
Effort	0.7668	0.9218	0.8934	0.6301	0.9901	0.9785	0.9722	0.9577	0.8607	0.8510	0.8822
DGS-Net	0.7693	**0.9766**	0.9429	0.5881	**0.9975**	**0.9928**	**0.9882**	**0.9563**	**0.9733**	**0.9550**	0.9140
NPR	0.5987	0.8743	0.9419	0.5376	0.9638	0.9616	0.9599	0.9655	0.9549	0.9406	0.8699
Video-level AIGC detectors											
ReStraV	0.6736	0.7441	0.5328	0.3515	0.2353	0.3071	0.3207	0.5199	0.6339	0.3944	0.4713
DeMamba	0.8117	0.9756	**0.9491**	0.8118	0.9750	0.9355	0.9398	0.9684	0.8682	0.8503	0.9085
DeCoF	0.6667	0.4904	0.6033	0.5198	0.3588	0.3681	0.3705	0.5646	0.3353	0.5785	0.4856
General video classification models											
TimeSformer	0.6188	0.8144	0.7296	0.5274	0.9194	0.8080	0.8036	0.8453	0.6431	0.7007	0.7410
VideoMAE	0.6496	0.7841	0.7161	0.5093	0.6381	0.8298	0.7498	0.7105	0.6824	0.7332	0.7003
UniFormerV2	0.7237	0.8522	0.8150	0.5658	0.8056	0.6997	0.7231	0.6921	0.6143	0.5043	0.6996
Deepfake transfer (self-synthesized supervision)											
SBI	0.5561	0.5945	0.5932	0.4797	0.5885	0.5732	0.5903	0.4444	0.5052	0.4949	0.5420
LAA-Net	0.5810	0.6339	0.6671	0.6313	0.5723	0.5976	0.5792	0.6319	0.6319	0.6180	0.6144
Ours	**0.8434**	0.9532	0.9154	**0.8573**	0.9760	0.9301	0.9374	0.9503	0.8573	0.9433	**0.9164**

Table 2. Cross-generator and cross-paradigm generalization in video-level ROC-AUC. The best result in each column is highlighted in bold.

Method	2D Implicit				2D Explicit			3D			Cross-domain
	HeyGen	HunyuanAvatar	LongCat 1.5	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
Image-level AIGC detectors											
Effort	0.7910	0.9540	0.8630	0.7014	0.8815	0.7737	0.7823	0.7860	0.7007	0.3309	0.7565
DGS-Net	0.6846	0.9518	0.8594	0.8181	0.8361	0.7059	0.7364	0.7452	0.7683	0.5183	0.7624
NPR	0.6489	0.8467	0.9119	0.7459	0.8348	0.7662	0.8186	0.6930	0.5799	0.5458	0.7392
Video-level AIGC detectors											
ReStraV	0.6721	0.7186	0.4977	0.2977	0.0558	0.3214	0.0395	0.4781	0.6042	0.3731	0.4058
DeMamba	0.7966	**0.9893**	**0.9607**	0.8437	0.8020	0.5114	0.6048	0.8123	0.7573	0.3717	0.7450
DeCoF	0.7557	0.3915	0.6444	0.5284	0.2246	0.2335	0.2557	0.5906	0.5061	0.5910	0.4722
General video classification models											
TimeSformer	0.6685	0.8247	0.7310	0.4164	0.3620	0.2365	0.2660	0.4042	0.4012	0.2150	0.4526
VideoMAE	0.7166	0.8322	0.7717	0.5665	0.4269	0.3814	0.4004	0.5294	0.5936	0.6236	0.5842
UniFormerV2	0.7162	0.8320	0.8071	0.5417	0.4744	0.3271	0.3589	0.5943	0.5440	0.4017	0.5597
Deepfake transfer (self-synthesized supervision)											
SBI	0.4883	0.5031	0.5018	0.4218	0.4989	0.5088	0.4969	0.3581	0.4440	0.4727	0.4694
LAA-Net	0.6126	0.6712	0.8275	0.6697	0.6483	0.7847	0.7531	0.5057	0.6040	0.4592	0.6536
Ours	**0.8874**	0.9867	0.9481	**0.9159**	**0.9461**	**0.8652**	**0.9123**	**0.9287**	**0.9038**	**0.8657**	**0.9160**

5.3 Cross-Generator and Cross-Paradigm Generalization

While the *Cross-Generator Protocol* evaluates generalization across unseen generators with diverse training coverage, the *Cross-Paradigm Protocol* considers a more challenging setting where synthetic training data only covers **2D Implicit-Driven Generation**. As reported in Table 2 (with ACC@0.5 results in Appendix Table C2), most existing detectors experience substantial performance degradation under cross-paradigm transfer, despite achieving competitive results when generation mechanisms are observed during training. This indicates that their learned evidence is closely tied to specific generation characteristics rather than transferable forensic patterns. In contrast, HuRF maintains comparable performance to the *Cross-Generator Protocol*, achieving a mean AUC of 0.9160 across ten held-out generators, demonstrating robust transfer across unseen generation mechanisms. Notably, HuRF achieves an AUC of 0.8657 on LAM, and Appendix Table C4 shows that HSRD with HARW substantially improves LAM detection, indicating the effectiveness of human-part relation modeling for geometry-driven 3D avatars.

5.4 In-the-Wild Evaluation

We further evaluate HuRF under real-world conditions using independently collected JDAvatar and BaiduAvatar videos, whose generation pipelines are unknown. Under the more challenging *Cross-Paradigm Protocol*, HuRF achieves AUCs of 0.8366 and 0.8387 on JDAvatar and BaiduAvatar, respectively, with a mean AUC of 0.8377, substantially outperforming the strongest baseline (0.6534). These results demonstrate that HuRF maintains robust generalization beyond controlled generators. Full results under both protocols are provided in Appendix Table C3.

5.5 External-Dataset Generalization

To further evaluate generalization beyond HuDigivid, we evaluate HuRF on the external DigiFakeAV dataset. As reported in Table 3, HuRF achieves an AUC of 0.8594 and ACC@0.5 of 0.7810, obtaining the best performance among all compared methods and demonstrating transferable human-centric forensic evidence beyond HuDigivid.

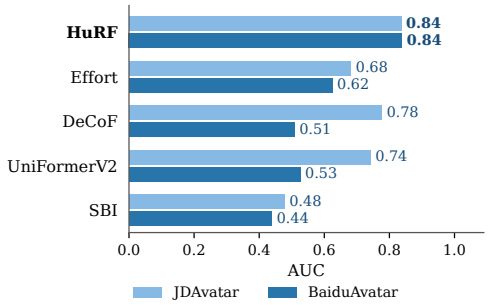


Figure 5: In-the-wild AUC under the *Cross-Paradigm Protocol* on JDAvatar and BaiduAvatar. We report the best in-the-wild baseline from each family together with HuRF.

Table 3. DigiFakeAV results using final *Cross-Paradigm Protocol* checkpoints (436 real and 436 fake videos). Best results are bold.

Method	AUC	ACC@0.5
Effort	0.8519	0.7534
DGS-Net	0.7906	0.6927
NPR	0.7996	0.7156
ReStraV	0.1199	0.1709
DeMamba	0.8497	0.7649
DeCoF	0.7261	0.6479
TimeSformer	0.5759	0.5459
VideoMAE	0.6293	0.5344
UniFormerV2	0.7560	0.6445
SBI	0.3966	0.4346
LAA-Net	0.5419	0.5195
HuRF	**0.8594**	**0.7810**

5.6 Ablation Study

Component ablation. Table 4(a) shows that the classification-only baseline achieves 0.8118 AUC. HSRD alone provides limited improvement, while introducing HARW further increases performance to 0.8337 AUC, demonstrating the benefit of emphasizing human-related semantic relations. TAC alone achieves 0.8586 AUC, confirming the effectiveness of adaptive temporal calibration. Combining semantic relation learning and temporal calibration achieves 0.9160 AUC, validating their complementary contributions. Detailed per-domain results are provided in Appendix Table C4.

Table 4: Ablation results.

(a) Components

HSRD	HARW	TAC	AUC
–	–	–	0.8118
✓	–	–	0.8115
✓	✓	–	0.8337
–	–	✓	0.8586
✓	✓	✓	**0.9160**

(b) Training strategy

Strategy	AUC
Stage 1 only (no TAC)	0.8337
End-to-end	0.7889
Two-stage	**0.9160**

Training strategy. Table 4(b) shows that end-to-end optimization achieves 0.7889 AUC, even lower than Stage-1-only training (0.8337), while the proposed two-stage strategy reaches 0.9160 AUC. This verifies that separating human-centric representation adaptation from temporal modeling leads to more effective optimization. Detailed per-domain results are provided in Appendix Table C5.

Additional ablations are provided in Appendix Tables C6 and C7. The backbone-scale study validates our choice of the 1B teacher–1B student configuration. The TAC placement study demonstrates the importance of applying temporal adaptive calibration across all 32 Transformer blocks.

6 Conclusion

We presented HuRF, a human-centric representation learning framework for generalizable native digital human video forensics. Motivated by the instability of generator-specific forensic cues across heterogeneous generation mechanisms, HuRF organizes forensic evidence around the shared human subject. It leverages pose-oriented human representations to capture appearance and structural information, HSRD to internalize human-part semantic relations while suppressing background interference, and TAC to adaptively calibrate temporal evidence across different generation mechanisms.

To systematically evaluate this problem, we constructed HuDigivid, covering 19 known generators across three generation paradigms together with in-the-wild digital human videos. Extensive experiments reveal that existing detectors can achieve competitive performance under cross-generator evaluation but suffer substantial degradation under cross-paradigm and in-the-wild scenarios, indicating their limited transferability across heterogeneous digital human generation mechanisms. In contrast, HuRF achieves strong cross-generator, cross-paradigm, in-the-wild, and cross-dataset generalization while remaining robust to common video degradations. These results demonstrate that human-centric representations provide transferable forensic evidence beyond generation-specific artifacts for native digital human video forensics.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript preparation, including organizing writing structure, improving language clarity, and translating text. AI tools were also used for code-related assistance during the research process. The authors take full responsibility for the final content of this paper, including all claims, experimental results, and conclusions.

ETHICS STATEMENT

The HuDigivid dataset is intended solely for research on media authenticity and digital human forensics. Data collection and release will respect the licenses and terms of the original datasets and platforms. We do not use the dataset for identity inference, and any public release will exclude restricted source media when redistribution is not permitted, providing derived metadata or access instructions instead.

REPRODUCIBILITY STATEMENT

The composition of HuDigivid, held-out generator assignments, evaluation protocols, model design, optimization settings, and additional ablations are documented in the main paper and appendix. Code, split metadata, and releasable HuDigivid dataset assets will accompany the final publication.

REFERENCES

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.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 7414–7423, 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.

Weiliang Chen, Wenzhao Zheng, Yu Zheng, Lei Chen, Jie Zhou, Jiwen Lu, and Yueqi Duan. Genworld: Towards detecting ai-generated real-world simulation videos. *arXiv preprint arXiv:2506.10975*, 2025a.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025b.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

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.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19207–19217. IEEE, 2025.

Yikang Ding, Jiwen Liu, Wenyuan Zhang, Zekun Wang, Wentao Hu, Liyuan Cui, Mingming Lao, Yingchao Shao, Hui Liu, Xiaohan Li, et al. Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. *arXiv preprint arXiv:2509.09595*, 2025.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. *arXiv preprint arXiv:2502.17796*, 2025.

HeyGen. HeyGen avatar video generation platform. `https://www.heygen.com/avatars`, 2026. Accessed: 2026-09-08.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 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.

Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12753–12762, 2021.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. In *International Conference on Learning Representations*, volume 2026, pp. 58484–58507, 2026.

Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, et al. Skyreels-v3 technique report. *arXiv preprint arXiv:2601.17323*, 2026a.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1632–1643. IEEE, 2023.

Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10747–10757. IEEE, 2025.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213. IEEE, 2020.

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. *arXiv preprint arXiv:2604.04634*, 2026b.

Jiaxin Liu, Jia Wang, Saihui Hou, Min Ren, Huijia Wu, Long Ma, Renwang Pei, and Zhaofeng He. Beyond face swapping: A diffusion-based digital human benchmark for multimodal deepfake detection. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14087–14091. IEEE, 2026.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting ai-generated video via frame consistency. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

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.

Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17395–17405. IEEE, 2024.

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*, volume 40, pp. 15582–15590, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 666–676. IEEE, 2024.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18699–18708. IEEE, 2022.

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.

Shuai Tan, Biao Gong, Zhuoxin Liu, Yan Wang, Xi Chen, Yifan Feng, and Hengshuang Zhao. Animate-x++: Universal character image animation with dynamic backgrounds. *arXiv preprint arXiv:2508.09454*, 2025.

Meituan LongCat Team, Xunliang Cai, Meng Cheng, Feng Gao, Zhe Kong, Jiamu Li, Le Li, Weiheng Li, Hongyu Liu, Shuai Tan, et al. Longcat-video-avatar 1.5 technical report. *arXiv preprint arXiv:2605.26486*, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21096–21106. IEEE, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer, 2025. URL https://arxiv.org/abs/2504.11289.

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*, 2025.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22601–22611. IEEE, 2023.

Zhipei Xu, Xuanyu Zhang, Qing Huang, Xing Zhou, and Jian Zhang. Avatarshield: Visual reinforcement learning for human-centric synthetic video detection. *International Journal of Computer Vision*, 134(9):402, 2026.

Jiazhen Yan, Ziqiang Li, Fan Wang, Boyu Wang, Ziwen He, and Zhangjie Fu. Dgs-net: Distillation-guided gradient surgery for clip fine-tuning in ai-generated image detection. *arXiv preprint arXiv:2511.13108*, 2025a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*, 2023.

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.

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12615–12625. IEEE, 2025b.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4212–4222. IEEE, 2023.

Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Kangjie Chen, Minghan Qin, Yu Li, and Haoqian Wang. Hravatar: High-quality and relightable gaussian head avatar. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26285–26296. IEEE, 2025a.

Jiaming Zhang, Shengming Cao, Rui Li, Xiaotong Zhao, Yutao Cui, Xinglin Hou, Gangshan Wu, Haolan Chen, Yu Xu, Limin Wang, et al. Steadydancer: Harmonized and coherent human image animation with first-frame preservation. *arXiv preprint arXiv:2511.19320*, 2025b.

Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5535–5545. IEEE, 2025c.

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.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3660–3669. IEEE, 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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1979–1989. IEEE, 2025.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

A Detailed Dataset Composition

A.1 Data Generation and Access

Reference sources. HuDigivid constructs the controlled digital-human subsets using reference sources matched to the input requirements of different generation paradigms. For 2D Implicit-Driven Generation, reference images and audio conditions are primarily obtained from selected public-platform videos and public datasets. For 2D Explicit-Guided Generation, we select public TikTok videos as driving sources due to their diverse human motions and large-amplitude movements. We extract frame-wise DWPose sequences from these videos and combine them with reference images to provide explicit motion and identity conditions for generation. For 3D Avatar-Rendered Generation, we use HDTF and other generator-specific public or officially provided assets, with HDTF providing high-quality frontal human videos with stable identity appearance for identity reconstruction and avatar rendering. Table A1 summarizes the generator-wise reference inputs, source data, and access policy.

Table A1: Reference inputs, source data, and planned access for the 19 controlled digital-human subsets. Public access remains subject to the licenses and redistribution policies of the original sources.

Generator	Paradigm	Identity / Reference Input	Input Form	Reference Source	Access
InfiniteTalk	2D implicit	Reference image	Image + audio	Public-platform real videos	Controlled
MagiHuman	2D implicit	Reference image	Image + audio	DFDC	Public
Kling Avatar	2D implicit	Reference image	Image + audio	Public-platform real videos	Controlled
SkyReels V3	2D implicit	Reference image	Image + audio	Public-platform real videos	Controlled
Wan 2.2	2D implicit	Reference image	Image + audio	Public-platform real videos	Controlled
HeyGen	2D implicit	Official example avatar	Avatar + speech	Official asset library	Public
HunyuanVideo-Avatar	2D implicit	Reference image	Image + audio	Public-platform real videos	Controlled
LongCat-Video-Avatar 1.5	2D implicit	Reference image / video	Image / video + audio	DFDC	Public
Wan-Animate	2D implicit	Reference image	Image + driving video	CelebV-HQ	Public
StableAnimator	2D explicit-guided	Reference image	Image + DWPose sequence	Public TikTok dataset	Public
SteadyDancer	2D explicit-guided	Reference image	Image + DWPose sequence	Public TikTok dataset	Public
Animate-X++	2D explicit-guided	Reference image	Image + DWPose sequence	Public TikTok dataset	Public
MimicMotion	2D explicit-guided	Reference image	Image + DWPose sequence	Public TikTok dataset	Public
UniAnimateDiT	2D explicit-guided	Reference image	Image + DWPose sequence	Public TikTok dataset	Public
SyncTalk	3D (NeRF)	Monocular video	Monocular-video modeling	HDTF	Public
HRAvatar	3D (3DGS)	Reference image	Image-based input	LipSynthesis public assets	Public
FateAvatar	3D (3DGS)	Monocular video	Monocular-video modeling	HDTF	Public
LAM	3D (3DGS)	Reference image	Single-image modeling	HDTF	Public
RGBAvatar	3D (3DGS)	Monocular video	Monocular-video modeling	HDTF	Public

Reference-source diversity. HuDigivid includes complementary reference conditions across generators. Several generators within the same paradigm draw from common source pools, allowing different synthesis methods to be compared under partially controlled source conditions. At the same time, the held-out test set also contains generators whose reference sources are independent of those used for *Cross-Paradigm Protocol* training. In particular, HeyGen uses officially provided avatar assets, whereas Wan-Animate uses CelebV-HQ (Zhu et al., 2022). Despite these distinct reference sources, HuRF achieves AUCs of 0.8874 and 0.9159 on HeyGen and Wan-Animate, respectively, under the *Cross-Paradigm Protocol* (Table 2). These results support transfer to held-out generators with different reference sources, although AUC alone cannot establish which source or identity factors drive the transfer.

Data access and privacy. HuDigivid adopts a hybrid release policy according to the provenance and redistribution conditions of its reference data. Subsets constructed from public datasets or officially provided assets are planned for public release when permitted by the corresponding licenses. For InfiniteTalk, Kling Avatar, SkyReels V3, Wan 2.2, and HunyuanVideo-Avatar, whose reference identities are derived from real human videos collected from public platforms, the generated subsets will be distributed through controlled access. All releases remain subject to the licenses and redistribution policies of the original data sources.

A.2 Dataset Composition

HuDigivid consists of controlled digital human videos generated by 19 known methods, two independently collected in-the-wild digital human subsets, and a multi-source real-video pool. Table A2 summarizes the digital-human subsets in terms of generation paradigm, source type, and video count, while Table A3 reports the detailed composition of the real-video pool.

Table A2. Composition of HuDigivid.

Video Source / Generator	Type / Paradigm	Source	CG	CP	Videos
Fake videos					
InfiniteTalk (Yang et al., 2025)	2D implicit	Open source	Train	Train	4,574
MagiHuman (Chern et al., 2026)	2D implicit	Open source	Train	Train	964
Kling Avatar (Ding et al., 2025)	2D implicit	API	Train	Train	429
SkyReels V3 (Li et al., 2026a)	2D implicit	Open source	Train	Train	2,349
Wan 2.2 (Wan et al., 2025)	2D implicit	Open source	Train	Train	3,196
HeyGen (HeyGen, 2026)	2D implicit	API	Test	Test	601
HunyuanVideo-Avatar (Chen et al., 2025b)	2D implicit	Open source	Test	Test	827
LongCat-Video-Avatar 1.5 (Team et al., 2026)	2D implicit	Open source	Test	Test	1,500
Wan-Animate (Cheng et al., 2025)	2D implicit	Open source	Test	Test	362
StableAnimator (Tu et al., 2025)	2D explicit-guided	Open source	Test	Test	336
SteadyDancer (Zhang et al., 2025b)	2D explicit-guided	Open source	Train	–	339
Animate-X++ (Tan et al., 2025)	2D explicit-guided	Open source	Test	Test	462
MimicMotion (Zhang et al., 2024)	2D explicit-guided	Open source	Train	–	837
UniAnimateDiT (Wang et al., 2025)	2D explicit-guided	Open source	Test	Test	457
SyncTalk (Peng et al., 2024)	3D (NeRF)	Open source	Test	Test	241
HRAvatar (Zhang et al., 2025a)	3D (3DGS)	Open source	Test	Test	511
FateAvatar (Zhang et al., 2025c)	3D (3DGS)	Open source	Train	–	493
LAM (He et al., 2025)	3D (3DGS)	Open source	Test	Test	398
RGBAvatar (Li et al., 2025)	3D (3DGS)	Open source	Train	–	399
In-the-wild digital human videos					
JDAvatar	Unknown	Platform-collected	Test	Test	481
BaiduAvatar	Unknown	Platform-collected	Test	Test	472
Overall Digital Human (19 generators + in-the-wild)			–	–	**20,228**

CG/CP: **Cross-Generator/Cross-Paradigm** Protocols. Train/Test/– indicate training/testing/unused. *Source* indicates generator availability or collection origin.

Table A3. Detailed composition of the real-video subset.

Video Source	CG	CP	Videos
JD Platform Real	Test	Test	3,409
YouTube News	Train	Train	2,655
Douyin	Train	Train	4,077
TikTok (platform-collected)	Train	Train	1,292
Bigo	Train	Train	1,035
CCTV News	Train	Train	535
DFDC	Train	Train	895
HDTF	Train	–	400
Public TikTok datasets	Train	–	1,073
Overall Real	–	–	**15,371**

CG/CP: **Cross-Generator/Cross-Paradigm** Protocols. JD Platform Real is used for testing; other sources are used for training, with HDTF and Public TikTok datasets excluded from CP. – denotes unused; counts cover the full real-video pool.

B Supporting Analyses and Results

B.1 Zero-Shot Transfer Analysis

Existing AIGV and deepfake detectors are developed for general synthetic videos or edited human videos. A natural question is whether native digital humans can be reliably detected by directly transferring these existing forensic paradigms without task-specific adaptation. This analysis investigates whether digital human forensics can be treated as a straightforward extension of existing detection tasks or requires dedicated representations for its unique generation mechanisms. We evaluate seven pretrained detectors from three families on the frozen, quality-reviewed ten-domain protocol without retraining on HuDigivid. Table B1 reports video-level AUC using public checkpoints and fixed calibration where applicable.

Image-level AIGC detectors. NPR (Tan et al., 2024), evaluated with its official ProGAN-trained checkpoint, reaches a ten-domain mean AUC of 0.3777. Its performance ranges from 0.2530 on HunyuanAvatar to 0.6385 on Wan-Animate, indicating limited zero-shot transfer and substantial variation across generators. UniversalFakeDetect (UnivFD) (Ojha et al., 2023), using its official ProGAN-trained CLIP ViT-L/14 checkpoint, achieves a higher mean AUC of 0.7444, but still varies from 0.5302 on LongCat to 0.9533 on LAM.

Video-level AIGC detectors. WaveRep G4 (Corvi et al., 2026), D3 (Zheng et al., 2025), and STALL (Ben Hayun et al., 2026) achieve mean AUCs of 0.5763, 0.7053, and 0.4973, respectively. D3 performs comparatively well on both 2D paradigms (0.7301 implicit and 0.7585 explicit-guided), but drops to 0.6190 on 3D videos. STALL exhibits a sharper contrast, reaching 0.9017 on 3D generation but only 0.1261 on explicit-guided generation; WaveRep G4 instead performs better on explicit-guided generation (0.6892) than on 3D generation (0.4692). Thus, video-level modeling alone does not ensure reliable transfer across generation mechanisms.

Deepfake detectors. LAA-Net (Nguyen et al., 2024) and ForAda (Cui et al., 2025) achieve mean AUCs of 0.6858 and 0.6766. Both are stronger on 3D generation (0.8408 and 0.8269) than on 2D implicit generation (0.6046 and 0.5593). One possible explanation is that rendering and compositing artifacts in some 3D videos resemble facial-manipulation cues encountered during pretraining, whereas such cues are less transferable to implicit generation. Overall, the results reveal family- and paradigm-dependent transfer, not a uniform failure on all digital humans.

Table B1. Zero-shot performance of seven pretrained detectors on the Shared Test Set, grouped by detector family (video-level AUC).

Method	HeyGen	HunyuanAvatar	LongCat	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
Image-level AIGC detectors											
NPR	0.3846	0.2530	0.3006	0.6385	0.3695	0.2831	0.2939	0.3736	0.3502	0.5296	0.3777
UnivFD	0.6875	0.7915	0.5302	0.5975	0.7802	0.7232	0.6413	0.8992	0.8398	0.9533	0.7444
Video-level AIGC detectors											
WaveRep G4	0.6667	0.4139	0.4325	0.7745	0.8292	0.6305	0.6079	0.3718	0.3611	0.6746	0.5763
D3	0.5089	0.8256	0.8350	0.7510	0.7421	0.7617	0.7717	0.6842	0.5028	0.6700	0.7053
STALL	0.5261	0.6395	0.1850	0.5384	0.1384	0.1257	0.1143	0.8415	0.9237	0.9399	0.4973
Deepfake detectors											
LAA-Net	0.4095	0.9118	0.3359	0.7612	0.5262	0.7257	0.6652	0.8241	0.9340	0.7641	0.6858
ForAda	0.3664	0.7251	0.5362	0.6094	0.5646	0.7676	0.7166	0.6949	0.9569	0.8288	0.6766

B.2 Evolution of Human-Centric Representation Design

HuRF is built on the hypothesis that human-specific structural information can provide useful forensic evidence beyond generator-dependent artifacts. To determine how such information should be represented, we progressively investigate explicit pose signals, RGB appearance representations, their direct fusion, and finally hierarchical representations learned by a pose-oriented human foundation model.

Explicit pose as a forensic cue. We first investigate whether human pose alone contains useful evidence for digital human forensics. As shown in Table B2, directly modeling 308-point pose coordinates yields a mean AUC of 0.4493, while converting them into Gaussian pose maps improves the mean AUC to 0.6050. More importantly, Gaussian pose maps achieve 0.7583, 0.7435, and 0.7329 AUC on HeyGen, HunyuanAvatar, and LongCat, respectively, showing that human structure and its temporal evolution can already provide substantial discriminative information for several 2D implicit generators. In contrast, performance drops to 0.5170, 0.5288, and 0.5404 on StableAnimator, Animate-X++, and UniAnimateDiT. This behavior is consistent with their explicitly-guided generation mechanism, where human motion is directly driven by pose sequences extracted from real videos and is therefore less discriminative for authenticity. Pose representations are also less effective on the 3D domains, particularly for head-centric avatars where full-body structural information is limited. These observations show that explicit pose provides meaningful human-centric forensic cues, but its effectiveness is constrained by both the generation mechanism and the information preserved by the pose representation.

RGB appearance and RGB–pose fusion. Explicit pose is a compressed geometric representation that discards substantial appearance, texture, local-shape, and rendering information. We therefore also examine the RGB representation from Sapiens2-Pretrain. It achieves 0.6875 mean AUC and performs strongly on several domains, including 0.7861, 0.8047, and 0.7943 on HeyGen, Hunyua-

nAvatar, and LongCat, respectively. Since pose and RGB contain fundamentally different types and amounts of information, we naturally ask whether their forensic evidence can be combined. We therefore fuse Gaussian pose maps with RGB features through either cross-attention or feature concatenation. However, the two variants achieve mean AUCs of 0.6671 and 0.6408, respectively, both below the RGB-only Sapiens2-Pretrain baseline. Although fusion improves several individual domains, it does not provide a consistent overall gain. One possible reason is that explicit pose is a lossy geometric abstraction and resides in a substantially different representation space from dense RGB features, making their forensic evidence difficult to integrate through straightforward fusion.

From explicit pose to pose-oriented human representations. These results motivate us to seek a representation that retains dense RGB information while internalizing human structural priors within the same feature space. We therefore adopt Sapiens2-Pose, which inherits a human-oriented visual representation from Sapiens2 pretraining and further acquires fine-grained pose and body-topology awareness through pose-oriented training while still taking RGB as input. As discussed in Sec. 4.2, an appropriate forensic representation should balance low-level appearance cues and high-level human semantics. We therefore use the intermediate block 32 representation as the forensic feature, which is expected to preserve dense appearance information while avoiding the stronger pose-task bias of the final layer. As shown in Table B2, Sapiens2-Pose block 32 achieves 0.8118 mean AUC, substantially outperforming the final-layer representation (0.7154), supporting the effectiveness of this intermediate representation for forensic transfer. After LoRA fine-tuning, this representation further improves over Sapiens2-Pretrain RGB (0.6875) and explicit RGB–pose fusion variants. These results indicate that pose-oriented intermediate representations provide a suitable balance between visual appearance and human structural awareness for digital human forensics.

Overall, this analysis motivates the representation design of HuRF. Explicit pose verifies that human structure itself contains useful forensic information, but its sparse and lossy representation limits transfer across heterogeneous generation paradigms. RGB retains richer visual evidence, yet straightforward RGB–pose fusion does not consistently improve performance. By internalizing pose-oriented human priors within a dense RGB representation, intermediate Sapiens2-Pose features provide a more effective foundation for subsequent forensic adaptation.

Table B2: Pose and RGB–Pose representation comparison on the Shared Test Set (video-level AUC).

Method	HeyGen	HunyuanAvatar	LongCat	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
Sapiens2-Pretrain (RGB)	0.7861	0.8047	0.7943	0.7007	0.6628	0.5137	0.5453	0.7333	0.7453	0.5893	0.6875
308-point pose coordinates	0.6303	0.6389	0.5616	0.4679	0.2057	0.2378	0.2076	0.5585	0.4840	0.5008	0.4493
Gaussian Pose Map	0.7583	0.7435	0.7329	0.5770	0.5170	0.5288	0.5404	0.5957	0.5580	0.4982	0.6050
RGB + Gaussian Pose (cross-attention)	0.7279	0.8098	0.7421	0.6091	0.8176	0.6337	0.6416	0.6833	0.5781	0.4282	0.6671
RGB + Gaussian Pose (feature concatenation)	0.7096	0.7796	0.6969	0.5829	0.7767	0.5927	0.5962	0.6654	0.5659	0.4421	0.6408
Sapiens2-Pose (RGB, layer 32)	0.8470	0.9474	0.8901	0.8059	0.8204	0.6612	0.7388	0.8284	0.8705	0.7079	0.8118
Sapiens2-Pose (RGB, final layer)	0.7882	0.8490	0.7929	0.7425	0.5803	0.4724	0.5399	0.7810	0.8521	0.7562	0.7154

C Detailed Performance Results

C.1 Detailed Generalization Results

Table C1. Held-out-generator generalization under the *Cross-Generator Protocol* in video-level class-balanced ACC@0.5. The best result in each column is highlighted in bold.

Method	2D Implicit				2D Explicit			3D			Cross-domain
	HeyGen	HunyuanAvatar	LongCat 1.5	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
Image-level AIGC detectors											
Effort	0.6639	0.8476	0.8213	0.5235	0.9137	0.9156	0.9083	0.8943	0.7614	0.7125	0.7962
DGS-Net	0.5973	**0.9220**	0.8390	0.4738	**0.9583**	**0.9535**	**0.9519**	0.8699	**0.8983**	0.8820	0.8346
NPR	0.5300	0.7757	0.8687	0.4751	0.9122	0.9058	0.9181	0.9051	0.8900	0.8695	0.8050
Video-level AIGC detectors											
ReStraV	0.6431	0.6729	0.5397	0.3660	0.2649	0.2944	0.3244	0.5470	0.6452	0.3373	0.4635
DeMamba	0.6714	0.9069	**0.8823**	0.6768	0.9286	0.8777	0.8820	**0.9295**	0.7303	0.6195	0.8105
DeCoF	0.5890	0.5048	0.5767	0.5166	0.4271	0.4459	0.4450	0.5431	0.4108	0.6114	0.5070
General video classification models											
TimeSformer	0.5674	0.7382	0.6543	0.4986	0.8065	0.7500	0.7301	0.7632	0.5830	0.6386	0.6730
VideoMAE	0.5541	0.6983	0.6283	0.4890	0.5610	0.7392	0.6644	0.6350	0.5560	0.6410	0.6166
UniFormerV2	0.6314	0.7594	0.7143	0.5345	0.6979	0.5465	0.5647	0.6125	0.5145	0.4725	0.6048
Deepfake transfer (self-synthesized supervision)											
SBI	0.5582	0.5713	0.5520	0.5083	0.5565	0.5433	0.5639	0.4667	0.5228	0.5017	0.5345
LAA-Net	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
Ours	**0.7188**	0.8845	0.8470	**0.7376**	0.9003	0.8561	0.8678	0.8806	0.7510	**0.9047**	**0.8348**

Table C2. Cross-generator and cross-paradigm generalization in video-level class-balanced ACC@0.5. The best result in each column is highlighted in bold.

Method	2D Implicit				2D Explicit			3D			Cross-domain
	HeyGen	HunyuanAvatar	LongCat 1.5	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
Image-level AIGC detectors											
Effort	0.6689	0.8930	0.7550	0.5787	0.7798	0.5909	0.6106	0.6614	0.6120	0.4486	0.6599
DGS-Net	0.5383	0.8761	0.7077	0.6671	0.6815	0.5119	0.5242	0.5744	0.6079	0.4374	0.6126
NPR	0.5682	0.7678	0.8423	0.6685	0.7693	**0.6699**	0.7399	0.5841	0.5083	0.4526	0.6571
Video-level AIGC detectors											
ReStraV	0.6281	0.6348	0.4743	0.3522	0.2426	0.4513	0.2533	0.4736	0.5892	0.2896	0.4389
DeMamba	0.6564	**0.9474**	**0.8947**	0.7169	0.6890	0.5130	0.5471	0.7339	0.6598	0.4888	0.6847
DeCoF	0.6772	0.4202	0.6023	0.5124	0.2976	0.2965	0.2937	0.5744	0.5124	**0.5923**	0.4779
General video classification models											
TimeSformer	0.6106	0.7430	0.6570	0.4420	0.4033	0.3647	0.3712	0.4344	0.4315	0.3559	0.4814
VideoMAE	0.6115	0.7479	0.6300	0.5110	0.4167	0.4297	0.4247	0.4648	0.5145	0.4763	0.5227
UniFormerV2	0.6057	0.7430	0.7017	0.5290	0.4777	0.4545	0.4575	0.5460	0.4793	0.4587	0.5453
Deepfake transfer (self-synthesized supervision)											
SBI	0.4809	0.5042	0.4920	0.4544	0.4881	0.4881	0.4929	0.3885	0.4481	0.4605	0.4698
LAA-Net	0.5666	0.5931	0.6047	0.5870	0.6146	0.6126	0.6081	0.5137	0.5705	0.4897	0.5760
Ours	**0.6897**	**0.9474**	0.8503	**0.7472**	**0.8467**	0.6396	**0.7430**	**0.7681**	**0.7365**	0.5390	**0.7508**

Table C3. Full in-the-wild video-level ROC-AUC under both protocols. The best result in each column is highlighted in bold.

	Cross-Generator			Cross-Paradigm		
Method	JDAvatar	BaiduAvatar	Mean	JDAvatar	BaiduAvatar	Mean
Image-level AIGC detectors						
Effort	0.7536	0.6084	0.6810	0.6819	0.6248	0.6534
DGS-Net	0.4629	0.5420	0.5024	0.4058	0.6509	0.5283
NPR	0.4254	0.5497	0.4876	0.4702	0.4680	0.4691
Video-level AIGC detectors						
ReStraV	0.5980	0.4830	0.5405	0.6030	0.4823	0.5427
DeMamba	0.6695	0.5647	0.6171	0.6085	0.6028	0.6057
DeCoF	0.7630	0.4323	0.5977	0.7766	0.5086	0.6426
General video classification models						
TimeSformer	0.6293	0.4393	0.5343	0.6184	0.4044	0.5114
VideoMAE	0.6683	0.4137	0.5410	0.6377	0.4098	0.5237
UniFormerV2	0.7648	0.5430	0.6539	0.7420	0.5263	0.6342
Deepfake transfer (self-synthesized supervision)						
SBI	0.5025	0.4528	0.4776	0.4779	0.4392	0.4586
LAA-Net	0.4675	0.5118	0.4897	0.4463	0.4383	0.4423
Ours	**0.8491**	**0.6945**	**0.7718**	**0.8366**	**0.8387**	**0.8377**

C.2 Detailed Ablation Results

Table C4: Component ablation on the Shared Test Set (video-level ROC-AUC).

HSRD	HARW	TAC	HeyGen	HunyuanAvatar	LongCat	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
–	–	–	0.8470	0.9474	0.8901	0.8059	0.8204	0.6612	0.7388	0.8284	0.8705	0.7079	0.8118
✓	–	–	0.8937	0.9192	0.8691	0.8459	0.7341	0.6018	0.6624	0.8474	0.8940	0.8474	0.8115
✓	✓	–	**0.9039**	0.9254	0.8796	0.8607	0.7935	0.6611	0.7193	0.8564	0.8877	0.8495	0.8337
–	–	✓	0.8443	0.9759	0.9235	0.8577	0.9337	0.7768	0.8464	0.8513	0.8556	0.7203	0.8586
✓	✓	✓	0.8874	**0.9867**	**0.9481**	**0.9159**	**0.9461**	**0.8652**	**0.9123**	**0.9287**	**0.9038**	**0.8657**	**0.9160**

Table C5: Training strategy ablation on the Shared Test Set (video-level ROC-AUC).

Training strategy	TAC	HeyGen	HunyuanAvatar	LongCat	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
Baseline	–	**0.9039**	0.9254	0.8796	0.8607	0.7935	0.6611	0.7193	0.8564	0.8877	0.8495	0.8337
End-to-End	✓	0.8007	0.9413	0.8797	0.7458	0.8581	0.6754	0.7552	0.8142	0.7871	0.6311	0.7889
Two-Stage	✓	0.8874	**0.9867**	**0.9481**	**0.9159**	**0.9461**	**0.8652**	**0.9123**	**0.9287**	**0.9038**	**0.8657**	**0.9160**

Table C6. Backbone-scale ablation on the Shared Test Set (video-level ROC-AUC).

Sapiens2-Seg Teacher	Sapiens2-Pose Student	HeyGen	HunyuanAvatar	LongCat	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
0.4B	1B	0.8473	0.9359	0.8213	0.8266	0.8526	0.6457	0.7344	0.8037	0.7760	0.7434	0.7987
0.8B	1B	0.8733	0.9773	0.8949	0.8871	0.8990	0.7562	0.8308	0.8774	0.8711	0.8034	0.8670
1B	0.4B	0.7446	0.9524	0.8528	0.7690	0.8661	0.6200	0.7210	0.7867	0.6987	0.5661	0.7577
1B	0.8B	0.7028	0.8946	0.7522	0.5976	0.7358	0.4508	0.5584	0.5723	0.5258	0.3170	0.6107
1B	**1B**	**0.8874**	**0.9867**	**0.9481**	**0.9159**	**0.9461**	**0.8652**	**0.9123**	**0.9287**	**0.9038**	**0.8657**	**0.9160**
5B	1B	0.8816	0.9538	0.9237	0.8088	0.8800	0.7497	0.8146	0.8313	0.8360	0.7369	0.8416
1B	5B	0.7493	0.9108	0.8902	0.7261	0.7567	0.6032	0.6736	0.7069	0.6741	0.4938	0.7185

Table C7: Ablation on the number of TAC modules on the Shared Test Set (video-level ROC-AUC).

TAC placement	# TAC	HeyGen	HunyuanAvatar	LongCat	Wan-Animate	StableAnimator	Animate-X++	UniAnimateDiT	HRAvatar	SyncTalk	LAM	Mean
Final block	1	**0.8982**	0.9152	0.8709	0.8467	0.7864	0.6490	0.7091	0.8394	0.8718	0.8312	0.8218
Last 6 blocks	6	0.8692	0.9711	0.8953	0.7879	0.9113	0.7257	0.8119	0.8195	0.7806	0.6919	0.8264
All 32 blocks	32	0.8874	**0.9867**	**0.9481**	**0.9159**	**0.9461**	**0.8652**	**0.9123**	**0.9287**	**0.9038**	**0.8657**	**0.9160**

D Robustness to Video Degradation

Digital human videos often undergo post-processing operations during distribution. We therefore evaluate the robustness of HuRF and representative baselines under four common degradation types, including H.264 compression, JPEG compression, downsampling, and upsampling. As shown in Fig. 6, HuRF maintains consistently high performance across different degradation levels, while several existing detectors exhibit larger degradation under severe compression and resizing. These results indicate that HuRF relies on more stable human-centric forensic evidence rather than fragile generation-specific artifacts, enabling robust detection after common video post-processing operations.

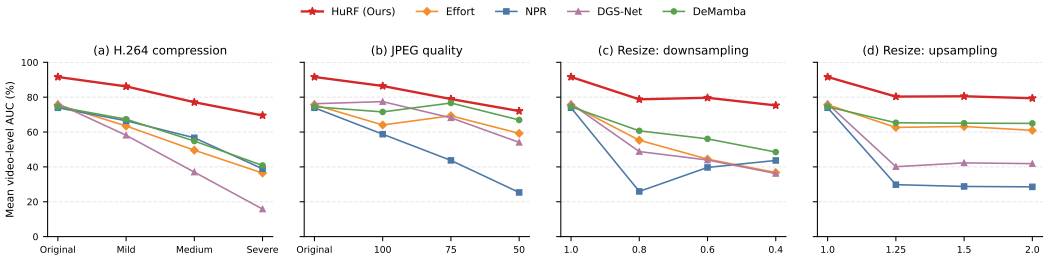


Figure 6: Robustness to H.264 compression, JPEG compression, downsampling, and upsampling on the Shared Test Set (mean video-level AUC).

E Human-centric Attention Visualization

Human-centric attention visualization. To qualitatively examine whether HuRF leverages human-centric representations for forensic prediction, we visualize the attention maps from the final pose-oriented representation. As shown in Fig. 7, HuRF consistently attends to informative human regions, including facial areas, hands, and body structures, while reducing irrelevant background responses. These observations are consistent with our design objective of organizing forensic evidence around the human subject rather than generation-specific or non-human cues.

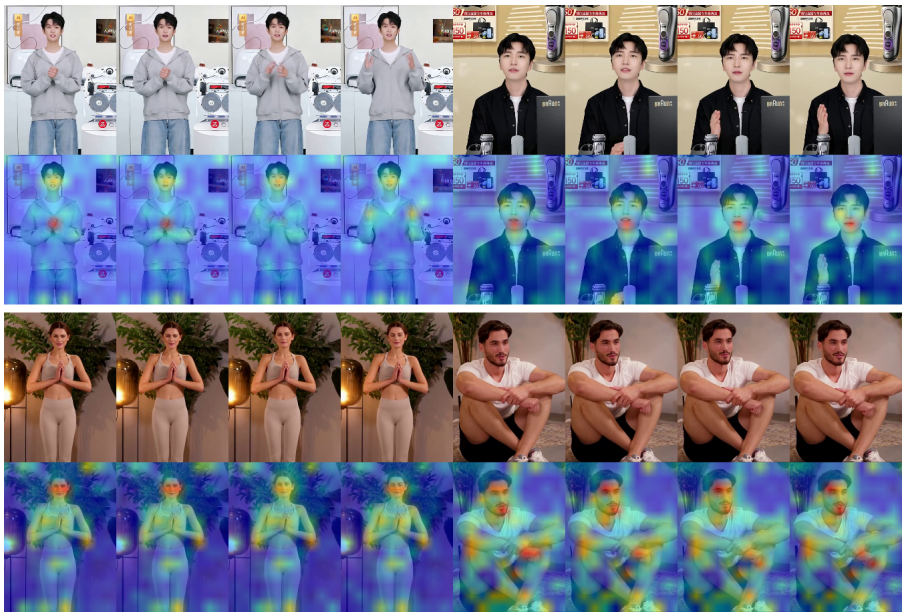


Figure 7: Qualitative visualization of block-32 patch-to-patch incoming attention on representative JD and HeyGen fake clips. Each sample shows four input frames followed by the corresponding attention overlays. The maps are head-averaged and normalized within each clip.

F VISUALIZATION OF HUDIGIVID

To provide an intuitive overview of HuDigivid, we visualize representative samples from the three digital human generation paradigms covered by our benchmark. As shown in Fig. 8, the examples exhibit diverse identities, appearances, backgrounds, and temporal motions, illustrating the range of digital human videos considered in HuDigivid and the different conditioning mechanisms used across the three paradigms. Each row contains eight frames sampled uniformly from the first five seconds of one video; all frames are center-cropped to 224×224 pixels for consistent visualization.

(a) 2D Implicit-Driven Generation



(b) 2D Explicit-Guided Generation



(c) 3D Avatar-Rendered Generation



Figure 8: Representative HuDigivid samples grouped by generation paradigm. Each row corresponds to one video and contains eight uniformly sampled frames; all frames are center-cropped to 224×224 pixels.