HumanSAM: Classifying Human-centric Forgery Videos in Human Spatial, Appearance, and Motion Anomaly

Supplementary Material

In this supplementary material, we offer further details on HumanSAM. Appendix A delves into the calculation of the three anomaly scoring mechanisms. Appendix B details the specifics of the experimental setup. Appendix C conducts a deeper analysis of the quantitative experimental results. Appendix D provides additional quantitative analyses to support the effectiveness of HumanSAM. Appendix E presents analyses of spatio-temporal information. Appendix F is dedicated to exploratory experimental analyses, focusing on the outcomes of training with various forgery data sources. Appendix G is limitations.

Additionally, we have included videos featuring human-centric anomalies as part of this supplementary material to demonstrate examples of each anomaly type.

A. Anomaly Scoring Mechanism

A.1. Spatial anomaly

We used Depth pro[2] to generate depth maps for each frame of the video. Subsequently, we employed a technique based on optical flow distortion error[6, 7, 11] to quantitatively evaluate these depth maps. This technique measures the consistency of motion by monitoring the trajectory of pixel movement. In the depth maps, the pixel values represent the spatial depth of the scene.By calculating the distortion error, we are able to assess the coherence between depth maps, thereby quantifying anomalies in spatial depth. The warping error is computed as follows:

Optical Flow Estimation: For two consecutive frames I_t and I_{t+1}, the optical flow $F_{t\to t+1}$ from frame t to frame $t+1$ is obtained using a optical flow estimation network [15].

Image Warping: Using the optical flow $F_{t\to t+1}$, frame I_t is warped to the coordinates of frame $t+1$, resulting in the warped image $\hat{I}_{t+1}$:

$$\hat{I}_{t+1} = W(I_t, F_{t\to t+1}), \tag{1}$$

where $W(\cdot,\cdot)$ represents the warping operation based on the optical flow.

Pixel-wise Difference Calculation: The pixel-wise difference between the warped image $\hat{I}_{t+1}$ and the predicted image I_{t+1} is computed using the L_2 norm:

$$E_t = \|\hat{I}_{t+1} - I_{t+1}\|_2^2. \tag{2}$$

Final Score: The warping error E_{warp} is calculated as the average of the pixel-wise differences over all consecutive frame pairs:

$$E_{\text{warp}} = \frac{1}{T-1}\sum_{t=1}^{T-1} E_t, \tag{3}$$

where T denotes the total number of frames.For specific examples, please refer to Fig. 3.

A.2. Appearance anomaly

We referred to the methods used in VBench[4] for calculating subject consistency and background consistency. Both calculations employ the same formula, which involves calculating the sum of the cosine similarities of image features between consecutive frames, as well as the sum of the cosine similarities between the first frame's image features and each subsequent frame, and then averaging these total similarity scores to determine the average consistency across the frames. For subject consistency, they utilized DINO[3], while for background consistency, they employed CLIP[6]. However, we found that this approach becomes ineffective when dealing with scene transitions.

To address this limitation, we abandoned the similarity calculation between the first frame and the subsequent frames and instead adopted a sliding window consistency approach. This method calculates the average similarity within a specified window, such as over a span of 5 frames. The specific calculation formula is as follows:

$$S_{\text{score}} = \alpha \cdot \frac{1}{T-1}\sum_{t=2}^{T}(f_{t-1}\cdot f_t) + \beta \cdot \frac{1}{N}\sum_{k=1}^{N} S_{\text{window},k} \tag{4}$$

where f_i represents the i^{th} frame, the $\langle\cdot\rangle$ operation denotes the calculation of the cosine similarity of image features, and α and β are the weights for the two terms, both of which are set to 0.5. The calculation formula for $S_{\text{window},k}$ is as follows:

$$S_{\text{window},k} = \frac{1}{W-1}\sum_{j=2}^{W}(f_{j-1}\cdot f_j) \tag{5}$$

where W denotes the window size, it is specified as 5.We finally calculate the score of a video using CLIP and DINOv2[10] respectively, and then take the average as the appearance anomaly score for that video.For specific examples, please refer to Fig. 4.

Table 1. Video forgery detection performance on the HFV dataset measured by **mapped binary classification** ACC (%) and AUC (%). [ACC/AUC in the Table; Key: **Best**; Avg.: Average].

Method	MiniMax	Gen-3	Vchitect-2.0 (VEnhancer)	Kling	CogVideoX-5B	Vchitect-2.0-2B	pika	Gen-2	Avg.
CNNDet[16]	93.7/93.0	93.8/93.2	94.1/92.8	77.6/79.1	93.0/92.4	89.6/89.5	92.3/91.8	93.8/93.1	91.0/90.6**(+4.3/+2.3)**
DIRE[18]	92.9/92.3	94.0/93.3	93.4/92.8	83.1/83.9	93.9/93.2	90.6/90.3	93.4/92.8	93.1/92.5	91.8/91.4**(+4.1/+2.4)**
F3Net[12]	92.0/90.7	92.0/90.7	88.0/84.0	83.5/85.0	91.0/89.3	89.5/85.7	91.0/88.7	89.5/89.7	89.6/88.0**(+5.2/+1.7)**
Uni-FD[9]	95.7/99.5	97.1/99.4	93.5/99.0	80.2/90.7	91.9/98.9	91.4/98.5	93.4/99.4	94.9/99.6	92.3/98.0**(+11.4/+2.7)**
TimeSformer[1]	95.6/99.6	95.6/99.4	96.0/99.7	87.6/95.8	96.0/99.7	95.5/99.4	96.1/99.8	96.0/99.7	94.8/99.1(-0.4/**+1.2**)
MM-Det[13]	98.1/99.8	**99.1/100**	**98.7**/99.9	73.4/94.6	99.0/99.9	98.7/99.9	99.0/99.9	99.0/100	95.6/99.3(-1.4/-0.4)
Ours	**99.1/100**	97.9/99.2	98.0/**100**	**90.0/99.6**	**99.3/100**	**99.3/100**	**99.7/100**	**99.4/100**	**97.8/99.9**(0/0)

A.3. Motion anomaly

The calculation method for motion anomaly is to directly compute the distortion error of the video frame images, which is the same method as the distortion error calculation for spatial anomalies.For specific examples, please refer to Fig. 5.

B. Implementation Details

B.1. Dataset

The dataset is organized according to the nine types of forged data sources shown in main text. The order from top to bottom also corresponds to the ranking by the VBench team[4]. This means that MinMax ranks first in the HFV, followed by Gen3, then Vchitect-2.0 (VEnhancer), with Gen-2 being the last.

B.2. Hyperparameters of HumanSAM

We train all parameters of the video understanding branch while freezing the parameters of the spatial depth branch. The video understanding branch selected is the distilled L version of the InternVideo2[17] single modality, with a patch size of 14×14. We choose the image encoder and patch encoder of Depth pro[2] as the spatial depth branch. The final output of the video understanding branch is $f_x \in \mathbb{R}^{2816}$. The vector from the spatial depth branch, after pooling and other operations, becomes $f_y \in \mathbb{R}^{1024}$. Therefore, f_x is passed through a linear layer to reduce its dimensionality to 1024. f_x and f_y are then combined through a trainable parameter α to form the final HFR.

B.3. Training and Inference

For the experimental resources used in training and inference, all experiments were conducted using a single NVIDIA RTX 3090 GPU with a maximum of 256G of memory.

During training, for each video, we performed segmented sampling, collecting a total of eight frames, which were then cropped to 224x224 as input.We used the AdamW optimizer with a learning rate of 2e-5 and ran for 100 epochs, selecting the best performance on the validation data from the training set.

For inference, we evaluated all models at the video level. For frame-level baselines, the final result was the average of all frame results. For video-level baselines, the results were obtained following their respective default frame sampling and evaluation settings.

C. Mapped Binary Classification Experiment

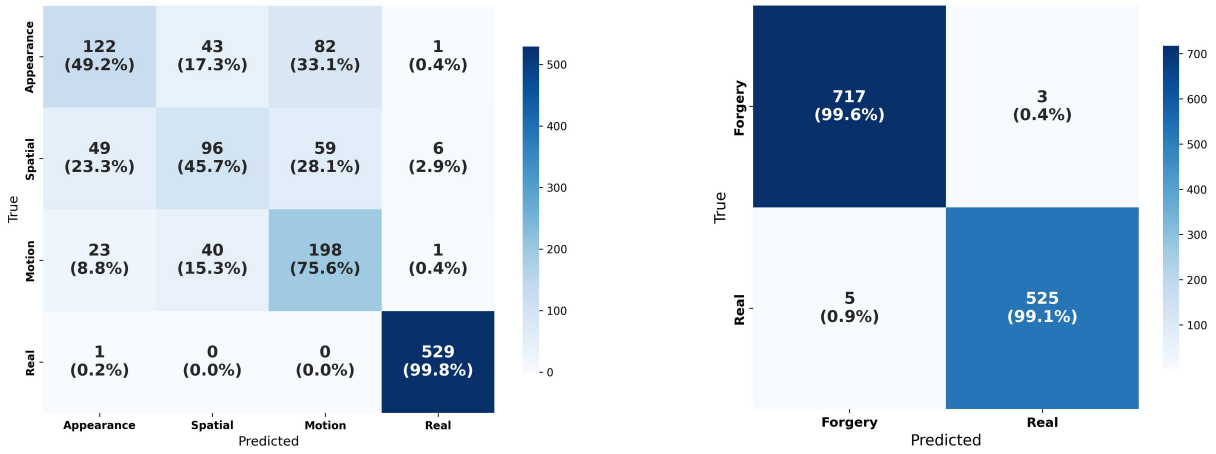
Due to space constraints in the main text, we supplement here the general binary classification experiments for TimeSformer and HFR. A comparison of Tab. 4 of main text and Tab. 1 reveals that methods with lower binary classification accuracy, such as CNNDet[16], DIRE[18], F3Net[14] and Uni-FD[9], can significantly improve their binary classification performance when trained using our proposed multi-class task. However, for models like TimeSformer[1], MM-Det[13] and ours, which already achieve high accuracy in binary classification, training with the new task has little impact on their binary classification performance.

Notably, HFR achieves an average ACC of 97.8% and an average AUC of 99.9%, further demonstrating that our method more effectively models video features, enabling it to distinguish between real and synthetic videos with greater precision.

D. Additional Quantitative Analyses

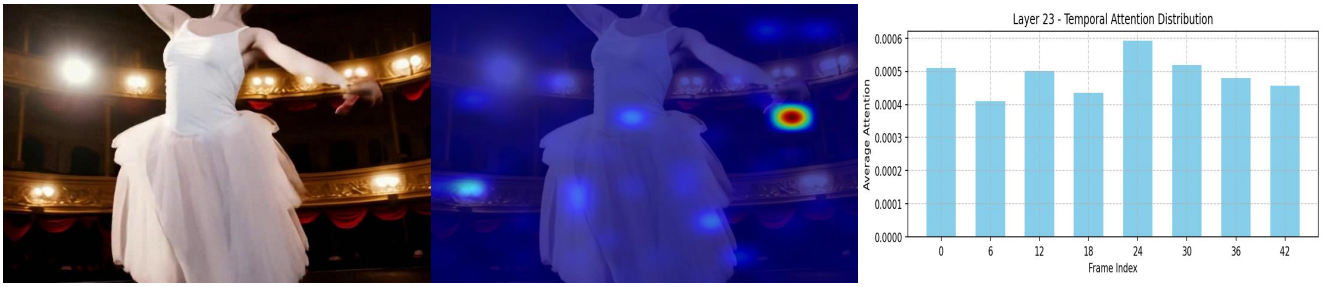
D.1. Confusion Matrices and Per-Class Performance

Fig. 1 shows the confusion matrices for both fine-grained (four-class) and binary classification. Our method achieves F1-scores of 0.5508 (appearance anomaly), 0.4936 (spatial anomaly), 0.6589 (motion anomaly), and 0.9916 (real). Most confusions occur between motion and spatial classes, partially due to the motion-sensitive video branch and the



(a) (b)

Figure 1. Confusion matrices on CogVideoX-5B dataset: (a) Multi-class, (b) Binary.



(a) (b) (c)

Figure 2. Spatial and temporal activation visualization on a CogVideoX-5B video. From the last Transformer layer of the video branch: (a) original frame at index 24; (b) spatial activation map of that frame; (c) temporal activation bar plot averaged over 8 sampled frames.

limitations of the frozen depth encoder. Nevertheless, the binary classification performance remains strong and stable.

D.2. Generalization on External Datasets

To evaluate generalizability beyond the HFV dataset, we test on the Sora dataset, which includes a broader distribution of scenarios. HumanSAM achieves 95.3% accuracy and 99.5% AUC, outperforming MM-Det[13] (81.0% / 98.4%). This demonstrates the framework's strong transferability even in the presence of diverse, non-human-specific generative content. Future work will explore methods with improved generalization across open-world forgery distributions.

E. Spatio-Temporal Information Analyses

Fig. 2 illustrates how our model leverages spatio-temporal cues to localize anomalies. Specifically, spatial attention highlights the girl's **distorted** right hand in frame 24, while the highest temporal score is also assigned to this frame. This demonstrates the model's potential for precise frame-level anomaly localization. Additionally, lighting inconsistencies—though not strictly human-centric—are considered part of the appearance anomaly when they impact human actions (e.g., uneven illumination on the left side in Fig. 2).

F. Exploratory Experimental Analysis

Due to the length constraints of the main text, some details of exploratory experiments are presented here. As shown in Tab. 1 of main text, the higher a synthetic data source ranks, the better its overall performance on the VBench benchmark[4]. MinMax[8] ranks first, while Kling[5] ranks fourth. As shown in Tab. 1, all methods experience a sudden performance drop on Kling. To further investigate this, while keeping the original experimental settings unchanged, we replaced the CogVideoX-2B forgery data in the training set with Kling and MinMax for four-class training. This adjustment was made to compare the results across the remaining seven forgery video sources.

As shown in Tab. 2, after replacing the training data with MinMax and Kling, the overall results still demonstrate that

Table 2. Comparison of **multi-class training** using **different** forgery video sources in the **HFV** dataset measured by ACC (%) and AUC (%).[ACC/AUC in the Table; Key: **Best**; Avg.: Average].

Video Source	MiniMax	Gen-3	Vchitect-2.0 (VEnhancer)	Kling	CogVideoX-5B	Vchitect-2.0-2B	pika	Gen-2	Avg.
CogVideoX-2B	**70.4/88.2**	**73.8/88.6**	**72.3/89.5**	65.8/87.5	**75.6/92.1**	66.5/**86.2**	**69.6/88.0**	**64.2/83.4**	**69.8/87.9**
MinMax	-	63.7/82.2	68.8/87.1	**84.5/96.4**	65.4/86.5	60.2/83.0	67.6 /85.1	55.6/80.6	66.5/85.6
Kling	69.5/87.2	63.7/82.2	68.8/87.1	-	64.2/83.5	60.2/83.0	67.0/85.1	55.6/80.6	64.1/84.1

the seventh-ranked CogVideoX-2B achieves the best performance. Upon closer examination of the CogVideoX-2B videos, we observed that they still exhibit noticeable gaps compared to realistic human behavior. This leads us to hypothesize that lower-ranked synthetic data sources may contain more of the three types of anomalies, which in turn benefits the proposed HFR in learning anomalous features. Kling and MinMax share identical metrics across most forgery data sources, indicating that their anomalous features are quite similar. When HFR trained with MinMax is used to predict Kling, the accuracy improves by 18.7%, and the AUC improves by 8.9%, compared to CogVideoX-2B. However, when HFR trained with Kling is used to predict MinMax, the performance metrics show a slight decline relative to CogVideoX-2B.

Based on an analysis of the synthetic video sources from Kling and MinMax, we speculate that while both exhibit visual consistency and logical patterns close to real-world videos, Kling has a grainy texture, making its video quality noticeably inferior to that of MinMax. This explains the significant improvement in metrics when MinMax is used to predict Kling, while using Kling to predict MinMax shows minimal change. Future work could explore experiments involving mixed synthetic data sources for detection.

G. Limitations

The rapid evolution of video generation models poses a fundamental challenge to the sustainability of existing detection frameworks. As generative techniques advance continuously, current artifact detection mechanisms may become outdated quickly, necessitating perpetual updates and adaptive strategies to maintain robustness against novel forgery patterns.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 2

[2] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 1, 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[4] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 1, 2, 3

[5] klingai. https://klingai.kuaishou.com/, 2024. 3

[6] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 1

[7] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33:1083–1093, 2020. 1

[8] MiniMax. https://hailuoai.com/video, 2024. 3

[9] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 2

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[11] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 1

[12] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020. 2

[13] Xiufeng Song, Xiao Guo, Xiaohong Liu, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, and Guangtao Zhai. On learning multi-modal forgery representation for diffusion generated video detection. In *Proceeding of Thirty-eighth Confer-*

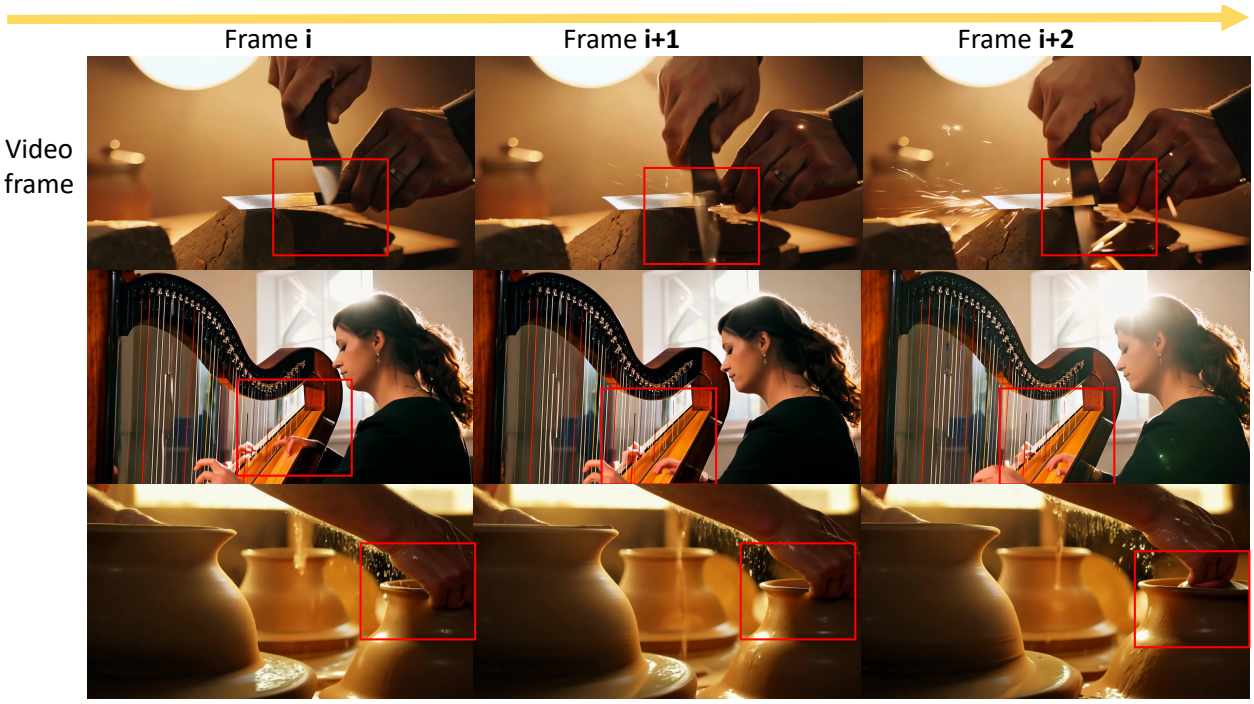


Figure 3. It can be seen that in the first row, two metal knives blur and pass through each other; in the second row, a woman's hand blurs as it reaches into the harp; in the third row, a person's hand passes through a clay pot under production without leaving any traces. Overall, this violates spatial logic and the normal rules of object interaction.

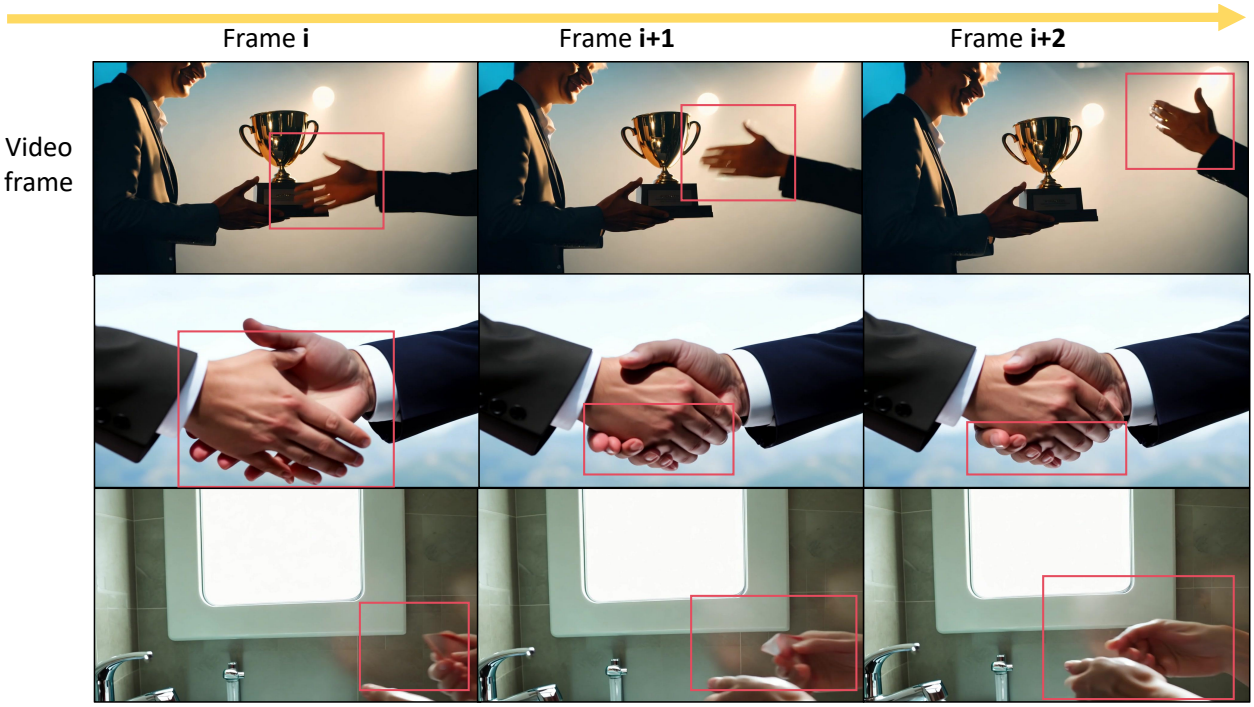


Figure 4. Some examples of appearance anomalies.It can be seen that in the first row, the hand on the right suddenly changes from an apparently left hand to a right hand. In the second row, the number of fingers on the right hand changes from six to five. In the third row, the object held in the hand gradually disappears. Generally speaking, the consistency in appearance cannot be maintained.

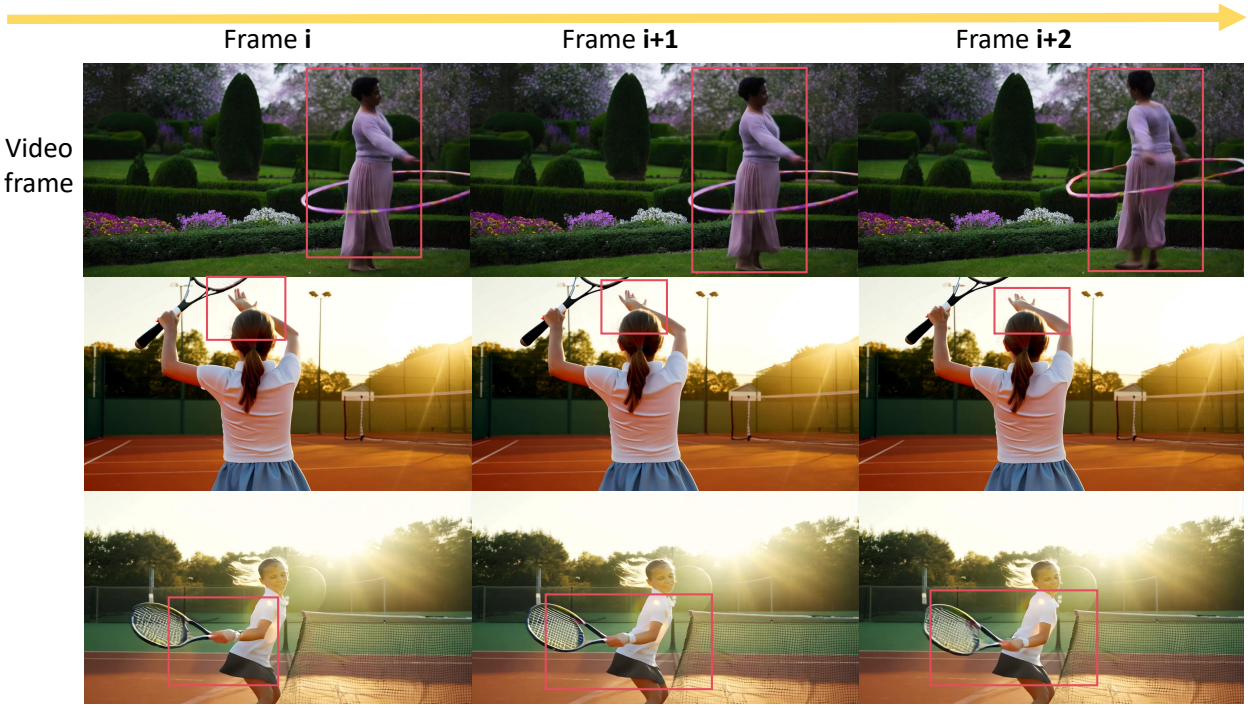


Figure 5. Some examples of motion anomalies.It can be observed that in the first row, the woman's body maintains a forward-leaning tendency, but her head suddenly rotates 180 degrees. In the second row, the girl's right hand takes on the shape of a left hand. In the third row, the girl's right hand assumes the posture of a left hand, which would be appropriate if the girl's body were rotated around. Generally speaking, the motion of the characters does not conform to normal biological motion patterns.

ence on Neural Information Processing Systems, Vancouver, Canada, 2024. 2, 3

[14] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2

[15] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 1

[16] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 2

[17] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 2

[18] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 2