Supplemental Meterials for ActivityForensics: A Comprehensive Benchmark for Localizing Manipulated Activity in Videos

Peijun Bao[1,2], Anwei Luo[3,4], Gang Pan[1], Alex C. Kot[5,6], Xudong Jiang[2]
[1]College of Computer Science and Technology, Zhejiang University
[2]School of Electrical and Electronic Engineering, Nanyang Technological University
[3]School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics
[4]Jiangxi Provincial Key Laboratory of Multimedia Intelligent Processing
[5]Shenzhen MSU-BIT University [6]VinUniversity
peijun001@e.ntu.edu.sg luoanwei@jxufe.edu.cn

1. Additional Dataset Details

1.1. Additional Dataset Statistics

As illustrated in Fig. 1, the training split is constructed to be balanced across the five available manipulation methods, ensuring that the model observes a diverse yet uniformly distributed set of manipulations during learning. In contrast, the testing split includes an additional commercial system that is never seen during training, while still maintaining a balanced distribution across all six mechanisms. Incorporating such commercial, black-box generators enables us to evaluate under a realistic open-world scenario.

1.2. Human Evaluation of Dataset Quality

To assess the visual perceptibility and overall quality of the forged activities in the ActivityForensics dataset, we conducted a human evaluation study. Participants were asked to watch a series of video clips and determine whether they contained AI-generated forged activities. Each video was rated using four options: "Definitely Real", "Likely Real", "Likely Forgery", and "Definitely Forgery". The results are shown in Fig. 2. For videos containing forgery activities (Fig. 2a), more than 60% of the samples were labeled as "real" or "likely real", indicating that most AI-generated activities exhibit high visual realism and strong perceptual deception. Ordinary observers can hardly notice any visual artifacts, and such activity-level forgeries could easily mislead the public in real-world scenarios such as social media or news broadcasting, posing potential risks to media credibility and social trust. In contrast, real videos (Fig. 2b) were mostly recognized as "real", yet about 8% were misclassified as "forgery", suggesting that complex natural dynamics or uncontrolled recording conditions can also cause confusion. These results show that the forged samples in the ActivityForensics dataset are not only highly realistic in appearance but also cover manipulation types with real-world harmful implications, thereby providing a more challenging and practically meaningful benchmark for evaluating model performance and generalization.

1.3. Video Manipulation Methods

We construct forged video segments using several state-of-the-art models for conditioned video generation and masked video editing. These approaches represent the latest progress in controllable and high-fidelity video synthesis. We consider the following state-of-the-art masked video editing methods, which perform localized edits conditioned on a textual prompt, start and end frames, and a spatial mask. Only the masked region is modified, while the unmasked areas remain unchanged, ensuring consistent appearance and motion outside the edited region.

Wan-2.1 [12] is a large-scale diffusion-transformer video foundation model trained on extensive video–text pairs. It supports both text- and image-conditioned synthesis and achieves strong temporal coherence through spatio-temporal attention and motion-aware latent modeling. The model's scaling strategy and high-capacity backbone enable realistic scene dynamics and consistent appearance across frames, making it well suited for prompt-guided generation between given start and end frames.

FCVG [14] performs frame-wise conditioned generation between two keyframes and leverages explicit human-pose conditioning to guide motion synthesis. Each intermediate frame is generated under pose supervision derived from the start and end frames, enabling accurate reconstruction of human motion trajectories and temporally smooth interpolation. This pose-driven formulation ensures fine motion controllability and natural continuity, making FCVG particularly effective for human-centric forged segments in our dataset.

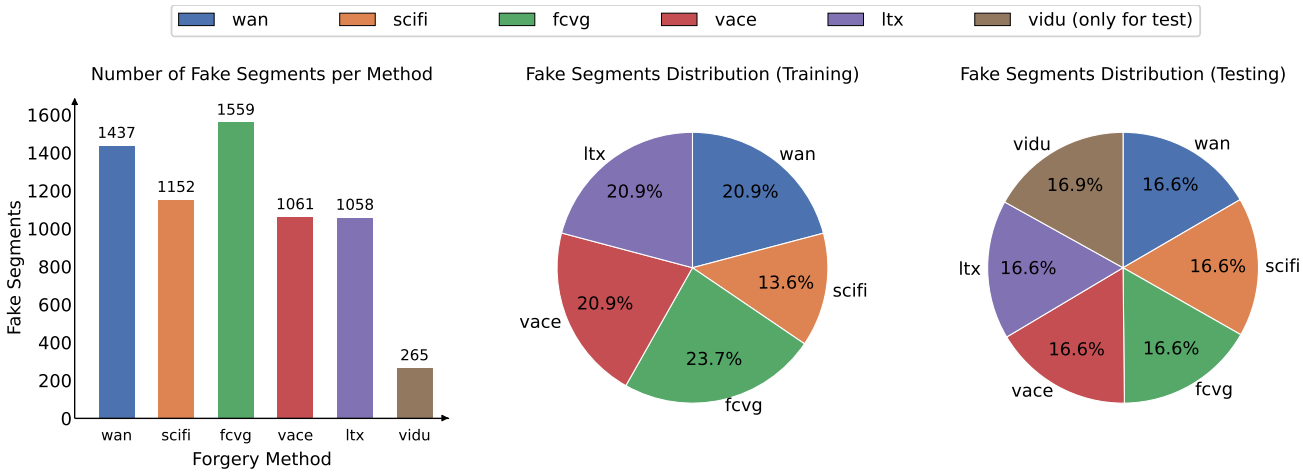

Figure 1. Data statistics of manipulation methods in ActivityForensics. The dataset covers diverse generation and editing mechanisms, highlighting its broad coverage and balanced distribution across different forgery types.

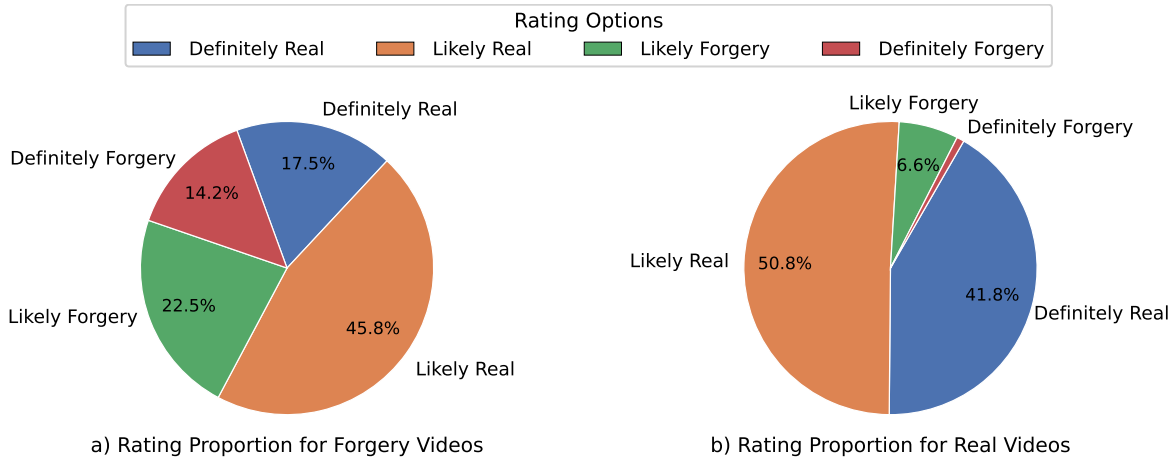

Figure 2. Human evaluation of data quality in the ActivityForensics dataset. For videos containing forgery activities (a), over 60% of the samples were labeled as "real" or "likely real", indicating that most AI-generated activities exhibit strong visual realism and perceptual deception. Such activity-level forgeries are difficult for human observers to distinguish intuitively and could potentially mislead audiences in real-world dissemination, posing risks to media authenticity and public trust.

Scifi [4] introduces symmetric conditioning on both start and end frames to enhance temporal stability in diffusion-based inbetweening. It employs balanced feature fusion to maintain semantic alignment and suppress directional bias during generation. The model produces continuous and visually coherent motion sequences, contributing realistic synthesized transitions under constrained frame guidance.

Vace is built upon the VACE framework [6], a unified architecture for conditioned video generation and masked video editing. It integrates the Wan-2.1 [12] backbone as the diffusion module to improve visual fidelity and motion consistency. The model utilizes a Video Condition Unit to encode prompt, frame, and mask conditions, and a Context Adapter for multimodal feature fusion. These components enable fine-grained region editing and smooth temporal transitions under explicit start- and end-frame guidance, producing visually realistic and temporally coherent manipulations.

LTX shares the same VACE architecture but replaces the backbone with LTX-Video [5], a lightweight yet high-performance diffusion model optimized for efficient controllable synthesis. LTX-Video employs a high-compression latent representation and temporal-aware self-attention to capture long-range dependencies at reduced computational cost. It further adopts multi-scale rendering for coarse-to-fine refinement, achieving near-real-time video generation while preserving spatial detail and temporal coherence. This variant serves as a complementary setting, providing efficient and diverse synthetic content for evaluation.

Please modify the original activity description occurring between {start} and {end} seconds so that the local activity within this interval is noticeably altered, while the visual state at both the start and end moments remains completely unchanged, ensuring that no abrupt or distracting transitions appear in the overall video. The goal is to introduce clear and perceivable variation inside the segment while preserving global continuity. Three types of modifications may be applied. 1) Activity Addition: introduce new and visually salient activities on top of the original ones, without adding any objects not explicitly mentioned, so that the viewer can clearly perceive richer local dynamics while the scene still transitions naturally. 2) Activity Deletion: remove certain segments of the original activity so the interval exhibits a clearly simplified or reduced pattern of movement, while maintaining logical continuity and preventing unnatural jumps. 3) Activity Replacement: substitute the original activity with a distinct and easily perceivable new sequence that relies only on existing subjects or objects, ensuring that the local change is evident yet still compatible with the preserved start and end states. All three modification strategies emphasize intentional and localized change while maintaining the principle of local variation without global disruption, so that the modified video segment appears both noticeably different and seamlessly integrated into the full sequence.

Figure 3. The prompt that guides LLMs to manipulate the activity descriptions.

1.4. Videos Sources

We use the videos from three widely adopted benchmarks in video grounding when constructing our ActivityForensics dataset, Charades-STA [7], ActivityNet Captions [8], and HC-STVG [15], whose raw videos are publicly available and cover a wide spectrum of visual scenes and diverse human activities. The details of the original datasets are listed as follows.

Charades-STA [7] contains 9,848 videos of daily indoor activities. The average length of a sentence query is 8.6 words, and the average duration of the video is 29.8 seconds. The dataset is originally designed for action recognition and localization [11], and later extended by Gao *et al.* [11] with language descriptions for temporal video grounding [1–3, 9]. Charades-STA contains 12,408 moment-sentence pairs in training set and 3,720 pairs in testing set.

ActivityNet Captions [8] consists of 19,290 untrimmed videos with diverse and open-domain content. The average duration of the video is 117.74 seconds and the average length of the description is 13.16 words. The training set contains 37,417 sentence queries while the testing set contains 17,031 sentence queries.

HC-STVG [15] is comprised of 5,660 untrimmed videos sourced from multi-person scenes, each annotated with one sentence query related to human attributes and actions. With 57.2% of video clips involving more than three individuals engaged in similar actions, this dataset presents crowded scenarios that make spatio-temporal grounding particularly challenging. There are 4,500 video-sentence pairs for training and 1,160 for testing in total.

2. Additional Implementation Details

We adopt a linear noise schedule by linearly interpolating the noise coefficients β_s from $\beta_{\text{start}} = 1 \times 10^{-4}$ to $\beta_{\text{end}} = 2 \times 10^{-2}$. The corresponding $\alpha_s = 1 - \beta_s$ values and their cumulative products $\bar{\alpha}_s = \prod_{i \leq s} \alpha_i$ are precomputed and used directly in the forward diffusion process. The total training epoch is set to 30. The hyperparameter η is set to 0.0. We set the number of frame to 192 for each video. The experiments are conducted on a single A100 GPU. The pre-trained CLIP ViT-L/14 [10] model is used to extract visual features for video frames. When constructing the dataset, we use the prompt shown in Fig. 3 with GPT-5 to manipulate the activity descriptions. All other implementation details follow ActionFormer [13] and can also be found in the attached code.

3. Additional Experiment Results

3.1. Failure Case Analysis

Fig. 4 presents two representative failure cases of TADiff in the temporal forgery localization task, covering both intra-domain and open-world scenarios. In the *intra-domain* case a), the model successfully identifies the major manipulated interval but misses a short forged segment at the beginning of the video. This forged portion occupies only a very small fraction of the untrimmed video and remains semantically and motion-wise consistent with the surrounding authentic frames. As a result, the subtle forgery signal is easily masked by strong contextual semantics over time. This observation reflects a key challenge in intra-domain localization: within long and semantically coherent videos, the model must discriminate extremely short and weak artifact cues from dominant real dynamics, which requires fine-grained temporal resolution and high sensitivity to subtle artifacts.

In the *open-world* case b), where the forged video is generated by an unseen commercial model, the model roughly localizes the manipulated interval but exhibits slight boundary drift, with the predicted segment extending beyond the actual forged region. This example contains noticeable illu-

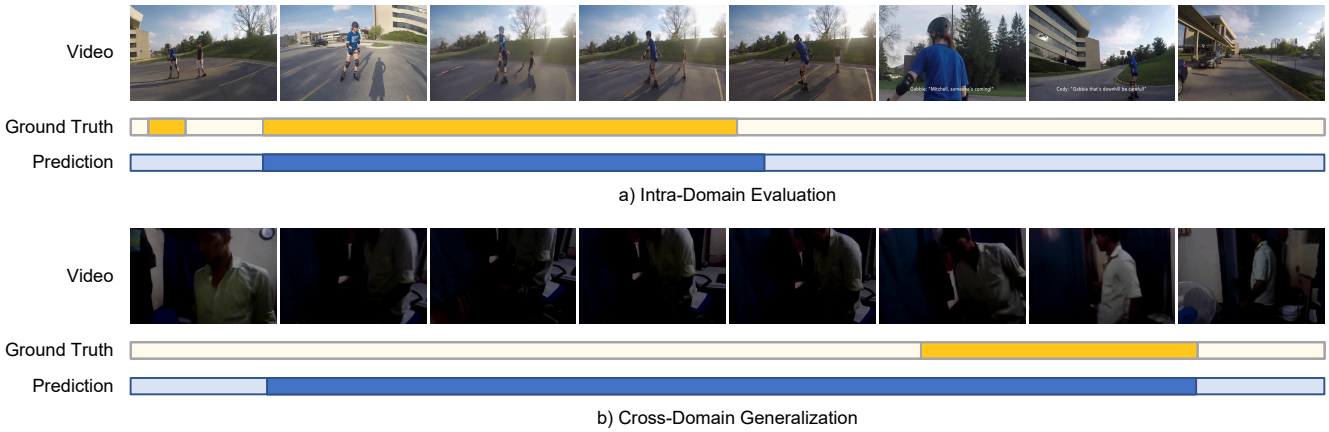

Figure 4. Failure case analysis. The darker yellow rectangle represents the ground-truth forgery segments, while the darker blue rectangle denotes the model's prediction.

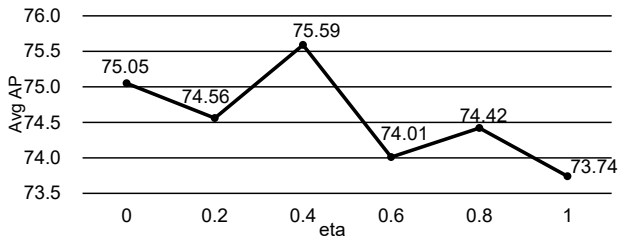

Figure 5. Influence of η.

mination variations and motion blur, causing non-stationary low-level temporal features that confuse boundary estimation. Such cases highlight a key difficulty in open-world generalization: when generation mechanisms, visual quality, or motion characteristics change significantly, the model must remain robust enough to distinguish genuine forgeries from natural video noise. Overall, TADiff achieves accurate localization in most cases, yet localizing short-duration manipulations and maintaining robustness across varying generation mechanisms remain the main challenges for temporal forgery localization. Future work may focus on improving temporal resolution and enhancing stability under varying video conditions.

3.2. Additional Ablation Studies

The Influence of η. We study the influence of the diffusion randomness coefficient η in the denoising process in Fig. 5, evaluated under the intra-domain setting. Overall, TADiff maintains stable performance across different η values. The results show that the performance peaks at $\eta = 0.4$ with an average AP of 75.59, while larger randomness ($\eta = 0.8, 1.0$) leads to a drop. Overall, TADiff maintains stable performance across different η values, and particularly strong results are achieved at $\eta = 0.4$ and $\eta = 0.0$, which obtain 75.59 and 75.05, respectively. It consistently outperforms the ActionFormer baseline (70.67), demonstrating the robustness of temporal artifact diffuser.

Table 1. The impact of β_{start} and β_{end}.

β_{start}	β_{end}	avg AP
0.5e-4	2.0e-2	74.93
1.0e-4		75.05
2.0e-4		75.52
3.0e-4		75.29
4.0e-4		75.91
5.0e-4		74.61
1.0e-4	0.5e-2	74.53
	1.0e-2	74.62
	1.5e-2	74.96
	2.0e-2	75.05
	2.5e-2	75.39
	2.5e-2	74.53

Table 2. Comparison of localization performance on AIGC manipulations and pure stitching within the modified test set.

Segment Type	AP@0.75	AP@0.85	AP@0.95
AIGC Manipulation ↑	77.63	71.98	50.85
Pure Stitching ↓	8.98	6.95	1.61

Impact of β_{start} and β_{end}. Table 1 reports the effect of different noise schedule parameters on model performance. Overall, the average AP stays within a narrow range around 74.5–75.9 across different combinations of β_{start} and β_{end}. When fixing $\beta_{\text{end}} = 2.0\text{e}{-2}$, increasing β_{start} produces small performance fluctuations, with the highest score 75.91 appearing near $4.0\text{e}{-4}$. A larger starting noise, such as $5.0\text{e}{-4}$, causes a slight drop, but the variation remains limited. When fixing $\beta_{\text{start}} = 1.0\text{e}{-4}$, changing β_{end} shows a similar trend. Mid-range values, including $2.0\text{e}{-2}$ and $2.5\text{e}{-2}$, give marginally better results, while

both smaller and larger terminal noise levels show mild decreases. In summary, different noise schedule settings lead to performance that remains closely clustered, and moderate increases in either the starting or ending noise often correspond to slightly better results.

Ablation study of boundary stitching bias. To investigate whether the model is biased toward detecting temporal boundary transitions rather than intrinsic AI-generated content (AIGC) artifacts, we construct a modified test set from ActivityForensics by randomly replacing a subset of AI-generated segments with unrelated real video segments. This results in a setting where AIGC and pure stitching segments are clearly distinguished, while AIGC-specific manipulation cues are removed from the stitched segments. As shown in Table 2, the proposed TADiff model achieves strong localization performance on AIGC manipulations, while exhibiting significantly degraded performance under pure stitching. The performance gap indicates that the model does not rely on boundary transitions as a shortcut.

References

[1] Triantafyllos Afouras, Effrosyni Mavroudi, Tushar Nagarajan, Huiyu Wang, and Lorenzo Torresani. Ht-step: Aligning instructional articles with how-to videos. In *NeurIPS*, 2023. 3

[2] Peijun Bao, Qian Zheng, and Yadong Mu. Dense events grounding in video. In *AAAI*, 2021.

[3] Peijun Bao, Zihao Shao, Wenhan Yang, Boon Poh Ng, Meng Hwa Er, and Alex C Kot. Omnipotent distillation with llms for weakly-supervised natural language video localization: When divergence meets consistency. In *AAAI*, 2024. 3

[4] Liuhan Chen, Xiaodong Cun, Xiaoyu Li, Xianyi He, Shenghai Yuan, Jie Chen, Ying Shan, and Li Yuan. Sci-fi: Symmetric constraint for frame inbetweening. *arXiv preprint arXiv:2505.21205*, 2025. 2

[5] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, David-Pur Moshe, Eitan Richardson, E. I. Levin, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2025. 2

[6] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 2

[7] Zhenheng Yang Jiyang Gao, Chen Sun and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 3

[8] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 3

[9] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 3

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, and et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[11] Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, Ivan Laptev, and Abhinav Kumar Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 3

[12] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2

[13] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, pages 492–510, 2022. 3

[14] Tianyi Zhu, Dongwei Ren, Qilong Wang, Xiaohe Wu, and Wangmeng Zuo. Generative inbetweening through frame-wise conditions-driven video generation. In *CVPR*, pages 27968–27978, 2025. 1

[15] Si Liu Guanbin Li Xiaojie Jin Hongxu Jiang Qian Yu Zongheng Tang, Yue Liao and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. In *TCSVT*, 2021. 3