ACHILLES' HEEL: UNVEILING TEMPORAL ARTIFACTS IN AI-GENERATED VIDEOS VIA VIDEO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

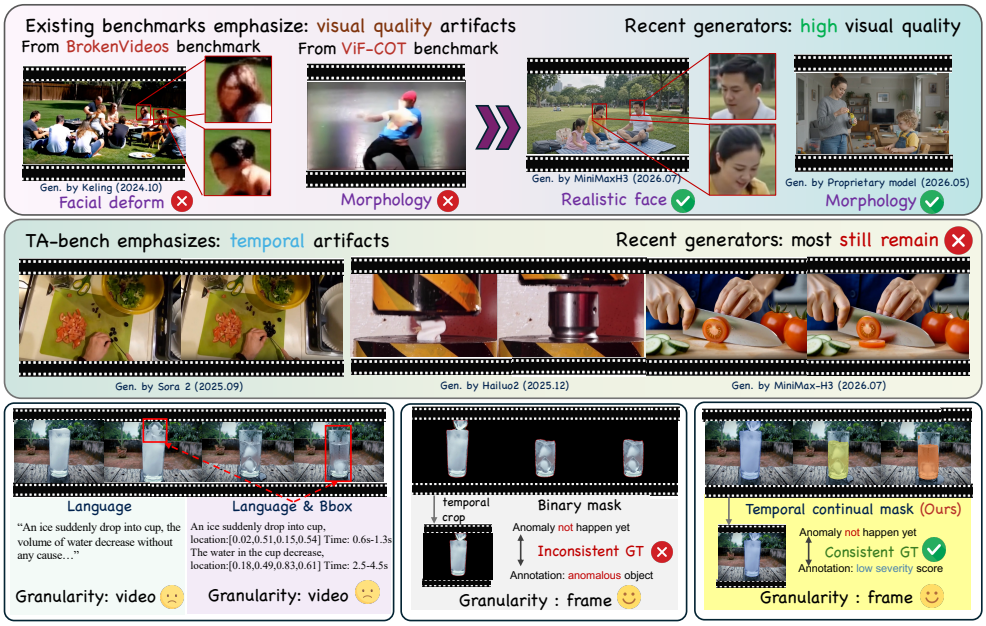

Figure 1: Existing benchmarks for diagnosing artifacts have limitations in both **Artifact perspective** and **Annotation perspective:** This motivate us to propose TA-Bench.

ABSTRACT

Despite rapid advances in visual quality, current video generation models still produce temporal artifacts, including appearance drift, unexplained object appearance or disappearance, and scene flicker. These artifacts may be difficult to detect in individual frames but become apparent over time. Existing benchmarks pay limited attention to such artifacts and provide only coarse-grained annotations. We introduce *temporally continuous masks* for fine-grained temporal artifact diagnosis and present **Temporal Artifact Bench (TA-Bench)**, comprising 6,945 videos from nine generative models, with 8,618 temporally continuous masks and 214.3K frame-level annotations. To construct TA-Bench at scale, our pipeline combines automatic triage, mask generation, controllable synthesis, and three-stage human verification. We conduct a comparative evaluation of general-purpose MLLMs, AIGC-detection-expert MLLMs, video segmentation models, and generative video models for temporal artifact localization. Remarkably, the comparison reveals that generative video models are most sensitive to temporal artifacts, while sensitivity remains limited in most evaluated models. Motivated by this finding, we propose **Video Diffusion Diagnoser (VDD)**, a *generation-for-detection* approach that repurposes a pretrained video diffusion model to predict temporally continuous masks directly from video latents in a single feed-forward pass. Experiments on TA-Bench show that VDD outperforms MLLM-based and segmentation-based methods, suggesting that generative video priors provide a strong foundation for fine-grained temporal artifact perception.

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.

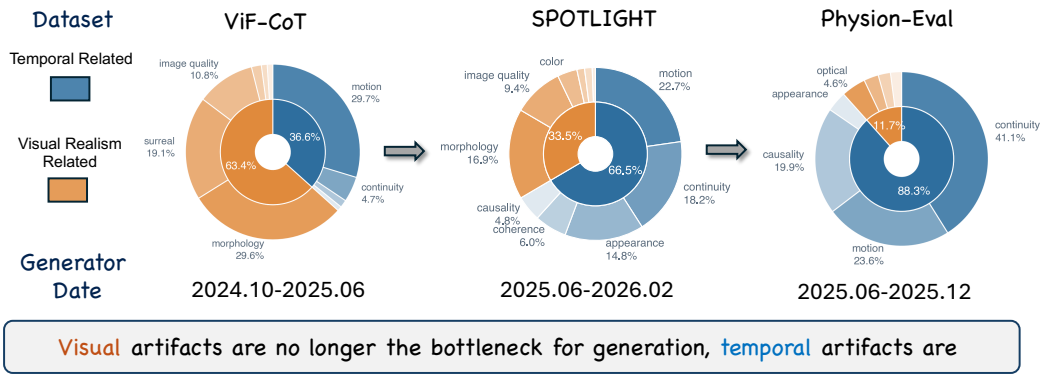

Figure 2: Evolution of artifacts in AI-generated video. The analysis is derived from human-feedback datasets used for AIGC Detection or generation evaluation. For each video, we use an LLM to categorize its associated artifact feedback in an fixed vocabulary.

1 Introduction

Recent advances in video generation have enabled the synthesis of visually realistic videos, supporting a wide range of applications in design, education, and entertainment. Yet high frame-level realism does not guarantee temporal coherence. As a video unfolds, object appearances may drift, entities may emerge or disappear without cause, and the entire scene may flicker. We collectively refer to these temporally evolving failures as **temporal artifacts**. As illustrated in Fig. 2, human feedback on recent text-to-video systems suggests that temporal artifacts remain a prevalent failure mode, even in state-of-the-art generators. Because these failures persist across model generations, they provide potentially generalizable evidence for identifying AI-generated videos. More importantly, explicitly localizing and characterizing temporal artifacts can turn an opaque binary prediction Chen et al. (2026); Ma et al. (2026) into an interpretable diagnosis.

These observations motivate a dedicated benchmark for temporal artifacts. Recent benchmarks annotate artifacts in AI-generated videos using textual descriptions Tang et al. (2026), text-grounded bounding boxes Wang et al. (2026a), or pixel-level masks Lin et al. (2025). However, as shown in Fig. 1, they still exhibit two important limitations. ***(1) Existing annotations do not jointly capture all diagnostic dimensions.*** Language, bounding boxes, and binary masks may describe an artifact, but none jointly specifies *where* it occurs, *when* it is visible, and *how severe* it is. Consequently, these annotations cannot directly supervise dense estimation and provide only limited support for fine-grained diagnosis. ***(2) Existing benchmarks underrepresent the failure modes of recent generators.*** Many benchmarks emphasize frame-level artifacts related to visual quality and realism. Although recent generators have substantially reduced such artifacts, temporal inconsistencies remain prevalent and are not adequately represented in current evaluation data.

To address these limitations, we introduce **Temporal Artifact Bench (TA-Bench)**, a benchmark dedicated to fine-grained temporal artifact diagnosis in AI-generated videos. **To provide multidimensional supervision**, TA-Bench introduces novel annotation representation, termed *temporal continuous mask*, $M \in [0,1]^{T\times H\times W}$. The mask assigns a severity value to every spatial location and time step of artifacts beyond classical binary mask, thereby jointly modelling *where* and *when* an artifact occurs and how severe it is. **To represent the failure modes of recent generators**, we curate videos from both open-source and proprietary video generation model and focus on challenging temporal artifacts. **To support benchmark construction**, we develop a data engine that combines controllable artifact synthesis, semi-automated mask generation, and 3-layer human verification. Starting from 17,569 candidate videos, automatic triage retains 8,357 videos likely to contain temporal artifacts. Ten experts then spent 210 hours verifying the annotations, yielding a 6,745 video main pool. Together with a separately curated 200-video OOD set from recent generators, TA-Bench contains 6,945 videos. We conclude several dimensions of comparison between TA-Bench and other benchmarks in Tab. 1.

Table 1: Comparison of benchmarks for diagnosing artifacts in AI-generated videos. **CC** denotes crop consistency.

Benchmark	Generator		# Videos	Annotation dimensions				Properties
	Num.	Latest		Form	Spatial	Temporal	Severity	CC
VideoPhy-2 Bansal et al. (2026)	9	Wan2.2-2025.03	3,940	Score	×	×	×	×
AEGIS Li et al. (2025a)	4	Sora-2024.10	5,199	Language	×	×	×	×
BrokenVideos Lin et al. (2025)	12	Kling1.5-2024.09	3,141	Binary mask	✓	✓	×	×
Geneva Kang et al. (2026)	3	Sora-2024.10	16,356	Lang. & box	✓	✓	×	×
GenVID Zhu et al. (2026)	3	Wan2.1-2025.02	80,000	Language	✓	×	×	×
Spotlight Chinchure et al. (2025)	3	Veo 3-2025.05	600	Language	×	✓	✓	×
CaC Wang et al. (2026a)	3	Kling2.5-2025.12	15,374	Lang. & box	✓	✓	×	×
Artifact-Bench Tang et al. (2026)	7	Kling2.5-2025.12	500	Language	×	×	×	×
Ours	9	MinimaxH3-2026.07	6,945	Continuous field	✓	✓	✓	✓

Among existing benchmarks, BrokenVideos Lin et al. (2025) is the most closely related to ours because it provides pixel-level artifact masks. However, its binary object-level annotations do not represent how artifact severity evolves over time. **This distinction is vital in temporal artifacts** as it may conflate the concept of "anomaly" and "anomalous objects". When an inconsistency is visible only during a short interval, labeling the affected object over a longer temporal track can blur the boundary between the actual artifact and other normal frames. In contrast, our continuous temporal mask assigns a value to every location and time step, with zero or low value indicating the visible artifact is not yet happen. The resulting annotation is *crop-consistent*: the annotation can be cropped in temporal alignment with the video while preserving label semantics, provided that sufficient diagnostic context remains. Furthermore, this property naturally enables label-preserving random temporal cropping as data augmentation in training models. Our experiments show that random temporal-crop augmentation lowers MSE and raises CC on both tested backbones.

Having established this benchmark, we next ask whether existing models can perform fine-grained temporal artifact diagnosis. We probe three representative model families: specialized AI-Generated video detector MLLM Tan et al. (2026), video segmentation models Hong & Do (2026), and general-purpose proprietary MLLM. We evaluate them using a unified protocol that checks each output in both textual and visual forms. The results reveal that none of these model families shows reliable sensitivity to temporal artifacts. They either lack pixel-level perception or understanding of the concept of artifacts. This motivates us to seek a model with both capabilities. Inspired by evidence that generative models can serve as general-purpose visual learners, we introduce the **Video Diffusion Diagnoser (VDD)**, which repurposes a pretrained video generation model for dense temporal artifact diagnosis. Video generation models are trained to jointly model appearance and motion across an entire video, providing a natural basis for fine-grained spatiotemporal prediction. We instantiate VDD at multiple model scales and show that it consistently outperforms strong MLLM-based and video segmentation baselines, including on videos from unseen generators. This finding reveals an "Achilles' heel" of video generation: the same generative priors used to synthesize videos can be repurposed to expose their temporal artifacts.

To summarize, our contributions are as follows:

- We introduce the task of *temporal artifact diagnosis* and formulate it as predicting a continuous temporal masks $M \in [0,1]^{T\times H\times W}$ that jointly captures *when* and *where* an artifact occurs, and *how severe* it is.
- We introduce **Temporal Artifact Bench (TA-Bench)**, a benchmark with 6,945 videos containing temporal artifacts and paired with temporal continuous masks. To construct it, we design a scalable data engine combining controllable synthesis, semi-automated annotation, and three layer human verification.
- We propose the **Video Diffusion Diagnoser (VDD)**, which repurposes a pretrained video diffusion model for temporal artifact diagnosis by exploiting the rich spatiotemporal representations learned during generative pretraining. VDD can be adapted to multiple generative backbone and the strongest variant achieves strong performance in ***TA-Bench***, demonstrating the transferability of generative priors to fine-grained temporal diagnosis.

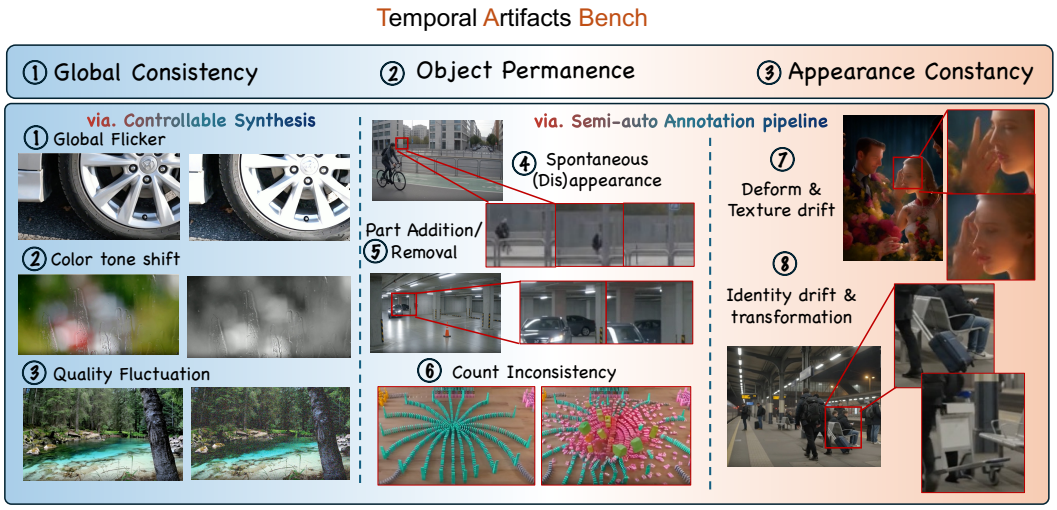

Figure 3: Hierarchical taxonomy of temporal artifacts and representative examples from TA-Bench. We partition temporal artifacts into global consistency, object permanence, and appearance constancy artifacts. The examples illustrate the characteristic temporal failures covered by each category.

2 Related Works

AI-Generated Video Detection Early studies formulate AI-generated video detection as binary classification and exploit low-level forensic cues Chen et al. (2026); Ma et al. (2025). Subsequent methods derive discriminative evidence from generation statistics, physical-rule violations, or feature-space trajectories Zheng et al. (2025); Internò et al. (2026); Zhang et al. (2026b). To improve generalization, recent approaches instead identify semantically meaningful visible artifacts: Skyra Li et al. (2025b), BusterX++ Wen et al. (2025), and DeeptraceReward Fu et al. (2025) equip MLLM-based detectors with structured chain-of-thought reasoning or reinforcement learning, while VideoVeritas Tan et al. (2026) introduces perceptual pretext tasks to strengthen artifact recognition. Nevertheless, these methods predominantly rely on static defects, such as low realism and malformed anatomy, and remain insensitive to artifacts that emerge only through temporal evolution.

Benchmarks and Evaluation of Video Generation Video-generation evaluation spans perceptual quality, text alignment, motion and temporal coherence, and physical or commonsense plausibility Huang et al. (2025); Sun et al. (2024); Duan et al. (2025); Liu et al. (2023). Comprehensive benchmarks decompose these aspects into automatic metrics or human-evaluation protocols, while specialized benchmarks target physical violations, generation artifacts, and localized errors Bansal et al. (2026); Motamed et al. (2026). Recent video reward models further aggregate such judgments into human-aligned scores, learning from quality ratings or pairwise preferences through multidimensional regression, ranking, or preference learning He et al. (2024); Liu et al. (2025); Wang et al. (2025). Nevertheless, existing evaluation data predominantly provide video-level preferences, text-alignment judgments, or sparse categorical labels, and often reduce temporal quality to motion smoothness or coarse consistency, leaving fine-grained temporal diagnosis insufficiently supported.

Diffusion Models as Vision Learners Trained on Internet-scale data, modern diffusion models acquire generalizable priors over visual geometry that can be repurposed for downstream perception. Marigold Ke et al. (2024) and DMP Lee et al. (2024) cast monocular depth estimation as conditional image generation, achieving strong performance with substantially less data compard with expert model trained on million-scale. Subsequent work extends diffusion priors to normal estimation, video segmentation, video matting, and other dense prediction tasks Xu et al. (2024); Bagchi et al. (2025); Lim et al. (2026). More recently, some works explore repurposing video diffusion models due to their strong spatiotemporal perception and emerging reasoning capabilities Wiedemer et al. (2025); Wang et al. (2026b). However, whether these generative priors can diagnose temporal artifacts in their own outputs remains largely unexplored.

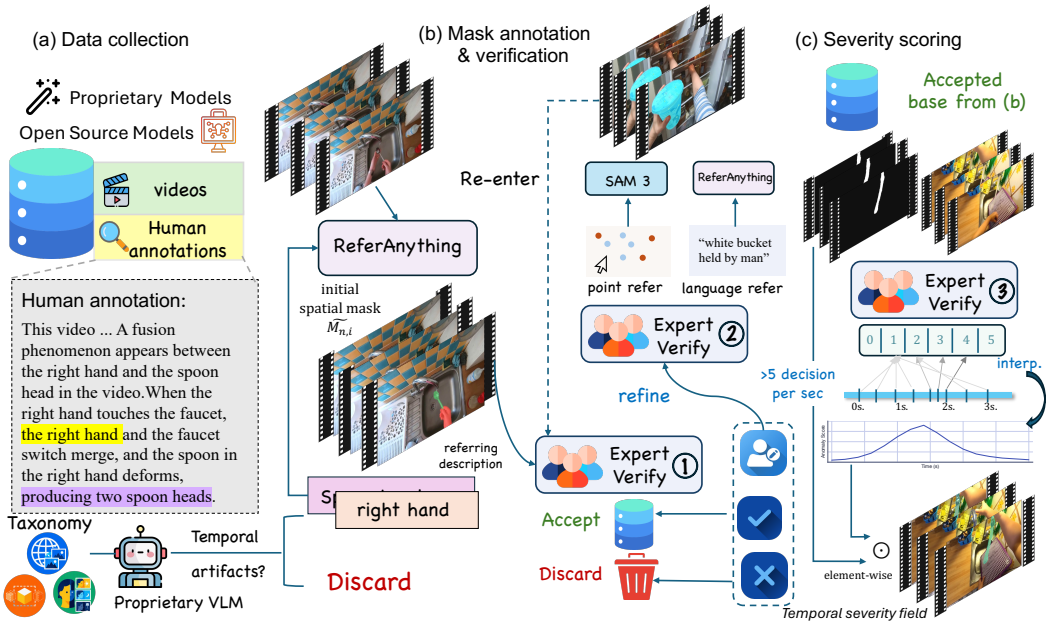

Figure 4: Construction pipeline of TA-Bench. **(a) Mask generation:** we collected and filtered human-annotated videos and produce candidate artifact masks. **(b) Mask verification:** annotators accept, discard, or refine each candidate mask via point prompts or re-written language prompts. **(c) Severity scoring:** annotators assign sparse severity scores on a 0–5 scale, which are interpolated and multiplied with the spatial mask to yield the final spatial continuous mask.

3 TA-BENCH

3.1 TAXONOMY OF TEMPORAL ARTIFACTS

To comprehensively cover temporal artifacts in AIGC videos, we collect viewer feedback on generated videos and distill the most frequently reported temporal failure patterns. After consolidating these patterns and delimiting the scope, we organize a taxonomy of the temporal artifacts,1 shown in Fig. 3.

At the top level, we partition temporal artifacts follow the principle of *whether the artifact can be attributed to a spatial referent, and, if so, whether the identity of that referent persists across the anomalous interval* and form three types. **Global Consistency Artifact** covers artifacts that happen in the entire frame with no spatial referent. Usually refer to failure imaging pattern in AIGC videos (*e.g.* blur, flicker and color style shift). **Object Permanence Artifact** denotes identity of an entity is *broken*, such as an entity appears, vanishes or changes in number without any visible mechanism. **Appearance Constancy Artifact** covers artifacts in which the identity *persists* while the appearance or surface attributes drift.

3.2 CONSTRUCTION OF TA-BENCH

3.2.1 DATA COLLECTION

To construct TA-Bench efficiently, we develop an semi-automated annotation pipeline with a 3-layer human verification principle,as shown in Fig. 4. Specifically, we collect 17,569 AI-Generated videos from existing human-annoatated video feedback and evaluation datasets Kang et al. (2026); Zhang et al. (2026a). Due to their annotations differ in format (e.g., text and bounding boxes) and types of artifacts, we use Large Language Models (LLMs) to classify the artifacts by their textual annotations under our Temporal Artifacts taxonomy and retain each matched video with its original description, yielding an initial corpus $\mathcal{D}_0 = \{(V_n, L_n)\}_{n=1}^{N}$, V and L denotes the video and its artifact description in natural language. In addition, we manually generate realism videos from current State-of-The-Art open source and proprietary video generation model, and manually annotate them with the same style of artifact description. We supplemented these videos serve as the OOD test set.

3.2.2 ARTIFACT ANNOTATION AND VERIFICATION

For each $(V_n, L_n) \in \mathcal{D}_0$, We use a proprietary LLM to ground every annotated artifact into a specific video instance using spatial, temporal, and relational cues, producing a concise referring expression $D_{n,i}$. Then, ReferAnything Bagchi et al. (2025) generates a candidate binary mask as

$$\widetilde{M}^b_{n,i} = \mathrm{ReferAnything}(V_n, D_{n,i}). \tag{1}$$

This automated stage enables scalable mask acquisition at low cost.

Subsequentlly, we introduce a three-layer human verification to ensure annotation accuracy. Ten experts inspect annotation by giving source video, original annotation, and candidate mask. They were ask to make decision between accept, reject, or refine for each of the masks. Each **accepted** artifact enters the final base as $\{(V_{n,i}, M^b_{n,i})\}_{i=1}$. **Rejected** mask will be discarded for incorrectly grounded masks. **Refine** is for inaccurate masks and they were regenerated from revised referring expressions. Moreover, for fine-grained parts (*e.g.* 2nd ankle of index finger), we offer annotaters SAM3 Carion et al. (2025) model with point prompts to generate mask and propagate across video. The regenerated masks re-enter the review pool to control re-generated quality.

Each accepted mask additionally undergoes temporal severity annotation. Annotators assign severity scores $I^{(j)}_{n,i} \in \{0, 1, 2, 3, 4, 5\}$ at timestamps t_j separated by at most $\tau = 0.2\,\mathrm{s}$, where larger values indicate stronger perceptual anomalies and greater degradation of realism. Then, we linearly interpolate the severity in $t \in [t_j, t_{j+1}]$, and conduct element-wise dot productwith the verified binary mask $M^b_{n,i}(t)$ to obtain the final continous spatial mask:

$$I_{n,i}(t) = \frac{t_{j+1} - t}{t_{j+1} - t_j} I^{(j)}_{n,i} + \frac{t - t_j}{t_{j+1} - t_j} I^{(j+1)}_{n,i}, \qquad M_{n,i}(t) = \frac{I_{n,i}(t)}{5} \odot M^b_{n,i}(t). \tag{2}$$

This lightweight human-in-the-loop stage mitigates model-induced annotation errors while retaining scalability, yielding the verified annotation set $\mathcal{A}_n = \{(V_{n,i}, M_{n,i})\}_{i=1}^{K_n}$.

3.2.3 CONTROLLABLE SYNTHESIS OF GLOBAL INCONSISTENCY

During annotation, we observed that global inconsistencies are rarely present in AI-generated video datasets, yet they severely degrade visual quality. To augment the data, we adopt a controllable synthesis strategy to generate global temporal artifacts. Specifically, we construct a suite of operators including **exposure gain**, **white-balance perturbation**, **Gaussian blur**, **Gaussian noise**, and **saturation modification**. For each video, we randomly sample $n_{\text{clip}} \sim \mathcal{U}[1, 4]$ clips and $n_{\text{frame}} \sim \mathcal{U}[3, 8]$ frames within each selected clip, and apply a single operator with varying magnitudes and temporal trajectories. Concurrently, an all-one mask is paired on the corresponding ground truth. We expand the amount by $4\times$ of the initial base data of global inconsistencies, while ensuring annotation fully controllable. More details are provided in the supplementary material Chap. B.

4 PROPOSED METHOD

4.1 EXISTING MODELS ARE TEMPORALLY BLIND

The next question we try to solve is *"What kind of model could be the solution for temporal artifact diagnosis"*. We then design a probe study to all kinds of methods, including an AI-generated video detection method Tan et al. (2026), a video segmentation model Hong & Do (2026), and general-purpose MLLMs. As existing models may generate heterogeneous outputs, we design a probing protocol that decouples *recognition* and *localization* and enables unified evaluation across modalities. As illustrated in Fig. 5, we construct a dataset $\{V, D, M^b\}$ containing 200 video samples from the intermediate outputs of TA-Bench, where V, D, and M^b denote the video, language description, and binary mask, respectively. For models producing textual outputs, we first employ an MLLM to determine whether the predicted description and the reference description D refer to the same object. We then use ReferAnything to generate a corresponding segmentation mask from the textual output and compute the IoU score. For models producing mask outputs, we first compute the IoU, and subsequently employ an MLLM to verify whether the predicted region corresponds to the same target region as specified by the reference description. Accordingly, the proposed protocol enables the joint validation of arbitrary model outputs from two complementary perspectives.

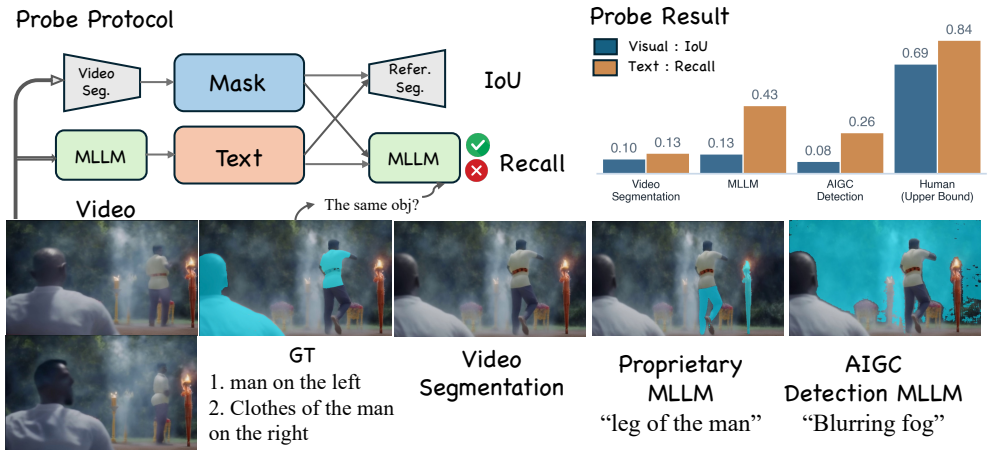

Figure 5: Probe protocol and its result on temporal artifact perception.

The results are shown in Fig. 5. We also provide an upper bound of this protocol fully validated by human to reveal errors in this protocol. These results reveal that none of the evaluated models exhibits meaningful perceptual sensitivity to temporal artifacts. Even the AI-generated video detection methods that make correct binary judgments rely predominantly on pixel-level anomalies and visual distortions rather than on genuine temporal inconsistencies. Video segmentation models perform even worse, producing almost no valid output. We therefore conclude that **existing models are temporally blind**. In light of recent work Moezzi et al. (2026), we attribute this limitation to two complementary factors: MLLMs lack fine-grained visual supervision and hence lack temporal perception ability, whereas video segmentation models lack the capacity to perceive abstract concepts. Addressing temporal artifacts thus calls for a visual backbone endowed with robust temporal perception ability.

4.2 Video Diffusion Diagnoser

Video Generation model conducts generative pretraining in internet-scale data. Hence, it captures rich spatiotemporal world knowledge and vision-language alignment that can serve as a promising backbone. Our model builds on a latent text-to-video diffusion model trained with conditional flow matching. The framework is shown in Fig. 6. Let $x_0 \sim p_{\text{data}}$ be a clean video latent, $x_1 \sim \mathcal{N}(0, I)$ be Gaussian noise, c be the text condition, and $t \sim \mathcal{U}[0, 1]$. Along the linear probability path $x_t = (1-t)x_1 + tx_0$, the model regresses the transport velocity from noise to data, the most commonly used loss for this paradigm is flow mathching v-loss:

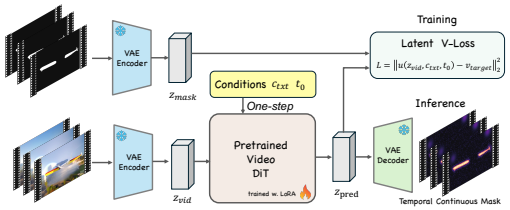

Figure 6: Overall framework of the proposed Video Diffusion Diagnoser (VDD) paradigm.

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{x_0, x_1, c, t}\left[\|u_\theta(x_t, c, t) - (x_0 - x_1)\|_2^2\right]. \tag{3}$$

Following principles of repurposing diffusion model Xu et al. (2025); Wang et al. (2026b), VDD repurposes this transport prior through a single-step feed forward formulation that maps a video latent directly to its latent of temporal continuous mask.

For each annotated video v, let m denote its temporal continuous mask, obtained by compositing the verified per-instance masks of Eq. 2, we use the element-wise maximum, $m(t) = \max_{1 \le i \le K_n} M_{n,i}(t)$). A variational autoencoder encodes the video and its temporal continuous mask $z_{\text{vid}} = \mathcal{E}(v)$ and $z_{\text{mask}} = \mathcal{E}(m)$. We define this video to mask transport velocity as

$$v_{\text{target}} = z_{\text{mask}} - z_{\text{vid}}. \tag{4}$$

Using a fixed timestep $t_0 = 0$ and text prompt c_{txt}, we fine-tune the DiT to regress this displacement in one step:

$$\mathcal{L}_{\text{VDD}}(\theta_{\text{DiT}}) = \mathbb{E}_{(v,m)\sim\mathcal{D}}\left[\|u_{\theta_{\text{DiT}}}(z_{\text{vid}}, c_{\text{txt}}, t_0) - v_{\text{target}}\|_2^2\right]. \tag{5}$$

Table 2: Comparison of all baselines on the in-distribution split of **TA-Bench**. ↑ indicates higher is better, ↓ indicates lower is better; best results among directly comparable methods are in **bold**. Loc. and Diag. denotes Localization and Diagnose, respectively.

Methods	Function		Localization			Diagnose		
	Loc.	Diag.	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$IoU\uparrow$	$MSE\downarrow$	$CC\uparrow$	$SIM\uparrow$
VIRST Hong & Do (2026)	✓	×	0.1679	0.1772	0.1108	–	–	–
ReferDINO Liang et al. (2025)	✓	×	0.2781	0.2884	0.2126	–	–	–
SAMWISE Cuttano et al. (2025)	✓	×	0.2851	0.2603	**0.2834**	–	–	–
VideoMAEv2 Wang et al. (2023)	✓	✓	0.0537	0.0350	0.0000	0.02503	0.0023	0.0903
V-JEPAv2 Assran et al. (2025)	✓	✓	0.0801	0.0362	0.0569	0.02511	0.0017	0.0902
VDD-Wan2.2-5B	✓	✓	0.2907	0.2390	0.2111	0.02248	0.2072	0.1424
VDD-Wan2.2-14B	✓	✓	**0.3918**	**0.3574**	0.2688	0.02107	0.2720	**0.2289**
VDD-ReferAnything	✓	✓	0.3611	0.3430	0.2804	**0.02047**	**0.2885**	0.1750

Because $m \in [0,1]^{T\times H\times W}$ is a temporal continuous mask rather than a binary mask or a video-level score, VDD is trained to output a dense, temporally resolved estimate of artifact severity.

Inference VDD performs one Euler step to reach the latent temporal continuous mask, and uses the decoder of the VAE to obtain the predicted artifact temporal continuous mask:

$$\widehat{m} = \mathcal{D}(z_{\text{vid}} + u_{\theta_{\text{DiT}}}(z_{\text{vid}}, c_{\text{txt}}, t_0)), \tag{6}$$

whose value at every spatiotemporal position lies in $[0,1]$ and encodes artifact severity. A binary artifact mask is recovered by thresholding $\widehat{m}$. Since in the annotation stage, an object with happening artifacts were labelled with at least a score of 0.2, we set to 0.1 as threshold for binarization in our evaluation protocol.

5 Experiments

Baselines and Experiment Setup We instantiate VDD on Wan2.2 at two scales (5B and 14B) as well as on ReferAnything; the latter has been pre-trained on video segmentation corpora, whereas the Wan2.2 variants are trained from their initial video generation checkpoints. We use a heldout of 10% of videos in TA-Bench as ID test set, and a OOD set with 200 videos from unseen generators. We decompose the task into two progressive sub-tasks, namely localization the artifact object (Localization) and regressing the temporal continuous mask (Diagnose): the former evaluates spatial localization, while the latter additionally requires temporal awareness of when artifacts occur. For localization, we compare against VIRST, ReferDINO, and SAMWISE, which produce binary masks and are evaluated by $\mathcal{J}$, $\mathcal{F}$, and $mIoU$ after binarization. For temporal diagnosis, which requires dense continuous values over time, we compare against dense-prediction backbones including VideoMAEv2 Wang et al. (2023) and V-JEPAv2 Assran et al. (2025). For metrics , we use Mean Square Error (MSE), correlation coefficient (CC), and similarity (SIM) in saliency heatmap evaluation Bylinskii et al. (2018), .

Implementation Details We fine-tune the video diffusion transformer with LoRA (rank 16, $\alpha = 32$) at a learning rate of 5×10^{-5} for 10 epochs. To satisfy the Video VAE constraint that the number of frames be $4k+1$, we uniformly sample each video to 73 frames, whereas baselines operate on 72 frames for both training and evaluation. For VideoMAEv2 and V-JEPAv2, we use a lightweight decoder to obtain dense output and fully finetune both encoder and decoder.

Main Results Tables 2 and 4 report results on the in-distribution (ID) and out-of-distribution (OOD) splits. Excluding degenerate dense-prediction baselines, VDD-Wan2.2-14B achieves the strongest ID region and boundary scores ($\mathcal{J} = 0.3918$, $\mathcal{F} = 0.3574$), while VDD-ReferAnything yields the best ID continuous mask regression ($MSE = 0.02047$, $CC = 0.2885$). On OOD videos, VDD-ReferAnything leads all localization metrics and achieves the best MSE and CC, whereas VDD-Wan2.2-14B obtains the highest SIM. In contrast, VideoMAEv2 and V-JEPAv2 approach the trivial all-zero solution, indicating that representations from understanding task or self-supervision offers no help for video artifacts diagnosis.

Table 4: Comparison of all baselines on the out-of-distribution split of **TA-Bench**

Methods	Function		Localization			Diagnose		
	Seg.	Diag.	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$IoU\uparrow$	$MSE\downarrow$	$CC\uparrow$	$SIM\uparrow$
VIRST Hong & Do (2026)	✓	×	0.1593	0.1707	0.0660	–	–	–
ReferDINO Liang et al. (2025)	✓	×	0.2222	0.2390	0.1205	–	–	–
SAMWISE Cuttano et al. (2025)	✓	×	0.1380	0.1443	0.1304	–	–	–
VideoMAEv2 Wang et al. (2023)	✓	✓	0.0428	0.0429	0.0000	0.01825	0.0009	0.0582
V-JEPAv2 Assran et al. (2025)	✓	✓	0.0527	0.0527	0.0433	0.01840	0.0015	0.0583
VDD-Wan2.2-5B	✓	✓	0.2530	0.2317	0.1193	0.01773	0.1287	0.0774
VDD-Wan2.2-14B	✓	✓	0.3090	0.2917	0.1455	0.01733	0.2185	**0.1623**
VDD-ReferAnything	✓	✓	**0.3171**	**0.3095**	**0.1785**	**0.01574**	**0.2287**	0.1113

Scaling Behaviour Scaling Wan2.2 from 5B to 14B consistently improves every metric on both splits; for example, ID IoU rises from 0.1593 to 0.2688 and CC from 0.2004 to 0.2720. Segmentation pretraining provides complementary benefits: relative to VDD-Wan2.2-14B, VDD-ReferAnything improves ID IoU, MSE, and CC, and further improves all OOD metrics except SIM. These results indicate that larger generative priors strengthen temporal-artifact perception, while segmentation priors improve localization and transfer to unseen generators.

Effect of Crop-Consistent Supervision Our continuous severity field remains semantically valid under temporal cropping because the video and its field can be restricted to the same interval, including intervals containing no visible artifact. This property can be exploited through label preserving random temporal cropping during training. We add this augment strategy with 50% and conduct experiments in various backbone. As shown in Table 3, random cropping substantially improves IoU from 0.1593 to 0.2369, while reducing MSE from 0.02260 to 0.02186 and increasing CC from 0.2004 to 0.2144. Although $\mathcal{J}$ decreases from 0.2907 to 0.2409, the gains in region overlap and continuous-field metrics show that crop-consistent supervision improves dense temporal diagnosis rather than uniformly improving every localization criterion. Applying the same strategy to a different backbone, VDD-ReferAnything, again lowers MSE (0.02047→0.01987) and raises CC (0.2885→0.2909); because these two improvements hold under both backbones, the proposed crop-consistent supervision is backbone-agnostic rather than tied to a particular architecture.

Table 3: Ablation of crop-consistent supervision across two backbones. Shaded rows apply temporal random crop augmentation.

Training strategy	$\mathcal{J}\uparrow$	$IoU\uparrow$	$MSE\downarrow$	$CC\uparrow$
VDD-Wan2.2-5B				
Uniform Sampling	**0.2907**	0.2111	0.02248	0.2072
+Temporal RandomCrop	0.2409	**0.2369**	**0.02186**	**0.2144**
VDD-ReferAnything				
Uniform Sampling	0.3611	**0.2804**	0.02047	0.2885
+Temporal RandomCrop	**0.3652**	0.2758	**0.01987**	**0.2909**
Mean Δ	−8.0%	+23.6%	−3.1%	+3.9%

6 Conclusion

We introduce the task of dense temporal artifact diagnosis in AI-generated videos. To enable this task, we build TA-Bench, a benchmark of 6,945 videos from nine open-source and proprietary generators, constructed through automatic triage, semi-automated mask generation, controllable synthesis, and three-stage human verification. Unlike existing artifact benchmarks, TA-Bench provides temporal continuous bench masks that jointly encode artifact location, temporal occurrence, and severity and remaining crop consistency. Constructing these annotations is challenging because temporal artifacts are often subtle, context-dependent, and visible only over short intervals; ten expert annotators spent 210 hours verifying and refining candidate masks and assigning severity scores, forming 214.3K annotations. We further propose VDD, which repurposes pretrained video diffusion models to predict these masks in a single feed-forward step. Together, TA-Bench and VDD establish a benchmark and baseline for localized, fine-grained, and interpretable diagnosis beyond real/fake classification.

Limitations and Future Work Our current study focuses on temporal-artifact diagnosis itself and has not yet transferred the learned representations or predictions to downstream tasks. Future work will investigate whether VDD can improve generalizable AI-generated video detection and provide artifact-aware feedback for evaluating or refining video generation models.

AI USE STATEMENT

In this work, we used generative AI tools for polish writing, Generate synthetic data sets,assist with translation, clean and reformat dataset, support qualitative and thematic data analysis,interpret results. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments,and formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, are not applicable to this work.Additionally, we used generative AI tools for create or modify scientific figures or images, create or edit software code.

We have reviewed all AI-assisted work. 'The primary use of AI was to support dataset construction. LLMs(MLLMs) were used to classify artifact descriptions and generate or refine referring expressions, We have set 3-layer human verification to mitigate hallucination. Probe study in Sec.4.1 are produced by an MLLM-based evaluator. We therefore report a human-validated oracle and treat the comparisons as estimate values. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Anurag Bagchi, Zhipeng Bao, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Refereverything: Towards segmenting everything we can speak of in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23221–23231, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HA8KSQW7SO.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

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.

Aditya Chinchure, Sahithya Ravi, Pushkar Shukla, Vered Shwartz, and Leonid Sigal. Spotlight: Identifying and localizing video generation errors using vlms, 2025. URL https://arxiv.org/abs/2511.18102.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3395–3405, June 2025.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025.

Google DeepMind. Gemini omni. https://deepmind.google/models/gemini-omni/, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.127. URL https://aclanthology.org/2024.emnlp-main.127/.

Jihwan Hong and Jaeyoung Do. Virst: Video-instructed reasoning assistant for spatiotemporal segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890.

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.

Jenna Kang, Maria Beatriz Silva, Patsorn Sangkloy, Kenneth Chen, Niall L. Williams, and Qi Sun. Geneva: A dataset of human annotations for generative text to video artifacts. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6174–6183, 2026. doi: 10.1109/WACV61042.2026.00597.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jieyu Li, Xin Zhang, and Joey Tianyi Zhou. Aegis: Authenticity evaluation benchmark for ai-generated video sequences. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 13346–13353, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758295. URL https://doi.org/10.1145/3746027.3758295.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: Ai-generated video detection via grounded artifact reasoning. *arXiv preprint arXiv:2512.15693*, 2025b.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. Referdino: Referring video object segmentation with visual grounding foundations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20009–20019, October 2025.

Sangbeom Lim, Seoung Wug Oh, Jiahui Huang, Heeji Yoon, Seungryong Kim, and Joon-Young Lee. Videomama: Mask-guided video matting via generative prior. *arXiv preprint arXiv:2601.14255*, 2026.

Jiahao Lin, Weixuan Peng, Bojia Zi, Yifeng Gao, Xianbiao Qi, Xingjun Ma, and Yu-Gang Jiang. Brokenvideos: A benchmark dataset for fine-grained artifact localization in ai-generated videos, 2025. URL https://arxiv.org/abs/2506.20103.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

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.

Minimax. Minimax h3. https://www.minimax.io/blog/minimax-h3, 2026.

Shayda Moezzi, Umer Saleem, Andong Deng, Chen Chen, and Sarah Ostadabbas. Position: Video LLMs must not ignore the pixel dynamics in plain sight. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026. URL https://openreview.net/forum?id=aLWz7CfnMc.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? pp. 948–958, 2026.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

Hao Tan, Jun Lan, Senyuan Shi, Zichang Tan, Zijian Yu, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Videoveritas: Ai-generated video detection via perception pretext reinforcement learning. *arXiv preprint arXiv:2602.08828*, 2026.

Yuqi Tang, Yang Shi, Zhuoran Zhang, Qixun Wang, Xuehai Bai, Yue Ding, Ruizhe Chen, Bohan Zeng, Xinlong Chen, Xuanyu Zhu, Bozhou Li, Yuran Wang, Yifan Dai, Chengzhuo Tong, Xinyu Liu, Yiyan Ji, Yujie Wei, Yuhao Dong, Shilin Yan, Fengxiang Wang, Yi-Fan Zhang, Haotian Wang, Yuanxing Zhang, and Pengfei Wan. Artifact-bench: Evaluating mllms on detecting and assessing the artifacts of ai-generated videos, 2026. URL https://arxiv.org/abs/2605.18984.

Jiyuan Wang, Huan Ouyang, Jiuzhou Lin, Chunyu Lin, Dewen Fan, Boheng Zhang, Haonan Fan, Fei Zuo, Jia Sun, Huaiqing Wang, Honglie Wang, Yiyang Fan, Zhenlong Yuan, Zijun Li, Yongrui Heng, Guosheng Lin, Fan Yang, and Tingting Gao. Cac: Advancing video reward models via hierarchical spatiotemporal concentrating, 2026a. URL https://arxiv.org/abs/2605.11723.

Letian Wang, Chuhan Zhang, Rishabh Kabra, Jasper Uijlings, Steven Waslander, Andrew Zisserman, Joao Carreira, Kaiming He, Misha Andriluka, Eduard Gabriel Bazavan, et al. Video generation models are general-purpose vision learners. *arXiv preprint arXiv:2607.09024*, 2026b.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, June 2023.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

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

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090*, 2024.

Guangkai Xu, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, Chunhua Shen, et al. What matters when repurposing diffusion models for general dense perception tasks? In *International Conference on Learning Representations*, volume 2025, pp. 6786–6799, 2025.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-eval: Evaluating physical realism in generated video via human reasoning. *arXiv preprint arXiv:2603.19607*, 2026a.

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, 2026b.

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.

Chen Zhu, Jiashu Zhu, Yanxun Li, Meiqi Wu, Bingze Song, Chubin Chen, Jiahong Wu, Xiangxiang Chu, and Yangang Wang. Artifact-aware evaluation for high-quality video generation, 2026. URL https://arxiv.org/abs/2601.20297.

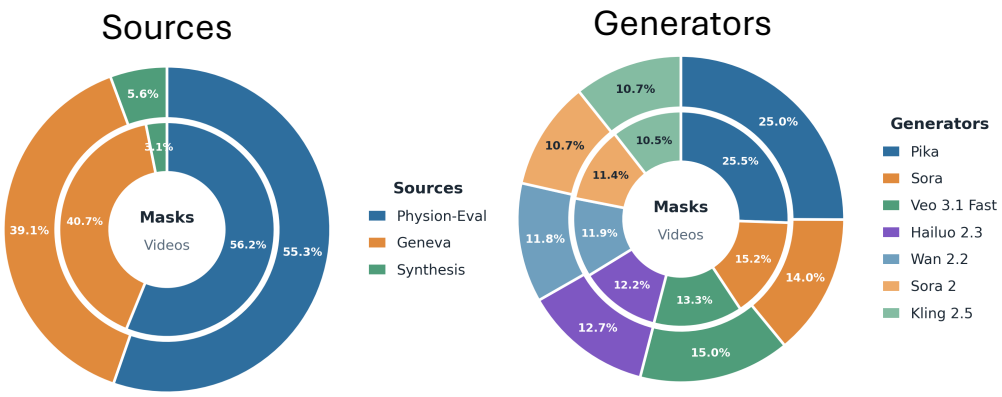

Figure 7: Composition of the main TA-Bench data pool. Left: source composition. Right: generator composition. In each nested donut chart, the inner ring denotes video share and the outer ring denotes artifact-mask share. The independently held-out OOD test set is excluded.

A Statistics of TA-Bench

A.1 Composition of the Data Pool

The main TA-Bench data pool contains 6,745 videos and 8,418 artifact masks collected from Physion-Eval, Geneva, and our controllable global synthesis pipeline. The independently curated out-of-distribution (OOD) test set is excluded from the following percentages so that its unseen generators are not conflated with the main data distribution.

Figure 7 summarizes the source and generator composition of the main data pool. The inner ring in each chart reports video share and the outer ring reports artifact-mask share.

The main pool covers seven video generators. Sora and Sora 2 denote different generator versions and are therefore reported separately. The ICV videos inherit the generator identities of their source videos.

In addition to the main data pool, we construct an independent 200-video OOD test set using Gemini Omni and MiniMax H3 Google DeepMind (2026); Minimax (2026). The two generators are unseen during training. Their videos are selected from two raw candidate pools containing 230 Gemini Omni generations and 670 MiniMax H3 generations, respectively.

A.2 Three-Layer Human Verification

We use a three-layer human verification protocol to convert model-generated candidate masks into artifact-aligned continuous severity fields:

1. **Candidate verification.** An annotator jointly inspects the source video, the original artifact description, and the candidate mask, and determines whether the candidate correctly identifies the visible artifact.
2. **Mask re-generation.** An inaccurate or incomplete mask is reconstructed using one of two interfaces. For language-referable objects, the annotator rewrites the referring expression and queries ReferAnything again. For fine-grained regions that are difficult to describe textually, the annotator supplies point references to SAM 3. Every reconstructed mask is returned to the annotation pool and reviewed by a different annotator.
3. **Severity annotation.** For every verified artifact, an annotator assigns a severity score from 0 to 5 at temporal intervals no longer than 0.2 seconds. The sparse scores are linearly interpolated along

Table 5: Mask-level outcomes of human verification and reconstruction.

Annotation outcome	Share
Pass	57.46%
Reject	24.45%
Refined mask	18.09%
Total	**100%**

time and normalized to $[0, 1]$. Multiplying the interpolated severity by the verified binary mask yields the final continuous field.

For consecutive annotated timestamps t_j and t_{j+1}, the temporal severity and final field are defined as

$$\begin{aligned} I_{n,i}(t) &= \frac{t_{j+1} - t}{t_{j+1} - t_j} I_{n,i}^{(j)} + \frac{t - t_j}{t_{j+1} - t_j} I_{n,i}^{(j+1)}, \\ M_{n,i}(t) &= \frac{I_{n,i}(t)}{5} \odot M_{n,i}^{b}(t). \end{aligned} \tag{7}$$

Table 5 reports the mask-level outcomes of the verification and reconstruction process. We report only percentages because these categories describe annotation decisions rather than the final dataset composition.

Ten expert annotators participate in the annotation process. To control annotation cost, each initial candidate is assigned to one annotator. Reconstruction nevertheless introduces cross-annotator verification because every reconstructed mask is routed to a different annotator's pool. To measure the consistency of severity judgments, all annotators additionally annotate a shared set of 50 videos while completing their assigned workloads. We measure inter-annotator agreement on this shared set with Krippendorff's $\alpha = 0.74$.

B Controllable Synthesis of Global Artifacts

B.1 Overview

Natural global-consistency artifacts are relatively sparse in existing annotated datasets, although transient exposure shifts, color casts, blur, noise, and saturation changes can substantially degrade temporal quality. We therefore construct a controllable synthesis pipeline that injects global imaging degradations into short temporal intervals. The perturbation type, temporal location, duration, trajectory, and magnitude are all recorded, yielding exact frame-level and pixel-level supervision.

We first select a pool of 500 videos. For each selected video, we sample one degradation operator, one temporal modulation function, and between one and four non-overlapping clips. Each clip contains three to eight frames and lies entirely within the video. Frames outside the selected clips remain unchanged. All 500 sampled videos are synthesized successfully; 212 of them are retained in the final main data pool.

B.2 Photometric Domain and Degradation Operators

For exposure, white-balance, and noise perturbations, we approximately invert the display gamma before applying the transformation:

$$L = \left(\max\left(\frac{I}{255}, 10^{-6} \right) \right)^{\gamma_{\text{enc}}}, \qquad \gamma_{\text{enc}} = 2.2, \tag{8}$$

where I is an input frame in the display domain and L is its pseudo-linear representation. After applying an operator, we map the result back to the display domain by

$$I' = \text{clip}(L', 0, 1)^{1/\gamma_{\text{enc}}} \cdot 255. \tag{9}$$

Table 6: Controllable operators used to synthesize global temporal artifacts.

Operator	Transformation	Domain
Exposure gain	$g_t = 1 + U(1.2, 2.8)m_t$, $L' = g_t L$	Pseudo-linear
White-balance perturbation	$g_{t,c} = 1 + s_c U(0.5, 1.0)m_t$, $s_c \in \{-1, 1\}$	Pseudo-linear
Gaussian defocus	$k_t = \mathrm{odd}(\mathrm{round}(5 + 20m_t))$	Spatial
Additive Gaussian noise	$L' = L + \mathcal{N}(0, \sigma_t^2)$, $\sigma_t = U(0.08, 0.20)m_t$	Pseudo-linear
Saturation modification	$S' = \mathrm{clip}(\alpha_t S, 0, 255)$, $\alpha_t = 1 \pm U(1.0, 2.5)m_t$	HSV

Table 7: Temporal modulation functions used within a sampled clip of length L.

Trajectory	Modulation	Interpretation
Step	$m_t = 1$	Abrupt change with constant magnitude
Ramp	$m_t = \mathrm{linspace}(0.3, 1.0, L)$	Gradually increasing drift
Pulse	$m_t = 0.4 + 0.6(1 - u_t^2)$, $u_t = \mathrm{linspace}(-1, 1, L)$	Transient flicker peaking near the clip center

This approximation avoids directly applying photometric gains in the gamma-compressed display domain.

Table 6 summarizes the five operators, where $m_t \in (0, 1]$ denotes the temporal modulation magnitude at frame t and $U(a, b)$ denotes uniform sampling over $[a, b]$.

B.3 Temporal Modulation and Ground Truth

We use three temporal trajectories within each sampled clip, as summarized in Table 7.

The perturbation always acts on the full spatial extent of an affected frame but only during the sampled temporal clips. Accordingly, the binary ground-truth mask is spatially all-one for an affected frame and all-zero otherwise:

$$M_t^b(x, y) = \begin{cases} 1, & t \text{ belongs to a sampled clip,} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Across the 500 synthesized videos, the sampled clips have a mean length of 5.5 frames, and each video contains between 3 and 32 affected frames, with a mean of 13.8. The final main data pool retains 212 synthesized videos. Their operator distribution is M3-defocus: 53, M3-noise: 43, M3-saturate: 41, white-balance perturbation: 41, and exposure gain: 34. Each retained video contributes one synthetic global-inconsistency mask and additionally retain object-level temporal continuous masks inherited from its source video.

C VDD Design Ablations

We conduct the following design ablations using the VDD-Wan2.2-5B backbone. All variants use the same training split, optimization schedule, checkpoint-selection rule, and ID evaluation protocol. The default configuration uses $t_0 = 0$ and a fixed task prompt; its reported ID result is included as the reference row.

As shown in Table 8, pretrained weights provide the clearest benefit: removing them degrades every localization and diagnosis metric, confirming that VDD relies on the spatiotemporal priors learned during video generation. In contrast, replacing the task prompt with random text produces only modest and mixed changes, indicating that the text condition contributes little to this task. The choice of t_0 similarly has no consistent effect: setting $t_0 = 1$ improves $\mathcal{J}$ and $\mathcal{F}$ but reduces IoU and all continuous-mask metrics. These results suggest that VDD's performance is driven primarily by pretrained visual representations rather than a particular text prompt or flow-matching timestep.

Table 8: Ablations of pretrained weights, text conditioning, and the fixed flow-matching timestep using VDD-Wan2.2-5B.

t_0	Text condition	Pretrained	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	IoU↑	MSE↓	CC↑	SIM↑
0.00	Task prompt	✓	0.2907	0.2390	0.2111	0.02248	0.2072	0.1424
0.00	Task prompt	×	0.2795	0.2136	0.2038	0.02371	0.1792	0.1246
0.00	Random text	✓	0.2985	0.2594	0.2066	0.02389	0.2003	0.1415
1.00	Task prompt	✓	0.3532	0.2926	0.1845	0.02560	0.1841	0.1233

D QUALITATIVE RESULTS

These examples are selected to illustrate two distinct temporal-artifact categories rather than to represent aggregate performance. They also reveal remaining errors, including inaccurate boundaries, occasional over-segmentation, and inconsistent responses across frames.

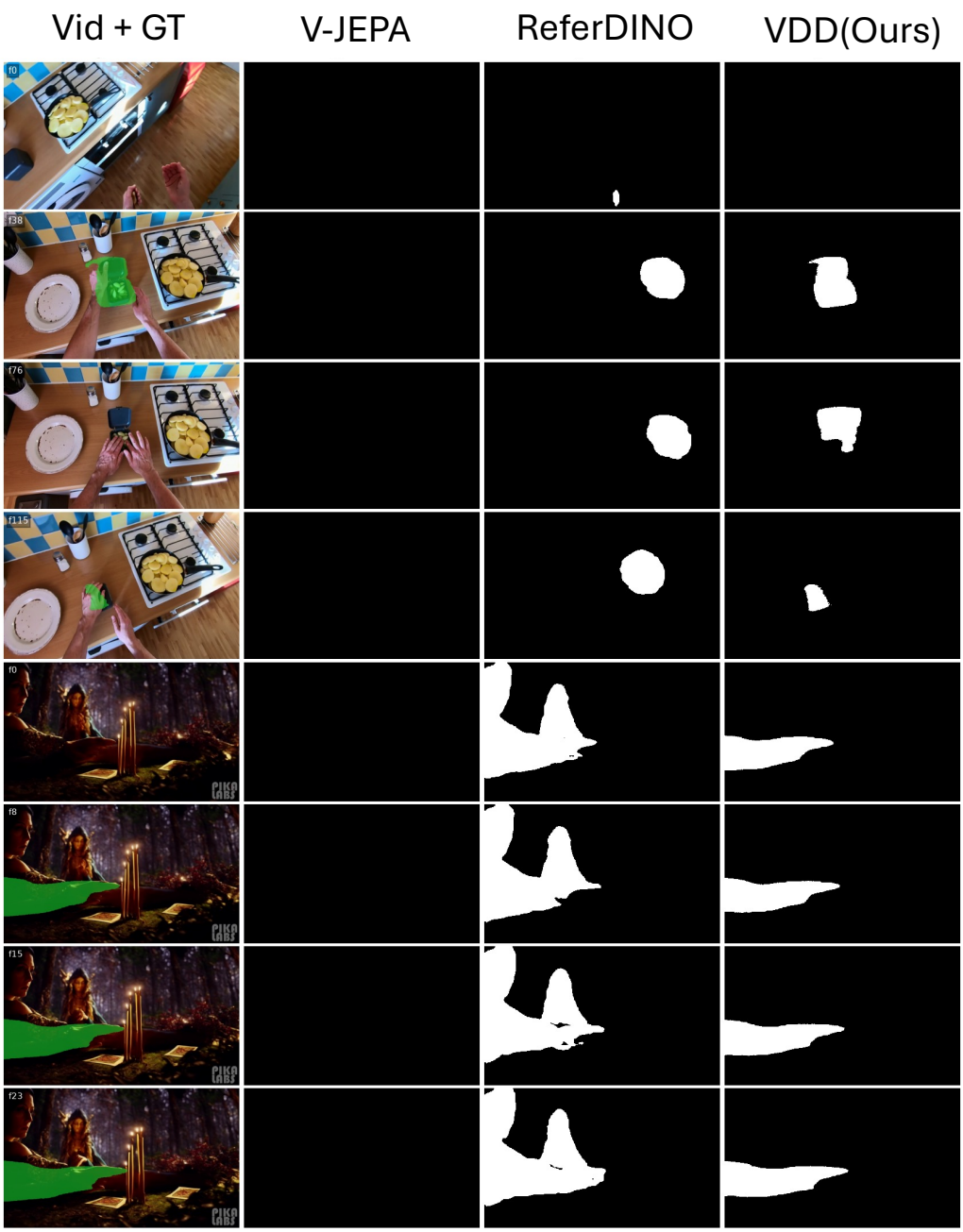

Figure 8: Qualitative comparison of baselines.