LoT-Pass: Long-term-robust Image Watermarking for Image to Video Generation

Guanjie Wang[1], Zehua Ma[1†], Han Fang[1†], and Weiming Zhang[1]

University of Science and Technology of China
wangguanjie@mail.ustc.edu.cn, {mzh045, fanghan, zhangwm}@ustc.edu.cn
† Corresponding author

Abstract. While Image-to-Video models unlock new creative possibilities, they also raise risks of unauthorized adaptation, deepfakes, and scenario manipulation, making robust authentication and traceability solutions urgent for image owners. While image watermarking offers a solution, existing methods expose a critical vulnerability: due to the inherent semantic drift of generated videos, images watermarks survive only in initial frames. This allows attackers to easily evade detection via *temporal cropping*—discarding early frames to obtain watermark-free derivative videos. Consequently, in an era of pervasive I2V models, *long-term robustness* is imperative for image watermarking. Since simply increasing forward embedding strength fails against severe generative divergence, we introduce **LoT-Pass**, a watermarking framework built on a new insight: only one-directional robustness enhancement is hard to achieve long-term robustness. We propose a two-way strategy: (1) expand the robustness distance by training under simulated temporal evolution in I2V generation, and (2) recover unextractable frames by reversing them through a temporal inversion module. This shift, from merely resisting distortions to actively rewinding the video to a watermark-friendly state, enables LoT-Pass to maintain extractability even under temporal drift. Experiments on mainstream open-source and commercial I2V models demonstrate that LoT-Pass achieves stronger long-term robustness while preserving imperceptibility, offering a new paradigm for image copyright protection in the era of widespread I2V adoption. Code

1 Introduction

With the rapid advancement of video generation models, image-to-video (I2V) generation techniques have become increasingly widespread (e.g., Stable Video Diffusion [4], Wan [32], Hunyuan [18]). However, this capability also introduces severe risks of malicious exploitation. For instance, copyrighted artworks or photos of public figures can be misappropriated as input conditions to synthesize unauthorized derivative works, deceptive misinformation, or deepfakes [1] [2]. Consequently, establishing reliable provenance tracing not just for images themselves,

[1] https://www.bbc.com/news/articles/cy402zx0w53o
[2] https://abcnews.go.com/US/video/fake-images-generated-ai-spreading-social-media-compounding-114824660

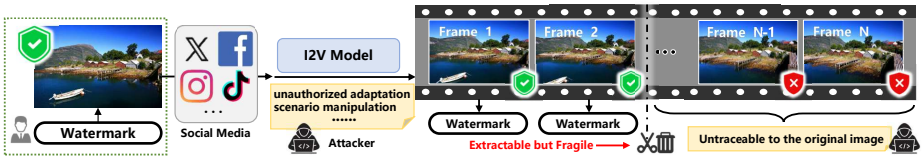


Fig. 1: IP laundering via temporal cropping in uncontrolled I2V generation. While source images are protected, attackers exploit the fragile, short-term robustness of existing watermarks during uncontrolled I2V generation. By discarding early extractable frames and isolating late-stage drifted frames, attackers can easily create untraceable derivative works.

but across the entire lifecycle of media generated from them, has become an urgent regulatory necessity.

To mitigate such misuse, recent regulations have called for mechanisms that restrict the generation of sensitive content and support reliable provenance tracing within synthetic media. Digital watermarking, which imperceptibly embeds ownership information into a source image and enables post-hoc verification, is widely regarded as a promising tool for this purpose [9,10,13,14,22,29,30,33,35,37]. However, when applied to uncontrolled, black-box I2V generation, existing image watermarks expose a critical security vulnerability. As illustrated in Figure 1, the inherent generative process introduces severe semantic and structural drift over time. While the watermark is reliably detected in the early frames that closely resemble the source image, extraction reliability sharply degrades in later frames as the generated content diverges from the original condition.

This short-term survival creates a trivial bypass for attackers: **IP Laundering via Temporal Cropping**. By intentionally driving the I2V model to produce temporally extended videos and subsequently discarding the initial frames (which contain both the watermark and obvious visual resemblance to the source), an attacker can easily obtain highly plausible, watermark-free derivative frames. Relying solely on early frames for detection fundamentally fails in this adversarial scenario, as the most valuable and misleading segments of the forged video often lie in the later, heavily drifted stages. Therefore, to ensure persistent provenance, watermarks must maintain *long-term robustness*—surviving deep into the temporal horizon of the generated video.

We argue that achieving such long-term robustness cannot rely exclusively on pushing the limits of forward embedding strength. As the non-linear semantic divergence grows during I2V generation, even the strongest watermarking methods encounter a saturation point beyond which the signal is destroyed. This motivates a fundamental shift in defense strategy: rather than solely forcing the watermark to endure extreme forward drift, we propose to rewind frames that fall outside this distance back into a watermark-recoverable state.

Based on this insight, we propose **LoT-Pass**, a bidirectional long-term robust watermarking framework designed for I2V generation. LoT-Pass combats temporal drift through two synergistic modules: (1) a *Temporal Evolution Simulation*

module that explicitly models universal generative distortions during training to proactively expand the intrinsic robustness boundary; and (2) a *Temporal Inversion Module* deployed during inference, which approximates backward geometric realignment to pull heavily drifted, late-stage frames back into the decoder's extractable manifold. We evaluate LoT-Pass against SOTA baselines under both open-source and commercial I2V models. Results demonstrate that LoT-Pass achieves significantly improved long-term robustness while preserving imperceptibility.

Our main contributions are summarized as follows:

- We identify the critical vulnerability of existing image watermarks under I2V generation, formulating the necessity of *long-term robustness* to defeat IP laundering and temporal cropping attacks in derivative synthetic works.
- We propose LoT-Pass, a long-term temporal robust watermarking framework based on an encoder-decoder architecture. By introducing a temporal evolution simulation module during training, LoT-Pass models generate-induced distortions and achieve strong robustness.
- We introduce a Temporal Inversion Module at the extraction stage, providing a plug-and-play mechanism to counteract severe temporal degradation and effectively retrieve watermarks from temporally distant frames.
- Experiments demonstrate that LoT-Pass outperforms baselines across four representative open-source and also effective on two commercial I2V models.

2 Related Work

2.1 Image Watermark

With the rapid advancement of deep learning, numerous studies have explored embedding invisible watermarks into images to establish content provenance. *Ma et al.* [26] proposed CIN, a combined reversible and non-reversible mechanism with a non-reversible attention module to address asymmetric extraction under noise attacks. Unlike mainstream architectures, SSLWM [11] embeds messages in the image features extracted by ResNet-50 [12] and extracts messages using a set of learnable secret keys. MuST [33] extends the traditional pipeline by incorporating a dedicated localization network to identify all watermarked regions within composite images. Alternatively, WAM [29] eliminates the need for an explicit localization module by leveraging SAM [17] 's zero-shot segmentation capabilities, thereby enabling simultaneous region detection and watermark extraction. Furthermore, TrustMark [6] specifically optimizes the embedding-extraction pipeline to handle diverse input resolutions seamlessly. Recent efforts have also targeted robustness against advanced generative editing. For instance, *Hu et al.* [13] introduced Robust-Wide, proposing a novel Partial Instruction-driven Denoising Sampling Guidance to withstand complex, instruction-guided image modifications. Similarly, VINE [25] provides an invisible watermarking framework meticulously tailored to resist malicious image editing. Some of these methods can retain limited effectiveness in generated videos, but in most cases, the watermark signal is completely removed in most frames of a generated video.

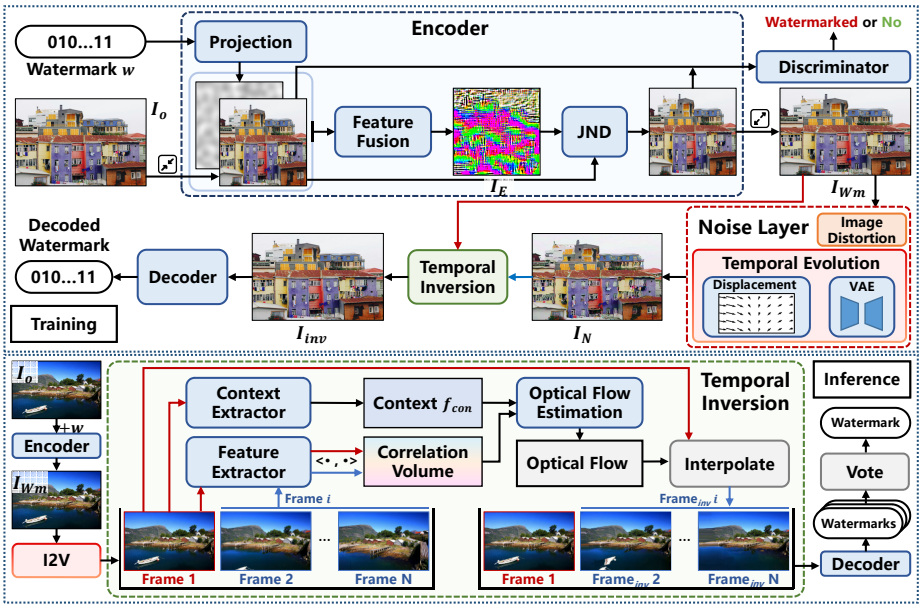


Fig. 2: The framework of the proposed LoT-Pass. The encoder embeds the watermark messages into I_o to generate I_{Wm}. An adversary discriminator is used to improve the visual quality of I_{Wm}. Then, the decoder extracts the watermark message from I_{Wm}, which is enhanced by the noise layer. During inference, for an image input, the watermark is directly extracted using the decoder. For a video input, the sequence is first processed by the temporal inversion module, after which each frame is fed into the decoder to extract its corresponding watermark. Finally, all extracted watermarks are aggregated through a voting scheme to obtain the final watermark.

2.2 Video Generation Model

Since Sora [5] demonstrated the potential of Diffusion Transformers (DiTs), video generation has advanced rapidly. Commercial models like Kling [19] and Hailuo [8] now achieve remarkable long-horizon temporal coherence, which inadvertently exacerbates IP misappropriation risks in extended videos. Concurrently, open-source models have introduced sophisticated architectures for complex spatiotemporal modeling. SVD [4] enables stable I2V generation by integrating temporal consistency layers into pre-trained image diffusion models. To efficiently support longer sequences, CogVideoX [36] employs a novel 3D causal VAE and an expert Transformer, while Wan [32] incorporates advanced spatiotemporal attention for diverse multimodal conditioning. Furthermore, Hunyuan [18] provides a reproducible ecosystem scaling these paradigms. Crucially, while these architectural innovations (e.g., 3D latent compression) dramatically improve video fidelity, they inherently introduce severe non-linear semantic drift, rendering image watermarks fragile over extended frames.

3 Method

In this section, we first present the architecture of LoT-Pass, followed by the details of its training and inference.

3.1 Overall

As illustrated in Figure 2, LoT-Pass consists of three trainable models and two auxiliary modules: a watermark encoder Enc_θ, a watermark decoder Dec_θ, a discriminator Dis_θ used only during training, a JND module for fusing the watermark with the original image, and a noise layer N for adversarial training to enhance robustness. θ denotes the learnable parameters of the model.

The watermark encoder Enc_θ is composed of a message projection module E_P and a feature fusion network E_{FF} based on MUNIT [15]. First, the watermark $w \in \{0,1\}^M$ (with M being the watermark length) is input into the projection module E_P, outputting a watermark latent $w_l \in \mathbb{R}^{256\times256\times3}$. w_l is concatenated with the original image $I_o \in \mathbb{R}^{256\times256\times3}$ then input into E_{FF} to generate the encoded image I_E. Finally, to generate the watermarked image I_{Wm}, we utilize the Just-Noticeable-Difference (JND) [7] module, which is a hand-crafted model of the minimum artifact perceivable by human at every pixel, to fuse I_E and I_o:

$$I_{Wm} = I_o + \lambda \times JND(I_E - I_o) \tag{1}$$

where λ is the JND scaling factor to control the strength of the watermark signal.

The discriminator Dis_θ, implemented based on ResNet [12], is designed to distinguish between watermarked and non-watermarked images. During the training of Enc_θ and Dec_θ, Dis_θ is incorporated into an adversarial training scheme to constrain the watermarked images generated by Enc_θ to be closer to the original images, thereby improving imperceptibility. I_o and I_{Wm} are fed into Dis_θ, yielding corresponding outputs $label_o$ and $label_{Wm}$. These outputs $label_o$ and $label_{Wm}$ will be compared against the ground truth to constrain Dis_θ and Enc_θ's loss.

Temporal evolution simulation is designed to simulate the key operation of the I2V generation process during the training of Enc_θ and Dec_θ, optimizing both models for generating and extracting robust watermark signals. The proposed temporal evolution simulation module includes two additional types of distortions, motivated by the characteristics of video generation, in addition to conventional geometric and non-geometric distortions. First, we employ the VAE of Stable Diffusion [28] to reconstruct watermarked images, efficiently simulating the inherent process of video generation, i.e., compression into latent space, noise addition and removal, and decompression into video frames. Second, we propose a random warping strategy to simulate pixel shifts that may occur during video generation. The formulations are as follows:

$$
\begin{aligned}
g &= \text{Upsample}(g_{\text{low}}), \quad g_{\text{low}} \sim \mathcal{N}(0, \sigma^2) \\
G_{\text{base}}(u, v) &= \left(\frac{2u}{H-1} - 1,\ \frac{2v}{W-1} - 1 \right), \text{for } u = 0, \ldots, H-1,\ v = 0, \ldots, W-1 \\
I_N &= \text{Interpolate}(I_{Wm},\ G_{\text{base}} + \alpha \cdot g)
\end{aligned}
\tag{2}
$$

We first generate a random two-dimensional grid flow field for low-resolution control points $g_{\text{low}} \in R^{grid_{size} \times grid_{size} \times 2}$ by sampling from a Gaussian distribution with a factor σ. The low-resolution grid flow field g_{low} is then upsampled to the target resolution using bicubic interpolation to obtain a smooth and continuous grid flow field g. In parallel, we construct a standard normalized sampling grid $G_{\text{base}} \in R^{H \times W \times 2}$ as the reference sampling coordinates without transformation. By adding the normalized displacement field g (with scale factor α) to the standard grid, we obtain the final sampling coordinates G. Finally, the input image I_{Wm} is resampled according to the final sampling coordinates to produce the randomly distorted image I_N. The noise layer N is composed of the temporal evolution simulation module and various common image distortions to enhance the robustness of watermarking algorithms.

The watermark decoder Dec_θ is constructed on the ConvNeXt [23] architecture, enabling it to extract the watermark signal from the input image effectively. The noised image I_N is first processed by a ConvNeXt backbone to extract features, which are then passed through an MLP to obtain a soft watermark vector. Finally, the vector is binarized to recover the decoded watermark w_{dec}.

3.2 Training Strategy

We jointly train Enc_θ, Dec_θ, and Dis_θ with a two-stage training strategy. In the first stage, the noise layer is not applied, allowing the model to converge to a stable state quickly. In the second stage, the noise layer is activated, and the frequency of each noise is adaptively adjusted at fixed intervals according to the validation. We reweight the different noise according to the following equation:

$$
w_i = \frac{\max(e_i, 0.01)}{\sum_{j=1}^{N} \max(e_j, 0.01)} \tag{3}
$$

We require all noise types to participate in the reweighting process, even if the watermarking scheme already achieves an error rate below 1% under a given noise. This ensures that the model does not discard previously learned robust features while attempting to acquire robustness against other types of noise. Besides, during training, I_N is fed into the temporal inversion module with probability $P = 0.5$. Using I_{Wm} as the reference, the module reconstructs I_N to obtain I_{inv}, which is then input into Dec_θ to recover the watermark. This design

enables the model to learn the signal features present in the aligned image. The overall training objective is formulated as follows:

$$Loss = \lambda_1 L_{enc} + \lambda_2 L_{lpips} + \lambda_3 L_{adv} + \lambda_4 L_{dec} \tag{4}$$

$$L_{enc} = L_2(I_{Wm}, I_o),\ L_{adv} = \log(1 - Dis_\theta(I_{Wm})) \tag{5}$$

$$L_{lpips} = lpips(I_{Wm}, I_o),\ L_{dec} = L_2(w_{dec}, w) \tag{6}$$

where λ_i $(i = 1, 2, 3, 4)$ represent the coefficients for each loss, specifically $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$ and $\lambda_4 = 50$. L_{lpips} is used to constrain the pixel domain's perceptual loss, in which $lpips(\cdot)$ follows *Zhang et al.* [38].

3.3 Inference Strategy

The inference process of LoT-Pass involves a single mode for Enc_θ and two modes for Dec_θ, as shown in Figure 2. For Enc_θ, LoT-Pass supports arbitrary common image resolutions through a residual rescaling and aggregation strategy, which can be formulated as follows:

$$I_{\text{low}} = \text{DownSample}(I_o) \tag{7}$$

$$I_{Wm} = I_o + \alpha \cdot \text{UpSample}(enc_\theta(I_{\text{low}}, w) - I_{\text{low}}) \tag{8}$$

where α indicates the strength factor of the watermark.

For Dec_θ, two modes are provided: an image extraction mode, where the watermark is directly extracted by feeding the image into the decoder, and a video extraction mode, which extends image extraction by incorporating three steps—temporal inversion, decoding, and voting.

Temporal inversion is based on the optical flow estimation model [31]. We first compute the dense optical flow F between the reference frame I_{ref} (by default, the first frame of the acquired video) and the target frame I_{tar} using an optical flow model, which represents the displacement of pixels from the reference frame to the target frame. The detailed calculation process is as follows: First, we extract the features f_{ref} and f_{tar} of I_{tar} and I_{ref}.

$$f_{\text{ref}} = \phi(I_{\text{ref}}), \quad f_{\text{tar}} = \phi(I_{\text{tar}}), \quad f_{\text{con}} = \psi(I_{\text{ref}}) \tag{9}$$

where $\phi(\cdot)$ is the feature extractor and $\psi(\cdot)$ is the context extractor. Then calculate the similarity between pixel pairs.

$$C(i, j, k, l) = \langle f_{\text{ref}}(i, j), f_{\text{tar}}(k, l) \rangle \tag{10}$$

where (i, j) denotes the reference frame feature coordinates, (k, l) denotes the target frame feature coordinates, $\langle \cdot, \cdot \rangle$ denotes the inner product. C represents the global correlation volume.

$$F = \mathcal{H}(\text{Sample}(f_{\text{con}}), C) \tag{11}$$

Table 1: Exhibition on the input image resolution of the model, capacity, bits per pixel (BPP), model parameters, visual metrics (PSNR, SSIM, and LPIPS), and inference cost (inference time and average GPU memory peak usage).

Method	Input Setting			# Params (M)			Visual Metric			Inference Cost		
	Resolution	Capacity (bits)	BPP(%) ↑	Encoder	Decoder	Total	PSNR ↑	SSIM ↑	LPIPS ↓	Encoder Time(s) ↓	Decoder Time(s) ↓	Memory(GB) ↓
SSLWM	224	30	0.06	-	26.42	26.42	36.15	0.9545	0.012	2.3	0.009	0.27
CIN	128	30	**0.18**	5.29	5.00	10.30	38.80	0.9866	0.008	0.025	0.014	**0.15**
MuST	256	30	0.05	0.3	0.63	0.9	36.33	0.9518	0.013	0.081	0.007	0.36
TrustMark	256	100	0.15	8.24	22.61	30.86	38.23	0.9819	0.008	**0.006**	**0.006**	0.41
WAM	256	32	0.05	1.1	96.0	97.1	39.43	0.9793	0.007	0.015	0.015	0.51
VINE	256	100	0.15	958.2	84.58	1042.79	36.74	0.9712	0.010	0.077	0.008	4.43
Robust-Wide	512	64	0.02	29.73	395.48	425.21	**39.62**	**0.9940**	**0.002**	0.027	0.017	2.81
LoT-Pass	256	32	0.05	3.7	47.4	51.1	38.58	0.9840	0.010	**0.006**	0.009	0.27

We utilize the optical flow estimation module $\mathcal{H}$ with C and sampled f_{con} to obtain the final optical flow $F \in \mathbb{R}^{H\times W\times 2}$.

$$X(u,v) = u + F_x(u,v),\ Y(u,v) = v + F_y(u,v) \tag{12}$$

where (u, v) denotes the pixel coordinates of the target frame, and (F_x, F_y) represent the horizontal and vertical components of the optical flow, respectively. Finally, by applying inverse sampling, we resample the target frame into the reference frame's coordinate system to obtain the aligned frame.

$$I_{\text{inv}}(u,v) = \text{Interpolate}(I_{\text{tar}}, (X,Y)) \tag{13}$$

Voting is based on a heuristic assumption: *as video frames move further away from the initial frame, the extracted watermarks increasingly resemble samples drawn from a binomial distribution.* For each frame I_i, obtain its I_{inv} and extract the corresponding $w^{(i)}$. The final watermark w can be obtained by a bit-majority vote in the N watermarks $w^{(i)}$.

$$w^{(1)}, w^{(2)}, ..., w^{(N)}, w^{(i)} \in \{0,1\}^M \tag{14}$$

where L denotes the length of a video.

$$w_j = \begin{cases} 1, & \text{if } \sum_{i=1}^{L} w_j^{(i)} \geq \frac{L}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

The final w of the input video is $(w_1, w_2, ..., w_M)$.

4 Experiments

4.1 Experiment Settings

Datasets. We adopt the widely used DIV2K [1] dataset as the test image set for watermarking methods. To construct the corresponding text prompts, we employ the Qwen2.5-VL-7B-Instruct [3] model to generate descriptions of each image. For I2V models supporting negative prompts, we uniformly apply a negative prompt shown in the Supplementary. In addition, we test the robustness against classic image distortions on the sampled validation set of MS COCO [21] and Flickr2K [20], a total of 200 images.

Metrics. We evaluate the imperceptibility of watermark models using standard metrics, including PSNR [2], SSIM [34], and LPIPS [38]. Watermark capacity is measured by bits per pixel (BPP). For watermark extraction, bit accuracy (ACC) serves as the most fundamental measure of extraction performance. To extend this evaluation to the image-to-video (I2V) setting, we further introduce two video-level robustness metrics: frame accuracy (***FACC***) and video accuracy ($\overline{VACC}$). Based on ***FACC***, we introduce the concept of **Robust Diffusion Distance (RDD)** to evaluate the robust persistence of watermark signals temporally under the modifications induced by Image-to-Video models. These metrics can be formally computed as follows.

$$\boldsymbol{FACC} = (\overline{ACC_{F_1}}, \ldots, \overline{ACC_{F_L}}) \tag{16}$$

$$\overline{ACC_{F_i}} = \frac{\sum_{j=1}^{T} ACC_{F_i}^j}{T}, \quad i = 1, \ldots, L \tag{17}$$

$$RDD = \arg\min_i \overline{ACC_{F_i}} < \tau, \quad \overline{ACC_{F_i}} \in \boldsymbol{FACC} \tag{18}$$

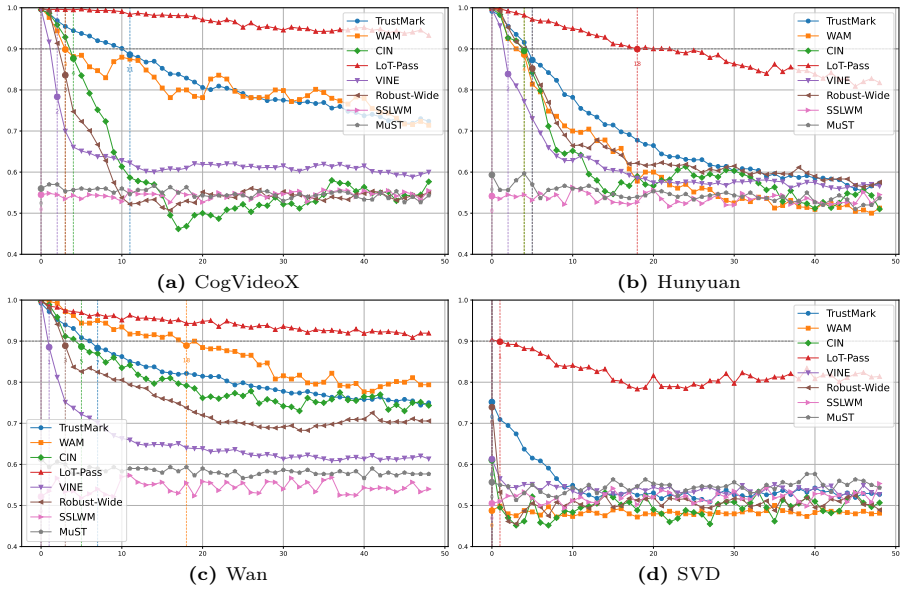
$$\overline{VACC} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{L} ACC_{F_i}^j}{N \times L} \tag{19}$$

where $\overline{ACC_{F_i}}$ denotes the average accuracy of i-th frame across test videos, T denotes the number of test videos, and L denotes the length of a video. Formally, RDD is defined as the first frame index at which the accuracy drops below the threshold $\tau = 0.9$, indicating the temporal boundary of robustness.

Training Implementation. We train LoT-Pass using the MS COCO dataset [21] at a resolution of 256×256. The whole framework is implemented by PyTorch [27] and all experiments executed on a single NVIDIA RTX A6000 GPU. We utilize AdamW [24] as the optimizer. The watermark messages are randomly generated sequences of 32 bits. The batch size in the training stage is set to 16, and the LoT-Pass models are trained for 250 epochs with an initial learning rate = 0.0001. Details are in the Supplementary.

Baselines. We compare LoT-Pass with seven watermark baselines, all utilizing their officially open-source implementations. These baselines include SSLWM [11], CIN [26], MuST [33], VINE-R [25], TrustMark [6], Robust-Wide [13], Watermark Anything Model(WAM) [29]. Although the baselines were trained at different fixed resolutions, we apply watermark residual scaling (using bilinear interpolation methods) to standardize all of them to the original shape of the input images. For fair comparison, all main-text metrics ($\overline{VACC}$ and RDD) are computed before voting; post-voting results are exhibited in the Supplementary.

Baseline Settings. In Table 1, we provide a comprehensive comparison between LoT-Pass and baseline methods, including input/output sizes, message capacity, perceptual quality metrics of the watermarked images, as well as inference time and computational cost for each method. It is worth noting that, to ensure consistent watermark strength, we adjust the PSNR of all watermarked images



(a) CogVideoX **(b)** Hunyuan

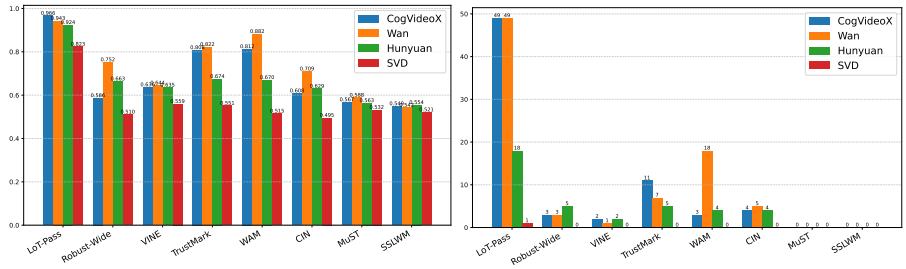
(c) Wan **(d)** SVD

Fig. 3: Frame accuracy (***FACC***) trends of different watermarking methods across four target models.

to be greater than 36 dB. This prevents unfair comparisons that could arise from sacrificing visual quality in exchange for robustness. Regarding the abnormally long encoding time of SSLWM, this can be explained as follows: SSLWM embeds watermarks by first fine-tuning a set of parameters for each batch of images, followed by data augmentation. When a new batch of images arrives, the parameters must be retrained. We argue that the data augmentation stage alone is insufficient for watermark embedding, and the preceding training stage plays a crucial role in this process.

4.2 Results and Analysis

Test Open-source I2V Model As shown in Figure 3 and Figure 4, we evaluate LoT-Pass and multiple baselines across four representative open-source image-to-video generation models: CogVideoX [36], Wan2.1 [32], and Hunyuan [18] as **prompt-guided** I2V models, and Stable Video Diffusion (SVD) [4] as a **non-prompt-guided** I2V model, each generating 50 480P videos for watermark verification. Detailed settings are in the Supplementary. Results demonstrate that LoT-Pass (triangular red line) achieves SOTA performance across all models. Although WAM and TrustMark were not trained with generative model components such as VAEs, they still exhibit stronger robustness than other baselines. The watermark patterns of different methods are visualized in Figure 5. Observation reveals that the watermark patterns of LoT-Pass, WAM, and TrustMark all exhibit distinct large-scale speckled patterns. As shown in Table 2, we compute



(a) $\overline{VACC}$ Comparison **(b)** RDD Comparison

Fig. 4: Comparison of video accuracy ($\overline{VACC}$) and robustness diffusion distance (RDD) across four target models.

Table 2: VAE latent difference across I2V robust methods.

Methods	LoT-Pass	Robust-Wide	VINE	TrustMark	WAM	CIN
Latent Diff.	0.0023	0.0009	0.0005	0.0015	0.0018	0.0010

the L2 distance between watermarked and unwatermarked images in the SD-v1.4 VAE latent space, where LoT-Pass, TrustMark, and WAM achieve higher values than other methods. We believe that such a speckled distribution allows watermark signals to remain more distinguishable after VAE compression, leading to greater retention during the generation process and, consequently, a larger RDD. Nevertheless, as the generation progresses, the influence of the input image on subsequent latents gradually weakens, leading all watermarking methods to exhibit a declining trend in $\boldsymbol{FACC}$ with increasing temporal distance from the initial frame. Furthermore, for frames that exhibit little to no visual similarity to the original image, we believe that their copyrights and dissemination should likewise be considered independent of the original image. Moreover, under SVD, most methods yield RDD values of zero. We argue this is due to SVD's lack of prompt conditioning, where video generation relies solely on the input image, resulting in uncontrolled degradation relative to the input image.

Test Commercial Model In addition to the four open-source models discussed above, we further evaluate LoT-Pass on two commercial models, including Kling Video 2.1 [19] and Minimax Hailuo 02 [8], generating ten videos each. The same image-prompt pairs used for the open-source models are employed to ensure comparability. As shown in Figure 6, compared with the results on open-source models, the performance on commercial models is better. By examining the quality of the generated videos, we find that commercial models generally produce higher-quality and more temporally coherent outputs, which suggests that improved video generation quality facilitates stronger preservation of watermark signals from the source image. Besides, we hypothesize that commercial models

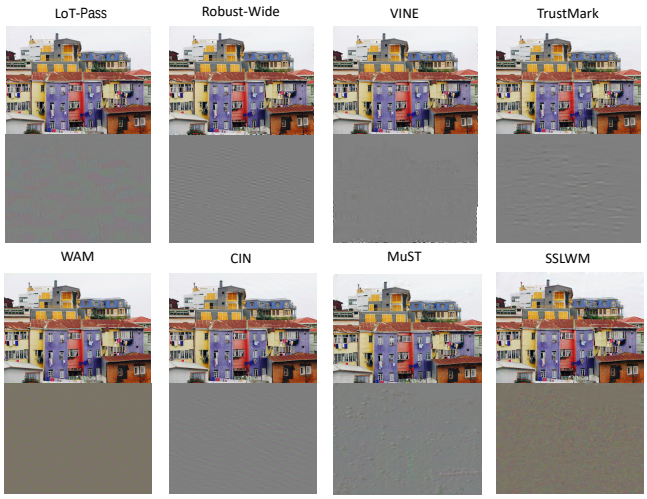


Fig. 5: Watermark pattern visualization of each method. $pattern = \alpha \times (I_{Wm} - I_o)$, where $\alpha = 3$ to make the pattern more easily recognizable.

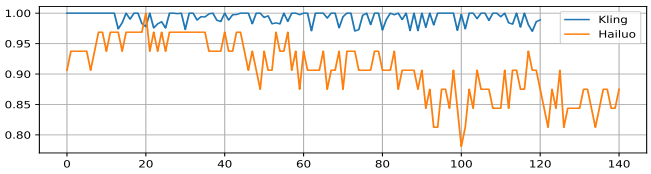


Fig. 6: ***FACC*** of Kling and Hailuo.

may incorporate certain specialized mechanisms that allow the input image to influence even temporally distant frames, thereby leading to consistently strong RDD performance ($RDD_{Kling} = 120, RDD_{Hailuo} = 47$).

Robustness against Frame Drop We further evaluate the robustness of LoT-Pass by discarding a certain proportion from the beginning of video frames on the same video used in Figure 3. As shown in Table 4, LoT-Pass demonstrates superior performance over the other baselines in most cases, even when these early frames, closely resembling the source image, are removed. This indicates that LoT-Pass exhibits inherent robustness to the I2V process regardless of temporal inversion, a conclusion further supported by our ablation studies.

Test Classic Noise We evaluated each method under classical distortion scenarios. As shown in Table 3, WAM, Robust-Wide, and LoT-Pass all demonstrated quite satisfactory performance. Watermarking methods that excel in the I2V task, such as TrustMark and WAM, also exhibit strong robustness against classical geometric distortions (e.g., cropping). LoT-Pass remains competitive

Table 3: The average $\overline{ACC}(\%)$ results under distortion settings. The parameters for each test image are randomly sampled from the range defined by the upper and lower bounds (represented by the two lines in the settings). **Bold** results indicate the best outcome among the comparisons, while underlined results signify the second-best. More results and details are in the Supplementary.

Type	Identity	RealCrop	Crop	RemoveLines	Resize	JPEG	GB	GN	Brightness	Contrast	Saturation	R.Bending
Settings	-	40%	40%	$freq_x = 2$	0.3	$Q = 50$	$\sigma = 0.5$	$s = 0.25$	0.5	0.5	0.5	$\sigma = 0.5$
	-	80%	80%	$freq_y = 2$	1.3	$Q = 80$	$\sigma = 1.5$	$s = 0.75$	2	2	2	$\sigma = 0.5$
SSLWM	98.22	50.16	73.48	94.27	94.50	89.18	96.18	92.09	84.90	76.58	83.60	56.82
CIN	99.67	50.23	90.98	98.68	99.50	82.36	98.47	96.70	98.63	**99.90**	99.03	59.47
MuST	96.73	88.97	83.90	93.30	92.37	73.03	94.37	65.93	75.70	79.87	84.34	65.00
TrustMark	99.67	65.26	92.13	98.45	98.79	98.75	99.18	88.82	92.99	94.68	99.06	59.26
WAM	99.22	**99.09**	95.59	99.72	96.00	80.91	99.50	99.31	94.69	95.78	97.00	74.78
VINE	99.57	50.38	51.82	99.39	99.13	96.03	99.52	97.00	97.39	99.01	99.38	56.65
Robust-Wide	**99.99**	50.82	89.03	**99.99**	**99.99**	98.90	**99.99**	98.76	98.18	99.39	**99.99**	60.88
LoT-Pass	99.98	98.09	**97.38**	99.91	99.91	**99.72**	**99.94**	**99.75**	**99.56**	99.06	99.91	**97.53**

Table 4: $\overline{VACC}$ and RDD of LoT-Pass under frame drop.

$\overline{VACC}/RDD$	0	5%	10%	15%	20%
CogVideoX	0.966/49	0.953/49	0.949/36	0.931/20	0.910/13
Hunyuan	0.924/18	0.895/14	0.874/8	0.858/2	0.795/0
Wan	0.943/49	0.928/36	0.912/21	0.894/15	0.874/7
SVD	0.823/1	0.698/0	0.629/0	0.625/0	0.582/0

under geometric transformations and demonstrates superior robustness to random bending attacks (random pixel displacements), achieving over 22% improvement over the second-best method. Meanwhile, it further demonstrates superior robustness to numerical distortions (e.g., Gaussian blur, Gaussian noise) compared to other methods. This observation indicates that the I2V process inherently combines variations in pixel values and motions, requiring robustness across both dimensions for resistance. Moreover, results validate the effectiveness of our noise layer, which jointly accounts for both distortion types.

4.3 Ablation Study

Effectiveness of Temporal Evolution Simulation. As shown in Figure 7, we compared two models trained for the same iterations: one with the temporal evolution simulation module in the noise layer and one without (both using temporal inversion during testing). It is evident that models trained without temporal evolution significantly underperform compared to LoT-Pass. This demonstrates that our temporal evolution simulation module effectively captures the key characteristics of I2V generation: VAE compression-reconstruction is an indispensable process, and random displacement simulates pixel motion.

Table 5: $\overline{VACC}$ with different motion intensity.

Models	CogVideoX	Hunyuan	Wan	SVD
$\overline{VACC}$ / Low Motion Video Num	0.971/34	0.936/15	0.956/19	0.883/2
$\overline{VACC}$ / High Motion Video Num	0.955/16	0.919/35	0.935/31	0.821/48

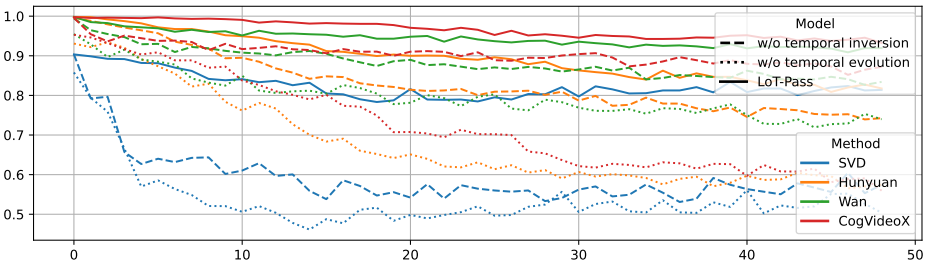


Fig. 7: ***FACC*** comparison of LoT-Pass, w/o temporal inversion (dashed line), and w/o temporal evolution (dotted line).

Effectiveness of Temporal Inversion. To assess the impact of temporal inversion, we evaluated both settings on the same generated videos and computed ***FACC***, as shown in Figure 7. Applying the temporal inversion module yields a significant improvement compared to direct extraction from the raw video. Temporal inversion remaps unknown temporal perturbations (motion shifts, pixel changes, et al) from video generation models back to the pixel domain; temporal evolution simulates temporal perturbations in the pixel domain during training. The combination of two approaches enables LoT-Pass to achieve strong performance.

Influence of Motion Intensity. Using VBench [16], we categorize videos into low- and high-motion subsets based on temporal flickering, motion smoothness, and dynamic degree. As shown in Table 5, although intense generative motions cause more severe spatial drift and slightly degrade the extraction accuracy ($\overline{VACC}$), LoT-Pass maintains exceptional robustness. Even in high-motion scenarios, it achieves a $\overline{VACC}$ exceeding 0.91 on most mainstream models, proving its strong resilience to complex temporal dynamics.

5 Conclusion

In this work, we introduce the cross-modal robust watermarking challenge in the image-to-video (I2V) generation for the first time. To systematically evaluate this challenge, we propose three effective evaluation metrics: RDD, ***FACC***, and $\overline{VACC}$. To address the challenge, we present LoT-Pass, which strengthens watermark signals along the forward temporal axis through a temporal evolution simulation module and enhances backward temporal recovery via the temporal inversion module. Experiments on multiple open-source and commercial video generation models demonstrate that LoT-Pass achieves SOTA performance.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 62402469, U2336206, U2436601, 62121002.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
2. Almohammad, A., Ghinea, G.: Stego image quality and the reliability of psnr. In: 2010 2nd International Conference on Image Processing Theory, Tools and Applications. pp. 215–220. IEEE (2010)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023)
5. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), `https://openai.com/research/video-generation-models-as-world-simulators`
6. Bui, T., Agarwal, S., Collomosse, J.: Trustmark: Universal watermarking for arbitrary resolution images. arXiv preprint arXiv:2311.18297 (2023)
7. Chou, C.H., Li, Y.C.: A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. IEEE Transactions on Circuits and Systems for Video Technology **5**(6), 467–476 (1995)
8. Creation, R.V.: Hailuo ai video generator (2024), `https://hailuoai.video/`
9. Fang, H., Jia, Z., Ma, Z., Chang, E.C., Zhang, W.: Pimog: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In: Proceedings of the 30th ACM international conference on multimedia. pp. 2267–2275 (2022)
10. Fernandez, P., Elsahar, H., Yalniz, I.Z., Mourachko, A.: Video seal: Open and efficient video watermarking. arXiv preprint arXiv:2412.09492 (2024)
11. Fernandez, P., Sablayrolles, A., Furon, T., Jégou, H., Douze, M.: Watermarking images in self-supervised latent spaces. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3054–3058. IEEE (2022)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hu, R., Zhang, J., Xu, T., Li, J., Zhang, T.: Robust-wide: Robust watermarking against instruction-driven image editing. In: European Conference on Computer Vision. pp. 20–37. Springer (2025)
14. Hu, R., Zhang, J., Zhao, S., Lukas, N., Li, J., Guo, Q., Qiu, H., Zhang, T.: Mask image watermarking. arXiv preprint arXiv:2504.12739 (2025)

15. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
16. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
18. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
19. Kuaishou: Kling (2024), https://kling.kuaishou.com/
20. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
22. Liu, X., Chang, H., Wei, J., Zhu, L., Liu, L., Li, L., Zhou, S., Li, C., Xu, D., Gao, W.: Implanting robust watermarks in latent diffusion models for video generation. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10888991
23. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
25. Lu, S., Zhou, Z., Lu, J., Zhu, Y., Kong, A.W.K.: Robust watermarking using generative priors against image editing: From benchmarking to advances. arXiv preprint arXiv:2410.18775 (2024)
26. Ma, R., Guo, M., Hou, Y., Yang, F., Li, Y., Jia, H., Xie, X.: Towards blind watermarking: Combining invertible and non-invertible mechanisms. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1532–1542 (2022)
27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
29. Sander, T., Fernandez, P., Durmus, A., Furon, T., Douze, M.: Watermark anything with localized messages. arXiv preprint arXiv:2411.07231 (2024)
30. Su, Z., Qiu, X., Xu, H., Jiang, T., Zhuang, J., Yuan, C., Li, M., He, S., Yu, F.R.: Safe-sora: Safe text-to-video generation via graphical watermarking. arXiv preprint arXiv:2505.12667 (2025)
31. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)

32. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
33. Wang, G., Ma, Z., Liu, C., Yang, X., Fang, H., Zhang, W., Yu, N.: Must: Robust image watermarking for multi-source tracing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5364–5371 (2024)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
35. Yang, J., Zhang, W., Yang, L., Zeng, S., Gao, T.: Ringet: A robust watermarking framework for diffusion-based video generation. IEEE Transactions on Circuits and Systems for Video Technology **36**(5), 7184–7195 (2026). https://doi.org/10.1109/TCSVT.2025.3643532
36. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
37. Ye, G., Gao, J., Wang, Y., Song, L., Wei, X.: Itov: Efficiently adapting deep learning-based image watermarking to video watermarking (2023)
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (2018)