This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

WaTeRFlow: Watermark Temporal Robustness via Flow Consistency

Utae Jeong[1] Sumin In[1] Hyunju Ryu[1] Jaewan Choi[1]
Feng Yang[2] Jongheon Jeong[1] Seungryong Kim[3] Sangpil Kim[1*]

[1] Korea University [2] Google DeepMind [3] KAIST AI

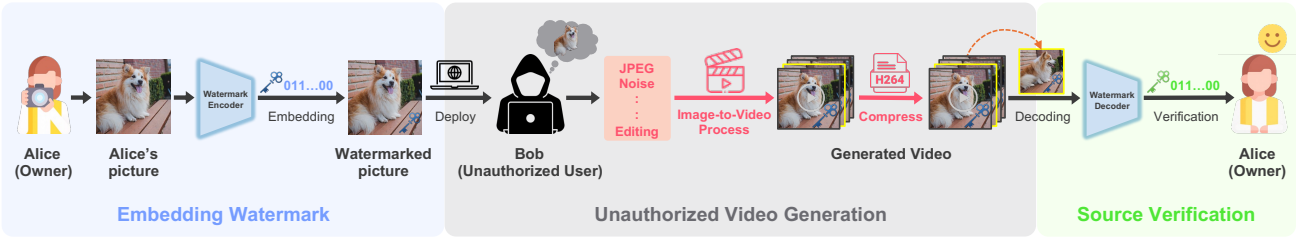


Figure 1. **Application scenario of WaTeRFlow.** Specifically, we consider watermark embedding for image copyright protection and verification from generated videos. Left: An owner, Alice, protects her picture by embedding a watermark. Middle: An unauthorized user, Bob, generates a video from Alice's picture. Right: On Alice's side, copyright is verified by recovering the watermark from the video frames using her watermark decoder, in order to identify the source used to generate the video and verify whether her picture was used.

Abstract

*Image watermarking supports authenticity and provenance, yet many schemes are still easy to bypass with various distortions and powerful generative edits. Deep learning-based watermarking has improved robustness to diffusion-based image editing, but a gap remains when a watermarked image is converted to video by image-to-video (I2V), in which per-frame watermark detection weakens. I2V has quickly advanced from short, jittery clips to multi-second, temporally coherent scenes, and it now serves not only content creation but also world-modeling and simulation workflows, making cross-modal watermark recovery crucial. We present **WaTeRFlow**, a framework tailored for robustness under I2V. It consists of (i) FUSE (Flow-guided Unified Synthesis Engine), which exposes the encoder–decoder to realistic distortions via instruction-driven edits and a fast video diffusion proxy during training, (ii) optical-flow warping with a Temporal Consistency Loss (TCL) that stabilizes per-frame predictions, and (iii) a semantic preservation loss that maintains the conditioning signal. Experiments across representative I2V models show accurate watermark recovery from frames, with higher first-frame and per-frame bit accuracy and resilience when various distortions are applied before or after video generation.*

*Corresponding Author

1. Introduction

Diffusion models [37] have made image editing, including text-guided edits [4, 10, 51], highly reliable. Digital watermarking embeds a message into an image to support authenticity and provenance while remaining imperceptible. Robustness requires accurate recovery after common transforms and edits. Recent watermarking methods are designed for robustness, and embedded watermarks often remain detectable even after editing with diffusion models. Unlike typical image distortions such as JPEG compression that perturb a fixed canvas, image-to-video (I2V) synthesizes temporally coherent frames from a single image, which often acts as a watermark-disrupting transformation by weakening the watermark signal across frequencies and inducing temporally varying subpixel misalignments that make per-frame recovery intrinsically difficult [23].

We study a provenance scenario shown in Fig. 1. Alice embeds a watermark in her picture. Bob, an unauthorized user, inputs that image to an I2V model and produces a video. He may further apply edits, regeneration, compression, or other manipulations before or after video synthesis to suppress the watermark signal. Using the generated video solely, Alice must recover the watermark with high average accuracy and consistently high per-frame accuracy across the frames to verify ownership of the source image.

To address this challenge, we introduce WaTeRFlow, a

watermarking and detection framework for I2V. Our key idea is to couple realistic video-induced effects with warping and temporal regularization when learning the encoder and decoder. We operationalize this via the Flow-guided Unified Synthesis Engine (FUSE), inserted between the encoder and decoder. FUSE unifies diffusion-based image edits and I2V within a single loop by integrating an instruction-driven image editing proxy [4] and a fast video diffusion proxy [42], exposing the system to realistic pixel transformations with low overhead. We warp each generated frame to the first frame using an optical-flow estimator [40] to control the strength of distortions induced by video generation seen by the encoder and decoder, thereby stabilizing learning. We feed edited images and flow-warped frames to the decoder. We apply a semantic preservation loss to keep the conditioning signal used by the video generator close to that of the original image while enforcing a Temporal Consistency Loss (TCL) on adjacent flow-warped frames to stabilize per-frame predictions.

We evaluate the transferability of robustness from video diffusion proxy to two representative I2V models. Specifically, we test on the U-Net based Stable Video Diffusion (SVD) [2] and the Diffusion Transformer (DiT) based CogVideoX [47], yielding consistent gains on both. WaTeRFlow improves both first-frame and per-frame bit accuracy while mitigating late-frame degradation in watermark detection. These outcomes are crucial for verifying content origin in real-world scenarios. For example, on SVD-XT the average bit accuracy increases from 73.92% with VINE [23] to 84.96%, and the first frame reaches 96.93%. Our main contributions are summarized as follows:

- We optimize the system end-to-end with FUSE, which inserts image editing proxy and video diffusion proxy between the encoder and decoder. In addition, we perform flow-guided warping of each generated frame to the first frame, enabling watermark recovery even when manipulations occur before or after video synthesis.
- We introduce a semantic preservation loss during watermark embedding to retain conditioning derived from the watermarked image that guides I2V generation, thereby improving first-frame accuracy.
- We apply a TCL to adjacent flow-warped frames, stabilizing frame-wise outputs and reducing variance over time.

2. Related Work

Steganography and Digital Watermarking. The rapid growth and easy distribution of digital media challenge both information confidentiality and intellectual property rights. Information hiding technologies evolved to address these needs, branching into steganography and digital watermarking, which differ in their primary objectives. Steganography [21, 39] prioritizes invisibility, covertly embedding information within digital content. In contrast, digital watermarking protects assets by embedding rights-identifying data, prioritizing robustness against various distortions and attacks. Early digital watermarking research focused on the spatial [41, 45] and frequency domains [1, 28]. Subsequently, deep learning-based methods [25, 53] jointly optimize invisibility and detection stability. These advancements have expanded watermarking applications from 2D images [8, 17, 26, 44] to videos [9, 16, 19, 49].

Robust Image Watermarking. Recent watermarking research has rapidly expanded, focusing on robustness against the diverse distortions and transformations in real-world distribution and editing environments. HiDDeN [53] proposed an approach that jointly learns watermarking and robustness by inserting a layer that simulates distortions such as Gaussian blur and JPEG. TrustMark [5] introduced a method robust to resolution changes by employing resolution scaling, while WAM [36] proposed one robust to partial edits through pixel-level detection and decoding. Following the advancement of diffusion-based generative models and reports that variational autoencoder (VAE) [22] and diffusion-based editing can easily remove watermarks [52], Robust-Wide [15] targeted instruction-driven image editing [4, 27, 48]. VINE [23] likewise addressed editing and proposed a benchmark integrating various distortions, including an I2V pipeline [2]. However, robustness to I2V remains a significant challenge for these methods.

Video Generation Models. With the recent advances in diffusion-based image generation models [6, 29, 34], video generation models have evolved by adopting U-Net [35] or the Diffusion Transformers (DiT) [31] as their backbones. The video diffusion model (VDM) [13] extended diffusion models to the video domain by modifying the 2D U-Net from image diffusion models into a spatial-temporal network. To reduce computational costs, Latent Video Diffusion Models (LVDMs) [3] compress videos into latent features. Stable Video Diffusion (SVD) [2] is a latent video diffusion model with a U-Net backbone, formalizing the video generation process by extending image diffusion temporally. In contrast, CogVideoX [47] leverages a DiT backbone with a 3D VAE and an expert transformer to enhance text-video alignment [20] and generate high-quality video. Despite these advancements, diffusion-based video generation models suffer from limitations in generation speed and memory efficiency due to their structural reliance on multi-step sampling. To address this, various methods using consistency models [38] have been proposed. Latent Consistency Model (LCM) [24] effectively improves memory usage and inference speed via consistency learning in the latent space, and AnimateLCM [42] extended this to the video domain, achieving fast sampling while maintaining strong temporal consistency and visual quality.

3. Preliminaries

Stable Video Diffusion (SVD) [2] performs image-to-video (I2V) generation in latent space as follows:

$$z_{\text{img}} = \text{enc}(I), \quad \mathbf{Z}^{(0)} \sim \mathcal{N}\big(\mathbf{0}, \sigma_{\max}^2 \mathbf{I}\big), \tag{1}$$

where $\text{enc}(\cdot)$ denotes the variational autoencoder (VAE) encoder, $z_{\text{img}} \in \mathbb{R}^{C_{\text{lat}} \times H_{\text{lat}} \times W_{\text{lat}}}$ is the encoded image latent representation, and $\mathbf{Z}^{(0)} \in \mathbb{R}^{T \times C_{\text{lat}} \times H_{\text{lat}} \times W_{\text{lat}}}$ is the initial latent noise. Here, T is the number of frames, C_{lat} is the number of latent channels, and H_{lat}, W_{lat} are the downsampled spatial sizes in latent space. Let c denote the conditioning signal derived from the input image and optional metadata (e.g., z_{img}, CLIP [33] image embedding, and frame rate). Using a finite-step sampler with a decreasing noise schedule $\sigma_0 = \sigma_{\max} > \sigma_1 > \cdots > \sigma_N \approx 0$, the denoiser D_θ iteratively updates the latent variables for $j = 1, \ldots, N$ as:

$$\begin{aligned} \mathbf{Z}^{(j)} = f_{\text{sampler}}\big(&\mathbf{Z}^{(j-1)}, \\ &D_\theta(\mathbf{Z}^{(j-1)}, \sigma_{j-1}, c), \sigma_{j-1}, \sigma_j\big), \end{aligned} \tag{2}$$

where f_{sampler} denotes the one-step update rule of the chosen sampler (e.g., Euler), which maps the current latent, the denoiser output at σ_{j-1}, and the pair (σ_{j-1}, σ_j) to the next latent. Finally, the latent variable at the last step $\mathbf{Z}^{(N)}$ is decoded to obtain the video as:

$$V = \{v_t\}_{t=0}^{T-1} = \text{dec}\big(\mathbf{Z}^{(N)}\big), \tag{3}$$

where $\text{dec}(\cdot)$ denotes the VAE decoder that projects the final latent variable $\mathbf{Z}^{(N)}$ back to pixel space.

AnimateLCM [42] adopts the same latent initialization and decoding as SVD in Eq. 1 and Eq. 3. The distinction lies in the denoising transition of Eq. 2, where the diffusion denoiser D_θ is replaced by a consistency-distilled denoiser C_ϕ, and the sampler in Eq. 2 uses only $K \ll N$ steps. Without classifier-free guidance (CFG) [11], each step is a single forward pass, yielding substantial speed and memory savings at high output quality. Style and acceleration are injected as lightweight residuals into a base weight η_0:

$$\eta' = \eta_0 + \alpha\, \tau_{\text{personalized}} + \beta\, \tau_{\text{acc}}^{\text{image}} + \gamma\, \tau_{\text{acc}}^{\text{video}}, \tag{4}$$

where $\tau_{\text{personalized}}$ is a LoRA residual [14], and $\tau_{\text{acc}}^{\text{image}}$ and $\tau_{\text{acc}}^{\text{video}}$ are acceleration residuals. The residuals are merged into η' without extra inference FLOPs.

4. Method

We aim to recover a k-bit watermark with high accuracy from frames synthesized using a single watermarked image. Let $I \in \mathbb{R}^{C \times H \times W}$ denote the original image and $w \in \{0,1\}^k$ the watermark message. The encoder E generates the watermarked image $I_w = E(I, w)$. An image-to-video (I2V) generator, such as Stable Video Diffusion (SVD) [2], conditions on I_w and produces frames $V = \{v_t\}_{t=0}^{T-1}$. The decoder D predicts the watermark from each frame v_t. As shown in Fig. 2, our methodology comprises three components: watermark encoding, a module for enhancing watermark robustness, and watermark decoding. We present them in that order for clarity and coherence.

4.1. Watermarking Encoder

The encoder uses a message-conditioned U-Net. We first interpret the message bits as a square bit grid with side length $m = \sqrt{k}$, where k is the message length. The bit vector is reshaped into a binary grid of size $1 \times m \times m$, then lifted to a learned feature map using a Convolutional Neural Network (CNN) and upsampling. This feature map is concatenated with the input image along the channel dimension and passed through a U-Net to produce watermark signal. To preserve perceptual quality, we modulate only the residual between the raw output and the original image. Specifically, we compute a data-dependent scale map from a just-noticeable difference (JND) [46] heatmap. This map has one channel and the same spatial resolution as the image, and it is bounded away from zero to prevent the watermark signal from disappearing. The final watermarked image is obtained by adding the element-wise product of the scale map and the residual back to the input image. This encourages the encoder to concentrate watermark signal in regions where changes are least visible while keeping the embedding aware of the image. We further enhance imperceptibility via adversarial training, as detailed in Sec. 4.4.

Semantic Preservation Loss. To keep high-level semantics stable, we regularize the CLIP [33] image embeddings of the watermarked image I_w and the original image I:

$$\mathcal{L}_{\text{sem}} = 1 - \cos(f_{\text{CLIP}}(I_w), f_{\text{CLIP}}(I)), \tag{5}$$

where f_{CLIP} denotes the pretrained, frozen CLIP image encoder [33] that maps an image I to a unit-norm embedding. As a term in the overall training objective, the semantic preservation loss keeps the watermarked image semantically consistent with the original and maintains the conditioning c used by a diffusion-based I2V model [2] during video generation. With c derived from the condition image, the loss further encourages c to be less affected by watermark signals when the video is generated from the watermarked image rather than from the original. Empirically, enforcing semantic preservation enables the watermark signal to survive the I2V process. In SVD, where the U-Net's attention uses the condition image's CLIP embedding as *keys* and *values*, the first frame v_0 remains close to I_w, yielding markedly higher bit accuracy of the first frame.

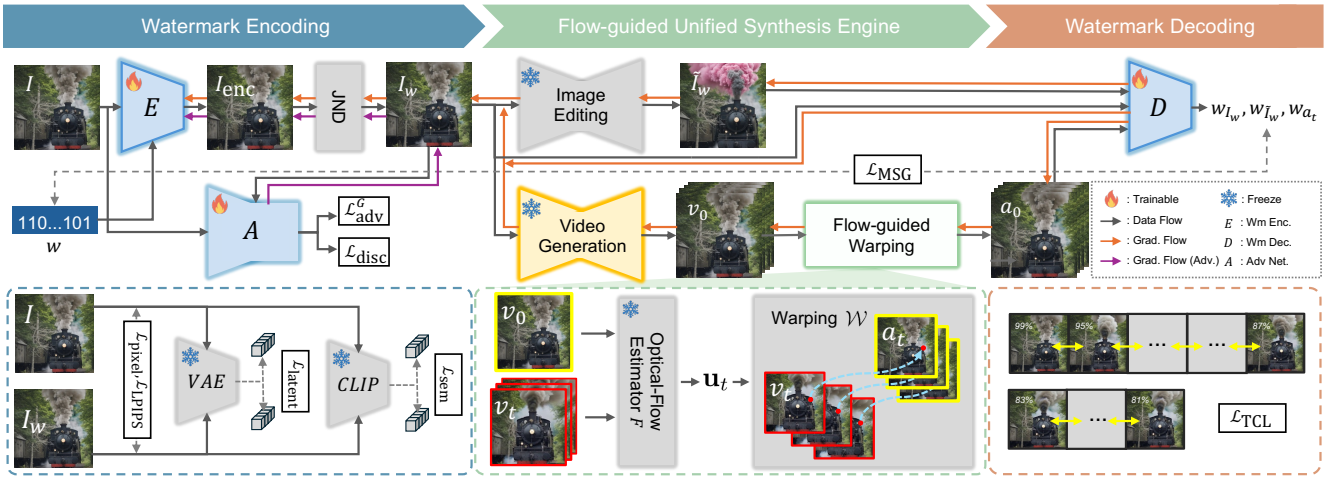


Figure 2. **The overview of WaTeRFlow.** Left: The watermark encoder is optimized to embed the watermark while preserving quality in both pixel and latent space, and it is trained to keep the watermarked image semantically close to the original. Middle: Image editing and video generation are performed by an image editing proxy and a video diffusion proxy, respectively, and the generated frames are then warped to the first frame. Right: The decoder processes the images produced by FUSE to decode the embedded watermark and compute the training loss. Overall, these components enable watermark insertion and detection that are robust to image-to-video generation.

4.2. FUSE: Flow-guided Unified Synthesis Engine

To train for robustness under both instruction-driven edits and I2V, we introduce a module named Flow-guided Unified Synthesis Engine (FUSE) between the watermark encoder and decoder. FUSE integrates an image editing branch and a video generation branch. During training, we optimize the encoder and decoder end-to-end with FUSE.

Image Editing Branch. Following [15], we employ InstructPix2Pix [4] as a proxy model for image editing. During training, this model edits the watermark-embedded image using text prompts, and the watermark decoder takes the edited image as input to recover the watermark.

Video Generation Branch. The video generation branch comprises a video diffusion proxy and an optical- flow estimator that warps the generated frames to a reference frame. We designate the first generated frame v_0 as the reference and warp each v_t toward v_0 with RAFT [40] to mitigate generation-induced distortions and stabilize learning:

$$a_t = \mathcal{W}(v_t, \mathbf{u}_t), \quad \mathbf{u}_t = F(v_0, v_t) \in \mathbb{R}^{H\times W\times 2}, \tag{6a}$$

$$[\mathcal{W}(v, \mathbf{u})](\mathbf{p}) = v(\mathbf{p} + \mathbf{u}(\mathbf{p})), \tag{6b}$$

where F is the optical-flow estimator and $\mathbf{u}_t$ is the forward optical-flow field from v_0 to v_t. a_t denotes the frame warped toward v_0 by the warping function $\mathcal{W}$. Here, $\mathcal{W}$ is a bilinear backward warper. The coordinate $\mathbf{p}$ lies on the reference grid of v_0, so $[\mathcal{W}(v_t, \mathbf{u}_t)](\mathbf{p})$ samples v_t at $\mathbf{p} + \mathbf{u}_t(\mathbf{p})$. During training, the decoder recovers the watermark from these generated and warped frames. For I2V generation, we adopt AnimateLCM [42] as the video diffusion proxy. Unlike models such as SVD [2], AnimateLCM attains high-quality videos in as few as two to four steps without classifier-free guidance (CFG) [11], eliminating extra unconditional and conditional passes and reducing memory usage. Because it avoids CFG and requires only a handful of diffusion steps, it keeps the additional compute minimal while preserving temporal coherence. We generate M key frames from the proxy, including the 0-th frame, and apply the above flow-guided warping before decoding.

4.3. Watermarking Decoder

The watermark decoder D is a stride downsampling residual CNN that maps an image to a k dimensional logit vector. It first encodes with strided convolutions until the spatial size matches the bit grid side $m = \sqrt{k}$, applies a stack of dilated residual blocks to aggregate mid and large contexts, and outputs a $1 \times m \times m$ map whose flattened responses form the logits. We recover the watermark k bits via:

$$w_q = \mathbf{1}\{\sigma(o_q) > \tfrac{1}{2}\} \in \{0,1\}^k, \quad o_q = D(q) \in \mathbb{R}^k, \tag{7}$$

where q denotes any input to the decoder D, $\sigma(\cdot)$ is the sigmoid function, and w_q is the recovered watermark.

4.4. Optimization Strategy

In our methodology we employ two main objective functions, one that optimizes the watermark encoder and decoder and another that optimizes the adversarial network.

Temporal Consistency Loss. I2V generation induces temporally varying subpixel drifts and misalignments that can persist even after flow-guided warping. These effects cause per-frame watermark predictions to fluctuate and degrade both average and late-frame bit accuracy. To stabilize decoding over time, we introduce a Temporal Consistency

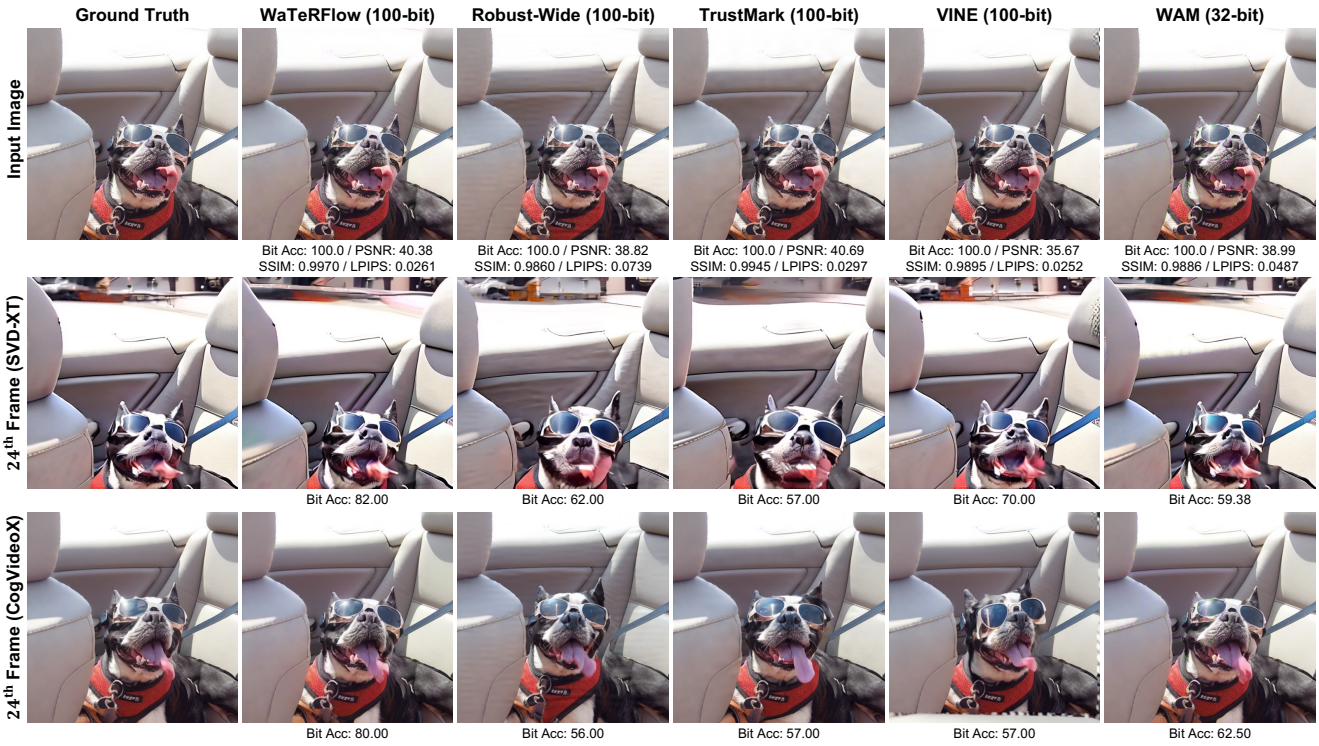


Figure 3. **Qualitative results.** Top: The original image and the watermarked images for each watermarking method. Middle: From left to right, the 24-th frames generated using SVD-XT are shown for the original image, our method, and the baselines. Bottom: Frames generated by CogVideoX. From left to right, we present the 24-th frames from the videos generated from the original image, then our method, followed by the baselines. Our method shows the highest bit accuracy for both video generation models in the given frames.

Loss (TCL) that encourages the decoder outputs for adjacent warped frames to be similar. Let a_ℓ be the ℓ-th frame warped toward the first frame v_0 using the optical-flow estimator described in Eq. 6a. The TCL is defined as:

$$\mathcal{L}_{\text{TCL}} = \frac{1}{M-1} \sum_{\ell=1}^{M-1} \left\| D(a_\ell) - D(a_{\ell-1}) \right\|_2^2, \tag{8}$$

where M is the number of key frames produced by FUSE, and D is the watermark decoder.

Encoder Loss. The watermark must be imperceptibly and stealthily embedded in the pixel space. Moreover, image editing and video generation operations are carried out in the latent space mapped by the VAE [22] encoder, so the watermark must remain stealthy there as well. Accordingly, we define the encoder loss as follows:

$$\begin{aligned} \mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{pixel}} + \lambda_{\text{latent}}\, \mathcal{L}_{\text{latent}} \\ + \lambda_{\text{LPIPS}}\, \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{sem}}\, \mathcal{L}_{\text{sem}}, \end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{pixel}}$ denotes the mean squared error (MSE) between the original and watermarked image. $\mathcal{L}_{\text{latent}}$ denotes the MSE between I and I_w computed in the latent space mapped by VAE encoder. $\mathcal{L}_{\text{LPIPS}}$ is the perceptual loss based on the LPIPS [50] evaluated between I and I_w. Finally, $\mathcal{L}_{\text{sem}}$ is the semantic preservation loss defined in Eq. 5.

Decoder Loss. The decoder is trained for per-bit message recovery on the watermarked image, the edited watermarked image, and flow-warped video frames. We adopt binary cross entropy (BCE) on per-bit probabilities:

$$\begin{aligned} \text{BCE}(x, y) = -\frac{1}{L} \sum_{i=1}^{L} \big[y_i \log x_i \\ + (1 - y_i) \log(1 - x_i) \big], \end{aligned} \tag{10}$$

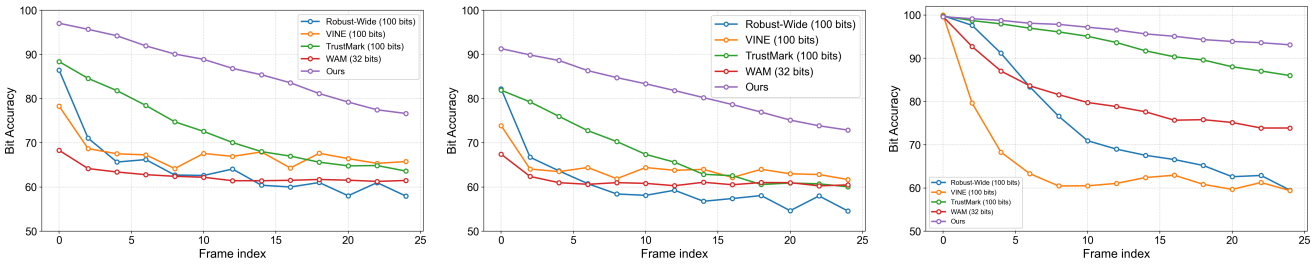
where $x \in \mathbb{R}^L$ is the predicted probability vector and $y \in \mathbb{R}^L$ is the binary target. The decoder loss is defined as:

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{TCL}} + \mathcal{L}_{\text{MSG}}, \tag{11a}$$

$$\begin{aligned} \mathcal{L}_{\text{MSG}} = \frac{1}{M} \sum_{\ell=0}^{M-1} \text{BCE}(p_{a_\ell}, w) \\ + \text{BCE}(p_{I_w}, w) + \text{BCE}(p_{\tilde{I}_w}, w), \end{aligned} \tag{11b}$$

where $\mathcal{L}_{\text{TCL}}$ is defined in Eq. 8 and p_{I_w}, $p_{\tilde{I}_w}$, and p_{a_ℓ} are the per-bit probabilities obtained by applying the sigmoid function to the decoder outputs for I_w, the edited image $\tilde{I}_w$ from FUSE's editing branch, and the ℓ-th flow-warped frame a_ℓ.

Adversarial Training. To enable the encoder to embed the watermark signal while preserving image quality, we adopt an adversarial strategy in which an adversarial network A



(a) I2V with SVD-XT (b) I2V with SVD-XT after image editing (c) I2V with CogVideoX

Figure 4. **Per-frame bit accuracy and I2V robustness.** Each plot visualizes bit accuracy on the even-numbered frames after image-to-video (I2V) generation. Across two representative I2V models, our method achieves the highest average bit accuracy compared to the baselines. It also shows the strongest robustness in the image-to-video (I2V) generation following image editing [51].

Method	Cap. (Bits)	Bit Acc.(%) ↑									
			Image-to-Video								
		None	None	Editing	H.264 CRF = 23	Regeneration $t_s = 150$	Contrast $\alpha = 1.2$	Brightness $s = 1.2$	JPEG $Q = 50$	Gaussian Noise $\sigma = 0.05$	Gaussian Blur $\sigma = 1.5$
Robust-Wide [15]	100	**99.99**	63.40	59.92	63.28	60.25	62.52	62.45	65.02	64.17	60.31
VINE [23]	100	**99.99**	73.92	66.56	73.75	71.83	79.72	78.08	75.15	79.40	77.66
TrustMark [5]	100	99.92	73.76	69.05	73.31	53.48	71.64	71.66	71.95	69.97	76.80
WAM [36]	32	99.90	63.87	61.55	63.27	60.72	62.64	62.80	62.34	61.12	63.00
WaTeRFlow (Ours)	100	99.86	**84.96**	**80.20**	**84.64**	**81.98**	**81.68**	**81.45**	**81.14**	**82.73**	**85.78**

Table 1. **Quantitative results.** The first None denotes the bit accuracy measured on the watermarked image with no preprocessing applied. All other columns report the *average bit accuracy over frames* generated by SVD-XT. The H.264 column reports average bit accuracy measured after H.264 compression to the generated videos. The entries from Editing [51] through Gaussian Blur indicate the preprocessing applied to the watermarked image before video generation. The best and second-best results are shown in **bold** and underlined, respectively.

is trained to distinguish original images from watermarked ones, and the encoder E learns to fool A, thereby encouraging quality aware watermark embeddings. Following prior work [7], we use a PatchGAN [18] adversarial network that outputs a patch-wise logit map, providing dense supervision and promoting imperceptible embeddings. The adversarial loss is defined as follows:

$$\mathcal{L}_{\text{adv}}^{G} = -\frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \log \sigma\big(A(I_w)_{i,j}\big), \tag{12}$$

where $A(I_w)_{i,j}$ denotes the (i,j)-th logit of the adversarial network A for the watermarked image, and H', W' are the adversarial network output map sizes. Meanwhile, the loss for the adversarial network A itself is defined as follows:

$$\mathcal{L}_{\text{disc}} = -\frac{1}{H'W'} \sum_{i,j} \Big[\log \sigma\big(A(I)_{i,j}\big) + \log\big(1 - \sigma\big(A(I_w)_{i,j}\big)\big) \Big]. \tag{13}$$

This objective encourages the adversarial network to assign high probabilities to patches from the original image and low probabilities to patches from the watermarked image.

Total Loss. The overall objective function for training our proposed watermark encoder and decoder is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{enc}} + \lambda_{\text{dec}} \mathcal{L}_{\text{dec}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}^{G}. \tag{14}$$

5. Experiments

This section presents experimental analysis, including details of the experimental setup, evaluation, comparisons with baseline methods, and ablation experiments.

5.1. Experimental Setting

Dataset. We train our models using the dataset proposed in InstructPix2Pix [4]. From this dataset, we use 20,000 original image–prompt pairs for training, and we use the paired textual instructions as editing prompts in the image editing branch of our framework. For evaluation, we use real-world images at a resolution of 512×512 drawn from UltraEdit [51], from which we randomly sample 500 to assess robustness in I2V generation for watermarked images. The graphs presented in Fig. 4 are produced using a randomly selected subset of 100 images from those 500. Meanwhile, the text prompts required for the CogVideoX experiments are generated with Qwen 2.5-VL-7B-Instruct [32]. Please refer to the Supplementary Material for further details.

Baseline. We use the officially released checkpoints and training code to standardize each model's capacity to 100-bit as much as possible. The only exception is WAM [36]. Its methodology explicitly states a 32-bit capacity limit, so we retain its original capacity. Additionally, because the baseline models were trained at different image resolutions, we follow the resolution-scaling procedure introduced in

VINE [23] to evaluate all models at a common resolution.

Evaluation Metrics. We evaluate image watermarking methods along two dimensions. *Robustness*: We measure robustness using bit accuracy, defined as the fraction of the embedded watermark that is correctly recovered by the watermark decoder from a watermarked image or from frames of a generated video. Beyond using watermarked images in the image-to-video (I2V) generation, we conduct robustness experiments that apply distortions immediately before or after video generation. The distortions include image editing [51], H.264 (CRF $= 23$), regeneration with a VP [12] scheduler ($t_s = 150$), contrast adjustment ($\alpha = 1.2$), brightness adjustment ($s = 1.2$), JPEG compression ($Q = 50$), Gaussian noise ($\sigma = 0.05$), and Gaussian blur ($\sigma = 1.5$). For example, when applying the regeneration, we first run the regeneration procedure on the watermarked image, then perform I2V generation, and measure bit accuracy. *Imperceptibility*: We assess image quality between the original and watermarked images using PSNR, SSIM [43], and LPIPS [50]. Using the same metrics, we compare frames generated from original images with those from watermarked images to assess video quality.

Implementation Details. We train on a single A100 GPU for 20,000 steps. In FUSE, all pretrained components remain frozen during training, as described in Sec. 4.2. FUSE generates four frames, including frame 0, with M denoting the count in Eq. 8. Starting at step 7,000, we activate FUSE's video generation branch, and in Eq. 11 the BCE terms and TCL are computed on flow-warped frames. We optimize Eq. 14 with $\lambda_{\text{latent}} = 10^{-3}$, $\lambda_{\text{LPIPS}} = 0.18$, $\lambda_{\text{sem}} = 10^{-3}$, $\lambda_{\text{dec}} = 1.3$, and $\lambda_{\text{adv}} = 0.004$. During training, FUSE uses InstructPix2Pix [4] and AnimateLCM [42] as proxy models. At evaluation, we generate frames with SVD-XT [2], a U-Net based model, and CogVideoX [47], a DiT based model, to demonstrate robustness across architectures. We then perform image edits with UltraEdit [51], treating it as one of the pre-I2V distortion scenarios. Specifically, we first apply UltraEdit to the watermarked image and then generate a video with SVD-XT [2].

5.2. Experimental Results

Image-to-Video Robustness. In this section, we evaluate the robustness of the watermark under I2V generation and compare our method against baselines. Using watermarked images as inputs, we generate videos with two models, SVD-XT [2] and CogVideoX [47]. As shown in Fig. 4a, our method achieves the highest bit accuracy across all evaluated frames, including the first generated frame. Likewise, as shown in Fig. 4c, our method attains the highest average bit accuracy over the evaluated frames for CogVideoX compared to the baselines. Overall, across both a U-Net based video generator [2] and a DiT based video generator [47], our method exhibits greater robustness than the baselines.

Method	Cap. (Bits)	Watermarked Image				Generated Video			
		PSNR ↑	SSIM ↑	LPIPS ↓	Bit Acc. (%)↑	PSNR ↑	SSIM ↑	LPIPS ↓	Bit Acc. (%)↑
Robust-Wide [15]	100	36.88	0.9698	0.0649	**99.99**	19.56	0.7228	0.2884	63.40
VINE [23]	100	33.88	0.9828	0.0334	**99.99**	17.86	0.6879	0.2930	73.92
TrustMark [5]	100	**40.33**	**0.9902**	0.0348	99.92	22.37	**0.7982**	0.1867	73.76
WAM [36]	32	37.91	0.9796	0.0337	99.90	20.51	0.7527	0.2232	63.87
WaTeRFlow (Ours)	100	38.83	**0.9902**	**0.0291**	99.86	**22.43**	0.7910	**0.1816**	**84.96**

Table 2. **Quality of watermarked images and generated videos.** The quality metrics for generated videos are evaluated against videos generated from images without embedded watermarks. In this setting, the videos are generated using SVD-XT.

FUSE		$\mathcal{L}_{\text{TCL}}$	$\mathcal{L}_{\text{sem}}$	Bit Acc.(%) ↑			Watermarked image		
I	V			First frame	Avg.	Avg. w/ Edit	PSNR ↑	SSIM ↑	LPIPS ↓
-	-	-	-	55.12	55.11	54.87	**52.65**	**0.9972**	**0.0002**
-	✓	✓	✓	93.26	81.65	77.47	36.28	0.9868	0.0422
✓	-	-	✓	91.89	66.32	63.56	39.25	0.9898	0.0262
✓	✓	-	✓	92.23	79.11	74.67	36.91	0.9860	0.0462
✓	✓	✓	-	89.63	76.82	72.79	40.44	0.9920	0.0299
✓	✓	✓	✓	**96.93**	**84.96**	**80.20**	38.83	0.9902	0.0291

Table 3. **Ablation of WaTeRFlow.** I and V denote the image-editing and video generation branches, and $\mathcal{L}_{\text{TCL}}$ and $\mathcal{L}_{\text{sem}}$ denote the loss terms in Sec. 4. Avg. indicates the average bit accuracy over frames generated by SVD-XT, while Avg. w/ Edit denotes the average bit accuracy after editing the watermarked image.

We conducted additional experiments and analyses on the relationship between the extent of motion in the generated videos and the bit accuracy in later frames. Please refer to the Supplementary Material for details and results.

More Challenging Scenarios. We evaluate robustness against distortions applied to the watermarked image before I2V and to the generated video after I2V. We consider a more demanding setting where the watermarked image is first modified by an image-editing model [51] and the edited image is then used as the input to the video generator. In addition, we include common degradations such as JPEG compression prior to video generation. We also evaluate robustness when videos generated from watermarked images are encoded with H.264 compression. As seen in Fig. 4b and Tab. 1, our method is more robust than the baselines for I2V under pre-I2V distortions including image editing, regeneration, contrast and brightness changes, JPEG compression, Gaussian noise, and Gaussian blur. It also remains robust under H.264 compression applied after I2V.

Imperceptibility Evaluation. We evaluate the quality of watermarked images and the frames generated from watermarked images. Even when considering the trade-offs among robustness, imperceptibility, and capacity, Fig. 3 and Tab. 2 show that our method yields competitive quality. In particular, as shown in Fig. 3, our method maintains motion close to the original while achieving higher bit accuracy.

5.3. Ablation Study

FUSE, TCL, and Semantic Preservation Loss. In this section, we evaluate the impact of FUSE, TCL, and the Semantic Preservation loss on robustness under I2V generation. With all other settings fixed, we ablate one component at a time during training. Using the resulting encoders,

we then generate SVD-XT videos from watermarked images and measure per-frame bit accuracy. The results are summarized in Tab. 3, and the average accuracy is highest when all three components are enabled. FUSE exposes the encoder and decoder to realistic distortions during training. TCL stabilizes frame-wise outputs after warping. The semantic preservation loss maintains the semantic information of the watermarked image and improves the accuracy of frame 0. Furthermore, Tab. 3 shows that removing FUSE's image editing branch leads to a decrease in bit accuracy for frames generated after editing the watermarked image. For further details, please refer to the Supplementary Material.

Video Diffusion Proxy Selection. We analyze how the choice of video diffusion proxy for FUSE's video generation branch affects performance and training cost. Tab. 4 summarizes results when we use SVD, SVD without CFG, and AnimateLCM in the video generation branch. AnimateLCM yields faster video generation and requires less peak VRAM during training than SVD without CFG. Considering the overall performance of SVD-based training, using SVD effectively requires enabling CFG. Taking this into account, AnimateLCM achieves competitive performance with lower peak VRAM and shorter training time.

Method	Bit Acc. (%) ↑	PSNR ↑	SSIM ↑	LPIPS ↓	VRAM (GB)	Train (h)	Gen. (s)
SVD [2] (w/o CFG)	72.93	38.55	0.9883	0.0298	67.98	23.09	2.6
SVD [2] (w/ CFG)	83.83	**38.99**	**0.9909**	**0.0254**	93.80	42.39	7.9
AnimateLCM [42] (Ours)	**84.96**	38.83	0.9902	0.0291	44.51	17.70	1.1

Table 4. **Ablation on video diffusion proxy choice.** We compare two SVD variants, one without CFG and one with CFG, together with AnimateLCM. Bit Acc. is measured on frames generated by SVD-XT, VRAM denotes peak memory usage during training, Train is total training time, and Gen. is per-video generation. Here, PSNR, SSIM, and LPIPS represent watermarked image quality.

Optical Flow and Training Stability. We analyze how the presence or absence of an optical-flow estimator in FUSE's video generation branch affects training. Fig. 5 visualizes the BCE loss measured on frames generated during training, which serves as an indicator of the decoder's recovery performance. The graph suggests that without an optical- flow estimator, the decoder does not adequately adapt to distortions introduced by image-to-video generation.

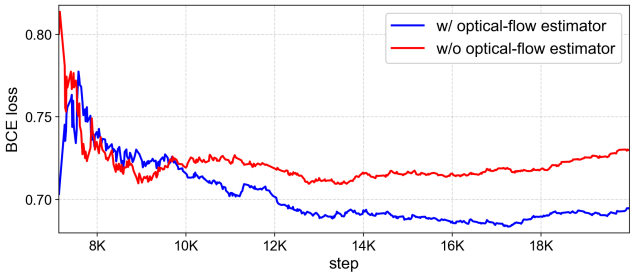


Figure 5. **Training stability.** Computed on frames from FUSE's video generation branch, comparing with and without an optical-flow estimator. The estimator yields a more stable trajectory, indicating improved robustness to image-to-video generation.

Semantic preservation with DINOv2. We replace the CLIP-based semantic preservation loss with a DINOv2 [30] based variant. It is defined in a representation space different from the CLIP conditioning encoder used by SVD. Despite this difference, constraining watermark insertion to preserve structural cues such as object layout, contours, and low frequency signals improves first-frame bit accuracy after image-to-video synthesis. The pixel-level and latent-level objectives in Eq. 9 already shrink the overall residual, and the DINOv2-based term then guides where the remaining change resides by reducing structure sensitive changes and keeping the remaining change in local details with smaller impact. Considering the first-frame bit accuracy of the w/o $\mathcal{L}_{\text{sem}}$ setting in Tab. 3, Tab. 5 shows that using DINOv2 also yields higher first-frame bit accuracy.

Method	Bit Acc. (%) ↑		Watermarked Image		
	First frame	Avg.	PSNR ↑	SSIM ↑	LPIPS ↓
DINOv2 [30]	95.84	83.49	38.27	0.9873	0.0406
CLIP [33]	96.93	84.96	38.83	0.9902	0.0291

Table 5. **Ablation on semantic preservation loss.** Even when the CLIP is replaced with DINOv2, it improves bit accuracy on the first frame after image-to-video synthesis. Avg. denotes the average bit accuracy across frames generated by SVD-XT.

6. Conclusion

In this work, we study an image-to-video provenance scenario that current image watermarking methods rarely consider. A single watermarked image can be turned into a video and the watermark must be recovered from generated frames. We introduce WaTeRFlow, which trains encoder and decoder under realistic image-to-video generation while preserving watermarked image quality. WaTeRFlow relies on three key components. The Flow-guided Unified Synthesis Engine exposes the system to instruction-driven image editing and fast video diffusion during learning. A Temporal Consistency Loss stabilizes frame-level decoding and a semantic preservation loss keeps the conditioning signal close to that of the original image. WaTeRFlow improves bit accuracy for the first frame and subsequent frames while minimally affecting the visual quality of watermarked images, making it a practical tool for provenance and ownership verification.

Limitations. We evaluate robustness on Stable Video Diffusion and CogVideoX. However, if the video generator substantially alters the initial frame or the semantic content of the conditioning signal, bit accuracy may drop. Robustness can also degrade for generators whose synthesis principles differ markedly from the video proxy models used during training. Although WaTeRFlow improves robustness in image-to-video scenarios, further discussion is needed regarding adversarial removal of the embedded watermark.

Acknowledgments

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025, 50%; Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469, 25%), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00521602, 23%), Institute of Information & communications Technology Planning & Evaluation (IITP) & ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT) (No.RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University), 1%; IITP-2025-RS-2024-00436857, 1%), and the NVIDIA Academic Grant Program using NVIDIA RTX PRO 6000 Max-Q Workstation Edition.

References

[1] Mauro Barni, Franco Bartolini, and Alessandro Piva. Improved wavelet-based watermarking through pixel-wise masking. *IEEE transactions on image processing*, 10(5): 783–791, 2001. 2

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3, 4, 7, 8

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. 1, 2, 4, 6, 7

[5] Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Robust watermarking and watermark removal for arbitrary resolution images. In *IEEE International Conference on Computer Vision (ICCV)*, 2025. 2, 6, 7

[6] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2021. Curran Associates Inc. 2

[7] Han Fang, Zhaoyang Jia, Hang Zhou, Zehua Ma, and Weiming Zhang. Encoded feature enhancement in watermarking network for distortion in real scenes. *IEEE Transactions on Multimedia*, 25:2648–2660, 2023. 6

[8] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023. 2

[9] Pierre Fernandez, Hady Elsahar, I Zeki Yalniz, and Alexandre Mourachko. Video seal: Open and efficient video watermarking. *arXiv preprint arXiv:2412.09492*, 2024. 2

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 4

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 7

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 2

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3

[15] Runyi Hu, Jie Zhang, Ting Xu, Jiwei Li, and Tianwei Zhang. Robust-wide: Robust watermarking against instruction-driven image editing. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. 2, 4, 6, 7

[16] Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Videoshield: Regulating diffusion-based video generation models via watermarking. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[17] Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37:3937–3963, 2024. 2

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[19] MinHyuk Jang, Youngdong Jang, JaeHyeok Lee, Feng Yang, Gyeongrok Oh, Jongheon Jeong, and Sangpil Kim. Lvmark: Robust watermark for latent video diffusion models. *arXiv preprint arXiv:2412.09122*, 2024. 2

[20] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 2

[21] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network.

2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 4713–4722, 2021. 2

[22] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2, 5

[23] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 6, 7

[24] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[25] Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. Distortion agnostic deep watermarking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13548–13557, 2020. 2

[26] Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 1532–1542. ACM, 2022. 2

[27] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text Inversion for Editing Real Images using Guided Diffusion Models . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2

[28] KA Navas, Mathews Cheriyan Ajay, M Lekshmi, Tampy S Archana, and M Sasikumar. Dwt-dct-svd based watermarking. In *2008 3rd international conference on communication systems software and middleware and workshops (COMSWARE'08)*, pages 271–274. IEEE, 2008. 2

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 2

[32] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 6

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 8

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[36] Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations (ICLR)*, 2025. 2, 6, 7

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 1

[38] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 2

[39] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2117–2126, 2020. 2

[40] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2, 4

[41] Ron G Van Schyndel, Andrew Z Tirkel, and Charles F Osborne. A digital watermark. In *Proceedings of 1st international conference on image processing*, pages 86–90. IEEE, 1994. 2

[42] Fu-Yun Wang, Zhaoyang Huang, Weikang Bian, Xiaoyu Shi, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. Animatelcm: Computation-efficient personalized style video generation without personalized video data. In *SIGGRAPH Asia 2024 Technical Communications*, pages 1–5, 2024. 2, 3, 4, 7, 8

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[44] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems*, pages 58047–58063. Curran Associates, Inc., 2023. 2

[45] Raymond B Wolfgang and Edward J Delp. A watermark for digital images. In *Proceedings of 3rd IEEE International Conference on Image Processing*, pages 219–222. IEEE, 1996. 2

[46] Jinjian Wu, Leida Li, Weisheng Dong, Guangming Shi, Weisi Lin, and C.-C. Jay Kuo. Enhanced just noticeable difference model for images with pattern complexity. *IEEE Transactions on Image Processing*, 26(6):2682–2693, 2017. 3

[47] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Representation Learning*, pages 83048–83077, 2025. 2, 7

[48] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: a manually annotated dataset for instruction-guided image editing. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 2

[49] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019. 2

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 7

[51] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In *Advances in Neural Information Processing Systems*, pages 3058–3093. Curran Associates, Inc., 2024. 1, 6, 7

[52] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative AI. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[53] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018. 2