Safe-Sora: Safe Text-to-Video Generation via Graphical Watermarking

Zihan Su[1] **Xuerui Qiu**[2,3] **Hongbin Xu**[4] **Tangyu Jiang**[1] **Junhao Zhuang**[1]
Chun Yuan[1*] **Ming Li**[5*] **Shengfeng He**[6] **Fei Richard Yu**[5]
[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Institute of Automation, Chinese Academy of Sciences
[3] Zhongguancun Academy [4] Bytedance
[5] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[6] Singapore Management University
zh-su24@mails.tsinghua.edu.cn

Abstract

The explosive growth of generative video models has amplified the demand for reliable copyright preservation of AI-generated content. Despite its popularity in image synthesis, invisible generative watermarking remains largely underexplored in video generation. To address this gap, we propose *Safe-Sora*, the first framework to embed graphical watermarks directly into the video generation process. Motivated by the observation that watermarking performance is closely tied to the visual similarity between the watermark and cover content, we introduce a hierarchical *coarse-to-fine adaptive matching mechanism*. Specifically, the watermark image is divided into patches, each assigned to the most visually similar video frame, and further localized to the optimal spatial region for seamless embedding. To enable spatiotemporal fusion of watermark patches across video frames, we develop a 3D wavelet transform-enhanced Mamba architecture with a novel *spatiotemporal local scanning strategy*, effectively modeling long-range dependencies during watermark embedding and retrieval. To the best of our knowledge, this is the first attempt to apply state space models to watermarking, opening new avenues for efficient and robust watermark protection. Extensive experiments demonstrate that *Safe-Sora* achieves state-of-the-art performance in terms of video quality, watermark fidelity, and robustness, which is largely attributed to our proposals. Code is publicly available at https://github.com/Sugewud/Safe-Sora

1 Introduction

Recent advances in video generation models have significantly transformed digital content creation [1–6]. VideoCrafter2 [2] delivers high-fidelity video generation results, while Open-Sora [7] enables efficient and scalable video generation. However, this rapid progress also raises growing concerns over copyright protection and ownership verification of generated videos.

Invisible watermarking has proven effective for copyright protection in image generation [8–15]. However, its extension to video generation remains relatively underexplored. Recent efforts such as VideoShield [16] and LVMark [17] embed watermarks by modifying latent noise or applying importance-based modulation strategies. Despite these advancements, existing approaches rely on embedding bitstring-based identifiers, which fall short of leveraging the high information capacity inherent in video content. Unlike static images, videos offer significantly greater embedding bandwidth, making them well-suited for graphical watermarks—e.g., logos or icons—that serve as more

*Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

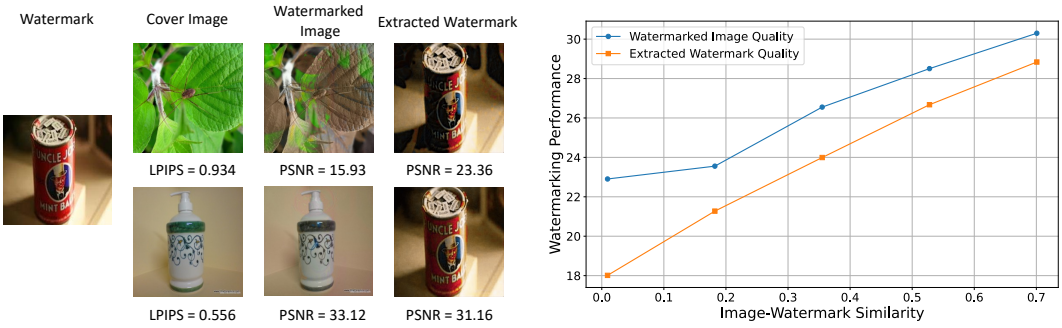

Figure 1: Impact of image-watermark similarity on watermarking performance. We used a pretrained classic image hiding network Balujanet [18] on 1,000 image pairs, each consisting of a graphical watermark from Logo-2k [19] and a cover image from ImageNet [20]. Image-Watermark similarity was quantified using 1-LPIPS and the quality of the watermarked image and extracted watermark was evaluated using PSNR. Higher PSNR and lower LPIPS indicate improved performance.

intuitive and visually recognizable evidence of ownership. Such designs enhance both the perceptual clarity and practical reliability of copyright verification.

Recognizing the untapped potential of graphical watermarking in video generation, we propose Safe-Sora, the first framework, to the best of our knowledge, that embeds graphical watermarks elegantly into the video generation process. As illustrated in Fig. 1, we observe that watermarking performance significantly correlates with the visual similarity between the watermark and cover images. In particular, embedding becomes significantly more effective when the cover image shares high visual similarity with the watermark content. Motivated by this, we propose a hierarchical coarse-to-fine adaptive matching mechanism, which first divides the watermark image into patches and assigns each patch to the most similar video frame through an *inter-frame* automatic selection strategy. Subsequently, an *intra-frame* localization is performed to embed the patch into the most visually similar region within the selected frame. To address the challenge of fusing and extracting watermark information distributed across spatiotemporal locations, we further propose a 3D wavelet transform-enhanced Mamba architecture with a tailored scanning strategy. This design enables bidirectional modeling across frequency subbands in the 3D wavelet transform, effectively and efficiently capturing long-range dependencies in both space and time. To the best of our knowledge, this is the first application of state space models to generative watermarking.

In our experiments, we utilize the widely-used Panda-70M [21] dataset as the video source due to its extensive scale and diverse video categories. For graphical watermarks, we employ the Logo-2K+ [19] dataset, which offers a wide variety of real-world logos. The quantitative and qualitative comparisons with existing methods demonstrate that the proposed Safe-Sora achieves state-of-the-art performance in terms of video quality, watermark fidelity, and robustness. For instance, our method achieves a Fréchet Video Distance of 3.77, far lower than the second-best baseline's 154.35, highlighting its superior temporal consistency. Our primary contributions can be summarized as follows:

- We introduce the first model specifically designed to embed graphical watermarks in video generation pipelines, directly addressing the pressing need for copyright protection of generated video content.
- We propose a hierarchical coarse-to-fine adaptive matching mechanism that strategically embeds watermark patches into visually similar frames and spatial regions, enhancing overall watermarking performance.
- We pioneer the application of state space models for watermarking through a novel 3D wavelet transform-enhanced Mamba architecture with a tailored scanning strategy, enabling enhanced fusion and extraction of watermark information across space and time.

2 Related Work

2.1 Video Diffusion Models

Recently, AI-generated content has been vibrant in the community [22–33]. Diffusion models [34–39] are a class of generative models that synthesize data through a gradual denoising process, beginning

from randomly sampled Gaussian noise. Latent Video Diffusion Models (LVDMs) [40] perform the diffusion process in the latent space to improve computational efficiency. VideoCrafter2 [2] builds high-quality video generation models by leveraging low-quality video data combined with synthesized high-quality images. Open-Sora [7] introduces the Spatial-Temporal Diffusion Transformer, an efficient video diffusion framework that separates spatial and temporal attention mechanisms. While LVDMs have shown strong performance in video generation, the integration of graphical watermarks into this framework has not been explored.

2.2 Generative Video Watermarking

Digital watermarking has emerged as an essential technique for copyright protection, content authentication, and ownership verification across various media types. However, watermarking for video diffusion models represents a relatively unexplored area. VideoShield [16] pioneered this space by modifying latent noise during the diffusion process to embed binary watermark information. More recently, LVMark [17] introduced an importance-based weight modulation strategy to minimize visual quality degradation. Nevertheless, these existing approaches primarily focus on embedding low-capacity binary strings, without taking advantage of the high-capacity nature of video media, which is well-suited for embedding richer information such as graphical watermarks.

2.3 State Space Models

State Space Models (SSMs) [41, 42] have emerged as efficient alternatives to transformers [43] for sequence modeling. The Mamba architecture [44] represents a significant advancement in SSMs by introducing selective state space modeling with data-dependent parameters, enabling dynamic resource allocation to important sequence elements while maintaining computational efficiency. Despite Mamba's remarkable success in language processing tasks [45, 46] and its growing adoption in computer vision applications [47, 48], its potential for watermarking techniques has remained entirely unexplored until now.

3 Graphical Watermarking for Video Generation

In this section, we present the pipeline of our Safe-Sora framework, which introduces a novel approach to embedding graphical watermarks directly within the video generation process (Fig. 2). We first partition the watermark image into patches and optimally assign them to appropriate video frames and regions (Section 3.1). These patches are then embedded and upsampled to generate the watermark feature map. To embed the watermark, this feature map is fused with multi-scale video features using a UNet built with 2D SFMamba blocks (Section 3.2), followed by a series of 3D SFMamba blocks that leverage our spatiotemporal local scanning strategy (Section 3.3), producing a watermarked video. To extract the watermark, the watermarked video is processed through an extraction network built with a degradation layer, a series of 3D SFMamba blocks, and position recovery. The training objectives are outlined in Section 3.4, while the preliminaries on latent video diffusion models, state space models, and wavelet transforms are detailed in Appendix A.

3.1 Coarse-to-Fine Adaptive Patch Matching

Motivated by the observation that greater similarity between the watermark and cover content enhances watermarking performance (as shown in Fig. 1), we propose a *coarse-to-fine adaptive patch matching* mechanism to systematically identify the most semantically similar spatial-temporal regions in a video for watermark embedding, as illustrated in the bottom-left corner of Fig. 2.

First, to enable accurate localization of each patch during the final watermark recovery, we propose a simple yet effective method: the position channel. Specifically, we represent patch positions using binary encoding (e.g., using 8 bits to represent 256 patch positions). This binary code is then replicated to form an additional channel, introducing redundancy that enhances robustness against spatial distortions and degradation. Finally, this position channel is concatenated with the patch content, embedding positional information directly into the input and eliminating the need for additional positional processing during subsequent training.

Then, we adopt a two-stage process to adaptively determine the most suitable embedding location for each patch. The first stage operates at the frame level. We extract features from both patches and

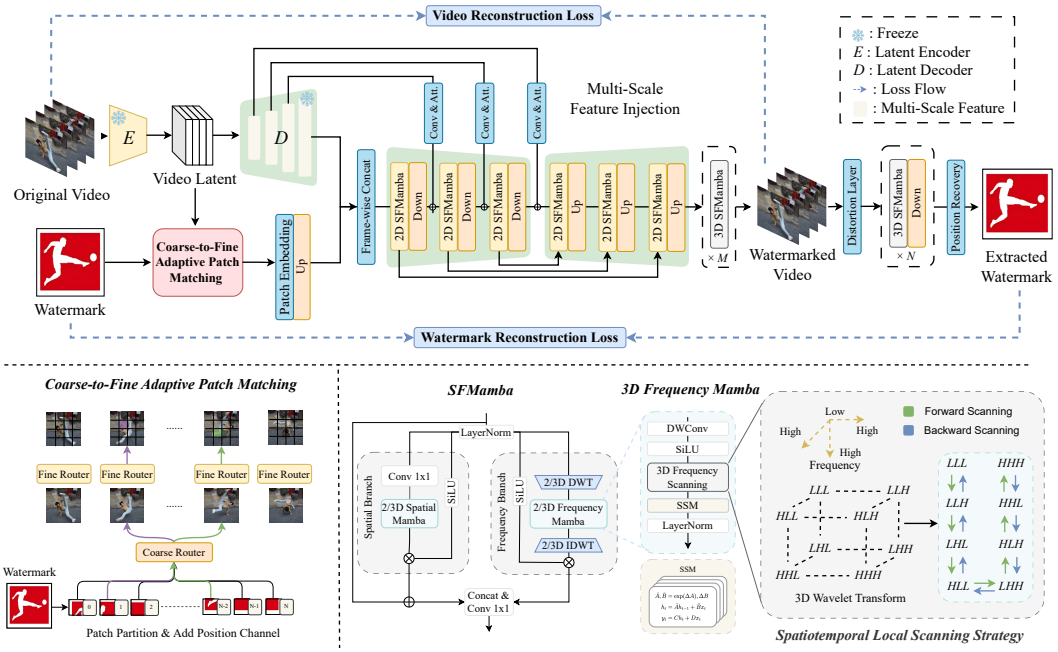

Figure 2: Overview of our Safe-Sora framework. Our method consists of three main components: (1) Coarse-to-Fine Adaptive Patch Matching: partitioning the watermark image into patches and optimally assigning them to appropriate video frames and regions, followed by patch embedding and upsampling to generate the watermark feature map; (2) Watermark Embedding: the watermark feature map is fused with multi-scale video features via a UNet with 2D SFMamba blocks, followed by a series of 3D SFMamba blocks that implement our spatiotemporal local scanning strategy, to produce the watermarked video; (3) Watermark Extraction: recovering the embedded watermark using an extraction network built with a distortion layer, a series of 3D SFMamba blocks, and position recovery. The difference between different types of Mamba blocks lies in their scanning strategies.

the latent representations of video frames using a convolution layer followed by ReLU and global average pooling (GAP). Similarity between each patch i and frame j is computed via dot product of these feature vectors, and normalized using Softmax:

$$\mathbf{w}_{i,j} = \text{Softmax}\left(\text{GAP}(\text{ReLU}(\text{Conv}(\mathbf{p}_i))) \cdot \text{GAP}(\text{ReLU}(\text{Conv}(\mathbf{z}_j)))\right). \tag{1}$$

Here, $\mathbf{w}_{i,j}$ denotes the similarity score between patch $\mathbf{p}_i$ and the latent representation $\mathbf{z}_j$ of frame j. Each patch is then assigned to the frame with the highest similarity score. To ensure balanced distribution, we impose a maximum capacity for each frame. If the top-ranked frame is full, the patch is redirected to the next highest available candidate. Having selected a frame, we proceed to the fine stage, which determines the optimal spatial position within that frame. Each frame is subdivided into spatial regions according to its patch capacity. Feature representations of these regions are computed similarly, and the similarity between patch i and region k in the assigned frame j is given by:

$$\mathbf{s}_{i,k} = \text{Softmax}\left(\text{GAP}(\text{ReLU}(\text{Conv}(\mathbf{p}_i))) \cdot \text{GAP}(\text{ReLU}(\text{Conv}(\mathbf{r}_{j,k})))\right), \tag{2}$$

where $\mathbf{s}_{i,k}$ is the similarity score between the i-th patch and the k-th region $\mathbf{r}_{j,k}$ in the latent representation of frame j. Note that we take full advantage of the inherent feature properties of latent variables in video generation models. Since latent variables can already be viewed as feature extractions of the original frames, we use only a single convolutional layer for feature extraction, which significantly reduces the computational overhead.

3.2 Spatial-Frequency Mamba for Spatial Fusion

Mamba [44] has demonstrated strong capabilities in modeling long-range dependencies with high efficiency, making it well-suited for spatiotemporal modeling in video tasks. Meanwhile, frequency domain information has been applied in various domains [49–51]. In watermark embedding, it has proven effective in capturing structural patterns and resisting distortions [50, 51]. To incorporate both advantages, we propose the Spatial-Frequency Mamba (SFMamba) block, as shown in Fig. 2.

SFMamba adopts a dual-stream design with separate spatial and frequency branches. It comes in two variants: a 2D version and a 3D version, differing primarily in the wavelet transform and scanning

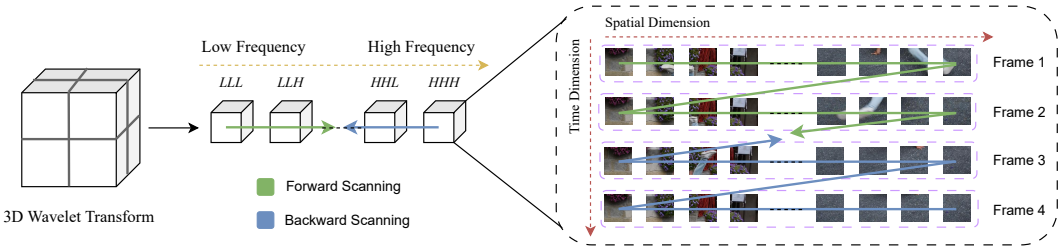

Figure 3: For 3D frequency scanning, we propose a spatiotemporal local scanning strategy for 3D wavelet transform, which processes the frequency components hierarchically from low frequency to high frequency and high frequency to low frequency.

strategy. The 3D SFMamba will be introduced in Section 3.3. We next introduce the 2D SFMamba block for efficient spatial fusion of watermark and video content. It consists of separate 2D spatial and 2D frequency branches.

2D Spatial Branch. The spatial processing begins with a LayerNorm operation on the input feature map $\mathbf{F}_{\text{in}}$, yielding normalized features $\mathbf{F}_{\text{N}}$. In the first path, $\mathbf{F}_{\text{N}}$ undergoes a simple SiLU activation function. In the second path, $\mathbf{F}_{\text{N}}$ passes through a 1×1 convolution layer, followed by our 2D spatial Mamba module. The 2D spatial branch output $\mathbf{F}_s$ is computed as:

$$\mathbf{F}_s = \text{SiLU}(\mathbf{F}_{\text{N}}) \odot \text{2DSpatialMamba}(\text{Conv}_{1\times1}(\mathbf{F}_{\text{N}})). \tag{3}$$

where $\odot$ denotes element-wise multiplication of the two pathway outputs.

2D Frequency Branch. For frequency domain processing, we transform $\mathbf{F}_{\text{N}}$ using a 2D Discrete Wavelet Transform (DWT), which decomposes the signal into four frequency subbands: *LL* (low-low), *LH* (low-high), *HL* (high-low), and *HH* (high-high). Each subband has spatial dimensions reduced by half compared to the original. Inspired by FreqMamba [52], we rearrange these components from top-left to bottom-right to restore the original resolution. The wavelet features are then divided into four blocks and scanned block by block. The output is projected back to the spatial domain via a 2D Inverse DWT (IDWT), followed by element-wise multiplication with $\text{SiLU}(\mathbf{F}_{\text{N}})$. The 2D frequency branch output $\mathbf{F}_f$ is computed as:

$$\mathbf{F}_f = \text{SiLU}(\mathbf{F}_{\text{N}}) \odot \text{IDWT}(\text{2DFreqMamba}(\text{DWT}(\mathbf{F}_{\text{N}}))). \tag{4}$$

The spatial branch output is enhanced with a residual connection from $\mathbf{F}_{\text{in}}$. Finally, we concatenate the outputs from both branches and apply a 1×1 convolution to produce the integrated output.

3.3 3D Frequency Scanning for Spatiotemporal Interaction

To address the challenges of fusing and extracting watermark information distributed across spatiotemporal locations, we propose an efficient architecture—3D SFMamba, a 3D Wavelet Mamba transform-enhanced design with a customized scanning strategy. This architecture enables bidirectional modeling across frequency subbands within the 3D wavelet transform, effectively capturing long-range dependencies in both spatial and temporal domains to accurately recover watermark information embedded in the temporal dimension. 3D SFMamba consists of separate 3D spatial and 3D frequency branches.

3D Spatial Branch. The 3D spatial branch employs a vanilla 3D scanning strategy, which processes features across all three dimensions (temporal, height, width) to capture both spatial and temporal dependencies effectively.

3D Frequency Branch. In the frequency domain branch, input features $\mathbf{F}_{\text{in}}$ undergo a 3D Discrete Wavelet Transform (3D DWT), decomposing them into eight subbands: *LLL*, *LLH*, *LHL*, *LHH*, *HLL*, *HLH*, *HHL*, and *HHH*. Each subband has half the original dimensions in frame, height, and width. To address the complexity of 3D wavelet-transformed features, we propose a novel spatiotemporal local scanning strategy as shown in Fig. 3. This approach first rearranges the eight subbands to restore the original video resolution, then divides them into eight distinct parts for separate scanning. For forward scanning, the order follows *LLL*, *LLH*, *LHL*, *HLL*, *LHH*, *HLH*, *HHL*, and *HHH*—progressing systematically from low to high frequencies. Additionally, we implement a reverse scanning mechanism that processes the subbands in the opposite direction—from *HHH* to *LLL*—enabling the model to capture information from high to low frequencies. Within each part, we employ a spatial-first, temporal-second scanning pattern. This spatiotemporal local scanning

strategy is specifically designed for 3D wavelet transforms, allowing the model to process frequency information hierarchically across multiple scales.

3.4 Training Objectives

Our training framework combines video reconstruction loss and watermark reconstruction loss. The video reconstruction loss uses mean squared error (MSE) to ensure the watermarked video $\hat{\mathbf{V}}$ closely resembles the original video $\mathbf{V}$:

$$\mathcal{L}_{\text{video}} = \text{MSE}(\mathbf{V}, \hat{\mathbf{V}}). \tag{5}$$

Similarly, the watermark reconstruction loss measures the extraction accuracy by comparing the extracted watermark $\hat{\mathbf{W}}$ with the original watermark $\mathbf{W}$:

$$\mathcal{L}_{\text{watermark}} = \text{MSE}(\mathbf{W}, \hat{\mathbf{W}}). \tag{6}$$

During training, we provide the correct positions to reconstruct the watermark image properly, while during testing, the model utilizes the embedded position channels to predict the correct arrangement of patches. The final loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{video}} + \lambda \, \mathcal{L}_{\text{watermark}}, \tag{7}$$

where the watermark weighting hyperparameter λ balances video quality against watermark fidelity.

4 Experiments

4.1 Experimental Setting

Datasets. For the video dataset, we use the Panda-70M [21] dataset for training, which is a large-scale dataset containing 70 million high-quality videos across diverse content types. Specifically, we randomly download 10,000 videos from Panda-70M, sample 8 frames from each video, and resize each frame to a resolution of 320×512 for training purposes. For the watermark dataset, we use the Logo-2K dataset [19], which contains 167,140 watermark images at a resolution of 256 × 256, spanning a wide range of real-world logo classes. For the evaluation of text-to-video generation, we employ the VidProm [53] dataset as the source of prompts. The prompts in VidProm are generated by GPT-4 [54], and we randomly select 100 prompts from the dataset for evaluation.

Implementation Details. We use VideoCrafter2 [2] as our backbone model to generate videos at a resolution of 320×512. Our method is compatible with various video generation backbones, with additional results provided in Appendix C. The patch size is set to 16×16. Patch Embedding maps each patch to a 1024-dimensional feature space. The model is trained for 30 epochs on 4 NVIDIA RTX 4090 GPUs. We adopt the AdamW optimizer [55], with the initial learning rate set to 5e-4, which is gradually decayed to 1e-6 following a cosine decay schedule. The watermark embedding network uses $M = 2$ 3D SFMamba Blocks, while the watermark extraction network uses $N = 4$ 3D SFMamba Blocks. The hyperparameter λ in Eq. 7 is set to 0.75. The distortion layer simulates various real-world distortions, including H.264 video compression, rotation, and other common transformations. Since H.264 is non-differentiable, we follow DVMark [56] and use a 3D CNN to mimic its effects. For position recovery, we propose a confidence-guided greedy assignment algorithm, with detailed descriptions provided in Appendix B.

Baselines. To the best of our knowledge, no existing method embeds graphical watermarks directly into video generation models. To provide a comprehensive comparison, we select five representative state-of-the-art methods spanning three distinct paradigms of graphical watermarking: (1) Post-processed image watermarking methods: **Balujanet**[18] – A classic image steganography network; **UDH**[57] – A classic graphical watermarking network; **PUSNet** [58] – A state-of-the-art image steganography network. (2) Generative image watermarking: **Safe-SD** [59] – A generative graphical watermarking approach. (3) Video steganography: **Wengnet** [60] – A method that hides one video within another. For a fair comparison, we retrain all baseline methods using the same training dataset as ours. For image-based methods, we embed a complete watermark image into each frame. For video-based methods, each frame of the secret video acts as a watermark and is embedded into the corresponding frame of the cover video.

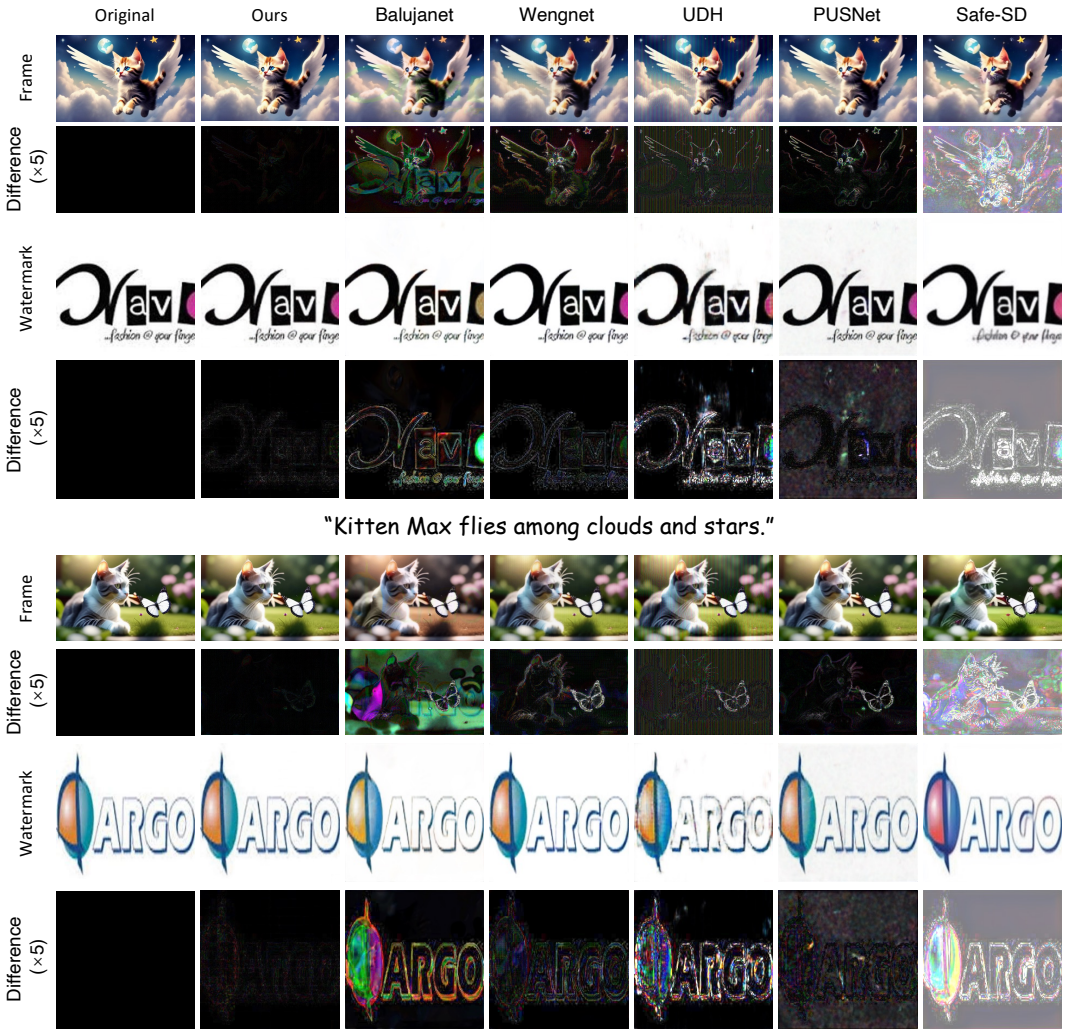

Figure 4: Qualitative comparison results on the first frame of each video. Difference maps show absolute differences between the watermarked and original videos, and between the recovered and original watermarks. More examples are shown in Fig. 10 of Appendix. **Best viewed with zoom in.**

4.2 Comparison with State-of-the-art Methods

Qualitative Comparison. Fig. 4 shows the qualitative comparisons on the first frame of each video, while Fig. 5 presents visual results of Safe-Sora across multiple frames. As illustrated, Balujanet introduces clearly visible artifacts in the watermarked video, UDH suffers from stripe-like distortions, and Safe-SD presents noticeable color shifts. From the difference maps, it is evident that both WengNet and PUSNet introduce considerable degradation to both video quality and watermark fidelity. In contrast, our method produces watermarked videos with high visual fidelity, exhibiting minimal differences from the original videos. Moreover, the recovered watermark images closely resemble the originals, demonstrating high reconstruction accuracy.

Quantitative Comparison. To evaluate the accuracy of watermark recovery and the invisibility of the watermark (i.e., video quality), we adopt standard metrics including PSNR, MAE, RMSE, SSIM [61], and LPIPS [62]. To assess temporal consistency in videos, we employ tLP [63] and Fréchet Video Distance (FVD) [64]. Quantitative results are summarized in Tab 1. As shown in the table, our method achieves state-of-the-art performance across all evaluation metrics. We observe that image watermarking methods inject watermarks by embedding them independently into each frame, which leads to poor temporal consistency and higher FVD scores. In contrast, our method leverages Mamba's long-range modeling capability across space and time, along with the proposed

Figure 5: Visual results of Safe-Sora on multiple frames. For each frame, we show the original image, the corresponding watermarked image, and their residual difference. **Best viewed with zoom in.**

Table 1: Quantitative results on watermark quality and video quality metrics. Watermark quality is measured by comparing the recovered watermark image with the original watermark, while video quality is evaluated by comparing the watermarked video with the original video.

Method	Watermark quality					Video quality						
	PSNR ↑	MAE ↓	RMSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑	MAE ↓	RMSE ↓	SSIM ↑	LPIPS ↓	tLP ↓	FVD ↓
Balujanet	25.28	9.61	15.10	0.91	0.11	25.26	10.09	14.58	0.87	0.25	1.32	512.22
Wengnet	33.18	3.71	5.82	0.96	0.06	28.09	6.27	10.69	0.85	0.21	1.27	265.82
UDH	22.90	11.29	19.29	0.77	0.24	27.75	8.16	10.72	0.73	0.32	2.09	1075.62
PUSNet	28.86	7.45	9.57	0.93	0.12	29.98	4.50	8.72	0.92	0.11	0.98	154.35
Safe-SD	24.24	9.78	17.39	0.84	0.11	22.32	11.65	20.64	0.75	0.24	1.87	849.83
Ours	**37.71**	**2.22**	**3.61**	**0.97**	**0.04**	**42.50**	**1.36**	**1.96**	**0.98**	**0.01**	**0.38**	**3.77**

spatiotemporal local scanning strategy, resulting in superior temporal consistency. Specifically, our method achieves an FVD of 3.77, significantly outperforming all baselines.

4.3 Robustness

To rigorously evaluate the robustness of our method, we apply a variety of distortion types. For random erasing, we randomly select an erasure ratio from the range [5%, 10%, 15%, 20%]. For Gaussian blur, we randomly choose a kernel size from 3, 5, 7. For Gaussian noise, we add noise with a standard deviation randomly sampled from a uniform distribution $\mathcal{U}(0, 0.2)$. For rotation, the degree is randomly sampled from the range $(-30°, 30°)$. Specifically for video, we adopt H.264 compression with a fixed CRF value of 24. We use PSNR, SSIM, and LPIPS to evaluate the robustness of watermark reconstruction under these distortions. As shown in Fig. 6, our method consistently achieves the best performance across all types of attacks, demonstrating strong robustness. In particular, under H.264 compression, all baseline methods suffer a significant drop in performance, whereas our method maintains high watermark quality.

4.4 Ablation Study

We conduct an ablation study on two key components— Coarse-to-Fine Adaptive Patch Matching and Spatiotemporal Local Scanning. Additional ablation studies can be found in Appendix D.

Impact of Coarse-to-Fine Adaptive Patch Matching. This strategy matches the most similar frame and spatial location for each watermark patch, based on similarity computed with the video latent representations. To evaluate the effectiveness of each component, we investigate three ablated variants of our method: **w/o CFAPM**, which completely removes the Coarse-to-Fine Adaptive Patch Matching mechanism; **w/o RtL**, which replaces the Routing by Latent strategy with a direct pixel-frame similarity computation; and **w/o FS**, which removes the Fine Stage responsible for spatial location refinement.

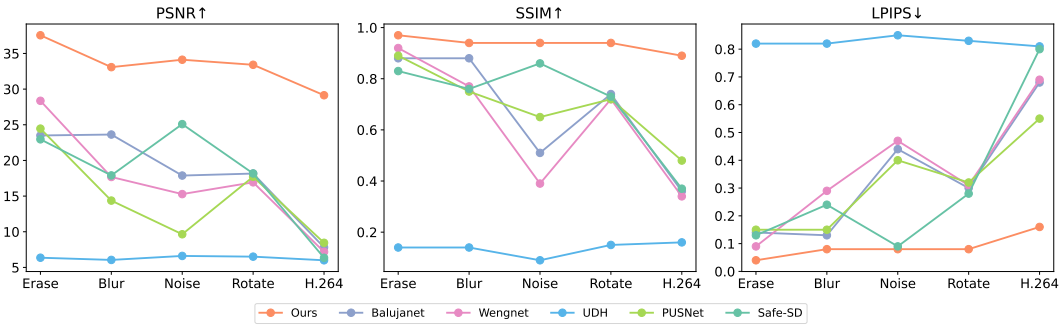

Figure 6: Watermark reconstruction quality under various distortions. Distortion settings include: Random Erasing (5%–20%), Gaussian Blur (kernel size 3/5/7), Gaussian Noise ($\sigma \sim \mathcal{U}(0, 0.2)$), Rotation (-30°, 30°), and H.264 Compression (CRF = 24).

Table 2: Comprehensive ablation study on key components of our method. CFAPM: Coarse-to-Fine Adaptive Patch Matching; RtL: Routing by Latent; FS: Fine Stage; SLS: Spatiotemporal Local Scanning; SFS: Spatial First Scanning within each subband.

Method	Watermark quality					Video quality						
	PSNR ↑	MAE ↓	RMSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑	MAE ↓	RMSE ↓	SSIM ↑	LPIPS ↓	tLP ↓	FVD ↓
w/o CFAPM	36.71	2.53	3.99	0.96	0.05	39.68	1.94	2.76	0.97	0.03	1.14	16.87
w/o RtL	36.36	2.67	4.13	0.96	0.05	40.23	1.79	2.54	0.97	0.04	1.30	6.37
w/o FS	36.88	2.45	3.94	**0.97**	**0.04**	41.25	1.58	2.26	0.97	0.03	1.17	4.82
w/o SLS	35.96	2.98	4.02	0.94	0.08	38.42	1.98	2.12	0.92	0.03	1.01	13.16
w/o SFS	36.41	2.59	4.17	0.96	0.05	42.21	1.38	2.05	**0.98**	**0.01**	**0.24**	5.24
Ours	**37.71**	**2.22**	**3.61**	**0.97**	**0.04**	**42.50**	**1.36**	**1.96**	**0.98**	**0.01**	0.38	**3.77**

The results in Tab. 2 clearly demonstrate that each component of the CFAPM strategy plays a critical role in enhancing overall performance. Computing the similarity between watermark patches and video latents leverages the compressed semantic information encoded in the latent space, enabling more accurate matching; the fine stage further refines this process by identifying the most visually similar spatial location for each patch. Overall, the Coarse-to-Fine Adaptive Patch Matching mechanism consistently improves both watermark fidelity and video quality.

Impact of Spatiotemporal Local Scanning. This strategy traverses the eight subbands of the 3D wavelet transform in a frequency-aware hierarchical order. Within each subband, patches are selected following a spatial-first, temporal-second scanning pattern. To evaluate the effectiveness of this design, we ablate two key components: **w/o SLS**, which replaces the structured traversal with a vanilla 3D scanning strategy; and **w/o SFS**, which applies a temporal-first scanning order within each subband instead of the proposed spatial-first policy.

Results in Tab. 2 demonstrate that the full SLS strategy significantly improves both watermark and video quality. While the temporal-first scanning achieves slightly better tLP, it consistently underperforms in watermark fidelity metrics. In summary, SLS enables more effective fusion and extraction of watermark signals distributed across spatiotemporal regions, thereby enhancing the overall performance of watermark embedding.

5 Conclusion

Our work introduces Safe-Sora, the first framework embedding graphical watermarks directly into generated video. We propose a hierarchical coarse-to-fine adaptive matching strategy that optimally maps watermark patches to visually similar frames and spatial regions. Our 3D wavelet transform-enhanced Mamba architecture with a novel spatiotemporal local scanning strategy, effectively models spatiotemporal dependencies for watermark embedding and retrieval, pioneering the application of state space models to watermarking. Experiments demonstrate that Safe-Sora achieves superior performance in video quality, watermark fidelity, and robustness. This work establishes a foundation for copyright protection in generative video and opens new avenues for applying state space models to digital watermarking.

Acknowledgments

This work is supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008), and the National Natural Science Foundation of China (62502317).

References

[1] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[2] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *CVPR*, 2024, pp. 7310–7320.

[3] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

[4] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao, "Latte: Latent diffusion transformer for video generation," *arXiv preprint arXiv:2401.03048*, 2024.

[5] J. Xing, M. Xia, Y. Zhang, H. Chen, W. Yu, H. Liu, G. Liu, X. Wang, Y. Shan, and T.-T. Wong, "Dynamicrafter: Animating open-domain images with video diffusion priors," in *ECCV*. Springer, 2024, pp. 399–417.

[6] C. Chen, J. Zhu, X. Feng, N. Huang, M. Wu, F. Mao, J. Wu, X. Chu, and X. Li, "S^2-guidance: Stochastic self guidance for training-free enhancement of diffusion models," 2025. [Online]. Available: https://arxiv.org/abs/2508.12880

[7] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You, "Open-sora: Democratizing efficient video production for all," *arXiv preprint arXiv:2412.20404*, 2024.

[8] Y. Zhao, T. Pang, C. Du, X. Yang, N.-M. Cheung, and M. Lin, "A recipe for watermarking diffusion models," *arXiv preprint arXiv:2303.10137*, 2023.

[9] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *ICCV*, 2023, pp. 22 466–22 477.

[10] R. Min, S. Li, H. Chen, and M. Cheng, "A watermark-conditioned diffusion model for ip protection," in *ECCV*. Springer, 2024, pp. 104–120.

[11] C. Xiong, C. Qin, G. Feng, and X. Zhang, "Flexible and secure watermarking for latent diffusion model," in *ACM MM*, 2023, pp. 1668–1676.

[12] L. Lei, K. Gai, J. Yu, and L. Zhu, "Diffusetrace: A transparent and flexible watermarking scheme for latent diffusion model," *arXiv preprint arXiv:2405.02696*, 2024.

[13] Z. Meng, B. Peng, and J. Dong, "Latent watermark: Inject and detect watermarks in latent diffusion space," *IEEE Transactions on Multimedia*, 2025.

[14] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu, "Gaussian shading: Provable performance-lossless image watermarking for diffusion models," in *CVPR*, 2024, pp. 12 162–12 171.

[15] H. Ci, P. Yang, Y. Song, and M. Z. Shou, "Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification," in *ECCV*. Springer, 2024, pp. 338–354.

[16] R. Hu, J. Zhang, Y. Li, J. Li, Q. Guo, H. Qiu, and T. Zhang, "Videoshield: Regulating diffusion-based video generation models via watermarking," *arXiv preprint arXiv:2501.14195*, 2025.

[17] M. Jang, Y. Jang, J. Lee, K. Kawamura, F. Yang, and S. Kim, "Lvmark: Robust watermark for latent video diffusion models," *arXiv preprint arXiv:2412.09122*, 2024.

[18] S. Baluja, "Hiding images within images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 7, pp. 1685–1697, 2019.

[19] J. Wang, W. Min, S. Hou, S. Ma, Y. Zheng, H. Wang, and S. Jiang, "Logo-2K+: a large-scale logo dataset for scalable logo classification," in *AAAI*, 2020.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255.

[21] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-w. Chao, B. E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang, and S. Tulyakov, "Panda-70m: Captioning 70m videos with multiple cross-modality teachers," in *CVPR*, 2024.

[22] T. He, L. Gao, J. Song, and Y.-F. Li, "Towards open-vocabulary scene graph generation with prompt-based finetuning," in *ECCV*. Springer, 2022, pp. 56–73.

[23] T. He, L. Gao, J. Song, J. Cai, and Y.-F. Li, "Semantic compositional learning for low-shot scene graph generation," *arXiv preprint arXiv:2108.08600*, 2021.

[24] T. He, L. Gao, J. Song, and Y.-F. Li, "Toward a unified transformer-based framework for scene graph generation and human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 6274–6288, 2023.

[25] M. Li, P. Zhou, J.-W. Liu, J. Keppo, M. Lin, S. Yan, and X. Xu, "Instant3d: instant text-to-3d generation," *IJCV*, 2024.

[26] H. Xu, C. Yu, F. Xiao, J. Xing, H. Ci, W. Chen, F. Wang, and M. Li, "Cyc3d: Fine-grained controllable 3d generation via cycle consistency regularization," *arXiv preprint arXiv:2504.14975*, 2025.

[27] S. Liu, J. Li, G. Zhao, Y. Zhang, X. Meng, F. R. Yu, X. Ji, and M. Li, "Eventgpt: Event stream understanding with multimodal large language models," in *CVPR*, June 2025, pp. 29 139–29 149.

[28] F. Zhao, M. Li, L. Xu, W. Jiang, J. Gao, and D. Yan, "Favchat: Unlocking fine-grained facial video understanding with multimodal large language models," *arXiv preprint arXiv:2503.09158*, 2025.

[29] J. Li, X. Qiu, L. Xu, L. Guo, D. Qu, T. Long, C. Fan, and M. Li, "Unif2ace: Fine-grained face understanding and generation with unified multimodal models," *arXiv preprint arXiv:2503.08120*, 2025.

[30] Y. Shi, W. Yan, G. Xu, Y. Li, Y. Chen, Z. Li, F. R. Yu, M. Li, and S. Y. Yeo, "Pvchat: Personalized video chat with one-shot learning," *arXiv preprint arXiv:2503.17069*, 2025.

[31] Z. Su, J. Zhuang, and C. Yuan, "Texturediffusion: Target prompt disentangled editing for various texture transfer," in *ICASSP*. IEEE, 2025, pp. 1–5.

[32] S. Tan, X. Qiu, Y. Shu, G. Xu, L. Xu, X. Xu, H. Zhuang, M. Li, and F. Yu, "Wmarkgpt: Watermarked image understanding via multimodal large language models."

[33] Y. Wei, Z. Hu, L. Shen, Z. Wang, C. Yuan, and D. Tao, "Open-vocabulary customization from clip via data-free knowledge distillation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[34] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[35] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*. PMLR, 2021, pp. 8162–8171.

[36] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[37] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*. PMLR, 2015, pp. 2256–2265.

[38] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *NeurIPS*, vol. 34, pp. 8780–8794, 2021.

[39] Y. Wei, Z. Hu, L. Shen, Z. Wang, L. Li, Y. Li, and C. Yuan, "Meta-learning without data via unconditional diffusion models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[41] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[42] J. T. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," *arXiv preprint arXiv:2208.04933*, 2022.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[44] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[45] J. Wang, T. Gangavarapu, J. N. Yan, and A. M. Rush, "Mambabyte: Token-free selective state space model," *arXiv preprint arXiv:2401.13660*, 2024.

[46] R. Waleffe, W. Byeon, D. Riach, B. Norick, V. Korthikanti, T. Dao, A. Gu, A. Hatamizadeh, S. Singh, D. Narayanan *et al.*, "An empirical study of mamba-based language models," *arXiv preprint arXiv:2406.07887*, 2024.

[47] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *NeurIPS*, vol. 37, pp. 103 031–103 063, 2024.

[48] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," in *ECCV*. Springer, 2024, pp. 237–255.

[49] X. Tan, H. Wang, X. Geng, and L. Wang, "Frequency-guided diffusion model with perturbation training for skeleton-based video anomaly detection," *arXiv preprint arXiv:2412.03044*, 2024.

[50] S. A. Al-Taweel and P. Sumari, "Robust video watermarking based on 3d-dwt domain," in *TENCON 2009-2009 IEEE Region 10 Conference*. IEEE, 2009, pp. 1–6.

[51] X. Li and R. Wang, "A video watermarking scheme based on 3d-dwt and neural network," in *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*. IEEE, 2007, pp. 110–115.

[52] Z. Zhen, Y. Hu, and Z. Feng, "Freqmamba: Viewing mamba from a frequency perspective for image deraining," *arXiv preprint arXiv:2404.09476*, 2024.

[53] W. Wang and Y. Yang, "Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models," in *NeurIPS*, 2024.

[54] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[56] X. Luo, Y. Li, H. Chang, C. Liu, P. Milanfar, and F. Yang, "Dvmark: a deep multiscale framework for video watermarking," *IEEE Transactions on Image Processing*, 2023.

[57] C. Zhang, P. Benz, A. Karjauv, G. Sun, and I. S. Kweon, "Udh: Universal deep hiding for steganography, watermarking, and light field messaging," *NeurIPS*, vol. 33, pp. 10 223–10 234, 2020.

[58] G. Li, S. Li, Z. Luo, Z. Qian, and X. Zhang, "Purified and unified steganographic network," in *CVPR*, 2024, pp. 27 569–27 578.

[59] Z. Ma, G. Jia, B. Qi, and B. Zhou, "Safe-sd: Safe and traceable stable diffusion with text prompt trigger for invisible generative watermarking," in *ACM MM*, 2024, pp. 7113–7122.

[60] X. Weng, Y. Li, L. Chi, and Y. Mu, "High-capacity convolutional video steganography with temporal residual modeling," in *Proceedings of the 2019 on international conference on multimedia retrieval*, 2019, pp. 87–95.

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[62] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[63] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey, "Learning temporal coherence via self-supervision for gan-based video generation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 75–1, 2020.

[64] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We mention in the abstract that we propose the first framework that integrates graphical watermarks directly into the video generation process.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention in the Section E that although our method can successfully embed image watermarks, embedding more information-rich video watermarks remains a limitation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the experimental details in Section 4.1 and include the code in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code are provided in the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 4.1 describes the experimental setups and parameters in detail.

Guidelines:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We followed the experimental setup of prior works in generative watermarking, which do not report statistical significance or error bars. As our results show consistent and large margins over all baselines, we believe statistical testing would not affect the validity of our main claims.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 describes the computational resources required to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This research complies with the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section F discusses the societal impacts of this work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers of the datasets and base models, and strictly comply with their licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code includes documentation with detailed information on training, licensing, and usage limitations.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Technical Appendices

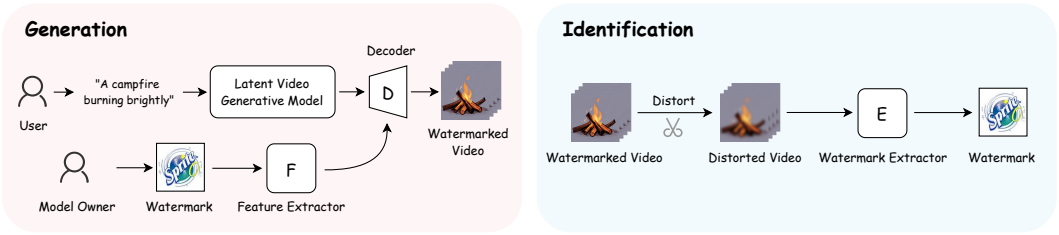

Figure 7: Application Scenario of Safe-Sora: A user provides a text prompt to a video generation model. The model owner's graphical watermark is embedded into the video through a feature extractor and decoder. Later, even if the video is distorted, a watermark extractor can recover the graphical watermark to verify authenticity and ensure copyright protection.

A Preliminaries

A.1 Latent Video Diffusion Models

Latent Video Diffusion Models (LVDMs) extend the concept of latent diffusion models to the video domain. These models operate in a compressed latent space rather than pixel space to improve computational efficiency while maintaining generation quality. The process can be described in three key steps:

First, a video encoder $\mathcal{E}$ maps the input video $x \in \mathbb{R}^{F\times H\times W\times 3}$ to a latent representation $z = \mathcal{E}(x) \in \mathbb{R}^{F\times h\times w\times c}$, where F is the number of frames, and the spatial dimensions are reduced: $h < H$ and $w < W$.

Second, a diffusion process gradually adds noise to the latent representation through a fixed Markov chain:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}), \tag{8}$$

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{9}$$

where β_t is the noise schedule, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Finally, a denoising network ϵ_θ is trained to predict the added noise at each time step. During generation, the reverse process starts from pure Gaussian noise $z_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises to produce z_0, which is then decoded to the final video $\hat{x} = \mathcal{D}(z_0)$ using a decoder $\mathcal{D}$.

For text-to-video generation, LVDMs incorporate a text encoder that processes a conditioning prompt, which guides the denoising process toward the desired content.

A.2 State Space Models

State Space Models (SSMs) are continuous dynamical systems defined by the following equations:

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \tag{10}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t), \tag{11}$$

where $\mathbf{x}(t)$ is the input, $\mathbf{h}(t)$ is the hidden state, $\mathbf{y}(t)$ is the output, and $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ are the parameters of the system.

For discrete sequence modeling, these continuous equations are discretized:

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t, \tag{12}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{h}_t + \mathbf{D}\mathbf{x}_t, \tag{13}$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are the discretized versions of $\mathbf{A}$ and $\mathbf{B}$.

Algorithm 1 Confidence-Guided Greedy Assignment for Watermark Position Recovery

Input: Watermark patches with position channel
Output: Reconstructed watermark image $\mathcal{W}$

Stage 1: Position Decoding
for each patch i **do**
    Normalize position channel to $[0, 1]$
    Compute probability vector p_i by averaging binary vectors in the position channel
    Compute confidence $c_i = \frac{1}{K} \sum_{j=1}^{K} |p_i^j - 0.5|$
    Convert p_i to binary $\hat{b}_i$ via thresholding
    Decode $\hat{b}_i \rightarrow$ position index $pos_i \in [0, N-1]$
end for

Stage 2: Confidence-Prioritized Assignment
Initialize watermark image $\mathcal{W} \leftarrow \emptyset$
Initialize unassigned patch pool $\mathcal{U} \leftarrow \emptyset$
for each patch i **do**
    if pos_i is unoccupied in $\mathcal{W}$ **then**
        Assign patch i to position pos_i in $\mathcal{W}$
    else if $c_i >$ confidence of current patch at pos_i **then**
        Replace patch at pos_i with i in $\mathcal{W}$
        Add the replaced patch to $\mathcal{U}$
    else
        Add patch i to $\mathcal{U}$
    end if
end for

Stage 3: Greedy Reassignment of Unassigned Patches
Sort $\mathcal{U}$ by descending c_i
for each patch j in $\mathcal{U}$ **do**
    Find nearest vacant position p_j to pos_j
    Assign patch j to position p_j in $\mathcal{W}$
end for
return $\mathcal{W}$

The Mamba architecture extends traditional SSMs by introducing input-dependent parameters:

$$\bar{\mathbf{A}}, \bar{\mathbf{B}} = \text{Projection}(\mathbf{x}), \tag{14}$$

$$\mathbf{h}_t = \bar{\mathbf{A}} \odot \mathbf{h}_{t-1} + \bar{\mathbf{B}} \odot \mathbf{x}_t, \tag{15}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{h}_t, \tag{16}$$

This input-dependent parameterization allows Mamba to dynamically adapt its processing based on input content, making it effective for modeling complex sequential dependencies.

A.3 Wavelet Transforms

Wavelet transforms decompose signals into multiple frequency components with localized time information, making them useful for frequency domain watermarking.

For images, the 2D Discrete Wavelet Transform (DWT) decomposes an image into four sub-bands: approximation (*LL*), horizontal detail (*LH*), vertical detail (*HL*), and diagonal detail (*HH*).

The 3D Discrete Wavelet Transform extends the 2D DWT to the temporal domain for video processing. A video sequence is decomposed into eight sub-bands: *LLL*, *LLH*, *LHL*, *LHH*, *HLL*, *HLH*, *HHL*, and *HHH*, with *L* and *H* representing low and high frequencies across the frame, height, and width dimensions. Each sub-band has half the resolution of the original video in all dimensions. The 3D DWT provides a multi-level representation of videos, capturing both spatial and temporal characteristics, which is beneficial for video watermarking by allowing embedding in specific frequency bands while preserving perceptual quality.

Table 3: Quantitative comparison on VideoCrafter2 and Open-Sora backbones.

Backbone	Watermark quality					Video quality						
	PSNR ↑	MAE ↓	RMSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑	MAE ↓	RMSE ↓	SSIM ↑	LPIPS ↓	tLP ↓	FVD ↓
VideoCrafter2	**37.71**	**2.22**	**3.61**	**0.97**	**0.04**	42.50	1.36	1.96	**0.98**	**0.01**	0.38	3.77
Open-Sora	35.42	2.93	4.70	0.96	0.06	**44.15**	**1.31**	**1.75**	0.97	**0.01**	**0.31**	**3.04**

Figure 8: Qualitative examples on Open-Sora backbone. Best viewed with zoom in.

B Robust Watermark Position Recovery Algorithm

To address rare cases where multiple watermark patches are decoded to the same spatial location due to distortion or attack, we propose a confidence-guided greedy assignment algorithm. This algorithm ensures reliable and unambiguous recovery of watermark positions by incorporating confidence estimation, conflict resolution, and greedy reassignment of unplaced patches.

The algorithm is as follows: first, compute the confidence score for each patch's predicted position. Then, assign each patch to its corresponding position; in case of conflicts, give priority to the patch with higher confidence. Finally, assign the remaining unplaced patches in descending order of confidence to the nearest available positions. The detailed procedure is illustrated in Algorithm 1.

The confidence-guided greedy assignment algorithm effectively handles noisy or partial position corruption and significantly improves the robustness of watermark extraction.

C More Backbones

While our main experiments are conducted using VideoCrafter2 [2], a UNet-based video generation model, we further evaluate our method using Open-Sora [7], a DiT-based video generation model. Quantitative results are shown in Tab. 3, and qualitative examples are provided in Fig. 8. As can be seen, Open-Sora achieves comparable performance to VideoCrafter2 and produces videos with higher visual quality, but slightly lower watermark fidelity. These results demonstrate that our method is effective across different video generation models.

Table 4: Additional Ablation Studies. MSFI: Multi-Scale Feature Injection.

Method	Watermark quality					Video quality						
	PSNR ↑	MAE ↓	RMSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑	MAE ↓	RMSE ↓	SSIM ↑	LPIPS ↓	tLP ↓	FVD ↓
w/o MSFI	36.56	2.56	4.06	0.96	0.05	39.39	2.02	2.84	0.97	0.03	1.19	14.11
Ours	**37.71**	**2.22**	**3.61**	**0.97**	**0.04**	**42.50**	**1.36**	**1.96**	**0.98**	**0.01**	**0.38**	**3.77**

Figure 9: Visual impact of Multi-Scale Feature Injection. We present difference maps (×5) between watermarked and original videos. After applying Multi-Scale Feature Injection, the differences are significantly reduced, leading to improved video quality.

D Additional Ablation Studies

To further assess the contribution of individual components in our framework, we perform extended ablation studies beyond the main experiments. In particular, we examine the impact of Multi-Scale Feature Injection, with quantitative results reported in Tab. 4 and qualitative comparisons shown in Fig. 9. The results demonstrate that incorporating the inherent multi-scale features of the VAE notably improves the visual quality of generated videos.

E Limitations

While our method demonstrates strong performance in embedding and recovering static graphical watermarks, it is currently limited to image-based watermarks such as logos or icons. Embedding more complex and information-rich video watermarks—e.g., animated sequences or temporally dynamic patterns—remains a challenge.

F Societal Impact

The ability to embed graphical watermarks directly into the video generation process carries important social and ethical implications. On the positive side, it provides a practical solution to the growing concerns over ownership verification and copyright protection in generative media. As synthetic content becomes increasingly widespread, methods like ours can help content creators assert their rights and trace misuse, thereby fostering accountability and transparency in digital media ecosystems.

However, like many watermarking techniques, our method may also be misused. For example, it could potentially be employed to falsely claim ownership over public material, or to embed unauthorized logos into generated videos. We strongly advocate for the responsible use of generative watermarking technologies and recommend that future research explores methods to verify the authenticity of embedded watermarks and prevent abuse.

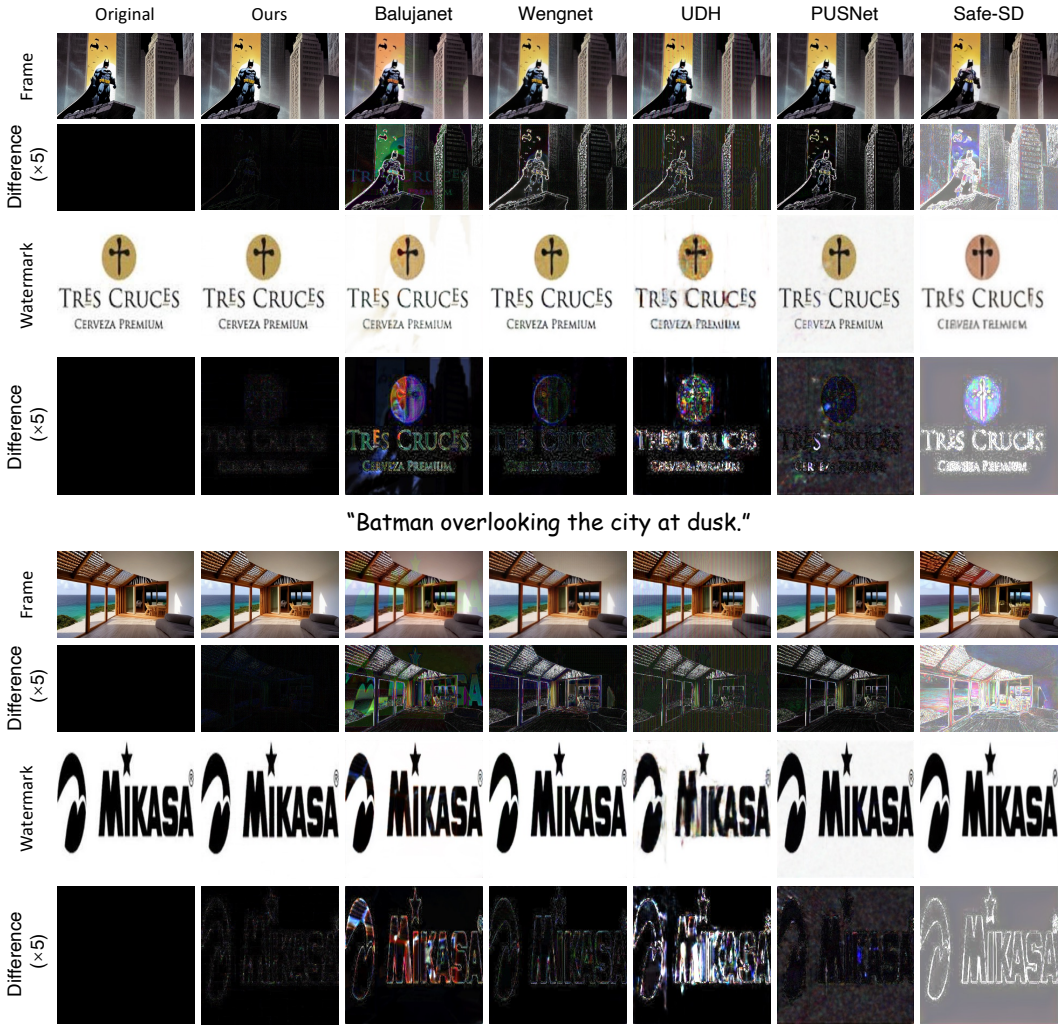

Figure 10: More qualitative examples on VideoCrafter2 backbone. Difference maps show absolute differences between the watermarked and original videos, and between the recovered and original watermarks. **Best viewed with zoom in.**