LD-RoViS: Training-free Robust Video Steganography for Deterministic Latent Diffusion Model

Xiangkun Wang[1,2] **Kejiang Chen**[1,2]* **Lincong Li**[1,2] **Weiming Zhang**[1,2] **Nenghai Yu**[1,2]

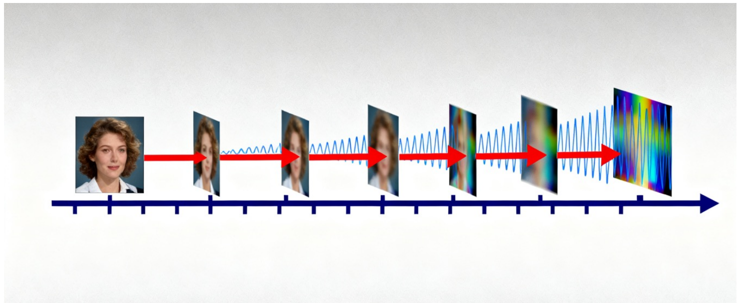
[1]University of Science and Technology of China, China
[2]Anhui Province Key Laboratory of Digital Security, China

wangxiangkun@mail.ustc.edu.cn chenkj@ustc.edu.cn

1 Motivation

The Dilemma of Traditional Video Steganography:

- Vulnerable to video compression:
  - **H.264 compression** coding and lossy processing on social platforms lead to **distortion drift** and low capacity.

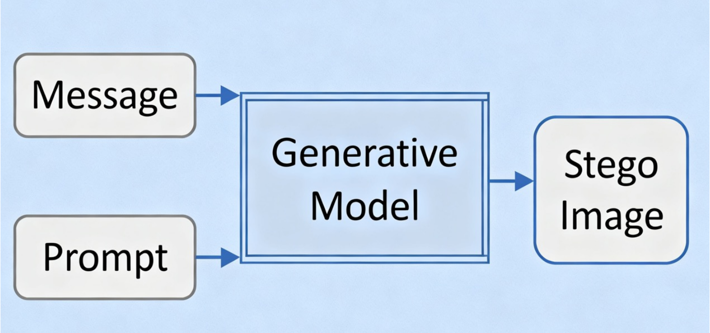


distortion drift

- Vulnerable to steganalysis:
  - To ensure the accuracy of extraction, steganographic embedding occurs in **low-frequency regions**, making it prone to detection.

A promising solution: Generative steganography

- **avoids direct modification** of the cover data, offers a promising solution.

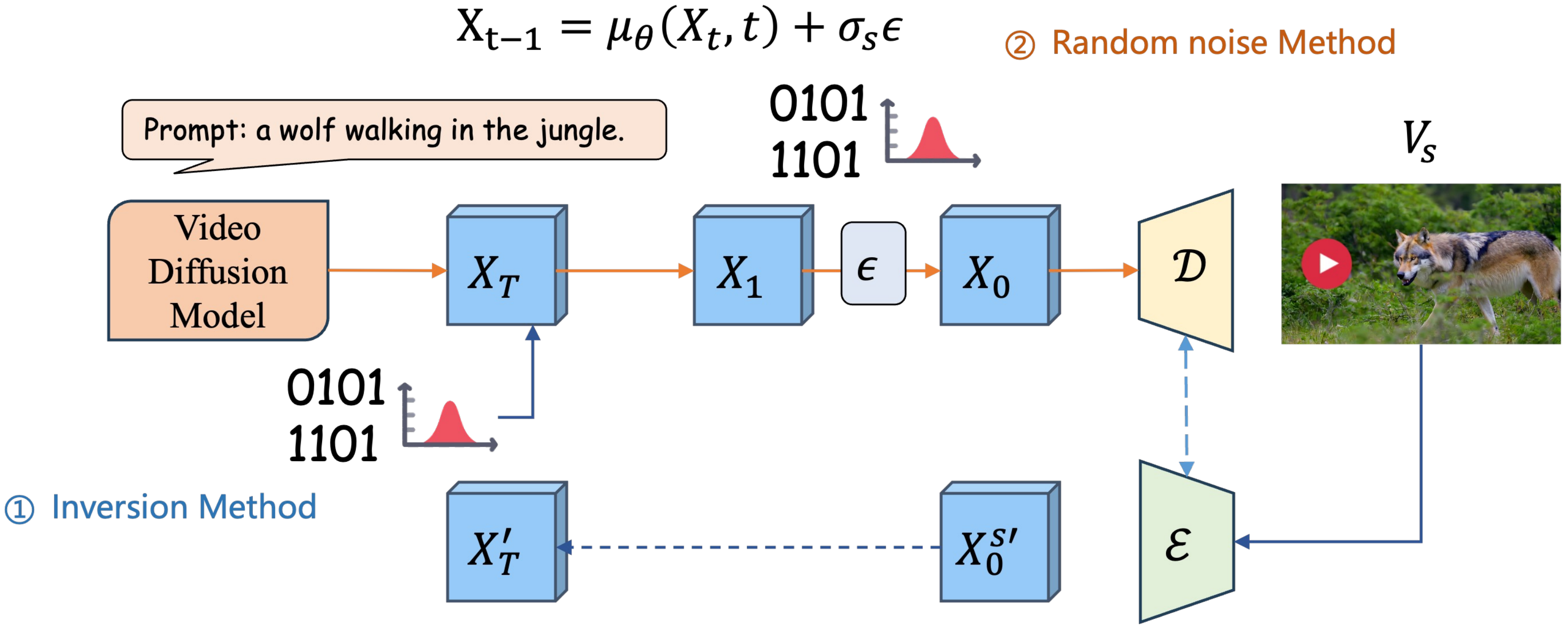


generative steganography

1 Motivation

The Dilemma of Generative Steganography Applied to Videos:

- Mainstream video generation models use **non-reversible samplers**, making precise inversion difficult.
- The lossy processing of **VAE encoding and decoding**, along with **deterministic sampling**, renders noise-based methods inapplicable.
- Processing on social platforms (such as **compression**) poses challenges to robustness.



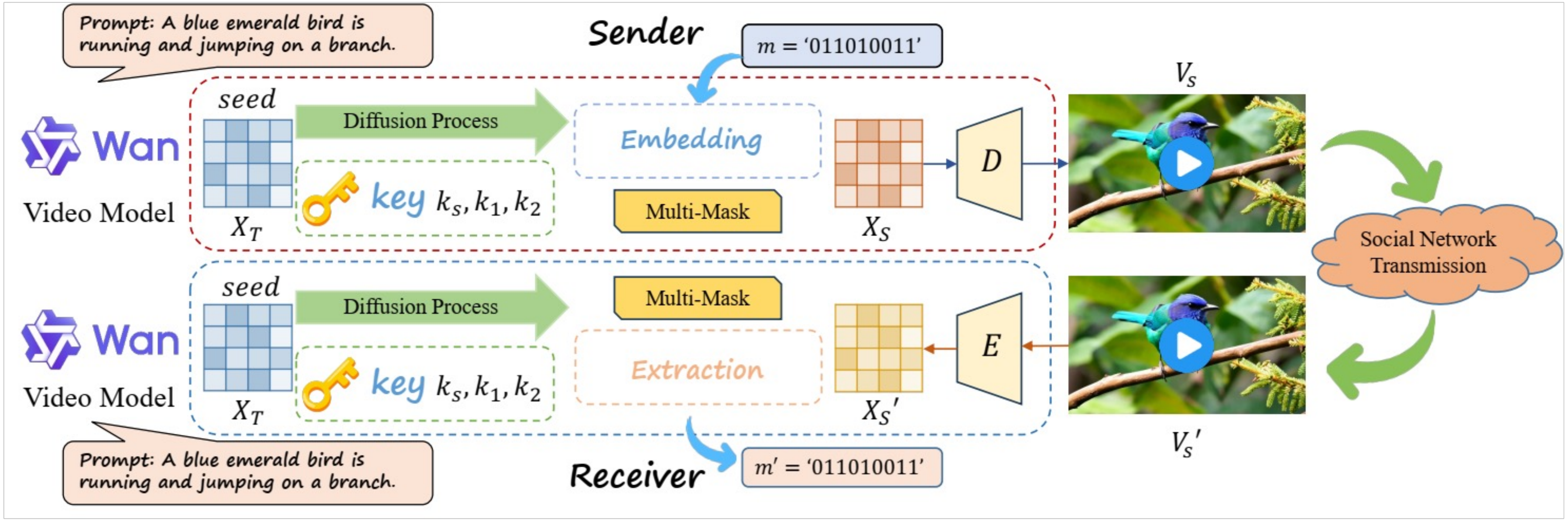
① Inversion Method: maps messages to the **initial latent variable (Gaussian noise)** and relies on **inverting the sampling process** for message extraction.

② Random noise Method: embeds messages into **the noise ε added at the final timestep**, but is **sensitive to VAE encoding–decoding distortions**.

2 Overview

➢ LD-RoViS: Training-Free and Robust Generative Video Steganography

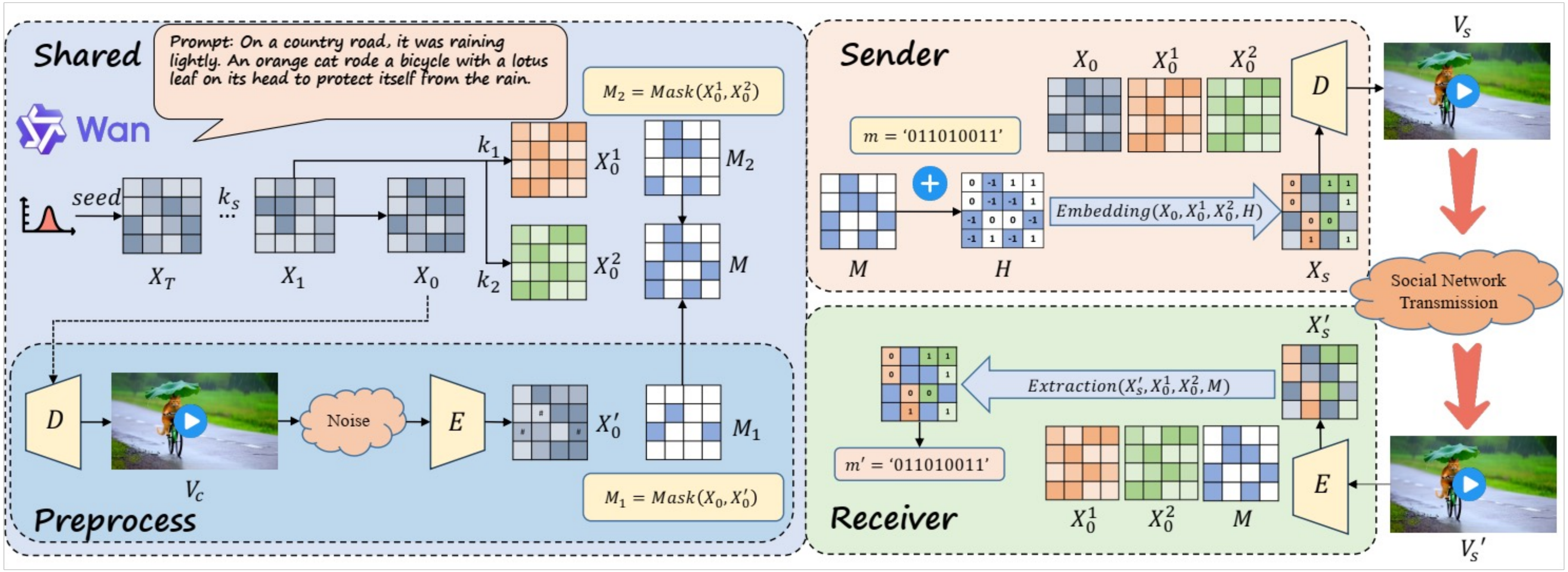
- Implemented based on Alibaba's Wan2.1 model
- Controlling video generation process and message embedding via shared keys
- Achieving robust region embedding through the Multi-Mask mechanism



2 Framework

➢ LD-RoViS：Message Embedding and Extraction

- In the diffusion model's final denoising step, **divergent variables** from varied parameter settings enable message embedding.
- Pre-encoding/decoding and discriminative processing identify **invariant and distinguishable regions,** generating multiple masks.
- The receiver uses shared parameters to **replicate the process** and extracts messages via divergent variable distances.



2 Framework

➢ LD-RoViS： Message Embedding

◆ Divergent Variable Acquisition: Construction of Steganographic Channel Based on Implicit Parameter Adjustment.

Divergent variables X_0^1, X_0^2 are obtained using k_1, k_2 , which are used for embedding steganographic messages.

$$\epsilon_\theta(\mathbf{x}_t, t) = pred_{uncond} + \mathrm{CFG} \cdot (pred_{cond} - pred_{uncond}),$$

$$X_i = \mathit{Diffuse}(G, X_1, k_i), \quad \text{where } X_i \in \{X_0, X_0^1, X_0^2\},\ k_i \in \{k_s, k_1, k_2\}.$$

◆ Multi-Mask Construction:

(1) Invariance mask M_1： Latent regions stable against VAE

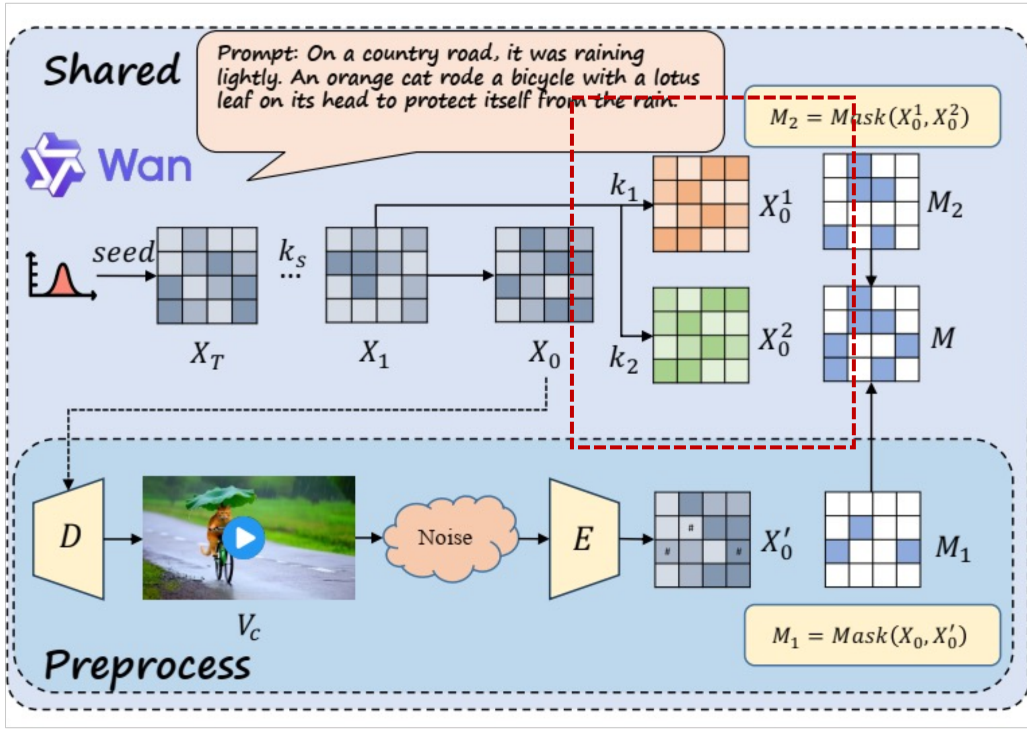
$$M_1(c, f, h, w) = \mathit{Mask}(d_1, \tau_1) = \begin{cases} 1 & \text{if } d_1(c, f, h, w) \in \text{top smallest } \tau_1, \\ 0 & \text{otherwise.} \end{cases}$$

(2) Distinguishability mask M_2： Regions with significant differences

$$M_2(c, f, h, w) = I - \mathit{Mask}(d_2, 1 - \tau_2) = \begin{cases} 1 & \text{if } d_2(c, f, h, w) \in \text{top largest } \tau_2, \\ 0 & \text{otherwise,} \end{cases}$$

(3) Final mask M is the dot product of the two masks

$$M = M_1 \odot M_2, M \in \mathbb{R}^{C' \times F' \times H' \times W'}.$$



2 Framework

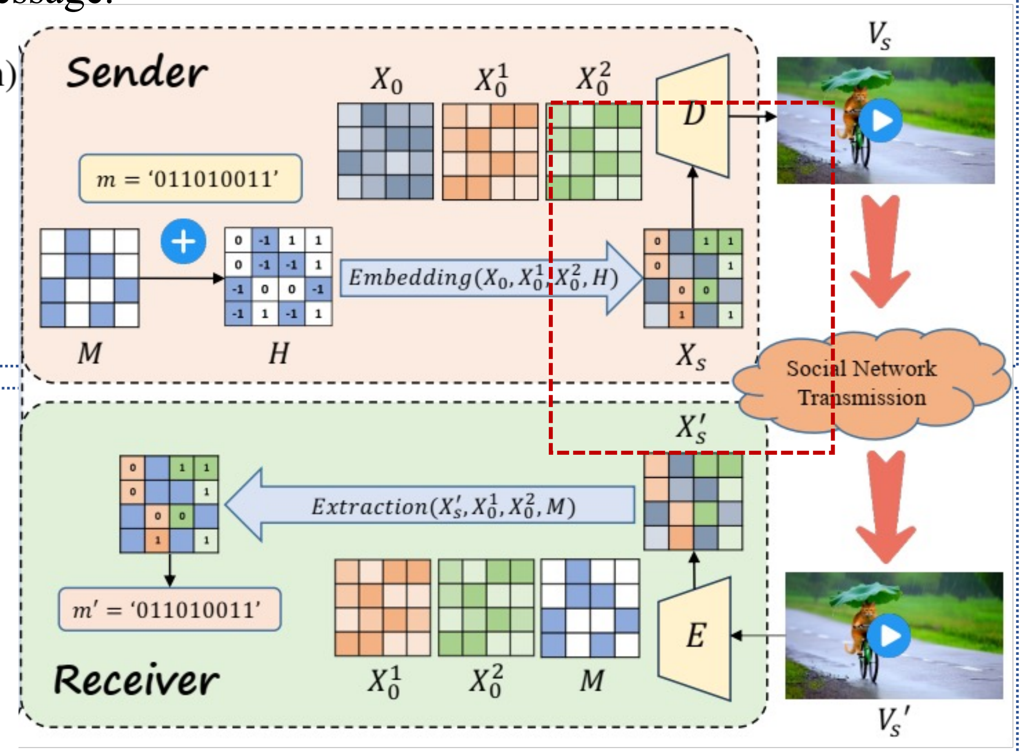
NEURAL INFORMATION PROCESSING SYSTEMS

➢ LD-RoViS： Message Embedding

◆ Matrix H, filled via message and mask, mixes divergent variables to hide the message.

$$H(c,f,h,w) = Transform(M,m) = \begin{cases} -1 & \text{if } M(c,f,h,w)=0, \quad \text{(non-embedding region)} \\ m_k & \text{if } M(c,f,h,w)=1, \quad \text{(embedding region)} \end{cases}$$

$$X_s(c,f,h,w) = Embedding(X_0, X_0^1, X_0^2, H) = \begin{cases} X_0(c,f,h,w) & \text{if } H(c,f,h,w)=-1, \\ X_0^1(c,f,h,w) & \text{if } H(c,f,h,w)=0, \\ X_0^2(c,f,h,w) & \text{if } H(c,f,h,w)=1. \end{cases}$$



➢ LD-RoViS： Message Extraction

◆ The message is recovered from the distance between restored mixed variables and divergent variables.

$$d_1'(c,f,h,w) = \|X_s'(c,f,h,w) - X_0^1(c,f,h,w)\|,$$

$$d_2'(c,f,h,w) = \|X_s'(c,f,h,w) - X_0^2(c,f,h,w)\|,$$

$$m_k' = \begin{cases} 0 & \text{if } d_1'(c,f,h,w) < d_2'(c,f,h,w) \text{ and } M(c,f,h,w)=1, \\ 1 & \text{if } d_1'(c,f,h,w) \geq d_2'(c,f,h,w) \text{ and } M(c,f,h,w)=1, \end{cases}$$

3 Experimental Results

➢ LD-RoViS：

- Visual Effects: 5-second 480p video, fps=16, with 12,000 bits of embedded message

3 Experimental Results

➤ LD-RoViS：

- Visual Quality Evaluation:
- PSNR and BRISQUE both remain optimal.
- Message extraction accuracy stays above 99%.

Table 1: Comparison of performance. The results are presented as the means and standard deviations.

Method	PSNR↑	BRISQUE↓	acc (%)↑	capacity↑
AQIM	34.81 ± 0.44	32.87 ± 6.06	**99.44 ± 0.27**	10000 (fixed)
MEC_AQIM	35.21 ± 0.47	32.71 ± 6.10	90.99 ± 5.90	10000 (fixed)
RoGVSN	–	49.53 ± 4.55	99.28 ± 0.38	729 (fixed)
Ours	**41.66 ± 1.52**	**28.90 ± 6.05**	99.17 ± 0.63	**11983 ± 1446**

- Compared with comparative methods, there are no block artifacts or resolution reduction.

3 Experimental Results

➢ LD-RoViS：

- Security Experiments: The error rates of three steganalysis methods are all around 50%, close to random guessing.
- Robustness Experiments: Tested lossy processes such as H.264 compression (CRF), Gaussian noise, salt-and-pepper noise, and brightness adjustment, showing strong robustness.

Table 2: P_E (%,↑) of steganalysis.

Method	SUPERB	CovNet	LWENet
AQIM	49.14	0.13	0.26
MEC_AQIM	47.32	0.01	1.07
RoGVSN	47.58	0.36	2.61
ours	**49.18**	**49.74**	**48.49**

Table 3: acc(%) under different compression and noise.

Method	-	CRF=18	CRF=23	CRF=27	noise	salt&pepper	brightness
AQIM	**99.44**	91.24	90.67	87.49	82.46	80.04	48.93
MEC_AQIM	90.99	82.83	82.29	78.87	72.83	71.60	50.31
RoGVSN	99.28	**97.42**	**97.06**	**97.04**	**96.20**	94.45	96.05
ours	99.17	95.89	93.70	91.67	92.82	**98.72**	**99.02**

- Ablation Experiments: Verified the effectiveness of the multi-mask mechanism.

Table 4: Ablation variants.

Method	Mask M_1	Mask M_2
variant#1	×	×
variant#2	✓	×
variant#3	×	✓
ours	✓	✓

Table 5: Performance of different variants.

Method	acc(%)↑	PSNR(db)↑	BRISQUE↓	capacity(bits)↑
variant#1	62.67	35.39	30.55	**1935111**
variant#2	75.46	37.49	29.47	617913
variant#3	88.59	40.53	29.01	41132
ours	**99.17**	**41.66**	**28.90**	11983

Thanks for listening!

Contact us: wangxiangkun@mail.ustc.edu.cn