中图法分类号:TP309　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)11-3465-14

论文引用格式:Li W Q, Zhou K Y, Hu X X, Zhang X P, Li S and Qian Z X. 2025. Generative video steganography via secret mapping in diffusion models. Journal of Image and Graphics, 30(11):3465-3478(李文琪, 周科扬, 胡宵宵, 张新鹏, 李晟, 钱振兴. 2025. 扩散模型中秘密信息映射的生成式视频隐写. 中国图象图形学报, 30(11):3465-3478)[DOI:10.11834/jig.250079]

扩散模型中秘密信息映射的生成式视频隐写

李文琪,周科扬,胡宵宵,张新鹏,李晟,钱振兴*

复旦大学计算机科学技术学院,上海　200433

摘　要:**目的**　视频隐写是一种以视频为载体隐藏秘密信息的技术。然而,现有方法通常修改视频中的冗余信息嵌入秘密信息,易在原始视频中留下痕迹,从而被隐写分析工具检测。为解决这一问题,首次提出了一种基于扩散模型的生成式视频隐写方案,可直接从秘密信息生成视频,有效提高隐写的安全性。**方法**　在文本提示词和参考图像的可控视频生成过程中嵌入秘密信息,生成与正常视频视觉质量相似的含密视频。在秘密信息嵌入阶段,通过对秘密信息进行重复、加密等预处理操作,借助频域变换生成接近高斯分布的含密隐变量,进而生成含密视频;在秘密信息提取阶段,通过扩散模型的逆向过程恢复含密隐变量,然后使用频域变换的逆运算还原出经过处理的密文,并使用后处理操作恢复原始秘密信息。**结果**　实验中使用了常见的视频处理操作模拟有损信道传输过程,并在WebVid-10M数据集上与两种方法进行了对比。实验结果表明,在嵌密容量显著提高的情况下,面对不同强度的视频处理,本文方法的秘密信息提取准确率均能保持在90%以上。此外,相关评估指标表明,本文生成的含密视频在视觉质量上与正常生成的视频相似。实验进一步验证了含密隐变量接近独立高斯分布的性质,同时使用3种隐写分析网络对含密视频进行检测的准确率均接近50%,充分验证了本文方法的安全性。**结论**　本文提出的生成式视频隐写方法具备更高的安全性和嵌密容量,并在面对视频压缩等视频处理场景时表现出良好的鲁棒性,展示了其在视频隐写领域的应用潜力。

关键词:隐写;扩散模型;文本到视频生成模型;鲁棒视频隐写;生成式视频隐写

Generative video steganography via secret mapping in diffusion models

Li Wenqi, Zhou Keyang, Hu Xiaoxiao, Zhang Xinpeng, Li Sheng, Qian Zhenxing*

School of Computer Science, Fudan University, Shanghai　200433, China

Abstract: **Objective** Video steganography is a technique used to hide secret information within videos, leveraging videos as carriers for secure communication. Existing video steganography methods primarily embed secret information through the modification of redundant information within cover videos. Although these approaches have been widely adopted, they still suffer from considerable limitations. First, modifying original videos often leaves detectable traces, increasing the vulnerability of these methods to detection via advanced steganalysis networks, which are adept at identifying unnatural patterns introduced by such modifications. Second, robustness is a critical challenge for steganographic methods because videos are often subjected to lossy processing during transmission, such as compression, bit errors, and frame rate conversion. Current approaches that address robustness typically train secret extractors on distorted video data generated by attack layers simulating known types of distortions. Despite their effectiveness in controlled environments, these methods heavily

收稿日期:2025-02-28;**修回日期**:2025-06-16;**预印本日期**:2025-06-23

*通信作者:钱振兴　zxqian@fudan.edu.cn

基金项目:国家自然科学基金(62572125);上海市自然科学基金(25ZR1401019)

Supported by: National Natural Science Foundation of China (62572125); Shanghai Municipal Natural Science Foundation (25ZR1401019)

depend on prior knowledge regarding the types and intensities of possible distortions that videos might encounter during transmission. This dependence limits their adaptability and effectiveness in real-world scenarios where transmission distortions are unpredictable and diverse. Aiming to overcome these limitations, this paper introduces, for the first time, a generative video steganography scheme based on diffusion models. Different from traditional methods that require modification of existing videos as carriers, the proposed approach directly generates videos from secret information, bypassing the need for carrier videos and substantially enhancing steganographic security. Additionally, the proposed method operates in a stable and compact latent space to help embed and extract secret information. This approach not only reduces the computational complexity but also enhances robustness against various video processing distortions commonly encountered in practical applications. Through its elimination of cover video modifications and leveraging of the properties of latent diffusion models, this approach represents a remarkable advancement in video steganography. **Method** The proposed method introduces a novel bijective mapping module that establishes a direct and reversible mapping between secret information and Gaussian noise. This module is specifically designed to facilitate seamless embedding and extraction of secret information within the latent space of text-to-video (T2V) generative models. During the embedding phase, the sender applies the bijective mapping module to facilitate the transformation of the secret information into a Gaussian-like noise distribution using an inverse discrete cosine transform (IDCT-2D). The transformed Gaussian noise, which contains the embedded secret, is then input into the latent space of a pre-trained T2V model. The T2V model iteratively denoises the latent Gaussian noise, producing the final steganographic video. This embedding process seamlessly integrates secret information into the video during generation, ensuring that the final output video appears visually natural and eliminates the need for modifying existing cover videos. During the secret extraction phase, the receiver processes the received steganographic video, which may have undergone lossy transmission or various distortions. The reverse denoising process of the T2V model is applied to recover the latent Gaussian noise from the received video. Once the Gaussian noise is retrieved, the bijective mapping module uses a discrete cosine transform (DCT-2D) to reconstruct the original secret information. The use of reversible transformations ensures accurate extraction of the embedded information, even under challenging conditions. Aiming to further improve extraction accuracy, the method employs a repeated embedding strategy, in which the same secret is embedded multiple times throughout the generated video. This strategy allows for majority voting during the extraction phase, effectively mitigating the effects of lossy transmission and increasing overall robustness. **Result** Aiming to simulate real-world scenarios, an attack layer was incorporated into the experiments to introduce various video processing distortions on the steganographic videos. This layer emulated common distortions such as video compression, bit errors, and frame rate conversion. For video compression, H. 264 and H. 265 codecs were tested under different intensity levels, including average bitrate (ABR) and constant rate factor (CRF) settings. Bit errors were then simulated to replicate imperfections in transmission channels, where specific bits are flipped during the process. Frame rate conversion scenarios were specifically designed to mimic frame rate adjustments caused by device or network constraints. Each distortion type was applied at two intensity levels, with high levels leading to substantial video degradation. Extensive experiments on the WebVid-10M dataset demonstrated the effectiveness of the proposed method compared to two baseline approaches. The proposed method achieved notably higher embedding capacities, with bit rates of 96, 192, and 384 bits per frame, surpassing the baseline methods. The proposed approach maintained high secret extraction accuracy, exceeding 90%, even under severe video distortions such as high-intensity H. 264 and H. 265 compression. In contrast, both baseline methods demonstrated substantial reductions in extraction accuracy under similar conditions, highlighting the robustness of the proposed method. The visual quality of the generated videos was also evaluated using the Fréchet video distance (FVD) and contrastive language-image pre-training(CLIP) metrics, which measured text alignment, domain similarity, and motion smoothness. Experimental results revealed that the steganographic videos maintained visual quality comparable to videos generated by the T2V model in the absence of secret embedding, demonstrating their practicality and usability in real-world applications. Additionally, the security of the proposed method was validated through theoretical analysis and steganalysis experiments. Theoretical evaluation involved visualizing the stego noise, which revealed a distribution closely resembling a Gaussian distribution. Furthermore, the autocorrelation function verified the independence of stego noise elements. For the steganalysis evaluation, three advanced steganalysis networks were used to detect hidden information in the generated steg-

anographic videos. Detection rates remained close to 50%, indicating the effective detection resistance of the proposed method. This resistance highlights the strong security of the approach, which avoids altering existing videos and instead operates within a generative framework to produce natural and undetectable outputs. **Conclusion** This paper presents a novel generative video steganography scheme that uses diffusion models to facilitate video generation from secret information, addressing the limitations of traditional approaches. By operating within a stable latent space, the proposed method ensures higher robustness against common video processing distortions, including compression, bit errors, and frame rate conversion. Experimental results demonstrate that the proposed method achieves substantially higher embedding capacities compared to state-of-the-art baseline methods, while maintaining over 90% secret extraction accuracy under varying levels of video distortion. Additionally, the proposed approach exhibits high visual quality and strong resistance to steganalysis. These results highlight the practicality and security of the proposed scheme, making it a promising solution for secure communication in real-world scenarios. This research not only introduces a new paradigm for video steganography but also establishes a foundation for further research in generative steganographic techniques.

Key words: steganography; diffusion model; text-to-video (T2V) generative model; robust video steganography; generative video steganography

0 引　言

隐写术是一种通过将秘密消息隐藏在载体信息中，使攻击者难以察觉其存在，从而实现隐蔽通信的技术（张新鹏 等，2016）。与密码学中依赖加密算法隐藏信息的方式不同，隐写术的核心特点是不可感知性，尤其在军事通信等高隐蔽性要求的场景中具有重要应用价值。常见的隐写载体包括图像（Zhou 等，2023）、文本（Zhou 等，2022）、音频（Gopalan，2003）及视频（Meng 等，2024）。视频相较于其他载体，存在更多冗余信息，因此在嵌入容量与安全性方面具有更优的表现。

传统的视频隐写技术主要通过对视频载体的改动实现秘密信息的嵌入，根据视频格式可以分为基于原始视频域和基于压缩域的隐写方法。在原始视频域中，经典的最低有效位（least significant bit，LSB）（Chan 和 Cheng，2004）隐写技术通过改变每个像素的最小有效位来表示秘密比特流，随后基于此方法衍生出多种 LSB 变体（Yadav 等，2013；Dasgupta 等，2012）。此外，离散小波变换（discrete wavelet transform，DWT）和离散余弦变换（discrete cosine transform，DCT）也广泛用于频域隐写。例如 Mstafa 等人（2017）将秘密数据嵌入视频帧的 DWT-DCT 域，Suresh 和 Sam（2018）则通过随机置乱 DCT 系数的最低有效位以提升安全性与鲁棒性。相较而言，基于压缩域的视频隐写术可以利用更多种类的隐写载体，如帧内预测模式（Nie 等，2018）、帧间预测模式（张新鹏 等，2016）、运动矢量（Zhai 等，2020）和量化离散余弦变换系数（Xue 等，2019）。同时，也有大量基于 HEVC（high efficiency video coding）编码的视频隐写方法，如 Li 等人（2023）通过修改预测单元划分方式嵌入秘密信息，Dong 等人（2023）则引入编码效率退化模型，自适应调整嵌入策略以降低编码开销。然而，上述传统的视频隐写方法依赖载体视频，易产生可检测痕迹，可能被隐写分析工具识别，存在一定的安全风险。

深度学习迅猛发展，其在视频隐写中的应用也逐渐兴起。Khare 等人（2014）首次采用反向传播神经网络提升了 LSB 隐写性能；Pan 等人（2020）提出基于 MobileNet v2 的无载体视频隐写，通过语义分割与哈希映射实现秘密信息嵌入；马宾等人（2024）提出了一种融合多重对抗训练与通道注意力的生成对抗网络隐写方法，以提升含密图像质量与安全性。

尽管修改式视频隐写方案能够成功嵌入秘密信息，但其改变了载体视频的原始统计特征，这使得嵌入的信息更容易被隐写分析工具检测到，存在较高的安全风险。

近年来，生成式模型的快速发展为隐写领域带来了新的契机。已有学者在生成式图像隐写中探索可证安全方案，如王健等人（2023）通过可逆图像处理网络改进拒绝采样方法，实现自然图像上的可证安全隐写。扩散模型的兴起进一步推动了该领域进展。Hu 等人（2024）提出基于扩散模型的鲁棒隐写方法，并数学证明其生成图像在统计上接近真实图像，显著提升了隐写的不可检测性。Yu 等人（2023）

则将扩散过程与加密机制结合，提出基于密钥控制的图像隐写方案，为生成式隐写的安全性奠定理论基础。在生成式视频隐写领域，Mao 等人（2024）提出了一种端到端的鲁棒生成式视频隐写网络 RoGVSN（robust generativevideo steganography network），通过修改视频语义特征实现秘密信息嵌入，并在换脸场景中完成了视频隐写。然而，该方法存在嵌密容量较小的问题，且依赖于针对特定视频攻击训练的秘密信息提取器，因此在面对未知攻击时，其秘密信息提取性能受到较大限制。

为了解决现有生成式视频隐写方法的不足，本文提出了一种基于扩散模型的视频隐写方法，通过在正常的文本到视频生成过程中嵌入秘密信息，实现了一种安全性和鲁棒性兼备的隐写方案。为增强用户对生成视频的控制力，引入 SparseCtrl 模块（Guo 等，2024），将参考图作为视频的第 1 帧，并根据文本提示词生成后续帧，从而生成符合用户需求的视频。具体而言，本文方法利用余弦系数将秘密信息扩散至视频生成模型的整个隐空间中，确保含密隐向量的分布非常接近独立高斯分布，与正常生成模型中的隐向量分布一致。这一特性使生成的含密视频与正常生成的视频在分布上无显著差异，提升了隐写方案的安全性。实验结果表明，在嵌入容量显著高于基线方法的情况下，本文方法在不同等级的视频有损处理后，仍能保持 90% 以上的秘密信息提取准确率，验证了方法的优越性。此外，通过定量评估含密视频的视觉质量及其与文本提示词的匹配度，表明生成的含密视频在视觉质量上与正常生成的视频几乎无差异。同时，还使用了 3 个隐写分析网络进行测试，进一步验证了本文隐写方法在安全性方面的优越性。

本文的主要贡献总结如下：

1）首次提出了一种基于扩散模型的视频隐写方案，该方法保持文本到视频生成模型的结构和参数不变，无需额外训练即可集成隐写功能。

2）基于余弦系数实现秘密信息在隐空间的嵌入与提取，确保含密隐向量分布符合独立高斯分布，从而增强了抗隐写分析的安全性。

3）实验表明，本文方法生成的含密视频在视觉质量上与正常生成的视频几乎无差异，且在高嵌密容量下，仍表现出优异的提取准确性、鲁棒性和安全性。

1 预备知识

1.1 去噪扩散概率模型

去噪扩散概率模型（denoising diffusion probabilistic model，DDPM）由 Ho 等人（2020）提出，是一种基于正向加噪声和反向去噪声确定性过程的生成模型，可以离散化为多个时间步 T 来完成。

在正向扩散过程中，逐步添加噪声将数据分布 $q(\boldsymbol{x}_0)$ 转化为标准高斯分布。给定任意时间步 t，正向扩散的条件概率分布定义为

$$q\left(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}\right)=N\left(\boldsymbol{x}_t;\sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1},\beta_t\boldsymbol{I}\right) \tag{1}$$

式中，β_t 是一个预定义的参数，$\boldsymbol{I}$ 表示单位协方差矩阵。进一步，正向扩散的多步计算可以通过重参数化采样公式直接从 $\boldsymbol{x}_0$ 推导到任意时间步 t

$$\begin{aligned}&q\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right)=N\left(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0,\left(1-\bar{\alpha}_t\right)\boldsymbol{I}\right),\\&\alpha_t=1-\beta_t,\bar{\alpha}_t=\prod_{s=1}^{t}\alpha_s\end{aligned} \tag{2}$$

式中，$\bar{\alpha}_t$ 表示经过 t 步之后，累计保留下来的原始信号比例，α_s 是从第 1 步到第 t 步每一步的信号保留系数。

根据 $t>1$ 时分布 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ 的均值和方差，得到反向过程中一个时间步的计算式

$$p_\theta\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t\right)=N\left(\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(\boldsymbol{x}_t,t\right)\right),\sigma_t^2\boldsymbol{I}\right)\approx q\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0\right) \tag{3}$$

式中，ε_θ 表示训练好的噪声预测模型。使得标准高斯噪声通过逐步采样，生成接近原始数据分布的样本 $\boldsymbol{x}_0$。

1.2 文本到视频生成

文本到视频生成（text-to-video，T2V）旨在根据文本描述自动生成与之匹配的动态视频，其关键在于理解文本中的场景、物体和动作，并生成连续且一致的视觉帧。其中一种常见做法是在文本到图像生成（text-to-image，T2I）模型的基础上扩展结构，通过引入时间建模机制以支持视频序列的生成。

在 T2V 任务中，为适配视频输入的数据维度，通常需要对网络结构进行膨胀。标准 T2I 模型处理的是二维图像，而 T2V 输入为 5 维视频张量（b，c，f，h，w）。一种常用方式是将时间维度 f 展开至批量

维度 b，实现逐帧处理；在时序建模阶段，再将空间维度信息合并至批量维度，在时间维度上整合特征信息。

此外，为生成连贯的视频，需要引入时间注意力机制以捕捉帧间依赖，确保时间维度上的一致性。具体而言，模型在扩散过程的多个阶段引入时间注意力层，将时间特征与空间特征联合处理，构建全局一致的动态表示，如图 1 所示。

设帧特征为 $\{z^1,\cdots,z^f;\ z^i\in\mathbf{R}^{(b\times h\times w)\times c}\}$，其中 f 为视频帧数，注意力输出计算为

$$z_{\text{out}}=Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{c}\cdot\boldsymbol{V}) \tag{4}$$

式中，$\boldsymbol{Q}=\boldsymbol{W}^{Q}z,\boldsymbol{K}=\boldsymbol{W}^{K}z,\boldsymbol{V}=\boldsymbol{W}^{V}z$ 是 3 个独立的映射。

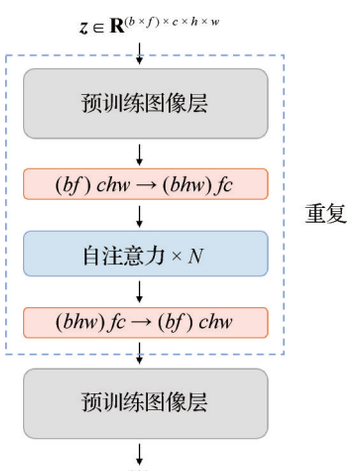

图 1　T2V 模型中的时间注意力机制

Fig. 1　Temporal attention mechanism in T2V model

本研究选用 AnimateDiff（Guo 等，2024）作为基础生成模型，主要原因在于其结构特性与隐写任务天然契合。具体而言，AnimateDiff 构建于扩散模型框架之上，具备正向扩散与逆向还原的对称结构，能够很好地适配隐写任务中“嵌入—提取”的双向过程。AnimateDiff 的优点在于支持图文联合控制，增强语义一致性；同时结合 MotionLoRA 技术，实现高适配性与轻量部署，便于多场景应用。

2　方　法

2.1　方案概述

本文提出的秘密信息隐藏与提取框架如图 2 所示，主要包括秘密信息的嵌入和提取两个过程。在嵌入阶段，发送方利用参考图像和文本提示词作为指导条件，在视频生成过程中嵌入秘密信息，生成含密视频，在提取阶段，接收方无需参考图像或文本提示词的指导，通过视频生成模型的逆向采样过程恢复含密高斯噪声，并进一步提取其中的秘密信息。具体流程如下所述。

2.2　秘密信息嵌入过程

秘密信息嵌入过程如图 2(a)所示，主要包括秘密信息预处理、二维逆离散余弦变换（inverse discrete cosine transform，IDCT）和正向采样视频生成 3 个环节。秘密信息嵌入过程的伪代码如算法 1 所示。

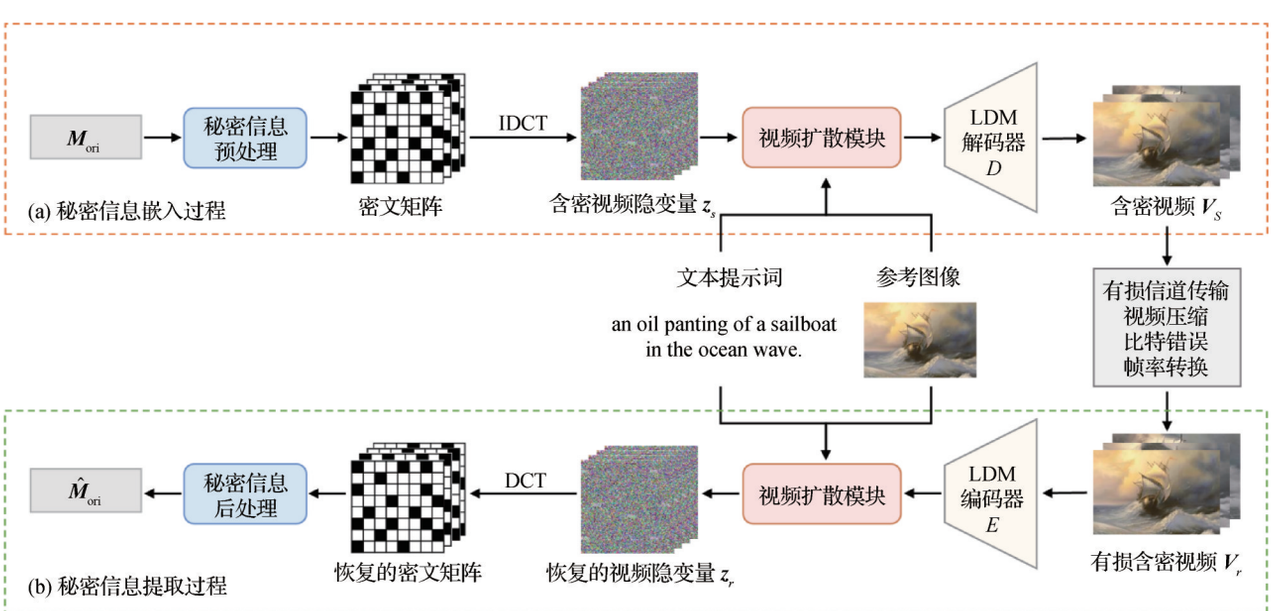

图 2　生成式视频隐写框架图

Fig. 2　Overview of generative video steganography framework ((a)secret information embedding; (b)secret information extracting)

算法1秘密信息嵌入过程

输入：秘密信息$\boldsymbol{M}_{\text{ori}}$、参考图像$\boldsymbol{I}_{\text{ref}}$、文本提示词$\boldsymbol{y}$、密钥$\boldsymbol{K}_{\text{enc}}$；

输出：含密视频$\boldsymbol{V}_s$；

1）$\boldsymbol{M}_{\text{repeat}} \leftarrow \boldsymbol{M}_{\text{ori}}$　　//秘密信息重复嵌入；

2）$\boldsymbol{M} \leftarrow \phi(\boldsymbol{M}_{\text{repeat}}, \boldsymbol{K}_{\text{enc}})$　　//生成密文；

3）$\boldsymbol{M}_{\text{reshape}} \leftarrow reshape(\boldsymbol{M})$；

4）$z_s \leftarrow latent_reshape(\text{IDCT_2D}(\boldsymbol{M}_{\text{reshape}}))$；

5）$\boldsymbol{V}_s \leftarrow D(VS_generation(z_s, \boldsymbol{I}_{\text{ref}}, \boldsymbol{y}))$。

秘密信息预处理的作用是将具有语义信息的原始秘密信息$\boldsymbol{M}_{\text{ori}}$转换为符合二项分布的密文矩阵，作为二维逆离散余弦（IDCT_2D）变换的输入。为了提高秘密信息的提取准确率，可以将原始秘密信息重复多次嵌入。具体而言，首先将原始秘密信息$\boldsymbol{M}_{\text{ori}}$重复$n$次，并对剩余容量填充随机比特串$\boldsymbol{R}$，生成$\boldsymbol{M}_{\text{repeat}}$

$$\boldsymbol{M}_{\text{repeat}} = \boldsymbol{M}_{\text{ori}}^{n} + \boldsymbol{R} \tag{5}$$

随后，发送方使用密钥为$\boldsymbol{K}_{\text{enc}}$的流加密算法$\phi$，对$\boldsymbol{M}_{\text{repeat}}$进行加密，生成符合二项分布的密文$\mathbf{M}$，即

$$\boldsymbol{M} = \phi(\boldsymbol{M}_{\text{repeat}}, \boldsymbol{K}_{\text{enc}}),\ \boldsymbol{M} \sim B(l, 0.5), 0 \leqslant l \leqslant L \tag{6}$$

式中，B表示二项分布，l表示密文的长度，L表示单个视频可以传输的最大密文长度。为了将符合二项分布的$\boldsymbol{M}$转换到服从高斯分布的隐变量z_s，需要将$\boldsymbol{M}$的形状调整为$C_r \times N \times N$，并将其值域映射到[−1, 1]，具体为

$$\boldsymbol{M}_{\text{reshape}} = secret_reshape(\boldsymbol{M}) \times 2 - 1 \tag{7}$$

二维逆离散余弦变换环节的作用是将符合二项分布的密文矩阵$\boldsymbol{M}_{\text{reshape}}$转换为接近独立高斯分布的噪声$z_s$，如式（8）所示。在本文中，基于IDCT的正交性，实现二项分布到高斯分布的变换。其理论基础在于：使用正交矩阵对二项分布矩阵进行线性变换，可以在数学上逼近标准高斯分布（Hu 等，2024）。与 Hu 等人（2024）的方法不同，该方法在每次隐写时均需生成随机正交矩阵，并在发送端与接收端之间共享密钥以保证正确解码，这在某些应用场景下引入了额外的密钥管理和传输开销。与 Hu 等人（2024）的方法不同，本文采用具有固定结构的 IDCT 正交矩阵，因此无需额外的密钥生成与传输。这样的设计大大简化了隐写系统的实现，提升了应用部署的便利性。

具体而言，在对$\boldsymbol{M}_{\text{reshape}}$使用IDCT变换的过程中，$\boldsymbol{M}_{\text{reshape}}$矩阵中的每个元素通过一组正交的余弦基函数加权，实现正交变换，得到对应位置的结果z_{kij}。其中，z_{kij}表示z_s第k帧中位置为(i, j)的元素值。实验结果（详见 3.2.5 节）进一步验证，隐变量z_s中的元素均符合高斯分布，且元素之间互相独立，经过维度调整后能够满足生成模型的输入要求。

$$z_s = latent_reshape(\text{IDCT_2D}(\boldsymbol{M}_{\text{reshape}})) \tag{8}$$

视频生成环节的作用是使用接近独立高斯分布的含密噪声z_s生成含密视频。具体而言，在视频生成过程中，从含密隐变量z_s开始，借助文本提示词$\boldsymbol{y}$和参考图像$\boldsymbol{I}_{\text{ref}}$这两个参考信息，利用视频扩散模块经过$T$步迭代去噪生成$z_0$。视频扩散模块的详细实现见 2.4 小节。最终，使用隐空间扩散模型（latent diffusion model，LDM）（Rombach 等，2022）中的解码器D，将z_0解码到图像空间，生成含密视频$\boldsymbol{V}_s$

$$\boldsymbol{V}_s = D(VS_generation(z_s, \boldsymbol{I}_{ref}, \boldsymbol{y})) \tag{9}$$

式中，z_s的维度为$C_r \times f \times h \times w$，$\boldsymbol{V}_s$的维度为$C \times f \times H \times W$，$H = h \times r$，$W = w \times r$，$r$是U-Net的下采样因子。

2.3　秘密信息提取过程

秘密信息提取过程如图2（b）所示，主要包含逆向采样恢复噪声、二维离散余弦变换（DCT_2D）和秘密信息后处理3个环节。

逆向采样恢复噪声的作用是将含密视频恢复到高斯噪声。具体而言，接收方在收到经过有损传输的含密视频$\boldsymbol{V}_r$后，利用LDM的编码器E将视频编码到隐空间$\hat{z}_0$。随后，无需参考信息，通过T步的迭代加噪过程即可还原得到z_r，即

$$z_r = VS_inversion(E(\boldsymbol{V}_r)) \tag{10}$$

二维离散余弦变换的作用是将z_r还原回密文矩阵。这一步依赖于DCT和IDCT变换的可逆性，为准确提取秘密信息提供了基础，具体为

$$\hat{\boldsymbol{M}}_{\text{reshape}} = \text{DCT_2D}(latent_reshape(z_r)) \tag{11}$$

式中，对于输入z_r，首先需要将其转换到$C_r \times N \times N$的维度，再进行DCT变换。

秘密信息后处理的作用是对恢复出的密文解密，并对n次恢复结果进行投票融合，以还原最终的原始秘密信息。具体而言，$\hat{\boldsymbol{M}}_{\text{reshape}}$的取值为−1或1，需先通过式（12）的计算，转换到[0, 1]的范围内以表示二进制比特，即

$$\hat{\boldsymbol{M}} = secret_reshape\left(\left[\frac{\hat{\boldsymbol{M}}_{\text{reshape}} + 1}{2}\right]\right) \tag{12}$$

然后，使用双方共享的流解密算法 ψ 和解密密钥 $\boldsymbol{K}_{\mathrm{dec}}$ 解密，得到提取出的密文矩阵

$$\boldsymbol{M}_{\mathrm{ext}} = \psi(\hat{\boldsymbol{M}}, \boldsymbol{K}_{\mathrm{dec}}) \tag{13}$$

式(12)中，[·]表示四舍五入到最接近的整数。因嵌入过程重复了 n 次原始秘密信息，所以每位秘密信息 $\boldsymbol{M}_{\mathrm{ori}}$ 中的第 i 位对应 $\boldsymbol{M}_{\mathrm{ext}}$ 中的 n 位信息。本文规定，当 n 位中超过一半为 1 时，$\boldsymbol{M}_{\mathrm{ori}}$ 的该位置取值为 1，否则为 0，具体为

$$\hat{\boldsymbol{M}}_{\mathrm{ori}} = \left[mean(reshape(\boldsymbol{M}_{\mathrm{ext}})) \right] \tag{14}$$

式中，*reshape* 的作用是将 $\boldsymbol{M}_{\mathrm{ext}}$ 转换为 $n \times l_{\mathrm{ori}}$ 的维度，l_{ori} 是 $\boldsymbol{M}_{\mathrm{ori}}$ 的长度。然后在第 1 个维度求平均，得到 l_{ori} 位恢复出的秘密信息 $\hat{\boldsymbol{M}}_{\mathrm{ori}}$。

2.4 视频扩散模块

由图 2 可知，本文设计的方案在嵌入和提取秘密信息的过程中都用到了视频扩散模块。视频扩散模块的结构如图 3 所示，其作用是在参考图像 $\boldsymbol{I}_{\mathrm{ref}}$ 和文本提示词 $\boldsymbol{y}$ 的条件指导下，使用 U-Net 网络预测当前时间步的噪声，实现去噪生成视频或加噪恢复噪声，其中，在加噪恢复噪声的过程中，可将参考图像与文本提示词均设为空，以实现无条件的噪声重建。

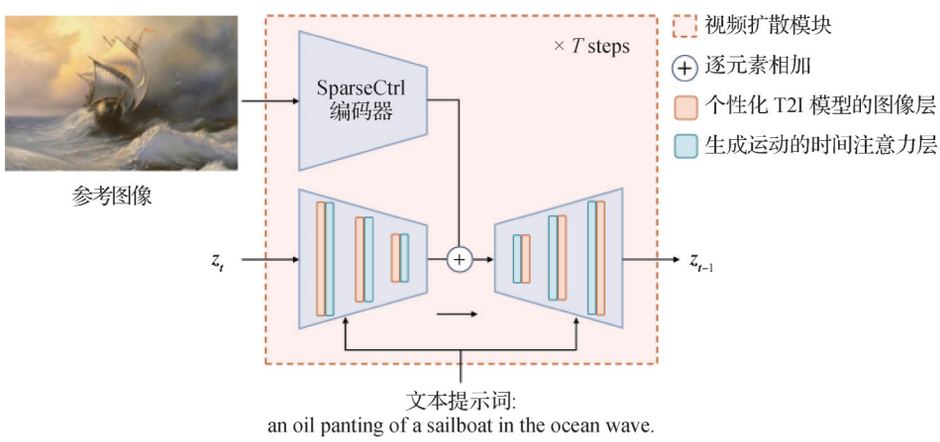

图 3　视频扩散模块

Fig. 3　Diffusion-based video sample module

具体而言，U-Net 网络中主要包含图像处理层和运动生成层。图像处理层为 T2I 的模块，利用注意力机制，可以将隐向量与文本提示词的特征融合，生成符合文本控制的图像。运动生成层的本质是时间维度上的注意力机制，这一步增加了不同视频帧之间的联系，有助于生成连贯的视频。

为使生成的视频更符合用户需求，还引入 SparseCtrl 网络（Guo 等，2024），支持用户在输入文本提示的基础上，输入参考图像，进一步约束生成视频的内容。SparseCtrl 提供了对视频关键帧的条件控制能力（例如：指定第 1 帧图像作为生成基础）。相比于文本到图像生成（T2I）模型中的条件控制网络 ControlNet（Zhang 等，2023a），SparseCtrl 专为视频生成任务设计，设定编码器部分的输入只有条件特征，同时在编码器内部增加了时间维度的注意力机制，从而实现条件信息在视频帧之间的有效传播。

3　实　验

3.1　实验设置

1）实现细节。本文设计的基于扩散模型的视频隐写方案可以嵌入到 T2V 架构中。实验中，以 Guo 等人（2024）提出的 AnimateDiff 模型为基础，开展相关实验。由于该模型生成的视频固定为 16 帧，因此本文生成的含密视频同样为固定的 16 帧。具体而言，使用在 WebVid-10M 数据集（Bain 等，2021）上训练过的 AnimateDiff 模型作为视频生成的核心框架。在实验中，设置隐向量的维度为 4 × 16 × 32 × 48，分别对应隐空间向量的通道数、帧数以及长和宽。此外，将 U-Net 的下采样因子 r 设置为 8，得到分辨率为 256 × 384 像素、长度为 16 帧的含密视频。所有的实验均基于 PyTorch 框架，在 4 张 NVIDIA RTX 4090 GPU 上完成。需要说明的是，本文方法具有良好的

可扩展性，若采用其他视频生成模型，也可根据其隐空间结构调整嵌密向量的维度与位置，从而适配不同架构的视频生成流程。

2）评估数据集。本文方法直接使用预训练完成T2V模型，无需微调或额外训练。因此，使用WebVid-10M数据集中的测试集部分构建评估数据集。具体而言，对于WebVid-10M测试集中每个视频，选取第1帧作为参考图像，并使用对应的视频文本描述作为文本提示词。最终，构建了包含100个文本—图像对的评估数据集。

3）对比方法。本文选择了REVMark（robustness enhanced video watermarking）（Zhang等，2023b）和RoGVSN（robust generativevideo steganography network）（Mao等，2024）作为对比实验。其中，REVMark是一种基于深度学习的鲁棒视频水印方案，主要思想是在训练过程中加入H.264视频压缩模拟层，增加水印对H.264视频压缩的鲁棒性。RoGVSN在利用换脸视频换脸修改视频语义特征的过程嵌入秘密信息，并且使用模拟视频处理的攻击层训练秘密信息提取器，增加隐写方案的鲁棒性。

4）评估指标。实验中使用了4个评估指标对本文方法和对比方法进行评估：视觉质量、嵌密容量、提取准确率和统计安全性。每个指标的具体描述和实验结果在3.2小节中详细介绍。

3.2 实验结果

3.2.1 视觉质量评估

视觉质量是衡量隐写算法感知隐蔽性的重要指标。图4展示了本文方法生成视频的定性评估结果。图中包含4个含密视频示例，每行的第1幅图像为生成视频所需的参考图像$\boldsymbol{I}_{\text{ref}}$，下方的文本为对应的视频文本提示词$\boldsymbol{y}$，每行的第2～5幅图像分别展示了含密视频的第5、9、13和16帧。从图中可以看出，生成的视频涵盖了多种场景，包括花朵、风景、人物以及动物等。生成结果很好地结合了参考图像和文本提示信息，展现了不同场景中高质量的动态视觉效果。本文方法生成的含密视频具有较高视觉质量，隐写过程几乎不影响视频视觉效果。

由于本文方法是一种在视频生成过程中隐藏秘

图4 使用不同文本提示词和参考图像生成的含密视频示例图（第1列是参考图像，其余4列是不同时刻的视频帧）

Fig. 4 Exampled stego videos generated using diverse text prompts and reference images
(the first column represents the reference image, the rest columns represent different frames in stego video)

密信息的方法，无需真实视频参与，因此没有任何视频作为参考视频，不能使用基于相似性度量的通用工具，如PSNR（peak signal-to-noise ratio）、SSIM（structural similarity）、VMAF（video multimethod assessment fusion）来衡量。因此，本文中使用FVD（Fréchet video distance）（Unterthiner 等，2019）和CLIP（contrastive language-image pre-training）分数（Radford 等，2021）这两个无参考视频质量评价工具进行衡量。

FVD是一种用于评估生成视频质量和多样性的指标。其核心原理是利用预训练的I3D网络（inflated 3D ConvNet）（Carreira 和 Zisserman，2017）提取视频的空间和时间特征，将生成视频与真实视频的特征分布表示为高维正态分布（由均值和协方差描述），然后计算这两个分布之间的 Fréchet 距离，以衡量生成视频与真实视频的接近程度。FVD 结合了帧内的视觉信息和帧间的时序一致性，已成为视频生成任务中广泛使用的评估指标。

CLIP分数是一种基于CLIP模型的评价指标，用于衡量图像和文本之间的语义一致性。本文关注3个关键维度：文本对齐、域相似性和视频一致性。

文本对齐度用于衡量生成视频与输入文本提示之间的语义一致性。具体方法为：提取生成视频中每一帧的 CLIP 图像特征，与文本提示的 CLIP 文本特征计算余弦相似度，并对所有帧的相似度取平均，得到文本对齐得分 S_{text}，计算为

$$S_{\text{text}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\boldsymbol{v}_i \cdot \boldsymbol{t}}{\|\boldsymbol{v}_i\| \cdot \|\boldsymbol{t}\|} \tag{15}$$

式中，$\boldsymbol{v}_i$表示第i帧视频的CLIP图像特征，$\boldsymbol{t}$表示提示词的CLIP文本特征，N表示视频帧数。

域相似性用于衡量生成视频与参考图像在视觉风格上的一致性。该指标计算每帧视频与参考图像之间的 CLIP 图像特征余弦相似度，并对所有帧的相似度进行平均，得到域相似性得分 S_{domain}，计算为

$$S_{\text{domain}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\boldsymbol{v}_i \cdot \boldsymbol{r}}{\|\boldsymbol{v}_i\| \cdot \|\boldsymbol{r}\|} \tag{16}$$

式中，$\boldsymbol{v}_i$表示第i帧视频的CLIP图像特征，$\boldsymbol{r}$代表参考图像的CLIP图像特征。

视频一致性则衡量生成视频帧间的时序连贯性。具体地，首先提取所有帧的 CLIP 图像特征，计算每对相邻帧之间的余弦相似度，再计算其相似度差值并取平均，得到视频一致性得分 S_{con}，具体计算为

$$S_{\text{con}} = 1 - \frac{1}{N-1}\sum_{i=1}^{N-1}\left(1 - \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_{i+1}}{\|\boldsymbol{v}_i\| \cdot \|\boldsymbol{v}_{i+1}\|}\right) \tag{17}$$

式中，$\boldsymbol{v}_i$表示第i帧视频的CLIP图像特征。

本文在Web10-Vid数据集上，计算了T2V模型生成的视频以及T2V模型生成的含密视频的FVD指标，并将得到的FVD结果与对比方法进行比较。由于对比方法并非基于 T2V 架构，故无法在其上计算 CLIP 文本对齐的相关指标。具体实验结果如表1所示。

表1　视频质量评估

Table 1　Video quality assessment

	FVD ↓	CLIP分数 ↑		
		文本对齐	域相似性	视频一致性
REVMark	1 352.99	–	0.943 2	0.976 3
RoGVSN	2 215.28	–	0.931 4	0.986 1
生成视频	1 463.73	0.291 6	0.774 6	0.932 4
含密视频	1 418.23	0.291 8	0.765 5	0.931 2

注："–"表示该方法没有对应的指标，"↓"表示指标数值越小越好，"↑"表示指标数值越大越好。

本文方法生成的含密视频在 FVD 指标上与原始 T2V 模型生成的视频仅存在微小差距，说明嵌入秘密信息后的视频在整体视觉质量上与正常生成视频基本一致。在CLIP分数的3个维度，计算出的分数都非常接近。这些实验结果证实，本文方法在嵌入秘密信息时不会大幅降低原始的视频生成质量。

3.2.2　嵌密容量

嵌密容量是指在秘密信息可以被准确提取的前提下，能够隐藏在载体视频中的秘密信息的数量。本文选择每帧含密比特数（bits per frame，bpf）作为衡量视频嵌密容量的指标。为适应真实场景中有损传输的要求，本文方法选择了一个更稳定的隐空间来隐藏和提取秘密信息。在实验中，隐空间由 $4 \times 16 \times 32 \times 48 = 98\,304$ 个元素组成。根据秘密信息重复次数不同，生成视频的bpf也不同。当秘密信息的长度为 l 时，秘密信息的最大嵌入重复次数 n 为 $\left\lfloor \frac{98\,304}{l} \right\rfloor$，对应的含密视频的bpf为

$$bpf = \frac{l}{f} \tag{18}$$

式中，f为视频帧数。具体而言，本文对n为16、32、64共3种情况进行了实验，对应生成视频的嵌入容量分别为384 bpf，192 bpf和96 bpf。

3.2.3 提取准确率与鲁棒性

在实际传输过程中，视频通常会受到各种处理操作的影响。为评估本文方法在有损传输下的鲁棒性，对以下常见的视频处理操作进行了测试：

1）视频压缩。视频压缩是为降低带宽占用而广泛采用的技术。本文测试了主流的H.264和H.265压缩方法。H.265相较于H.264提升了压缩效率并降低了计算复杂度，更适用于超高清视频。两种压缩方式包括平均比特率（average bitrate，ABR）压缩和恒定速率因子（constant rate factor，CRF）压缩。ABR动态调整每帧比特率以达到目标平均比特率，CRF则通过量化参数控制帧细节以保持视觉质量。

2）比特错误。由于信道传输的不完美，接收端可能发生比特翻转，产生误码。这些误码可能在传输链路或解码中扩散，进一步加重错误影响。通常使用传输过程中出错的比特数占总比特数的比例衡量比特错误。

3）帧率转换。帧率转换指接收端的视频帧率与发送端不一致，可能导致视频模糊或卡顿，影响视觉质量。

在实验中，对每个视频处理都设置了两个强度等级，等级高者会导致更高的视频失真。表2展示了实现的具体细节。

针对生成的含密视频，在以下3种情况下进行了秘密信息提取准确率的定量评估，并与基线方法进行对比：1）不经过视频处理；2）经过等级1的视频处理；3）经过等级2的视频处理。实验结果如表3和表4所示。

表2 视频处理的实现细节

Table 2 Implementation details of video process

视频处理	参数	视频处理强度等级	
		1	2
H.264 ABR	码率比例因子	1/2	1/4
H.264 CRF	恒定码率因子	27	33
H.265 ABR	码率比例因子	1/2	1/4
H.265 CRF	恒定码率因子	27	33
比特错误	比特错误率	1/100 000	1/50 000
帧率转换	视频帧率	8	5

表3 视频处理等级为1的情况下，秘密信息提取准确率比较

Table 3 Comparison results on extraction accuracy at video process level 1

方法	未处理	H.264 ABR	H.264 CRF	H.265ABR	H.265CRF	比特错误	帧率转换
REVMark	**0.999 7**	0.899 1	0.862 5	0.755 8	0.821 3	0.959 1	0.824 4
RoGVSN	0.973 7	0.950 2	0.959 9	0.925 7	0.927 9	0.963 3	0.932 9
本文方法(16)	0.979 6	0.961 7	0.931 0	0.960 2	0.921 0	0.943 8	0.921 8
本文方法(32)	0.994 7	0.984 4	0.975 1	0.979 4	0.949 0	0.985 4	0.954 3
本文方法(64)	0.997 6	**0.991 5**	**0.987 2**	**0.996 6**	**0.984 1**	**0.993 1**	**0.984 3**

注：括号中数字表示秘密信息重复次数。加粗字体表示每列最优结果。

表4 视频处理等级为2的情况下，秘密信息提取准确率比较

Table 4 Comparison results on extraction accuracy at video process level 2

方法	未处理	H.264 ABR	H.264 CRF	H.265ABR	H.265CRF	比特错误	帧率转换
REVMark	**0.999 7**	0.825 7	0.764 4	0.695 3	0.697 2	0.953 9	0.853 2
RoGVSN	0.973 7	0.880 5	0.891 5	0.836 1	0.799 0	0.948 6	0.949 9
本文方法(16)	0.979 6	0.927 9	0.866 9	0.928 3	0.853 6	0.931 2	0.927 0
本文方法(32)	0.994 7	0.963 2	0.923 7	0.953 8	0.893 2	0.953 6	0.964 0
本文方法(64)	0.997 6	**0.986 3**	**0.962 4**	**0.963 2**	**0.932 5**	**0.981 3**	**0.982 4**

注：括号中数字表示秘密信息重复次数。加粗字体表示每列最优结果。

REVMark（Zhang 等，2023b）的嵌密容量为 12 bpf，由于该方法针对 H. 264 视频压缩攻击进行了训练，因此在经过 H. 264 视频处理后，秘密信息提取准确率相对其他处理方式更高，但仍未达到 90%。

RoGVSN（Mao 等，2024）的嵌密容量最高可达 18 bpf，在视频处理等级 1 的情况下，秘密信息提取准确率超过 90%，但在等级 2 的情况下，提取准确率明显下降。

相比之下，本文方法在嵌密容量和鲁棒性上均优于上述对比方法。在实验中，针对 n 为 16、32、64 的 3 种设置进行了评估，生成视频的嵌密容量分别为 384 bpf、192 bpf 和 96 bpf。实验结果表明，即使在经历等级 2 的视频处理后，嵌密容量为 96 bpf 的含密视频依然能够保证秘密信息提取准确率达到 90% 以上，表明本文方法在嵌密容量和提取准确率方面具有显著优势，能够从有损的视频中成功提取秘密信息。

3.2.4 消融实验

图 5 展示了在不同重复嵌入次数下，秘密信息提取准确率的测试结果。当秘密信息的重复嵌入次数小于 20 时，增加重复次数能够显著提升秘密信息的提取准确率；当重复嵌入次数超过 20 时，提取准确率已稳定在 0.975 以上，进一步提升的效果不明显。因此，从嵌密容量与提取准确率的平衡角度来看，可以选择重复嵌入 20 次。

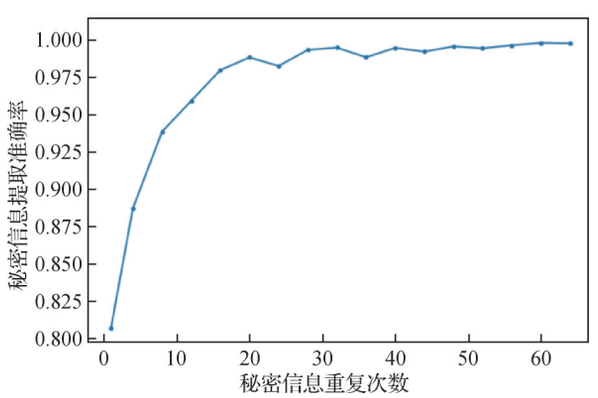

图 5　秘密信息重复嵌入次数的消融实验

Fig. 5　Ablation study on embeddings of secret information repeat times

3.2.5 安全性分析

由于本文是基于视频生成模型的隐写方案，从理论上分析，只要含密噪声和正常生成视频的隐变量服从相似的独立高斯分布，则本文的隐写算法能抗隐写分析检测。为验证这一点，设计了两组实验，从分布特性和元素独立性两个方面进行分析。首先，可视化了含密噪声的分布特性，以验证其是否符合高斯分布。图 6 展示了含密噪声 z_s 和正常高斯噪声 z_c 的分布对比。考虑到 z_s 中元素是有限个离散值，因此使用直方图绘制了 z_s 中各元素数值的概率分布。由图可见，z_s 的分布十分接近标准高斯分布 z_c，验证了 z_s 具备高斯分布的特性。其次，使用自相关函数检验 z_s 中元素的独立性，选择默认的滞后阶数为 50 进行计算。实验结果表明，z_s 在所有滞后阶数上的自相关性均接近于 0，说明 z_s 各元素之间独立。

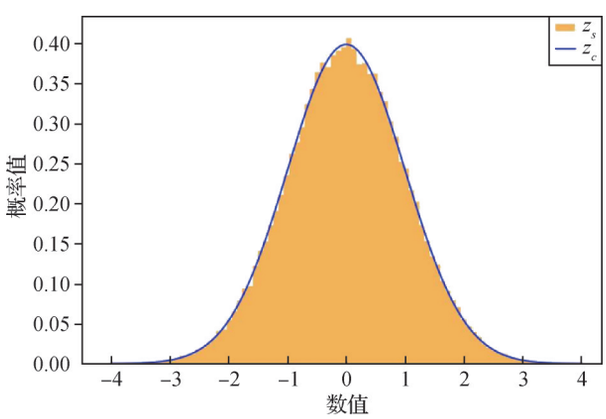

图 6　含密隐变量 z_s 与标准高斯分布 z_c 的分布对比

Fig. 6　Comparison of the distribution between the stego latent z_s and the standard Gaussian distribution z_c

此外，本文还通过隐写分析网络对方案的安全性进一步验证。具体而言，采用 3 种主流的视频隐写分析方法，包括 MVC（motion vector consistency）（Zhai 等，2020）、IDFB（in-loop deblocking filter block）（Zhang 等，2020）和 SUPERB（subtractive prediction error block）（Cao 等，2021），对生成视频和含密视频分别提取特征，然后使用分类器对视频分类，以评估隐写方案的抗检测能力。需要特别说明的是，MVC、IDFB 和 SUPERB 这 3 种隐写分析方法本质上并非端到端的视频分类网络，即它们本身不直接输出视频是否含有秘密信息的判定结果，而是作为特征提取工具使用。它们通过对视频编码结构、运动矢量、帧间/帧内预测等低层编码特性进行分析，提取能够表征隐写修改痕迹的统计特征向量。在这些方法的原始论文中，研究者均是先利用其方法提取特征，再将这些特征输入到常规分类器中进

行训练和测试，从而完成隐写检测。因此，本文在安全性分析实验中延续了这一标准流程：先用3种分析方法提取生成视频和含密视频的特征，再通过训练支持向量机（support vector machine，SVM）分类器评估隐写方案的抗检测能力。

具体而言，MVC基于宏块内子块运动矢量在隐写修改后趋于一致的现象，提取了一组反映运动矢量一致性的12维通用特征。IDFB针对H. 264视频隐写算法，通过分析修改变换系数对H. 264去块滤波操作的影响，提取亮度像素的平均修改量构建768维特征。SUPERB则通过构建3种类型的预测误差块差分，结合其一阶马尔可夫过渡概率和帧内预测模式变化，提取统计特征集合，用于隐写分析。

将上述3种方法提取的特征用于训练SVM分类器，并按照训练集与测试集8∶2的比例进行隐写分析测试。表5展示了各方法的性能。实验结果表明，本文方法相较于对比方法，在所有测试中均表现出较强的抗检测能力，进一步验证了所提隐写方案的安全性。

表5　隐写安全性定量分析

Table 5　Quantitative steganography security analysis

方法	MVC	IDFB	SUPERB
REVMark	0.444 4	0.450 0	0.641 0
RoGVSN	0.611 0	0.625 4	0.632 1
本文方法(16)	0.468 8	0.452 1	0.514 3
本文方法(32)	0.516 1	0.467 3	0.514 3
本文方法(64)	**0.500 0**	**0.492 5**	**0.500 0**

注：括号内数字表示秘密信息重复次数。加粗字体表示每列最优结果。

4　结　论

本文提出了一种基于扩散模型的生成式视频隐写方法，可在文本提示词和参考图像的指导下，在生成视频的过程中嵌入秘密信息，实现对生成视频内容的精准控制，保障隐蔽通信的行为安全性。同时，本文方法在提取秘密信息时无需参考图像与提示词文本，有利于实际应用。

实验结果表明，本文方法生成的视频在视觉质量上与正常生成的视频差异较小，并能够准确匹配文本提示词中的语义信息，充分体现了其在视觉隐蔽性方面的优越性。此外，本文方案在384 bpf的嵌密容量下，经过不同等级的视频攻击，秘密信息提取准确率均可保持在90%以上，表明本文方法在嵌密容量、秘密信息提取准确率以及隐写安全性分析等方面，均显著优于对比方法，展现出良好的综合性能。

但是，由于本文方法基于AnimateDiff模型，该模型生成的视频长度固定为16帧。若需生成更长的视频，则需替换为支持更长输出的T2V模型，并同步调整隐空间的高斯噪声维度，以适配新的结构。因此，未来工作将重点研究适用于有损通信环境的更鲁棒的视频隐写方法，探索针对长视频的生成式隐写算法，以进一步提升视频隐写技术的实用性和应用范围。

参考文献（References）

Bain M, Nagrani A, Varol G and Zisserman A. 2021. Frozen in time: a joint video and image encoder for end-to-end retrieval//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1708-1718 [DOI: 10.1109/ICCV48922.2021.00175]

Cao Y, Zhang H, Zhao X F and He X L. 2021. Steganalysis of H.264/AVC videos exploiting subtractive prediction error blocks. IEEE Transactions on Information Forensics and Security, 16: 3326-3338 [DOI: 10.1109/TIFS.2021.3078822]

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Chan C K and Cheng L M. 2004. Hiding data in images by simple LSB substitution. Pattern Recognition, 37(3): 469-474 [DOI: 10.1016/j.patcog.2003.08.007]

Dasgupta K, Mandal J K and Dutta P. 2012. Hash based least significant bit technique for video steganography (HLSB). International Journal of Security, Privacy and Trust Management, 1(1): 37-45

Dong Y, Jiang X H, Li Z H, Sun T F and He P S. 2023. Adaptive HEVC steganography based on steganographic compression efficiency degradation model. IEEE Transactions on Dependable and Secure Computing, 20(1): 769-783 [DOI: 10.1109/TDSC.2022.3144139]

Gopalan K. 2003. Audio steganography using bit modification//Proceedings of 2003 International Conference on Multimedia and Expo. Baltimore, USA: IEEE: I-629 [DOI: 10.1109/ICME.2003.1220996]

Guo Y W, Yang C Y, Rao A Y, Agrawala M, Lin D H and Dai B. 2025.

SparseCtrl: adding sparse controls to text-to-video diffusion models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 330-348 [DOI: 10.1007/978-3-031-72946-1_19]

Guo Y W, Yang C Y, Rao A Y, Liang Z Y, Wang Y H, Qiao Y, Agrawala M, Lin D H and Dai B. 2024. AnimateDiff: animate your personalized text-to-image diffusion models without specific tuning [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2307.04725.pdf

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851

Hu X X, Li S, Ying Q C, Peng W L, Zhang X P and Qian Z X. 2024. Establishing robust generative image steganography via popular stable diffusion. IEEE Transactions on Information Forensics and Security, 19: 8094-8108 [DOI: 10.1109/TIFS.2024.3444311]

Khare R, Mishra R and Arya I. 2014. Video steganography using LSB technique by neural network//Proceedings of 2014 International Conference on Computational Intelligence and Communication Networks. Bhopal, India: IEEE: 898-902 [DOI: 10.1109/CICN.2014.189]

Li Z H, Jiang X H, Dong Y, Meng L J and Sun T F. 2023. An anti-steganalysis HEVC video steganography with high performance based on CNN and PU partition modes. IEEE Transactions on Dependable and Secure Computing, 20(1): 606-619 [DOI: 10.1109/TDSC.2022.3140899]

Ma B, Li K, Xu J, Wang C P, Li J and Zhang L W. 2024. High-security image steganography with the combination of multiple competition and channel attention. Journal of Image and Graphics, 29(2): 355-368 (马宾, 李坤, 徐健, 王春鹏, 李健, 张立伟. 2024. 联合多重对抗与通道注意力的高安全性图像隐写. 中国图象图形学报, 29(2): 355-368) [DOI: 10.11834/jig.230134]

Mao X Y, Hu X X, Peng W L, Gan Z L, Qian Z X, Zhang X P and Li S. 2024. From covert hiding to visual editing: robust generative video steganography//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 2757-2765 [DOI: 10.1145/3664647.3681149]

Meng L J, Jiang X H, Sun T F, Zhao Z Y and Xu Q. 2024. A robust coverless video steganography based on the similarity of inter-frames. IEEE Transactions on Multimedia, 26: 5996-6011 [DOI: 10.1109/TMM.2023.3344357]

Mstafa R J, Elleithy K M and Abdelfattah E. 2017. A robust and secure video steganography method in DWT-DCT domains based on multiple object tracking and ECC. IEEE Access, 5: 5354-5365 [DOI: 10.1109/ACCESS.2017.2691581]

Nie Q K, Xu X B, Feng B W and Zhang L Y. 2018. Defining embedding distortion for intra prediction mode-based video steganography. Computers, Materials and Continua, 55(1): 59-70 [DOI: 10.3970/cmc.2018.055.059]

Pan N, Qin J H, Tan Y, Xiang X Y and Hou G M. 2020. A video coverless information hiding algorithm based on semantic segmentation. EURASIP Journal on Image and Video Processing, 2020: #23 [DOI: 10.1186/s13640-020-00512-8]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtually: PMLR: 8748-8763

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Suresh M and Sam I S. 2018. High secure video steganography based on shuffling of data on least significant DCT coefficients//Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems. Madurai, India: IEEE: 877-882 [DOI: 10.1109/ICCONS.2018.8662920]

Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M and Gelly S. 2019. Towards accurate generative models of video: a new metric and challenges [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1812.01717.pdf

Wang J, Chen K J, Zhang W M and Yu N H. 2023. Image processing network-inverted identifiable secure natural steganography. Journal of Image and Graphics, 28(3): 749-759 (王健, 陈可江, 张卫明, 俞能海. 2023. 面向可逆图像处理网络的可证安全自然隐写. 中国图象图形学报, 28(3): 749-759) [DOI: 10.11834/jig.220529]

Xue Y M, Zhou J, Zeng H, Zhong P and Wen J. 2019. An adaptive steganographic scheme for H.264/AVC video with distortion optimization. Signal Processing: Image Communication, 76: 22-30 [DOI: 10.1016/j.image.2019.04.012]

Yadav P, Mishra N and Sharma S. 2013. A secure video steganography with encryption based on LSB technique//Proceedings of 2013 IEEE International Conference on Computational Intelligence and Computing Research. Enathi, India: IEEE: 1-5 [DOI: 10.1109/ICCIC.2013.6724212]

Yu J W, Zhang X Y, Xu Y M and Zhang J. 2023. Cross: diffusion model makes controllable, robust and secure image steganography//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 80730-80743

Zhai L M, Wang L N and Ren Y Z. 2020. Universal detection of video steganography in multiple domains based on the consistency of motion vectors. IEEE Transactions on Information Forensics and Security, 15: 1762-1777 [DOI: 10.1109/TIFS.2019.2949428]

Zhang H, You W K and Zhao X F. 2020. A video steganalytic approach against quantized transform coefficient-based H.264 steganography

by exploiting in-loop deblocking filtering. IEEE Access, 8: 186862-186878 [DOI: 10.1109/ACCESS.2020.3030685]

Zhang L M, Rao A Y and Agrawala M. 2023a. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3813-3824 [DOI: 10.1109/ICCV51070.2023.00355]

Zhang X P, Qian Z X and Li S. 2016. Prospect of digital steganography research. Journal of Applied Sciences, 34(5): 475-489 (张新鹏, 钱振兴, 李晟. 2016. 信息隐藏研究展望. 应用科学学报, 34(5): 475-489) [DOI: 10.3969/j.issn.0255-8297.2016.05.001]

Zhang Y L, Ni J Q, Su W K and Liao X. 2023b. A novel deep video watermarking framework with enhanced robustness to H.264/AVC compression//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 8095-8104 [DOI: 10.1145/3581783.3612270]

Zhou X J, Peng W L, Yang B Y, Wen J, Xue Y M and Zhong P. 2022. Linguistic steganography based on adaptive probability distribution. IEEE Transactions on Dependable and Secure Computing, 19(5): 2982-2997 [DOI: 10.1109/TDSC.2021.3079957]

Zhou Z L, Su Y C, Li J, Yu K P, Wu Q M J, Fu Z J and Shi Y Q. 2023. Secret-to-image reversible transformation for generative steganography. IEEE Transactions on Dependable and Secure Computing, 20(5): 4118-4134 [DOI: 10.1109/TDSC.2022.3217661]

作者简介

李文琪，女，硕士研究生，主要研究方向为生成式视频隐写。
E-mail：wenqili23@m.fudan.edu.cn
钱振兴，通信作者，男，教授，主要研究方向为多媒体信息安全、隐写、数字水印、图像取证和密文域信号处理。
E-mail：zxqian@fudan.edu.cn
周科扬，男，硕士研究生，主要研究方向为视频隐写。
E-mail：zhouky24@m.fudan.edu.cn
胡宵宵，女，博士研究生，主要研究方向为生成式图像信息隐藏、多媒体安全。E-mail：23110240079@m.fudan.edu.cn
张新鹏，男，教授，主要研究方向为多媒体信息安全、AI安全和图像处理。E-mail：zhangxinpeng@fudan.edu.cn
李晟，男，副教授，主要研究方向为生物数据隐私保护、多媒体内容安全和图像识别。E-mail：lisheng@fudan.edu.cn