中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

《语音内容的可追溯保护：音频水印研究》

中科院自动化所
周俊佐 任勇

音频水印的概念

向语音中嵌入加密信息，并由相应的检测器从信号内容中解码还原信息

版权保护

语音合成的主动溯源

当局对语音合成内容进行监管备案

……

四部门联合发布《人工智能生成合成内容标识办法》

音频水印的特点：人耳不可察觉

图像水印：更偏向鲁棒性，允许并鼓励检测与验证

图像隐写：更偏向隐蔽性，信息保密

音频水印的评估

相互制约的属性

- 不可感知性(Imperceptibility)：信噪比与 PESQ等语音质量指标
- 容量(Capacity)：平均每秒声音可以嵌入的比特数，单位BPS (bit per second)
- 鲁棒性(Robustness)：解码出的比特序列与原始比特序列计算错误率BER (bit error rate), 平均各数位的准确率(accuracy rate)，ROC曲线下面积(AUC)，TPR@FPR=0.01等

满足“不可感知性”要求后，在鲁棒性与容量之间的取舍取决于最终的应用需求

音频水印的发展

传统语音水印基于专家知识，经验设计，泛化性和鲁棒性不足

神经语音水印

2023.12
WavMark

2024.3
TracebleSpeech

2024.8
WMCodec
……

2022.9
DNN-WM

2023.6
Dear

2023.9
Collaborator、
MaskMark

2023.12
Timbre WM

2024.3
AudioSeal

2024.4
Groot

2024.9
SSR-Speech
SynthID
…

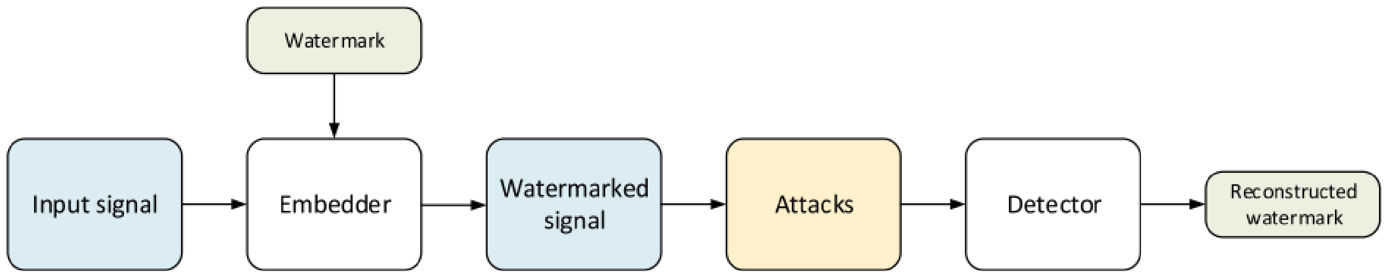
1：通用的事后音频水印

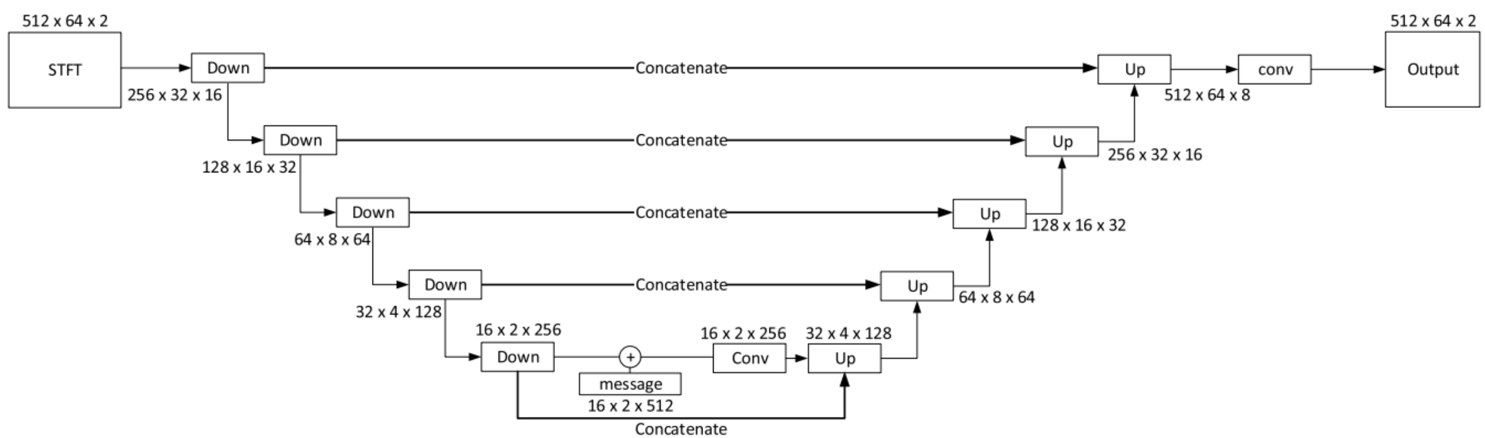
2：任务融合驱动的音频水印

3：开源模型音频水印

通用型水印

DNN-WM





1：STFT频域上执行嵌入

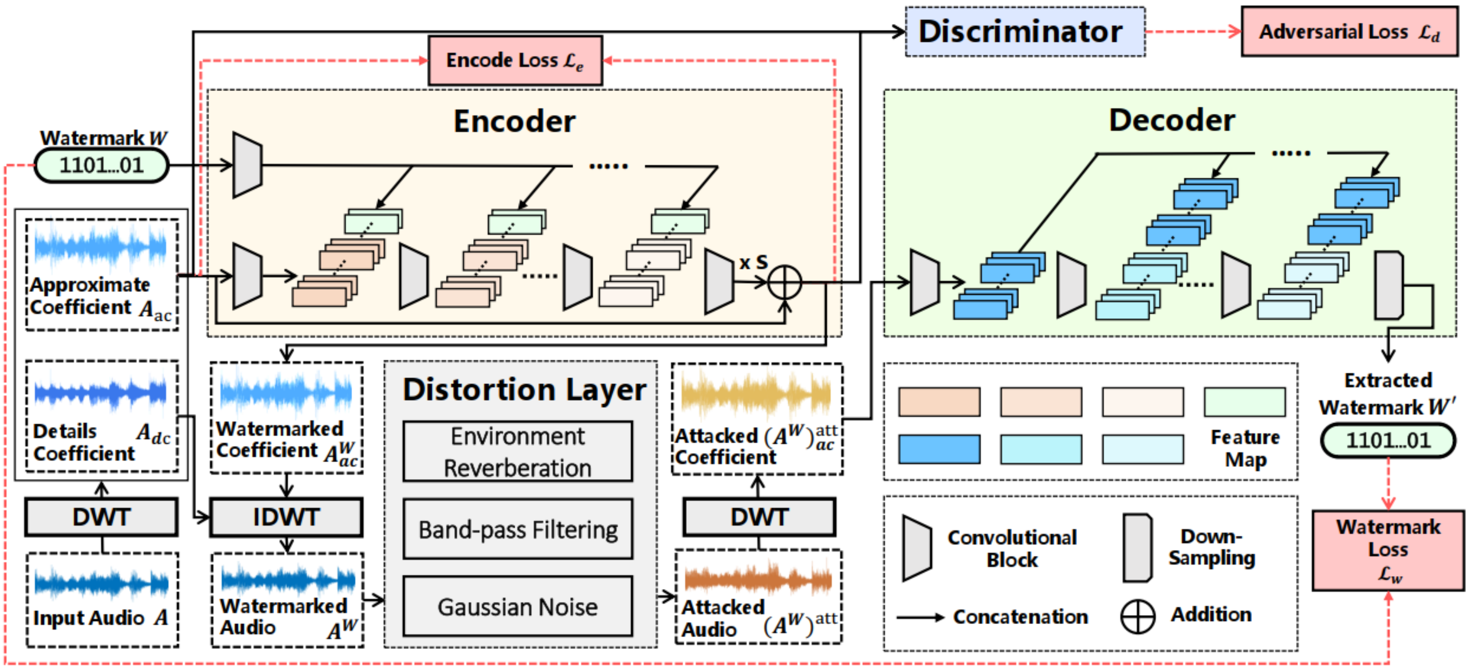
2：实现对三种攻击类型的鲁棒性（Dropout、随机噪声、高通滤波）

3：嵌入容量较低（2.5 bit / 2s）：

Pavlović K, Kovačević S, Djurović I, et al. Robust speech watermarking by a jointly trained embedder and detector using a DNN[J]. Digital Signal Processing, 2022, 122: 103381.

通用型水印

DeAR



1：DWT频域上执行嵌入

2：水印通过Encoder融入语音时采用残差设计，调整水印-语音比例

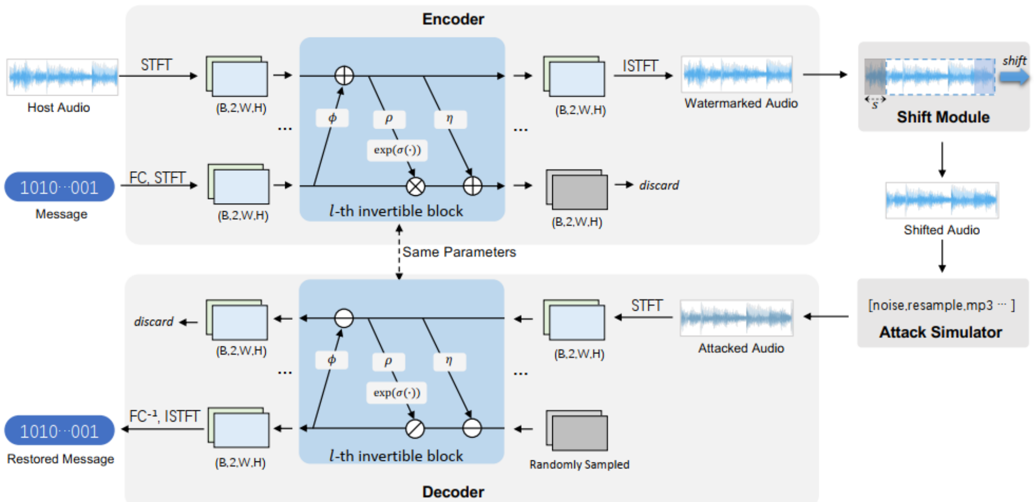
3：考虑音频转录环境作为模拟攻击

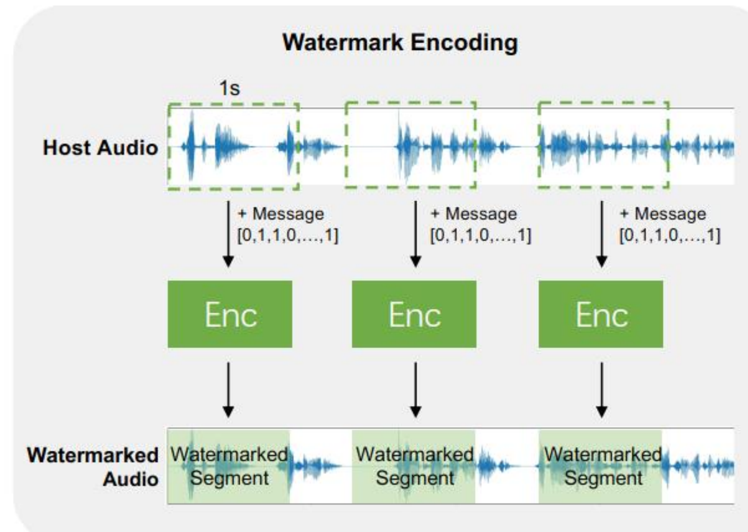
4：嵌入容量进一步提高(100bit / 11s)

Liu C, Zhang J, Fang H, et al. Dear: A deep-learning-based audio re-recording resilient watermarking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 13201-13209.

通用型水印

WavMark



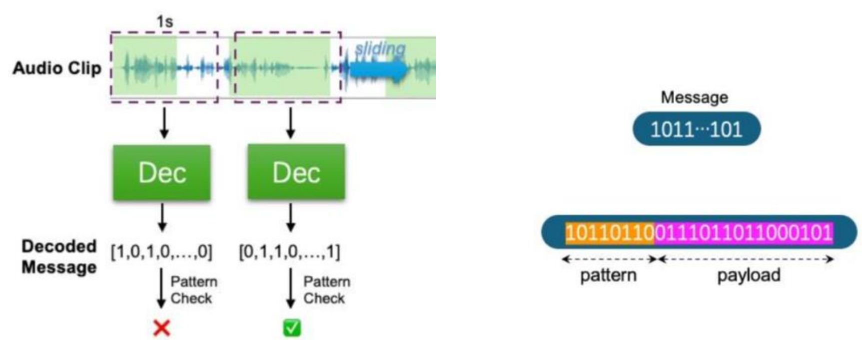


1：采用可逆网络的设计编码和解码:
- $y = f(x), x = f^{-1}(x)$

2：采用了9种模拟攻击
- 随机噪声、滤波器、重采样、幅度缩放、回声......

3：嵌入容量进一步提高(32bit / 1s)

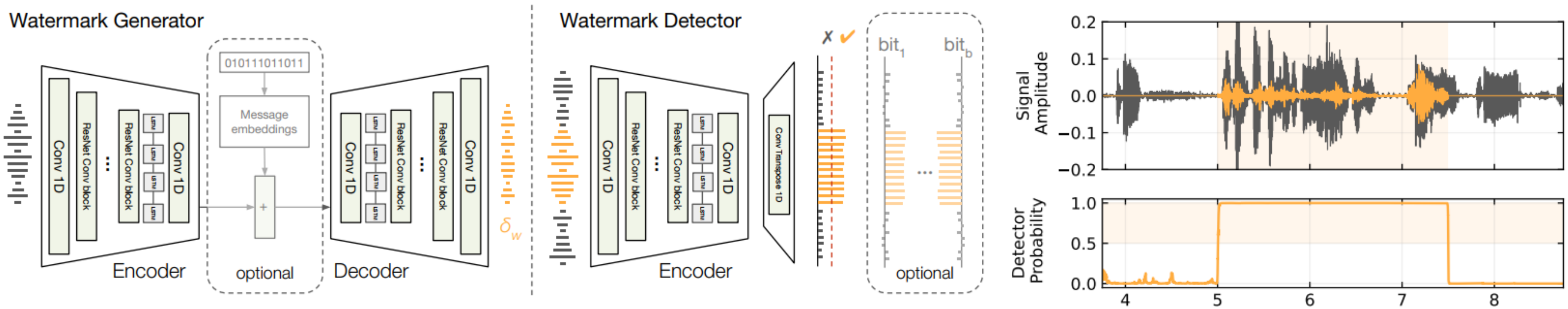
4：长语音下的水印段定位问题
- 滑动探测窗口暴力匹配, 兼顾定位和解码
- pattern(16bit) + payload(16bit)



Chen G, Wu Y, Liu S, et al. Wavmark: Watermarking for audio generation[J]. arXiv preprint arXiv:2308.12770, 2023.

通用型水印

AudioSeal



1：水印嵌入不涉及频谱

2：水印存在段的帧级别定位

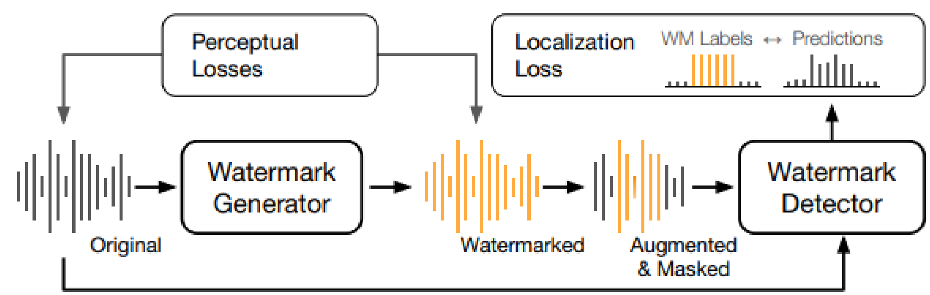
- 精度高达 1/16k 秒

3：水印检测与内容位提取的结构统一

4：仅需单次前向传播

- 整段音频的水印检测或内容提取无需滑动窗口

5：保持了嵌入容量 (16bit / 1s) 和鲁棒性



San Roman R, Fernandez P, Elsahar H, et al. Proactive Detection of Voice Cloning with Localized Watermarking[C]//ICML 2024-41st International Conference on Machine Learning. 2024, 235: 1-17.

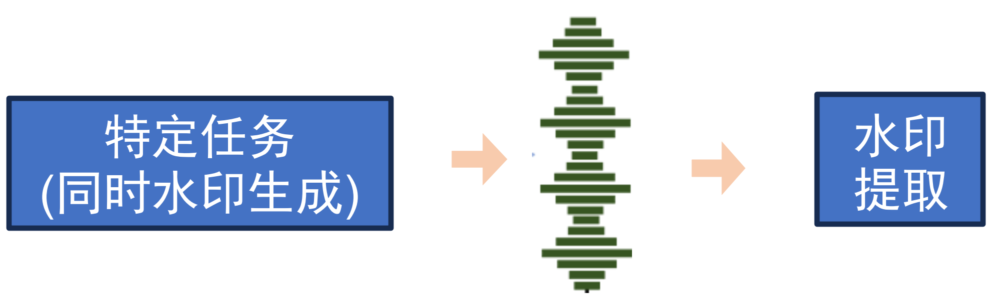
任务驱动型水印

通用水印是事后的、分阶段的、级联式的非端到端系统

合成后需要被溯源的语音、
需要被版权保护的音乐、
……

附加水印后的语
音波形

为什么会有任务驱动型的水印？



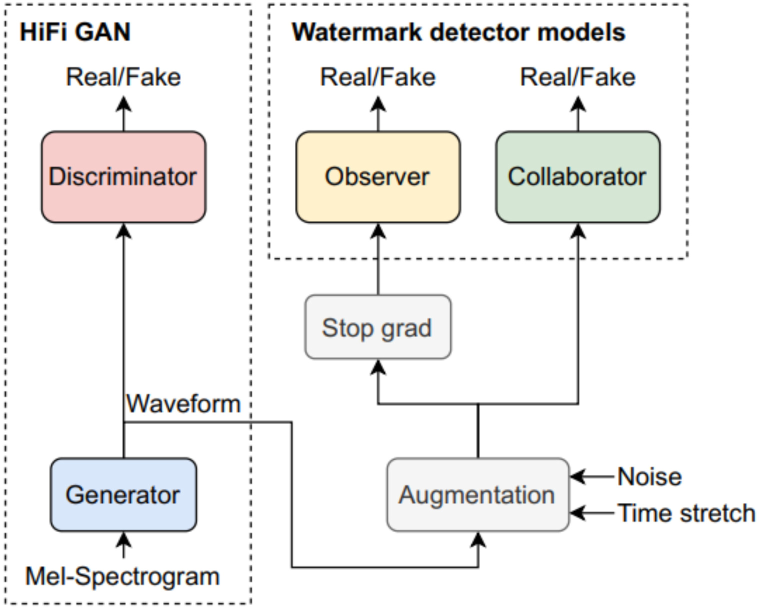
附加水印后的语
音波形

任务驱动型水印

Collaborator Watermarking

1：语音合成时强化真假标签的可检测性
水印标识纳入声码器训练

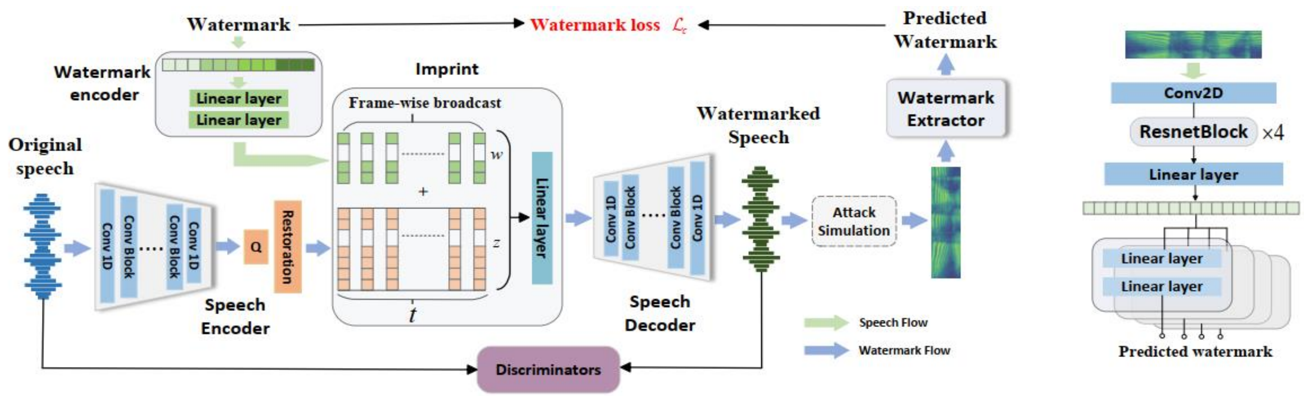
2：水印检测器直接采用语音鉴伪模型
标识仅反映真假，不涉及水印内容的还原



Juvela L, Wang X. Collaborative watermarking for adversarial speech synthesis[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 11231-11235.

任务驱动型水印

TraceableSpeech



(a) Overall architecture. (b) Watermark Extractor module.

Figure 1: *The first stage: Watermarking mechanism integrate into neural codec.*

1：语音合成时嵌入水印内容，提升不可感知性。

- 第一阶段：水印在Codec解码端侧与语音特征融合
- 第二阶段：VALL-E 语言模型合成语音

2： 逐时域帧广播水印内容

- 提供全时段保护，提升对于合成语音剪辑的鲁棒性
- 更灵活地支持可变时长的推理

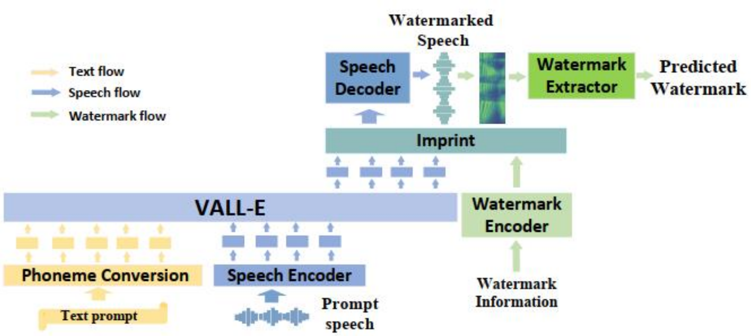


Figure 2: *The second stage: Watermarking mechanism integrate into language model of VALL-E.*

Zhou J, Yi J, Wang T, et al. TraceableSpeech: Towards Proactively Traceable Text-to-Speech with Watermarking[C]//Proc. Interspeech 2024. 2024: 2250-2254.

任务驱动型水印

Imperceptibility

Table 1: *Watermark Imperceptibility Metrics in Speech Reconstruction*

Model	PESQ ↑	STOI ↑	ViSQOL ↑
HiFicodec + WavMark(16bit)	3.197	0.947	3.880
TraceableSpeech(4@10)	**3.641**	**0.950**	**4.060**
TraceableSpeech(4@16)	3.569	0.948	3.985

[1] @ denotes the watermarking capacity. For example, 4@16 indicates 4-digit base-16, equivalent to the 16-bit capacity of WavMark used in the baseline. This annotation is applicable to other tables as well.

Table 2: *Speech Quality in Zero-Shot Speech Synthesis*

Model	WER(%) ↓	MOS ↑
VALL-E + WavMark(16bit)	10.80	3.554 ± 0.19
TraceableSpeech(4@10)	**9.61**	**3.959** ± 0.18
TraceableSpeech(4@16)	10.47	3.905 ± 0.17

- 波形重建实验和语音合成实验的不可感知性均获提升

Robustness

- 即使随机移除2/3的语音段落依旧能准确提取

Table 3: *Watermark extraction accuracy (%) under various attacks*

Model \ Attack	Respliscing	Normal	RSP-90	Noise-W35	SD-01	AR-90	EA-0315	LP5000
VALL-E + WavMark(16bit)	No	100.00	99.76	91.41	100.00	100.00	94.53	100.00
TraceableSpeech(4@10)	No	100.00	**100.00**	**100.00**	100.00	100.00	**100.00**	100.00
TraceableSpeech(4@16)	No	98.97	98.82	98.95	99.12	99.46	97.71	98.84
VALL-E + WavMark(16bit)	Once	91.10	91.46	63.53	95.95	93.61	88.58	89.66
TraceableSpeech(4@10)	Once	**100.00**	**100.00**	**100.00**	**99.90**	**100.00**	**100.00**	**100.00**
TraceableSpeech(4@16)	Once	100.00	99.82	99.83	98.78	99.50	99.57	99.62
VALL-E + WavMark(16bit)	Twice	76.65	77.74	49.14	79.47	85.46	68.19	75.32
TraceableSpeech(4@10)	Twice	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
TraceableSpeech(4@16)	Twice	99.58	99.20	99.58	99.56	99.00	99.65	98.83

Ours 16bits

[1] The respliscing column mean the times of respliscing attack

Flexibility and limitations

- 0.3s的语音片段负载4位64进制水印信息依旧可以恢复95%+

Table 4: *Watermark extraction accuracy (%) of larger capacity models under various speech durations (s)*

Model \ Duration	1.0	0.8	0.5	0.3	0.2	0.175	0.15	0.125	0.1
TraceableSpeech(4@32)	100.00	100.00	99.74	99.23	94.13	86.22	77.29	57.14	50.51
TraceableSpeech(4@64)	100.00	100.00	99.86	95.57	80.59	66.79	53.90	27.47	17.01

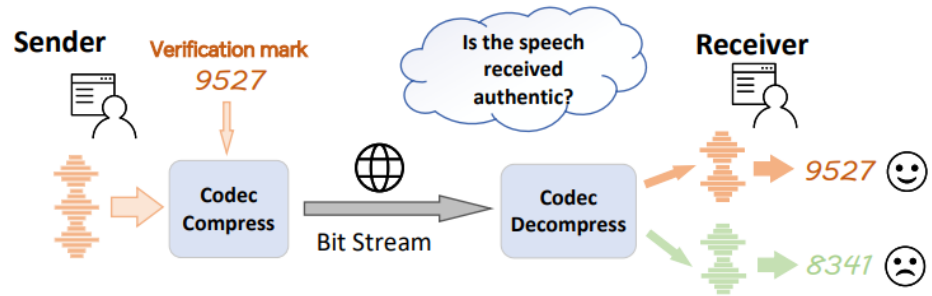
任务驱动型水印

WMCodec

1：语音Codec 传输前后的水印嵌入与提取

- 发送端在压缩语音前嵌入水印
- 接收端解压语音后依旧实现提取

Fig. 1. Example of Watermark as Verification Marking for Codec Protection

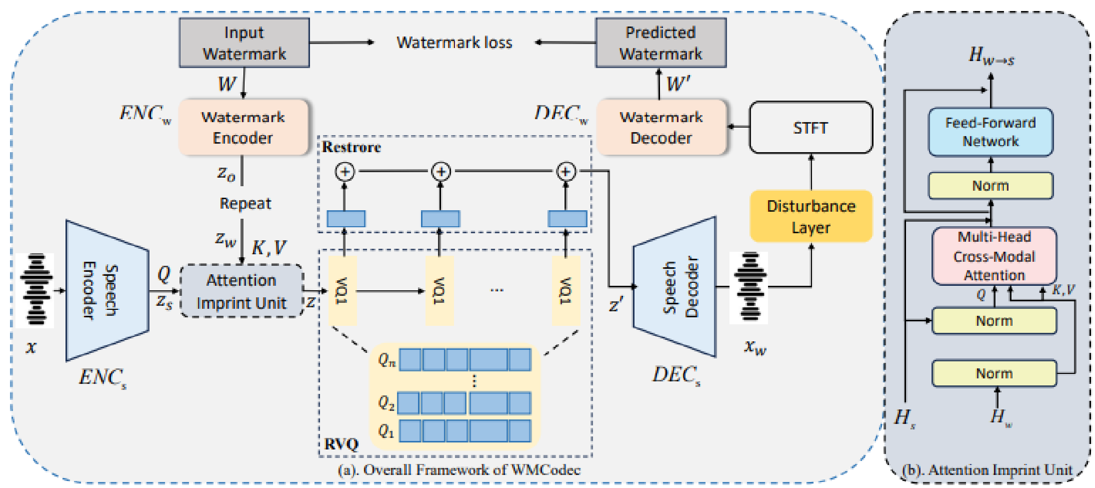


2：任务驱动端到端训练

• 过去的方法仅将Codec视为事后攻击的一种，或者语音合成的中间过称，水印机制并未处理量化器压缩引发的失真。

2： 水印迭代地Cross-Attention嵌入

- 过去的嵌入方法均基于cat或addition
- 消融实验证明：注意力的融合方式有损于不可感知性，但进一步促进了可提取性



(a). Overall Framework of WMCodec

(b). Attention Imprint Unit

Zhou J, Yi J, Ren Y, et al. WMCodec: End-to-End Neural Speech Codec with Deep Watermarking for Authenticity Verification[C]//ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE

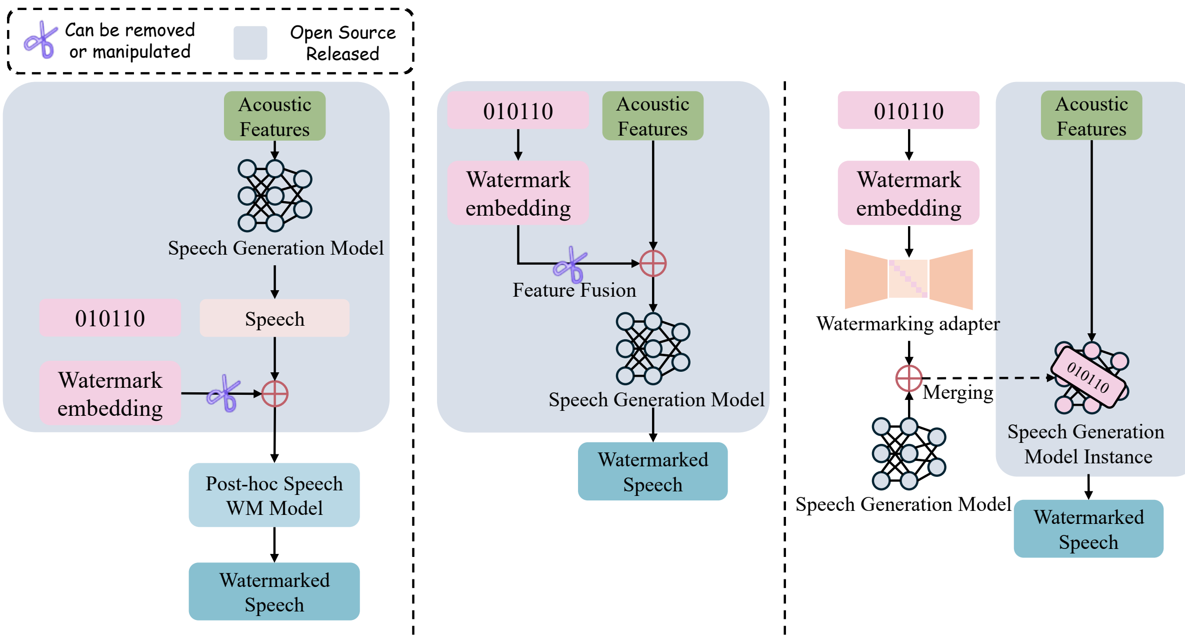
再谈音频水印的评估

- 不可感知性(Imperceptibility)：基本需求
- 容量(Capacity)：更大的追求
- 鲁棒性(Robustness)：域外泛化性与实用

模型参数级水印

为什么要做模型参数级水印？

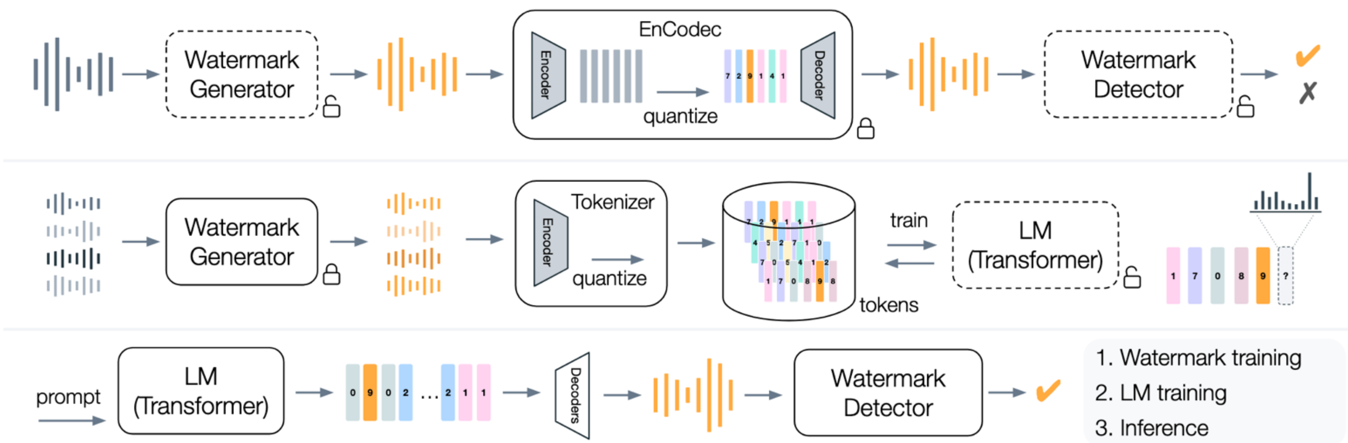
- Audio-level Watermarking (Post-Hoc Watermarking)
  - AudioSeal, WavMark, etc.
  - 在音频中添加水印
- Feature-level Watermarking
  - TraceableSpeech, WMCodec, etc.
  - 水印特征和声学特征进行特征级融合，然后输入生成模型生成带有水印的音频
- Parameter-level Watermarking
  - Latent Watermarking, HiFiGANw, P2Mark
  - 水印嵌入在模型参数里
  - 可用于代码和模型开源的场景



(a) Audio-level Watermarking (b) Feature-level Watermarking (c) Parameter-level Watermarking

模型参数级水印

Latent Watermarking of Audio Generative Models



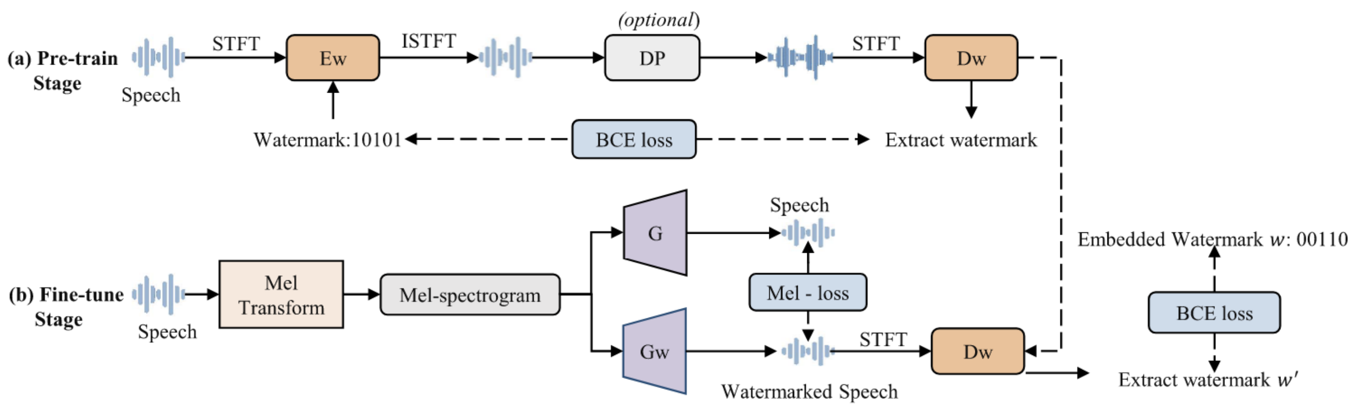
- 基于AudioSeal训练水印生成器和检测器，模拟EnCodec攻击来增强对EnCodec的鲁棒性；
- 对训练数据集添加水印，在加了水印后的数据集上训练MusicGen；
- 推理时生成生成的音频可检测到水印；

不足：
- 需要从头开始训练模型，难以对大的模型进行版本迭代或适应已经训练的模型；
- 训练数据进行水印处理，降低了训练数据的质量；
- 水印的鲁棒性增强需要针对生成模型来设计。

San Roman R, Fernandez P, Deleforge A, et al. Latent Watermarking of Audio Generative Models[J]. 2024.

模型参数级水印

HiFi-GANw: Watermarked Speech Synthesis Via Fine-Tuning of HiFi-GAN



- 预训练水印编码器Ew和解码器Dw，以提取二进制水印；
- 用固定的水印微调Hifi-GAN的生成器G，使得所有合成的语音都嵌入了此水印。

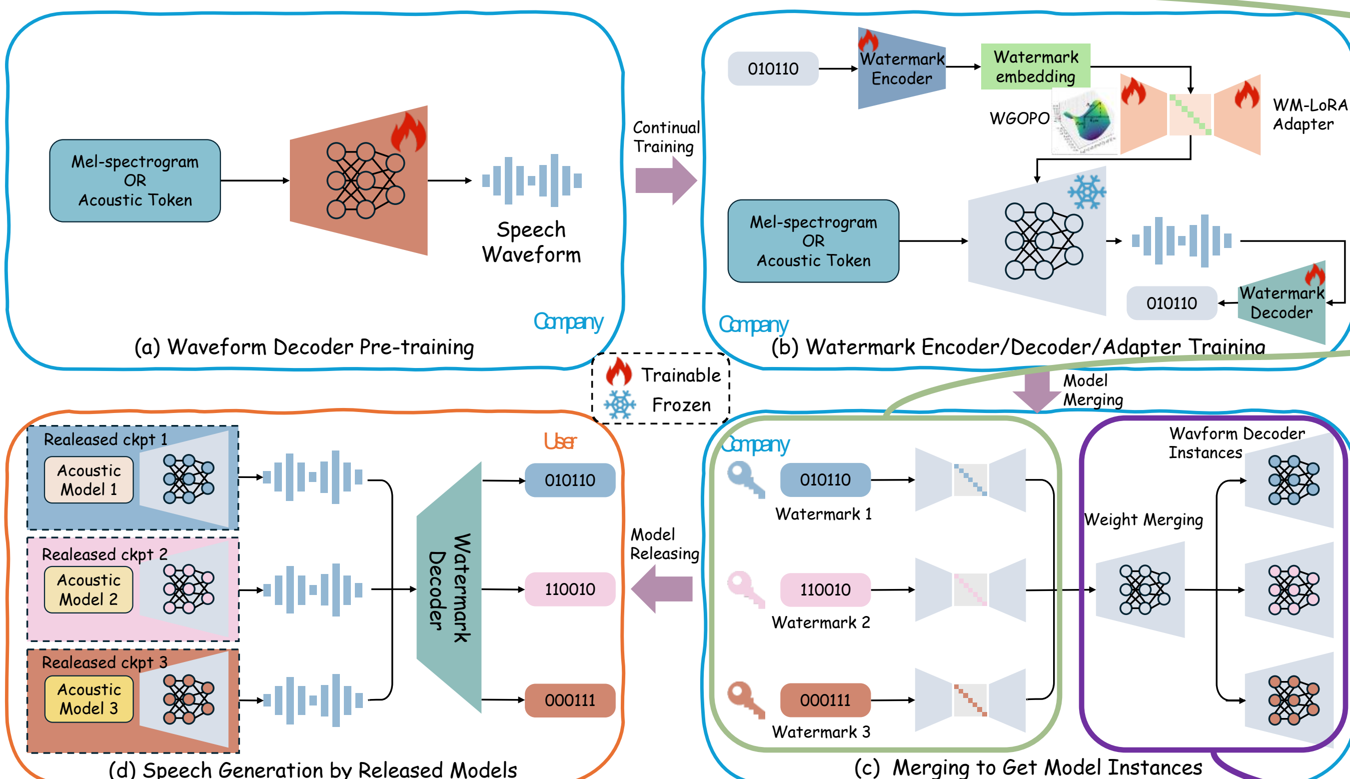
不足:

- 在微调过程中嵌入的水印是固定的，要改变模型中嵌入的水印需要重新微调，缺乏灵活性；

Cheng X, Wang Y, Liu C, et al. HiFi-GANw: Watermarked Speech Synthesis Via Fine-Tuning of HiFi-GAN[J]. IEEE Signal Processing Letters, 2024.

模型参数级水印

P2Mark: Plug-and-play Parameter-intrinsic Watermarking for Neural Speech Generation



Watermark Adapter

灵活性：发布前易更改

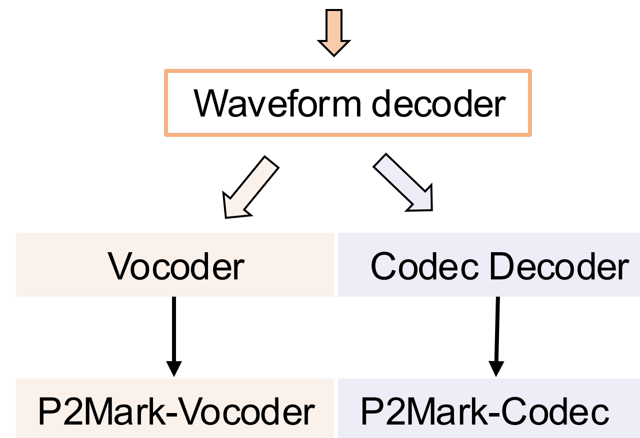
Parameter-level Merging

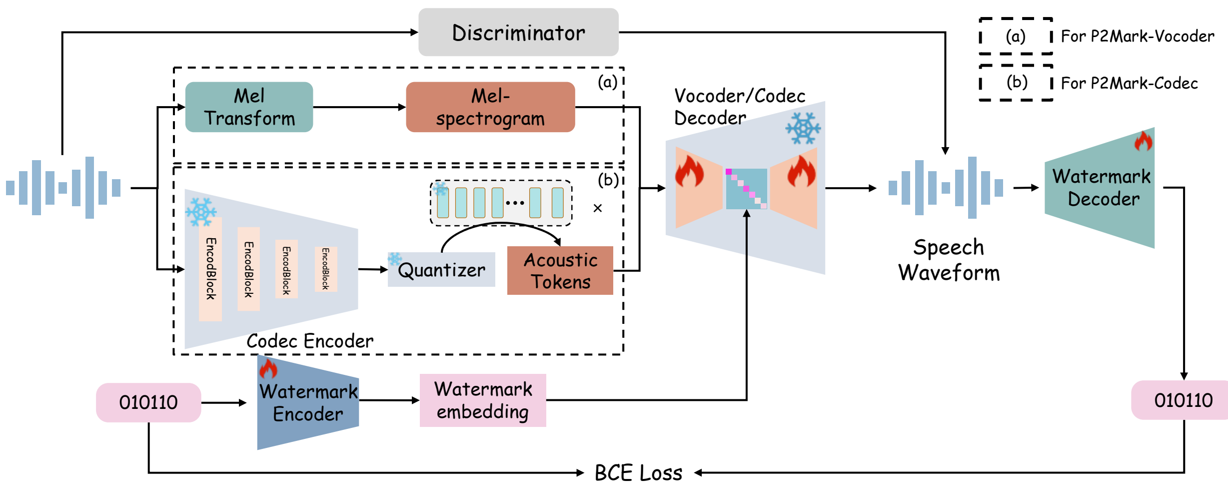
安全性：发布后不易更改

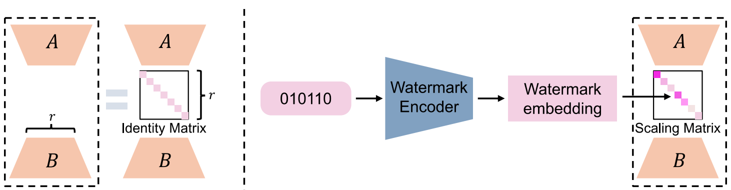
Ren Y, Yi J, Wang T, et al. P2Mark: Plug-and-play Parameter-intrinsic Watermarking for Neural Speech Generation[J]. arXiv preprint arXiv:2504.05197, 2025.

模型参数级水印

P2Mark: Plug-and-play Parameter-intrinsic Watermarking for Neural Speech Generation







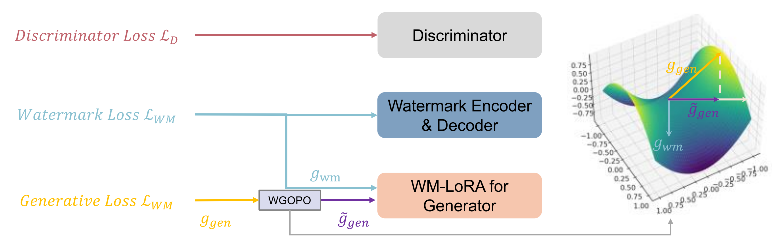
(a) LoRA Adapter

(a) WM-LoRA Adapter

$$h = W_0x + BSAx.$$

$$E^i_{wm}(w_i) = \begin{cases} emb_i, & \text{if } w_i = 1,\\ \mathbf{0}, & \text{otherwise.}\end{cases}$$

$$S = \mathrm{diag}\left(\mathbf{1} + \frac{1}{\sqrt{l}}\sum_{i=1}^{l} E^i_{wm}(w_i)\right)$$



$$minimize_{\tilde{g}_{gen}} \frac{1}{2}\|g_{gen} - \tilde{g}_{gen}\|_2^2 \quad s.t. \quad \tilde{g}_{gen}^{\top} g_{wm} \geq 0.$$

$$\tilde{g}_{gen} = g_{gen} - \frac{g_{gen}^{\top} g_{wm}}{g_{wm}^{\top} g_{wm}} g_{wm},$$

Ren Y, Yi J, Wang T, et al. P2Mark: Plug-and-play Parameter-intrinsic Watermarking for Neural Speech Generation[J]. arXiv preprint arXiv:2504.05197, 2025.

模型参数级水印

P2Mark: Plug-and-play Parameter-intrinsic Watermarking for Neural Speech Generation

Task	Method	Type	WB-P	Audio quality metrics PESQ↑	STOI↑	Mel Dis↓	STFT Dis↓	ACC↑
Vocoder	HiFi-GAN			3.25	0.966	3.26	3.10	–
	WavMark[11]	Audio-level	✗	3.09	**0.964**	3.94	3.20	**1.00**
	AudioSeal[12]	Audio-level	✗	**3.17**	**0.965**	**3.40**	**3.12**	**1.00**
	P2Mark-Vocoder	Parameter-level	✓	**3.21**	**0.965**	**3.46**	**3.19**	**1.00**
Codec	HiFi-Codec			3.52	0.966	3.02	2.71	–
	WavMark[11]	Audio-level	✗	3.32	0.963	3.69	2.82	**1.00**
	AudioSeal[12]	Audio-level	✗	**3.45**	**0.964**	3.20	**2.73**	**1.00**
	TraceableSpeech[14]	Feature-level	✗	3.11	0.959	3.53	2.89	**1.00**
	WMCodec[15]	Feature-level	✗	3.43	**0.961**	**3.13**	2.77	**1.00**
	P2Mark-Codec	Parameter-level	✓	**3.48**	**0.964**	**3.09**	**2.74**	**1.00**

Table 1: Performance comparison between two variants of P2Mark on speech generation models' decoders: P2Mark-Vocoder and P2Mark-Codec, against baseline audio watermarking models. WB-P indicates whether the method can provide white box protection in the source code and weights open source scenario. The **red** denotes the highest result, and the **blue** denotes the second highest result.

Attack Type	Subtype	Description
Noise	Pink	Adds pink noise to audio signal (std=0.1)
	White	Adds Gaussian noise to audio signal (std=0.05)
Filtering	Lowpass	Applies lowpass filter with 500 Hz cutoff
	Bandpass	Applies Bandpass filtering in 500 Hz - 1.5 kHz
	Highpass	Applies highpass filter with 1.5 kHz cutoff
Volume	Boost	Amplifies audio by factor 10
	Duck	Reduces volume by factor 0.1
Compression	MP3	MP3 codec at 128 kbps bitrate
	AAC	AAC codec at 128 kbps bitrate
Others	Resample	Upsamples from 24 kHz to 44.1 kHz then downsamples back
	Echo	Adds 0.5s delay with 0.5 decay factor
	Crop	Keeps only the first half of waveform

Table 3: Detailed description of audio attack types and their settings.

Task	Variant	Bits	Audio quality metrics PESQ↑	STOI↑	Mel Dis↓	STFT Dis↓	ACC↑
Vocoder	HiFi-GAN		3.25	0.966	3.26	3.10	–
	P2Mark-Vocoder	16	3.21	0.965	3.46	3.19	1.00
	- w/o WGOPO		3.18(-0.03)	0.959(-0.006)	3.60(+0.14)	3.22(+0.03)	1.00(-0.00)
	P2Mark-Vocoder	32	3.04	0.955	3.80	3.29	1.00
	- w/o WGOPO		2.94(-0.10)	0.947(-0.008)	3.98(+0.18)	3.32(+0.03)	0.97(-0.03)
Codec	HiFi-Codec		3.52	0.966	3.02	2.71	–
	P2Mark-Codec	16	3.48	0.964	3.09	2.74	1.00
	- w/o WGOPO		3.36(-0.12)	0.960(-0.004)	3.21(+0.12)	2.78(+0.04)	0.98(-0.02)
	P2Mark-Codec	32	3.42	0.963	3.14	2.75	1.00
	- w/o WGOPO		3.29(-0.13)	0.957(-0.006)	3.33(+0.19)	2.81(+0.06)	0.99(-0.01)

Table 2: The ablation study on the efficiency of WGOPO and the watermark capacity.

Attack Type	Subtype	Method WavMark	AudioSeal	P2Mark-Vocoder	P2Mark-Codec
None		1.00	1.00	1.00	1.00
Noise	Pink	0.98	0.99	0.98	0.99
	White	0.50	0.62	0.60	0.55
Filtering	Lowpass	0.50	0.50	0.50	0.50
	Bandpass	0.50	1.00	0.76	0.72
	Highpass	1.00	0.49	0.99	1.00
Volume	Boost	1.00	1.00	1.00	1.00
	Duck	1.00	1.00	1.00	1.00
Compression	MP3	1.00	1.00	0.98	0.99
	AAC	1.00	0.63	1.00	1.00
Others	Resample	1.00	1.00	1.00	1.00
	Echo	0.97	1.00	1.00	1.00
	Crop	0.96	1.00	1.00	1.00

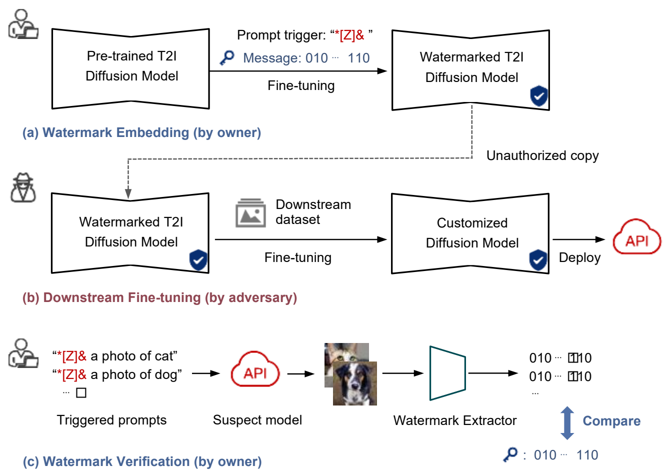
Table 4: Robustness comparison under various attacks. The underline indicates a watermark extraction accuracy below 0.90.

Ren Y, Yi J, Wang T, et al. P2Mark: Plug-and-play Parameter-intrinsic Watermarking for Neural Speech Generation[J]. arXiv preprint arXiv:2504.05197, 2025.

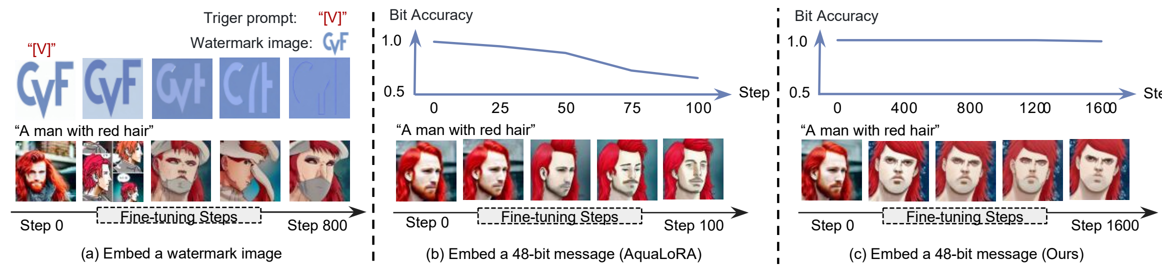
模型参数级水印

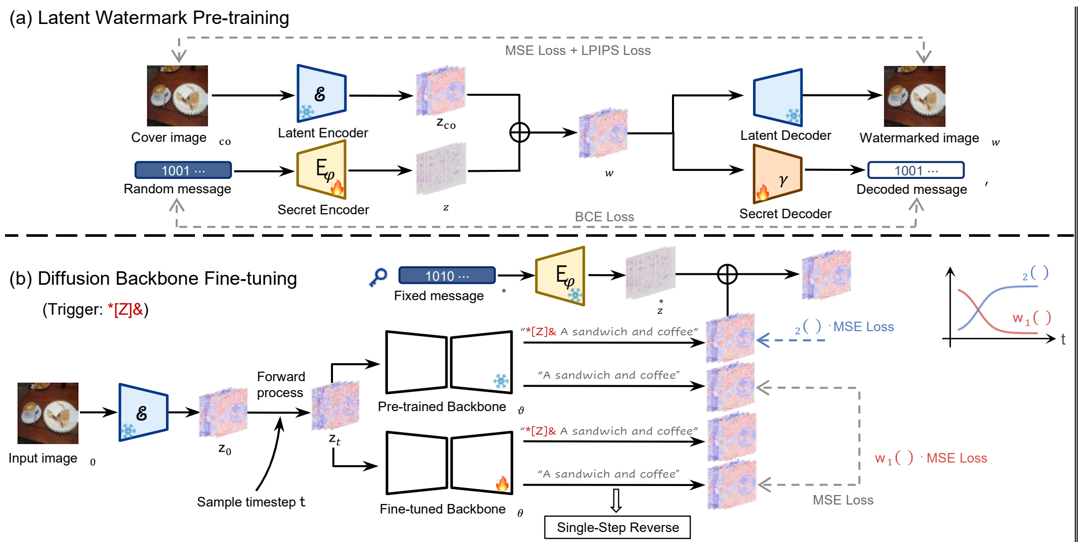
问题——微调能否保留水印？

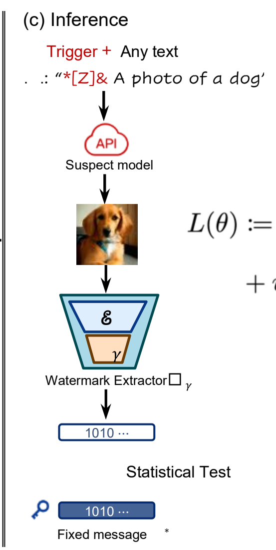
SleeperMark: Towards Robust Watermark against Fine-Tuning Text-to-image Diffusion Models



Trigger-based watermarking





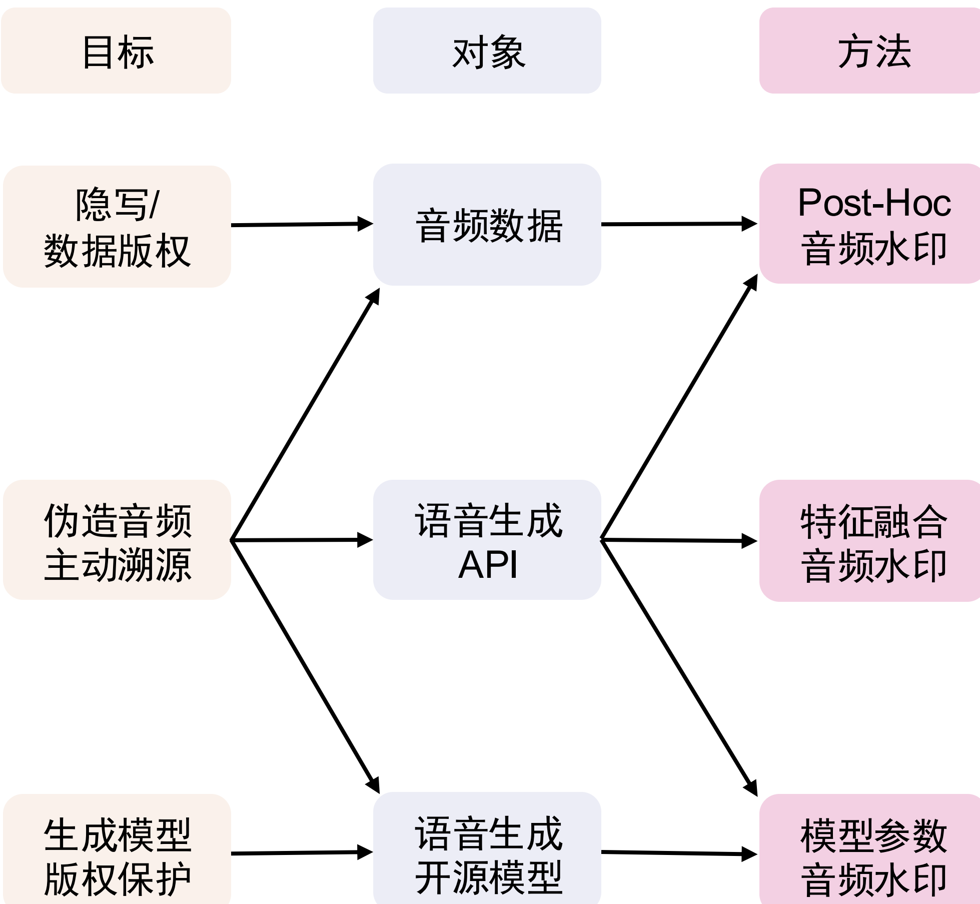


Triggered Prompts

$$L(\theta) := \mathbb{E}_{t,y,z_0,\epsilon}\Big[\eta \cdot w_1(t) \cdot \big\|\hat{z}_{0_\theta}^{t,y_{tr}} - (\hat{z}_{0_\vartheta}^{t,y_{tr}} + \delta_z^*)\big\|^2 + w_2(t) \cdot \big\|\hat{z}_{0_\theta}^{t,y_{tr}} - \hat{z}_{0_\vartheta}^{t,y_{tr}}\big\|^2 + \big\|\hat{z}_{0_\theta}^{t,y} - \hat{z}_{0_\vartheta}^{t,y}\big\|^2\Big]$$

Wang Z, Guo J, Zhu J, et al. SleeperMark: Towards Robust Watermark against Fine-Tuning Text-to-image Diffusion Models[J]. arXiv preprint arXiv:2412.04852, 2024.

挑战和未来可能的方向



- Audio-level
- 关注点：水印容量
- 优势：通用性
- 未来研究：更大容量；嵌入文本/图像；抵御域外攻击

- Feature-level
- 关注点：不可感知性、提取准确性
- 优势：端到端，速度快
- 未来研究：更加深度有效的融合

- Parameter-level
- 关注点：不可感知性、提取准确性
- 优势：参数级融合，开源白盒保护
- 未来研究：抵御微调；针对对抗攻击的防御；基于触发的水印；灵活（易用）性

Thank you !