LOCKEY: A Novel Approach to Model Authentication and Deepfake Tracking

Mayank Kumar Singh
Sony AI, Japan

Naoya Takahashi
Sony AI, Europe

Wei-Hsiang Liao
Sony AI, Japan

Yuki Mitsufuji
Sony AI, Japan

Abstract

This paper presents a novel approach to deter unauthorized deepfakes and enable user tracking in generative models, even when the user has full access to the model parameters, by integrating key-based model authentication with watermarking techniques. Our method involves providing users with model parameters accompanied by a unique, user-specific key. During inference, the model is conditioned upon the key along with the standard input. A valid key results in the expected output, while an invalid key triggers a degraded output, thereby enforcing key-based model authentication. For user tracking, the model embeds the user's unique key as a watermark within the generated content, facilitating the identification of the user's ID. We demonstrate the effectiveness of our approach on two types of models—audio codecs and vocoders—utilizing the SilentCipher watermarking method. Additionally, we assess the robustness of the embedded watermarks against various distortions, validating their reliability in various scenarios.

1 Introduction

The potential misuse of deep learning-based generative models has garnered significant attention in recent years due to the substantial advancements in generative AI, which have produced outputs nearly indistinguishable from real data (1; 2; 3; 4; 5; 6). Such generated content, commonly referred to as deepfakes, can be exploited for malicious purposes. Consequently, there has been significant research aimed at detecting deepfakes, with most approaches relying on identifying discrepancies between the statistical distributions of generated samples and that of real data (7; 8; 9). However, as generative models continue to improve, these discrepancies diminish, rendering traditional detection methods less effective. This necessitates an active approach to track generated content, with watermarking emerging as a widely adopted solution (10; 11; 12; 13).

In cases where generative models are provided as a service, such as through a cloud-based platforms, the service provider can embed watermarks containing user-specific metadata into the generated output, facilitating the tracking and detection of deepfakes. However, when users have complete access to the model parameters, it becomes easier to circumvent the embedding of the user-id into the generated output via the watermarking process, posing a significant challenge to ensuring the integrity and traceability of generated content.

To address this issue, we propose a key-based authentication method designed to prevent users from bypassing the watermarking process. In our approach, the generative model produces a degraded output if an invalid key is provided, while a valid key results in the implicit embedding of the user's unique ID as a watermark in the generated output. We demonstrate the generalizability of our method on two classes of models, audio codecs and vocoders, utilizing SilentCipher (12), a deep learning-based watermarking technique. Specifically, we employ the Encodec model (14) for audio codecs and the HiFi-GAN model(15) for vocoders. Although audio codecs and vocoders are not strictly generative models, our motivation for enabling key-authentication for them stems from the growing trend of generative AI models that operate in latent spaces and use latent decoders to convert

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

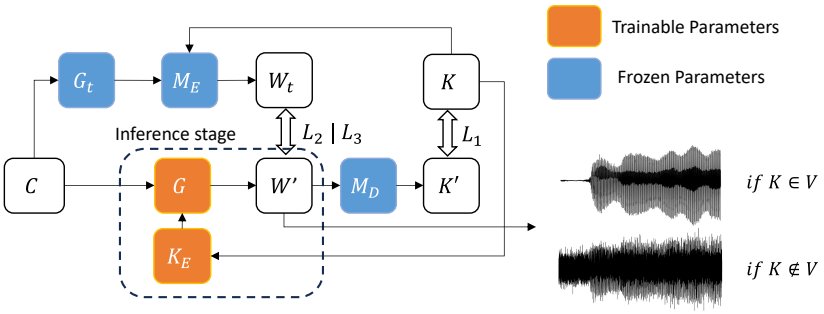

Figure 1: Model Training & Inference Flow

the latent representations to the data domain outputs (2; 3; 4; 5). Demo samples for our proposed method can be found at [1]

2 Related Works

Previous works have focused on ensuring generative AI models produce signature watermarks that enable the identification of the model used to generate a sample (16; 17; 18). However, to the best of our knowledge, this is the first work to propose a key-based authentication mechanism in a white-box scenario, where users have access to both the model parameters and inference script, enabling the tracking of individual users.

Typical deepfake detection methods have primarily relied on passive approaches, such as training classifiers to distinguish between the distributions of generated and real samples (7; 8; 9). These approaches, however, face limitations due to diminishing differences between real and generated samples and the limited information that can be extracted from the classifiers. To address this, active methods like watermarking have been employed, increasing the capacity of the embedded message without being constrained by the indistinguishability of real and generated samples (10; 11; 12). While earlier watermarking techniques suffered from perceptible noises, recent advancements have enabled watermarks that remain imperceptible and are robust against various distortions (12). These advancements allow user-tracking in cloud-based scenarios, where the generative models' watermark user-specific signatures in the generated outputs. However, when models are available locally, users can bypass *post-hoc* watermarking process, making it difficult to track malicious activity.

Our method addresses this challenge by combining key-based authentication with *in-model* watermarking, making it more difficult to bypass the watermarking techniques even in local environments.

3 Proposed Method

To address the ease of bypassing post-hoc watermarking when users have access to model parameters, we propose an in-model watermarking technique. Unlike existing methods that embed a constant key, our approach enables user-specific watermarks by conditioning the model on the user's unique key. To prevent misuse, the model is trained to distinguish between real and fake keys, generating degraded output if a fake key is detected. An overview of our method is shown in Figure 1.

First, we uniformly sample a set of valid keys V from the set of all possible keys A. During training, a key K is sampled from either V or $V \setminus A$ with equal probability. The key, $K \in Z_2^T$, where $Z \in \{0, 1\}$, is projected to learnable embeddings, $L \in \mathbb{R}^{T \times M}$ where T is the key size and M is the embedding dimensions, and fed to the key encoder K_E. The output of K_E, along with the input condition C is fed to the trainable generator G to get W'.

Valid key losses As illustrated in Figure 1, during training we feed the generated waveform to the frozen pretrained message decoder M_D to get K'. To ensure that the respective models learn to embed the key as a watermark, we apply the cross entropy loss between K and K' if the key is sampled from V as per the equation 1. For the gradients to propagate to G, M_D must be differentiable.

[1] `https://interspeech2024.github.io/model_auth/`

$$\mathcal{L}_1 = -\sum_{i=1}^{N}\sum_{j=1}^{T} K_{ij}\log(K'_{ij}) \tag{1}$$

To ensure that the fine-tuned model does not have degradation due to the watermarking loss, we introduce a perceptual loss when $K \in V$. The perceptual loss is defined as the MSE loss between the watermarked output W_t of the frozen pre-trained model G_t and W'. We get W_t by providing C to G_t and feeding the output of G_t, along with K, to the pre-trained message encoder M_E. Please refer to the Figure 1 for the notations. Our initial experiments suggested that the perceptual loss does not succeed in removing the perceptual distortions. To further improve upon it, we introduce the MSE loss between the log normalized magnitude spectrogram of W_t and W' as per equation 2.

$$\mathcal{L}_2 = ||W_t - W'||_2 + ||\log(|\mathrm{STFT}(W)|) - \log(|\mathrm{STFT}(W')|)||_2 \tag{2}$$

Invalid key losses When the sampled key, $K \notin V$, we minimize the negative MSE between W' and W_t. For stability, we introduce a curriculum learning method wherein the invalid loss is restricted to a lower bound, B, which is increased as per the details mentioned in Section 4.

$$\mathcal{L}_3 = ReLU(B - ||W_t - W'||_2) \tag{3}$$

Key Verification Loss To make it easier for the model to distinguish the valid and invalid keys, we feed the output of K_E to a two-layer fully connected neural network with ReLU activation which is trained with cross-entropy loss $\mathcal{L}_4$ to output zero if $K \notin V$ and one if $K \in V$.

Total Loss During training we combine the losses as given in equation 4, where λ_1, λ_2 and $\lambda_3 \in \mathbb{R}$, $\mathbb{I}(K \in V)$ is an indicator function that equals 1 if $K \in V$ and 0 if $K \notin V$ and $\mathbb{I}(K \notin V)$ is an indicator function that equals 0 if $K \in V$ and 1 if $K \notin V$.

$$\mathcal{L} = (\lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2) \times \mathbb{I}(K \in V) + \lambda_3\mathcal{L}_3 \times \mathbb{I}(K \notin V) + \lambda_4\mathcal{L}_4 \tag{4}$$

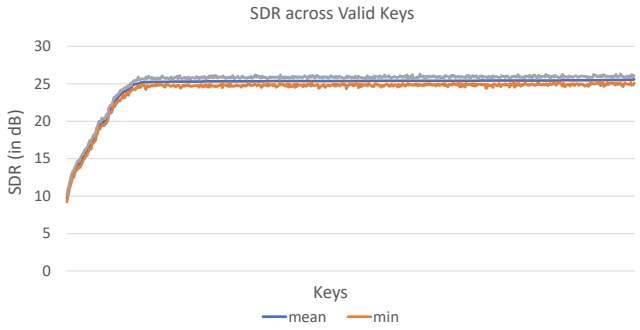

Figure 2: SDR across valid keys. The keys are sorted ascendingly based on their mean SDR on 200 samples

4 Experiments

We apply our key-based authentication method to two models: HiFi-GAN (15), trained at 22.05 kHz, and Encodec 32 kHz (14). The input condition, C, for HiFi-GAN model is MEL spectrogram, while for Encodec's decoder, it is latent codes. K_E is composed of five alternating convolution and ReLU layers. The output of K_E is added to the output of the second layer of G_t for both HiFi-GAN and Encodec. We train the message encoder M_E and message decoder M_D based upon SilentCipher (12), a deep learning based watermarking technique. Separate SilentCipher models are trained for a

Table 1: Base Silent Cipher Model Objective Scores. We compare the baselines using objective test scores by simulating various attacks. SDR: SDR between watermarked and original signal, eq: random equalization, gaus: additive Gaussian noise of 40dB, quant: 16-bit floating-point Quantization, time_jit: time-jittering, resamp: random resampling from 6.4kHz to 16kHz and orig: No attacks.

					mp3			ogg						
Models	Dataset	SDR	eq	gaus	64k	128k	256k	64k	128k	256k	quant	resamp	time_jit	orig
SC	VCTK	31.03	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HiFi-GAN+SC	VCTK	30.26	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SC	MTG	32.39	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
Encodec+SC	MTG	30.78	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Objective Test Results. SDR Valid: SDR (in dB) of the generated sample when conditioned on a valid key, SDR Invalid: SDR (in dB) of the generated sample when conditioned on an invalid key. For other notations, refer to Table 2 captions

					mp3			ogg						
Models	SDR Valid	SDR Invalid	eq	gaus	64k	128k	256k	64k	128k	256k	quant	resamp	time_jit	orig
HiFi-GAN	25.95	1.28	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Encodec	23.20	3.72	0.91	0.94	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.88	0.97	0.97

Table 3: Subjective Scores with 95% confidence intervals

Encodec	MOS	HiFi-GAN	MOS
Real	4.06 ± 0.22	Real	4.31 ± 0.20
Watermarked	3.92 ± 0.22	Watermarked	4.35 ± 0.16
Valid	3.76 ± 0.22	Valid	3.17 ± 0.19
Invalid	1.11 ± 0.07	Invalid	1.50 ± 0.13

sampling rate of 22.05kHz and 32kHz. Unless otherwise stated, the models use a 16-bit key with 655 valid keys randomly selected from the possible 2^16 keys. Evaluations are conducted on six-second samples. For invalid loss, we employ curriculum training where the lower bound of the invalid loss is gradually decreased. We double B after every 5000 iterations, starting from 0.005.

Datasets For training the HiFi-GAN model we use the VCTK dataset (19) compromising of 44 hours of speech data. For the Encodec model we use the MTG-Jamendo dataset (20) which contains 55k full music audio tracks. The train, validation and testing set are split in the ratio 0.8:0.1:0.1. We process the data for HiFi-GAN by extracting the MEL spectrogram of the waveform with the size of the fourier transform being 1024, window length being 1024, hop size being 256 and nuumber of mels being 80. For Encodec, we process the waveform and extract the latent codes using Encodec's encoder.

Training All our methods were fine-tuned using the Adam optimizer with a learning rate of 1e-4 for a total of 25k iterations for HiFi-GAN and 80k iterations for Encodec. The audio duration during training is fixed to 10 seconds. Although we don't introduce any distortions in the watermarked output during the training of the models, we evaluate our models on various distortions like Gaussian noise, random equalization of frequency bands and audio compression algorithms. We iterate over the bit-rates 64, 128 and 256 kbps across two compression methods, MP3 and OGG.

5 Results

Objective Results We evaluate the accuracy of the decoded watermarks when the model is conditioned on valid keys, following distortion of the encoded signal. The applied distortions include additive Gaussian noise at 40 dB (gaus), random band-limited equalization of 15 dB at 35 Hz, 200 Hz, 1000 Hz, and 4000 Hz (eq), 16-bit floating point quantization (quant), random resampling between 40% and 100% of the original sampling rate (resamp), time-jittering (time_jit), and MP3/OGG compression at 64, 128, and 256 kbps. Since there are no established baselines, we compare our method to *post-hoc* watermarking techniques. Table 1 summarizes the objective results of applying

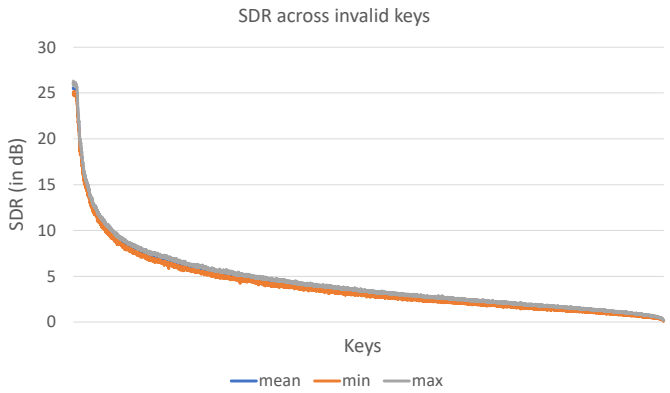

Figure 3: SDR across invalid keys. The keys are sorted descendingly based on their mean SDR on 200 samples

the SilentCipher (SC) watermarking technique to both real and reconstructed samples generated by pretrained models on the VCTK and MTG datasets. The results demonstrate the robustness of the SilentCipher model in withstanding various distortions, consistently achieving an average SDR exceeding 30 dB. For the VCTK dataset, we use SilentCipher trained at a sampling rate of 22.05 kHz, while for the MTG dataset, the model trained at a sampling rate of 32 kHz is utilized. Table 2 presents the objective results for HiFi-GAN and Encodec model after fine-tuning for key-based authentication and watermarking. Signal-to-distortions ratios (SDRs) are computed between W' and W_t, where SDR_valid represents conditioning on valid keys and SDR_invalid represents conditioning on invalid keys. As shown in Table 2, SDR_valid is significantly higher than SDR_invalid, indicating the model's ability to distinguish between valid and invalid keys while largely preserving watermarks despite distortions.

Subjective Results We also conducted a subjective mean opinion score (MOS) test to assess the perceptual quality of the waveforms generated with valid and invalid keys. Sixteen audio engineers rated the audio on a scale of 1-5, 1 being completely unnatural and 5 representing completely natural samples. The results are presented in Table 3. For the Encodec model, "Real" refers to unaltered samples from the MTG dataset and "Watermarked" refers to samples watermarked using the SilentCipher-32kHz model. Similarly, for the HiFi-GAN model, "Real" and "Watermarked" represent unaltered and watermarked samples of the VCTK dataset using SilentCipher-22.05kHz. "Valid" and "Invalid" refer to the generated samples conditioned upon valid and invalid keys, respectively. The subjective results align with the objective metrics, demonstrating the model's ability to generative high-quliaty or degraded samples based on key validity.

Probing the HiFi-GAN model We evaluated HiFi-GAN on 200 samples across all keys and plotted the minimum, average, and maximum SDR for valid and invalid keys as shown in Figure 2 and Figure 3, respectively. Approximately 12% of the valid keys achieved an average SDR below 25dB and around 1% of the invalid keys achieve an average SDR above 20dB. As the SDR for a specific key across the samples does not vary much, shown by the small gap between the maximum and minimum SDR for each key, it is easy to verify if a key works well by evaluating the SDR on a few samples and discarding them if their SDR's lie beyond a certain threshold.

6 Conclusions

We present a novel approach for authenticating generative AI models in white-box scenarios, where users have full access to model parameters. The effectiveness of the proposed method is demonstrated through comprehensive objective and subjective evaluations on the HiFi-GAN and Encodec models. Future work will focus on minimizing perceptible distortions in the generated outputs for valid cases and expanding both the number of valid keys and the total key size.

References

[1] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon, "Consistency trajectory models: Learning probability flow ODE trajectory of diffusion," in *The Twelfth International Conference on Learning Representations*, 2024.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[3] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024, NIPS '23, Curran Associates Inc.

[4] Zach Evans, Cj Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons, "Fast timing-conditioned latent audio diffusion," in *Proceedings of the 41st International Conference on Machine Learning*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, Eds. 21–27 Jul 2024, vol. 235 of *Proceedings of Machine Learning Research*, pp. 12652–12665, PMLR.

[5] Naoya Takahashi, Mayank Kumar Singh, and Yuki Mitsufuji, "Robust one-shot singing voice conversion," 2023.

[6] Mayank Kumar Singh, Naoya Takahashi, and Naoyuki Onoe, "Iteratively Improving Speech Recognition and Voice Conversion," in *Proc. INTERSPEECH 2023*, 2023, pp. 206–210.

[7] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato, "Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models," 2023.

[8] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li, "Dire for diffusion-generated image detection," 2023.

[9] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee, "Towards universal fake image detectors that generalize across generative models," 2024.

[10] Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei, "Wavmark: Watermarking for audio generation," 2023.

[11] Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, and Hady Elsahar, "Proactive detection of voice cloning with localized watermarking," 2024.

[12] Mayank Kumar Singh, Naoya Takahashi, Weihsiang Liao, and Yuki Mitsufuji, "Silentcipher: Deep audio watermarking," 2024.

[13] Naoya Takahashi, Mayank Kumar Singh, and Yuki Mitsufuji, "Source mixing and separation robust audio steganography," 2022.

[14] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023, Featured Certification, Reproducibility Certification.

[15] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 17022–17033, Curran Associates, Inc.

[16] Robin San Roman, Pierre Fernandez, Antoine Deleforge, Yossi Adi, and Romain Serizel, "Latent watermarking of audio generative models," 2024.

[17] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein, "A watermark for large language models," in *Proceedings of the 40th International Conference on Machine Learning*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, Eds. 23–29 Jul 2023, vol. 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084, PMLR.

[18] Lauri Juvela and Xin Wang, "Collaborative watermarking for adversarial speech synthesis," 2024.

[19] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[20] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019.

A Appendix

A.1 White-box attack on the authentication enabled HiFi-GAN model

We conducted a white-box attack on the HiFi-GAN model to remove the embedded watermark by adding a small Gaussian noise to the output of each layer of the model, refer to Figure 4. We vary the standard deviation of the added Gaussian noise and plot the accuracy of the watermark as a function of SDR of the generated samples. Although high-energy Gaussian noise degraded the accuracy of the embedded watermark to zero, this occured alongsize significant degradation in the SDR of the generated sample. For low-energy noise, the embedded watermark is detectable with a high accuracy.

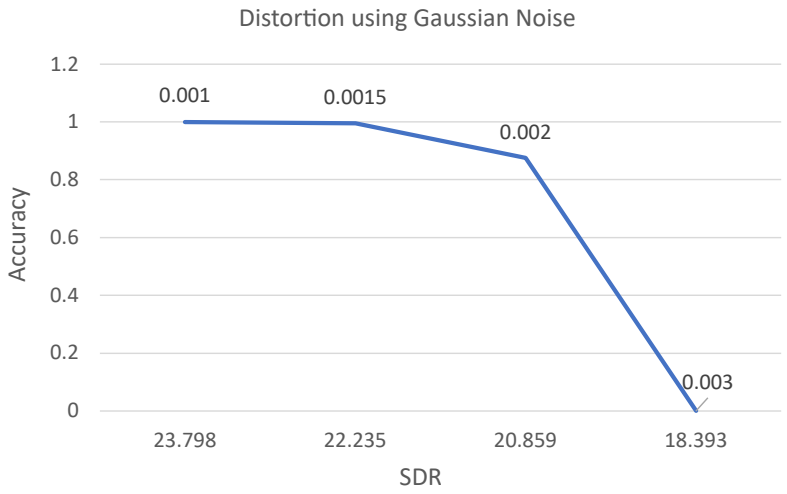

Figure 4: Distortions Using Gaussian Noise. The numbers on corresponding to each data point denote the standard deviation of the added gaussian noise.

A.2 Scalability of the authentication enabled HiFi-GAN model

We explored the HiFi-GAN model's scalability by plotting SDR_valid and SDR_invalid as a function of the total number of keys while keeping the number of valid keys fixed at 655 (Figure 5). The results indicate that SDR_valid and SDR_invalid remain distinct as the total key size increases.

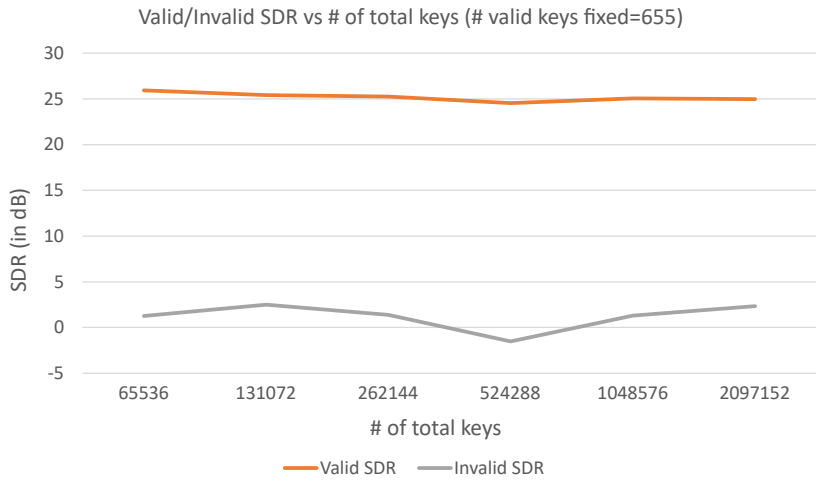

Figure 5: Valid-Invalid SDR across no of total keys