Ownership authentication and integrity verification of digital images using generative models and custom signature

Jonathan Flores-Monroy, Manuel Cedillo-Hernandez, Mariko Nakano-Miyatake, and Hector Perez-Meana
Seccion de Estudios de Posgrado e Investigacion SEPI ESIME Culhuacan
Instituto Politecnico Nacional
Mexico City, Mexico
0000-0002-2467-3600

***Abstract*—In this paper we present an alternative distortion-free algorithm to ownership authentication and integrity verification of digital images that use Contrastive Language-Image Pre-Training (CLIP), Vector Quantized Generative Adversarial Network (VQGAN) and a custom signature based on cryptographic message digest. Authentication and integrity of digital media are evaluated by the naked eye, as shown in the output of the generative architecture, as well as numerical metrics such as PSNR and SSIM, respectively. The code is publicly available at https://github.com/JonathanFlores2503/NoiseHashGAN.**

***Index Terms*—ownership authentication; integrity verification; generative models; artificial intelligence**

I. Introduction

In today's digital age, multimedia content is stored in various formats, including images, audio, video, and text. Artificial intelligence tools can now generate fake content without leaving traces, leading to significant information security issues. This makes proving the authenticity and verifying the integrity of content essential tasks and major challenges for the scientific community in the coming years. Over the past decades, scientific literature has proposed a repertoire of algorithms based on digital watermarking, reversible data hiding, and zero-watermarking, which, along with cryptographic techniques, have been used to address various information security issues related to copyright protection, ownership authentication, and integrity verification of digital multimedia [1]–[3]. Depending on the application scenario, some solutions are more suitable than others. For example, digital watermarking may be suitable if the host signal can tolerate moderate distortion, but if distortion is not acceptable, reversible data hiding or zero-watermarking may be more appropriate solutions.

For instance, in [2], the authors proposed a dual watermarking scheme combining invisible watermarking and zero-watermarking to prevent detachment between medical images and electronic patient records. The authors of [4] presented a method to detect tampering and reconstruct the original image using a self-embedding scheme. In [5], a robust watermarking method was proposed, maintaining watermark imperceptibility and high payload using principal component analysis, singular value decomposition, and the human visual system.

The authors thank the Instituto Politecnico Nacional (IPN) as well as the Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) for the support provided during the realization of this research.

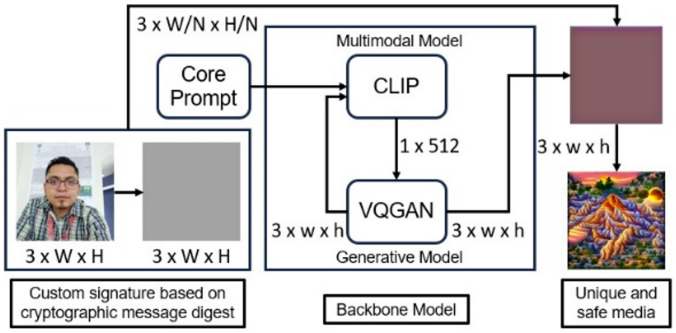

Fig. 1. Architecture of the CLIP + VQGAN model with the integrated signature mechanism.

In this paper, we present an alternative algorithm for ownership authentication and integrity verification of digital images. Unlike traditional methods, our algorithm does not rely on the existence of an original image. Instead, it generates an image from an initial message ("core prompt") and a unique seed ("digital signature"), ensuring that the image can be exactly recreated only if these input data, which we have termed "digital signatures," are used. Our approach leverages Contrastive Language-Image Pre-Training (CLIP) [6], [8], Vector Quantized Generative Adversarial Network (VQ-GAN) [7], [8], and a custom signature based on a cryptographic message digest. Our main contributions are summarized as follows:

- We generate images from a prompt and a seed encode from input data that serves as a digital signature (whether it be an image, audio, text, etc.), instead of starting with an existing image. This means there is one "original" image; the generated image is the original and can only be reproduced with the exact input data for replication.
- We leverage the capabilities of CLIP and VQ-GAN to

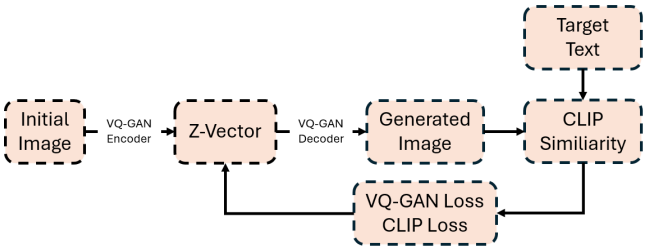

Fig. 2. General operation diagram of the VQ-GAN + CLIP architecture

jointly generate and encode images. In traditional techniques, VQ-GAN uses an "initial image" with random pixels. However, in our approach, CLIP comprehends and encodes the semantic content of the initial message ("core prompt"), while VQ-GAN uses the predefined seed ("digital signature") as the "initial image."

- We implement a unique cryptographic signature for each created image, based on a hash that converts digital data (whether image, audio, text, etc.) into a unique representation, serving as our author signature. This signature guarantees authenticity and prevents alterations. The author's signature, along with the initial message, allows the generation and precise replication of the final image.

The remainder of this paper is organized into five sections as follows. Section 2 provides brief descriptions of related works. Section 3 describes the proposed methodology in detail. Section 4 presents the experimental results of the proposed method. Finally, Section 5 provides the conclusions of this work.

II. Related Works

Vector Quantized Generative Adversarial Networks (VQ-GAN) [7], [8] represent a significant advancement in generative models, combining the high perceptual quality of GANs with the discrete latent space of vector quantization. VQ-GAN uses a convolutional autoencoder structure composed of an encoder E and a decoder D. The encoder transforms an **input image** into a latent representation z that is quantized using a codebook of discrete embeddings z_q. The decoder utilizes this quantized representation to reconstruct the image, maintaining fine details and complex structures. Vector quantization allows for efficient high-resolution image synthesis, improving coherence and visual quality through a patch-based discriminator D_p and transformers that model the composition of visual parts. This approach is particularly effective in modeling long-range dependencies and complex textures within images.

Contrastive Language–Image Pretraining (CLIP) [6], [8] is a model developed by OpenAI that addresses the challenge of linking visual and textual representations through contrastive learning. CLIP jointly trains text encoders T and image encoders I to map textual descriptions and images to a shared representation space. Utilizing a large collection of image-text pairs, CLIP maximizes the similarity of corresponding pairs and minimizes that of non-corresponding pairs, allowing for the effective evaluation of semantic correspondence between an image and a text. One of CLIP's most notable features is its ability to perform zero-shot tasks, where it can classify and generate visual content based on textual descriptions without being explicitly trained on those specific examples. This is due to CLIP's highly generalizable representations that capture the semantics of both language and images, making it effective in a variety of computer vision applications.

The combination of VQ-GAN and CLIP in the VQ-GAN+CLIP methodology [8] leverages the strengths of both models for the generation and editing of images guided by textual descriptions. VQ-GAN generates high-quality images while CLIP evaluates the semantic correspondence between text and image. The iterative process begins with an **initial image**, which is processed by VQ-GAN encoder to generate a **quantized latent representation** z_q. CLIP evaluates the similarity between the generated image and the textual description, adjusting the latent representation z to improve semantic correspondence (Fig. 2). This method does not require additional training, using pretrained models to guide the generation and editing of images efficiently and adaptively to various tasks.

III. Proposed Method

The key idea of the proposed method revolves around the generation of an authenticity signature capable of safeguarding the copyright of the content. This method employs a dual-layer security approach, as illustrated in Fig. 1. The primary layer is constituted by a specific prompt, which acts as the fundamental element for the subsequent creative process. Following this, the second layer incorporates a digital signature, where multimedia such as audio, text, images, and more can be employed. It's worth noting that this article exclusively focuses on implementing the proposed model using an image as the second layer of security. By merging these two parameters, a generative model is utilized to create a distinctive image. This process is based on the fusion of CLIP [6] and VQ-GAN [7], both serving as the primary models. These models take the two layers of security as inputs, culminating in the generation of a singular image. This approach ensures the preservation of originality authorship and, effectively preventing duplication. The intricate details of these processes are meticulously elucidated in this section.

A. First lock "Core prompt"

Following the fundamental concepts outlined in [6], [7], [10], the initial purpose of the "core prompt" is to provide the generative model with instructions on how we want our resulting image to take shape, enabling it to be generated in accordance with our guidelines. However, for our approach, the "core prompt" will not only serve as a guide for image generation but also as an essential requirement to produce the resulting image we aim to safeguard. This "prompt" must be written exactly as it was originally implemented. Unlike

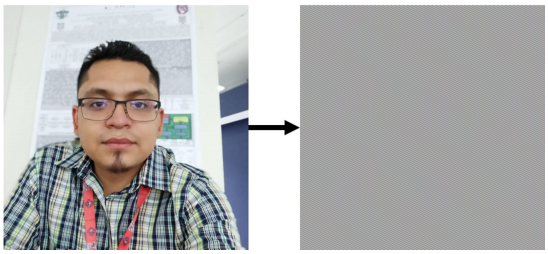

Fig. 3. Original image and its corresponding hashed version.

classical approaches, this prompt plays a more pivotal role in image generation.

B. Second lock "Digital signature"

In the realm of generative models, it is customary that, to initiate the image generation process, a randomly generated **initial image**, such as a random pixel image, is used. This initial image undergoes a coding process to transform it into a vector z [5], [9], [10]. This vector z goes through multiple stages in an iterative cycle to finally generate the desired image (Fig. 2).

In the combined VQ-GAN+CLIP architecture, VQ-GAN uses an encoder E to transform this randomly generated input image into a latent vector z, which is then quantized into z_q using a codebook of discrete embeddings. The decoder D reconstructs the image from z_q, maintaining fine details and complex structures. CLIP, which has been pretrained to understand the semantic correspondence between images and text, evaluates the similarity between the generated image and a provided textual description. Using text encoders T and image encoders I, CLIP projects both inputs into a shared feature space and calculates the cosine similarity between the representations. This similarity value is used to backpropagate the error and adjust the latent vector z, iteratively refining z so that the image generated by VQ-GAN better aligns with the provided textual description.

However, due to the random nature of the initial image, there is a risk that the generated images will be different in each run, which can affect the consistency and reproducibility of the results.

To address this issue, our approach takes a different path by replacing the random input image with our unique digital signature. This leads us to the following process:

- **From Image to Hash Value:** To begin with, we utilize a piece of multimedia data as a foundation for creating the digital signature. In this context, we employ an image $\boldsymbol{F} \in \mathbb{R}^{3\times W\times H}$ as the basis. This image is subjected to the BLAKE2S algorithm [12], which generates a unique hash value with a constant length of 16 bytes. This hash value distinctively captures the properties of the image.
- **From Hash Value to Base Image (Fig. 3):** The hash value obtained previously is processed using the MD5 hash algorithm [13]. The result of this process is a sequence of bytes that encapsulates the unique characteristics of the hash value. For each pixel in the resulting image, color components (red, green, and blue) are selected from this byte sequence. This selection occurs cyclically, ensuring that the image is as representative as possible of the hash value. This iterative process involves assigning the corresponding bytes to the color components of the pixels. Once the color components have been collected and assigned, they are used to populate the previously created image. As a result, we obtain our initial image $\boldsymbol{F_H} \in \mathbb{R}^{3\times W\times H}$.

C. Initial image to Z-vector

Up to this point, we have established the core idea of our approach. However, there remains an essential step: adapting our initial image $\boldsymbol{F_H} \in \mathbb{R}^{3\times W\times H}$ to the Z-vector $\boldsymbol{Z} \in \mathbb{R}^{w\times h\times k}$ (Fig. 4), necessary for generating the image described by our prompt (Fig. 2). This adaptability process unfolds in several stages. Initially, our initial image undergoes a two-step post-processing. The first step is to resize the initial image to a dimension that is divisible by N (e.g. $w = W/N$ and $h = H/N$), where N represents the number of codes in the VQ dictionary, in line with the requirements of the VQ-GAN model. This resizing step is imperative as the VQ-GAN model segments the input image into blocks of size $N \times N$, encoding them into discrete codes using the VQ dictionary. If the size of the input image is not divisible by N, some blocks will not be encoded correctly, which can affect the quality of the generated image. Therefore, it is crucial to ensure that the size of the input image is divisible by N to achieve the best possible results with the VQ-GAN model. The second step involves reducing the dimension of channels; we convert the initial image to grayscale, simplifying the image representation to a single channel, which facilitates its handling in later processes. Once our processed initial image $\boldsymbol{F_z} \in \mathbb{R}^{w\times h}$ is obtained, it is normalized for optimal data handling. Subsequently, we perform a process of expansion and replication in K specific dimensions to meet the input requirements of the generative model, resulting in a tensor shaped appropriately. This tensor represents the Z-vector $\boldsymbol{Z} \in \mathbb{R}^{w\times h\times k}$ (Fig. 4) necessary for the generation of the desired image.

D. Backbone Integration

Having the Z-vector $\boldsymbol{Z} \in \mathbb{R}^{w\times h\times k}$ already, we move on to the last step: the integration or use of the Backbone, which

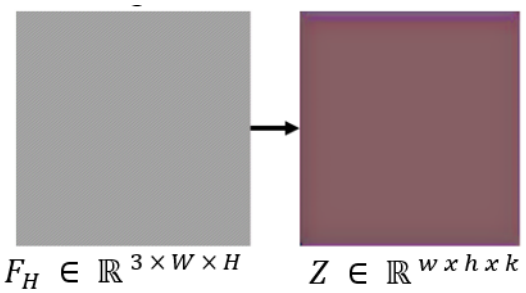

Fig. 4. Initial image and its corresponding Z-vector representation.

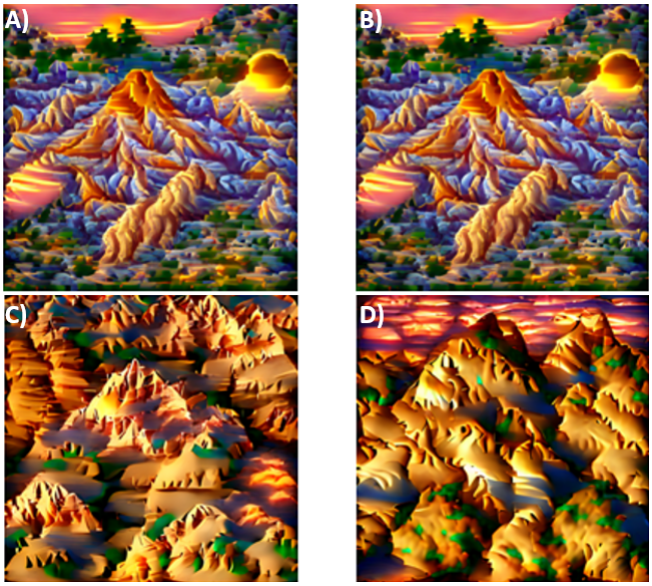

Fig. 5. Images generated by VQ-GAN+CLIP; (a) First image generated using the security keys (FI); (b) Octave image generated using the security keys (OI); (c) Image generated in a conventional way (IC); (d) Image generated from the altered signature (IGA).

in this case involves the combined use of CLIP + VQ-GAN, following the procedures established in [6], [7], [10]. In our approach, the vector $\boldsymbol{Z}$ will serve as the initial input to the VQ-GAN's generative network. This input, which houses our digital signature, is introduced into the VQ-GAN generative network, initiating the creation of the image requested by the promp. This process benefits significantly from the vector quantization structure of VQ-GAN, allowing for precise and high-resolution image reconstruction.

With the $\boldsymbol{Z} \in \mathbb{R}^{w \times h \times k}$ in hand, the CLIP model comes into play to provide semantic guidance. The descriptive text supplied by the prompt, which in our case acts as one of our security layers, is processed by CLIP, resulting in a semantic representation. This semantic representation will be essential in subsequent iterations, steering the image generation towards a visual representation that resonates with the semantics of the text from the prompt. In each iteration, the generated image is adjusted to maximize the semantic correspondence with the text description, according to the evaluation by CLIP, while remaining within the confines of the latent space defined by VQ-GAN. This iterative and collaborative cycle between VQ-GAN and CLIP is carried out until satisfactory coherence is reached between the textual description and the generated visual representation, culminating in the production of the final image $\hat{F} \in \mathbb{R}^{3 \times w \times h}$.

IV. Experimental Results

To evaluate the efficacy of our proposal, we implemented a VQ-GAN+CLIP Backbone with a pre-trained VQ-GAN model, in which a reduction factor of $f = 1/16$ and a vocabulary of 16386 tokens are utilized. Our goal is to demonstrate the feasibility of our security signatures or keys; no additional training was conducted on the Backbone. Herein, the key parameters employed are presented:

- **Loss Function**: We employed a loss function named Inclusion and Exclusion similarity loss function. This function consists of two terms:
  1) **Inclusion Term**: Aims to maximize the semantic similarity between the generated images and the desired text descriptions.
  2) **Exclusion Term**: Minimizes the similarity with specified undesired features.

$$Loss = -\alpha \cdot \text{InclusionTerm} - \beta \cdot \text{ExclusionTerm} \quad (1)$$

Where α and β are hyperparameters that weight the importance of the inclusion and exclusion terms respectively (in this case we set $\alpha = 1$ and $\beta = 0.5$).

- **Stochastic Noise Factor:** A stochastic noise factor was introduced during the optimization to promote the exploration of the latent space. This factor modulates the intensity of random noise applied to the cutouts of the generated images in each iteration; for our proposal, 32 random cutouts and a noise factor of 0.22 were implemented.
- **Optimizer:** *AdamW* optimizer with a learning rate of 0.5 and a weight decay of 0.1 was used.
- **Tag Exclusion:** The generation of images with the tags "watermark", "cropped", "confusing", "incoherent", "cut", and "blurry" were excluded.

These configurations were sufficient to validate the efficacy of the proposed signatures in the realm of generation and verification of secure signatures. Detailed results will be presented in the following.

A. Comparison between generated images

Beginning with the result analysis, in this section, we will illustrate the effectiveness of our proposal when integrating our signature into the VQ-GAN+CLIP architecture. For this analysis, we highlight some crucial points. First, the presented experiments focus on generating resulting images from a single prompt, specifically, ***"Mountainous landscape at sunset"***. This is done with the objective of demonstrating that authenticity protection is viable only when all original parameters are retained, and any modification, no matter how minor, would lead to substantially different results.

TABLE I
SIMILARITY COMPARISON BETWEEN FIRST IMAGE (FI) AND GENERATED IMAGES

Compared image	SSIM	PSNR (dB)
OI	1.00	100
IC	0.06	27.95
IGA	0.08	28.02

In Fig. 5, the images generated by the architecture are shown. Fig. 5-(a) was created from our initial image $\boldsymbol{F_H} \in \mathbb{R}^{3\times W\times H}$ and, consequently, by our Z-vector $\boldsymbol{Z} \in \mathbb{R}^{w\times h\times k}$. A personal context with distinctive textures and colors is clearly appreciated. In Fig. 5-(b), we present an image generated from the same context and the same initial image $\boldsymbol{F_H} \in \mathbb{R}^{3\times W\times H}$, illustrating that the image context does not undergo changes and the comparison with Fig. 5-(a) reveals notable consistency. This is corroborated through Table I, where the Structural Similarity Index (SSIM) value is 1.0 and the Peak Signal-to-Noise Ratio (PSNR) presents a value of 100 dB, evidencing that, by keeping all original parameters, it is possible to accurately determine the original authorship.

Proceeding with Fig. 5-(c), an image was generated with the same prompt, but allowing the architecture to create the image from a random vector z, as would be conventionally done [6], [7], [10]. This figure underscores that, despite having a security mechanism, it is imperative to possess both parameters to accurately replicate the image we wish to protect. Visually, the images no longer bear any relation, and Table 1 shows a SSIM value of 0.06, an extremely low value, indicating that, although both images refer to the same theme, they significantly differ in representation.

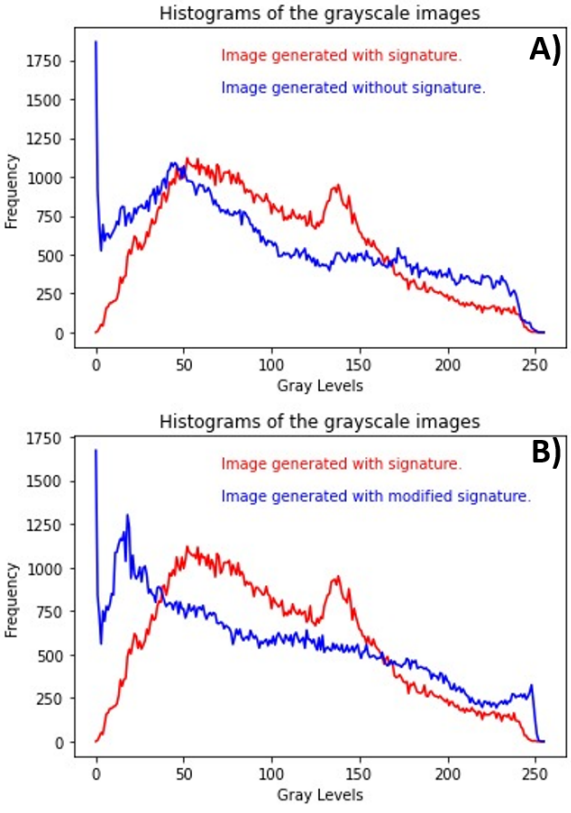

Fig. 6. Comparative histograms; A) Comparative histogram in gray scale between the image generated using the security keys and the image generated conventionally; B) Comparative grayscale histogram between the image generated using the security keys and the image generated from the altered signature.

In Fig. 6-(a), we conducted a comparative analysis of grayscale histograms between the image protected with our signatures and the image generated using a random vector z. This initial analysis emphasizes the importance of maintaining the integrity of the original parameters to ensure authenticity and consistency in image generation. The experiments reiterate the necessity of our signature proposal in the context of image generation through VQ-GAN+CLIP.

B. Signature protection

As we have evidenced, our proposal fulfills the intention of safeguarding original copyright; additionally, our method encloses an additional benefit. Although it is essential to have both security keys that protect our creation, it is imperative to keep the original parameters intact. In other words, any modification, however minimal, would result in entirely different images, as clearly illustrated in Fig. 5-(d), derived from the generation of the resultant image from an altered initial image Fig. 7-(b). In Fig. 7-(b), it's perceived that the input image (our second key) appears to be the same at first glance (Fig. 7-(a)), with no apparent modification or alteration; however, internally there was a minimal alteration of random pixels. This minimal alteration impacted both the initial image and the Z-vector, resulting in a completely different resultant image. To corroborate this, in Table II, it is observed that the PSNR value differs by 26.15 dB between the original and altered image. Similarly, with even more drastic changes, the initial images show a PSNR difference of 72.25 dB, thus explaining why in Fig. 5-(d) the contexts between the resultant images are so disparate, resembling more a resultant image with a random Z-vector. Finally, in Fig. 6-(b), a comparative histogram in grayscale is presented between the resultant image protected with our keys and the random one.

TABLE II
COMPARATIVE METRICS IN DIGITAL SIGNATURES

Compared image	**SSIM**	**PSNR (dB)**
Unaltered vs Altered Input Image	0.99	73.85
Unaltered vs Altered Initial image	-0.07	27.75

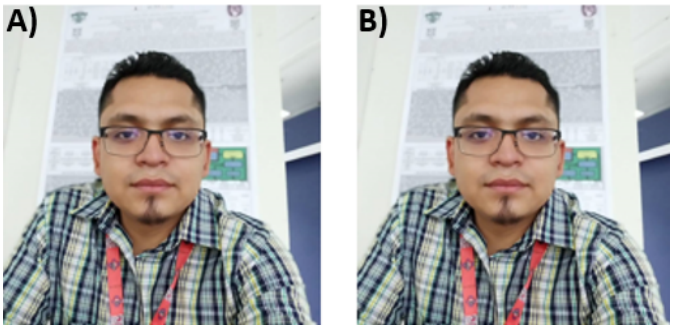

Fig. 7. Input images; A) Original image without alterations; B) Altered image.

This analysis reinforces the critical importance of preserving the original parameters to ensure the authenticity and protection of the generated creations. Likewise, it underscores the added value of our proposal in the preservation of copyright in the realm of data generation through generative methods, establishing a solid precedent for future research and improvements in this direction.

V. Conclusion

In this paper, we present a novel approach to ownership authentication and integrity verification of digital images. By leveraging the combined strengths of VQ-GAN and CLIP, we generate high-quality images guided by textual descriptions, ensuring that each image can be exactly recreated only if the specific digital signature and prompt are used. This guarantees authenticity and prevents unauthorized alterations of the digital content.

The results demonstrated how our approach is capable of differentiating between any variation and manipulation of the security layers. We observed that even the slightest alteration can drastically change the resulting image. Despite having a 0.99 SSIM, indicating that the input images are almost identical, the resulting image generated with the same prompt but with an altered digital signature shows a PSNR value of 73.85. This demonstrates that, although it maintains similar characteristics (due to the prompt), the original image can never be generated without the two unaltered keys, making the system extremely sensitive, which is precisely the desired outcome. This reinforces the integrity of the information and the authenticity of the generated content.

For future work, we plan to explore the extension of our approach to other types of multimedia data, such as audio and video. Additionally, knowing that the image will never be altered once the final image is obtained, we aim to develop a reversible process to retrieve the original creator's signature and, thereby, validate or provide evidence of who created the image. This will not only ensure the integrity of the image but also provide irrefutable proof of the digital content's authorship.

References

[1] M. Barni, F. Bartolini, "Applications," in *Watermarking Systems Engineering: Enabling Digital Assets Security and Other Applications*, Boca Raton, FL: CRC Press, 2004.

[2] M. Cedillo-Hernandez, A. Cedillo-Hernandez, M. Nakano-Miyatake, H. Perez-Meana, "*Improving the management of medical imaging by using robust and secure dual watermarking*," *Biomed. Signal Process. Control*, vol. 56, 101695, 2020. " Available: https://doi.org/10.1016/j.bspc.2019.101695

[3] I. Cox, M. Miller, J. Bloom, Ma. Miller, "Applications and properties," in *Digital Watermarking*, Morgan Kaufmann Publishers, USA, 2002. [Online]. https://www.elsevier.com/books/digital-watermarking/cox/978-1-55860-714-9

[4] N. Daneshmandpour, H. Danyali, and M. S. Helfroush, "Scalable image self-embedding based on dual-rate SPIHT-LDPC reference generation scheme,"*Radioeng*, vol. 28, pp. 199–206, 2019. Available: http://doi.org/10.13164/re.2019.0199

[5] M. Imran and B. Harvey, "A Blind Adaptive Color Image Watermarking Scheme Based on Principal Component Analysis, Singular Value Decomposition and Human Visual System,"*Radioengineering*, vol. 26, pp. 823-834, Sep. 2017. Available: https://doi.org/10.13164/re.2017.0823

[6] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," 2021. Available: https://doi.org/10.48550/arXiv.2103.00020

[7] P. Esser, R. Rombach, B. Ommer, "Taming Transformers for High-Resolution Image Synthesis," 2020. Available: https://doi.org/10.48550/arXiv.2012.09841

[8] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance,"in *European Conference on Computer Vision*, Springer, 2022, pp. 88–105.

[9] C. Paar, J. Pelzl, *Understanding Cryptography: A Textbook for Students and Practitioners*. Berlin: Springer-Verlag, 2010.

[10] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, E. Raff, "CLIP: Open Domain Image Generation and Editing with Natural Language Guidance," 2022. Available: https://doi.org/10.48550/arXiv.2204.08583

[11] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors (Online Source Style). *IEEE Trans. Plasma Sci.* [Online]. *21(3)*. pp. 876-880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

[12] J.-P. Aumasson, W. Meier, R. C.-W. Phan, and L. Henzen, "Blake2,"*The Hash Function BLAKE*, Springer, pp. 165–183, 2014.

[13] R. Rivest, "The MD5 message-digest algorithm,"1992. Available: https://www.rfc-editor.org/rfc/rfc1321