Visual Paraphrase Attack Safe and Distortion Free Image Watermarking Technique for AI-Generated Images

Shreyas Dixit[1]* **Ashhar Aziz**[2]* **Shashwat Bajpai**[3]*
Vasu Sharma[4] **Aman Chadha**[5,6]† **Vinija Jain**[5] **Amitava Das**[7]

[1]VIIT Pune, India [2]IIIT Delhi, India [3]BITS Pilani Hyderabad, India
[4]Meta AI, USA [5]Stanford University, USA [6]Amazon GenAI, USA
[7]AI Institute, University of South Carolina, USA

Abstract

A report by the European Union Law Enforcement Agency predicts that by 2026, up to 90% of online content could be synthetically generated EUROPOL (2022), raising concerns among policymakers, who cautioned that "Generative AI could act as a force multiplier for political disinformation. The combined effect of generative text, images, videos, and audio may surpass the influence of any single modality" Janjeva et al. (2023). In response, California's Bill AB 3211 mandates the watermarking california legislature (2023) of AI-generated images, videos, and audio. However, concerns remain regarding the vulnerability of invisible watermarking techniques to tampering and the potential for malicious actors to bypass them entirely. Generative AI-powered de-watermarking attacks, especially the newly introduced visual paraphrase attack Barman et al. (2024), have shown an ability to fully remove watermarks, resulting in a paraphrase of the original image. This paper introduces PECCAVI, the first visual paraphrase attack safe and distortion free image watermarking technique. In visual paraphrase attacks, an image is altered while preserving its core semantic regions, termed Non-Melting Points (NMPs). PECCAVI strategically embeds watermarks within these NMPs and employs multi-channel frequency domain watermarking. It also incorporates noisy burnishing to counter reverse-engineering efforts aimed at locating NMPs to disrupt the embedded watermark, thereby enhancing durability. PECCAVI is model-agnostic. All relevant resources and codes will be open-sourced[3].

1 Introduction - the Necessity & Urgency

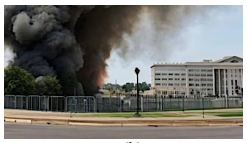
The proliferation of text-to-image generative AI models like Stable Diffusion(s) (Rombach et al., 2022), DALL-E(s) (Ramesh et al., 2021, 2022; Betker et al.), Midjourney (Midjourney, 2024), and Imagen (Saharia et al., 2022) has revolutionized visual content creation, unlocking unprecedented creative potential. However, this rapid evolution and widespread accessibility presents significant challenges, particularly concerning the misuse of AI-generated images.

*Equal contribution.

†Work does not relate to position at Amazon.

[3]PECCAVI is protected under a USA patent. In accordance with the Patent Act, adoption or research involving this technology for non-profit purposes is strictly prohibited.

(a) An AI-generated image of Pope Francis wearing a gigantic white puffer jacket went viral on social media platforms in March 2023, sparking discussions about the potential misuse of Generative AI technologies.



(b) A fake image purporting to show an explosion near the Pentagon was shared by multiple verified Twitter accounts, causing confusion and leading to a brief dip in the stock market.

(c) A screenshot from a video showing Taylor Swift holding a flag reading "Trump Won" went viral in February 2024, garnering over 4.5 million views on X. However, the content was confirmed to be AI-generated.

Figure 1: Some examples of the negative impacts of AI-generated images: (a) *"That Viral Image Of Pope Francis Wearing A White Puffer Coat Is Totally Fake"* (Forbes story), (b) *'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion* (CNN's cover story), and (c) *Viral Video Of Taylor Swift 'Endorsing' Donald Trump Is Completely Fake* (Forbes story).

In March 2023, an open letter (Future of Life Institute, 2023) signed by numerous AI experts and industry leaders called for a six-month halt on the development of AI systems more advanced than GPT-4. The central concern noted in the letter (Future of Life Institute, 2023) is "*Should we let machines flood our information channels with propaganda and untruth?*". While individual viewpoints on the notion of a moratorium may vary, the raised concern cannot be ignored. The findings of the latest (7^{th}) evaluation of the European Commission's Code of Conduct (Commission, 2022) that seeks the eradication of mis/dis-information online reveals a decline in companies' responsiveness. The percentage of notifications reviewed by companies within 24 hours decreased, falling from 90.4% in 2020 to 64.4% in 2022.

This decline likely reflects the increased accessibility of Gen AI models, leading to a notable influx of AI-generated content on the web. Approximately 3.2 billion images and 720,000 hours of video are uploaded to social media platforms daily (Thomson et al., 2020) (as of 2020). With all existing image watermarking techniques proving brittle against Gen AI-powered de-watermarking attacks (Barman et al., 2024), the need for developing robust, attack-resistant watermarking methods is more critical than ever.

AI-generated misinformation stands as one of the most formidable challenges in advancing responsible AI for society, as emphasized by leading figures including Geoffrey Hinton (Guardian, 2023), Bill Gates (CNBC, 2024), and Sundar Pichai (Kharpal, 2018), among others. Figure 1 illustrates recent instances of AI-generated misinformation that have caused significant disruptions.

2 Dewatermarking Attacks - Related Works

Digital watermarking has been a focus in computer vision research for 3–4 decades, primarily divided into two categories : (i) static, learning-free methods: such as DwtDctSVD (Navas et al., 2008), IA-DCT (Podilchuk and Ramchandran, 1998), and IA-W (Podilchuk and Ramchandran, 1998), among others; and (ii) learning-based methods, which represent the more contemporary, state-of-the-art approaches such as Stable Signature (Fernandez et al., 2023), Tree-Ring Watermark (Wen et al., 2023), Watermark Anything Model (WAM) (Sander et al., 2024), and ZoDiac (Zhou et al., 2024) etc. In addition to developing and evaluating various watermarking methods, researchers have also explored de-watermarking techniques, including classical image-altering methods such as (a) *brightness adjustment* (Loch, 2010), (b) *JPEG compression* (Yadav et al., 2012), and (c) *Gaussian noise addition* (Cox et al., 1997). More recently, advanced Generative AI-powered techniques have been introduced, such as (d) regeneration attacks (Zhao et al., 2023) and (f) adversarial purification (Nie et al., 2022).

While watermarking originated in computer vision, advancements in large language models (LLMs) have spurred interest in text watermarking. OpenAI, for instance, hinted at watermarking techniques for ChatGPT (Business Standard, 2024). Early LLM watermarking models by (Kirchenbauer et al., 2023) faced criticism after studies by (Sadasivan et al., 2024) and (Chakraborty et al., 2023) showed that paraphrasing could effectively remove these watermarks. This has spurred interest in visual paraphrase attacks (Barman et al., 2024) on image watermarks, a technique enabled by advances in text-to-image and image-to-text systems.

2.1 Visual Paraphrase attack

The concept of *visual paraphrasing attack*, first introduced in (Barman et al., 2024), refers to generating variations of an image that retain the same semantic content while altering visual presentation. An illustration is visible in Figure **??**, taken from the original paper. Unlike linguistic paraphrasing in natural language processing (e.g., "*What is your age?*" vs. "*How old are you?*"), visual paraphrasing utilizes image-to-image diffusion (Rombach et al., 2022) system, to adjust an image's visual representation while preserving its meaning. The authors examine two primary parameters for creating effective visual paraphrases: Strength and Guidance Scale. Their findings suggest that a specific range of Strength, paired with an optimal range of Guidance Scale, yields satisfactory dewatermarked paraphrased images. An example can be found in Figure 2. For a more detailed explanation, please refer to (Barman et al., 2024). Further details are reported in the Appendix: Visual Paraphrase.

The visual paraphrase attack is comparable to other generative AI-powered attacks, such as regeneration attacks (Zhao et al., 2023) and adversarial purification (Nie et al., 2022). Although the recently proposed ZoDiac (Zhou et al., 2024) has demonstrated resilience against regeneration and adversarial purification attacks, the visual paraphrase attack remains the most effective method for completely removing watermarks.

The visual paraphrasing attack, introduced by (Barman et al., 2024), takes a watermarked image and uses an image-to-image diffusion system to produce a watermark-free visual paraphrase. An alternative approach could involve: $watermarked\ image \Rightarrow image\ captioning\ system \Rightarrow image\ caption \Rightarrow text\ to\ image\ system \Rightarrow watermark\ free\ image$, which we term *open-ended visual paraphrase*. However, generating a near-identical image this way is nearly impossible due to the challenge of crafting a precise text prompt, resulting in countless variations and making the exact reproduction improbable, with potential seed values reaching 2^{64} (PyTorch Discussion Forum, 2022). Additionally, text-to-image (T2I) systems are inherently stochastic, producing different outputs each time, even with the same input prompt. Our image-to-image approach, however, ensures reliable adherence to the original image's appearance and meaning, even within this variable parameter space, delivering a consistent and structurally faithful visual paraphrase. Further discussion is in Appendix: Visual Paraphrase.

3 PECCAVI: Visual Paraphrase Attack Safe & Distortion-Free Image Watermarking Technique

To the best of our knowledge, there is no existing work on visual paraphrase attack-safe watermarking, making it challenging to directly compare PECCAVI with similar techniques. In designing PECCAVI, we considered several fundamental questions: (i) *where to place the watermark*, (ii) *which watermarking technique to use*, (iii) *the need for a more sophisticated detection mechanism*, (iv) *how to assess resistance to visual paraphrase attacks*, and (v) *whether the watermarking process distorts the original concept excessively*. We describe PECCAVI in detail with these guiding questions in mind to aid the reader's understanding. The overall pipeline of PECCAVI is illustrated in Figure 4.

3.1 Where to add watermark? - Non-Melting Points (NMPs)

Visual paraphrasing creates alternate visual representations of an image while preserving its core meaning. The key concept here is to identify regions that remain largely unaffected by paraphrasing. These relatively stable regions are ideal for embedding watermark signals, as they are less likely to be altered. We refer to these areas as Non-Melting Points/Regions (NMPs). Detecting NMPs involves two main steps:

Saliency detection: Salient region detection in image processing identifies the most "*salient*" or visually prominent areas within an image, based on unique features like color contrast, texture, or edges. Saliency detection is a well-established sub-discipline, with various methods; empirically, we found that (Kapishnikov et al., 2019) XRAI performed best in our experiments, followed by MSI-Net (Kroner et al., 2020) and hence we will be using these for our experiments in the rest of the paper. Refer to Figure **??** for a detailed illustration.

Non-Melting Points (NMPs): For a given image, we generate five automatically paraphrased versions using the method described in (Barman et al., 2024). In each paraphrased image, we identify key

regions and use Intersection over Union (IoU) to find the most stable areas—regions that consistently appear in similar locations across variations. These stable areas are referred to as Non-Melting Points (NMPs). To refine our NMP selection, we apply Non-Maximum Suppression (NMS), which eliminates redundant overlapping boxes and retains only the most representative regions. The final set of NMPs is then mapped onto a predefined patch grid, where each patch is evaluated for stability. Each region receives a stability score, which reflects how frequently it appears across the paraphrased images. Lower scores indicate greater consistency across variations, making those regions more reliable as NMPs. If no sufficiently stable regions are found, we include a default box to ensure robustness.

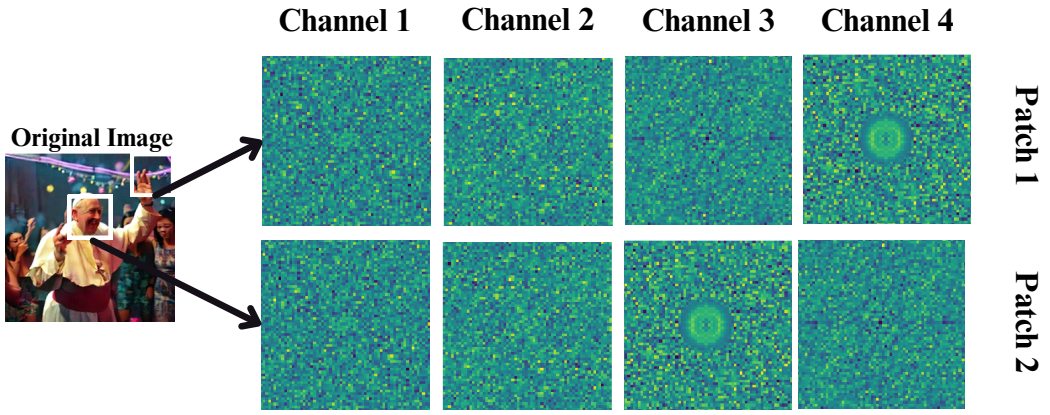


Figure 2: Illustration of multipatch watermarking, where watermark patterns are embedded in different channels of selected patches from the original image.

3.2 How to Watermark? - Strategies

Once NMPs are identified, the next essential step is determining how to embed watermark signals within them. We experimented with four strategies: (i) Baseline watermarking with SoTA methods like ZoDiac (Zhang et al., 2024) etc, (ii) watermark strength, (iii) single-channel strength watermarking, and (iv) multi-channel watermarking strength. Our results showed that multi-channel strength watermarking was the most effective approach.

Baseline watermarking with SoTA methods: After identifying NMPs, our first step was to embed watermarks using state-of-the-art techniques like ZoDiac (Zhang et al., 2024), Stable Signature (Fernandez et al., 2023) and WAM (Sander et al., 2024). We are particularly interested in evaluating which method provides more resilient watermarking against visual paraphrase attacks. **Watermark strength**: As argued by (Barman et al., 2024), stronger paraphrasing removes watermarks more effectively. Therefore, using a higher watermark strength in these NMPs should make PECCAVI more resilient.

Watermark strength is determined by the distance between rings within the watermark, with smaller distances indicating greater strength. For example, Channel 4 (Figure 2) shows a smaller ring distance (0.5), while Channel 3 reflects a larger distance (0.75). Strength values range from 0 to 1.0, depending on the number of paraphrases containing the NMP: $W_s = \max(0.1, 1 - 0.25 \cdot (n-1)), \quad n \in \{1, 2, 3, 4, 5\}$. Here n represent the number of regions that an NMP appears in out of the 5 paraphrases. **Single-Channel Watermarking vs. Multi-Channel Watermarking**: Fourier Space Watermarking (Gourrame et al., 2022) embeds watermarks into the frequency domain of an image rather than the traditional spatial domain (pixels). This approach increases the watermark's resilience to image manipulations, as the image undergoes Fourier Transform decomposition before the watermark is added to its components. We can embed watermarks across same or different channels for each patch. Fig. 2 provides an illustration adding watermarks in different channels.

Noisy Burnishing: Attackers may attempt to identify salient regions of an image to remove the watermark. This can be countered by adding adversarial noise to the watermarked image, which disrupts the detection of these salient regions, as proposed in (Gao et al., 2022). Fig. 3 demonstrates how salient regions become distorted following a noisy burnishing attack.

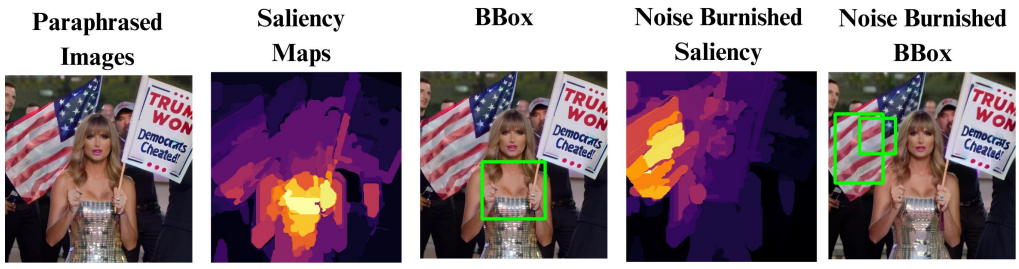


Figure 3: Noisy burnishing disrupts saliency detection in watermarked images, hindering attackers from locating NMPs or altering watermarked areas. This technique preserves the frequency-based watermark, ensuring high detectability while enhancing security against tampering.

3.3 Paraphrase Attack Safety

NMP-based watermarking faces two key challenges: (i) assessing the resilience of NMP-embedded watermarks against further paraphrasing, and (ii) anticipating potential countermeasures from attackers who may reverse-engineer methods to detect and distort NMPs, reducing watermark detectability. To address these, we propose two strategies: (a) random patching to embed additional watermarks, and (b) noisy burnishing to prevent NMP detection. Details of these techniques are provided below.
Random Patching: Since NMP detection relies on widely recognized saliency detection methods, such as XRAI (Kapishnikov et al., 2019), attackers could potentially reverse-engineer these methods to locate the salient regions where watermarks are embedded. To enhance security, we introduce a technique called random patching. This technique is simple yet effective: once all NMPs are detected and saved, we identify the smallest one among them and generate an additional NMP of the same shape at a random, non-overlapping location. The selection can be randomized using a vendor-specific pseudo-random algorithm. Watermarks are then embedded in these randomly placed patch, similar to the original patches, using either single-channel or multi-channel approaches.

3.4 PECCAVI - Watermark Detection

The PECCAVI watermark detection process, illustrated in Figure 4, employs a brute-force approach to detect watermarks across all patches of an image. The highest detection score among all processed patches is selected as the final score. Additionally, the method scans multiple channels to identify watermark signals and ultimately computes a Watermark Detection Probability for the image.

3.5 Adaptive Enhancement - Minimizing Distortion

We apply adaptive image enhancement from (Zhang et al., 2024) to improve watermarked image quality by blending it with the original: $\bar{x}_0 = \hat{x}_0 + \gamma(x_0 - \hat{x}_0)$ where $\gamma \in [0, 1]$ balances quality and watermark strength. The goal is to find the smallest γ such that similarity $S(\bar{x}_0, x_0) \geq s^*$, typically using SSIM. An illustration is shown in the Figure **??**.

4 Efficacy of PECCAVI

To assess PECCAVI's robustness, we evaluate two key dimensions: distortion and detectability. For image quality distortion, we use metrics such as PSNR and SSIM to measure the watermark's impact on visual fidelity, both perceptually and structurally. For watermark detectability, we analyze PECCAVI's resilience against classical attacks like brightness adjustments, Gaussian noise, JPEG compression, and varying paraphrasing strengths using the Average WDP. A summary of these metrics, presented in Table 1, highlights PECCAVI's effectiveness in preserving high image quality while ensuring robust watermark retention under diverse attack scenarios.

Meta's Watermark Anything Model (WAM) (Sander et al., 2024) (*released on 11th Nov*) enables imperceptible, localized image watermarking, embedding, locating, and decoding multiple watermarks

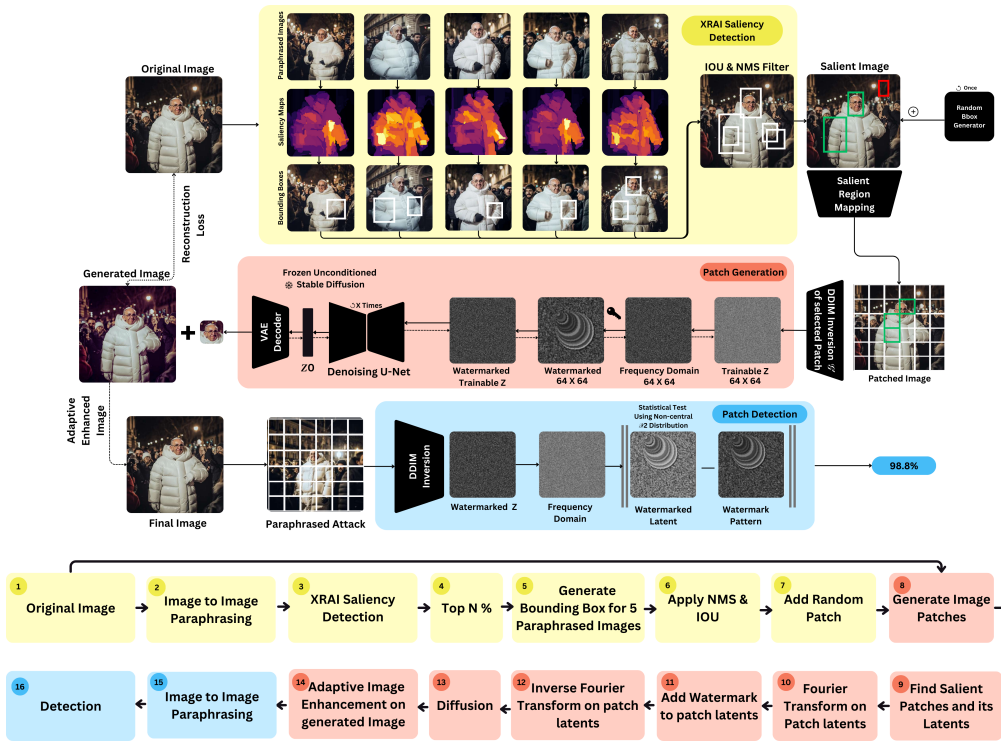


Figure 4: The PECCAVI pipeline for image watermarking encompasses NMP detection, multi-channel watermark embedding, adaptive enhancement. These components collectively ensure robust, low-distortion watermarks that resist paraphrase attacks, safeguarding AI-generated images from unauthorized alterations.

in small regions of high-resolution images. Our evaluation shows that PECCAVI outperforms WAM in resisting visual paraphrasing attacks.

4.1 Choice of T2I models

To evaluate the efficacy of PECCAVI, we tested it across diverse T2I models, including Stable Diffusion 3 (SD 3) (Esser et al., 2024; AI, 2024), Stable Diffusion XL (SDXL) (Podell et al., 2023), Stable Diffusion 2.1 (SD 2.1) (Rombach et al., 2022; AI, 2023), DALL-E 3 (Betker et al.), and Midjourney 6 (Midjourney, 2024). This process produced a dataset we call MS COCO_{AI}, where captions and images from the original MS COCO dataset (Lin et al., 2014) were fed into these models to generate and store corresponding images. A snapshot of the data can be viewed here. Results present in Table 1 present an average over all the images in MS COCO_{AI}.

4.2 Measuring Paraphrase Attack Safety

Paraphrased images were generated at different strengths s, with lower s-values keeping more original details and higher values allowing greater alteration. WDP assesses watermark retention, while SSIM measures similarity to the original image. PECCAVI shows high WDP at lower strengths, retaining watermark integrity even through moderate paraphrasing. Detection gradually decreases with higher s but remains effective, as shown in Table 1.

4.3 Measuring Distortion

A key requirement in image watermarking is minimal distortion of the original content. We assess this distortion using metrics like PSNR (Baig et al., 2019) and SSIM (Wang et al., 2004). Further details are reported in the Appendix: Distortion vs. Detectability. Together, these metrics provide a balanced view of pixel-level (PSNR, SSIM) distortion, helping us assess overall image quality. For distortion metric results refer to Table 1.

Figure 5: Comparison of two sets of images before and after adaptive enhancement, which optimizes watermark detectability while minimizing visual distortion.

4.4 Results

We compare the performance of PECCAVI watermarking scheme with various post-process image watermarking methods such as (Zhang et al., 2024), (Tancik et al., 2020), (Fernandez et al., 2023), (Navas et al., 2008) and (Ma et al., 2022). The methods were compared under the following attack schemes: (i) Brightness Enhancement with a factor of 0.5, (ii) Gaussian Noise with a std of 0.05, (iii) JPEG compression with a quality factor of 50, and (iv) Visual Paraphrasing (Barman et al., 2024), using `stable-diffusion-xl-base-1.0` with image captions and paraphrase strengths of 0.1 and 0.2. We further test our method on VAE-based image compression model (Ballé et al., 2018) and (Cheng et al., 2020) with a quality setting of 3, Stable Diffusion-based image regeneration model (Zhao et al., 2024) with 60 denoising steps using stable-diffusion-2-1-base; results provided in supplementary material.

The results were produced on 100 images randomly sampled from the COCO Dataset (Lin et al., 2014).

		Image Quality		Avg. Watermark Detection Probability (WDP)					
				Pre-Attack	Post-Attack				
Method	λ	PSNR	SSIM		Brightness	Gaussian Noise	JPEG	Paraphrase (s=0.1)	Paraphrase (s=0.2)
DwtDctSVD	-	41.04	0.988	0.98	0.01	0.14	0.65	0.00	0.00
Stable Signature	-	42.91	0.98	0.99	0.75	0.73	0.65	0.59	0.51
WAM	-	46.05	1.00	1.00	0.62	0.61	0.58	0.63	0.56
ZoDiac	-	28.47	0.92	1.00	0.92	0.90	0.89	0.81	0.70
PECCAVI with different saliency methods									
PECCAVI (Vanilla Integrated)	Top 30	31.50	0.95	0.96	0.96	0.95	0.96	0.72	0.69
	Top 40	31.26	0.94	0.95	0.97	0.97	0.95	0.72	0.68
	Top 50	31.31	0.95	0.96	0.96	0.97	0.97	0.73	0.68
PECCAVI (MSI Net)	Top 30	30.64	0.94	0.98	0.94	0.95	0.95	0.84	0.79
	Top 40	30.57	0.94	0.99	0.98	0.98	0.97	0.83	0.80
	Top 50	30.71	0.94	0.99	0.98	0.98	0.98	0.87	0.83
PECCAVI (XRAI)	Top 30	29.56	0.93	0.99	0.98	0.99	0.99	0.92	0.87
	Top 40	29.87	0.93	0.99	0.99	0.99	0.98	0.90	0.84
	Top 50	29.84	0.93	0.99	0.99	0.99	0.99	0.90	0.85

Table 1: Watermarked image quality is compared in terms of PSNR and SSIM scores. Watermark robustness is compared based on Average WDP before and after attacks on the MS-COCO dataset.

5 Conclusion

This paper introduces PECCAVI is the first visual paraphrase attack-safe, distortion-free image watermarking technique. With the rise of AI-generated misinformation, we believe PECCAVI will contribute significantly to the greater social good. It surpasses existing watermarking techniques like ZoDiac and WAM in performance, though it requires substantial computational resources. Due to space constraints, comparative analysis is provided in the appendix.

References

Stability AI. Stable diffusion 2.1, 2023. Available at: `https://huggingface.co/stabilityai/stable-diffusion-2-1`.

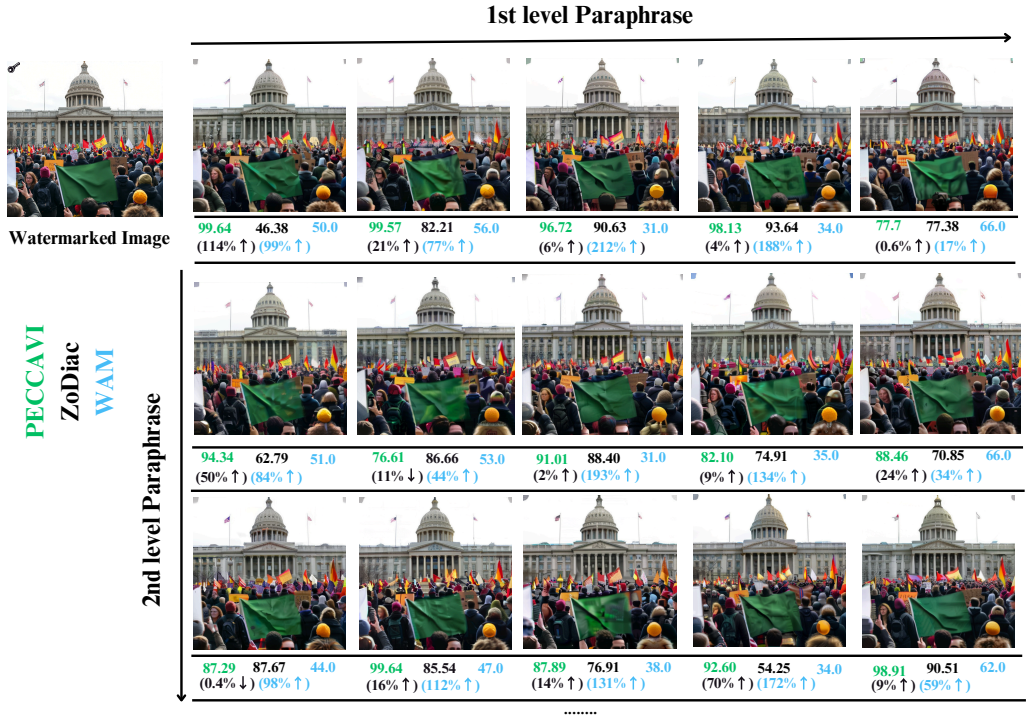


Figure 6: The cumulative impact of successive paraphrasing attacks on PECCAVI (green) and ZoDiac (black) watermarked images is depicted, with detection scores. PECCAVI shows superior resilience, maintaining stable scores under high-strength paraphrasing attacks, demonstrating its durability over ZoDiac.

Stability AI. Stable diffusion 3: Research paper. 2024.

Md Amir Baig, Athar A. Moinuddin, and E. Khan. Psnr of highest distortion region: An effective image quality assessment method. In *2019 International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–4, 2019.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior, 2018.

Niyar R Barman, Krish Sharma, Ashhar Aziz, Shashwat Bajpai, Shwetangshu Biswas, Vasu Sharma, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. The brittleness of ai-generated image watermarking techniques: Examining their robustness against visual paraphrasing attacks, 2024.

James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions.

Business Standard. Openai mulls watermarking chatgpt generated text, but treads with caution. *Business Standard*, 2024.

california legislature. Ab-3211 california digital content provenance standards. 2023.

Megha Chakraborty, S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Shreya Gautam, Tanay Kumar, Krish Sharma, Niyar Barman, Chandan Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. Counter Turing test (CT2): AI-generated text detection is not as easy as you may think - introducing AI detectability index (ADI). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2206–2239, Singapore, 2023. Association for Computational Linguistics.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules, 2020.

CNBC. Bill gates says this is the no. 1 unsolvable problem facing today's young people: 'the harm is done'. 2024.

European Commission. Eu code of conduct against online hate speech: latest evaluation shows slowdown in progress. 2022.

Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, 1997.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

EUROPOL. Facing reality?: Law enforcement and the challenge of deepfakes. 2022.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models, 2023.

Future of Life Institute. Pause giant ai experiments: An open letter, 2023.

Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Huazhu Fu, Wei Feng, Yang Liu, and Song Wang. Can you spot the chameleon? adversarially camouflaging images from co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2150–2159, 2022.

Khadija Gourrame, Frederic Ros, Hassan Douzi, Rachid Harba, and Rabia Riad. Fourier image watermarking: Print-cam application. *Electronics*, 11(2):266, 2022.

The Guardian. Godfather of ai' geoffrey hinton quits google and warns over dangers of misinformation. 2023.

Ardi Janjeva, Alexander Harris, Sarah Mercer, Alexander Kasprzyk, and Anna Gausen. The rapid rise of generative ai: Assessing risks to safety and security, 2023.

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. Xrai: Better attributions through regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4947–4956, 2019.

Arjun Kharpal. Google ceo sundar pichai: Fears about artificial intelligence are very legitimate, he says in post interview. *The Washington Post*, 2018.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, 2023.

Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Francis G. Loch. Image processing algorithms part 4: Brightness adjustment. *The Crypt Mag*, 2010.

Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1532–1542, 2022.

Midjourney. Midjourney: A new era of ai-generated art, 2024.

A. Navas et al. Digital watermarking techniques: A review. *Journal of Information Technology*, 23(4):345–360, 2008.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16805–16827. PMLR, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

C. Podilchuk and K. Ramchandran. Image watermarking: Capacity issues and applications. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 445–448. IEEE, 1998.

PyTorch Discussion Forum. What is the max seed you can set up?, 2022. Accessed: 2024-11-13.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected?, 2024.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages, 2024.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

T.J. Thomson, Daniel Angus, and Paula Dootson. 3.2 billion images and 720,000 hours of video are shared online daily: Can you sort real from fake? *The Conversation*, 2020. Accessed: 2024-11-04.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Rajesh K. Yadav et al. Study and analysis of wavelet based image compression techniques. *International Journal of Engineering, Science and Technology*, 4(1):1–7, 2012.

Lijun Zhang, Xiao Liu, Antoni Viros Martin, Cindy Xiong Bearfield, Yuriy Brun, and Hui Guan. Attack-resilient image watermarking using stable diffusion, 2024.

Xuandong Zhao, Kexun Zhang, Yu-Xiang Wang, and Lei Li. Generative autoencoders as watermark attackers: Analyses of vulnerabilities and threats. *arXiv preprint arXiv:2306.01953*, 2023.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai, 2024.

Yuan Zhou, Peng Zhang, Mengya Song, Alice Zheng, Yiwen Lu, Zhiheng Liu, Yong Chen, and Zhaohan Xi. Zodiac: A cardiologist-level llm framework for multi-agent diagnostics, 2024.