ICONMARK: ROBUST INTERPRETABLE CONCEPT-BASED WATERMARK FOR AI IMAGES

Vinu Sankar Sadasivan, Mehrdad Saberi & Soheil Feizi
Department of Computer Science
University of Maryland, College Park, USA
{vinu, msaberi, sfeizi}@cs.umd.edu

ABSTRACT

With the rapid rise of generative AI and synthetic media, distinguishing AI-generated images from real ones has become crucial in safeguarding against misinformation and ensuring digital authenticity. Traditional watermarking techniques have shown vulnerabilities to adversarial attacks, undermining their effectiveness in the presence of attackers. We propose IConMark, a novel in-generation robust semantic watermarking method that embeds interpretable concepts into AI-generated images. Unlike traditional methods, which rely on adding noise or perturbations to AI-generated images, IConMark incorporates meaningful semantic attributes, making it interpretable to humans and hence, resilient to adversarial manipulation. This method is not only robust against various image augmentations but also human-readable, enabling manual verification of watermarks. We demonstrate a detailed evaluation of IConMark's effectiveness, demonstrating its superiority in terms of detection accuracy and maintaining image quality. Moreover, IConMark can be combined with existing watermarking techniques to further enhance and complement its robustness. We introduce IConMark+SS, a hybrid approach combining IConMark with StegaStamp, to further bolster robustness against multiple types of image manipulations.

1 INTRODUCTION

With the rapid advancements in generative AI, distinguishing between AI-generated and real images has become a critical challenge. The proliferation of deepfake technologies and synthetic media has raised concerns about misinformation, copyright infringement, and digital authentication (Helmus, 2022). Traditional watermarking techniques, which add imperceptible noise or frequency domain modifications to images, have been widely used to establish provenance and protect intellectual property (Cox et al., 2007; Tancik et al., 2020; Bui et al., 2023; Wen et al., 2023; Sander et al., 2024; Fernandez et al., 2023). However, recent research has demonstrated the vulnerability of existing watermarking methods to adversarial attacks, including diffusion purification and model substitution adversarial attacks, which can remove or spoof watermarks with minimal image alterations (Saberi et al., 2023).

Watermarking methods can broadly be categorized into post-hoc and in-generation approaches. Post-hoc watermarks modify an image after the generation, embedding signals through additive noise or frequency-space perturbations (Tancik et al., 2020; Fernandez et al., 2023; Cox et al., 2007; Bui et al., 2023). In contrast, in-generation watermarks such as TreeRing (Wen et al., 2023) embed information during the image generation process, modifying the distribution of generated samples rather than making explicit post-hoc changes. While in-generation watermarking methods tend to be theoretically more robust, they are still susceptible to attacks, particularly adversarial strategies that attempt to remove or obfuscate the watermark (Saberi et al., 2023).

In this work, we introduce a robust Interpretable Concept-based Watermark (IConMark), a novel in-generation semantic watermarking approach that embeds interpretable concepts into AI-generated images. Unlike traditional watermarks, which primarily rely on additive noise, our method integrates meaningful semantic attributes, making it robust against various image perturbations and adversarial purification attacks. Since these concepts are interpretable to humans, we can also manually check

if an image generated with IConMark is watermarked. This novelty makes IConMark interpretable and robust to adversarial attacks, potentially making it a strong technique for manual image forensics with the help of human experts. IConMark can also automate detection using a visual language model, which queries the presence of these embedded concepts. This ensures both machine verifiability and human interpretability. Furthermore, we demonstrate that IConMark can be effectively combined with existing watermarking approaches since it only perturbs the input prompt given to the image generation model. We demonstrate that this ability of our IConMark to complement existing watermarking approaches can result in strong robust image watermarking techniques without much image quality degradation.

Our contributions are as follows:

- We propose IConMark (Section 3), a novel semantic watermarking method that embeds interpretable concepts rather than additive noise, improving the robustness and interpretability of watermarking.
- We demonstrate that IConMark can be effectively combined with post-hoc watermarking techniques to further enhance robustness against image augmentations (Section 4).
- We show that our method is resistant to various image augmentation attacks, including diffusion purification attacks (Section 5). IConMark being interpretable makes it easier to be detected by human experts, hence making our method resilient to adversarial attacks.
- We evaluate the quality of images watermarked using IConMark, showing that it maintains high visual quality while ensuring watermark detection integrity.

2 Related Works

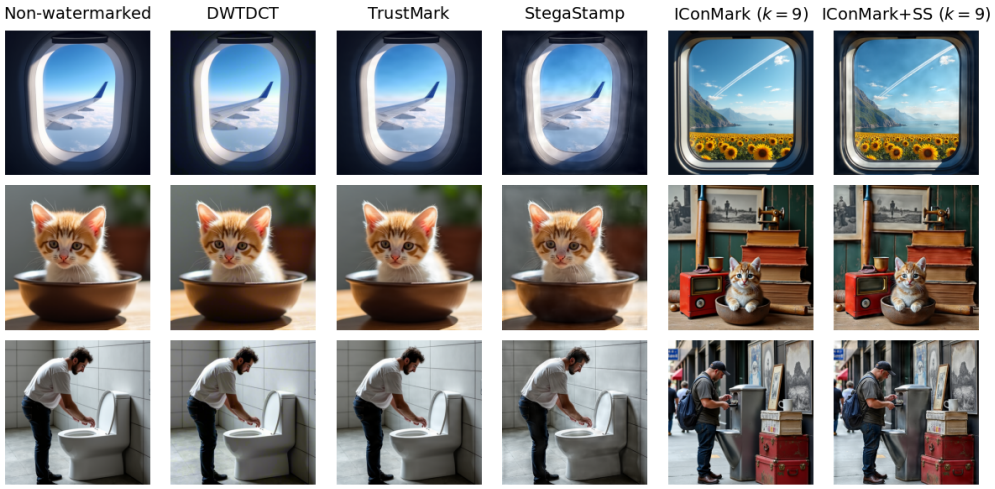
Traditional and deep learning-based watermarking has long been a crucial tool for copyright protection, content authenticity, and AI-generated media detection (Honsinger, 2002; Swanson et al., 1998). Classical techniques embed signals in the spatial or frequency domain, leveraging transformations such as Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) (Al-Haj, 2007; Cox et al., 2007). Deep learning-based watermarking methods, such as StegaStamp (Tancik et al., 2020), StableSignature (Fernandez et al., 2023), TrustMark (Bui et al., 2023), and Watermark Anything Model (Sander et al., 2024) have further improved robustness by training neural networks to encode and extract watermarks from images. However, these methods remain susceptible to sophisticated removal techniques, including adversarial attacks and diffusion purification (Saberi et al., 2023).

Saberi et al. (2023) demonstrated that diffusion purification attacks, which leverage denoising diffusion models, can effectively remove low-perturbation watermarks by reconstructing images with minimal modifications. Similarly, WAVES (An et al., 2024), a benchmarking framework, revealed that watermark detection accuracy significantly degrades under adversarial and regeneration attacks, highlighting the need for more robust watermarking solutions. Additionally, high-perturbation watermarking methods, such as TreeRing (Wen et al., 2023), have been shown to be more resistant to removal but are still vulnerable to adversarial model substitution attacks (Saberi et al., 2023).

Furthermore, research has shown that watermarking systems are not only vulnerable to removal but also to spoofing attacks, where real images are falsely classified as watermarked, leading to false attributions and reputational risks (Saberi et al., 2023). Adversarial perturbations designed to mimic the statistical properties of watermarked images have been shown to effectively fool detection systems, raising questions about the reliability of current watermarking schemes. Our work builds on these findings by introducing an approach that is inherently more interpretable and robust to such attacks.

The concept of semantic watermarking, which embeds meaningful and interpretable information into images, has been relatively unexplored. Existing methods primarily focus on imperceptibility and robustness but often lack interpretability. Our work builds on these foundations by integrating interpretable concepts into AI-generated images, ensuring both human and machine verifiability while maintaining robustness against adversarial attacks. Our approach aligns with recent trends in AI provenance tracking and content authentication, presenting a compelling direction for the future of watermarking technologies.

(a) IConMark generation for various numbers of selected concepts k.



(b) Comparing images with various watermarking techniques.

Figure 1: Comparing images generated with different watermarking techniques. The images in the first column are non-watermarked AI-generated images from the Flux model. For each of the rows of the images, the main user prompts p for generation are "A view from a window on board an airplane flying in the sky.", "A small kitten is sitting in a bowl.", and "A man getting a drink from a water fountain that is a toilet.", respectively. In Figure 1a, we show different ablations of IConMark over the number of sampled concepts from the concept database. Figure 1b shows the comparison of our methods IConMark and IConMark+SS with baseline techniques such as DWTDCT (Cox et al., 2007), TrustMark (Bui et al., 2023), and StegaStamp (Tancik et al., 2020).

3 ICONMARK: INTERPRETABLE CONCEPT-BASED WATERMARKING

In this section, we describe our proposed method IConMark in detail. For every user prompt for image generation, IConMark samples related concepts from a private database. IConMark augments the user prompt with the sampled concepts and uses this augmented user prompt to query the image generator. The AI-image generator performs in-generation watermarking by adding these interpretable syntactical signatures or concepts to the generated images. At detection time for a candidate image, IConMark checks for the presence of various concepts from the private concept database with the aid of a visual language model. If the candidate image has more than a threshold number of concepts from the private concept database, it is classified as watermarked. Below, we describe various steps for the proposed watermarking pipeline.

3.1 INITIALIZATION: CONCEPT DATABASE GENERATION

We prompt ChatGPT (OpenAI, 2023) to generate multiple image concepts that can be added to an image. We instruct the model to generate concepts describing simple objects with a unique detail. For example, " brass table lamp" or "a metal blue street sign". We also instruct the model to generate concepts that can occur in various settings such as indoors, nature, sky, forest, streets, etc. We then manually craft a diverse database $\mathcal{D}$ with N concepts. Note that the concept database generation needs to be performed pre-hoc only once for our setup. This $\mathcal{D}$ can be later used to automatically generate or detect any new IConMark watermarked image without any manual intervention as shown in the following subsections.

3.2 WATERMARKED IMAGE GENERATION

For every user prompt p, IConMark samples top-k related concepts from the private concept database $\mathcal{D} = \{c_1, c_2, ..., c_N\}$. IConMark prompts Llama-3.1-8B-Instruct model, $\mathcal{L}$, to sample the related concepts. Llama gets the database $\mathcal{D}$, the user prompt p, and the number of concepts k to be sampled as inputs using a custom system prompt template to sample concepts $c_p^1, c_p^2, ..., c_p^k \in \mathcal{D}$. Here is the custom prompt template for $\mathcal{L}$:

System prompt: Here is a database of N concepts: \n \n c_1 \n c_2 \n ... \n c_N \n \n
In an image of 'p', what are the top k related concepts from this database that can very likely occur in the background of this image? Consider only concepts that are related to this given image. For example, an image of a lion cannot have a basketball or a table in the background, whereas an image of a bird can have a tree or a mountain in the background. The concepts should be ONLY from the database of concepts given above. You should NOT generate new concepts.
User: Print each of the k related concepts verbatim between <a >and .

IConMark then uses a prompt template to generate an augmented user prompt $\tilde{p}_k$. IConMark passes $\tilde{p}_k$ to the image generator $\mathcal{G}$ to generate a watermarked image with the syntactical concept-based signatures. Here is the augmented prompt template for $\tilde{p}_k$:

p in the foreground. add following: \n c_p^1 \n c_p^2 \n ... \n c_p^k. \n \n sharp, detailed.

3.3 WATERMARK DETECTION

For a candidate image x, IConMark has access to the private concept database $\mathcal{D}$ and $\mathcal{V}$, a visual language model, IDEFICS3-8B-Llama3. IConMark prompts $\mathcal{V}$ to check for the presence of each of the N concepts in $\mathcal{D}$ given the image x using a custom prompt template. Here is the prompt template for prompting $\mathcal{V}$ to detect the presence of a concept $c_i \in \mathcal{D}$:

Image input: x
Text input: Print yes or no. Is there something like 'c_i'?

The detection score is the number of objects in $\mathcal{D}$ that were detected in x by $\mathcal{V}$. If the detection score is greater than a threshold τ, the candidate image x is classified as watermarked. Else, the image is labeled as non-watermarked.

4 ICONMARK+SS: HARNESSING STEGASTAMP AND ICONMARK

In this section, we harness the combined strength of StegaStamp (Tancik et al., 2020) and our method IConMark. An et al. (2024) endorse StegaStamp for its resilience to various image manipulations. StegaStamp stood out as a robust watermark among other popular techniques such as TreeRing (Wen et al., 2023) and StableSignature (Fernandez et al., 2023) in the WAVES benchmarking by An et al. (2024).

However, unlike IConMark these prior watermarking techniques lack interpretability and are not robust to adversarial attacks (Saberi et al., 2023). Since IConMark is a complementary technique with any other watermarking technique, we propose to combine the powers of both StegaStamp and our method IConMark. We name this method IConMark+SS.

IConMark+SS generates IConMark watermarked images and then applies post-hoc watermarking using StegaStamp. At detection time, IConMark+SS labels an image as watermarked if either the IConMark or StegaStamp detectors label it as watermarked. If both the detectors label the image as non-watermarked, only then the image is labeled as non-watermarked by IConMark+SS. This makes IConMark+SS robust to all the attacks that both StegaStamp and IConMark are resilient to.

5 EXPERIMENTS

In this section, we provide the experiments to demonstrate the effectiveness of our watermarks. Section 5.1 provides the experimental setup of our work describing the baselines, dataset, models, and metrics used. In Section 5.2, we show the performance of IConMark and IConMark+SS when compared to other baselines. We also perform ablation studies over hyperparameters used for IConMark and quantify the effect of changes in image quality due to watermarking. See Figure 1a for images generated via IConMark with different values of k. Section 5.3 shows the experiments testing various watermarks in the presence of different image manipulation techniques to study their robustness to such modifications. Our experiments show that IConMark, being the only interpretable watermark, is also the clear winner in the presence of certain image modifications when compared to other baselines. We use this to our advantage to complement the power of IConMark with StegaStamp to demonstrate the robustness of IConMark+SS. In our experiments, IConMark+SS obtains the best detection performance without trading-off the image quality.

5.1 EXPERIMENTAL SETTINGS

Baselines. We compare IConMark to several recent watermarking baselines, including StegaStamp (Tancik et al., 2020), TrustMark (Bui et al., 2023), and DwtDctSVD (Cox et al., 2007). All methods employ 100-bit binary watermark keys and encode the watermark by computing additive noise patterns that are applied to the images. Due to the additive nature of these watermarks, they can be easily integrated with IConMark for higher robustness (see Section 4).

Dataset and Models. We use 108 captions from the MS-COCO dataset (Lin et al., 2014) to generate AI images using the FLUX.1-dev model (Labs, 2024). Throughout the paper, we refer to the image generation model $\mathcal{G}$ as Flux. For sampling the top related image concepts for IConMark, we use the Llama-3.1-8B-Instruct model (Dubey et al., 2024). Throughout the paper, we refer to the language model $\mathcal{L}$ as Llama. We use IDEFICS3-8B-Llama3 (Laurençon et al., 2024) as the visual language model for measuring IConMark detection score. Throughout the paper, we refer to the visual language model $\mathcal{V}$ as Idefics. In all our experiments, we generate 10 different images per prompt with $\mathcal{G}$. Hence, we use 1080 non-watermarked images and 1080 watermarked images for our main experiments, totaling 2160 images in our evaluation. For all our detection robustness experiments, we halve our dataset size to 1080 images. For IConMark, we use a concept database of size $N = 100$ and, by default, sample $k = 9$ concept to augment the user prompt.

Detection Metrics. We plot the Receiver Operating Characteristic (ROC) curves or the True Positive Rates (TPR) vs. False Positive Rates (FPR) for the detection tasks. We measure the area under the ROC curves (AUROC) and the accuracy of the detection methods. We also measure TPR at 5% FPR (T@5%F) and TPR at 1% FPR (T@1%F). It is quite straightforward to measure these detection metrics for any watermarking detector that provides a continuous range of detection scores. However, for IConMark+SS, we do not have an explicit detection score since it performs detection conditioned on the output of both IConMark and StegaStamp detectors. Therefore, we measure the TP values of IConMark+SS at different FP values by varying their detection thresholds. This gives us the TPR and FPR for IConMark+SS. To obtain the ROC curves, we then sample the Pareto-optimal detection threshold pairs of IConMark and StegaStamp, ensuring that no selected point in the curve has a higher FPR for the same or lower TPR. After obtaining the ROC curve in this manner, we measure the detection metrics for IConMark+SS.

Image quality metrics. We use Clip Score (Radford et al., 2021; Hessel et al., 2021), Ratings (aesthetic score), Artifacts (Xu et al., 2024), and Diversity (Astolfi et al., 2024) metrics to measure the image generation qualities. Clip score measures the similarity of the generated image with the input prompt text with respect to the CLIP model (Radford et al., 2021). Ratings and Artifacts measure the changes in aesthetic and artifact features of images using a fine-tuned CLIP image

reward model. Diversity quantifies the capability of an image generator to produce diverse images for a prompt.

5.2 Watermark Detection And Image Quality

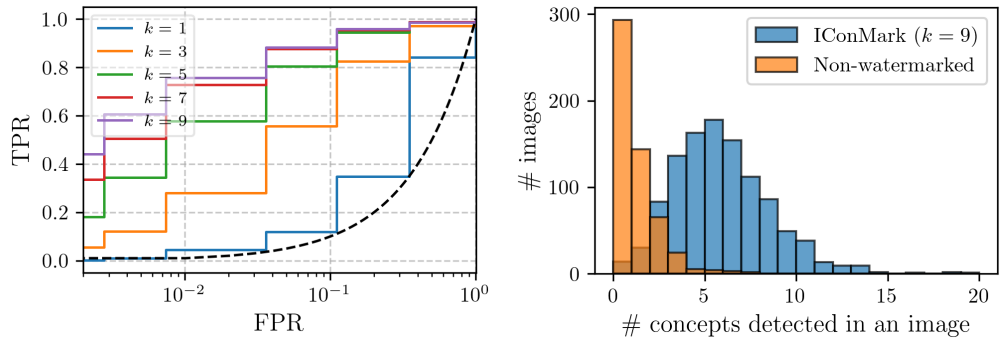


Figure 2: (Left) ROC curves for ablations of k for our proposed method IConMark. The black dashed lines indicate the ROC curve of a random detector. (Right) Number of concepts detected in watermarked (IConMark $k = 9$) and non-watermarked images by the IDEFICS3 visual language model.

	Detection (%) ↑				**Quality**			
k	AUC	Accuracy	T@5%F	T@1%F	Clip Score ↑	Ratings ↑	Artifacts ↓	Diversity ↑
1	76.05	74.58	11.85	4.44	0.026 ± 0.030	6.073 ± 1.016	2.156 ± 0.631	0.285 ± 0.014
3	92.04	85.65	55.65	28.06	0.029 ± 0.031	6.069 ± 1.042	2.164 ± 0.633	0.307 ± 0.018
5	96.47	91.67	80.46	57.69	0.031 ± 0.032	6.123 ± 0.998	2.139 ± 0.622	0.325 ± 0.019
7	97.31	92.18	87.69	72.78	0.032 ± 0.032	6.250 ± 1.028	2.102 ± 0.613	0.342 ± 0.021
9	97.46	92.41	88.24	75.65	0.033 ± 0.032	6.282 ± 1.084	2.100 ± 0.632	0.349 ± 0.022

Table 1: Comparison of IConMark watermark detection and quality metrics for different ablations of the number of sampled concepts k.

	Detection (%) ↑				**Quality**			
Methods	AUC	Accuracy	T@5%F	T@1%F	Clip Score ↑	Ratings ↑	Artifacts ↓	Diversity ↑
DWTDCT	100.00	100.00	100.00	100.00	0.036 ± 0.028	5.431 ± 1.045	2.368 ± 0.646	0.278 ± 0.012
TrustMark	100.00	100.00	100.00	100.00	0.032 ± 0.028	5.860 ± 1.054	2.264 ± 0.639	0.274 ± 0.013
StegaStamp	100.00	100.00	100.00	100.00	0.029 ± 0.027	5.475 ± 1.059	2.452 ± 0.630	0.284 ± 0.013
IConMark	97.46	92.41	88.24	75.65	0.033 ± 0.032	6.282 ± 1.084	2.100 ± 0.632	0.349 ± 0.022
IConMark+SS	100.00	100.00	100.00	100.00	0.028 ± 0.029	6.158 ± 1.016	2.303 ± 0.578	0.361 ± 0.021

Table 2: Comparison of watermark detection and quality metrics across different methods.

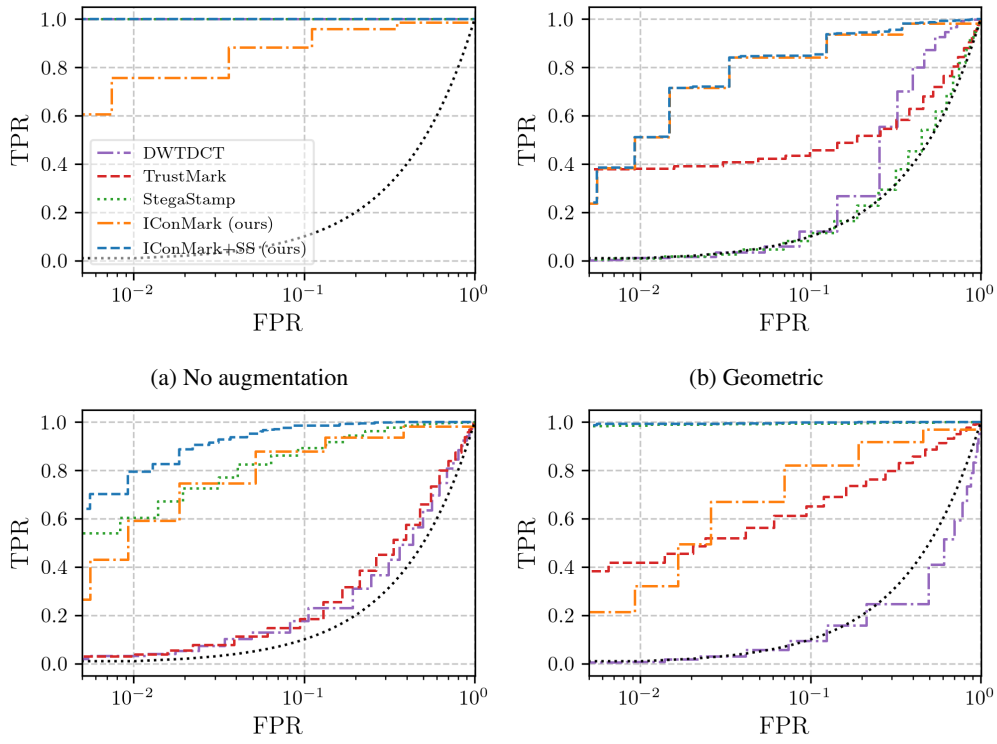
In this section, we demonstrate the watermark detection capability of our method when compared to the baseline approaches. We also measure the image quality of the generated images. See example generated images we use in Figure 1 (more in Appendix Figures 4–9).

In Figure 2, we provide the ROC curves for IConMark with different ablations of top-k or the number of concepts sampled from the database $\mathcal{D}$. As shown in the plot, the detection performance of IConMark improves as the number of sampled concepts increases. The AUROC and T@1%F values rise over 21% and 71%, respectively, as k changes from 1 to 9. Figure 2 also shows the frequency histogram of the number of concepts detected in the IConMark (k=9) watermarked and non-watermarked (k=0) images. As shown in the histogram, the watermarked images have a much higher likelihood of having concepts from the database than the non-watermarked images.

We also measure the detection and image quality metrics for IConMark for various values of k and provide them in Table 1. As shown in the table, the quality of the watermarked images does not degrade with different ablations of k, although it shows an increasing trend for the quality of images

as k increases. In the rest of the analysis, we fix $k = 9$ for IConMark since that gives the best detection results without a degradation in the image quality. Table 2 compares the detection and quality metrics of our watermarks IConMark and IConMark+SS with baselines. As shown in the table, IConMark has lower detection scores than other methods. However, IConMark+SS performs at par with the baselines. We also note that our methods offer a higher image quality Rating and Diversity, while other quality metrics are comparable with respect to the baselines.

5.3 Robustness of Watermarks



(a) No augmentation (b) Geometric

(c) Regen (d) Valuemetric

Figure 3: ROC curves of various watermarking techniques in the presence of various image augmentations. Black dotted curves indicate the performance of a random detector. As shown, our IConMark+SS performs the best.

	No augmentations				Geometric				Regen				Valuemetric			
Methods	AUC	Acc	T@5%F	T@1%F	AUC	Acc	T@5%F	T@1%F	AUC	Acc	T@5%F	T@1%F	AUC	Acc	T@5%F	T@1%F
DWTDCT	100.00	100.00	100.00	100.00	73.52	70.32	3.43	1.20	60.29	56.90	12.87	3.24	45.50	51.76	5.74	0.83
TrustMark	100.00	100.00	100.00	100.00	70.00	68.89	42.31	38.15	63.07	59.35	11.30	2.96	85.75	78.75	56.20	41.94
StegaStamp	100.00	100.00	100.00	100.00	55.41	54.58	4.72	1.02	96.54	90.09	82.50	60.46	99.83	98.98	99.26	98.52
IConMark	97.46	92.41	88.24	75.65	96.27	90.65	84.07	51.11	96.14	91.39	74.63	59.26	93.40	87.50	67.04	32.22
IConMark+SS	**100.00**	**100.00**	**100.00**	**100.00**	**96.32**	**90.65**	**84.63**	**51.11**	**99.05**	**95.46**	**95.19**	**79.44**	**99.93**	**99.35**	**99.44**	**99.26**

Table 3: Comparison of watermark detection across different augmentation settings. As shown, our IConMark+SS performs the best.

In this section, we evaluate the robustness of our watermarks in the presence of various image manipulation techniques (Sander et al., 2024). We employ three types of composite modifications to evaluate watermarks' robustness. Each composite modification, as shown below, comprises one or more elementary perturbations designed to emulate typical real-world degradations:

Geometric. Random rotation (in range $[-20°, 20°]$) and random cropping retaining 70%–95% of the original area.

Valuemetric. Photometric distortions, including brightness and contrast adjustments (factors in $[1.4, 1.7]$), Gaussian blur (radius between 1 and 3 pixels), additive Gaussian noise (standard deviation in $[0.05, 0.15]$), and JPEG compression (quality in the range $[40, 70]$).

Regen. Image regeneration (Saberi et al., 2023; An et al., 2024) using a diffusion model with 300 diffusion steps.

For examples of these types of augmented images, see Figure 10. We display our results to demonstrate the robustness of our watermarks in Table 3 and Figure 3. As shown in the results, our method IConMark+SS is the clear winner regarding the detection performance. IConMark+SS gets its strength from combining the resilience of IConMark to geometric and regen augmentations, and the resilience of StegaStamp to valuemetric augmentations.

6 Conclusion

IConMark represents a significant advancement in AI image watermarking, offering both interpretability and robustness against a wide range of adversarial threats. By embedding semantic concepts, it enables watermark detection by both humans and machines, making it a reliable tool for image forensics and authentication. Integrating IConMark with existing techniques like StegaStamp (IConMark+SS) further strengthens its resilience, establishing it as the most robust method compared to baselines. Our experimental results demonstrate that our watermarks achieve high detection accuracy while preserving image quality, ensuring practical usability. This work lays the foundation for more secure and interpretable watermarking methods in the era of generative AI.

However, IConMark relies on AI models for concept generation, sampling, and detection, which may introduce performance bottlenecks. A well-curated and diverse concept database is crucial for maintaining high detection rates and ensuring alignment with user requests. As AI models continue to advance, IConMark is expected to become even more effective. Future research should investigate the influence of different datasets and models on their performance to further enhance their robustness.

Additionally, IConMark may be less effective when users generate images with highly specific prompts, potentially limiting its practicality in certain scenarios. This challenge aligns with theoretical findings on the limitations of AI-generated content detection in low-entropy output spaces, as discussed in Sadasivan et al. (2023) and Saberi et al. (2023). Nevertheless, we believe our novel approach to interpretable watermarks opens up an exciting new avenue for research in this field.

Acknowledgements

The authors thank Matthijs Douze, Jakob Verbeek, and Pierre Fernandez for their support and mentorship throughout the project. This project was supported in part by a grant from an NSF CAREER AWARD 1942230, ONR YIP award N00014-22-1-2271, ARO's Early Career Program Award 310902-00001, Army Grant No. W911NF2120076, the NSF award CCF2212458, NSF Award No. 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS), a MURI grant 14262683, an award from meta 314593-00001 and an award from Capital One.

References

Ali Al-Haj. Combined dwt-dct digital image watermarking. *Journal of computer science*, 3(9): 740–746, 2007.

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Benchmarking the robustness of image watermarks. *arXiv preprint arXiv:2401.08573*, 2024.

Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models. *arXiv preprint arXiv:2406.10429*, 2024.

Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Universal watermarking for arbitrary resolution images. *arXiv preprint arXiv:2311.18297*, 2023.

Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edition, 2007. ISBN 9780080555805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models, 2023. URL https://arxiv.org/abs/2303.15435.

Todd C Helmus. Artificial intelligence, deepfakes, and disinformation. *RAND Corporation*, pp. 1–24, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Chris Honsinger. Digital watermarking. *Journal of Electronic Imaging*, 11(3):414, 2002.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

OpenAI. Chatgpt: An ai language model, 2023. URL https://openai.com/chatgpt. Accessed: YYYY-MM-DD.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. *arXiv preprint arXiv:2310.00076*, 2023.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. *arXiv preprint arXiv:2411.07231*, 2024.

Mitchell D Swanson, Mei Kobayashi, and Ahmed H Tewfik. Multimedia data-embedding and watermarking technologies. *Proceedings of the IEEE*, 86(6):1064–1087, 1998.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix

A.1 Additional Examples of Generated Images

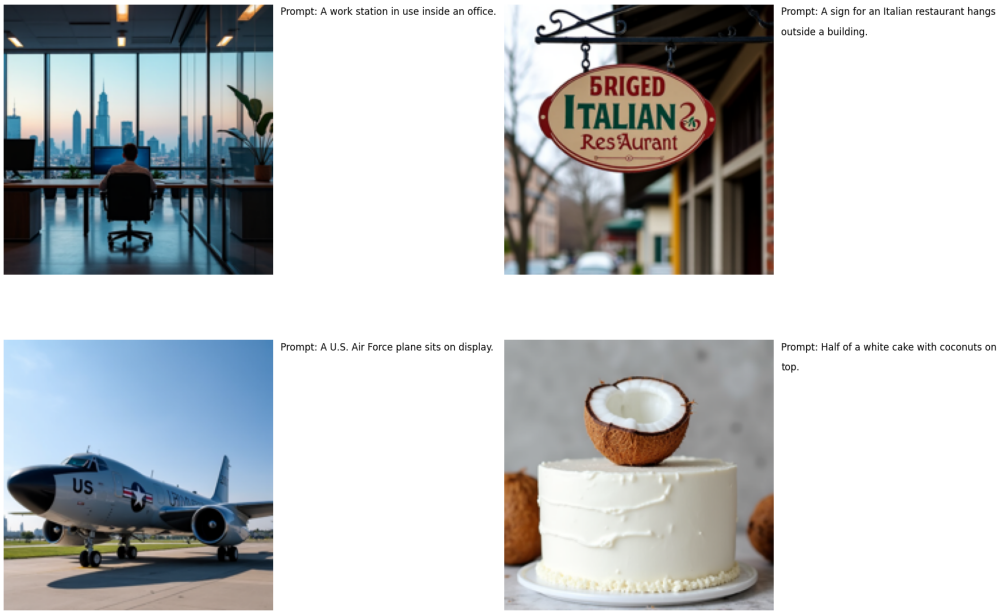


Figure 4: Non-watermarked images with their corresponding prompts for image generation using the Flux model.

Figure 5: IConMark watermarked images ($k = 1$) with their corresponding prompts for image generation using the Flux model and detected concepts from the concept database $\mathcal{D}$ using the IDEFICS3 visual language model.

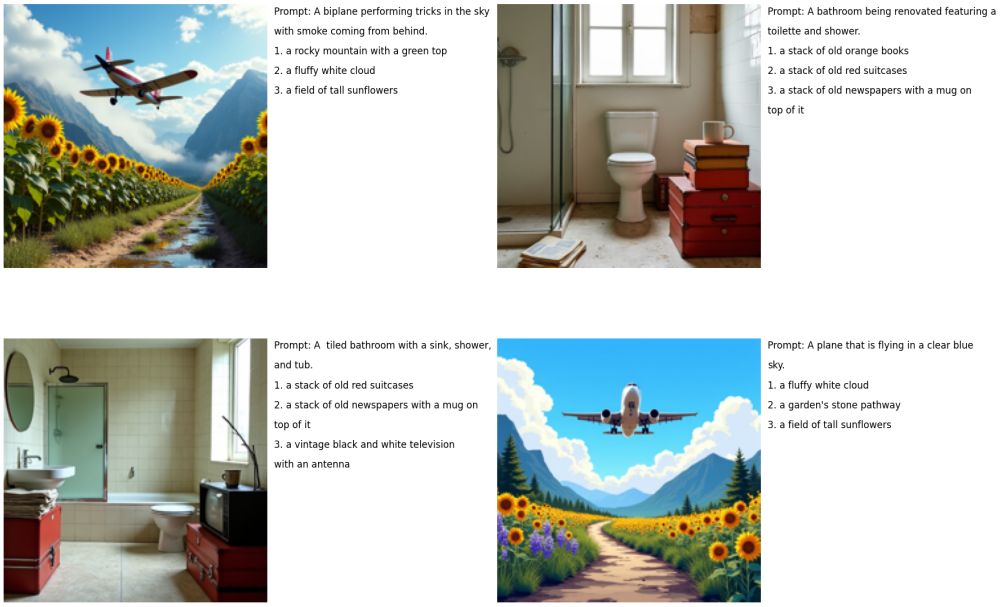


Figure 6: IConMark watermarked images ($k = 3$) with their corresponding prompts for image generation using the Flux model and detected concepts from the concept database $\mathcal{D}$ using the IDEFICS3 visual language model.

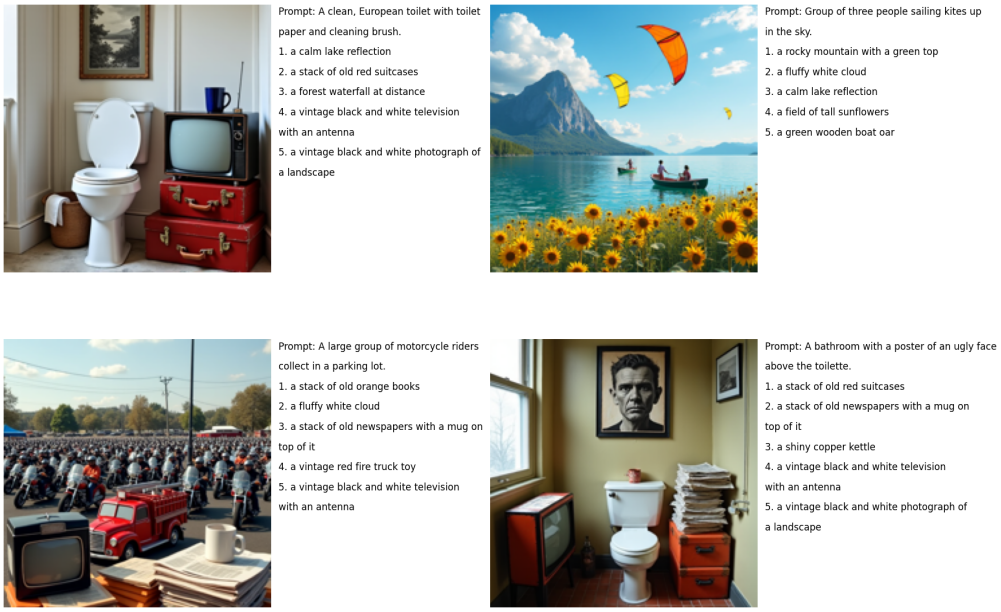


Figure 7: IConMark watermarked images ($k = 5$) with their corresponding prompts for image generation using the Flux model and detected concepts from the concept database $\mathcal{D}$ using the IDEFICS3 visual language model.

Prompt: A children play area absent of any children.
1. a worn blue leather armchair
2. a stack of old orange books
3. a vintage sewing machine
4. a stack of old newspapers with a mug on top of it
5. a vintage black and white photograph of a landscape
6. a vintage metal harmonium
7. a small potted venus flytrap plant

Prompt: Group of three people sailing kites up in the sky.
1. a rocky mountain with a green top
2. a fluffy white cloud
3. a plane contrail
4. a sailboat on the horizon
5. a beachside lifeguard tower
6. a calm lake reflection
7. a field of tall sunflowers

Prompt: A small child climbs atop a large motorcycle
1. a rocky mountain with a green top
2. a fluffy white cloud
3. a calm lake reflection
4. a stack of old red suitcases
5. a vintage copper microscope
6. a vintage astronomical globe
7. a vintage metal harmonium

Prompt: A desk with multiple computer screens forming one large screen.
1. a stack of old orange books
2. a stack of old newspapers with a mug on top of it
3. a shiny copper kettle
4. a vintage copper microscope
5. a vintage red fire truck toy
6. a vintage black and white photograph of a landscape
7. a vintage astronomical globe

Figure 8: IConMark watermarked images ($k = 7$) with their corresponding prompts for image generation using the Flux model and detected concepts from the concept database $\mathcal{D}$ using the IDEFICS3 visual language model.

Prompt: A man getting a drink from a water fountain that is a toilet.
1. a silver tea set
2. a metal blue street sign
3. a cratered moon surface
4. a stack of old red suitcases
5. a street artist's canvas of a portrait
6. a stack of old newspapers with a mug on top of it
7. a set of red vintage postcards
8. a vintage black and white photograph of a landscape
9. a vintage blue leather-bound journal

Prompt: A half eaten dessert cake sitting on a cake plate.
1. a brass table lamp
2. a silver tea set
3. a volley of seashells
4. a cratered moon surface
5. a stack of old red suitcases
6. a stack of old newspapers with a mug on top of it
7. a red metal lantern post
8. a vintage red fire truck toy
9. a vintage black and white photograph of a landscape

Prompt: A fancy bathroom with his and her mirrors and sinks next to a toilette.
1. a stone garden statue of buddha
2. a calm lake reflection
3. a cratered moon surface
4. a martian landscape
5. a mythical dragon statue
6. a stack of old red suitcases
7. a mountain hiking trail with puddles
8. a vintage black and white photograph of a landscape
9. a beautifully crafted stone inca statue

Prompt: A desk with multiple computer screens forming one large screen.
1. a brass table lamp
2. a calm lake reflection
3. a stack of old red suitcases
4. a forest waterfall at distance
5. a stack of old newspapers with a mug on top of it
6. a vintage copper microscope
7. a vintage black and white television with an antenna
8. a vintage black and white photograph of a landscape
9. a vintage astronomical globe

Figure 9: IConMark watermarked images ($k = 9$) with their corresponding prompts for image generation using the Flux model and detected concepts from the concept database $\mathcal{D}$ using the IDEFICS3 visual language model.

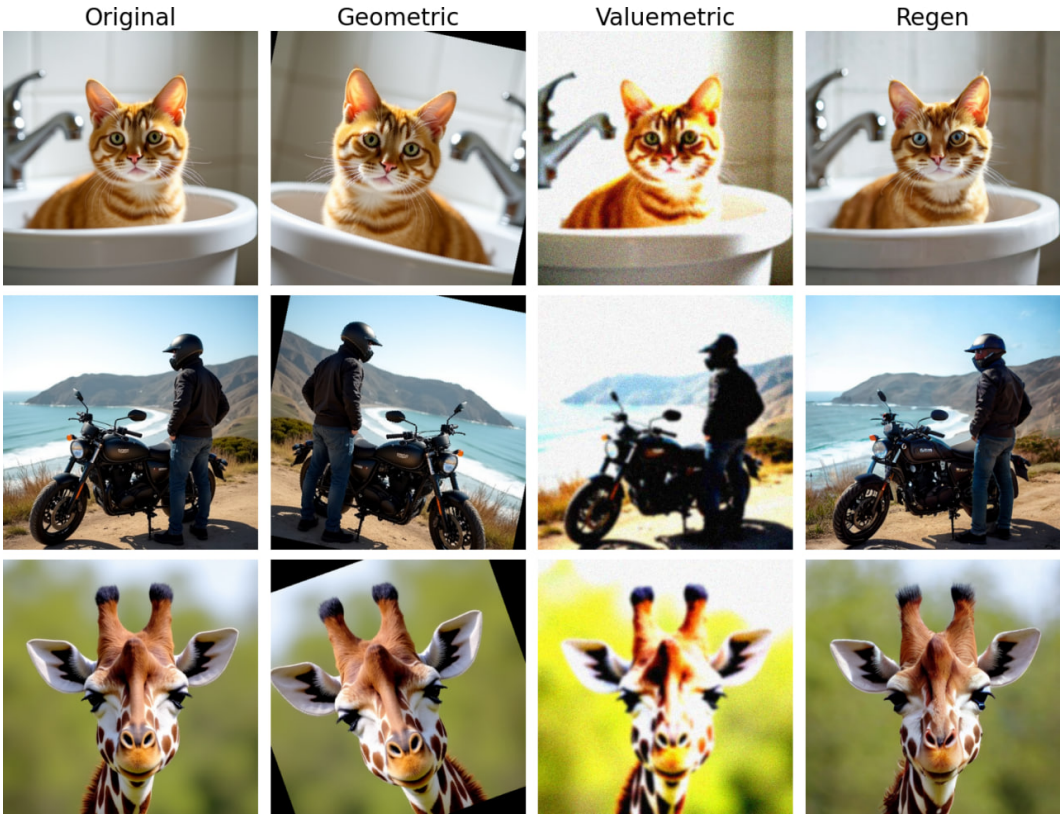


Figure 10: Examples of modified images for each modification type used in Section 5.3.