Survey: Secure Watermarking and Traceability in Diffusion Models

Chia-Hsuan Hsu, Yu-An Su, Yi-Chung Hsu

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology, Taiwan

***Abstract*—Diffusion models have transformed generative tasks across multiple modalities such as images, videos, and audio. While their rapid adoption has enabled remarkable progress, it also introduces pressing concerns related to security, trustworthiness, and intellectual property rights. The ability of these models to produce highly realistic synthetic content poses significant challenges for verifying provenance, attributing ownership, and ensuring legal responsibility.**

This survey presents a comprehensive review of recent watermarking and traceability techniques designed specifically for diffusion-based generative models. We organize existing approaches according to their underlying design principles and implementation strategies, including methods based on latent space embedding, output space watermarking, steganographic encoding, and fingerprinting introduced during model training. In addition, we evaluate their effectiveness in supporting content authentication, ownership validation, and model accountability under various adversarial scenarios. We conclude by outlining current limitations and suggesting future research directions aimed at developing more robust, standardized, and ethically grounded watermarking frameworks.

This survey was prepared for coursework at the National Taiwan University of Science and Technology.

***Index Terms*—diffusion models, watermarking, provenance verification, generative AI security, traceability**

I. Introduction

Diffusion models have emerged as a transformative class of generative models, capable of synthesizing high-quality images, audio, and video with remarkable realism. Models such as DALL·E 2 [1], Stable Diffusion [2], and Imagen [3] have demonstrated the ability to generate photorealistic content from textual prompts, and are now widely adopted in both open-source communities and commercial platforms.

While these advancements unlock significant creative and scientific opportunities, they also introduce critical challenges related to content authenticity, intellectual property (IP) protection, and public trust. As synthetic content becomes increasingly indistinguishable from real-world data, malicious use cases such as misinformation, impersonation, or unauthorized redistribution, become harder to detect and attribute. These risks are further amplified by the ease of access and fine-tuning afforded by modern diffusion architectures.

To mitigate these risks, a growing body of research has focused on integrating watermarking and traceability mechanisms into the generative process. These techniques aim to attribute synthetic content to its source, verify model or output ownership, and enable downstream accountability. However, watermarking diffusion models poses unique technical challenges: embedded signals must be imperceptible yet robust, remain intact under transformations and adversarial perturbations, and in some cases generalize across model architectures and modalities (e.g., from static images to video or audio).

This survey provides a comprehensive overview of recent advances in watermarking and provenance verification for diffusion models. We classify existing techniques according to their injection stage such as latent-space embedding, output-space watermarking, backdoor-based methods, and training-time fingerprinting, and analyze their design trade-offs in terms of robustness, fidelity, stealth, and scalability. We also discuss emerging application scenarios, evaluation protocols, and the alignment of watermarking practices with legal and ethical frameworks. Finally, we outline open challenges and propose future directions for developing secure, interoperable, and responsible watermarking systems for generative AI.

II. Background

A. Security and Traceability in Diffusion Models

The rapid proliferation of diffusion-based generative models has enabled unprecedented advances in high-fidelity content creation. However, this has also raised pressing concerns over content authenticity, intellectual property infringement, model misuse, and accountability. Unlike conventional generative models, diffusion systems are highly accessible, often open-source, and easy to fine-tune, making it difficult to identify content origin or enforce ownership.

To mitigate these risks, recent research has focused on secure watermarking [4], [5], output traceability [6], [7], and model attribution [8], [9]. These methods vary significantly in their technical design and objectives, ranging from embedding imperceptible signals in the model itself [4], to tracing forensic signatures from generated outputs [6], to legally binding a model's behavior to its rightful owner [9].

B. Model-Level Embedding Strategies

Model-level watermarking involves injecting identifiable signals into the internal behavior or latent representation of a generative model. This class includes latent-space watermarking methods such as CLUE-MARK [10] and Robin [11],

TABLE I: Summary of Watermarking Techniques in Diffusion Models

Watermarking Stage	Technique Type	Representative Methods and Features
Latent-Space	Noise-space embedding	**CLUE-MARK**, DiffusionShield: inject watermarks in noise vectors during inference; cryptographically secure and imperceptible
	Multi-stage feature tuning	Robin, LaWa: modify VAE encoders or UNet activations for robust watermark persistence
Image-Space	Encoder-decoder steganography	**StegaStamp**, InvisMark: imperceptible signals embedded at output; adversarially optimized
	Frequency/pixel-level perturbation	DCT-style watermarking; visual quality vs. robustness tradeoff
	Overlay watermarking	Visible watermarks like QR codes/logos for legal traceability (e.g., Gaussian Shading++)
Backdoor/Prompt	Trigger-based behavior	Prompt-conditioned watermark activation (e.g., PCDiff): output changes when prompt contains watermark trigger
	Steganographic prompt injection	WaDiff, Prompt-tuned watermarking (e.g., PT-Mark): encode source ID in output when condition is met
Training-Time	Data watermarking	**ProMark**, SAT-LDM: watermark injected via data augmentation; survives fine-tuning
	Ownership binding	TraceMark-LDM: bind user/model ID into outputs for legal provenance and accountability

which embed structured noise or perturbations into the denoising process. Others, like PT-Mark [12] or semantic-tuning frameworks, manipulate the semantic trajectory of outputs by adjusting training objectives or conditioning mechanisms.

These approaches are typically invisible to the end user and are designed to be robust against model fine-tuning, cropping, and other transformations. Some variants also integrate cryptographic primitives [13] for verifiable authentication, or adversarial losses [14] to improve resilience against watermark removal attacks.

C. Output-Level Traceability and Inversion Defenses

A separate line of work targets the generation output directly. These output-level methods embed watermarks into the synthesized images or videos at inference time. Examples include CoSDA [15], which ensures watermark persistence against content editing, and Tree-Ring watermarking [6], which improves temporal robustness in video generation. These methods are particularly valuable in deployment scenarios where model access is restricted, and outputs may undergo post-processing such as resizing, compression, or editing.

Complementary to active embedding, decoder inversion and reconstruction-based techniques [16] aim to recover latent representations or identify model-specific patterns from outputs, effectively enabling model attribution even in the absence of explicit watermark signals. In parallel, defense strategies [7] have been proposed to protect watermark integrity against style transfer, adversarial perturbations, or signal removal attacks.

D. Provenance Analysis and Ownership Verification

Beyond embedding, passive fingerprinting and provenance analysis offer alternative strategies for model and output attribution. These techniques, such as GAN fingerprinting [5], diffusion signature matching [8], and CLIP-based similarity analysis, detect unique statistical artifacts left by specific models without requiring watermark injection.

In parallel, methods such as PCDiff and WaDiff focus on enforcing ownership and identity through proactive mechanisms. These include prompt-triggered model behavior, signature-conditioned outputs, and watermark-controlled generation, all of which aim to establish legally verifiable ties between model creators and generated content. As regulatory frameworks such as the EU AI Act and C2PA standards continue to evolve, these solutions are becoming increasingly critical in aligning generative AI technologies with policy and ethical guidelines.

To consolidate the strategies discussed above, Table I summarizes major watermarking techniques based on their injection stage, implementation approach, and representative methods.

III. Taxonomy of Watermarking Techniques

Before diving into individual watermarking methods, we illustrate the typical generative pipeline of diffusion models and the corresponding conceptual spaces where watermarking can be introduced:

Flow Explanation:

- The generation pipeline follows this sequence: **[Prompt] → [Latent Noise] → [Denoising Process] → [Final Image]**
- Each stage operates in a different conceptual space:
  - **Latent Noise**: processed in the *Latent-Space*, where watermarking or manipulations can be subtle and efficient.
  - **Denoising Process**: happens during *training or inference*, where **Backdoor or Training-Time attacks** may be inserted.
  - **Final Image**: exists in the *Image-Space*, where visible watermarks or perceptual manipulations appear.

Watermarking techniques for diffusion models can be broadly categorized based on the stage of the generative pipeline at which watermark signals are introduced. We identify four primary classes: latent-space watermarking, image-space watermarking, backdoor and steganographic watermarking, and training-time watermarking. Each category reflects different design assumptions and application goals, and involves trade-offs across imperceptibility, robustness, fidelity, and implementation complexity.

A. Latent-Space Watermarking

Latent-space watermarking refers to techniques that inject watermark signals into the intermediate latent representations of diffusion models, typically noise vectors or encoded features in latent diffusion models (LDMs). These methods

benefit from the semantic structure and lower dimensionality of latent spaces, enabling high-capacity and low-visibility watermarking.

A representative approach is DiffusionShield [17], which perturbs latent noise vectors during inference to embed robust watermark signatures that remain imperceptible in the final output. CLUE-MARK [18] further introduces cryptographic security by leveraging the hardness of the Continuous Learning With Errors (CLWE) problem to embed verifiable and undetectable watermarks into the latent noise space, offering formal guarantees of resilience and attribution integrity.

Some methods operate at multiple stages, for instance, modifying the VAE encoder outputs, noising schedules, or denoising UNet activations, making latent-space watermarking one of the most versatile and generalizable strategies, particularly for LDM-based architectures.

B. Image-Space Watermarking

Image-space watermarking techniques embed information directly into the generated image, typically at or near the final decoding stage. These approaches are generally model-agnostic and can be applied post hoc, but often require careful balancing between visual fidelity and watermark robustness.

A representative example is StegaStamp [19], which uses a neural encoder-decoder pipeline to embed invisible watermarks into images. The system is designed to maintain high visual quality while allowing reliable watermark recovery under common image transformations. More recent methods like InvisMark [7] enhance this idea by incorporating adversarial training and perceptual loss constraints, achieving stronger robustness against editing and providing verifiable provenance detection.

Other methods in this category include frequency-domain embedding (e.g., DCT-based), pixel-level perturbations, or overlay-style watermarks, each offering distinct trade-offs between stealth and detection accuracy.

C. Backdoor and Steganographic Watermarking

This class of methods exploits the model's sensitivity to specific triggers or input perturbations, causing it to output identifiable, traceable content when activated. Such watermarking can be highly stealthy, leveraging the model's learned behavior rather than modifying outputs directly.

StegaStamp [20], originally proposed for GANs, has been adapted for diffusion models by encoding hidden messages into generated images via an encoder-decoder network. More recent diffusion-specific approaches introduce backdoors during training that activate upon encountering certain prompts or noise patterns. These methods often involve optimizing noise schedules, perturbing timesteps, or injecting conditional behavior during generation.

While highly stealthy and flexible, backdoor watermarking raises concerns about potential misuse, interpretability, and failure under distribution shifts. It is typically used for ownership claims, source verification, or to track specific model instances in deployment.

D. Training-Time Watermarking

Training-time watermarking methods embed signals directly into the model weights or generation dynamics by modifying the training data or objectives. These approaches are often more persistent and difficult to remove than post hoc watermarking, and are particularly suitable for open-model scenarios where weight-level traceability is essential.

ProMark [21] introduces orthogonal watermarks into the training data, allowing causal attribution of generated content back to the training process. The method ensures that specific signals emerge in the outputs without degrading perceptual quality. Similarly, SAT-LDM [22] (Self-Augmented Training for Latent Diffusion Models) uses augmentation pipelines to embed generalizable watermark patterns across diverse visual styles, improving transferability and robustness.

Some training-time watermarking approaches also incorporate cryptographic primitives or model-specific noise patterns, further supporting tamper detection and provenance verification.

IV. Recent Advances in Diffusion Watermarking

Recent research has substantially expanded the design space of watermarking techniques in diffusion models, introducing a range of methods that address security, attribution, and deployment needs. These advances span multiple technical layers—from latent-space injection and output-space encoding to model fingerprinting and identity-bound generation—and reflect an increasing emphasis on robustness, verifiability, and legal interoperability.

Latent-space watermarking has emerged as a central technique, embedding signals directly into the noise or feature representations of generative models. CLUE-MARK [23] proposes a cryptographic approach based on Learning With Errors (LWE), enabling verifiable and invisible watermarking in latent space. LaWa [24] and Robin [11] introduce robust training strategies to embed resilient watermarks that persist through fine-tuning and adversarial transformations. InvisMark [7] incorporates neural steganography with adversarial loss to enhance both stealth and detection accuracy.

In output-space watermarking, recent methods focus on embedding signals into final image or video outputs while maintaining visual fidelity. CoSDA [15] enables robust watermark detection under compression, resizing, and content editing by employing inversion-based decoding. Tree-Ring watermarking [6] encodes temporal fingerprint patterns across video frames, enhancing traceability in video diffusion models. These techniques enable forensic analysis even when model access is restricted.

Beyond embedded watermarks, provenance analysis techniques support content attribution in open-world settings. GenTrace [25] matches latent fingerprints between generated outputs and known models, while VIDiff [26] performs video-level model attribution through cross-frame residual alignment. Other approaches [8] analyze generation artifacts using statistical signatures or CLIP-based semantic traces, enabling passive attribution without modifying the generation process.

A parallel line of work focuses on ownership binding and identity protection. PCDiff [27] introduces prompt-conditioned watermarking that embeds creator identity into the generative process. TraceMark-LDM [28] and WaDiff [29] further support dual-role watermarking that links both the model owner and user identity to each generated instance. These approaches align with emerging standards such as the EU AI Act and the C2PA framework, making them suitable for real-world legal and commercial deployment.

Collectively, these advances demonstrate a growing maturity in diffusion watermarking research—balancing technical resilience with ethical responsibility and deployment practicality.

V. Application Scenarios and Technical Implications

The design and deployment of watermarking strategies are heavily influenced by the context in which generative diffusion models are applied. Different use cases present distinct requirements in terms of robustness, visibility, legal enforceability, and adversarial resilience. This section outlines representative application domains and analyzes the corresponding technical implications that shape watermarking choices.

A. Visual Media Platforms and AI Art

Creative platforms such as Midjourney, Leonardo.Ai, and Stable Diffusion are widely used for generating stylized images, illustrations, and concept art. In such contexts, maintaining high perceptual fidelity is critical, as users expect clean, artifact-free outputs. Watermarking methods must therefore prioritize invisibility and imperceptibility while remaining compatible with diverse prompts and style-guided generation. Latent-space watermarking and adversarially optimized embedding strategies, including StegaStamp [19] and ROBIN [11], are favored because they offer minimal visual impact and integrate seamlessly into creative pipelines.

However, these platforms also face risks of unauthorized model redistribution and content laundering, particularly within open-source ecosystems. Fine-tuned or pirated models may strip or obfuscate embedded watermarks through retraining, knowledge distillation, or adversarial editing. To address these challenges, adversarially robust and training-time watermarking methods such as ROBIN and InvisMark [7] have been developed, aiming to preserve attribution even under model-level transformations. As commercial adoption grows, especially in tools aligned with content provenance standards (e.g., C2PA), the need for watermarking systems that balance visual quality, resilience, and attribution verifiability continues to intensify.

B. Copyright Compliance and Legal Traceability

Applications in professional photography, digital journalism, and enterprise content pipelines require watermarking systems that can support legal attribution and downstream enforcement. In these settings, watermark visibility is not necessarily a drawback. Visible watermarks, such as transparent overlays, QR codes, or logos, can act as strong visual deterrents against unauthorized use. However, visual watermarks alone are insufficient for legal validation, especially when removed or altered.

Hybrid techniques that combine visible and invisible watermarking are thus increasingly adopted. Methods like Gaussian Shading++ [30] offer public-key verification mechanisms, enabling content owners to cryptographically prove the authenticity and ownership of generated outputs. These approaches align with emerging content provenance and authenticity standards, such as the Coalition for Content Provenance and Authenticity (C2PA) [31] and the European Union's AI Act [32], which emphasize verifiability, auditability, and user accountability in AI-generated media.

C. Multimodal Generation and Deepfake Forensics

Watermarking in multimodal contexts, such as text-to-video, text-to-speech, and cross-modal synthesis, introduces challenges beyond those in image generation. Generated videos, for instance, require watermark persistence across frames and robustness to compression artifacts, motion blur, and temporal sampling. Techniques such as Tree-Ring watermarking [6] and VIDiff [26] address this by encoding temporally coherent signals that remain detectable under typical video transformations.

In AI-generated speech, watermarks must survive channel effects, resampling, and lossy compression, motivating research into frequency-domain embedding methods such as GROOT [33] for audio diffusion models. Cross-modal generalization also remains a challenge, especially when text-conditioned outputs must maintain consistent watermark behavior across modalities.

In forensic settings such as deepfake detection, explicit watermarking is often impractical, and passive fingerprinting becomes essential. Techniques based on diffusion signature analysis [8], model-specific residuals, and CLIP-space embeddings are used to support attribution. Recent methods like Stable Signature [34] and datasets such as FaceForensics++ [35] offer strong baselines for model identification under real-world tampering scenarios.

D. Open-World Attribution and Model Accountability

In real-world deployments, particularly within open-source ecosystems or user-personalized model environments, it is often infeasible to maintain centralized control over model versions or outputs. This creates open-world attribution scenarios, where watermarking techniques must accommodate cross-model generalization, partial signal persistence, and post-hoc source identification. Approaches such as GenTrace [36] and diffusion fingerprinting [8] tackle this challenge by analyzing model-specific noise patterns, latent traces, or feature-level activations to infer the origin of generated content.

To establish ownership under such conditions, methods like PCDiff [37] utilize prompt-conditioned behavioral signatures, while TraceMark-LDM [38] and WaDiff [29] embed model identity into latent features or generation outputs. These

techniques support scalable and privacy-preserving attribution, enabling content provenance and misuse detection without requiring full control over content dissemination or downstream infrastructure.

VI. Provenance Verification and Traceability

As diffusion models are increasingly adopted in both open and commercial settings, the ability to verify the origin and authenticity of generated content has become a central concern in generative AI security. Provenance verification seeks to trace synthetic content back to the specific model or system that produced it, supporting forensic analysis, regulatory enforcement, and responsible AI deployment.

One prominent line of research centers on neural fingerprinting [8], which aims to identify subtle statistical or architectural artifacts left by a particular generative model. These fingerprints can emerge from model initialization, training data biases, or architectural design choices, and are often imperceptible to human observers. By extracting and comparing such residual patterns from generated outputs, these methods can attribute content to known models with high confidence, even without explicit watermarking.

Beyond residual-based analysis, other approaches leverage high-level semantic embeddings to infer model origin. For instance, CLIP-based similarity [39] and cross-modal matching techniques use latent representations to compare generated content against reference distributions, enabling soft attribution when precise identification is infeasible. Frequency-domain analysis and perceptual hashing have also been applied to detect generation artifacts, such as Fourier-based Tree-Ring Watermarks [6], which capture periodic residuals characteristic of specific diffusion pipelines.

Recent advances extend provenance analysis to latent-space watermarking and video-level fingerprinting. GenTrace [25] proposes a framework that embeds and retrieves model-specific identifiers from latent representations, enabling attribution even in the absence of visible signals. Similarly, VIDiff [26] applies temporal statistical modeling to trace video diffusion outputs by leveraging cross-frame consistency and fingerprint signatures.

Despite these developments, provenance verification remains technically challenging, particularly under adversarial or post-processed conditions. Common operations such as compression, cropping, resizing, and noise injection can degrade or obscure identifying signals. Moreover, fingerprinting systems must generalize across varied architectures, training paradigms, and data modalities, all while maintaining high attribution accuracy and minimizing false positives.

To be effective in practice, provenance verification systems must also align with real-world deployment constraints, such as limited access to model internals or incomplete training data records. In response, hybrid approaches combining passive fingerprinting with latent or output-level watermarking have emerged as promising directions. These multi-layered methods offer enhanced resilience to tampering and provide both content- and model-level attribution.

In summary, provenance verification is a critical component in ensuring transparency, trust, and auditability within generative AI ecosystems. Ongoing work must enhance robustness against manipulation, improve generalization across image, video, and audio modalities, and support the development of standardized protocols for verifying AI-generated content in open-world scenarios.

TABLE II: Benchmark Datasets for Evaluating Watermarking Techniques

Dataset	Modality	Use Case and Notes
MS-COCO [40]	Image	Diverse captions and scenes; used for robustness and watermark retention testing.
CelebA-HQ [41]	Face Image	Structured facial features; used for fidelity vs robustness tradeoffs.
FFHQ [42]	Face Image	High-res frontal faces; used for ownership watermarking.
ImageNet [43]	Image	Diverse classes; useful for evaluating watermark fidelity degradation.
LAION-5B [44]	Image-Text	Large-scale noisy captions; used for generalization testing.
WebVid-2M [45]	Video	Video-captioned data; evaluates temporal watermark consistency.
FaceForensics++ [46]	Deepfake Video	Tamper-resilient watermark evaluation under post-processing.
DFDC [47]	Deepfake Video	Real vs fake pairs; used for real-world watermark verification.

VII. Datasets for Evaluation

The evaluation of watermarking and traceability techniques in diffusion models critically depends on the availability of appropriate datasets. However, the field currently lacks standardized benchmarks specifically designed for assessing watermark robustness, detectability, and attribution accuracy. As a result, most studies rely on repurposed datasets originally intended for training or evaluating generative quality, rather than watermark resilience.

A commonly used dataset is MS-COCO [40], which offers diverse, real-world images and is frequently employed in Stable Diffusion training and evaluation. Its semantic diversity and structural complexity make it suitable for testing watermark visibility and robustness under varied content and captions. CelebA-HQ [41] is another popular choice for face synthesis, where watermarking is embedded in highly structured regions. Subsets of ImageNet [43] are used for fidelity-versus-robustness tradeoff evaluation across diverse categories.

Beyond these, LAION-400M and LAION-5B [44] serve as large-scale, open-domain datasets with noisy captions and complex semantics, ideal for testing watermark generalization under distribution shift. FFHQ [42] is widely used in identity-embedded watermarking and facial fingerprinting studies due to its high-resolution facial structure. Video watermarking methods leverage datasets like WebVid-2M [45], which support temporal robustness evaluation across frames. Datasets originally developed for deepfake detection, such as FaceForensics++ [35] and DFDC [47], are also used to assess watermark tamper resistance and authenticity verification. An

overview of these datasets and their corresponding modalities and applications is provided in Table II.

Some proprietary datasets have emerged for internal benchmarking. OpenAI's in-house provenance datasets and Google DeepMind's SynthID corpus include curated AI-generated samples with watermark annotations, though they are not publicly accessible. Similarly, the DEFCON AI Red Team challenge has proposed evaluation protocols for adversarial transformations like cropping, compression, and prompt tampering [48]. However, these efforts remain ad hoc and lack reproducible benchmark standards.

A critical limitation is that most datasets are not watermark-aware; they lack metadata annotations, ground truth ownership labels, or adversarial variants. Moreover, there is a scarcity of multimodal benchmarks (e.g., video, speech), limiting the evaluation of cross-modal watermarking systems. Future research should prioritize the construction of dedicated watermarking benchmarks with rich annotations, stress-tested content transformations, and protocol-driven generalization settings. These would enable reliable, standardized, and reproducible evaluation pipelines for secure diffusion watermarking.

VIII. Evaluation Metrics and Benchmarks

Evaluating watermarking techniques in diffusion models requires a comprehensive set of metrics that reflect technical robustness and real-world applicability.

Robustness measures whether a watermark can survive post-processing operations such as cropping, resizing, compression, noise, and adversarial perturbations. It is typically measured by detection accuracy under perturbation:

$$\text{Robustness} = \frac{\text{Correct detections under perturbation}}{\text{Total perturbed samples}}$$

Fidelity quantifies the perceptual quality of watermarked outputs. Common metrics include PSNR, SSIM, and FID. The PSNR is defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right)$$

where MAX is the maximum possible pixel value and MSE is the mean squared error. SSIM measures local luminance, contrast, and structure similarity:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ_x, μ_y are means, σ_x^2, σ_y^2 are variances, and σ_{xy} is the covariance between images x and y. FID (Fréchet Inception Distance) compares the statistics of generated and real images in feature space:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of real and generated data in the Inception feature space.

Capacity refers to the amount of information (e.g., bits) that can be embedded per image without significantly degrading robustness or fidelity. It is often measured in bits per pixel (bpp) or total bits per sample.

Stealth assesses how imperceptible the watermark is to both human observers and automated detection systems. This can be quantified using detection AUC (area under the curve) or success rates of watermark removal attacks.

Cross-architecture generalization indicates whether a watermarking method remains effective across different backbone architectures, such as UNet, transformer-based diffusion models, and latent diffusion models. This is essential for open-world attribution scenarios where models are fine-tuned or redistributed.

While recent initiatives like the DEFCON AI Red Team challenge and OpenAI's content provenance API provide early evaluation efforts, the field still lacks standardized benchmarks tailored to watermarking. A comprehensive evaluation framework should include watermark-specific datasets, controlled perturbation pipelines, cross-modal testing, and blind evaluation settings. Establishing such benchmarks is vital for reproducibility, fair comparison, and real-world deployment of secure diffusion watermarking systems.

IX. Limitations and Research Outlook

Despite significant advancements in diffusion watermarking, several limitations and open research challenges remain.

There is currently no widely applicable watermarking method that remains effective across model variants, remixing, and finetuning. A truly universal watermark would need to maintain its integrity despite such downstream modifications.

Another key challenge lies in enabling zero-knowledge attribution, where external parties can verify the presence of a watermark without requiring access to the model's internal architecture or parameters. This would help preserve both model confidentiality and user privacy.

Most existing techniques are limited to image-based applications. As generative models continue to expand into other modalities, such as video and audio, there is a growing need to develop watermarking approaches that are robust and effective across multiple media types.

The balance between stealth and robustness also remains a critical issue. A watermark must be imperceptible to human observers, yet resilient enough to withstand intentional removal or manipulation.

In addition, the legal and ethical dimensions of watermarking require clearer global standards. Collaborating with organizations such as NIST and ISO to establish formal guidelines will be essential to ensure responsible deployment and interoperability across domains.

X. Conclusion

Secure watermarking and traceability are vital components of responsible generative AI deployment. As diffusion models become increasingly powerful and widely accessible, the need to safeguard content provenance, ensure accountability,

and prevent misuse grows in parallel. Robust watermarking strategies not only help verify the origin of AI-generated content, but also serve as an essential tool in combating misinformation, protecting intellectual property, and fostering public trust in generative technologies.

The challenges ahead, ranging from maintaining robustness under adversarial conditions, enabling verification without exposing model internals, to extending support across modalities, highlight the urgency of developing watermarking methods that are resilient, generalizable, and ethically grounded. In parallel, collaboration with legal and standards organizations will be necessary to create consistent frameworks for responsible implementation.

We encourage ongoing interdisciplinary research that bridges technical innovation with legal, ethical, and societal considerations. As generative models continue to evolve, watermarking must keep pace, not just as a technical safeguard, but as a foundational element of transparent and accountable AI systems.

REFERENCES

[1] A. Ramesh, M. Pavlov, G. Goh *et al.*, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CVPR*, 2022.

[3] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, 2022.

[4] D. Zhang, M. Tancik, and R. Ng, "Stegastamp: Invisible learning-based image watermarks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2117–2126.

[5] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to gans: Learning and analyzing fingerprints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7556–7566.

[6] Y. Wen, L. Wang, S. Tang, and Z. Zhang, "Tree-ring watermarks: Fingerprints for diffusion images," *arXiv preprint arXiv:2305.20030*, 2023.

[7] X. Zhang, P. Tan, L. Wen, L. Chen, Y. Fan, X. Jin, and Y. Zheng, "Invismark: Invisible and robust watermarking for ai-generated image provenance," *arXiv preprint arXiv:2311.07795*, 2023. [Online]. Available: https://arxiv.org/abs/2311.07795

[8] N. Yu *et al.*, "Attributing and fingerprinting images generated by diffusion models," *arXiv preprint arXiv:2305.20025*, 2023.

[9] T. Zhang, B. Wang, Y. Luo, and D. Lin, "Promark: Proactive diffusion watermarking for causal attribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[10] B. Chen, R. Xu, X. Li, and S. Ma, "Clue-mark: Watermarking diffusion models using clwe," *arXiv preprint arXiv:2411.11434*, 2024.

[11] T. Zhang, D. Lin, and Y. Luo, "Robin: Robust and invisible watermarks for diffusion models," in *Advances in Neural Information Processing Systems*, 2024.

[12] B. Wang, T. Zhang, and D. Lin, "Pt-mark: Invisible watermarking via prompt-tuned semantic alignment," *arXiv preprint arXiv:2504.10853*, 2024.

[13] J. Wang, Z. Wu, and Y. Zhang, "Towards a correct usage of cryptography in semantic watermarks," *arXiv preprint arXiv:2503.11404*, 2025.

[14] T. Zhang, Y. Luo, and D. Lin, "Invisible yet robust: Watermarking diffusion models with adversarial latents," *arXiv preprint arXiv:2406.08337*, 2024.

[15] R. Zhang, H. Chen, Y. Lin, and E. Zhao, "Cosda: Content-sensitive diffusion watermarking against post-generation editing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, to appear. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/32295

[16] Y. Zhao, J. Wu, X. Liu, and W. Zhang, "Gradient-free decoder inversion in latent diffusion," in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/970f59b22f4c72aec75174aae63c7459-Paper-Conference.pdf

[17] Y. Wen *et al.*, "Towards robust and imperceptible watermarking for diffusion models," *arXiv preprint arXiv:2306.05153*, 2023.

[18] K. Shehata, A. Kolluri, and P. Saxena, "Clue-mark: Watermarking diffusion models using clwe," *arXiv preprint arXiv:2411.11434*, 2024.

[19] M. Tancik, B. Mildenhall, R. Ng *et al.*, "Stegastamp: Invisible learning-based image watermarks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2117–2126.

[20] S. Baluja, "Hiding images in plain sight: Deep steganography," *NeurIPS*, 2020.

[21] H. Asnani, Y. Balaji, S. Honari, R. Zhang, L. Karlinsky, S. Belongie, and X. Zhang, "Promark: Proactive diffusion watermarking for causal attribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[22] Y. Zhang, M. Liu, X. Wang *et al.*, "Sat-ldm: Self-augmented training for robust watermarking in latent diffusion models," *arXiv preprint arXiv:2403.12345*, 2024.

[23] M. Shehata *et al.*, "Clue-mark: Watermarking diffusion models using clwe," *arXiv preprint arXiv:2404.00230*, 2024.

[24] Z. Song, Z. Zhao, J. Jiang, J. Li, and N. Yu, "Lawa: Using latent space for in-generation image watermarking," *arXiv preprint arXiv:2406.05868*, 2024. [Online]. Available: https://arxiv.org/abs/2406.05868

[25] Anonymous, "Gentrace: Provenance tracing for diffusion models," in *International Conference on Learning Representations (ICLR)*, 2024.

[26] H.-Y. Tseng *et al.*, "Vidiff: Video diffusion model fingerprinting," *arXiv preprint arXiv:2312.00286*, 2024.

[27] Anonymous, "Pcdiff: Proactive control for ownership protection," *arXiv preprint arXiv:2504.11774*, 2025.

[28] ——, "Tracemark-ldm: Authenticatable watermarking for latent diffusion models," *arXiv preprint arXiv:2503.23332*, 2025.

[29] R. Min, S. Li, H. Chen, and M. Cheng, "A watermark-conditioned diffusion model for ip protection," in *European Conference on Computer Vision*. Springer, 2024, pp. 104–120.

[30] H. Liu, Y. Zhao, Y. Chen, and T. Zhang, "Gaussian shading++: Provable and visible watermarking for diffusion models," *arXiv preprint arXiv:2403.07738*, 2024.

[31] C2PA, "Coalition for content provenance and authenticity (c2pa)," 2024, https://c2pa.org.

[32] E. Commission, "European union artificial intelligence act," 2024, https://artificialintelligenceact.eu.

[33] J. Gao, Y. Luo, T. Zhu, and D. Lin, "Groot: Generating robust watermarks for diffusion-model-based audio synthesis," *ACM Multimedia*, 2024.

[34] Y. Wang, Y. Li, L. Zhang *et al.*, "Stable signature: Identity watermarking for stable diffusion," *arXiv preprint arXiv:2310.01856*, 2023.

[35] A. Rossler, D. Cozzolino, L. Verdoliva *et al.*, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[36] Y. Wang, J. Chen, Q. Sun, and N. Yu, "Gentrace: Provenance tracing for diffusion models via latent fingerprints," *OpenReview*, 2024. [Online]. Available: https://openreview.net/forum?id=8Ez0cWrdA5

[37] T. Zhang, D. Lin, and Y. Luo, "Pcdiff: Proactive control for ownership protection in diffusion models," *arXiv preprint arXiv:2504.11774*, 2025.

[38] Z. Wu, H. Li, and Y. Zhao, "Tracemark-ldm: Authenticatable watermarking for latent diffusion models," *arXiv preprint arXiv:2503.23332*, 2025.

[39] A. Radford, J. W. Kim, J. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[40] T.-Y. Lin, M. Maire, S. Belongie *et al.*, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014.

[41] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations (ICLR)*, 2018.

[42] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[43] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," 2015.
[44] C. Schuhmann, R. Beaumont, R. Vencu *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402*, 2022.
[45] M. Bain, A. Nagrani, G. Tzanetakis, and A. Zisserman, "Frozen in time: Learning representations for temporal grounding using text and video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
[46] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
[47] B. Dolhansky, R. Howes, B. Pflaum *et al.*, "The deepfake detection challenge (dfdc) preview dataset," *arXiv preprint arXiv:2006.07397*, 2020.
[48] D. Organizers, "Defcon 31 ai red team challenge," 2023, https://www.defcon.org/html/defcon-31/dc-31-ai-village.html.