COMPUTER DESIGN SYSTEMS. THEORY AND PRACTICE

Vol. 7, No. 1, 2025

Davyd Telenko[1], Oksana Oborska[2], Dariya Rebot[3]

[1]Computer Design Systems Department, Lviv Polytechnic National University, S. Bandery street 12, Lviv, Ukraine, E-mail: davyd.telenko.mknsp.2023@lpnu.ua, ORCID 0009-0004-8625-7516
[2]Computer Design Department, Lviv Polytechnic National University, S. Bandery street 12, Lviv, Ukraine, E-mail: oksana.v.oborska@lpnu.ua, ORCID 0009-0001-0825-1379
[3]Computer Design Department, Lviv Polytechnic National University, S. Bandery street 12, Lviv, Ukraine, E-mail: dariya.p.rebot@lpnu.ua, ORCID 0000-0002-3583-0800

A UNIFIED SYSTEM FOR AI-GENERATED IMAGE AUTHENTICATION AND MANAGEMENT

Received: March 03, 2025 / Revised: March 12, 2025/ Accepted: March 20, 2025

https://doi.org/

Abstract. This article presents the development of an image generation system that employs digital watermarking and metadata embedding technologies to determine whether an image has been generated by an AI model. The system acts as an intermediary service between providers (web services with generation models) and end users, ensuring seamless integration of these technologies. With the growing volume of AI-generated content, distinguishing such images from authentic ones has become increasingly challenging. Additionally, the lack of universal tools for managing generated assets and embedding metadata creates inefficiencies and risks related to authenticity and intellectual property. This article attempts to create a viable centralized solution that integrates protection measures into any user-generated image, regardless of the originating service. The system operates as a middleware solution compatible with existing generation models, providing a unified interface for users. Developed pipeline facilitates both addition of watermarking into the generative process as well as embedding metadata. The intuitive interface enhances usability, while the centralized repository enables users to manage and verify their generated content.

This approach is innovative, combining digital watermarking, metadata integration, and centralized management within a single platform. Unlike existing tools tailored to specific platforms, this system offers cross-service functionality. The solution is highly relevant for content authenticity, intellectual property management, and user convenience. It enhances trust in digital content and provides a scalable architecture adaptable to diverse platforms and needs. Future research could extend this approach to broader areas of information technology, ranging from non-image generation models to operating system-level modules for protecting against generated products.

Keywords: image generation, artificial intelligence, generative models, digital watermarking, web-service, information safety.

Introduction

The widespread adoption of generative tools today has made it increasingly clear that modern systems display a variety of problems that emerge across multiple levels. This article aims to present a sample implementation designed to address these challenges.

Currently, there are two primary methods for end-users to generate images: the first involves utilizing publicly available generative models distributed through platforms such as Hugging Face, while the second leverages use of online services provided by large-scale corporations, such as OpenAI's ChatGPT, Microsoft's Bing Image Creator, and Google's Gemini. Each of these approaches has its own set of advantages and disadvantages.

The primary argument for using custom-trained open-source models is that they typically lack overt censorship and are executed locally, thereby offering greater freedom in terms of content and ownership. On the other hand, the use of free models provided by large corporations offers the significant advantage of being entirely independent of the user's local machine, meaning that no computational resources are required from the end-user to generate an image.

Additionally, there exists a third approach to running generative models, that involves greater financial investment. This method requires the user to rent a generative model provider, followed by either pay-per-use or monthly fees to host the model on a cloud computing platform, where generation tasks can be executed.

In summary, each of these approaches ultimately yields the same product: the desired generated image. Typically, this image is delivered as a file, most commonly in JPEG format, or in other modern formats such as AVIF or WEBP. The critical point is that the resulting image is merely a data file, lacking any metadata or information pertaining to its generative origin [17].

Problem Statement

The primary issue which arises in all of the previously named generative system workflows for end users is inability to view the source of an image via it's metadata. This means that when you send the image over the wire the other party cannot view the actual origin of this image unless you explicitly and manually provide it in description. This can lead to several major issues.

First, the user themselves may struggle to correlate the prompt used to generate the image with the resulting output. This necessitates manual storage of both prompts and images, as well as a method to associate them, which can be cumbersome and error-prone.

Second, the network or platform hosting the image will store it without any indication of its generative origin. This can result in the image being ingested by other generative learning processes, which may then train on already generated data, potentially leading to a degradation in the overall quality of generated outputs. A further complication is the challenge of copyrighting such images without clear documentation of their origin. While this issue is broader in scope, some of the solutions proposed in this paper may also address this concern.

Another set of issues arises when examining the implementation of user interfaces (UI) and the general infrastructure of these systems. Most existing approaches fall short in one way or another. A significant limitation is that the majority of systems do not store or organize images alongside their corresponding prompts. This creates a substantial inconvenience for users, as retrieving and preserving prompts is critical — prompts often require extensive discovery process and experimentation to achieve the desired result. From the author's personal experience, losing the prompt associated with an image effectively means losing access to the generative process that produced it, leaving the user with only the final output, which is merely a byproduct of the model.

Additionally, the UI and accessibility of these systems are often lacking. Many systems are designed to attract casual users with flashy features rather than prioritizing usability and user experience (UX). For instance, basic functionalities such as selecting the number of images to generate, partially retaining prompts, utilizing a global database for prompt storage, and enabling manual image management are frequently overlooked. These shortcomings are particularly prevalent in web-based services, where user control is limited. In contrast, local model management and execution are handled by the user, and there are several open-source frontends for image generation that offer robust features. However, even these local solutions often lack built-in mechanisms for syncing and preserving images, requiring users to manage these tasks manually.

Review of Modern Information Sources on the Subject of the Paper

The challenges associated with image generation can be addressed through three primary solutions, each with its own technical and practical implications. These solutions aim to mitigate issues such as

provenance tracking, content authenticity, and user experience, which are critical in the context of generative systems.

C2PA Metadata. The Coalition for Content Provenance and Authenticity (C2PA) is a standardized metadata framework developed through a collaboration between industry leaders such as Adobe, Arm, Intel, Microsoft, and Truepic [16]. The primary objective of C2PA is to establish a verifiable chain of custody for digital content, including images generated by AI systems. This is achieved by embedding a cryptographic history of edits, transformations, and generative processes directly into the metadata of the image file.

- Provenance Tracking: C2PA metadata records the entire lifecycle of an image, from its creation (e.g., the generative model used, the prompt, and parameters) to any subsequent modifications. This is stored in a tamper-evident format, ensuring the integrity of the data.
- Validation Mechanism: Software that supports C2PA can read and validate the metadata using cryptographic signatures. This allows users to verify the authenticity of the image and its generative origin.
- User Trust: By providing transparency about the image's history, C2PA empowers users to make informed decisions about the content they consume or share. For instance, it can help distinguish between human-created and AI-generated content, which is increasingly important in contexts such as journalism, academia, and intellectual property.

C2PA leverages advancements in cryptographic hashing and digital signatures, which are well-established in cybersecurity. The use of tamper-evident metadata ensures that any unauthorized alterations to the image or its provenance record can be detected, aligning with principles of data integrity and trustworthiness [10-12].

Digital Watermarking. Digital watermarking is a technique designed to embed imperceptible information within an image, which can be used to identify its origin or status. While this method does not inherently solve the problem of associating prompts with images, it provides a robust mechanism for marking AI-generated content and tracking its distribution.

- Invisible Watermarks: Watermarking algorithms embed a low-visibility pattern of pixel data into the image during the generation process. This pattern is typically undetectable to the human eye but can be extracted using specialized software or neural networks.
- Payload Capacity: The embedded payload is usually small (e.g., 48 bits), which is sufficient to encode information such as the author's identity, the generative model used, or a flag indicating that the image is AI-generated.
- Detection and Verification: Companies like Google and Facebook have developed watermarking solutions that integrate directly into their generative pipelines. These watermarks can be detected even after the image undergoes compression, resizing, or other common transformations.

Digital watermarking relies on signal processing and machine learning techniques to ensure robustness against image manipulations. Recent research has explored the use of neural networks for both embedding and extracting watermarks, improving resilience to adversarial attacks. However, challenges remain, such as ensuring that watermarks are not easily removed by malicious actors while maintaining the visual quality of the image [1-9].

Centralized Data Store and accessible UI. A centralized data store combined with an intuitive user interface (UI) addresses the organizational and accessibility challenges faced by users of generative systems. This solution focuses on improving the user experience by streamlining the storage, retrieval, and management of generated images and their associated metadata.

- Centralized Storage: A centralized database stores images alongside their corresponding prompts, parameters, and other metadata. This eliminates the need for users to manually manage files and associations, reducing the risk of data loss or disorganization.
- Search and Retrieval: Advanced indexing and search functionalities allow users to quickly locate specific images or prompts based on criteria such as generation date, model used, or content type.

A clean and intuitive UI also enhances usability by providing features such as:

- Prompt Retention: The ability to save and reuse prompts, either partially or in full, for iterative refinement.
- Batch Generation: Options to generate multiple images simultaneously, with customizable parameters.
- Image Management: Tools for organizing, tagging, and exporting images in a structured manner.

The design of centralized data stores and UIs draws on principles from human-computer interaction (HCI) and database management. Research in HCI emphasizes the importance of reducing cognitive load and improving task efficiency, which is particularly relevant for users who engage in extensive prompt engineering. Additionally, centralized storage systems can leverage distributed database technologies to ensure scalability and reliability, especially in cloud-based environments [13-15].

Main Material Presentation

Developed system mainly consists of six critical layers: External Services, Adapter Modules, Generalized API, Database and Server Functions, REST API, Representational Layer (Fig. 1.). The system is organized this way for several reasons, the main of which are: ease of modification, module independence, module adaptability and dynamicity.

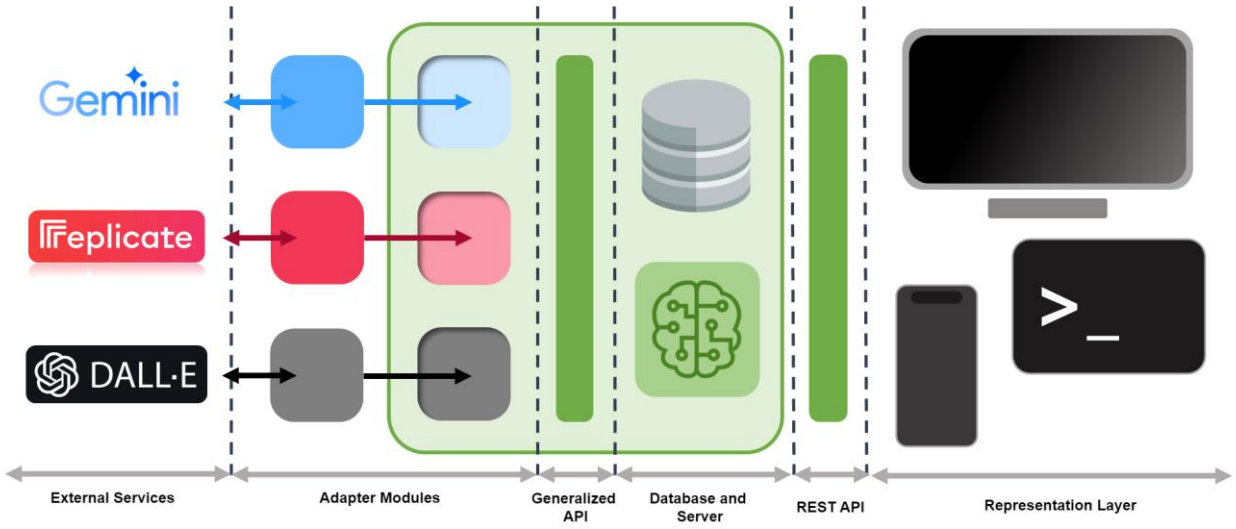

Fig. 1. Generalized structural diagram of the system.

The core idea behind this approach is an ability to organize several External services under one Generalized API through the use of Adapter Modules which act as a bridge between external API's and internal system Generalized API designed to generate, save in Database, retrieve, modify and present images to user through Representation layer. System uses centralized server, which is easily refactorable to serverless functions on demand. System also facilitates the ability to create and connect multiple clients independently of implementation framework and platform.

The system's data flow and processes are best illustrated by the example shown in Fig. 2. The user initiates a request through any of the representation layer's interfaces, specifying details such as the service, model, number of images, and most importantly, the generation prompt. This request is then processed by the server functions and routed to the appropriate handler, which validates and converts it into the required format for the designated service.

At this stage, any of the generation methods described earlier in the article can be utilized. Users may opt for a pay-per-use API service, leverage large-scale service providers, or deploy a custom model on their own server. Regardless of the chosen approach, the system ensures that generated content is processed, stored, and made available through the server's functionality and database.

Since image generation occurs externally, watermarking cannot be natively integrated into the server's functionality. The system receives only the final output from the model, which may or may not contain a watermark. To address this, the service verifies the presence of a digital watermark and, if absent,

applies a steganography-based watermark for security. This aspect of the system remains an open area for improvement in generalized architectures.

Following this, the system embeds a metadata header into the image, containing details such as the model used, the user's system identifier, and the generation prompt. The image, along with its associated metadata and ownership information, is then stored in a relational database, enabling users to view, save, and organize their images into virtual collections.

At the end of the user's workflow, images are presented with options to delete them immediately, save them to specific collections, or regenerate them on demand. This design enhances user experience by allowing seamless modifications to the workflow without unnecessary context switching or interruptions.

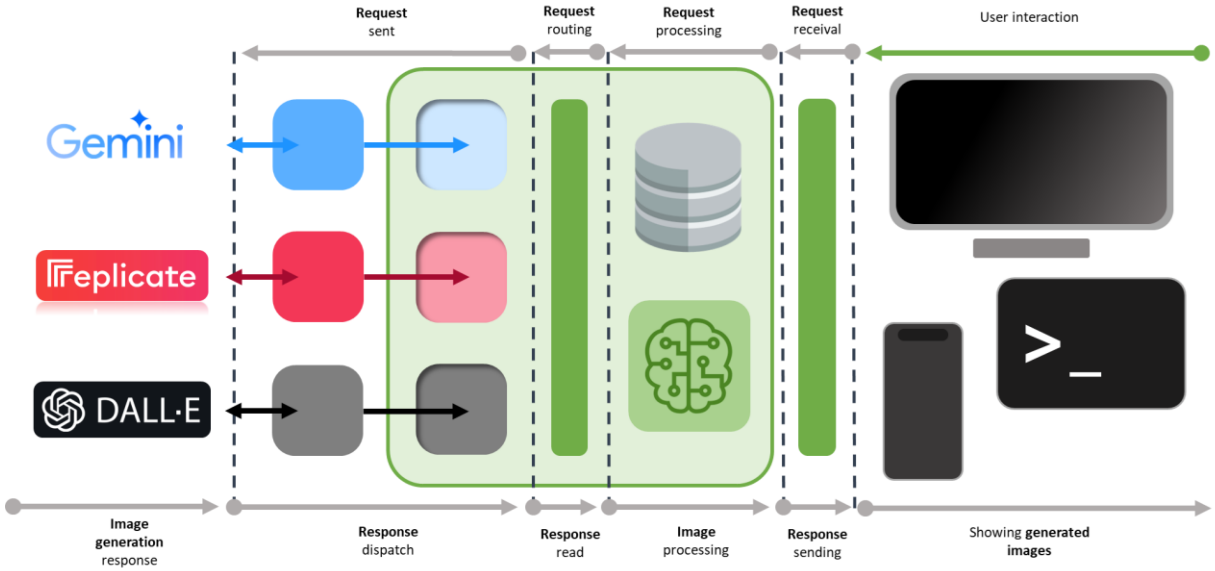

Fig. 2. Dataflow in the system.

This system is developed using TypeScript, with Node.js as the backend runtime and React as the frontend framework. The overall design emphasizes dynamic operations, leveraging TypeScript's popularity and flexibility to facilitate seamless development and rapid extensibility. Notably, the system is highly customizable and modular, imposing no strict requirements on the implementation details of its components. This allows individual modules to be easily restructured or optimized as demand increases. A crucial factor in this adaptability is the HTTPS server data channel between the system's core components, which introduces an additional layer of abstraction and decouples system dependencies.

The main frontend of the system is a reactive single page web application. Most of the features of the system is covered in its frontend representations and most of the UI/UX decision were made in accordance to the main downsides of existing systems: inefficient screen space usage, intrusiveness, inability to save local state (Fig. 3.).

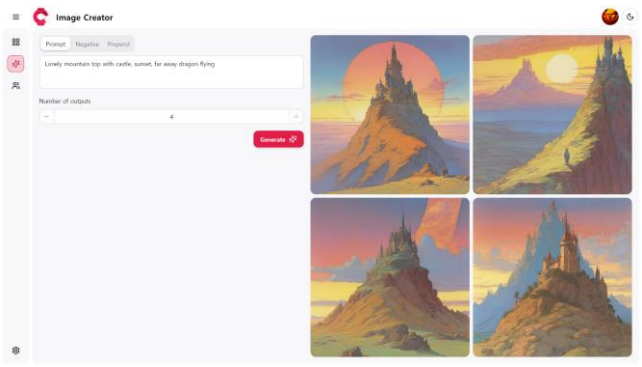

Fig. 3. "Generation" page of web application.

From a monetization perspective, the system is currently not profitable and is unlikely to become so in its current form. The main challenge lies in sustaining, maintaining, and hosting the server to support database and cloud computing features. Since users interact with generative services using their own credentials and payment models, they effectively face a double payment issue — paying both for our platform's data retention and security layer, as well as for the actual image generation service. This places our system in a precarious market position, as users may not find the added

features compelling enough to justify the additional cost and may instead opt for their own local image management solutions.

There are three potential solutions to this challenge. The first is to reposition the system as a self-hosted application for home use, allowing users to choose their own database providers and manage their own demand. The second is to directly rent and manage all external services, charging users accordingly. The third, albeit the least desirable, is to shift to an advertisement-based monetization model. As this article showcases the system in its entirety, we intend to make the code free and open-source in the foreseeable future, once the development of additional mainstream model adapters is complete. Source code of the system will be available in the author's GitHub repository[1].

Conclusions

The developed system serves as a strong foundation for a wide range of potential implementations and use cases. While it is impractical to fully integrate all major protective measures against digital fraud and AI misuse within a centralized system, it provides the flexibility to incorporate state-of-the-art security solutions into intermediary modules and enforce them for external services. Digital watermarking, which requires integration at the generative stage of external services, is particularly critical. Additionally, the centralized system allows for secondary digital watermarking on already generated content, along with metadata inclusion to enhance user awareness of an image’s authenticity.

Beyond security, the system offers extensive opportunities for both users and system owners to build a thriving community and implement intelligent data management solutions. It can be positioned as a platform for sharing, storing, and interacting with generated content—enabling features such as comments, discussions, and idea exchange. This opens the possibility of evolving into a fully-fledged social network for image generation enthusiasts.

Looking ahead, the system has the potential to expand beyond image generation, supporting 3D geometry modeling, audio, and video generation. This would create a unified platform for managing diverse types of generative content while maintaining a strong focus on security and data integrity.

References

[1] Fernandez P, Couairon G, Jégou H, Douze M, Furon T. The stable signature: rooting watermarks in latent diffusion models. arXiv.org. https://arxiv.org/abs/2303.15435. Published March 27, 2023.

[2] Xu R, Hu M, Lei D, et al. InvisMark: Invisible and robust watermarking for AI-generated image provenance. arXiv.org. https://arxiv.org/abs/2411.07795. Published November 10, 2024.

[3] Jiang Z, Guo M, Hu Y, Gong NZ. Watermark-based attribution of AI-Generated content. arXiv.org. https://arxiv.org/abs/2404.04254. Published April 5, 2024.

[4] Fernandez P. Watermarking across Modalities for Content Tracing and Generative AI. arXiv.org. https://arxiv.org/abs/2502.05215. Published February 4, 2025.

[5] Li G, Chen Y, Zhang J, et al. Warfare:Breaking the watermark protection of AI-Generated content. arXiv.org. https://arxiv.org/abs/2310.07726. Published September 27, 2023.

[6] Padhi, S. K., Tiwari, A., & Ali, S. S. (2024). Deep learning-based dual watermarking for image copyright protection and authentication. IEEE Transactions on Artificial Intelligence, 1–12. https://doi.org/10.1109/tai.2024.3485519

[7] Fairoze, J., Ortiz-Jiménez, G., Vecerik, M., Jha, S., & Gowal, S. (2025, February 7). On the Difficulty of Constructing a Robust and Publicly-Detectable Watermark. arXiv.org. https://arxiv.org/abs/2502.04901

[8] Simmons, J. C., & Winograd, J. M. (2024, May 20). Interoperable Provenance Authentication of Broadcast Media using Open Standards-based Metadata, Watermarking and Cryptography. arXiv.org. https://arxiv.org/abs/2405.12336

[9] SynthID. (2025, February 25). Google DeepMind. https://deepmind.google/technologies/synthid/

[10] OpenAI joins C2PA Steering Committee - C2PA. (n.d.). https://c2pa.org/post/openai_pr/

[11] Balan K, Agarwal S, Jenni S, Parsons A, Gilbert A, Collomosse J. EKILA: Synthetic Media Provenance and Attribution for Generative Art. arXiv.org. https://arxiv.org/abs/2304.04639? Published April 10, 2023.

[12] Longpre S, Mahari R, Obeng-Marnu N, et al. Data Authenticity, Consent, & Provenance for AI are all

[1] https://github.com/DavidTelenko

broken: what will it take to fix them? arXiv.org. https://arxiv.org/abs/2404.12691v1? Published April 19, 2024.

[13] Bieniek J, Rahouti M, Verma DC. Generative AI in Multimodal User Interfaces: Trends, challenges, and Cross-Platform Adaptability. arXiv.org. https://arxiv.org/abs/2411.10234. Published November 15, 2024.

[14] Tolomei G, Campagnano C, Silvestri F, Trappolini G. Prompt-to-OS (P2OS): Revolutionizing Operating Systems and Human-Computer Interaction with Integrated AI Generative Models. arXiv.org. https://arxiv.org/abs/2310.04875. Published October 7, 2023.

[15] Luera R, Rossi RA, Siu A, et al. Survey of user interface design and Interaction Techniques in Generative AI Applications. arXiv.org. https://arxiv.org/abs/2410.22370v1. Published October 28, 2024.

[16] C2PA specifications :: C2PA specifications. https://c2pa.org/specifications/specifications/2.1/index.html.

[17] Bond-Taylor S, Leach A, Long Y, Willcocks CG. Deep Generative Modelling: a comparative review of VAEs, GANs, normalizing flows, Energy-Based and autoregressive models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021;44(11):7327-7347. https://doi.org/10.1109/tpami.2021.3116668

Давид Теленько[1],Оксана Оборська[2], Дарія Ребот[3]

[1]Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", вул. С. Бандери, Львів, Україна, E-mail: davyd.telenko.mknsp.2023@lpnu.ua, ORCID 0009-0004-8625-7516

[2]Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", вул. С. Бандери, Львів, Україна, E-mail: oksana.v.oborska@lpnu.ua, ORCID 0009-0001-0825-1379

[3]Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", вул. С. Бандери, Львів, Україна, E-mail: dariya.p.rebot@lpnu.ua, ORCID 0000-0002-3583-0800

УНІФІКОВАНА СИСТЕМА АУТЕНТИФІКАЦІЇ ТА УПРАВЛІННЯ ЗОБРАЖЕННЯМИ, ЗГЕНЕРОВАНИМИ ШТУЧНИМ ІНТЕЛЕКТОМ

Отримано: Березень 03, 2025/ Переглянуто: Березень 12, 2025/ Прийнято: Березень 20, 2025

Анотація. У цій статті представлено розробку системи генерації зображень, яка використовує технології цифрового водяного знака та вбудовування метаданих для визначення того, чи було зображення створене моделлю штучного інтелекту. Система виступає як проміжний сервіс між провайдерами та кінцевими користувачами, забезпечуючи безшовну інтеграцію цих технологій.

Зі зростанням обсягу контенту, створеного штучним інтелектом, стає все складніше відрізняти такі зображення від автентичних. Крім того, відсутність універсальних інструментів для керування згенерованими активами та вбудовування метаданих створює неефективність і ризики, пов'язані з автентичністю та інтелектуальною власністю. Тому важливо створити життєздатне централізоване рішення, яке інтегрує заходи захисту в будь-яке зображення, створене користувачем, незалежно від сервісу-джерела.

Система працює як проміжне програмне забезпечення, сумісне з наявними генеративними моделями, надаючи уніфікований інтерфейс для користувачів. Розроблений конвеєр спрощує як додавання водяних знаків у процес генерації, так і вбудовування метаданих. Інтуїтивний інтерфейс покращує зручність використання, а централізоване сховище дозволяє користувачам керувати та перевіряти створений контент.

Цей підхід є інноваційним, оскільки поєднує цифрове водяне маркування, інтеграцію метаданих і централізоване керування в єдиній платформі. На відміну від наявних інструментів, орієнтованих на конкретні платформи, ця система забезпечує міжсервісну функціональність. Рішення є надзвичайно актуальним для забезпечення автентичності контенту, управління інтелектуальною власністю та зручності користувачів. Воно підвищує довіру до цифрового контенту та забезпечує масштабовану архітектуру, яка може адаптуватися до різних платформ і потреб.

Майбутні дослідження можуть розширити цей підхід на ширші сфери інформаційних технологій — від неграфічних генеративних моделей до модулів на рівні операційних систем для захисту від згенерованих продуктів.

Ключові слова: генерація зображень, штучний інтелект, генеративні моделі, цифрове водяне маркування, веб-сервіс, інформаційна безпека.