IDENTIFAI
FIND ORIGIN

Watermarking Provenance and AI Image Detection



A comparative evaluation of SynthID MetaSeal C2PA and identifAI Revelio

Author	identifAI Team
Date	24 Nov 2026
identifAI - Whitepaper	

identifAI - https://identifai.net/ - Via Fratelli Gabba, 1/A, 20121 Milano (MI) - info@identifai.net

Watermarking Provenance and AI Image Detection

1 – Introduction

Digital watermarking embeds a machine-readable signal in media to support identification or attribution. Invisible image watermarks alter the image signal in ways intended to remain unobtrusive to viewers. A compatible detector searches for the mark or recovers an associated message. Watermarking therefore creates evidence that a participating system has marked the content; it does not inspect every possible image for signs of AI generation.

Signed provenance addresses a related need by attaching verifiable statements about an asset's origin and history. Both approaches are valuable, but their usefulness depends on whether the relevant evidence is available and interpretable when an image is assessed. This study evaluates that dependency through eight related datasets of Gemini-generated images and compares the resulting coverage with identifAI Revelio.

1.1 - Synth ID

Google's SynthID embeds imperceptible watermarks into AI-generated media, designed to resist common modifications [1]. This study uses the SynthID Detector portal (synthid.com), preserving its original outcome categories: Detected, Not detected, and Unsure. While a positive result confirms the presence of a supported watermark, a negative or inconclusive outcome does not prove human authorship. To attempt the removal of SynthID we used the open-source tool remove-ai-watermarks available on github [4]

1.2 - MetaSeal and open source VideoSeal

Meta's VideoSeal repository provides open-source watermarking models for images and video [2]. "MetaSeal" is used throughout this white-paper as the experiment's

label for the tested open-source implementation. The study does not evaluate Meta's proprietary Content Seal deployment, and its results must not be extrapolated to that system. A post-hoc watermark can mark an existing image regardless of how the image was created; here, the source images are independently known to be synthetic.
To attempt the removal of MetaSeal we used an open-source tool wmforger included in the official project [2].

1.3 - C2PA

C2PA is a signed provenance standard, rather than a pixel watermark. Its Content Credentials bind assertions to an asset, potentially including a declaration of AI generation. Credential presence, signature validity, signer trust and assertion meaning are distinct considerations [3]. In this study, C2PA-based identification refers to the test workflow recognising the synthetic-origin information associated with the applied credentials, not to visual classification by C2PA.

To remove the metadata containing the C2PA signature we used the open-source tool ExifTool [5].

1.4 - Content based detection

identifAI Revelio analyses images to classify artificial content without requiring a provider-applied label as an input prerequisite. Its role in this comparison is to assess whether synthetic images remain identifiable after changes to their watermark or provenance signals. The distinction is operational: verification asks what evidence accompanies an image, while content analysis evaluates the image itself.

2 - Datasets and test procedure

All source images were generated with Gemini and initially carried SynthID. The study derives eight datasets from this source collection. Each transformation preserves the known synthetic origin of the image: adding a watermark, signing a file or suppressing a signal does not turn an AI-generated image into a human-created one.

Set	Parent	Transformation
001	Source	Gemini-generated images carrying SynthID
002	001	Treatment intended to remove SynthID
003	001	Application of the open-source MetaSeal watermark
004	003	Treatment intended to remove MetaSeal
005	002	Application of MetaSeal after the SynthID-removal treatment
006	005	Treatment intended to remove MetaSeal
007	001	Application of C2PA credentials
008	002	Application of C2PA credentials after the SynthID-removal treatment

Table 1 - *Dataset construction. The parent column specifies the processing sequence, not merely a similarity between datasets.*

2.1 - Execution and recorded outcomes

Images in each dataset were assessed using the SynthID Detector portal, the watermark check for the tested open-source MetaSeal implementation, the C2PA verification workflow and Revelio. Results were aggregated by dataset. SynthID's three outcome categories were kept separate. For MetaSeal, the reported measure is the proportion with a positive watermark result. For Revelio, it is the proportion classified as artificial. The C2PA measures records identification through the applied provenance information.

A separate metadata-removal control checked C2PA identification after the credentials had been stripped using ExifTool [5]. This control is reported separately

from sets 001–008: sets 007 and 008 represent signed files before that removal. The control does not redefine any of the eight dataset labels.

2.2 - Comparison logic

Sets 001 and 002 isolate the SynthID-removal treatment. Sets 003 and 004 compare MetaSeal before and after its removal treatment on the SynthID-bearing branch. Sets 005 and 006 repeat that comparison on images already treated for SynthID removal. Sets 007 and 008 examine C2PA signing on the two branches. This structure allows both targeted effects and changes in other signals to be examined.

“Removal” denotes the treatment’s purpose. A negative result establishes non-detection by the tested verifier, not proof that every physical trace has disappeared. Rates are reported to two decimal places; differences are expressed in percentage points. The analysis uses dataset-level percentages, without reconstructing sample counts or claiming statistical significance from rounded rates. No result from a metadata-removal control is attributed to Revelio unless it belongs to the eight measured sets.

3 - Results across the eight datasets

Table 2 presents the complete SynthID distribution. Blank cells in the source tabulation are represented as 0.00%. “Unsure” remains a separate outcome and is not recorded as either a detection or an explicit non-detection.

Set	Detected	Not detected	Unsure
001	**100.00%**	**0.00%**	0.00%
002	5.00%	67.50%	27.50%
003	**100.00%**	**0.00%**	0.00%
004	99.17%	0.83%	0.00%
005	5.00%	71.67%	23.33%
006	2.50%	95.83%	1.67%
007	**100.00%**	**0.00%**	0.00%
008	5.00%	67.50%	27.50%
Total	52.08%	37.92%	10.00%

Table 2 - *SynthID Detector outcomes. Each row totals 100.00%. The total describes the experimental collection and is not an estimate of detection across arbitrary internet images.*

Set	SynthID detected	MetaSeal positive	C2PA identification	identifAI artificial
001	**100.00%**	**0.00%**	**0.00%**	**100.00%**
002	5.00%	**0.00%**	**0.00%**	**100.00%**
003	**100.00%**	89.17%	**0.00%**	**100.00%**
004	99.17%	6.67%	**0.00%**	**100.00%**
005	5.00%	93.33%	**0.00%**	**100.00%**
006	2.50%	4.17%	**0.00%**	**100.00%**
007	**100.00%**	**0.00%**	**100.00%**	**100.00%**
008	5.00%	**0.00%**	**100.00%**	**100.00%**

Table 3. *Positive outcomes by technology. The columns measure different evidence types: watermark detection, provenance-based identification and content classification. Their juxtaposition compares coverage on known-synthetic images, not equivalent confidence scores.*

In the separate C2PA control, identification through the embedded credentials was 100.00% before metadata removal and 0.00% afterwards. Files without applied C2PA credentials also produced 0.00% identification through that mechanism. Revelio classified 100.00% of images as artificial in every numbered set.

4 - SynthID results and interpretation

4.1 - Strong identification on the original branch

SynthID Detector returned detected for 100.00% of the original images in set 001. The rate remained 100.00% after MetaSeal application in set 003 and after C2PA signing in set 007. These results show that the tested addition of a second watermark or a signed credential did not reduce SynthID's positive detection rate on this branch.

Set 004, following the MetaSeal-removal treatment, retained a 99.17% SynthID detection rate, with 0.83% not detected. This is a decrease of 0.83 percentage points relative to set 003. Within this workflow, a treatment targeting MetaSeal therefore left SynthID detectable in almost all images that had not first undergone SynthID removal.

4.2 - A large reduction after targeted treatment

The transition from set 001 to set 002 reduces detected outcomes **from 100.00% to 5.00%, a decrease of 95.00 percentage points.** The remaining 95.00% consists of 67.50% explicit not_detected outcomes and 27.50% Unsure outcomes. It would be inaccurate to describe all 95.00% as explicit non-detections, but none of those outcomes provides an affirmative detected result.

Applying MetaSeal to the treated images leaves SynthID detection at 5.00% in set 005. The non-detected share rises to 71.67% and the uncertain share falls to 23.33%. Applying C2PA instead produces set 008, whose aggregate distribution matches set 002 exactly. Neither operation restores SynthID detection to its original level. Matching percentages do not establish that every individual file received the same verdict.

4.3 - Sequential treatment produces the lowest positive rate

Set 006, obtained by treating set 005 for MetaSeal removal, returns **2.50% detected, 95.83% not_detected and 1.67% Unsure.** Relative to set 005, detected decreases by 2.50 percentage points, not_detected increases by 24.16 points and Unsure decreases by 21.66 points. These are aggregate changes, not measured individual transitions between verdict categories.

The sequence illustrates why uncertainty matters. A workflow that forwards Unsure cases for review would retain some investigative opportunity in sets 002, 005 and 008. In set 006, however, the great majority receives an explicit not_detected outcome. Treating that verdict as proof of human creation would misclassify known-synthetic material.

4.4 - What the overall rate means

Across the experimental collection, 52.08% of outcomes are detected, 37.92% are not_detected and 10.00% are Unsure. This total reflects the mixture of intact and treated sets. The more informative operational comparison is between the 99.17–100.00% detected rates on the branch without SynthID-removal treatment and the 2.50–5.00% rates on the treated branch. The findings establish a limitation under the tested treatment, not a claim that SynthID has no value or fails under every transformation.

5 - MetaSeal results and interpretation

5.1 - Coverage after watermark application

The tested open-source MetaSeal implementation returned positive results for 89.17% of set 003, where the watermark was applied to the original SynthID-bearing images, and 93.33% of set 005, where it was applied after SynthID-removal treatment. Positive detection was therefore incomplete even before the dedicated MetaSeal-removal step: 10.83% and 6.67%, respectively, did not produce a positive watermark result.

These rates describe the full embedding and verification workflow. They do not, by themselves, isolate whether a non-positive result arose from embedding effectiveness, detection thresholds or another processing factor. The defensible conclusion is that watermark application did not guarantee subsequent positive identification under the tested configuration.

5.2 - Coverage after the removal treatment

Comparison	Before treatment	After treatment	Decrease
003 to 004	89.17%	6.67%	82.50 points
005 to 006	93.33%	4.17%	89.16 points

Table 4. *Changes in MetaSeal positive detection. Differences use the reported two-decimal rates. These are decreases in dataset-level detection, not per-image conditional removal-success estimates.*

After treatment, only 6.67% of set 004 and 4.17% of set 006 remain positive. Accordingly, 93.33% and 95.83% of those datasets do not yield positive MetaSeal identification. The residual positives show that the treatment does not suppress the check on every image. The overall reduction nevertheless leaves most known-synthetic images without a usable positive signal from this watermark.

5.3 - Differences between the two processing branches

Before MetaSeal removal, the SynthID-treated branch had a 4.16-percentage-point higher positive rate than the original branch: 93.33% versus 89.17%. After MetaSeal removal, it has a 2.50-point lower positive rate: 4.17% versus 6.67%. The observed decline is consequently larger on the branch that had already undergone SynthID treatment.

This difference is consistent with processing history affecting the subsequent watermarking workflow. It does not establish that the presence of SynthID itself causes the difference: the branches also differ in image processing, and aggregate rates cannot separate those effects. No statistical significance is claimed. The common finding is the substantial loss of positive MetaSeal coverage on both branches.

5.4 - Unmarked images and the scope of the finding

Sets 001, 002, 007 and 008 return 0.00% MetaSeal identification, consistent with no MetaSeal application in those sets. This is expected behaviour for a verifier seeking its own watermark; it also illustrates why such a verifier cannot identify synthetic content from sources that do not carry that mark. These findings apply only to the tested open-source implementation, not to Meta's proprietary Content Seal system.

6 - C2PA and Revelio results

6.1 - C2PA provides identification while the credentials are available

The C2PA workflow identifies 100.00% of the signed images in sets 007 and 008 through their applied synthetic-origin information. It **returns 0.00% for the unsigned sets**. The separate metadata-removal control **also returns 0.00% after the embedded credentials are removed**. The result demonstrates a dependency on the availability of provenance information in the tested file-based workflow.

The distinction is important: C2PA does not infer artificial origin from pixels. The reported 100.00% refers to recognition through the credentials used in this experiment. A generic C2PA signature on an unrelated file would not, by itself, establish AI generation; the assertion and its trust context determine what the credential supports [3].

Metadata removal does not break the cryptographic signature or prove that the provenance statement was false. It removes the evidence from the path used by this verifier. C2PA also supports durable provenance approaches, including rediscovery through external storage and soft bindings [3]. The experiment's removal result concerns embedded metadata and does not evaluate such recovery mechanisms.

6.2 - identifAI Revelio retains detection across all eight conditions

Revelio classified 100.00% of the images as artificial in every numbered dataset. It retained that result on the original images, after SynthID treatment, after MetaSeal application and removal, and after C2PA signing. Its observed synthetic-image sensitivity is therefore 100.00% within each of the eight evaluated conditions.

The most consequential comparisons are sets 002 and 006. In set 002, SynthID returns detected for only 5.00%, while neither MetaSeal nor C2PA provides a

positive result. In set 006, SynthID returns detected for 2.50%, MetaSeal for 4.17%, and C2PA for 0.00%. Revelio nevertheless identifies every image in both sets as artificial. This is direct evidence of complementary coverage on the evaluated synthetic images.

6.3 - Why combined watermark checks are still insufficient

The marginal rates do not reveal which images are positive for both watermarks. For set 006, the proportion with at least one affirmative watermark result can be no greater than approximately 6.67%, the sum of 2.50% and 4.17%. Thus at least approximately 93.33% has no affirmative result from either watermark, regardless of their overlap. This bound treats Unsure as non-affirmative, not as an explicit negative. Exact combined coverage requires per-image verdicts.

6.4 - The meaning of the comparison

The results support Revelio's role as a content-analysis layer where provenance checks leave gaps. They do not imply perfect performance on all images or attacks. Because the corpus contains synthetic images, these rates measure sensitivity rather than false-positive rate or precision on mixed real-world traffic. The study also does not establish the model's causal features or its resistance to attacks designed specifically against Revelio.

7- Conclusions

The results show both the value and the limits of watermarking and signed provenance. SynthID identifies 100.00% of the original images, but positive detection falls to 5.00% after the targeted treatment and to 2.50% in the sequential-treatment set. MetaSeal identifies 89.17–93.33% immediately after application, falling to 4.17–6.67% after its removal treatment. C2PA-based identification reaches 100.00% on signed images and falls to 0.00% when the embedded credentials are stripped in the metadata-removal control.

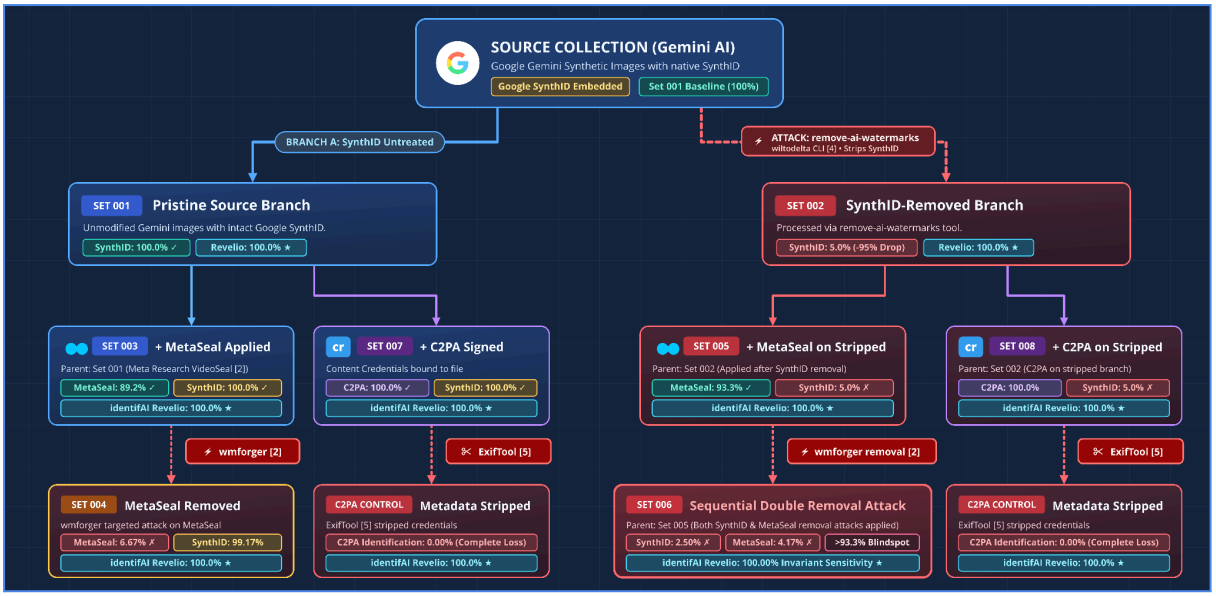


Figure 1. *From datasets to Detection: Revelio Technique Maximizes detection*

Revelio identifies 100.00% of the images as artificial in each of the eight datasets. The practical contribution is its continued coverage in conditions where supported watermarks are largely undetectable and C2PA credentials are absent. This demonstrates why **provenance verification and content-based detection should be used together.**

7.1 - A necessary complementary capability

Watermarks remain useful for supported-source attribution, and signed credentials remain useful for carrying verifiable provenance. Their absence, however, cannot resolve whether an image is synthetic. Systems that must assess unlabelled or deliberately treated images therefore need an additional capability that does not require those signals to survive. Image-analysis models such as identifAI Revelio provide that complementary layer.

An effective workflow preserves and verifies available provenance, checks supported watermarks, and analyses the image content. Uncertain or conflicting outcomes should trigger further assessment rather than an automatic declaration of authenticity. The objective is to combine evidence with different dependencies, while avoiding the assumption that an undetected marker means a genuine image.

7.2 - Scope of the conclusions

The study evaluates Gemini-generated still images and the specified processing paths. It does not measure performance across every generator, visual-quality impact, false positives on non-AI images or the robustness of proprietary Meta systems. The percentages are descriptive results for these datasets; sample counts and paired raw outcomes are not inferred from rounded values. The study is conducted by identifAI and includes its own detector. These boundaries qualify generalisation without changing the observed loss of provenance-based coverage after treatment.

References

[1] Google DeepMind. Google SynthID. Official overview of watermarking and verification.
https://deepmind.google/models/synthid/

[2] Meta Research. VideoSeal. Official repository for open-source image and video watermarking.
https://github.com/facebookresearch/videoseal

[3] C2PA. C2PA and Content Credentials Explainer, version 2.4. Provenance assertions, trust and durable credentials.
https://spec.c2pa.org/specifications/specifications/2.4/explainer/Explainer.html

[4] wiltodelta. remove-ai-watermarks. Open-source library and CLI for removing AI watermarks and metadata.
https://github.com/wiltodelta/remove-ai-watermarks

[5] Phil Harvey. ExifTool. Read, Write and Edit Meta Information.
https://exiftool.org/

Technical references accessed 6 November 2026. All experimental percentages in this white-paper are results of the identifAI evaluation.

identifAI | Watermarking and AI image detection

Thank you

For more information

Visit our website:

Contact: info@identifai.net