

The Recipe Is the Fingerprint: Creator Deanonymization from Default AI-Image Metadata at Scale

Daniel Okwor

Orygn LLC daniel@orygn.tech ORCID: 0009-0005-2546-181X

Preprint, July 2026

DOI: [10.5281/zenodo.21500861](https://doi.org/10.5281/zenodo.21500861)

Abstract

Every mainstream AI image generator writes the full recipe of how an image was made into the image itself. The base model, the LoRAs and embeddings, the sampler and its settings, the negative prompt, and the software build all go into the PNG as cleartext, by default, and without the user’s name. We show that this recipe is a behavioral fingerprint precise enough to name its creator. Across 15.7 million public AI images spanning four datasets and two platforms, and about 92,000 creators on Civitai, one of the largest platforms for openly shared AI images, a single image identifies its creator 55.8% of the time against the full population, and 54% even when near-duplicate reposts are forbidden from matching, using only the recipe and no account or server context. Chance is one in ninety thousand. This is closed-set identification against a known population. The open world, where the target may be absent, is harder and splits by distribution: an image from a different platform or toolchain matches almost no one and is rejected cleanly, while against a same-platform stranger not in the database about 40% of real creators are still named at a 10% false-accusation rate. Accuracy falls only gently as the pool of suspects grows and holds across a full year and a change of tooling generation. The signal is behavioral, not a proxy for tool choice: holding the base model fixed, a single image still names the individual 69 to 82% of the time, and a fingerprint built from structural fields alone, with no free text of any kind, still reaches 35%. It reaches ordinary users who own no custom assets, of whom 51% are identifiable, most often by the negative prompt they reuse without thinking. The magnitude of the leak is set by the platform: on DiffusionDB, drawn from the official Stable Diffusion Discord where the recipe is nearly identical for everyone, the same method reaches only 1.8%, and hashing usernames adds nothing. We demonstrate the full chain from an anonymous image to a named, living account, at scale: on 10,000 creators drawn at random, a single bare image recovers the correct account 55.9% of the time and 91% of the recovered accounts are live. We quantify the fix: only removing all free text, all resource identifiers, and the build string together drives identification below one percent, roughly a seventyfold drop though not fully to chance; partial redaction fails. We disclosed to the tool developers and the platform before publication. The working data is the deanonymization itself, the real accounts behind the images and the chains that link them, so it is kept private; we release only detectors, synthetic fixtures, and aggregate results.

1 Introduction

Every mainstream local image generator writes a record of how an image was made into the image itself. Automatic1111 and its forks write a text block of the prompt, negative prompt, and sampler settings; ComfyUI writes the complete node graph as JSON. Both are stored in the PNG’s ancillary

text chunks, unsigned and uncompressed. The feature is useful and intentional. It lets a creator or a viewer recover the recipe, and it powers the recipe-sharing culture of sites like Civitai. It is also on by default, and most users do not know it is there.

That the metadata exists, and that it can carry a prompt someone would rather not share, is not a new observation. The tool communities have raised it for years, and every major tool offers some way to turn it off. What has not been measured is how *identifying* the default metadata is once the prompt text is set aside. The concern has been treated as “my prompt is visible.” We show the risk is a different and larger one: the default recipe is a behavioral fingerprint precise enough to pick a single creator out of tens of thousands from one image, to reach ordinary users who share nothing custom, and to resolve, for public creators, to a named and live identity.

One fact frames the whole problem. The username is never written into the file. Only the recipe is. An image can travel to any platform, be reposted with no attribution, or leak from a breach, and it carries no account and yet enough to point back at one person. Recovering the creator from that recipe is therefore genuine deanonymization, not the reading of a handle the person already attached.

We make the following contributions:

- **The largest corpus assembled for metadata-based creator attribution:** 15.7 million public SFW AI images across four datasets and two platforms, roughly 92,000 Civitai creators, with a normalization and cross-source deduplication pipeline (Section 4).
- **Single-image identification at scale, and its floor:** one image names the correct creator 55.8% of the time out of about 92,000, with confidence intervals and graceful decay in population size, and a structural-only fingerprint with no free text that still reaches 35% (Sections 6, 7).
- **Evidence that the signal is behavioral, not a tool-choice proxy:** it survives holding the base model fixed (69 to 82% within a single checkpoint) and blinding the model feature, and it reaches ordinary users primarily through their reused negative prompt (Sections 7, 8).
- **The platform-verbosity thesis:** identification tracks how much a platform lets the recipe vary; the same method yields 55.8% on Civitai and 1.8% on DiffusionDB, and username hashing does not close the gap (Section 12).
- **A naming existence-proof at scale, responsibly bounded:** the chain from an anonymous image to a named live account, demonstrated on 10,000 creators drawn at random, of whom 55.9% are recovered and 91% of the recovered accounts are live, with a measurement of how often that account reaches an off-platform identity, and with ordinary users and all per-person chains kept private (Section 11).
- **An honest open-world map:** when the target may be absent, an image from a different platform is rejected almost perfectly, a same-platform stranger is the hard case (about 40% still named at a 10% false-accusation rate), aggregating a few images of one source reaches 75%, and a database a tooling generation out of date is nearly useless (Section 9).
- **Persistence and defenses:** a stale reference database still identifies 23% of images a year and a tooling generation later, and a mitigation matrix showing that only complete redaction of free text, resource identifiers, and the build string reduces identification to under one percent (Sections 12, 14). We disclosed to the tool developers and platform before publication.

2 Threat model

We consider an adversary who wishes to link an AI image to the person who made it, using nothing but the file.

Adversary capabilities. The adversary holds a reference database of AI images labeled with creator accounts, built once from public sources, and can extract the cleartext metadata from any file. This is a modest assumption: the datasets we use are themselves such databases, freely downloadable, and the metadata reader is a hundred lines of code. The adversary has no access to accounts, servers, upload logs, or any information beyond the target file and the reference database.

Target. The target is a single AI image that reached the adversary stripped of account context: reposted without attribution, shared as a bare file, leaked, or scraped. The image carries its generation metadata, because the platforms and channels where AI images circulate largely preserve it (Section 13). The username is not in the file.

Goal and success. The adversary succeeds by naming the creator, which we operationalize in two standard biometric framings. In *identification* (one to many), the adversary ranks the reference database and the top candidate is correct. In *verification* (one to one), the adversary decides whether two images share a creator. We report both. Identification against a known population is the primary setting an adversary with a reference database faces; we also report the harder open-world setting, where the target may not be in the database at all (Section 9).

Out of scope. We do not consider active tampering with metadata, side channels beyond the file, or attacks on the generation tools themselves. We study people who share AI images online, which is a large population but not the whole world; individuals who never publish are outside our data by construction, though the mechanism does not depend on publishing. We scope the corpus and all published analysis to SFW content.

3 Related work

Our result sits at the intersection of three lines of work, and its novelty is the intersection itself.

Fingerprinting and anonymity sets. Device and browser fingerprinting established that a small set of innocuous configuration values, taken together, can uniquely identify a user, and framed that leakage in bits of identifying information and anonymity-set size [1]. We adopt that framing directly: an AI image’s recipe is a configuration fingerprint, and we report the identifying information it carries in bits (Section 6). The difference is the surface. A browser fingerprint is read by a server the user chooses to contact; an image fingerprint travels inside a file that outlives any session and can be read by anyone who obtains the file.

Authorship attribution and stylometry. A long tradition recovers the author of a text or a program from stylistic regularities they are not aware of leaving, including work that de-anonymizes programmers from coding style at scale [2] and de-anonymization of ostensibly anonymous records more broadly [3]. Our fingerprint is stylometry of a different kind: not of the words a person writes or the code they author, but of how they configure a generative tool. The regularity is the same in spirit, a stable personal habit the person does not think of as identifying, except that here it is recorded verbatim in the artifact rather than inferred from it.

File metadata and AI-image provenance. Metadata leaks are old, from EXIF geolocation in photographs to authorship fields in office documents. The generation recipe is a new member of that family and a uniquely rich one, because it encodes behavior rather than a single datum.

On the provenance side, the C2PA standard [5] cryptographically signs how an image was made; the metadata we study is its unsigned, unmanaged predecessor, present by default in the very files the signing standard was built to improve on. A separate line actively attributes diffusion outputs to a user by planting a per-user watermark in the model or its weights [8, 9]; that requires the generator’s cooperation and a signal inserted in advance, whereas we read the identifying information the default pipeline already writes for free, with no cooperation from anyone. Closest to our subject, the DiffusionDB dataset [4], built from the official Stable Diffusion Discord server, distributed only SHA-256 hashes of its contributors’ usernames out of an explicit concern that “explicitly linking them to their creators can cause harm,” and the 2026 UpGuard analysis [6] recovered prompts and generated images from internet-exposed ComfyUI instances configured for anonymous access. Both anticipate the risk; neither measures it. DiffusionDB pre-empted a linkage it did not demonstrate, and UpGuard needed an exposed server instance rather than merely a file. We show the linkage is real, is achievable from the file alone, reaches about 92,000 creators, and puts a name to specific people.

Awareness in the tool communities. The generation tools have fielded privacy requests about embedded metadata for years, and each offers an opt-out. This work does not claim to discover that the metadata exists. It measures, for the first time, how identifying the default metadata is, and it is that measurement, not the existence of the field, that should decide the default. The reply that this was already known does not weaken the result; it sharpens it. The field was known to be present and still ships enabled by default, precisely because no one had shown that it names people. Awareness of the field was never the missing piece; a measurement of its power was.

4 Corpus

We assemble, to our knowledge, the largest corpus gathered for metadata-based creator attribution: **15,694,612 public SFW AI images** across four datasets and two platforms. We use only already-public datasets and perform no new scraping.

Sources. Three datasets originate on Civitai, where users run their own tools and the full recipe reaches each file: hanruijiang (a 2023 snapshot, 2.5M images), thecraft (an earlier 2023 snapshot, 325K), and Civiverse [7] (a modern 2024 snapshot of SDXL- and Flux-era generations, 4.0M). The fourth, DiffusionDB (14M images, 10.46M after SFW filtering), is drawn from the official Stable Diffusion Discord server and serves throughout as a control: its usernames are pre-hashed by the dataset’s own authors and its settings are near-uniform. Every source is normalized to one common record, and every username is reduced to a stable hashed key for analysis while the working tier retains the real handle privately.

Deduplication and honest denominators. The two older Civitai sets overlap heavily: of thecraft’s roughly 185,000 SFW images, 84% (about 155,000) are the same Civitai images as hanruijiang, so naively summing the sets would double-count those. We deduplicate on the Civitai image identifier and keep only the 29,752 records unique to thecraft. Civiverse, from a different time window, overlaps hanruijiang on only 0.88% of a content signature and is effectively disjoint. After deduplication the corpus is 15.69M images, or 14.69M once exact content-repeats are also removed.

The split that matters. Scale here is not a single number. The **Civitai portion is 5.23M images and about 92,000 distinct creators** after SFW filtering and cross-source deduplication, and it carries every attribution result in this paper. **DiffusionDB is 10.46M images and about 10,000 creators**, and it functions as the negative control that establishes the null at scale

(Section 12). We report the 15.7M figure as what it is, the total corpus analyzed, and we never let the DiffusionDB images inflate an identification number.

Coverage. Not every image can be fingerprinted. About 3% of Civitai records carry no usable recipe at all; these are excluded from attribution and reported as coverage rather than silently dropped. We scope the entire corpus and all published analysis to SFW content, using each platform’s own rating, so the finding cannot be dismissed as an artifact of adult material.

5 Method

The fingerprint. We reduce each image to a set of features drawn only from the generation recipe, never from the pixels and never from the descriptive positive prompt. A feature is one of: the base model or its hash, a LoRA or embedding by name, a hypernetwork, a hash of the normalized negative prompt, the settings tuple of sampler, steps, guidance, and clip-skip, and the software build string. LoRA and embedding names are resource identifiers, nearer to a filename than to content, and we extract only the identifier even when a tool records it inside the prompt string. We define two tiers so the reader can see exactly where the signal comes from. **Tier 1** is the structural fingerprint: everything above except the negative prompt, containing no free text of any kind. **Tier 2** adds the negative-prompt hash. The positive prompt enters neither.

Similarity and weighting. Not all features are equally telling. A sampler shared by half the corpus says little; a private LoRA used by one person says almost everything. We weight each feature by its inverse document frequency and score two fingerprints by their IDF-weighted cosine similarity, treating features present in more than ten percent of the corpus as stopwords for candidate generation. An inverted index lets one query be scored against millions of records, so identification runs against the full population rather than a sample of it.

Tasks. We evaluate two standard biometric tasks. In *identification*, we hold out one image of a creator, rank every other record in the reference set, and record a success when the top-ranked record shares the creator; chance is one over the number of creators. In *verification*, we score pairs and decide same-creator or not, reported as area under the ROC curve and as precision at a decision threshold. We report accuracy with bootstrap confidence intervals, and, because a single number can mislead, under four honest cuts: pooled, content-deduplicated, capped at twenty images per creator, and per source.

Reproducibility. All sampling is deterministic, seeded by a hash of the record identifier, so every number regenerates exactly. Where we subsample queries for tractability we state the count, and the headline numbers move by less than one point across seeds.

6 One image against ninety-two thousand creators

The central result is simple to state. Given a single image with its account removed, and a reference set of about 92,000 Civitai creators, the top-ranked match is the true creator **55.8%** of the time. This is what we report throughout as top-1 accuracy: out of the whole population, the single best-ranked candidate, our one guess, is the right person. The estimate carries a 95% confidence interval of 55.5 to 56.2, computed over all 71,272 held-out queries. Random chance at that scale is 0.0011%, roughly one in ninety thousand, so the method is about fifty thousand times better than guessing. On the older and smaller hanruijiang set alone, against 34,834 creators, the figure is **68.6%** (confidence interval 68.0 to 69.1). Weighting features by rarity is what carries the result

at scale: without it the pooled number falls from 55.8 to 49.3%, because a private LoRA and a universal sampler would otherwise count the same.

The number survives every honest cut. A single accuracy figure can hide a repost artifact or a handful of prolific users, so we test for both. To rule out that the method merely re-finds the same image posted twice, we forbid it from matching any candidate whose content is identical to the query, on the full population and with no change to which creators are evaluated: top-1 falls by less than a point, to 54.2%. It is finding *different* images by the same person, not copies of the same one. The four cuts of the Civitai population (Table 1) round this out: capping every creator at twenty images, so that no prolific account can dominate, moves the number by half a point, and content-deduplication raises it to 63.2%, though because that cut also removes about 7,700 creators the rise partly reflects the smaller population, not reposts alone. The signal is not an artifact of duplication or of a few heavy users.

Table 1: Single-image top-1 identification on the Civitai population under four cuts.

Cut	Top-1	Creators
Pooled	55.8%	91,741
Content-deduplicated	63.2%	84,032
Capped at 20 images per creator	56.0%	91,741
Per source (hanruijiang)	68.6%	34,834

Accuracy decays gently as the crowd grows. The natural question is how the number scales with the size of the suspect pool. It falls far more slowly than one might expect (Figure 1). A ninety-fold increase in the size of the crowd, from about a thousand creators to ninety-two thousand, costs about twelve points. The fingerprint is individuating enough that adding distractors barely dilutes it, which means the result does not depend on a conveniently small database.

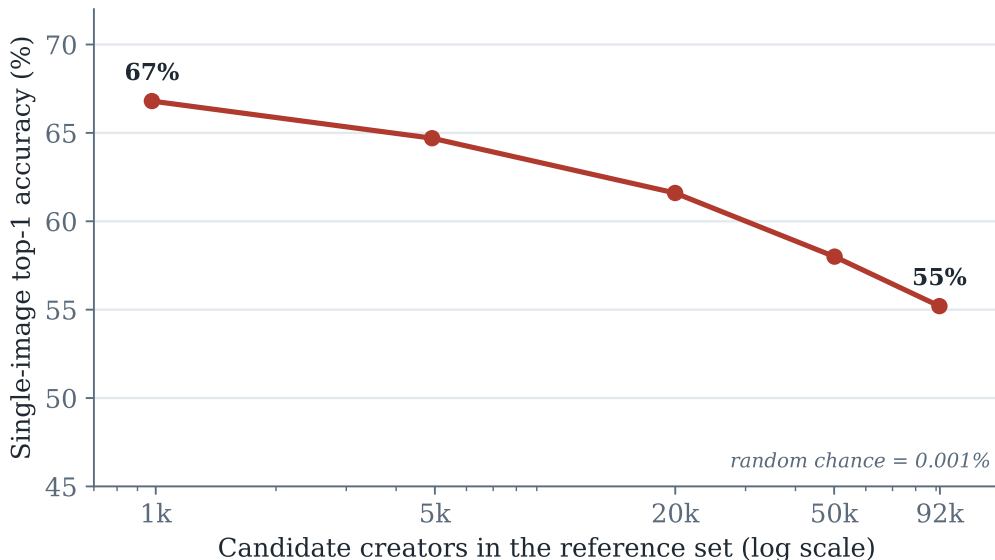


Figure 1: Single-image identification accuracy against the size of the candidate pool, on the unified Civitai population. Accuracy falls only gently as the reference set grows from about a thousand creators to ninety-two thousand; random chance is roughly one in ninety thousand.

One reference image is enough. Accuracy also barely depends on how many images of the target the adversary already holds. Capping each creator’s gallery at a single reference image beyond the query, the hardest version of the task, still yields **63.1%** on hanruijiang; more reference images add only a few points. The adversary does not need a rich history of the target. One prior image suffices.

In bits. Framed as information, identity among 91,741 creators is 16.5 bits. The average image leaks **13.97** of those bits, and the median image leaks all 16.5, which is to say the median image ranks the true creator first out of 91,741. For 82.9% of images the true creator falls within the top 22, a leak of at least 12 bits. These figures are the true creator’s rank, not the size of an anonymity set, so they measure correct placement, not mere narrowing.

7 The fingerprint is behavioral, not a proxy for tool choice

The most natural skeptical reading is that we have built an elaborate detector for “which base model does this person use.” Two experiments rule that out.

Blinding and holding the model fixed. Removing the model feature from the fingerprint entirely lowers pooled accuracy by only 1.9 points. Conversely, restricting both the query and the reference set to a single popular checkpoint, so that everyone under comparison used the identical model and only individual habits remain, leaves the individual identifiable **69 to 82%** of the time. Among 2,353 people all running PonyDiffusionV6, one image finds the right person 79.4% of the time; among 1,391 on AutismMix, 80.8%. These same-model pools are far smaller than the full population, so the higher rate partly reflects the smaller crowd; the cleaner evidence that the signal is personal is the blinding above, where removing the model costs only 1.9 points across all 91,741 creators. Holding the model constant does not defeat the attack, because what remains is individual habit.

Where the signal lives. Decomposing the fingerprint into families, on a per-creator-capped slice of hanruijiang where the full fingerprint scores 68.9%, no single family is the whole story (Table 2). The identity is distributed and redundant. Remove the negative prompt and 54.9% remains; remove every resource and 50.1% remains. Adding the widely-shared base model to the personal resources even lowers accuracy slightly, from 52.3 to 47.5%, because a common checkpoint matches many candidates and dilutes the rarer features under cosine weighting: the individuating signal is in the rare resources, not the model. No one field is load-bearing, which is why partial redaction fails (Section 14).

Table 2: Single-image top-1 using restricted feature families (hanruijiang, capped at 20 images per creator; the full fingerprint scores 68.9%).

Feature family	Top-1
LoRA and embeddings	52.3%
Negative prompt	46.9%
What-you-run (model, LoRA, embedding, hypernetwork)	47.5%
Model or hash only	15.6%
Settings only	6.6%
Everything except the negative prompt	54.9%
Everything except what-you-run	50.1%

The no-text floor. Because the strongest single family for many users is the negative prompt, which is text, we report the structural fingerprint separately: model, resources, settings, and build string, with no free text of any kind (Figure 2). With no free text at all, a single image still identifies its creator 35.4% of the time, and 75.6% for anyone who uses a custom asset. On the modern Civiverse subset, where resource names come from structured fields and are never parsed from a prompt, the no-text figure is 33%, confirming that this floor does not secretly depend on reading prompt text. The negative prompt roughly doubles the number for ordinary users and is treated in Section 8; the behavioral core stands without it.

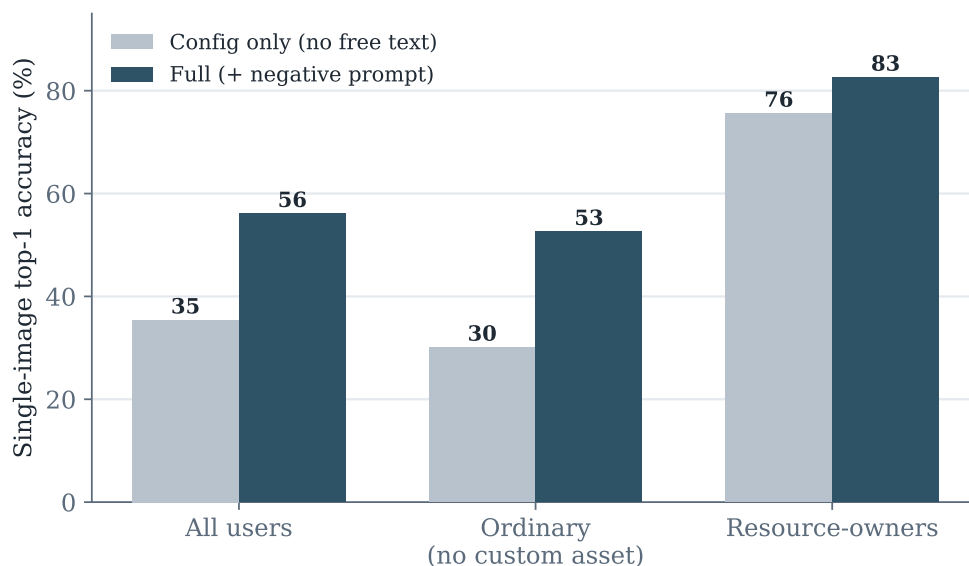


Figure 2: The structural fingerprint with no free text of any kind (Tier 1) against the full fingerprint that adds the negative prompt (Tier 2), by user category, on the unified Civitai population. Even with no text at all, ordinary users are identified 30% of the time and resource-owners 76%.

8 The ordinary user and the negative prompt

The people most exposed are not the ones with rare custom models. They are ordinary users, and the field that gives them away is one almost no one thinks of as identifying.

Ordinary users are identifiable. Restricting to the 25,245 creators who own no globally unique resource, who use only mainstream shared models and LoRAs, **51.0%** are still single-image identifiable, with the rate landing between 51 and 53% depending on exactly how the ordinary user is drawn. Verifying a representative random sample of 2,025 of them against their public profiles, **90%** are live accounts, **68% publish no models at all**, and only **13%** link any off-platform identity, against 57% for the commercial creators of Section 11. These are hobbyists, and they are both the most exposed group and the most private beyond the account, which is to say the group with the least awareness and the least recourse.

The negative prompt is a signature. For 71% of these ordinary identifications, the deciding feature is the negative prompt. This holds up under scrutiny. Of the 490,909 distinct negative prompts in the corpus, **90.4% are used by exactly one creator**, and only 89 are used by more than a hundred. People develop a personal negative prompt, some particular arrangement of quality and anatomy tokens with their own edits, and paste it into every generation. It feels like boilerplate; it functions as a signature. When we delete every negative prompt shared by more than twenty creators, removing any chance that a viral template is doing the work, ordinary-user accuracy falls only from 52.2 to 50.7%. Sharper still: of the identifications the negative prompt drives, 76% rest on a negative used by exactly one creator in the entire corpus, and deleting every negative shared by more than one creator leaves single-image identification at 48%. Nor is the uniqueness an artifact of loose hashing: under an aggressive normalization that collapses token order, weight syntax, punctuation, and duplicate tokens, so that any two negatives with the same token set become one, uniqueness barely moves (89.2% used by a single creator) and ordinary-user identification is unchanged within a point. The tell is an individual token set, a personal habit, not a downloaded template.

Who is safe. The mirror of the result is instructive. Among sampled creators, **22.9%** are hard to identify, with single-image accuracy at or below 25%, and their profile is uniform: small fingerprints, no custom assets (3.3% versus 30.2% among the exposed), and, decisively, no rare feature. The average most-distinctive feature of a safe user is common; the average most-distinctive feature of an exposed user is rare. Safety here is a synonym for generic, and most users are not generic. A further 3% of records carry no usable recipe at all and are safe by omission.

9 Verification and the open world

The headline numbers assume the target is in the database. This section relaxes that: first the targeted question, is this image my specific suspect, then the full open world, where the person behind an image may be in no database at all.

Verification. Asked instead to decide whether two images share a creator, the method reaches an area under the ROC curve of **0.83** on hanruijiang, where 0.5 is chance and 1.0 is perfect separation. More useful for an adversary is the precision available: at a similarity threshold of 0.2, deciding “same creator” is correct **99.9%** of the time while still recovering 53% of true pairs, and at 0.3 it is essentially perfect precision at 47% recall. An adversary who wants near-certainty before naming anyone can have it, at the cost of recall.

The open-set dragnet. In a true open-set test we enroll most creators and hold the rest out as unknowns whose images have no correct answer and must be rejected. A naive similarity threshold is a poor rejecter: an unknown person’s image still resembles *someone* in a database of ninety thousand, and false accusation stays near 58%. The fix is to accept a match only when it is both strong and rests on a rare feature, rejecting on the quality of the evidence rather than its raw similarity. Under that rejecter, and against the hardest possible unknowns, held-out creators from the same platform who are statistically identical to the enrolled, about **40%** of real creators are still correctly named at a 10% false-accusation rate, and roughly half at 20%. Naming a stranger cold from a single image is genuinely harder than the closed-set number, and we report it as such; the targeted question, “is this image my specific suspect,” is the verification result above, and it is strong.

An image from elsewhere matches almost no one. Those same-platform unknowns are the worst case, drawn from the very distribution the database was built on. A more realistic

open-world query is an image from a *different* platform and toolchain, which describes most of the open internet. We test this directly, querying DiffusionDB images, built on a different pipeline with different models and settings, against the Civitai database: they share almost no features with it, the rarity-weighted rejecter discards them, and full recall is retained at essentially zero false accusation. The risk of a false accusation comes not from the wider internet but specifically from other creators on the same platform who configure their tools similarly; an out-of-distribution image is rejected almost perfectly. One caveat keeps this honest: DiffusionDB’s recipes are unusually homogeneous, so it is a favorable foreign test, and a more varied foreign source could produce somewhat more false matches. No dataset offers ground truth for an arbitrary image from the open internet, so we bound the claim to what can be measured.

More than one image. A determined adversary rarely holds just one anonymous image; a leaked folder or an anonymous account carries several. Aggregating the recipes of a few images believed to share a source sharpens the attack steeply (Table 3): a single image names the creator 56% of the time, five images 75%. This is closed-set recall, though, not a free lunch in the open world: more evidence also makes a same-platform look-alike match more confidently, so it does not by itself ease the rejection problem above. One stray file is the floor of the attack, not its ceiling.

Table 3: Closed-set recall as the adversary aggregates more anonymous images believed to share one source.

Anonymous images aggregated	1	2	3	5
Correctly named	56%	65%	68%	75%

A stale database in the open world. The open world is far bleaker for an out-of-date adversary. Turning the 2023 database on 2024 images, across a tooling-generation gap (Section 12), open-set recall against unknowns collapses to about **2%** at a 10% false-accusation rate: a cross-generation match rests on a single persistent habit, too weak to tell from coincidence once the target may be absent. A database a tooling generation old still identifies people known to be in it (Section 12), but as an open-world dragnet it is nearly useless.

10 Unsupervised clustering

Everything above uses labels. To ask whether the metadata organizes into creators on its own, we cluster the 5.07M fingerprintable Civitai records with no labels, joining any two that share a rare feature and taking connected components, then score the clusters against the true creators. At the rare-feature threshold, the clustering reaches a **V-measure of 0.742 with homogeneity 0.997**: when the metadata groups two images together, they are almost never different creators. Completeness is bounded near 0.59 because generic images, with no rare feature to bind them, remain singletons, which is the same population that Section 8 found safe. Loosening the threshold chains unrelated creators into a single giant component and homogeneity collapses, so the rare-feature operating point is the meaningful one. The metadata partitions into people almost cleanly, and the only images it leaves unclustered are the ones it cannot identify anyway.

11 Naming: from an anonymous image to a real person

Attribution accuracy is abstract until it has a name attached. This section makes the leap, and does so within a strict ethical boundary: we resolve only to accounts a person chose to make public, we stop at what they published about themselves, and we name no one in this paper. The real handles and full chains are held privately.

Signatures. For **8,191** Civitai creators, some LoRA or embedding in their fingerprint appears under exactly one account across all 5.2 million Civitai images. That resource is a personal signature. For **201** of them, the signature is self-documenting in the plainest way: the resource filename contains the creator’s own handle, so the recipe names them with no lookup at all.

A worked case. Consider one prolific and entirely public creator, pseudonymized here. **390** of the resources in their fingerprint are unique to them across the whole corpus, and **356** of those are models they published on the platform under their account. A single one of their images identifies them **99.7%** of the time, 299 of 300 in a held-out test against the full population. The chain is short and fully mechanical: an anonymous image, a LoRA filename in its metadata, a search on that name, their live public profile.

The demonstration on random people. The worked case invites the obvious objection that we chose a distinctive person, so we stopped choosing. We drew **10,000 Civitai creators at random** from those with at least two images, the minimum for a held-out test, and with no selection for how identifiable or prominent they are; we took a single image from each, stripped its account context to a bare file, and matched it against the roughly 92,000-creator reference set. This is the closed-set case by construction: the target posts on the platform and is therefore in the reference population, which is also the realistic one, since the people an adversary can hope to name are exactly those who left a public footprint. A single bare image recovered the correct account for **5,589 of them, 55.9%**, essentially the rate Section 6 reports as an accuracy, now realized as the recovery of a specific account on a fresh random draw of people. What gives a random person away is rarely a rare asset: the deciding feature was the negative prompt for 65% of them, a LoRA for 22%, the base model for 11%, and a setting or the build string for the remainder, so for the ordinary person the tell is most often the very list of things they told the model to avoid.

They were not trying to be found. To confirm the recovered accounts are real and current, we checked every one of the **5,589** recovered accounts against the platform, reading only each account’s own public profile and opening nothing further. **91.3%** resolved to a live account, 5,103 of 5,589, with 486 gone and none left undetermined, matching on a separate sample of identifiable creators the live share we report next. Yet only **11.5%** of these live accounts publish any off-platform link, and 3% a payment account, against 57% for the signature-identifiable creators below. That gap is the finding, not a discrepancy: a random draw of people who post AI images is overwhelmingly ordinary hobbyists who advertise nothing and keep no trail elsewhere, and they were still named to a live account more than half the time from a single file that never carried their name. For them the exposure is the naming itself. The per-person chains, with real handles, are held privately; only these aggregates are reported.

How far the account reaches. We then asked what the recovered account leads to, using only the creators’ own public profiles. Of 1,000 identifiable creators, 905 resolve to live accounts, **90.8%** publish models, and **57% self-publish at least one off-platform link**, a quarter of them a payment account such as Patreon or Ko-fi. For a majority, then, the full path runs from the bytes of one image to a persistent identity the person maintains elsewhere, and every step past the account is something they chose to make public. We performed no correlation, no reverse-image

search, and no lookup beyond a profile page, and we did none of this for ordinary users.

The spectrum. Among active creators, those with at least eight images, **63.2%** are highly identifiable, named by a single one of their images at least three times out of four, but only **17.6%** owe it to a unique resource. The remaining **45.6%** are identified by behavior alone, with no custom asset, which is the ordinary-user population of Section 8. The exposure is not a property of prominent creators. It is close to universal among people who post, and it reaches hardest into those who never intended to be found.

12 Persistence and the platform

A stale database still works. An adversary does not need a current reference set. We built one from the 2023 snapshot alone and used it to identify images from 2024, a gap that spans not just a year but a change of tooling generation, from the Stable Diffusion 1.x era to SDXL and Flux. For the 27,351 creators present in both snapshots, a single 2024 image was matched to the right 2023 identity **23.4%** of the time, against a chance of 0.0028%, one in the 35,748 creators in the 2023 database. A database built once keeps working.

How the fingerprint ages. Directly measuring how a creator’s fingerprint drifts, we find it is highly stable in the short term and decays slowly (Table 4). Self-similarity stays two to four times above the different-creator baseline for a full year before fading into it. This reconciles with the temporal result: the average fingerprint drifts toward noise over a year, yet cross-year identification still succeeds a quarter of the time, because it needs only a single persistent habit rather than a stable whole. The attack’s shelf-life is months for the average creator and longer for anyone who keeps one distinctive element.

Table 4: Average fingerprint similarity between two images by the same creator, as a function of the time between them, against the different-creator baseline.

Time between two of a creator’s images	Fingerprint similarity
Under a day	0.62
One to seven days	0.37
One to four weeks	0.29
One to three months	0.20
Three months to a year	0.12
Over a year	0.06
(two different creators)	0.07

The platform decides the anonymity. The magnitude of the leak is not a constant of AI images; it is a property of how much a platform lets the recipe vary. DiffusionDB is our control: 14 million images drawn from the official Stable Diffusion Discord server, where a shared bot pushed almost everyone to the same settings. Its sampler takes one value 98.4% of the time. Running the full attack on it, and granting it every structured field including image dimensions, identification reaches only **1.8%**. The recipe is present, but it is nearly the same recipe for everyone, so it identifies almost no one. Civitai, where users bring their own tools and settings, yields 55.8% under the identical method. The instructive part is that DiffusionDB hashed its usernames out of concern for exactly this risk, and it did not matter, because the homogenized recipe protected people regardless. On Civitai the same hashing would protect no one. Anonymity here is a function of metadata verbosity, not of whether the name field is hidden.

13 The recipe is in the file, and it survives

Everything above is computed over dataset metadata. This section grounds it in the file itself.

The recipe lives in the PNG’s ancillary text chunks, unsigned and uncompressed. Automatic1111 and Forge write a text block; ComfyUI writes the node graph as JSON. A hundred-line reader walks the chunks and returns the fingerprint. To confirm this is current behavior and not an artifact of old files, we generated three fresh images in the current ComfyUI release, giving each a different prompt while holding one thing constant, the negative prompt. The three images look nothing alike; read from their bytes, their fingerprints were identical. We then uploaded all three to Discord and downloaded the served copies, which came back byte for byte identical to the originals, confirmed by matching hashes, with the full recipe intact.

That result sets the reach of the threat, and it rests on a single mechanism. Because PNG is lossless, a platform keeps the recipe by default and has to actively re-encode the file to remove it, so the only question for any channel is whether it re-encodes. We tested both ends of that question directly: the Discord round-trip above preserved the recipe, and an image we pulled from 4chan came back with every text chunk gone, re-encoded on the way in. For the channels in between we reason from the mechanism rather than a per-platform test. Those where AI images actually circulate mostly do not re-encode: the generation platforms, direct file hosts, chat attachments saved as files, model repositories, and any breach of stored files serve the original bytes with the recipe intact. Those that strip it are the ones that re-encode by design, the major social networks and the privacy-minded platforms such as Mastodon and Bluesky, which re-encode every upload specifically to protect their users. The recipe therefore survives precisely where this content concentrates, and is washed out mainly on the platforms people pass through rather than the ones where they build a presence.

14 Defenses and coordinated disclosure

The problem is fixable, and the fix is specific enough to state precisely. We measured identification under a series of stripping policies applied to every image, since a real mitigation removes fields from all files, not one (Table 5). No partial measure works. Removing resource names alone leaves ordinary users almost untouched; removing the negative alone leaves resource-owners at 76%. Only removing all free text and all resource identifiers together brings identification to roughly 3%, and the small residual is carried by the settings and the software build string, so a thorough mitigation strips those too. The tools already offer ways to do this, but they are off by default, sometimes incomplete, and rarely found, and the platforms do not tell users the metadata is there. Civitai’s

own terms and privacy policy never disclose that the recipe is embedded in the downloaded file, that it survives when the file is shared, or that it can identify the creator [10]. A default that ships enabled, is hard to find, and is never disclosed is not an opt-out most people can meaningfully exercise.

Table 5: Single-image top-1 identification after each stripping policy is applied to every image.

What is stripped	All	Ordinary users
Nothing (default)	55.8%	52.2%
Negative prompt	34.3%	28.9%
All resource names	46.5%	43.8%
All free text and all resource names	3.1%	2.7%
Everything but settings	0.8%	0.8%

We disclosed these findings and this recommendation to the tool developers, ComfyUI and Forge, and to Civitai as the platform that stores and serves the metadata, ahead of publication. We framed it not as a vulnerability but as a privacy-by-default question, and asked each to consider making stripping the default, making the existing opt-out complete and reliable, and warning users that a shared image can identify them. We sent these disclosures on 16 July 2026, a week before this preprint was made public.

15 Ethics

This study analyzes only already-public datasets and performs no new scraping. It is scoped to SFW content throughout. It operates on a two-tier model: the working data, which retains real handles and the full resolution chains, is kept private on local storage and is never published; everything released is aggregate, hashed, or synthetic. That private tier exists only to verify the results and the disclosures; it is stored securely, shared with no one, and used for nothing else. In the naming analysis we resolve only to accounts the person made public, we stop at links they published about themselves, and we name no individual in the paper. We deliberately study ordinary, non-commercial users only in aggregate, because they are the group least able to absorb exposure. The fingerprint is computed from the recipe, not the pixels, so it is blind to what an image depicts, so the exposure it measures applies uniformly across content categories, sensitive or explicit material included; our SFW-only scope is a limit we imposed, not a property of the mechanism. Civitai requires generation metadata on its sexualized content and hides posts that lack it [10], so that category carries the recipe by mandate. We deliberately did not turn the method on it, and did not need to: the content-independence of the mechanism already establishes the reach, so running the attack on real people there would add no finding, only harm. The warning is the contribution; running the attack ourselves would only enlarge the harm we document. The work is dual-use, and we treat that seriously: we pair the finding with a concrete, tested mitigation and with coordinated disclosure to the affected tools and platform, so that the net effect is to make defense easier rather than attack. In weighing publication we followed the established principles for security research, beneficence, respect for persons, and respect for law and the public interest: harm is minimized by resolving only to self-published information and reporting people in aggregate, and it is outweighed by the benefit of documenting an on-by-default privacy leak and handing defenders a concrete remedy. The author declares no competing financial interests.

16 Scope and limitations

We state the boundaries of the result plainly. It concerns people who share AI images online, a large and growing population but not everyone; individuals who never publish are outside our data by construction, though nothing in the mechanism depends on publishing. The primary numbers are closed-set identification against a known population, the setting an adversary with a reference database faces; the open-world case, where the target may be absent, is harder, and we report it directly in Section 9: an out-of-distribution image is rejected almost perfectly, a same-platform stranger is still named about 40% of the time at a 10% false-accusation rate, and a database a tooling generation out of date is nearly useless as a dragnet. The strongest single feature for ordinary users is the negative prompt, which is text, though it is reused configuration text rather than a description of the image, and the structural fingerprint with no text at all still identifies a third of people. An account is not always one person; we confirmed the headline holds, and in fact rises slightly, after removing automated batch accounts that make up 3.9% of accounts and 28.7% of records. Identifiability rises with a creator’s output, from 42% for those with only two or three images to 72% for the most prolific, but even a two-or-three-image footprint is identified 42% of the time, some forty thousand times chance, so the exposure is not confined to heavy users. The temporal result rests on two snapshots of a single platform. The findings apply to open tools that expose the recipe and much less to closed services that homogenize it, as DiffusionDB demonstrates. Our sources span two platforms and end in 2024; extending the measurement to more platforms and to current data, including additional datasets we have requested access to, is natural future work.

17 Reproducibility

We release the detector that reads the recipe from a PNG, a set of synthetic fixtures with planted metadata for testing it, and the aggregate result tables. We do not release the corpus images, the extracted metadata, real handles, or a form of others’ datasets that reconstructs private data; reproduction is intended to run on one’s own files or on the public source datasets, which remain available. So that the identification numbers can be checked without our raw data, we additionally release the corpus in opaque form: every feature and every creator becomes a meaningless integer index, with feature types retained but no values and creator labels hashed. Because the attribution depends only on which features co-occur and which records share a creator, this artifact reproduces the pooled rate, the four cuts, and the clustering exactly, while revealing no identity, no resource, and nothing a dictionary attack could invert. All evaluation sampling is deterministic and seeded, so every figure regenerates exactly.

18 Conclusion

Default AI-image metadata is a behavioral fingerprint. The recipe that every mainstream generator writes into every file, by default and without a name, is enough to pick a single creator out of about ninety thousand from one image, to reach ordinary users who share nothing custom, to resolve for public creators to a named and living identity, and to keep working a year after the database was built. The leak is not inherent to AI images; it is a function of how much a tool lets the recipe vary and whether it strips the recipe before the file leaves the machine. The remedy is equally concrete: remove the identifying metadata by default, completely, and tell people it was ever there.

The name was never in the file. It did not need to be.

Data and code availability

The detector, synthetic fixtures, aggregate result tables, and an opaque-indexed corpus that reproduces the identification statistics will be deposited in an archival repository with a DOI accompanying this preprint. To protect the individuals in the corpus, the working data (real handles, extracted metadata, and the resolution chains behind Section 11) is withheld; aggregate statistics are reported in full. Contact: daniel@orygn.tech.

References

- [1] P. Eckersley. *How Unique Is Your Web Browser?* Privacy Enhancing Technologies (PETS 2010), LNCS vol. 6205, pp. 1–18. Springer, 2010. https://doi.org/10.1007/978-3-642-14527-8_1
- [2] A. Caliskan-Islam, R. Harang, A. Liu, A. Narayanan, C. Voss, F. Yamaguchi, and R. Greenstadt. *De-anonymizing Programmers via Code Stylometry*. 24th USENIX Security Symposium, pp. 255–270, 2015.
- [3] A. Narayanan and V. Shmatikov. *Robust De-anonymization of Large Sparse Datasets*. IEEE Symposium on Security and Privacy, pp. 111–125, 2008.
- [4] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau. *DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models*. Proceedings of ACL 2023. ACL Anthology 2023.acl-long.51.
- [5] Coalition for Content Provenance and Authenticity (C2PA). *Content Credentials: Technical Specification*. <https://c2pa.org>
- [6] UpGuard Research. *Detecting Generative AI Data Leaks from ComfyUI*. 2026. <https://www.upguard.com/blog/detecting-generative-ai-data-leaks-from-comfyui>
- [7] M.-T. De Rosa Palmini, L. Wagner, and E. Cetinic. *Civiverse: A Dataset for Analyzing User Engagement with Open-Source Text-to-Image Models*. arXiv:2408.15261, 2024. <https://arxiv.org/abs/2408.15261>
- [8] C. Kim, K. Min, M. Patel, S. Cheng, and Y. Yang. *WOUAF: Weight Modulation for User Attribution and Fingerprinting in Text-to-Image Diffusion Models*. CVPR 2024. <https://arxiv.org/abs/2306.04744>
- [9] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon. *The Stable Signature: Rooting Watermarks in Latent Diffusion Models*. ICCV 2023, pp. 22466–22477.
- [10] Civitai. *Terms of Service, Privacy Policy, and Content Policy*. <https://civitai.com/content/tos>; <https://civitai.com/content/privacy> (accessed July 2026).