NAMESAKES: Probing Identity Memorization in Text-to-Image Models

Anonymous ACL submission

Abstract

Text-to-image (T2I) models generate realistic likenesses of some individuals when prompted with their names, raising privacy concerns. However, distinguishing whether a generated face is memorized or fabricated currently requires ground-truth photos, access to training data, or white-box access to model internals, limiting applicability. We introduce a fully black-box behavioral probe that distinguishes between these regimes while requiring no reference photos or prior knowledge of training data. To benchmark this task, we present the NAMESAKES dataset of over one thousand names and faces of public figures spanning a wide range of fame levels, along with perturbed, less famous names. Experiments on state-of-the-art T2I models show that our probe substantially predicts identity memorization and separates memorized from unrecognized names, with further insights into differences across model families.

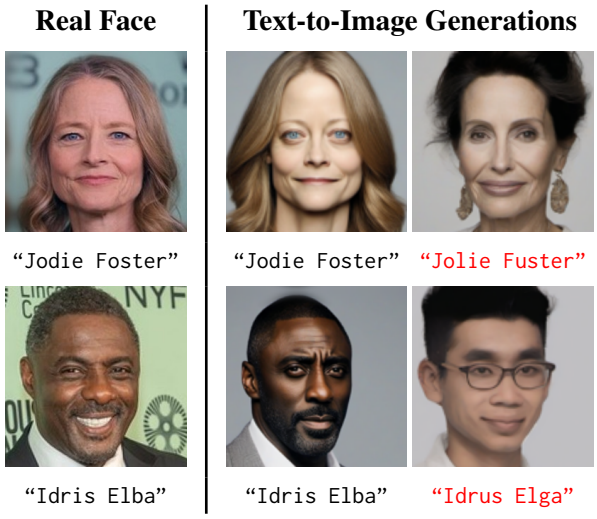


Figure 1: T2I models memorize the faces of some individuals (left) and synthesize their likenesses when prompted with their names (middle). However, when prompted with unfamiliar names (right, red) these models fabricate plausible faces. Our probing method distinguishes between these cases without access to training data or ground-truth photos.

1 Introduction

Text-to-image (T2I) models can now generate nearly photorealistic images from text prompts. What happens when a prompt contains a personal name? As seen in Figure 1, models often reproduce celebrity faces accurately, but for unfamiliar names, they often fabricate plausible faces from demographic cues. This difference reflects two distinct regimes. *Identity memorization* refers to cases where a model reproduces a specific identity learned during training (Gu et al., 2023; Hintersdorf et al., 2024b; Ma et al., 2025). We term the converse case *identity fabrication*, where the model generates a plausible but non-grounded face based on demographic or semantic priors inferred from the name, related to observations of T2I demographic stereotyping (Bianchi et al., 2023; Luccioni et al., 2023).

Distinguishing these regimes in vision-language settings matters for privacy, model auditing, and evaluating unlearning methods. In particular, an individual may wish to know whether a model can generate their likeness, and auditors may need to assess memorization at scale for purposes such as regulatory compliance. Yet existing approaches require either (a) ground-truth (GT) photos of candidates, (b) access to training data, or (c) white-box access to model weights. These requirements are often unmet: State-of-the-art (SOTA) T2I models' training data is not fully disclosed, their architectures may be closed or non-standard, a comprehensive gallery of GT photos does not exist, and a user may be unwilling to upload their own photos.

In this work, we introduce a fully black-box behavioral probe designed to distinguish between memorized and fabricated generations for personal names in T2I models. This is structured without architecture-specific assumptions, and does not re-

Shakira	George Clooney	Shah Rukh Khan	Hasan Minhaj	Donna Air	Remo Ruffini
vs. Shakiba	*vs. Georg Clowney*	*vs. Shaha Rush Kōan*	*vs. Hassan Minhai*	*vs. Danna Airy*	*vs. Remco Ruffin*
log-views: 11.5	log-views: 10.4	log-views: 9.0	log-views: 7.0	log-views: 5.3	log-views: 2.3

Figure 2: Samples from the 1,269 items in NAMESAKES. Each item consists of a public figure's name and ground-truth face (sourced from open data on Wikipedia), and fame as measured by pageview counts (log-scaled). Each real name (first line) is accompanied by a perturbed name (second line, after "vs.") designed to orthographically resemble it. Figures are chosen to span a spectrum of fame—from the average point of view of English Wikipedia users—from highly well-known (left) to relatively obscure (right).

quire access to any GT photos, model internals, or prior information about training data. In order to benchmark this method, we introduce the NAMESAKES dataset of over one thousand paired names and faces of public figures from open Wikipedia data spanning a full spectrum of fame levels, along with *perturbed* names. We show that this benchmark effectively enables measuring our probe's effectiveness across names and models, with additional insights about the span of names recognized and identities fabricated by popular SOTA T2I models.

We will release our code and data to enable future work on privacy and interpretability of T2I models. Our release will use non-commercial licensing, adhering to relicensing and attribution requirements, as well as stipulating ethical use requirements.

2 NAMESAKES

We present NAMESAKES, a dataset of 1,269 names and paired multimodal data extracted from Wikipedia designed to serve as a benchmark for identity probing in vision-language models. We present core details here with further details provided in the appendix.

Dataset Contents. As shown in Figure 2, each entry in NAMESAKES consists of: (1) the personal name of a public figure (from their Wikipedia entry), (2) a GT photo portraying their face, (3) an estimate of the individual's fame based on pageviews, and (4) a perturbed name (defined below). Names are chosen to span a wide range of fame levels, from extremely famous individuals to relatively obscure figures (still deemed public with respect to Wikipedia's inclusion criteria). Additional metadata is provided with details such as image licensing and attribution. Limitations regarding global coverage and diversity are discussed in the limitations section.

Dataset Construction. Data is sourced from English Wikipedia entries. About 190K entries have structured person information; among these, pageview counts v are approximately log-normally distributed, making $f := \log(v) \sim \mathcal{N}(\cdot, \cdot)$ a natural scale. We bin f and use stratified uniform sampling over f (binned), filtering for entries with freely-licensed photos. This yields 1,269 entries including names of very famous individuals likely to be memorized by models, as well as obscure names likely to produce fabricated faces. The resulting distribution of fame levels is shown in Figure 3.

We validate that fame is a meaningful proxy for memorization likelihood in Figure 4, which compares f with reference similarity s_{gt} (Section 3.2). Pearson correlation values are moderate ($r = 0.36$ to 0.53 depending on model; all $p \ll 0.001$), confirming that more famous individuals are more likely to have their identities memorized. The stratified sampling in NAMESAKES is thus designed to ensure sufficient coverage across this spectrum. Note that fame is only used to guide dataset construction; our probing method (Section 3) operates solely on model outputs without using external metadata.

Name Perturbation. For each name in NAMESAKES, we also construct a *perturbed name*: a plausible but fictitious name that orthographically resembles the original. As seen in Figure 1, paired real and perturbed names may differ minimally on the textual level while producing significantly different generations, as the latter is not grounded in a potentially memorized identity. Perturbed names are constructed by perturbing each name compo-

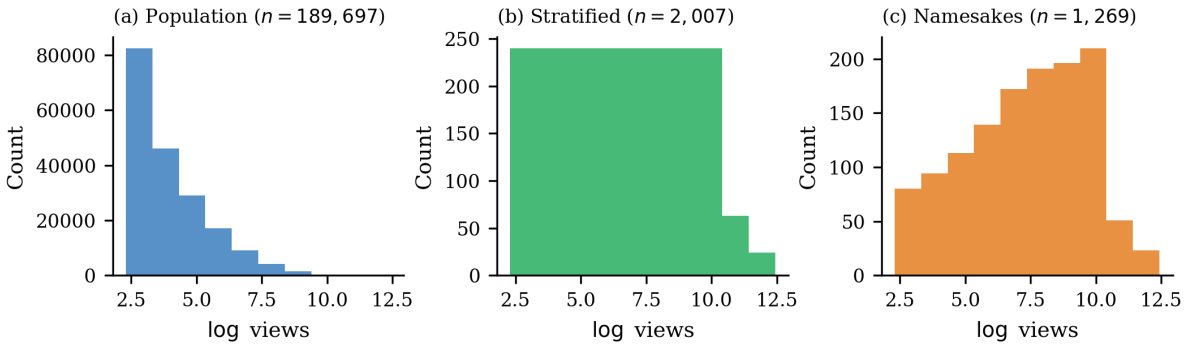


Figure 3: Distribution of fame levels (log-pageviews) in multiple stages of constructing NAMESAKES: in the initial population of entries with infoboxes (left), after stratified sampling (center), and after filtering for pages with freely-licensed infobox images (right). Overall, the final distribution covers fame levels more evenly than the initial skewed distribution. Histograms above use ten evenly-spaced bins.

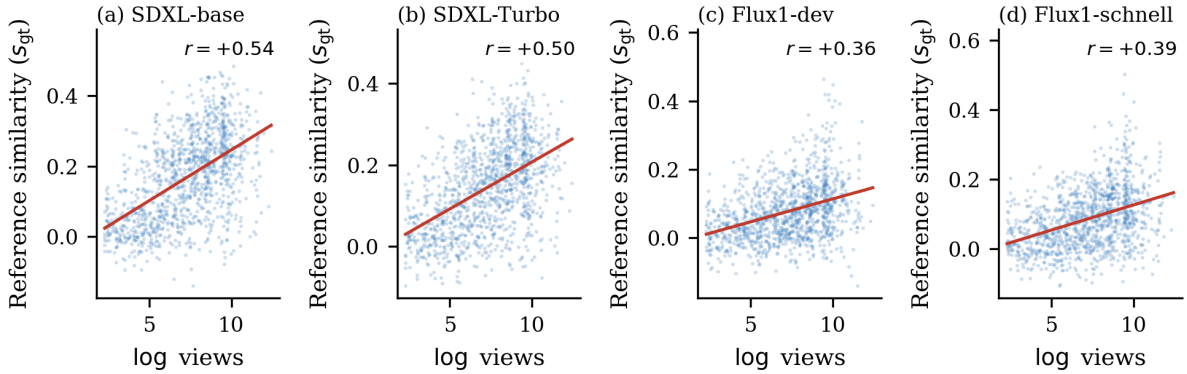


Figure 4: Plots of fame (log-pageviews) and reference similarity s_{gt} for various T2I models under consideration, with Pearson correlation (r) values and best-fit lines. All settings show significant ($p \ll 0.001$) correlation, indicating that identities of more famous individuals are generally more likely to be memorized, justifying our stratified sampling in constructing NAMESAKES.

nent while preserving its first letter (e.g., *Jodie Foster* → *Jolie Fuster*), by searching a large pool of names from Wikipedia for distinct names with minimum Levenshtein edit distance from the original. These new names seldom collide with existing celebrity names (see appendix); as such, they are unlikely to refer to real memorized identities, enabling our separability analysis (Section 4).

As a consistency check, we compare perturbed names against 100 randomly constructed names (formed by independently shuffling first and last names from the same Wikipedia pool) on SDXL-Base using the probing scores defined in Section 3.3. Perturbed names produce more dispersed generations (δ: 0.58 vs. 0.55, p=0.008) and less similar centroids (s_{cen}: 0.54 vs. 0.56, p=0.046), supporting their use representing the unmemorized regime.

Train-Test Split. NAMESAKES is designed to be used with cross-validation (CV). Our evaluation involves fitting lightweight predictive models (an OLS linear regression) to predict reference similarity s_{gt} from probe scores, and a logistic regression classifier for real-vs.-perturbed name separability (Section 4)—separately for each T2I model, since models differ substantially in their memorization behavior and generation diversity (see Table 2). All metrics are calculated on held-out data folds to avoid data leakage: in particular, centroid similarity s_{cen} for a test item is computed using only train-fold centroids (Appendix C), ensuring that the test name's own generations do not influence its score. Cross-validation also mitigates potential issues of demographic imbalance and sensitivity to noise that could arise from a single fixed test split given our dataset size.

Demographic Breakdown. As our dataset inherits subject-matter bias from the English Wikipedia, we report the demographic makeup of NAMESAKES for full transparency. We use Gemini 3.1

Pro (Google, 2026) with web search grounding to annotate coarse gender and ethnic categories based on online knowledge of the public figure (for internal use only; not released), and validate accuracy on a 100-item subset (99% correct).

As shown in Table 1, NAMESAKES is majority White and male, with moderate gender balance (though lacking in non-binary representation) and larger imbalance of ethnic categories. This highlights the importance of careful analysis of demographic equity in methods using NAMESAKES, and the potential for expansion with data sources that could enhance diverse representation.

Gender	Count
Male	757
Female	511
Non-binary	1

Ethnicity	Count
White	938
Asian	163
Black	73
Hispanic/Latino	45
Multiracial	47
Indigenous	3

Table 1: Estimated Demographic Breakdown of NAMESAKES (Gender and Ethnicity)

3 Method

Given a T2I model and name, our reference-free probing method aims to predict whether the model has memorized the name's associated identity, i.e., whether generations will resemble the individual's likeness. We formalize this by defining a ground-truth measure of memorization (reference similarity) and then designing behavioral probes that predict it without access to any reference photos. We emphasize that reference photos are not used in the probe itself; reference similarity is only used in our experiments for a post-hoc evaluation of probe effectiveness.

3.1 Preliminaries

To probe for knowledge of identities, our method requires a similarity metric: given facial images I_i and I_j, the similarity score should be high when the same person is depicted in both images and low otherwise. We implement this using a facial recognition encoder which assigns embeddings to these images and calculates similarity via their cosine similarity. This serves as a proxy for identity similarity, while inheriting the limitations of current facial recognition methods as discussed in Section 6.

Given a name with GT reference image G and a given T2I model we generate k images using a fixed textual prompt template and input noise seeds. The respective L2-normalized face embeddings are denoted by e_{gt} and $e_1, \cdots, e_k$ (generations). We also denote the centroid of generation embeddings as $\bar{e} := \frac{1}{k}\sum_i e_i$, L2-normalized to $\hat{e} := \frac{1}{\|\bar{e}\|}\bar{e}$.

3.2 Reference Similarity (s_{gt})

We define reference similarity as the mean similarity between the ground-truth (GT) and generated facial images:

$$s_{\text{gt}} := \frac{1}{k}\sum_{i=1}^{k} e_i \cdot e_{\text{gt}} = \bar{e} \cdot e_{\text{gt}}. \tag{1}$$

Reference similarity directly operationalizes identity memorization: when s_{gt} is high, the model reproduces the real person's likeness (identity memorization); when low, it generates a generic face unrelated to the individual (identity fabrication). Our goal is to predict this quantity *without access to the ground-truth appearance*, enabling black-box memorization detection. Accordingly, the subsequent probing scores do not have access to e_{gt}.

3.3 Probing Scores

We define two complementary scores, δ and s_{cen}, designed to predict reference similarity, and thereby identity memorization, without access to ground-truth photos, illustrated in Figure 5. These are both entirely black-box and reference-free; in addition, δ only uses a single name as input, while s_{cen} also requires comparisons to generations for other names in NAMESAKES. We will later (Section 4) show that δ can be used alone with substantial performance; alternatively, s_{cen} may be added for an additional boost.

Dispersion (δ). This quantity measures inter-generation consistency, i.e., to what degree multiple generations of a name depict the same face. This is defined as the mean squared distance

$$\delta := \frac{1}{k}\sum_{i=1}^{k} \|e_i - \bar{e}\|^2 = \text{tr}(\Sigma), \tag{2}$$

where Σ is the sample covariance matrix of $(e_1, \cdots, e_k)$. Intuitively, memorized identities should result in more similar generated faces, negatively correlating δ with s_{gt}.

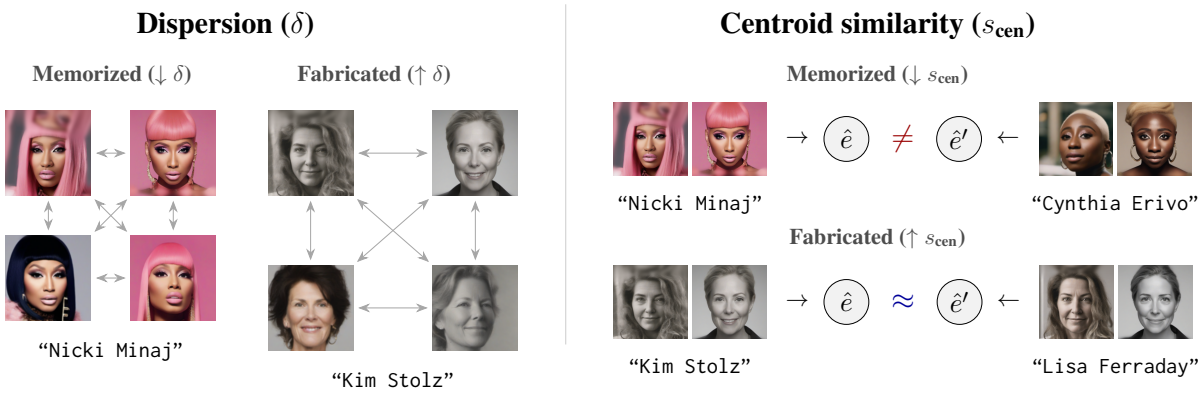


Figure 5: The two probes comprising our method, shown on memorized vs. fabricated names (SDXL-Base). **Left — Dispersion (δ):** Four generations of a memorized name form a tight cluster (low δ), while generations of a fabricated name spread apart (high δ). Lines show pairwise comparisons; spacing reflects embedding distance. This comparison is *within* a single name's generations. **Right — Centroid similarity (s_{cen}):** Each name's generations are summarized by an embedding centroid $\hat{e}$. Memorized names have distinctive centroids ($\hat{e} \neq \hat{e}'$), while fabricated names with similar demographic associations converge to nearly identical centroids ($\hat{e} \approx \hat{e}'$). This comparison is *between* names.

Centroid similarity (s_{cen}). Let $\hat{\mathcal{N}}$ denote the set of all L2-normalized centroids of other names in NAMESAKES, and define

$$s_{\text{cen}} := \max_{\hat{e}' \in \hat{\mathcal{N}}} \hat{e} \cdot \hat{e}'. \tag{3}$$

Intuitively, unfamiliar names are expected to produce generic outputs that closely resemble those of other names, while a memorized identity should only be produced by its associated name. Hence, s_{cen} is expected to negatively correlate with s_{gt}, consistent with findings that T2I models converge to default images for unknown prompts (Simonen et al., 2026).

4 Experiments

Experimental Setup We evaluate four T2I models covering recent text-conditioned denoising diffusion models, in both base and distilled variants: Stable Diffusion XL-base (SDXL) (Podell et al., 2024), SDXL-Turbo (Sauer et al., 2024), Flux1-Dev, and Flux1-Schnell (Labs, 2024). We generate $k = 4$ images per name and model for all real and perturbed names in the NAMESAKES benchmark of $n = 1,269$ items. We ablate this choice and the name pool size used to calculate s_{cen} in the appendix. Generated faces undergo standard preprocessing including facial alignment, and are then embedded with ArcFace (Deng et al., 2019). All metrics are calculated on held-out data using 5-fold CV. Additional ablations on the choice of face embedding model are shown in the appendix.

Predicting Reference Similarity We test the effectiveness of our probe scores to predict reference similarity via ordinary least squares (OLS) linear regression. Results for all models are shown in Table 2 (left columns). The two-probe model achieves R^2 values above Cohen's large-effect benchmark (Cohen, 2013), with 0.58 for SDXL-Base and between 0.33–0.44 for the remaining models, indicating that our reference-free probes capture a substantial portion of the memorization signal that would otherwise require ground-truth photos. We note that reference similarity is itself a noisy proxy for true identity memorization (limited by face embedding accuracy and generation variance), which places an inherent ceiling on achievable R^2.

The variation in probe effectiveness between models is consistent with known differences between them: distilled models (SDXL-Turbo, Flux1-Schnell) and model families optimized for aesthetic quality (e.g., Flux) are known to have reduced diversity in generations (Gandikota and Bau, 2026; Adamkiewicz et al., 2026), potentially weakening the effectiveness of our dispersion probe. In addition, Flux models have been observed to underperform on generating celebrities, leading to community speculation on whether they were trained with methods such as image recaptioning or targeted redaction of personal names (GitHub, 2024; Reddit, 2024).

We also see that both probes contribute to overall performance, although for SDXL-Base dispersion

Model	Predicting Reference Similarity (R^2)			Real vs. Perturbed Names Separability			
	δ only	s_{cen} only	Both	AUC_{all}	Acc_{all}	AUC_{high}	Acc_{high}
SDXL-Base	0.547 $\pm$0.031	0.128 $\pm$0.030	0.581 $\pm$0.040	0.859 $\pm$0.015	0.791 $\pm$0.014	0.947 $\pm$0.011	0.902 $\pm$0.014
SDXL-Turbo	0.288 $\pm$0.050	0.171 $\pm$0.052	0.438 $\pm$0.044	0.772 $\pm$0.036	0.723 $\pm$0.029	0.835 $\pm$0.024	0.777 $\pm$0.019
Flux1-Dev	0.218 $\pm$0.054	0.150 $\pm$0.007	0.349 $\pm$0.060	0.773 $\pm$0.015	0.735 $\pm$0.018	0.879 $\pm$0.013	0.828 $\pm$0.013
Flux1-Schnell	0.137 $\pm$0.078	0.187 $\pm$0.036	0.325 $\pm$0.056	0.785 $\pm$0.016	0.748 $\pm$0.011	0.844 $\pm$0.018	0.792 $\pm$0.015

Table 2: Per-model probe results on $n = 1{,}269$ items in NAMESAKES (excluding failed face alignment; see appendix). Left: R^2 from OLS linear regression predicting reference similarity s_{gt}, using each probe alone or both. Right: Real vs. perturbed separability using logistic regression on both probe features (*all*: all names; *high*: high-fame names with log-pageviews above median). Full fame stratification in appendix. All results use cross-validation (mean $\pm$ std across folds).

dominates. This is supported by observing that δ and s_{cen} have Pearson correlation 0.26 for SDXL-Base and are nearly orthogonal (between -0.02 and -0.06) for other models. We hypothesize that in cases where generations are less diverse, δ provides much weaker signal, increasing the relative utility of s_{cen}.

Real-vs.-Perturbed Separability The more practically relevant question is whether our probe can categorically separate memorized from unrecognized names: the key capability for privacy auditing and model assessment. To test this, we apply our probe to the real and perturbed names in NAMESAKES (Section 2), fitting a logistic regression classifier to real vs. perturbed names in each CV train split and evaluating AUC and binary accuracy on held-out data. Note that the two classes are balanced by design, as each real name has a corresponding perturbed name. Results are shown in Table 2 (right column), finding that the probe generally separates real from perturbed names across all models. While SDXL-Base again shows the strongest performance ($\text{AUC} = 0.86, \text{Acc} = 0.79$), separability remains moderate ($\text{AUC} > 0.77, \text{Acc} > 0.72$) for models where continuous prediction of s_{gt} is weaker, suggesting that this probe also serves as a binary predictor of identity. This aggregate score reflects the fame distribution of NAMESAKES *by design*: low-fame names are rarely memorized, so a well-calibrated probe should treat them as indistinguishable from their perturbations, the intended behavior, as the probe targets memorization rather than name familiarity. Restricting to high-fame individuals (the regime where memorization is plausible) confirms this, yielding substantially stronger results (e.g., AUC of 0.95 and $>90\%$ binary accuracy for SDXL-Base; see Table 2).

Qualitative results Figure 6 shows qualitative examples of generations (SDXL-Base) conditioned on names, illustrating different regimes for memorized and unknown names–in the former case, generations are mutually consistent and resemble the GT face, while in the latter case, they reflect more dispersed default images unrelated to the GT face, reflected in s_{gt} and probe values. Similar results for all models tested, with more names, are shown in the appendix.

In Figure 7, we compare results between T2I models for a celebrity name, illustrating the patterns seen in our quantitative results. SDXL-Based models reproduce the individual's identity more faithfully ($s_{\text{gt}} > 0.37$), while Flux-based models fabricate generic faces that do not match the ground-truth face ($s_{\text{gt}} < 0.07$). This matches our findings that SDXL models (particularly SDXL-Base) exhibit more identity memorization. We also see that probe values inversely correlate with s_{gt}, as expected.

In the appendix, we also introduce an interpretability technique for visualizing overall visual associations with unfamiliar names.

Human Evaluation and Demographic Analysis To validate our automatic metrics (reference similarity s_{gt} and probes) against human perception, we conduct a face similarity study on Amazon Mechanical Turk. We evaluate 100 celebrities from NAMESAKES, stratified to be demographically balanced as well as covering the spectrum of fame levels, asking annotators to indicate the similarity between the GT photo and SDXL-Base generation for a given name on a 5-point Mean Opinion Score (MOS) scale. We collect 9 annotations per item (after filtering for control tests). Further annotation task details are provided in the appendix.

In general, we find a substantial correlation be-

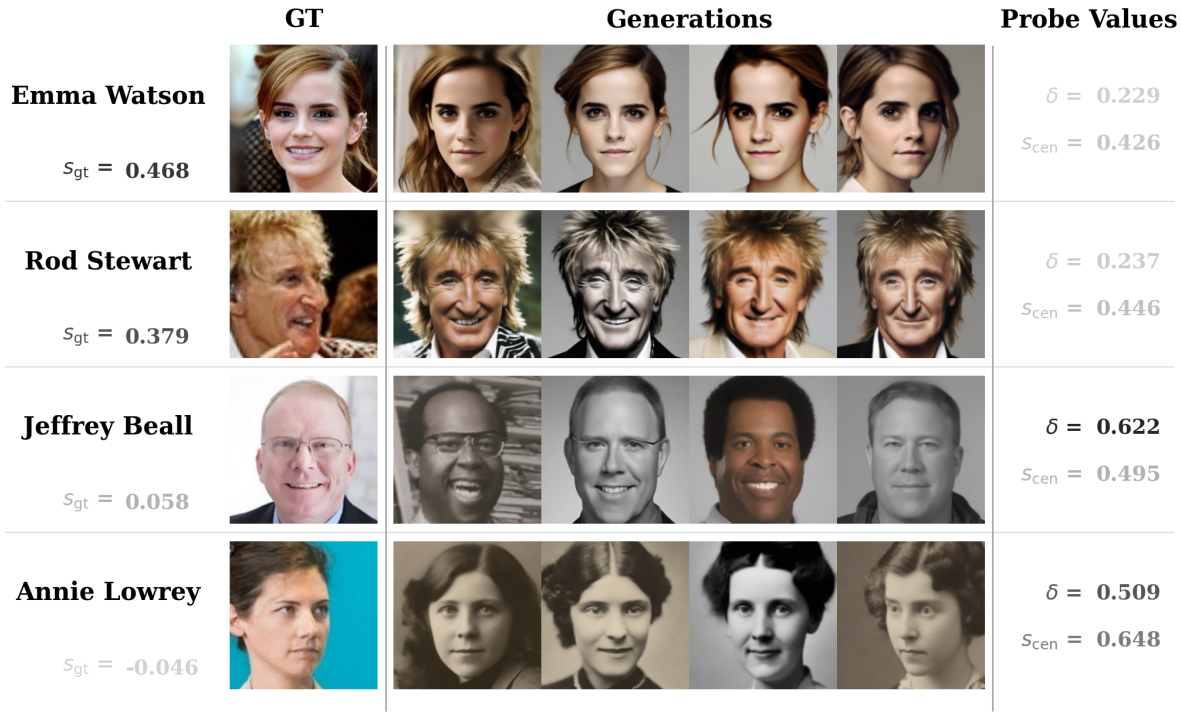


Figure 6: Qualitative examples spanning the memorization spectrum for SDXL-Base. Each row shows a name's Wikipedia GT photo and four generated images, followed by probe values δ (dispersion) and s_{cen} (centroid similarity); the prediction target s_{gt} (reference similarity to GT) appears in the name label at left. All values are shaded darker with greater magnitude. Rows are ordered from high to low s_{gt}: top rows show memorized identities whose generations resemble the GT and each other, while bottom rows show fabricated identities. Additional examples across all T2I models are shown in the appendix.

tween s_{gt} and human similarity judgments (e.g., Pearson $r = 0.56$, considered a large effect by Cohen's conventions (Cohen, 2013)), with large differences between ethnic (Asian $r = 0.36$; Black $r = 0.44$; Hispanic/Latino $r = 0.72$; White $r = 0.74$) and gender (male $r = 0.65$; female $r = 0.47$) categories. The correlation between our OLS scores (Section 4) and human similarity judgments shows the same general pattern shown in Figure 8 (i.e., lower on Black, Asian, and female faces). While our automatic metrics still have a moderate correlation with similarity judgments across categories, they inherit unequal performance from the facial recognition models used for similarity calculations (Buolamwini and Gebru, 2018; Yucer et al., 2024). This highlights the importance of mitigating bias in these models for downstream methods such as ours. Users should account for this context in interpreting results of our system, as we discuss in our ethical considerations section.

5 Related Work

Our work is most related to studies on identity inference in vision-language models: methods identifying leakage of data related to an individual's identity, where we focus on data relating to names and facial imagery. Prior works have explored this in the context of discriminative CLIP models (Hintersdorf et al., 2024a; Li et al., 2025) and image generation models (Webster et al., 2021; Vora et al., 2025), but these works require either GT photos, access to training data, or white-box access to gradients. By contrast, our probe targeting generative T2I models is fully black-box and reference-free.

A related line of work studies data memorization in image generation models such as diffusion models. Works on inference attacks (Hu and Pang, 2023; Duan et al., 2023; Dubiński et al., 2024) study whether a given image was seen during training, and works on memorization (Carlini et al., 2023; Somepalli et al., 2023a,b) identify memorized images, but these do not address whether a holistic identity was memorized. Wen et al. (2024) find that memorized prompts steer gener-

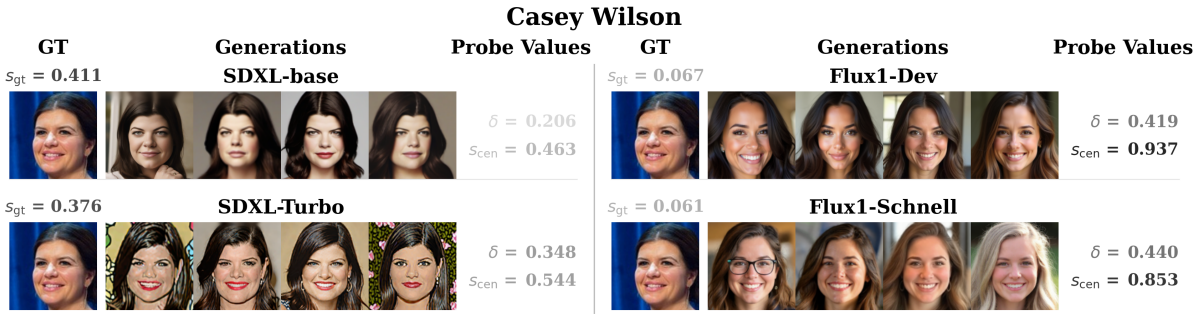


Figure 7: Cross-model comparison for a single celebrity name (Casey Wilson). For each model, we show the GT face from NAMESAKES, generated faces and their reference similarity s_{gt}, and our probe values (dispersion δ and centroid similarity s_{cen}). All values are shaded darker with greater magnitude. SDXL models (left) reproduce the identity more faithfully than Flux-based models (right)—reflected in reference similarity s_{gt} values—and probe values inversely correlate with s_{gt}, as expected.

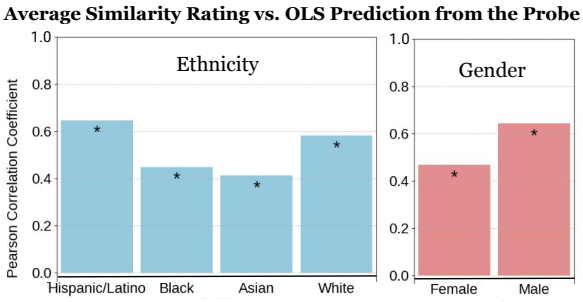


Figure 8: Comparison across single demographic attributes (Gender, Ethnicity) for the association between predicted memorization (OLS predictions on the probing scores) and the average human-rated similarity scores. $*$ indicates significance of the Pearson r, with $p < .05$.

ation more strongly towards seed-independent directions, supporting our dispersion probe. However, their method requires access to intermediate model predictions during denoising and targets *image-level* memorization (reproducing a specific training image), whereas our dispersion probe is fully black-box and targets *identity-level* memorization (learning a coherent identity).

Our work also builds on research using behavioral signals to probe vision-language models for latent knowledge, including via input-level pseudoword probes (Alper and Averbuch-Elor, 2023), embedding geometry (Alper and Averbuch-Elor, 2024), and character-level prompt perturbations (Struppek et al., 2024). Yuan et al. (2023) find that celebrity identities span a text embedding subspace; this supports our geometric intuitions, although our behavioral probe does not depend on input embedding geometry. Luccioni et al. (2023) characterize the average face associated by T2I models to prompts, parallel to our visualization method using blended prototype faces (appendix).

6 Conclusion

In this work, we have introduced a black-box behavioral probe to distinguish between identity memorization and identity fabrication in T2I models applied to personal names. To benchmark performance, we have contributed the NAMESAKES dataset of identities stratified by fame, including names and ground-truth faces as well as perturbed names. Our results on NAMESAKES have shown that overall our probe can effectively diagnose identity memorization, revealing substantial differences between SOTA models in use today. Moreover, as this probe requires neither access to GT images nor model internals, it may generalize to any future T2I model agnostic of architecture, making it suitable for future privacy auditing.

We foresee various additional applications and extensions of our methodology: Future work could evaluate the effectiveness of identity-unlearning methods and whether memorization is unequally distributed among demographic groups. Another promising direction is mechanistic probing, for example, by finding interpretable circuits which encode identity recognition.

Limitations

A central limitation of our current methodology is the reliance on data and models which may underrepresent diverse faces and identities. Our dataset reflects the English Wikipedia's demographic skew, with names and faces over-representing demographic groups such as English-speaking males.

Future work could increase representation by collecting data from more diverse open sources, expanding beyond our limited 1,269 sample size, as well as grounding facial similarities in human annotations. In addition, facial similarity calculations use neural face embeddings, which have been documented to exhibit gender and racial bias (Buolamwini and Gebru, 2018; Yucer et al., 2024). As our framework is modular with respect to the face embedding model, the use of future, more fair models should transfer to our probe.

We have observed our probe to perform less strongly on distilled models, likely related to their known lack of diversity in generations. This is mitigated by an increased contribution of s_{cen} in our combined probe; as seen in Table 2, the combined probe still achieves meaningful performance (R^2=0.33−0.44 and AUC>0.77). We also acknowledge that while our probe significantly predicts reference similarity (our continuous measure of identity memorization), this statistical relationship is not guaranteed to hold for a single sample, and our method should not be used as an exclusive diagnostic without human validation in performance-critical settings. Finally, we do not account for public figures with identical names (e.g., "Michael Jordan" refers to both a famous basketball player and a well-known machine learning researcher), name changes, or those whose appearance has changed dramatically over time.

Ethical Considerations

The NAMESAKES uses only openly-licensed images from Wikipedia, adhering to their respective licensing and attribution terms. While the individuals depicted did not explicitly consent to inclusion, we note that (1) all subjects are public figures according to Wikipedia's notability criteria, (2) the images were published under licenses permitting reuse, (3) the benchmark has protective intent for privacy auditing, such as flagging memorized identities in T2I models to inform unlearning or regulatory compliance, and (4) we include a datasheet (Gebru et al., 2021) stipulating ethical use restrictions. We will also honor opt-out requests for removal from individuals included in the benchmark. Computer vision datasets embed disciplinary values and politics through curation choices (Stevens and Keyes, 2021), like over-representing certain demographics; we explicitly report NAMESAKES's skew (Table 1: majority White/male) as inherited from English Wikipedia's population. During benchmarking, generated images are used solely for calculating probe statistics and are not publicly released.

The facial recognition and T2I generation systems used in our pipeline are dual-use technologies. While we use them to evaluate identity recognition and privacy in existing models, the same existing tools can be used for surveillance, propagation of fake information, and stereotype perpetuation. Image databases often constructed with skewed race/gender categories, often amplifying biases in downstream facial analysis (Scheuerman et al., 2021) – issues our probes inherit via embedding models (Buolamwini and Gebru, 2018; Yucer et al., 2024). Use of these technologies must follow responsible guidelines for human-centered technologies, and we only condone use of our dataset and probe in this framework. In particular, as our probe may detect specific identities memorized by models, this information could in principle be maliciously exploited (e.g., by generating deepfake images of memorized identities); we recommend using this probe for audit purposes only, while highlighting its importance in flagging such privacy concerns. To address gaps in data traceability, our release also includes metadata on image sourcing, fame stratification, and demographic estimates (via Gemini 3.1 Pro, validated on 100 items), enabling data subjects or auditors to trace origins and request exclusions.

Our study involved manual annotation of image similarity using a crowdsourcing platform (Amazon Mechanical Turk). The task presented only images and annotation instructions; no personally identifiable information was collected or retained about the crowdworkers. Annotator identifiers were used only for task administration and quality control and were not included in the research dataset. We did not use any annotator attributes (e.g. demographics or individual-level metadata) in our analysis, and no attempts were made to re-identify individuals. Because the research uses de-identified annotation data and does not involve analysis of identifiable private information about living individuals, it qualifies as exempt from human subjects research oversight under our institutional policies.

Our benchmark and probe may inherit demographic biases from the data and models used. In particular, both facial recognition and T2I models exhibit demographic bias (Buolamwini and Gebru, 2018; Yucer et al., 2024; Luccioni et al., 2023),

which may affect probe reliability unevenly. We do not recommend using our system as a conclusive determination of memorization for a given name and model in isolation, and encourage users to report demographic stratification in their results.

References

Krzysztof Adamkiewicz, Brian Moser, Stanislav Frolov, Tobias Christian Nauen, Federico Raue, and Andreas Dengel. 2026. When pretty isn't useful: Investigating why modern text-to-image models fail as reliable training data generators. *arXiv preprint arXiv:2602.19946*.

Morris Alper and Hadar Averbuch-Elor. 2023. Kiki or bouba? sound symbolism in vision-and-language models. *Advances in Neural Information Processing Systems*, 36:78347–78359.

Morris Alper and Hadar Averbuch-Elor. 2024. Emergent visual-semantic hierarchies in image-text representations. In *European Conference on Computer Vision*, pages 220–238. Springer.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1493–1504.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pages 5253–5270.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. 2023. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pages 8717–8730. PMLR.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. 2024. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4860–4869.

Rohit Gandikota and David Bau. 2026. Distilling diversity and control in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1304–1313.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

GitHub. 2024. Flux doesn't understand specifics! https://github.com/lllyasviel/stable-diffusion-webui-forge/discussions/1494.

Google. 2026. Gemini 3.1 Pro: A smarter model for your most complex tasks.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. 2023. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*.

Dominik Hintersdorf, Lukas Struppek, Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. 2024a. Does clip know my face? *Journal of Artificial Intelligence Research*, 80:1033–1062.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. 2024b. Finding nemo: Localizing neurons responsible for memorization in diffusion models. *Advances in Neural Information Processing Systems*, 37:88236–88278.

Hailong Hu and Jun Pang. 2023. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*.

Black Forest Labs. 2024. Flux.1 [dev]. https://github.com/black-forest-labs/flux.

Songze Li, Ruoxi Cheng, and Xiaojun Jia. 2025. Tuni: A textual unimodal detector for identity inference in clip models. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 1–13.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhe Ma, Qingming Li, Xuhong Zhang, Tianyu Du, Ruixiao Lin, Zonghui Wang, Shouling Ji, and Wenzhi Chen. 2025. An inversion-based measure of memorization for diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16959–16969.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Reddit. 2024. What happened here, and why? (flux-dev). https://www.reddit.com/r/StableDiffusion/comments/1ejuuzm/what_happened_here_and_why_fluxdev/.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2024. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer.

Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. 2021. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–37.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Hannu Simonen, Atte Kiviniemi, Hannah Johnston, Helena Barranha, and Jonas Oppenlaender. 2026. An exploration of default images in text-to-image generation. In *ACM CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. ACM.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023a. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6048–6058.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023b. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803.

Nikki Stevens and Os Keyes. 2021. Seeing infrastructure: Race, facial recognition and the politics of data. *Cultural Studies*, 35(4-5):833–853.

Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2024. Exploiting cultural biases via homoglyphs intext-to-image synthesis (abstract reprint). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8486–8486.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14.

Jayneel Vora, Nader Bouacida, Aditya Krishnan, Prabhu Shankar, and Prasant Mohapatra. 2025. Identity-focused inference and extraction attacks on diffusion models. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 1522–1530.

Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. 2021. This person (probably) exists. identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018*.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. 2024. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*.

Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. 2023. Inserting anybody in diffusion models via celeb basis. *Advances in Neural Information Processing Systems*, 36:72958–72982.

Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby Breckon. 2024. Racial bias within face recognition: A survey. *ACM Computing Surveys*, 57(4):1–39.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and 1 others. 2021. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502.

A Dataset Construction Details

A.1 Data Source and Filtering

Our data is sourced from the official Wikimedia data dumps, using data from the English Wikipedia. We use archived page contents from January 2026, and pageview counts from January 1, 2026. We select articles transcluding `Template:Infobox_person`, using only pages with at least 10 views, yielding 189,697 pages. We

use string heuristics to omit pages with malformed titles such as those containing punctuation, multi-person entries, and disambiguation suffixes (e.g., *John Smith (actor)*)). After stratified sampling over fame levels (Section 2) to obtain about 2K pages, we further filter by requiring a freely-licensed infobox image (CC*, Public Domain, or GFDL) with a successfully detected and aligned face, yielding the final 1,269 entries in NAMESAKES. Images are resized to have longest side length 256px and saved in PNG format.

A.2 Construction of Perturbed Names

To construct perturbed names, we first create a pool of unique name components (such as first names and surnames) using the ∼190K names associated with selected Wikipedia infobox entries (Section A.1), split with word tokenization. Subsequently, in order to create a perturbed name for an existing name in NAMESAKES, we replace each name component with a candidate from the pool with minimum nonzero Levenshtein edit distance, subject to matching first letter. In cases of ties, one is selected at random. Random examples from NAMESAKES include the following: (shown in format *real* → *perturbed*)

- *Greg Abel* → *Gren Adel*
- *Katie Holmes* → *Kathie Holes*
- *J. Paul Gerry* → *Jr. Poul Gatty*
- *Diana Ross* → *Dyana Rossi*
- *Paul Whitehouse* → *Paulo Whithouse*
- *Shreyas Talpade* → *Shreyans Talmage*
- *James Nesbitt* → *Jayes Nesbit*
- *Whitney Blake* → *Whitley Blaže*
- *David Mickey Evans* → *Davud Mickley Evins*
- *Ali Zafar* → *Adi Zafer*

None of the perturbed names in NAMESAKES collide with real names from the ∼190K-name list of public figure entries with infoboxes and sufficient pageviews in Wikipedia. While this does not preclude that a perturbed name could match the name of a celebrity without a Wikipedia infobox, or that it could refer to a real person without a significant presence on Wikipedia, this confirms that these names are unlikely to correspond to public figures whose identities may have been memorized by T2I models trained on web-scale imagery. However, we leave open the possibility that additional identities may have been memorized by these models beyond the coverage of Wikipedia, which we foresee probes like ours detecting.

B Image Generation and Postprocessing

B.1 Computation Details

All models are run on a single NVIDIA RTX A5000 GPU.

B.2 Text-to-Image Settings

All images are generated using the fixed prompt template:

`"a photo of the person {name}"`

where `{name}` is substituted with the personal name to be generated.

All T2I models are run at half or 8-bit precision, using checkpoints (from Hugging Face Hub) and settings listed in Table 3. Generation settings use recommended defaults for compute-restricted settings (e.g., using the minimum number of recommended denoising steps). For SDXL-Base, the optional refiner module omitted.

For each name and model, we generate $k = 4$ images in a single batch, with noise seeds fixed between generations. This value was chosen to balance between robustness and computational limitations; in particular, the larger T2I models tested would not be feasible to run on the entire NAMESAKES dataset within our time and computation budget if k were significantly increased.

B.3 Face Detection and Alignment

All generated images and GT photos are postprocessed via facial detection, alignment, and resizing. We use dlib's HOG-based face detector and landmark detector[1] followed by warping with a similarity transformation into a canonical reference frame. The resized output is a 256×256 px RGB image.

Edge cases where detection fails return a fully black image. This never occurs for GT images in NAMESAKES; for generated images, failure rates are low across models ($< 1.8\%$ of images). Rates are higher for perturbed names ($5.6\% - 6.6\%$ of images for SDXL models; $\leq 1.5\%$ for Flux models)—expected since fictional and obscure names are more likely to produce less face-like images. As

[1] `shape_predictor_68_face_landmarks.dat`

Model	Checkpoint	#Params	Precision	Steps	CFG	Resolution
SDXL-Base	`sd_xl_base_1.0`	3.5B	fp16	20	7.0	1024×1024
SDXL-Turbo	`sd_xl_turbo_1.0_fp16`	3.5B	fp16	4	1.0	512×512
Flux1-Dev	`flux1-dev-fp8`	12B	fp8	20	3.5	1024×1024
Flux1-Schnell	`flux1-schnell`	12B	bf16	1	—	1024×1024

Table 3: Generation settings per model. Flux1-Schnell uses timestep distillation; CFG is not applicable).

black images would bias metric values (e.g., a name where all images are black would have zero dispersion, falsely signalling memorization) we apply a strict exclusion policy, excluding names from all analyses when any generated image fails face alignment. This leaves $n = 1{,}215$–$1{,}224$ names for OLS and $n = 996$–$1{,}224$ names for separability probes (for the latter, we exclude corresponding real and perturbed names if any generation for either of them fails face alignment).

B.4 Face Embeddings

After undergoing alignment, we embed all facial imagery to calculate similarity scores, used for both reference similarity and probe calculations. For our embedding model, we use ArcFace (Deng et al., 2019) with the `w600k_r50.onnx` model provided by InsightFace: a ResNet-50 backbone trained on WebFace600K (a.k.a. WebFace12M) (Zhu et al., 2021). This produces 512-dimensional embeddings which we L2-normalize before use.

B.5 StyleGAN Inversion

For the face blending technique used to illustrate centroid similarity, we encode each of the k generated faces (after alignment) with an E4E model (Tov et al., 2021) into a latent code of shape 18×512. These k latent codes are averaged elementwise and then passed through the frozen StyleGAN2 generator (Karras et al., 2020) to produce the blended face image. We use frozen E4E and StyleGAN2 models that were trained on FFHQ (Karras et al., 2019).

C Experimental Details

Probes using linear regression probes fit an OLS regression model (`numpy.linalg.lstsq`) with an intercept term on the scores being used (δ and s_{cen} when using our full probe). Probes using logistic regression use defaults from scikit-learn (`sklearn.linear_model.LogisticRegression`) and 1000 maximum iterations. These are fit and evaluated separately for each cross-validation fold.

For each cross-validation fold, centroid similarity s_{cen} is computed inductively by only using names from the train split in calculations. In other words, s_{cen} for a test item yields the similarity of the closest train set centroid.

D Computational Cost Analysis

k=4 generations per name requires only modest resources. Per-name probe computation takes seconds for fast models (e.g., SDXL-Turbo) to a few minutes for the largest (Flux1-Dev) on a single NVIDIA RTX A5000 GPU. Image generation is the main bottleneck; probe computation adds negligible overhead.

E Additional Experiments and Results

E.1 Ablations

Face Embedding. To ablate the effect of the chosen ArcFace face embedding model, we compare to results using our methodology using FaceNet (Schroff et al., 2015)–with an Inception-ResNet-V1 backbone pretrained on VGGFace2 (Cao et al., 2018), loaded via the `facenet-pytorch` library. A comparison of results between these embedding models is provided in Table 4. Overall performance is similar, though ArcFace is more performant, particularly for OLS probing with SDXL models, justifying our choice of ArcFace as our primary embedding model.

Number of generations (k). Subsampling existing generations to k=2 yields R^2 of 0.49 vs. 0.58 (k=4) for SDXL-Base and 0.22 vs. 0.35 for Flux1-Dev. Separability (AUC) is more consistent, dropping only 0.01–0.05 across models. k=4 thus provides a meaningful improvement while remaining in our feasible with our computational resources.

Reference pool size for s_{cen}. Restricting the pool of names used to calculate s_{cen} to 500 and 250 names (3 random subsamples each), R^2 drops $\leq$ 0.03 and AUC drops $\leq$ 0.01 across all models, demonstrating robustness to this pool size.

Model	Predicting Reference Similarity (R^2)			Real vs. Perturbed	
	δ only	s_{cen} only	Both	AUC	Acc
ArcFace (primary)					
SDXL-Base	0.547 ±0.031	0.128 ±0.030	0.581 ±0.040	0.859 ±0.015	0.791 ±0.014
SDXL-Turbo	0.288 ±0.050	0.171 ±0.052	0.438 ±0.044	0.772 ±0.036	0.723 ±0.029
Flux1-dev	0.218 ±0.054	0.150 ±0.007	0.349 ±0.060	0.773 ±0.015	0.735 ±0.018
Flux1-schnell	0.137 ±0.078	0.187 ±0.036	0.325 ±0.056	0.785 ±0.016	0.748 ±0.011
FaceNet					
SDXL-Base	0.478 ±0.012	0.045 ±0.032	0.493 ±0.017	0.848 ±0.012	0.771 ±0.015
SDXL-Turbo	0.208 ±0.070	0.066 ±0.021	0.294 ±0.063	0.780 ±0.038	0.733 ±0.033
Flux1-dev	0.275 ±0.059	0.082 ±0.037	0.328 ±0.065	0.777 ±0.014	0.732 ±0.007
Flux1-schnell	0.252 ±0.039	0.068 ±0.022	0.311 ±0.042	0.811 ±0.007	0.734 ±0.005

Table 4: Embedding model ablation: ArcFace vs. FaceNet. All columns as in Table 2. Results are comparable between models, with ArcFace generally outperforming FaceNet, justifying our choice of ArcFace for our main results.

E.2 Fame-Stratified Separability

Table 5 reports real-vs.-perturbed name separability separately for high- and low-fame names, defined as those with fame (log-pageview) scores above or below the median. Calculations use cross-validation following our main results. High-fame names show substantially higher separability across models, expected since both low-fame and perturbed names are unlikely to be memorized. Low-fame names exhibit low separability as expected, since these names are unlikely to have been memorized; the probe correctly assigns them scores comparable to their perturbations. This stratification supports the intended interpretation of our method: the probe identifies memorization where it occurs and reports its absence elsewhere, rather than producing spurious positives for unfamiliar names.

E.3 Additional Qualitative Results

Figures 9 to 12 show qualitative results in the format of Figure 6 for all T2I models under consideration. Comparing across figures illustrates the model differences discussed in Section 4: SDXL models exhibit more identity memorization than Flux models, with SDXL-Base showing the strongest effect.

E.4 Prototype Blending

We introduce an interpretability technique, *prototype blending*, that visualizes overall visual associations with unfamiliar names (Figure 13). For each name, we invert a set of generated faces into the latent space of a pretrained StyleGAN2 (Karras et al., 2020) using an E4E encoder model (Tov et al., 2021) and decode their mean via the StyleGAN2 generator to produce a *prototype face* image illustrating the model's overall association for that name. As seen in the figure, prototypes for unknown names reflect aggregate demographic associations such as gender, ethnicity, and age, providing a visual tool for understanding T2I models' stereotypical associations with names.

F Human Evaluation Details

F.1 Platform and Participants

For our human evaluation, we recruit participants on Amazon Mechanical Turk, without geographic limitations.

Each participant received $0.50 in compensation for a single HIT (approximately 3 minutes in duration), which covered 25 names (one-fourth of the set under consideration). We collected a total of 38 HITs across four such splits, targeting at least 9 responses per name.

Our survey involved collecting image similarity annotations from crowdworkers, without collecting personally identifiable information, sensitive data, or research questions regarding annotators themselves.

F.2 Survey Methodology

Each survey item follows a two-step design: annotators first indicate whether they recognize the person by name (shown alongside a single GT photo), then rate the similarity between the GT photo and a generated image on a 5-point Mean Opinion Score (MOS) scale (1 = Different person, 5 = Same person). Each survey includes a hidden control pair (two real photos of the same person); annotators rating the control below 4 are excluded—in total, excluding two out of 38 HITs. After filtering, we retain 9 valid annotators per item (items were split between multiple HITs).

Figure 9: Additional qualitative examples (SDXL-Base) spanning the memorization spectrum. Layout as in Figure 6. Values are shaded darker with greater magnitude.

Figure 10: Additional qualitative examples (SDXL-Turbo). Same names and layout as Figure 9.

Figure 11: Additional qualitative examples (Flux1-Dev). Same names and layout as Figure 9.

Figure 12: Additional qualitative examples (Flux1-Schnell). Same names and layout as Figure 9.

Model	AUC		Acc	
	High Fame	Low Fame	High Fame	Low Fame
SDXL-Base	0.947 ±0.011	0.760 ±0.032	0.902 ±0.014	0.697 ±0.040
SDXL-Turbo	0.835 ±0.024	0.650 ±0.047	0.777 ±0.019	0.606 ±0.044
Flux1-Dev	0.879 ±0.013	0.644 ±0.031	0.828 ±0.013	0.623 ±0.042
Flux1-Schnell	0.844 ±0.018	0.681 ±0.043	0.792 ±0.015	0.651 ±0.039

Table 5: Fame-stratified separability (real vs. perturbed): High-fame names (log-pageviews above the median) are more separable than low-fame names, consistent with stronger identity memorization for well-known individuals.

Zayden Chuckson **Gertrude Schwartz** **Mehmet Demir**

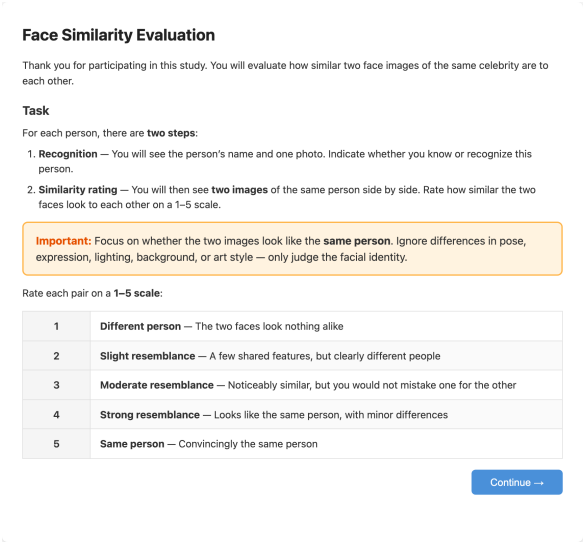
Figure 13: *Prototype blending*, an interpretability technique for visualizing associations with unfamiliar names. For each name, generated faces (faded, periphery) are inverted into StyleGAN2 latent vectors, and their mean code is decoded to produce a *prototype face* (center, full color). Despite diverse generations, this yields prototype images that reflect their shared aggregate characteristics, such as demographic associations (e.g., gender, ethnicity, and age). Note that this technique is not used as part of our probe; it is a post-hoc interpretability visualization. All generations use SDXL-Base.

F.3 Survey Interface

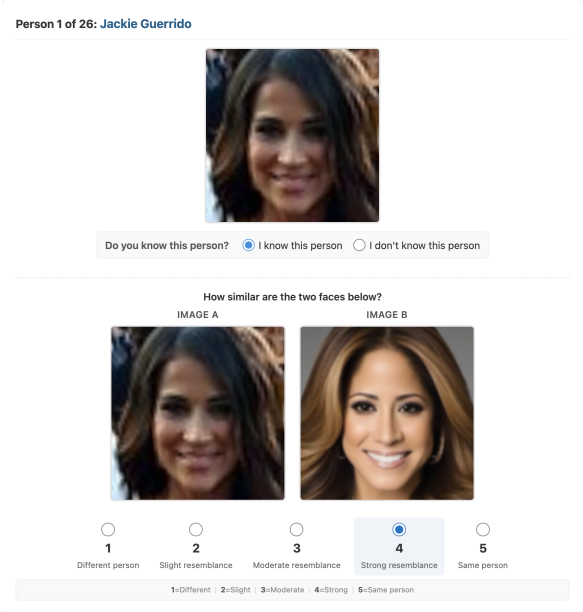
Figure 14 shows the annotation interface used in our human evaluation study. Each annotator is first presented with task instructions and the rating scale (fig. 14a). They then evaluate each celebrity in a two-step card layout (fig. 14b): first indicating whether they recognize the person, then rating the similarity between a ground-truth photo and a generated image on a 1–5 Mean Opinion Score (MOS) scale.

G Generative AI Disclosure

Generative AI tools were used for polishing wording and formatting in this manuscript, and for coding assistance. Gemini was used for demographic annotation as discussed in Section 2; these annotations are not included in the dataset for release.



(a) Task instructions shown to annotators.



(b) Evaluation card: recognition question followed by side-by-side similarity rating.

Figure 14: User study interface. (a) Annotators first read the task guidelines and rating scale. (b) For each celebrity, they indicate whether they recognize the person, then rate the similarity between two images on a 1–5 scale.