DIAGNOSING AND MITIGATING MEMORISATION IN GENERATIVE MODELS WITH METRIC SPACE MAGNITUDE

Anonymous authors
Paper under double-blind review

ABSTRACT

Memorisation in deep generative models is often reduced to the question of whether a similarity threshold is crossed. Current successful approaches typically analyse sample-wise comparisons between generated outputs and individual training examples. In this work, we propose a framework based on metric-space magnitude, an invariant that measures the effective number of distinct samples across multiple scales, providing a notion of diversity. For example, in an image embedding space, similar images contribute less effective diversity than distinct ones. By comparing the effective diversity of the training and generated sets and their combination, we shift from pointwise copy detection to a set-level description of how strongly and in what form the generated distribution is absorbed into the training set. We apply our framework to image, 3D, and neural-network parameter generation, where it provides a richer quantitative account of how memorisation is expressed in these settings, for example, whether it is broadly distributed over the training set, concentrated in a small part of it, or accompanied by self-collapse. This broader view remains consistent with the main findings of previous analyses. Finally, we use our approach to construct post-training interventions for image and face generation, substantially reducing memorisation while avoiding undesirable geometric side effects.

1 INTRODUCTION

Memorisation in deep generative models occurs when a model reproduces training examples rather than learning reusable structures from the training distribution. This behaviour has been observed across several generative settings and raises concerns about privacy, copyright, and model reliability (Chen et al., 2026; Carlini et al., 2023; Somepalli et al., 2023; Usynin et al., 2024; Wei et al., 2025).

Quantifying memorisation, however, remains difficult. Most existing approaches compare a generated sample with examples from the training set and decide whether it is a copy when its similarity to an example exceeds a threshold (Somepalli et al., 2023; Zeng et al., 2026). Other approaches partition the sample space (Meehan et al., 2020) or use nearest-neighbour tests (Alaa et al., 2021).

A binary decision only gives a partial description of the geometry of a generative model. Generated samples can become concentrated around training examples without reproducing them exactly, while a model can lose output diversity without moving toward the training set. Conversely, a small set of well-separated modes can maintain large pairwise distances even when it provides poor coverage of the target distribution. We therefore want to distinguish training-set overlap from other geometric failures, such as mode dropping and self-collapse, as seen in figure 1.

In this work, we view memorisation as *excess geometric redundancy with the training set*. A well-generalising model may generate samples on the same data manifold as the training data, but these samples should not be unusually redundant with training examples compared with fresh samples from the same distribution. This motivates a relative measurement focused on how much effective geometric structure the generated set shares with the training set beyond what is expected from a ”healthy” set.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

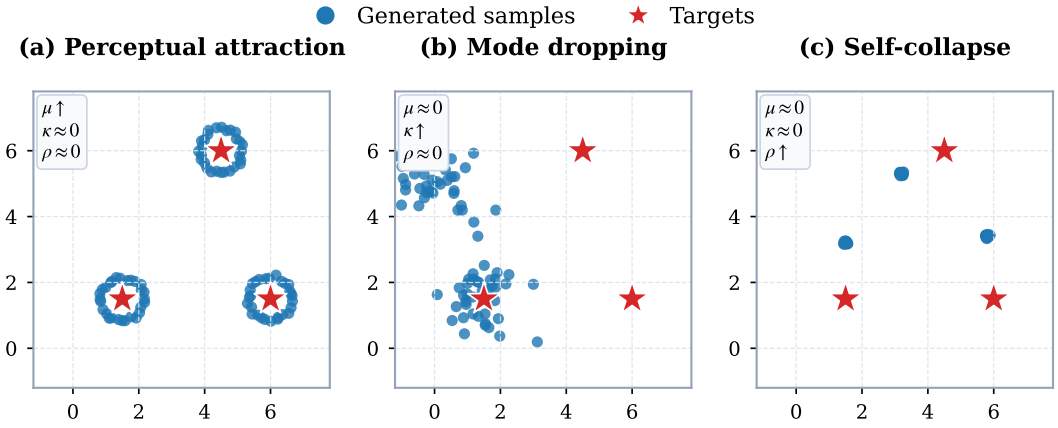

Figure 1: Diagnostic failure modes of the generative distribution relative to a target set. **(a)** *Perceptual attraction*: generated mass orbits targets without covering support. **(b)** *Mode dropping*: missing modes induce high coverage deficit. **(c)** *Self-collapse*: internal sample diversity collapses independently of targets.

We build this comparison using the **magnitude of metric spaces** (Leinster, 2013), a multi-scale measure of the effective size of a metric space. We define complementary quantities that measure the excess absorption of generated samples by the training set (μ), the relative coverage deficit (κ), and the relative internal redundancy (ρ) of the generated set. Together, these quantities describe different aspects of generative failure without reducing the analysis to a single nearest-neighbour threshold. The construction only requires samples to be represented in a common metric space, allowing the same analysis to be applied across different data modalities and representations.

We first study the framework in controlled experiments where the amount and type of memorisation are known. We then revisit previous analyses of memorisation in text-to-image diffusion models (Carlini et al., 2023), 3D generative models (Pu et al., 2026), and neural-network parameter generators (Zeng et al., 2026). Our measurements recover the main qualitative findings of these specialised analyses while revealing how apparently similar memorisation signals can arise from different geometric effects. In particular, the framework distinguishes attraction towards the training set from repeated generation in the same regions and separates both from coverage failures. This distinction matters because these behaviours can require different forms of mitigation rather than a single generic notion of "more" or "less" memorisation.

Finally, we use these diagnostic signals to design post-training strategies that target the identified geometric failure and use "healthy" samples to determine how far the intervention should extend. In our image generation analysis, we reduced detected copies from the base model by approximately 73% on held-out prompts, with only a small additional cost to ordinary generation diversity. In face generation, our final magnitude-based post-training reduces training-set absorption by approximately 51% for identities never seen during post-training.

Our main contributions are:

1. We introduce a multiscale, counterfactual-relative framework for diagnosing memorisation using metric-space magnitude. The resulting fingerprint separates three complementary geometric effects: training-set absorption (μ), coverage deficit (κ), and internal redundancy (ρ), using the same construction across different metric representations.
2. We empirically validate the framework across image diffusion, text-to-3D generation, and neural-network generation. The proposed diagnostics recover the main qualitative findings of previous analysis while revealing additional structure that is hidden by sample-wise proximity or single-statistic comparisons.
3. We show that the diagnostic geometry can directly inform post-training mitigation. In image and identity generation, the resulting interventions substantially reduce memorisation,

generalise to held-out memorising examples, and allow geometric side effects of naive repulsion to be identified and corrected.

The remainder of the paper is organised as follows. Section 2 contains related work on memorisation and metric-space magnitude. In Section 3 we introduce the proposed framework. Finally, we present the diagnostic case studies and the post-training experiments in Section 4.

2 Related Work

Memorisation In generative models, memorisation is commonly evaluated by comparing generated samples with training examples, either through similarity thresholds, nearest-neighbour statistics, or distributional tests (Meehan et al., 2020; Alaa et al., 2021; Somepalli et al., 2023; Carlini et al., 2023). Early work on data copying formalised the problem as distinguishing generated samples that are unusually close to the training set from those expected under a held-out population (Meehan et al., 2020). Related generative-model evaluation frameworks have also separated fidelity, coverage, and sample authenticity; for example, Alaa et al. (2021) combines precision–recall analysis with an authenticity criterion intended to detect copying from the training data.

In diffusion models, training data extraction and replication have been demonstrated using image-similarity and retrieval-based criteria (Carlini et al., 2023; Somepalli et al., 2023). Subsequent work has developed methods to detect memorising and to reduce replication during generation or fine-tuning (Wen et al., 2024; Chen et al., 2024; Ren et al., 2024). MemBench (Hong et al., 2024) provides a collection of prompts and associated training images known to trigger reproduction, making it suitable for comparing our set-level diagnostic with established copy-detection signals.

Memorisation diagnostics have also been extended beyond 2D image generation. Pu et al. (2026) study text-to-3D models using nearest-training distances and a Mann-Whitney statistic, showing that memorisation can vary non-monotonically with guidance strength. In generated neural networks, Zeng et al. (2026) compares generated checkpoints with training checkpoints. Identity memorisation in text-to-image models has similarly been studied through face-representation similarity (Alper et al., 2026). These settings provide complementary test cases because they use different data modalities, representations, and specialised memorisation diagnostics.

A more fundamental limitation is that a memorisation score need not reveal which failure mode drives it. Stein et al. (2023) show that several memorisation metrics can be dominated by properties such as mode shrinkage or image fidelity. This motivates a diagnostic that goes beyond detecting abnormal similarity and instead separates the geometric effects that can produce it.

Magnitude in Deep Learning Magnitude of Metric Spaces (Leinster, 2013) was introduced as a measure for biological diversity (Solow & Polasky, 1994) and has recently been incorporated as a valuable tool in machine learning.

Specifically, Limbeck et al. (2024) develops multi-scale magnitude measures for the diversity of latent representations. Their approach integrates differences between magnitude functions and is shown to detect mode collapse and mode dropping in generative models. Torkamani et al. (2026) instead constructs a distance between two finite datasets from the magnitudes of the individual sets and their union, thereby using the non-additivity of magnitude to quantify geometric dissimilarity across scales and use this distance as a learning objective for a generative model.

Other work has explored magnitude as a mechanism for controlling diversity. Huntsman (2023) uses magnitude to increase diversity in multi-objective evolutionary optimisation, while concurrent work by Hu et al. (2026) studies scale-dependent diversity profiles for generative AI. These works establish magnitude as both a multi-scale descriptor and a useful optimisation signal. However, their focus is on diversity, dataset dissimilarity, or diversity enhancement rather than memorisation itself.

3 Method

This section introduces metric-space magnitude and our counterfactual-relative framework for diagnosing memorisation.

3.1 PRELIMINARIES: METRIC-SPACE MAGNITUDE

Magnitude assigns an effective size to a metric space (Leinster, 2013). Let $X = \{x_1, \ldots, x_n\}$ be a finite metric space with pairwise distances $d_{ij} = d(x_i, x_j)$. At scale $t > 0$, define

$$Z_{ij}(t) = \exp(-t d_{ij}). \tag{1}$$

When $Z(t)$ is invertible, its weighting vector $w(t)$ satisfies

$$Z(t)w(t) = \mathbf{1}, \qquad \mathrm{Mag}(X, t) = \mathbf{1}^\top w(t). \tag{2}$$

Magnitude can be interpreted as an effective number of points: geometrically redundant points contribute less independent effective size and increasingly separated points contribute more.

Crucially, magnitude is scale-dependent. Rescaling the metric by t gives the *magnitude function*

$$\mathrm{Mag}_X(t) = \mathrm{Mag}(X, t). \tag{3}$$

At low resolution t, nearby observations become indistinguishable, while every point becomes different at higher t. For a finite metric space, $\mathrm{Mag}_X(t) \to |X|$ as $t \to \infty$. This multiscale behaviour is particularly useful for analysing generated data. Exact duplicates remain redundant, but magnitude also responds continuously to near-duplicates and broader clusters. At the same time, taking t too large is uninformative for our purpose: for any pair of distinct points with positive distance, $\exp(-td) \to 0$, so even very similar points are eventually resolved as independent. Interestingly, relevant redundancy appears over an intermediate range of resolutions.

Limbeck et al. (2024) exploit this observation for representation diversity by summarising magnitude functions by using the area under a magnitude function and compare two spaces through the corresponding area between their magnitude functions. Integrating across scales retains information that would be lost by evaluating magnitude at a single chosen resolution.

Finally, for two finite metric spaces in which all cross-set distances are infinite, the magnitude of their union is additive:

$$\mathrm{Mag}(A \sqcup B, t) = \mathrm{Mag}(A, t) + \mathrm{Mag}(B, t). \tag{4}$$

For point clouds at finite separation, the loss of additivity depends on the scale and the points' proximity. This motivates us to use this deficit as a measure of cross-set geometric redundancy.

3.2 OUR FRAMEWORK

We consider three sets represented in a common metric space: R, the reference or training set; G, the generated samples; and P, a fresh counterfactual comparable to R and not used for training, for example, a validation set. The role of P is to give a referent of what we expect healthy generations to look like. Hence, in our framework, we ask whether G is unusually redundant with R compared to a "healthy" baseline.

Shared magnitude. Our construction is inspired by the classical view of overlapping sets, but operates on geometric redundancy. From the additive baseline in Eq. 4, we define the magnitude deficit between the reference and a set X as

$$O_{R,X}(t) = \left[\mathrm{Mag}(R, t) + \mathrm{Mag}(X, t) - \mathrm{Mag}(R \cup X, t)\right]_+, \qquad X \in \{G, P\}. \tag{5}$$

If X contributes geometry independent of R, their effective sizes add and $O_{R,X}$ is small. If points in X are redundant with R, less effective size is gained by combining the two spaces and the deficit grows. Unlike an exact-copy count, this quantity can respond continuously to partial geometric similarity.

The same shared magnitude can be viewed from either side. We define

$$s_X(t) = \frac{O_{R,X}(t)}{\mathrm{Mag}(X, t)}, \qquad c_X(t) = \frac{O_{R,X}(t)}{\mathrm{Mag}(R, t)}. \tag{6}$$

Here s_X measures the fraction of the effective mass of X that is shared with the reference, whereas c_X measures how strongly X reaches the reference. Intuitively, the distinction is analogous to viewing the same overlap from opposite sides of a Venn diagram.

Absorption. We first compare the shared fractions of generations and fresh samples,

$$\delta(t) = s_G(t) - s_P(t). \tag{7}$$

Positive δ means that generations are more redundant with the training set than fresh data are at the same resolution. We summarize this excess across resolutions up to $t = t_C$ as

$$\mu = \frac{\int_0^{t_C} \delta(t)\, dt}{\int_0^{t_C} (1 - s_P(t))\, dt}. \tag{8}$$

We call μ the *absorption degree*. The denominator measures the novel effective mass available to the fresh counterfactual, so μ quantifies how much of this expected novelty is lost by G through excess redundancy with the training set.

Coverage. A generator may also fail by covering too little of the reference geometry. We therefore compare the reference-side fractions,

$$\kappa = \frac{\int_0^{t_C} (c_P(t) - c_G(t))\, dt}{\int_0^{t_C} c_P(t)\, dt}. \tag{9}$$

We call κ the *coverage deficit*. Values $\kappa > 0$ indicate that G reaches less of the reference geometry than fresh data do, as under mode dropping or over-correction; $\kappa \approx 0$ indicates comparable reach; and $\kappa < 0$ indicates that G hugs the reference more strongly than the counterfactual.

Internal redundancy. We measure redundancy within the generated set itself by comparing its normalised magnitude profile with that of the healthy counterfactual. For $X \in \{G, P\}$, let $d_X(t) = \mathrm{Mag}(X, t)/|X|$. We then define

$$\rho = \frac{\int_0^{t_C} (d_P(t) - d_G(t))\, dt}{\int_0^{t_C} d_P(t)\, dt}. \tag{10}$$

Thus, $\rho > 0$ indicates excess internal redundancy relative to the counterfactual.

Our proposed quantities separate behaviours that can produce similar nearest-neighbour or diversity scores, but correspond to different expressions as can be seen in Figure 1: μ measures excess absorption by the training set, κ measures loss of reference coverage, and ρ measures collapse within the generated set.

4 Experiments

4.1 Diagnostics

We evaluate our framework on three generative settings: text-to-image diffusion, text-to-3D generation, and neural-network parameter generation. The aim is to test whether the same diagnostic construction remains consistent with previous memorisation findings and to complement it with further insights from our framework. Complete definitions of R, G, and P, representation-specific distances, comparison criteria, and full numerical results are reported in Appendix B.

For **image diffusion**, we evaluate `Stable Diffusion v1.4` (Rombach et al., 2021) on MemBench (Hong et al., 2024), generating samples for each from prompts known to memorise. We compare the complete population with the prompts for which SSCD (Pizzi et al., 2022) detects at least one copy, and repeat the analysis using DINOv2 (Oquab et al., 2024), and CLIP (Radford et al., 2021) spaces. For **3D generation**, we evaluate the text-conditioned VecSet model of Pu et al. (2026) at guidance strengths $w \in \{1, 3, 5, 10\}$, where guidance controls the strength of the text condition. This tests whether our approach recovers the previously reported non-monotonic dependence of memorisation on guidance. Finally, for **neural-network generation**, we evaluate Hyperrepresentations (HR) (Schürholt et al., 2021; 2022), G.pt (Peebles et al., 2022), and P-diff (Wang et al., 2024) as in Zeng et al. (2026), considering parameter, functional, and internal representation spaces. Across all three settings, only the underlying metric space changes.

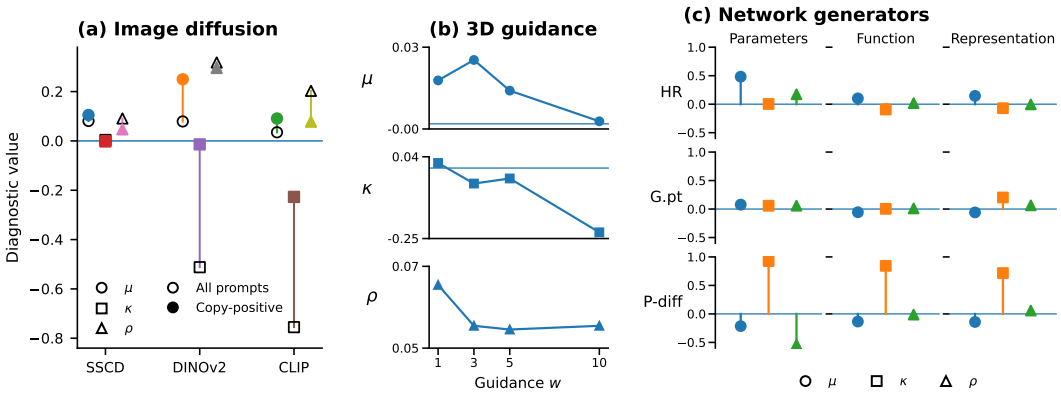

Figure 2: **Our framework across three generative settings. (a)** Stable Diffusion on MEMBENCH: hollow markers show the complete prompt population and filled markers the ones with SSCD-detected copies. Absorption μ increases on the copy-positive subset, while positive ρ shows that internal redundancy is also present. **(b)** Text-to-3D generation across guidance strengths: μ peaks at moderate guidance, while ρ remains nearly unchanged, and κ varies separately. **(c)** Neural-network generation across parameter, functional, and representation spaces. HR shows strong μ, G.pt a weaker effect, and P-diff a large κ. Full details are in Appendix B.

Overall, the framework recovers the qualitative conclusions of the previous analyses. In 3D generation, memorisation remains weak and μ retains the reported moderate-guidance maximum. In neural network generation, HR shows the strongest parameter-space replication signal, G.pt a substantially weaker one, and P-diff no positive absorption from individual training checkpoints, consistent with the behaviour identified by Zeng et al. (2026).

The image generation experiment shows directly that the detected memorisation is caused by two different effects. In DINOv2, the proximity criterion detects more than double the images than in SSCD as copies, but absorption remains almost unchanged. In our analysis, these copies instead appear as internal redundancy, with $\rho = 0.318$ compared with $\rho = 0.091$ in SSCD. Thus, a pointwise criterion can count many nearby generations without distinguishing whether they correspond to additional absorption or repeated generations occupying the same memorised regions.

The analysis on the subset of prompts with detected copies further supports this interpretation (Figure 2(a)). Restricting the analysis to the 36 prompts where copying is observed increases μ in all three representations, most strongly in DINOv2, from 0.079 to 0.250, while ρ remains positive. Memorising prompts are therefore associated with both stronger training-set absorption and an internally redundant generated population, two effects that a single proximity count does not separate.

In 3D generation, the moderate-guidance peak occurs in μ while ρ remains nearly constant (0.053–0.065), showing that the effect is not simply a loss of generated-set diversity. In neural-network generation, figure 2(c) shows that HR's strong parameter-space absorption ($\mu = 0.483$) weakens substantially in functional and representation space, while G.pt exhibits only weak parameter-space absorption. P-diff is qualitatively different: μ is negative in all three spaces, but κ is very large (0.928, 0.843, and 0.717). This means that the characteristic of learned generations is a restricted coverage of the reference space, consistent with the narrow interpolation behaviour identified by Zeng et al. (2026).

Taken together, these experiments show that the same framework can be applied to different modalities, and it is interpretable and effective, as shown by being consistent with the qualitative findings of specialised diagnostics. More importantly, it separates training-set absorption, internal redundancy, and restricted coverage, thereby providing the insights required to diagnose memorisation in a model.

4.2 POST TRAINING

Based on our magnitude measures, we develop post-training interventions whose targets are defined from matched healthy counterfactual populations rather than from memorised training examples. In both settings, we begin with a simple distance-based surrogate for magnitude redundancy and progressively develop a magnitude-native post-training objective. Here we present the resulting geometric and behavioural changes; loss objectives are explained in detail in Appendix D, while measured thresholds, training protocols, and implementation details are provided in Appendix E

4.2.1 STABLE DIFFUSION IMAGE GENERATION

Memorising prompts produces abnormally tight groups of generations across independent diffusion seeds. We train a LoRA only within prompt pairs lying inside this abnormal region, stopping at a distance floor measured from matched non-memorising prompts. We use 24 of the 36 memorising prompts for training and reserve the remaining 12 for held-out evaluation; full training details are given in Appendix D and Appendix E. We evaluate our results using different encoders. The number of detected copies on both the training and the held-out set prompts falls substantially, as seen in Table 9.

Table 1 shows that for all encoders absorption μ, and internal redundancy ρ decreases substantially under repulsion, with ρ moving toward the healthy counterfactual level of zero. In contrast, κ remains close to its original value. The consistency across sets and encoders evaluated shows that the reduction in copying is accompanied by a consistent change in independently measured image geometry. This effect can also be observed in Figure 3.

Table 1: **Replicated geometric effect of post-training.** Values are mean $\pm$ standard deviation on three independent runs and controls across two runs without our loss applied.

Representation	Axis	Untouched	+ Repulsion	Control
SSCD	μ	0.130	$.061 \pm .009$	$.124 \pm .010$
	κ	-0.002	$-.0022 \pm .0002$	$-.0001 \pm .0003$
	ρ	0.060	$.026 \pm .004$	$.058 \pm .005$
DINOv2	μ	0.267	$.160 \pm .006$	$.224 \pm .018$
	κ	-0.051	$-.0495 \pm .0007$	$-.0464 \pm .0021$
	ρ	0.447	$.299 \pm .009$	$.396 \pm .025$
CLIP	μ	0.135	$.020 \pm .004$	$.095 \pm .019$
	κ	-0.273	$-.2759 \pm .0022$	$-.2665 \pm .0011$
	ρ	0.269	$-.017 \pm .011$	$.202 \pm .048$

Finally, we evaluate ordinary prompts to verify that the reduction in memorisation is not due to degraded generation, and we find no substantial changes in generation quality (Table 10). Overall, the same geometric quantity that identifies the abnormal within-prompt concentration also supplies a stopping criterion for post-training. Moving only the pathological regions to the level measured on prompts without copies detected substantially reduces memorisation, transfers to held-out memorising prompts, and produces a replicated movement toward healthy geometry across independent image representations, without a major impact on the model's performance.

4.2.2 FACE GENERATION

For identity memorisation, we follow a progressive intervention strategy. Direct same-name repulsion removes the memorised identity but introduces a new contraction across post-trained identities. We therefore introduce magnitude-based self- and cross-name constraints to preserve identity separation, followed by a plausibility constraint. All intervention targets are measured from perturbed-name counterfactuals. Training uses AdaFace (Kim et al., 2022) for face recognition, and magnitude results are independently evaluated with ArcFace (Deng et al., 2019). Complete objectives and training details are provided in Appendix D and Appendix E.

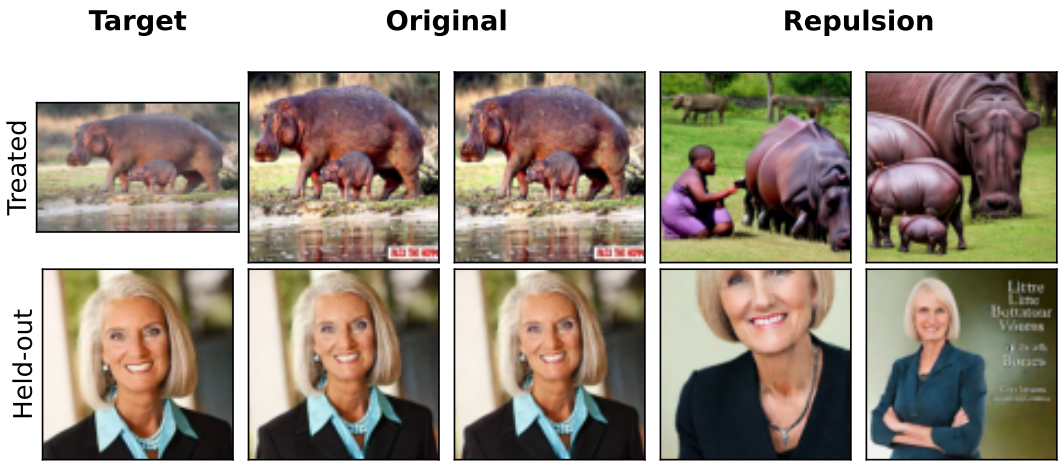

Figure 3: **post-training samples on memorising prompts.** We show a prompt (top) used in the post-training and a held-out memorising prompt (bottom). MemBench training image followed by generations from the original model and the post-trained model. The corresponding prompts are "mothers influence giving gabriel a hippo" and "little beyond other voices with ann graham lotz"

Table 2: Face-generation post-training results on training and held-out identities. Training identities are evaluated with both AdaFace and ArcFace. Held-out results are evaluated using ArcFace on 24 memorising identities that were never used during post-training. Δsim denotes the median ArcFace similarity change to the reference identity on treated identities.

		Treated identities						**Held-out identities**		
		AdaFace			**ArcFace**			**ArcFace**		
Variant	Δsim	μ	κ	ρ	μ	κ	ρ	μ	κ	ρ
Original	–	0.255	0.021	0.155	0.232	0.016	0.134	0.204	0.028	0.102
Retention control	-0.019	0.215	0.011	0.131	0.199	0.007	0.116	0.215	0.029	0.106
Distance repulsion	-0.447	0.144	0.104	0.197	0.137	0.099	0.186	0.118	0.090	0.150
Magnitude self+cross	-0.409	0.096	0.033	0.083	0.095	0.025	0.070	0.099	0.036	0.080
Magnitude + plausibility	-0.373	0.061	0.013	0.063	0.062	0.007	0.053	0.100	0.038	0.086

Following the procedure in Alper et al. (2026), we prompt `Stable Diffusion v1.4` with real names and compare the resulting faces with a reference photograph of the corresponding person. For each name, we construct a one-character perturbation of the name as a matched counterfactual. Across a population of names, we find that the names produce generations that are substantially closer to the corresponding identity than their perturbed-name controls. We use the strongest cases for post-training and reserve a separate set of high-memorisation identities for held-out evaluation.

The loss is implemented in the identity space. During training, the predicted image is decoded, its face is detected and aligned using an SCRFD detector (Guo et al., 2022), and the face is embedded by an identity encoder. During training, the predicted image is embedded with AdaFace. We train a rank-8 LoRA adapter on the UNet attention projections and keep the remaining model, face detector, and identity encoder frozen.

Results are shown in Table 2. Although direct repulsion strongly reduces identity similarity and absorbed mass, the coverage deficit and internal redundancy increase. While the intervention has removed the original identity, it created excess redundancy elsewhere in identity space, as can also be observed in Figure 4.

This contraction motivates an intervention that operates directly on magnitude weights. We replicate the repulsion effect, and add another repulsion effect to other samples generated from other names that are affected during the post-training. As Table 2 shows, adding the self- and cross-name magni-

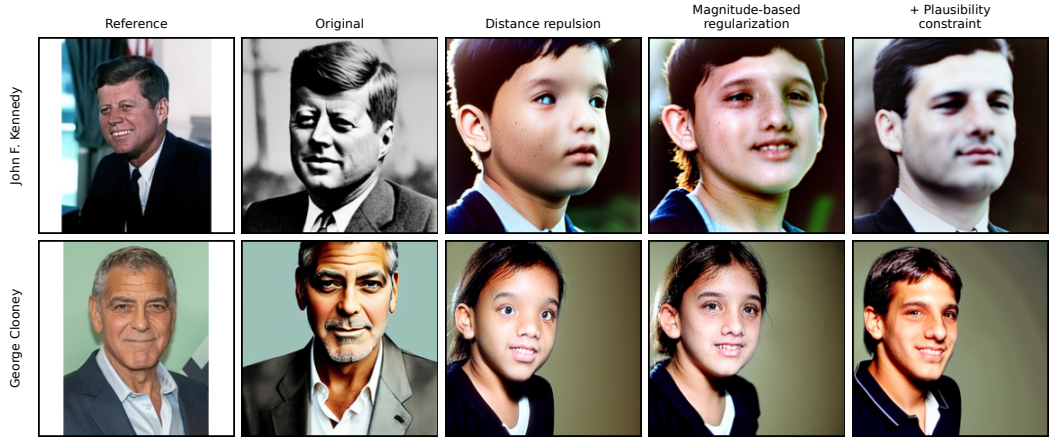

Figure 4: **Progressive intervention for identity memorisation.** Each row shows one identity: the reference photograph, the original model output, distance repulsion, the magnitude-based intervention, and the plausibility-constrained variant. Direct repulsion removes the memorised identity but can introduce a shared degraded appearance; the magnitude constraints restore separation across identities, while the plausibility ceiling reduces the remaining common drift.

tude constraints preserves strong identity removal while repairing the failure introduced by distance repulsion: κ returns close to its original level and ρ moves toward the healthy counterfactual.

Distinct identities can remain separated from one another while drifting together away from the healthy face population. We add a plausibility term that constrains a candidate whose magnitude weight relative to H is unusually high, thereby pulling it back toward a 'healthy' range. This also reduces the geometric memorisation signal (Table 2) and reduces alignment with the measured child-like drift direction. The trade-off is a moderate reduction in identity-removal strength.

Our results are consistent with the held-out names. Table 2 uses the same independent ArcFace evaluator and fresh generation seeds. Direct distance repulsion again reduces absorption but produces coverage and redundancy failure, whereas the magnitude-based variants retain low absorption without the same contraction. Results are consistent in the held-out ArcFace identity similarity, showing that the learned adapter transfers beyond the identities used for post-training with only a modest effect on ordinary generation(Table 11). Figure 4 shows the same progression qualitatively.

5 Conclusions

We introduced a framework for diagnosing memorisation in generative models using metric-space magnitude. Rather than reducing memorisation to pointwise proximity, our construction provides a set-level comparison that separates training-set absorption, coverage deficit, and internal redundancy.

Our framework recovers the main qualitative findings of previous memorisation analyses while revealing additional structure. In particular, it distinguishes training-set attraction from repeated generations occupying the same regions, and identifies restricted coverage in cases where direct absorption is absent.

We further showed that the same magnitude geometry can be used to reduce memorisation. In image diffusion, a local surrogate for magnitude redundancy substantially reduces copying on held-out prompts while decreasing both internal redundancy and training-set absorption. In face generation, where this local correction introduces a new collapse, direct magnitude-weight regularisation restores healthier population geometry while retaining strong memorisation reduction. These results establish magnitude as both a diagnostic and an optimisation principle for generative memorisation.

AI USE STATEMENT

We used ChatGPT (OpenAI) to assist with drafting and editing parts of the manuscript to improve clarity and organisation. We used Claude (Anthropic) to assist with the implementation, debugging, and modification of experimental and evaluation code. All AI-assisted text and code were reviewed by the authors, experimental outputs were checked for correctness and consistency, and all reported results and claims were verified by the authors. The authors take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The definitions of the proposed diagnostics, including the scale selection procedure and numerical evaluations, are given in Section 3 and Appendix B. Appendices D and E provide the post-training objectives, calibration procedures, optimisation settings, training protocols, evaluation procedures, and matched controls used in the post-training experiments. We intend to upload the repository after publication.

REFERENCES

Ahmed M. Alaa, Boris van Breugel, Evgeny S. Saveliev, and Mihaela Vanderschaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231942787.

Morris Alper, Vasudha Varadarajan, Moran Yanuka, Angelina Wang, and Hadar Averbuch-Elor. Namesakes: Probing identity memorization in text-to-image models, 2026. URL https://arxiv.org/abs/2606.20155.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Chen Chen, Daochang Liu, and Chang Xu. Towards Memorization-Free Diffusion Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8425–8434, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00805. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00805.

Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. *Computer Graphics Forum*, 22(3):223–232, 2003. doi: https://doi.org/10.1111/1467-8659.00669. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8659.00669.

Yunhao Chen, Shuejie Wang, Difan Zou, and Xingjun Ma. Side: surrogate conditional data extraction from diffusion models. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i1.36972. URL https://doi.org/10.1609/aaai.v40i1.36972.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RhB1AdoFfGE.

Chunsan Hong, Tae-Hyun Oh, and Minhyuk Sung. Membench: Memorized image trigger prompt dataset for diffusion models. *ArXiv*, abs/2407.17095, 2024. URL https://api.semanticscholar.org/CorpusID:271404527.

Xiuyuan Hu, Xuege Hou, Guoqing Liu, Yang Zhao, Jieran Li, Dongbiao Sun, José Miguel Hernández-Lobato, Hao Zhang, and Xue Liu. Evaluating the diversity of ai-generated content with diversity profiles, 2026. URL https://arxiv.org/abs/2608.17731.

Steve Huntsman. Diversity enhancement via magnitude. In Michael Emmerich, André Deutz, Hao Wang, Anna V. Kononova, Boris Naujoks, Ke Li, Kaisa Miettinen, and Iryna Yevseyeva (eds.), *Evolutionary Multi-Criterion Optimization*, pp. 377–390, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-27250-9.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18729–18738, 2022. doi: 10.1109/CVPR52688.2022.01819.

Tom Leinster. The magnitude of metric spaces. *Documenta mathematica*, 18:857–905, 2013. ISSN 1431-0643.

Katharina Limbeck, Rayna Andreeva, Rik Sarkar, and Bastian Rieck. Metric space magnitude for evaluating the diversity of latent representations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 123911–123953. Curran Associates, Inc., 2024. doi: 10.52202/079017-3937. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/dfc24bd3ec5d74960e104268bbb52849-Paper-Conference.pdf.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. ISSN 00034851. URL http://www.jstor.org/stable/2236101.

Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. *CoRR*, abs/2004.05675, 2020. URL https://arxiv.org/abs/2004.05675.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

William Peebles, Ilija Radosavovic, Tim Brooks, Alexei A. Efros, and Jitendra Malik. Learning to learn with generative models of neural network checkpoints, 2022. URL https://arxiv.org/abs/2209.12892.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. *Proc. CVPR*, 2022.

Shu Pu, Boya Zeng, Kaichen Zhou, Mengyu Wang, and Zhuang Liu. Memorization in 3d shape generation: An empirical study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 1828–1838, June 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine*

Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII*, pp. 340–356, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72979-9. doi: 10.1007/978-3-031-72980-5_20. URL https://doi.org/10.1007/978-3-031-72980-5_20.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Konstantin Schürholt, Dimche Kostadinov, and Damian Borth. Self-supervised representation learning on neural network weights for model characteristic prediction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16481–16493. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/89562dccfeb1d0394b9ae7e09544dc70-Paper.pdf.

Konstantin Schürholt, Boris Knyazev, Xavier Giró-i Nieto, and Damian Borth. Hyper-representations as generative models: Sampling unseen neural network weights. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27906–27920. Curran Associates, Inc., 2022. doi: 10.52202/068431-2024. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b2c4b7d34b3d96b9dc12f7bce424b7ae-Paper-Conference.pdf.

Andrew R. Solow and Stephen Polasky. Measuring biological diversity. *Environmental and Ecological Statistics*, 1(2):95–103, jun 1994. ISSN 1573-3009. doi: 10.1007/BF02426650. URL https://doi.org/10.1007/BF02426650.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2023. doi: 10.1109/CVPR52729.2023.00586.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sahel Torkamani, Henry Gouk, and Rik Sarkar. Magnitude distance: A geometric measure of dataset similarity. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=9ylPoHEKed.

Dmitrii Usynin, Moritz A Knolle, and Georgios Kaissis. Memorisation in machine learning: A survey of results. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://api.semanticscholar.org/CorpusID:273145923.

Kai Wang, Dongwen Tang, Boya Zeng, Yida Yin, Zhaopan Xu, Yukun Zhou, Zelin Zang, Trevor Darrell, Zhuang Liu, and Yang You. Neural network diffusion, 2024. URL https://arxiv.org/abs/2402.13144.

Jiaheng Wei, Yanjun Zhang, Leo Yu Zhang, Ming Ding, Chao Chen, Kok-Leong Ong, Jun Zhang, and Yang Xiang. Memorization in deep learning: A survey. *ACM Comput. Surv.*, 58(4), October 2025. ISSN 0360-0300. doi: 10.1145/3769076. URL https://doi.org/10.1145/3769076.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=84n3UwkH7b.

Boya Zeng, Yida Yin, Zhiqiu Xu, and Zhuang Liu. Generative modeling of weights: Generalization or memorization? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41974–41984, June 2026.

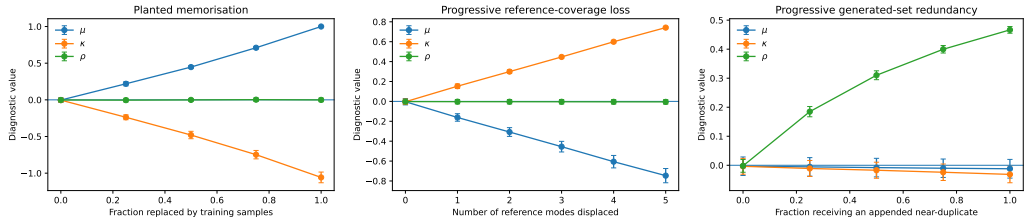

Figure 5: **Controlled response of the final geometric fingerprint. Left:** replacing generated samples with training samples produces a monotonic increase in absorption μ, reaching $\mu \simeq 1$ under complete replacement. **Centre:** progressively removing reference coverage increases κ, while the rigid within-mode translations leave ρ approximately unchanged. **Right:** adding near-duplicate generations increases ρ while leaving μ and κ approximately unchanged. Error bars denote 95% confidence intervals over 30 independent seeds.

Table 3: **Endpoint response in the controlled sanity check.** Values are mean $\pm$ 95% confidence interval over 30 independent seeds. Full dose–response curves are shown in Fig. 5.

Perturbation	**Control**	μ	κ	ρ
Planted memorisation	$\alpha : 0 \to 1$	$-0.003 \pm 0.032 \to +1.000 \pm 0.000$	$-0.004 \pm 0.027 \to -1.058 \pm 0.073$	$-0.002 \pm 0.022 \to -0.001 \pm 0.025$
Reference coverage	covered modes: $6 \to 1$	$-0.003 \pm 0.032 \to -0.746 \pm 0.072$	$-0.004 \pm 0.027 \to +0.742 \pm 0.011$	$-0.002 \pm 0.022 \to -0.005 \pm 0.022$
Internal redundancy	near-duplicate fraction: $0 \to 1$	$-0.003 \pm 0.032 \to -0.012 \pm 0.032$	$-0.004 \pm 0.027 \to -0.031 \pm 0.029$	$-0.002 \pm 0.022 \to +0.466 \pm 0.012$

A Controlled Sanity Checks

We verify the behaviour of the final three-axis fingerprint on a single controlled metric space. The base distribution is a two-dimensional mixture of six isotropic Gaussian modes arranged on a circle. For every run, the reference set R, healthy counterfactual P, and initial generated set G_0 are independent samples from the same mixture, with eight samples per mode and standard deviation 0.5. We use the same reference-anchored t_C, 40-point scale grid, and definitions of (μ, κ, ρ) as in Section 3.2. Results are averaged over 30 independent random seeds and error bars show 95% confidence intervals.

We introduce three controlled perturbations of G_0, each targeting a different geometric property. First, for *planted memorisation*, a fraction α of generated samples in each mode is replaced by exact samples from R. Second, to reduce *reference coverage* without introducing generated-set collapse, an increasing number of generated modes is translated radially away from its corresponding reference mode while preserving all within-mode distances. Third, to induce *internal redundancy* while leaving the existing generated support unchanged, we append near-duplicates of a controlled fraction of generated samples; the perturbation applied to each duplicate has standard deviation 0.02, compared with the within-mode standard deviation 0.5.

The controlled responses match the intended interpretation of the three axes. Planted copies increase μ and simultaneously make κ negative, reflecting the expected increase in reference hugging. Removing reference support drives κ upward while preserving the internal geometry measured by ρ; the accompanying negative μ indicates that the displaced generations are less absorbed by R than the healthy counterfactual. Finally, near-duplicate outputs increase ρ without changing the support occupied by G, leaving the two reference-relative axes approximately unchanged. The axes are therefore not forced to vary independently, but each responds predictably to the geometric property it was designed to measure.

Table 4: **Stable Diffusion memorisation on the complete MEMBENCH evaluation.** All magnitude quantities are computed on the same generations from all trigger prompts. The threshold column reports the generated rows/prompts crossing the corresponding criterion. SSCD uses the established copy threshold; DINOv2 and CLIP use healthy proximity nulls.

Representation	Threshold	μ	κ	ρ
SSCD	99 / 36	0.081	0.004	0.091
DINOv2	248 / 70	0.079	−0.512	0.318
CLIP	117 / 47	0.035	−0.755	0.203

B DIAGNOSTICS

In this section, we evaluate three cases of memorisation: the first is the generation of images, the second is the generation of 3d surfaces, and finally, we evaluate neural network parameter generation.

Across all diagnostic experiments, the multiscale range is selected using the same reference-anchored rule. For the corresponding reference set R, with $n_R = |R|$, we define

$$q_R = Q_{1/n_R}\left(\{D_{ij} : i < j,\ D_{ij} > 0\}\right), \qquad t_C = \frac{\log(n_R - 1)}{q_R}. \tag{11}$$

We evaluate all magnitude-based quantities on a fixed grid of 40 scales: $t = 0$ together with 39 logarithmically spaced values between $t_C/40$ and t_C. Thus, the same scale-selection procedure is used throughout the diagnostic experiments, with t_C determined only by the geometry of the corresponding reference set.

B.1 TEXT-TO-IMAGE DIFFUSION

We first evaluate `Stable Diffusion v1.4` (Rombach et al., 2021) on MEMBENCH (Hong et al., 2024), which provides 150 prompts together with training images that those prompts are known to reproduce. For each prompt, we generate $k = 8$ images using fixed seeds, for a total of 1,200 generations. Following prior work (Somepalli et al., 2023), we use SSCD (Pizzi et al., 2022) similarity to the benchmark-provided training image paired with each prompt as the reference copy detector. A generation with similarity above 0.5 is counted as a detected copy.

We then apply our framework to the set of generated images. The reference set R contains 63 unique benchmark-provided training images associated with the 150 prompts, G contains the generated images, and P consists of ordinary `Stable Diffusion v1.4` generations obtained using COCO captions (Lin et al., 2014), providing a healthy model-output counterfactual. We evaluate the same R, G, and P construction in three representation spaces: SSCD, DINOv2 (Oquab et al., 2024), CLIP (Radford et al., 2021). In every space we use

$$D_{ij} = \left(\frac{\|x_i - x_j\|_2}{\mathrm{median}_{a<b}\|x_a - x_b\|_2}\right)^2, \tag{12}$$

and evaluate the fingerprint over the reference scale range $[0, t_C]$ defined above.

For comparison with threshold-based decisions beyond SSCD, we also construct representation-specific proximity thresholds for DINOv2 and CLIP. Since these representations do not have an established copy threshold, we define the threshold as the 99.9th percentile of the maximum similarity between a healthy counterfactual generation and any image in R. This yields thresholds of 0.729 for DINOv2 and 0.873 for CLIP.

Table 4 first provides a direct comparison with the established SSCD criterion. The binary detector labels $99/1200 = 0.0825$ generations as copies, while our threshold-free multiscale analysis gives an absorption degree of $\mu = 0.081$. It is worth noting that the detected absorbed mass is approximately equal to the total detected memorisation in the SSCD case.

On DINOv2 and CLIP, the results are not as straightforward, where the memorisation is coherent with the absorbed mass μ, as it can be seen by the detection of 248 copies, and a similar μ then with

Table 5: **Geometry of prompts with detected memorisation.** The fingerprint is evaluated on the 36 prompts that detect copies under the reference SSCD criterion.

Representation	μ	κ	ρ
SSCD	0.105	−0.002	0.047
DINOv2	0.250	−0.014	0.296
CLIP	0.091	−0.227	0.078

SSCD. But the high ρ makes the nature of the copies detected clear, while our method absorbed mass suggests the existence of less diverse copies, it correctly detects a relevant internal redundancy on the generated set. Thus, many detected copies correspond to generations occupying redundant regions rather than to an equally large degree of persistent training-set absorption. A similar decomposition appears in CLIP, which indicates that in these encoders, there is a collapse of the generated samples in close regions absorbed by the training samples.

To further characterise the geometry of the prompts on which copying is actually observed. We evaluated the 36 prompts for which the reference SSCD detector identifies at least one copy. Here, R contains benchmark images associated with these prompts, G contains the generations in the prompt with copies detected, and P is a matched set of ordinary model generations. Table 5 reports the same magnitude induced metrics.

This analysis shows that the memorisation is substantially more pronounced. Relative to the complete population in Table 4, absorption rises significantly, especially in DINOv2, verifying the previous diagnostic, meaning a relevant self-collapse concentrated on the region being absorbed by the training set, while CLIP now show levels closer to SSCD results, with the difference being the existence of this self-collapse. Nevertheless, all three representations show a strong prompt-conditional concentration effect.

The resulting picture for this setting is therefore more specific than the binary copy decision alone. On prompts where copies are observed, the framework finds increased training-set absorption and abnormally tight repetition across diffusion seeds. This local concentration provides the geometric mechanism used in the post-training intervention of Sec. 4.2.

B.2 3D Generation

We evaluate in 3D generation using the setting of Pu et al. (2026). Their controlled experiments use a text-conditioned VecSet diffusion model. We use their released Text-Top16 checkpoint and sampling pipeline, and use Light-Field Distance (LFD) (Chen et al., 2003) for shape dissimilarity. Our experiments aim to evaluate two main insights of the setting. First, the model evaluated exhibits limited overall memorisation, and the effect of the guidance scale (the extent of the text prompt's influence on generation) is non-monotonic: low or high guidance yields lower memorisation, whereas intermediate levels are higher. Additionally, we study per-class behaviour to evaluate if the memorisation per class reflects the global pattern.

We perform the analysis on a subset of the categories by sampling the generating prompts, specifically on seven categories from the benchmark data. Within each category, generation prompts are sampled from the corresponding training captions. For the category-level analysis, we use 300 samples per category. For the pooled guidance experiment, we use 150 prompts per category and keep the same prompts fixed across guidance strengths to ensure a paired comparison.

For each of the used label we take R as the corresponding training shapes, G contains the generations produced from training captions, and P contains shapes from the held-out split. We evaluate using our framework on the pooled cases and the separated classes. On the pooled categories, we also evaluate guidance scales $w = 1$, $w = 3$ and $w = 5$, $w = 10$.

As an external reference, we compare the memorisation statistic proposed by Pu et al. (2026). They compare the LFDs on training samples nearest to the generated samples and on held-out samples using a Mann-Whitney statistic (Mann & Whitney, 1947), Z_U, following Meehan et al. (2020) 's work. Negative Z_U indicates that generated shapes tend to lie closer to the training set than held-out

Table 6: Comparison with the nearest-neighbour memorisation diagnostic of Pu et al. (2026). Z_U denotes the statistic evaluated on our seven-category LFD bed, while Z_U^{Pu} gives the values reported by Pu et al. (2026) for their controlled guidance experiment. The two columns, therefore, differ in evaluation population and are compared in terms of their guidance trend rather than their absolute values. $Z_U < 0$ indicates a shift towards closer training neighbours. μ measures excess set-level absorption, κ the coverage deficit, and ρ the internal redundancy relative to the counterfactual.

Condition	Z_U	Z_U^{Pu}	μ	κ	ρ
Pooled, $w = 1$	-1.810	$+1.40$	$+0.017$	$+0.018$	$+0.065$
Pooled, $w = 3$	-1.660	-5.11	$+0.025$	-0.055	$+0.054$
Pooled, $w = 5$	-0.930	-3.12	$+0.013$	-0.037	$+0.053$
Pooled, $w = 10$	-0.220	$+0.07$	$+0.001$	-0.228	$+0.054$
Rifle	-5.770	—	$+0.180$	$+0.080$	$+0.254$
Gun	-3.440	—	$+0.086$	-0.129	$+0.152$
Figurine	-3.030	—	$+0.010$	-0.162	$+0.227$
Toy	-1.780	—	-0.090	-0.266	$+0.105$
Vase	$+2.290$	—	-0.055	-0.273	$+0.060$
Table lamp	$+2.430$	—	-0.040	-0.328	$+0.031$
Chair	$+1.620$	—	-0.011	-0.171	$+0.047$

shapes, with more negative values corresponding to a stronger nearest-neighbour shift. Unlike our framework, this comparison reduces each shape to its single nearest training distance.

The results provided by our framework in 6 are consistent with the broader conclusion of Pu et al. (2026) that modern large-scale 3D generators exhibit limited overall memorisation. At the same time, the guidance sweep recovers their main non-monotonic finding: excess absorption is strongest at moderate guidance and decreases again as guidance is increased. The Z_U statistic evaluated on our smaller seven-category bed captures the subsequent decrease but does not resolve the increase from $w = 1$ to $w = 3$. In contrast, μ preserves the moderate-guidance peak observed by Pu et al. (2026).

The controlled experiment of Pu et al. (2026) evaluates a substantially larger, class-balanced population, whereas our pooled bed contains only seven categories and fewer samples. This matters particularly for Z_U. Our analysis C indicates that the guidance effect is expressed under the reduced category composition of our setting. The larger population evaluated by Pu et al. (2026) exhibits a net shift towards smaller nearest-training distances at moderate guidance, which is directly captured by Z_U. In our seven-category subset, however, this nearest-neighbour response is strongly heterogeneous: rifle and gun move in the memorisation direction from $w = 1$ to $w = 3$, whereas the remaining categories move in the opposite direction, so the pooled nearest-distance shift largely cancels and Z_U does not recover the peak. The effect nevertheless remains present in the joint geometry of the generated distribution. Thus, on our reduced population, the moderate-guidance effect is expressed primarily through the cross-category organisation of the generated set rather than through a systematic change in each sample's nearest-training distance. Because magnitude depends on the full similarity structure of G and $R \cup G$, it retains this pooled geometric effect even when the marginal nearest-neighbour statistic does not.

The remaining magnitude axes show that the guidance effect is not simply a change in generated-set diversity. ρ remains nearly unchanged across guidance strengths, while κ varies independently, particularly under strong guidance. Thus, the peak in absorption is not explained by increased self-redundancy, and reference reach can change substantially even when excess absorption is small.

The category decomposition reveals further structure hidden by the small pooled effect. Rifle and gun show the clearest excess absorption, whereas figurine exhibits little set-level absorption despite a marked nearest-neighbour shift and substantial internal redundancy. Toy provides an even sharper separation: its nearest-neighbour statistic indicates greater proximity to training examples, while its negative μ indicates no excess set-level absorption relative to the counterfactual. Vase and table lamp show little evidence of either effect. These results show that memorisation is strongly

Table 7: Memorisation diagnostics for neural-network generation across parameter, functional, and representation spaces. Z_U is an auxiliary nearest-training Mann–Whitney statistic computed on the same distance matrices; negative values indicate that generated networks lie closer to training checkpoints than the held-out reference, while positive values indicate the opposite shift.

Generator	Space	Z_U	μ	κ	ρ
HR	Parameters	-12.1	$+0.483$	$+0.004$	$+0.174$
	Function	-6.8	$+0.102$	-0.093	$+0.019$
	Representation	-4.6	$+0.146$	-0.070	-0.004
G.pt	Parameters	-5.4	$+0.078$	$+0.058$	$+0.059$
	Function	-2.2	-0.056	$+0.004$	$+0.011$
	Representation	-0.2	-0.059	$+0.203$	$+0.065$
P-diff	Parameters	$+13.6$	-0.214	$+0.928$	-0.539
	Function	$+13.6$	-0.134	$+0.843$	-0.012
	Representation	$+13.6$	-0.141	$+0.717$	$+0.059$

category-dependent and can appear as nearest-neighbour attraction, set-level absorption, or internal redundancy without these phenomena necessarily coinciding.

Overall, in this setting, we find that our framework preserves the principal model-level findings from the previous literature, establishing the reliability of our approach, while exposing substantial heterogeneity beneath the pooled result, showing versatility for diagnostics.

B.3 Neural Network Parameter Generation

Finally, we study memorisation in models that generate neural-network parameters. We evaluate the released Hyper-Representations (HR) (Schürholt et al., 2021; 2022), G.pt (Peebles et al., 2022), and P-diff(Wang et al., 2024) generators studied by Zeng et al. (2026). Following their analysis, we consider memorisation not only in parameter space but also in the behaviour and internal representation of the generated networks. We again found coherence with the qualitative insights from previous research, while generating insights that show naturally from our framework. Moreover, in this analysis, we demonstrate the suitability of our framework across different spaces and modes, assuming a reasonable and informative metric space.

For each generator, we evaluate three spaces. In addition to the parameter space and functional space from the original setting of Zeng et al. (2026), we also analyse the embedding space. In the parameter space, we use the weight alignment, as done in each method's source code, and compute the Euclidean distance over their flattened representations. Functional space corresponds to the predictions by the generated networks, where we take the Jaccard distance between sets of test examples misclassified by each network. Finally, for the embedding space, we explore a representation distance that captures the connectivity structure of network activations. For each checkpoint, we compute the minimum-spanning tree of the activation cloud, and we use the Euclidean distance between the sorted edge-length vectors. We do this because these representations don't need alignment and make spaces comparable.

We construct P as an equally sized held-out subset of the checkpoint population associated with each generator. For HR and G.pt, these held-out checkpoints come from independently trained models. P-diff is different: its checkpoint population consists of consecutive states from a single training trajectory, so P is necessarily drawn from that same trajectory. Thus, in each case, P represents the expected variation within the generator's native checkpoint population.

HR exhibits the clearest replication signature, showing that nearly half of the generated mass is absorbed by the training set, with a relative amount also being redundant. The effect is also detected in the function and representation space, although the amount is significantly less severe, suggesting that strong parameter-space replication does not imply a comparable loss of behavioural or representational diversity. The nearest-neighbour rank statistic gives the same ordering, in parameter space and smaller negative shifts in the other two spaces.

G.pt shows a weaker and strongly space-dependent effect. Parameter space gives modest positive absorption. In the function and representation space, signed absorption becomes negative. The resulting picture is therefore one of small-parameter-space attraction with a mild loss of coverage in the representation space.

P-diff exhibits a different geometry. Absorption is negative in all three spaces, ruling out the idea that the generated parameters are actually absorbed by the training data. Moreover, we see a negative ρ in the parameter space, indicating far less redundancy than in the counterfactual set.

We also observe that the rank statistic saturates at its sample-size-dependent upper bound in all three spaces because the held-out trajectory checkpoints are almost uniformly closer to the reference than generated checkpoints; this ceiling effect prevents Z_U from resolving further differences in the strength of separation. It is worth noting that, as stated by Zeng et al. (2026) this pathological observation is reasonable within the context of the training data being samples of the same network across training iterations, explaining the similarity saturation.

What is particularly unique about this network, is the high κ, meaning that while on all spaces the signal of μ and ρ doesn't suggest a pathological learning compared to the counterfactual set, it suggests a coverage of the training distribution significantly collapsed and is evidence of a highly bounded interpolation of the training data, with little to none generalisation, as it is suggested in the original analysis.

Overall, the measurements agree with the nearest-neighbour analysis of Zeng et al. (2026) while separating several effects that a single distance ranking cannot distinguish. HR combines strong parameter-space absorption; G.pt shows a weak parameter-space memorisation reported previously, but this does not translate into comparable set-level absorption in behaviour or representation space, and P-diff does not show a significant absorption from the individual training checkpoints while remaining internally dispersed relative to its unusually tight trajectory-derived counterfactual, which is evidence consistent with the bounded interpolation hypothesis.

Table 8: Cross-category affinity ablation for the pooled guidance experiment. $\Delta\mu_{3-1} = \mu_{w=3} - \mu_{w=1}$. Removing cross-category R–G affinities preserves most of the original positive ordering, whereas removing cross-category G–G affinities reverses it. Removing cross-category R–G affinities preserves approximately 82% of the original $w = 1$ to $w = 3$ increase. In contrast, removing cross-category G–G affinities reverses the ordering, as does removing both interaction types.

Intervention	$\mu_{w=1}$	$\mu_{w=3}$	$\Delta\mu_{3-1}$
Original geometry	+0.0168	+0.0246	+0.0078
Remove cross-category R–G	-0.1684	-0.1620	+0.0064
Remove cross-category G–G	+0.0626	+0.0488	-0.0138
Remove both	-0.1712	-0.1901	-0.0189

C Analysis of the 3D Guidance Discrepancy

The pooled guidance experiment in B.2 presents an apparent discrepancy between our seven-category analysis and the one reported by Pu et al. (2026); the magnitude reproduces the non-monotonic guidance ordering reported by, with

$$\mu_{w=1} = 0.0168, \qquad \mu_{w=3} = 0.0246,$$

whereas the nearest-neighbour statistic changes from

$$Z_U(w = 1) = -1.81 \qquad \text{to} \qquad Z_U(w = 3) = -1.66.$$

Thus, Z_U does not reproduce the increase in memorisation from weak to moderate guidance on our reduced evaluation population.

Z_U is computed globally, but each generated shape contributes only its nearest-training distance,

$$d_R(g) = \min_{r \in R} d(g, r).$$

The absence of the pooled $w = 1$ to $w = 3$ shift is explained by substantial heterogeneity across the seven categories. Figure 6 reports

$$\Delta Z_U = Z_U(w = 3) - Z_U(w = 1)$$

for each category. Negative values correspond to the direction reported by Pu et al. (2026). Rifle and gun move in this direction, with $\Delta Z_U = -1.009$ and -0.595, respectively, whereas figurine, toy, table lamp, and vase all move in the opposite direction. Consequently, their contributions largely cancel when the categories are pooled, producing $\Delta Z_U = +0.157$ on the complete seven-category bed.

This composition sensitivity also appears when the evaluated population is perturbed. Among the seven individual categories, only two reproduce the direction of the Z_U guidance effect, and only two of the seven leave-one-category-out populations do so. We therefore do not interpret the discrepancy as a difference in the implementation of Z_U, nor simply as reduced statistical power. Rather, the nearest-training-distance distribution itself responds differently under our restricted category mixture. The larger evaluation population of Pu et al. (2026) produces a net nearest-neighbour shift at moderate guidance, whereas this marginal shift is cancelled in our seven-category population.

The disappearance of the nearest-neighbour shift does not imply that the pooled geometry is unchanged. To identify the component responsible for the increase in μ, we perform controlled affinity ablations on the pooled $w = 1$ and $w = 3$ distance matrices and rescore each intervention using the same magnitude estimator used for the main results. We separately suppress cross-category affinities between reference and generated samples (R–G), between generated samples (G–G), and both simultaneously. The reference geometry and all within-category relationships are otherwise preserved.

The original geometry gives

$$\Delta\mu_{3-1} = +0.0078.$$

After suppressing cross-category R–G affinities, the difference remains positive at

$$\Delta\mu_{3-1} = +0.0064,$$

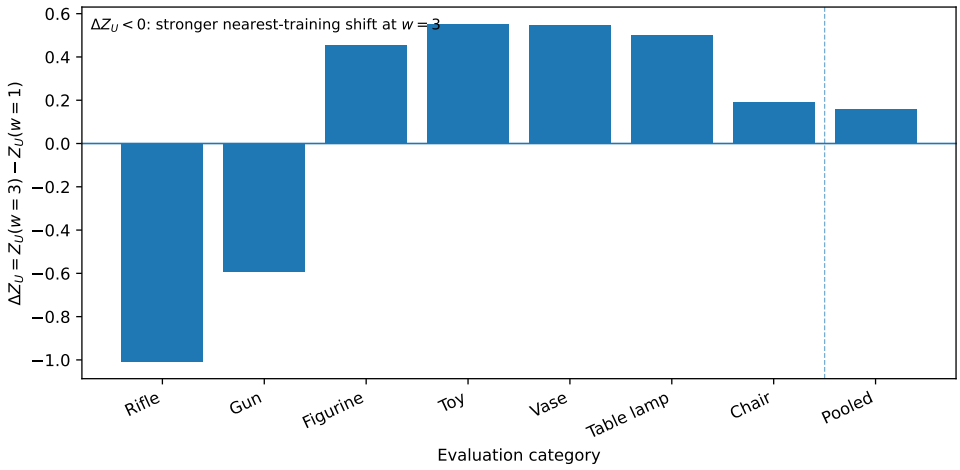

Figure 6: Change in the Pu-style nearest-neighbour statistic from $w = 1$ to $w = 3$ for the seven evaluated categories and their pooled population. Negative ΔZ_U corresponds to a stronger nearest-training shift at moderate guidance. Only rifle and gun move clearly in this direction; the remaining categories oppose it, causing the pooled nearest-neighbour response to cancel.

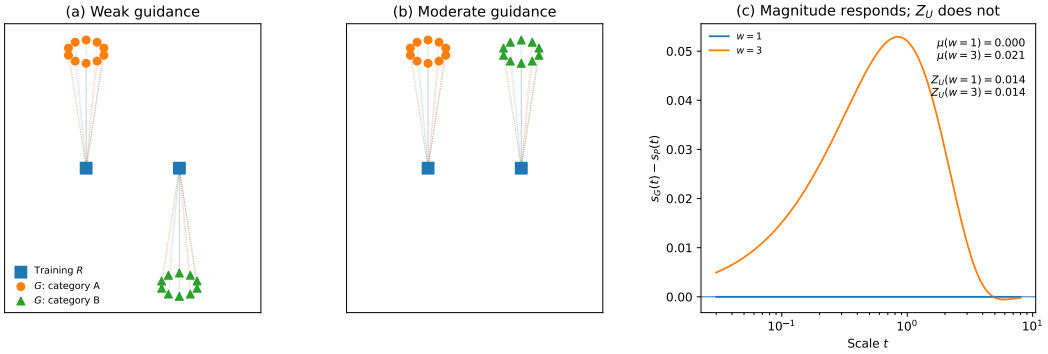

Figure 7: Illustrative construction showing how a guidance effect can be visible to magnitude while remaining invisible to a nearest-training statistic. In panels (a) and (b), the complete R–G geometry and the within-category G–G geometry are unchanged. The only modification is the cross-category organisation of the generated set. Consequently, Z_U is unchanged because every generated sample retains the same nearest-training distance, whereas the magnitude-based overlap profile $s_G(t) - s_P(t)$ changes because magnitude depends on the full geometry of G and $R \cup G$.

retaining approximately 82% of the original increase. In contrast, suppressing cross-category G–G affinities changes the difference to

$$\Delta\mu_{3-1} = -0.0138,$$

and removing both interaction types gives

$$\Delta\mu_{3-1} = -0.0189.$$

The moderate-guidance ordering therefore depends on the cross-category organisation of the generated distribution rather than primarily on cross-category proximity between generated and training shapes.

This distinction is directly relevant to the difference between the diagnostics. Z_U reduces every generated sample to the scalar $d_R(g)$ before comparing the generated and held-out populations. Relationships between generated samples, including cross-category G–G structure, are therefore absent from the statistic. Magnitude instead depends on $\mathrm{Mag}(G, t)$ and $\mathrm{Mag}(R \cup G, t)$ and consequently retains these relationships. On our reduced setting, the $w = 1$ to $w = 3$ effect is therefore present in

a component of the pooled geometry that contributes to μ but is discarded by the nearest-neighbour reduction.

The category analysis further supports this interpretation. The positive pooled change in μ is not reproduced by a simple aggregation of consistent within-category effects: only rifle individually shows a positive $w = 1$ to $w = 3$ change in μ, and the positive pooled ordering disappears in all seven leave-one-category-out evaluations. This means that the seven-category mixture contains an emergent cross-category geometric interaction that remains consistent with the moderate-guidance effect even though the corresponding marginal nearest-training shift is absent. This explains how the same guidance phenomenon can be detected by Z_U on the larger population studied by Pu et al. (2026), yet appear only through the richer set geometry measured by μ on our reduced population.

D POST-TRAINING OBJECTIVES

This appendix describes the post-training procedures used in Section 4.2. The two applications use different representation spaces, but follow the same principle: a geometric failure is first identified by the diagnostic analysis, a healthy operating range is measured from an appropriate counterfactual population, and the model is penalised only while its generations remain outside that range.

The notation in this appendix is local to the intervention and does not replace the diagnostic notation (R, G, P) from Section 3.2. In particular, the reference set R used to evaluate memorisation is never used by the post-training objectives below.

D.1 MEASURED INTERVENTION FRAMEWORK

Let z denote the representation of the current generated candidate being updated through the model. For each condition, such as a text prompt or an identity name, we maintain a finite bank B of generations produced under that same condition. We further use counterfactual H, which we expect to serve as a referent of what a healthy distribution looks like, to determine what range of the relevant geometric statistic should be considered normal.

Our post-training procedure distinguishes three roles: the current generation z, the condition-specific bank B that describes the geometry to be corrected, and the healthy population H used to calibrate the correction. Finally, we can add a fourth object, C, containing generations associated with other conditions. This allows the loss to detect a failure in which each condition escapes its original concentration, but several conditions collapse toward the same replacement region.

All interventions are combined with a retention objective,

$$\mathcal{L} = \mathcal{L}_{\text{retain}} + \mathcal{L}_{\text{intervene}}, \tag{13}$$

where $\mathcal{L}_{\text{retain}}$ preserves ordinary model behaviour. The intervention is one-sided: once the measured healthy range is reached, its contribution vanishes. No memorised reference image or identity target is required by any intervention.

D.2 DISTANCE-BASED REPULSION

We first use nearest-neighbour distance as a local surrogate for magnitude redundancy. Since magnitude is computed from $Z_{ij}(t) = \exp(-td_{ij})$, abnormally small distances produce the strongest similarities and therefore the strongest local redundancy in the magnitude calculation. Repelling only samples whose distances fall below the range observed in healthy counterfactuals directly weakens these terms without requiring a magnitude-system solve at every training step. For a candidate z and a condition-specific bank B, define its nearest-bank distance

$$d_{\min}(z; B) = \min_{b \in B} d(z, b). \tag{14}$$

A distance d_{floor} is measured from the corresponding counterfactual population. We then use

$$\mathcal{L}_{\text{dist}} = \lambda \left[\frac{(d_{\text{floor}} - d_{\min}(z; B))_+}{d_{\text{floor}}} \right]^2. \tag{15}$$

This objective acts only when the candidate lies closer to its own bank than the measured healthy floor. Once $d_{\min}(z; B) \geq d_{\text{floor}}$, the penalty is zero, so that the objective only removes the excess concentration identified relative to the counterfactual population.

This is the intervention used for the Stable Diffusion experiment in Section 4.2.1. It also serves as the first intervention in the face-generation study. The latter experiment exposed an important limitation: a nearest-distance objective specifies what a generation should move *away from*, but does not constrain where it moves *to*. Different treated conditions may therefore satisfy their individual distance floors while contracting toward a common replacement region.

D.3 MAGNITUDE-BASED REDUNDANCY CONTROL

The contraction observed after direct distance repulsion in 4.2.2 motivates replacing the nearest-neighbour statistic with a quantity that depends on the geometry of the full comparison set.

At a fixed scale $t^\star$, let $w_{t^\star}(z; S)$ denote the magnitude weight assigned to candidate z when magnitude is computed on $S \cup \{z\}$. Following Eq. 2, this is the component associated with z in

$$\left(Z_{S\cup\{z\}}(t^\star) + \epsilon I\right) w = \mathbf{1}, \tag{16}$$

where the small diagonal term ϵI is used for numerical stability.

A candidate that is highly redundant with S receives a smaller effective contribution, whereas a candidate that contributes more independent geometry receives a larger weight. This gives a per-generation quantity that can be optimised during post-training.

Within-condition redundancy. We first replace the distance limit from the previous approach with a magnitude-weight limit against the candidate's own condition-specific bank. Let B_n^{-i} denote the bank for condition n with the source element excluded. We define

$$\mathcal{L}_{\text{self}} = \lambda_{\text{self}} \left[\frac{\left(f_{\text{self}} - w_{t^\star}(z; B_n^{-i})\right)_+}{f_{\text{self}}} \right]^2. \tag{17}$$

The threshold f_{self} is measured from healthy counterfactual generations. As with the distance intervention, the term vanishes after the healthy range is reached. Unlike Eq. 15, however, the signal depends on redundancy with the complete bank rather than only the single nearest member.

D.4 Preventing Cross-Condition Collapse

Due to the results in 4.2.2, where generations seemed to collapse to a common identity, we introduce a comparison set C_{-n} containing recent generations other than n. We define f_{cross} from the healthy counterfactual population using the same reference-set size employed during training. For each healthy counterfactual candidate z, we draw a fresh set of embeddings from the counterfactual population and compute its magnitude weight $w_{t^\star}(z; S)$. We then set

$$f_{\text{cross}} = Q_{0.10}\left(\{w_{t^\star}(z; S)\}\right), \tag{18}$$

Thus, $w_{t^\star}(z; C_{-n}) < f_{\text{cross}}$ means that the candidate is more redundant with generations from other conditions than the lower tail observed under the counterfactual geometry.

The cross-condition penalty is

$$\mathcal{L}_{\text{cross}} = \lambda_{\text{cross}} \left[\frac{\left(f_{\text{cross}} - w_{t^\star}(z; C_{-n})\right)_+}{f_{\text{cross}}} \right]^2. \tag{19}$$

During training, C_{-n} contains the most recent embeddings generated from other than n. Matching this cardinality to the calibration sets allows the same healthy threshold to be used as an anti-collapse constraint, where one term repulses the generations from others' prompt-generated samples, while the other term limits how much this is done.

Combining Eqs. 17 and 19 gives

$$\mathcal{L}_{\text{mag}} = \mathcal{L}_{\text{self}} + \mathcal{L}_{\text{cross}}. \tag{20}$$

This constitutes the magnitude self+cross variant evaluated in Section 4.2.2.

D.5 Healthy-Population Plausibility

While previous terms successfully generate non-memorised generations, all generations can remain mutually separated yet drift together away from the region occupied by healthy generations. Therefore, we introduce a term to pull the generations to the expected manifold.

Using the same distribution employed for f_{cross}, we set

$$c_H = Q_{0.90}\left(\{w_{t^\star}(z; S)\}\right), \tag{21}$$

where each S is a set drawn from the counterfactual embeddings. Thus, f_{cross} and c_H are, respectively, the lower and upper healthy thresholds of the same magnitude-weight statistic.

During training, we compare the candidate with a healthy counterfactual set H. If $w_{t^\star}(z; H) > c_H$, the candidate contributes more independent geometry relative to the healthy population than is typical under the counterfactual distribution, indicating excessive drift away from that population. We therefore define

$$\mathcal{L}_{\text{plaus}} = \lambda_{\text{plaus}} \left[\frac{(w_{t^\star}(z; H) - c_H)_+}{1 - c_H} \right]^2 . \tag{22}$$

The term is inactive while the candidate remains within the healthy upper range and becomes active only when it is unusually distinct from the counterfactual population. It therefore constrains common population-level drift without using, or pulling the candidate toward, the memorised reference example.

The complete magnitude loss term is

$$\begin{aligned} \mathcal{L}_{\text{mag+plaus}} = \lambda_{\text{self}} & \left[\frac{\left(f_{\text{self}} - w_{t^\star}(z; B_n^{-i})\right)_+}{f_{\text{self}}} \right]^2 \\ & + \lambda_{\text{cross}} \left[\frac{(f_{\text{cross}} - w_{t^\star}(z; C_{-n}))_+}{f_{\text{cross}}} \right]^2 \\ & + \lambda_{\text{plaus}} \left[\frac{(w_{t^\star}(z; H) - c_H)_+}{1 - c_H} \right]^2 . \end{aligned} \tag{23}$$

The three objectives therefore form an incremental ablation: nearest-distance correction, magnitude-based within- and cross-condition control, and finally a healthy-population plausibility constraint that limits residual common drift.

Table 9: **Copy reduction after post-training.** Copies are defined by the reference SSCD > 0.5 criterion. Repulsion results are the mean $\pm$ standard deviation across three independent 400-step runs; controls use two matched retention-only runs.

Prompt set	Untouched	+ Repulsion	Control
Treated (24×8)	85	36.3 ± 6.0	84.5 ± 9.2
Held out (12×8)	34	9.3 ± 2.3	25.5 ± 4.9

E Post-Training Implementation Details

This appendix describes the specific post-training and instantiation procedures used in Section 4.2 following the notation of D.

E.1 Stable Diffusion Instantiation

For the Stable Diffusion experiment in 4.2.1, we mitigate memorisation of prompts known to generate copies. For each prompt, we generate a bank of $k = 8$ latent representations using the original model. During training, one bank member is lightly noisified and denoised to produce an estimate $\hat{x}_0$, which serves as the current candidate z. Its source element is excluded from the comparison bank, giving

$$d_{\min} = \min_{j \neq i} \|\hat{x}_0 - x_j\|_2 . \tag{24}$$

Excluding the source prevents self-reconstruction from contributing to the distance statistic.

The healthy population is formed from prompts that do not exhibit detected memorisation. We measure

$$d_{\text{floor}} = 58.74, \tag{25}$$

the 10th percentile of their leave-one-out minimum latent distance. The post-training intervention is therefore

$$\mathcal{L}_{\text{rep}} = 0.25 \left[\frac{(58.74 - d_{\min})_+}{58.74} \right]^2 . \tag{26}$$

The intervention is combined with a retention loss consisting of noise-prediction MSE on cached ordinary generations and their captions:

$$\mathcal{L} = \mathcal{L}_{\text{retain}} + \mathcal{L}_{\text{rep}}. \tag{27}$$

We train a rank-8 LoRA adapter on the UNet attention projections while keeping the remaining model parameters frozen.

Of the 36 prompts detected by the reference SSCD criterion, 24 are used as treated prompts and 12 are reserved completely for held-out evaluation. We train three independent 400-step intervention runs.

All post-training effects reported in Section 4.2.1 are evaluated using fresh generations. The external evaluation uses SSCD, DINOv2, and CLIP.

We also evaluate ordinary, non-memorising prompts to check whether the intervention broadly degrades generation. We measure seed-fair coverage, within-prompt diversity, and CLIP caption adherence, together with matched retention-only controls. As shown in Table 10, ordinary generation is largely preserved, with stable coverage and caption adherence and only a small additional reduction in diversity attributable to the repulsion objective.

E.2 Face-Generation Instantiation

For face generation, the prompts contain real names, and we aim for the model not to generate the person associated with that name. The healthy counterfactual for each name is obtained from its one-character perturbed name using the same prompt template.

Table 10: **Preservation of ordinary generation after Stable Diffusion post-training.** Seed-fair κ_{ord} measures the coverage change on ordinary prompts. Diversity is the mean within-prompt pairwise-distance ratio relative to the untouched model, and caption adherence is measured by CLIP relative to the untouched model. The control consists of matched retention-only post-training runs.

Measure	+ Repulsion	Control
Seed-fair κ_{ord}	0.015 ± 0.002	0.015 ± 0.005
Within-prompt diversity	0.919 ± 0.003	0.938 ± 0.017
CLIP adherence change	$+0.003$–$+0.005$ across trained variants	

The loss is calculated in the identity-embedding space. During each engaged training step, a cached generation from the treated name is noised and denoised, decoded to image space, detected and aligned using an SCRFD face detector, and embedded with an AdaFace network. The resulting identity embedding is the candidate z.

Distance baseline. The first strategy uses Eq. 15. The bank contains the other cached generations of the same name and the distance constrain is measured from perturbed-name generations. The measured value is

$$d_{\text{floor}} = 1.0015. \tag{28}$$

This defines the distance-repulsion variant reported in Section 4.2.2.

Magnitude scale and thresholds. The magnitude-based objectives require a per-candidate weight at a single resolution, rather than the multiscale integrals used by the diagnostic framework. We therefore select one training scale $t^\star$ before post-training. Using the estimator's fixed 39-scale grid, we compute leave-one-out magnitude weights for denoised memorising generations and for their healthy perturbed-name counterparts. At each scale, we compare the 10th percentiles of the two weight distributions and select the scale at which has the largest separation:

$$t^\star = \arg\max_t \left[Q_{0.10}(w_{\text{healthy}}(t)) - Q_{0.10}(w_{\text{mem}}(t)) \right]. \tag{29}$$

This gives $t^\star = 2.669$. The same $t^\star$ is then used for every magnitude-based training term.

The thresholds are calibrated from the perturbed-name population at this scale. For the within-name term, we set

$$f_{\text{self}} = 0.7728, \tag{30}$$

The 10th percentile of healthy leave-one-out weights relative to the candidate's own name bank.

For the cross-name and plausibility terms, we construct a second healthy calibration distribution by measuring each healthy candidate against a random 64-member set of clean perturbed-name embeddings. We set

$$f_{\text{cross}} = 0.3533 \tag{31}$$

to the 10th percentile of this distribution, and set c_H to its 90th percentile. Thus, f_{cross} defines the minimum healthy level of separation from a crowd of other identities, whereas c_H defines the upper healthy limit for separation from the counterfactual population. All scales and thresholds are fixed before post-training.

The magnitude self+cross model uses

$$\mathcal{L}_{\text{intervene}} = \mathcal{L}_{\text{self}} + \mathcal{L}_{\text{cross}}, \tag{32}$$

whereas the final plausibility-constrained model additionally includes $\mathcal{L}_{\text{plaus}}$ from Eq. 22. We use

$$\lambda_{\text{self}} = \lambda_{\text{cross}} = \lambda_{\text{plaus}} = 0.25, \tag{33}$$

with $\lambda_{\text{plaus}} = 0$ for the self+cross variant.

The same-name bank contains the remaining cached generations of the current name. The cross-condition set C_{-n} is a rolling buffer of recent embeddings from other treated names, using the most recent 64 once the buffer contains at least 16 entries. The plausibility set H is a fixed 64-member draw from the clean perturbed-name population. Matching these reference-set sizes to the calibration procedure allows the corresponding healthy thresholds to be used directly during training.

A diagonal stabiliser $\epsilon = 10^{-6}$ is added when solving for the per-candidate magnitude weights.

Table 11: **Preservation of ordinary generation after face-generation post-training.** Results use 60 name-free prompts and 480 generations per model with matched seeds. Diversity is the mean within-prompt CLIP pairwise-distance ratio relative to the untouched model, and adherence is the CLIP similarity change relative to the untouched model.

Variant	Diversity	Adherence change
Untouched	1.000	0.000
Retention control	0.898	+0.003
Distance repulsion	0.955	+0.001
Magnitude + plausibility	0.936	+0.002

Training population and model. We begin from a population of 240 real names and their corresponding one-character perturbations. The 24 strongest memorising names are used for post-training, while the next 24 are held out entirely from optimisation.

All face interventions train a rank-8 LoRA adapter on the UNet attention projections while the remaining diffusion model, face detector, and identity encoder remain frozen. Each training step combines an ordinary retention example with an intervention example. The adapter is trained for 400 steps.

E.3 Independent Evaluation and Controls

For Stable Diffusion, the post-training loss is defined in the model's latent space, whereas the principal memorisation and geometric results are measured independently in SSCD, DINOv2, and CLIP.

For face generation, AdaFace supplies the differentiable identity representation used during optimisation. All paper-facing identity similarity and primary magnitude results are additionally evaluated with ArcFace, a separate identity encoder that never appears in the training loss. Evaluation generations use seeds disjoint from the cached generations used for post-training.

Matched retention-only controls in Table 11 use the same LoRA post-training procedure without the intervention objective, and find that the post-training has only a modest effect on ordinary generation.

Across both settings, the reference example used to establish the presence of memorisation remains external to the post-training objective. The intervention, therefore, uses only the model's own generated geometry and statistics measured from healthy counterfactual populations.