FROM GIST TO VERBATIM: THE CONTROLLABLE PATH OF ONE IMAGE THROUGH A DIFFUSION PRIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern generative models are typically studied for their ability to generalize, while memorization is treated as a failure mode. We invert this perspective and ask how a generative model can recall a single datum. We frame memorization as a control problem: a learned conditioning signal steers a frozen diffusion model toward a target image, and we follow the formation of this memory throughout training. Before converging to exact reproduction (*verbatim*), memory passes through a regime of *abstraction*: samples preserve the target's semantic content (*gist*) while remaining diverse in appearance, leaving predictable detail to the generative prior. This semantic organization emerges despite the absence of labels, captions, or any explicit semantic objective. In the abstraction regime, the learned memory approaches ordinary text conditioning and supports editing, one-shot recognition, and retrieval more effectively than in the near-perfect reconstruction regime. These results suggest that memorization can serve as a probe of the relevance structure encoded by a generative model: what must be stored about a datum, what can be reconstructed from prior knowledge, and how learned memories interpolate between reconstructive recall and literal data storage.

1 INTRODUCTION

A scientific textbook condensing decades of research can occupy less disk space than a blurry vacation photograph. This fact may feel alien from a human perspective, because of the fundamental difference between data storage and memory. A disk stores any object literally, without a notion of relevance or an interpretative prior. Memory, instead, does not preserve an observation in its entirety: we make a choice about what to retain, and reconstruct the rest at recall from prior knowledge. How does the interaction between new information and prior shape a memory as it forms?

This question sits at the intersection of two longstanding lines of work. The reconstructive view of memory distinguishes the stored trace from the prior knowledge used to rebuild an experience (Bartlett, 1932; Schacter & Addis, 2007; Spens & Burgess, 2024). Coding and minimum-description-length perspectives, similarly, make description length relative to a given model: once the model is fixed, only the unexplained residual needs to be encoded (Rissanen, 1978; Grünwald & Roos, 2019). Both views suggest that memory should be understood relative to what is already known. What remains unanswered, however, is how the information in a single datum is ordered by relevance: what is committed to memory first, and what can initially remain implicit in the prior.

Modern generative models provide us with an experimental playground in which we can explore this question. Recent work on memorization in diffusion models has investigated when training examples are reproduced and how such replication can be detected or mitigated (Carlini et al., 2023; Somepalli et al., 2023). In parallel, multiple personalization methods have been proposed to adapt pretrained diffusion models to new subjects or concepts so that they can be regenerated in novel contexts (Gal et al., 2023; Ruiz et al., 2023; Kumari et al., 2023). In this work, we ask a different question: if a strong generative prior is deliberately made to **memorize one image**, what does it learn along the way? How does memory formation unfold? We freeze the prior, train only a conditioning signal from a single image, and follow the resulting memory throughout training.

The resulting trajectory is surprising yet familiar. Before approaching exact reconstruction, the memory passes through an *abstraction regime*: generated samples incorporate the target's semantic content (*gist*) while retaining diversity in pose, appearance, background, and other instance-specific

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.

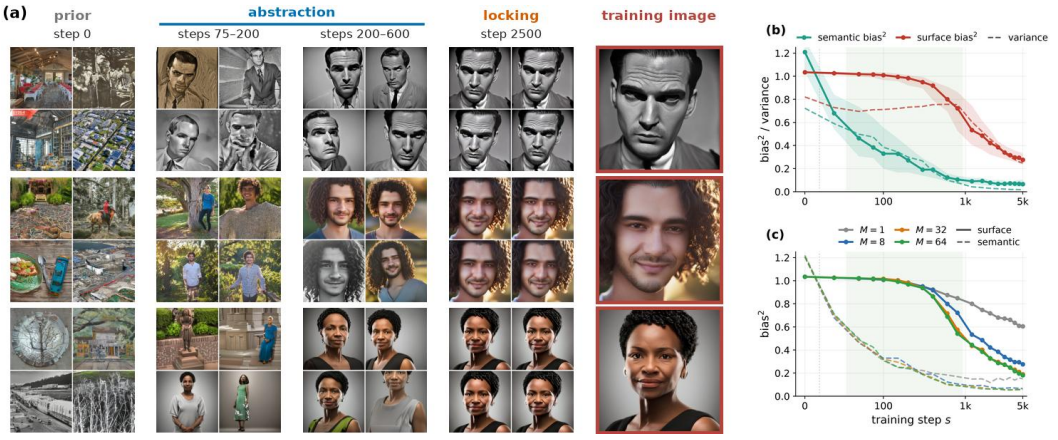

Figure 1: **One image, two regimes: abstraction before locking. (a)** Three memories (M=8 vectors each), each trained on a single image (rightmost, red frame) and used to condition a frozen diffusion model. Each 2×2 block shows four samples at different input noise realizations. At initialization, samples follow the prior (first column). During an extended *abstraction* regime, the memory first generates other subjects sharing the target's style (second column), then the target's identity with varying pose, clothing and lighting (third column). By step 2500, *locking* sets in and the training image is generated regardless of the input noise realization (fourth column). **(b)** The same trajectory across 30 single-image memories (10 animals, 10 vehicles, 10 faces; median±IQR). Solid: squared distance between the mean sample and the target (bias^2), measuring progress toward it. Dashed: dispersion of the samples around their mean (variance), measuring remaining variability. Semantic distance (DINOv2 embedding) collapses within $\sim 10^2$ steps, while surface distance (VAE latent) collapses an order of magnitude later. Green band: the abstraction window, between the median semantic and surface half-decay steps (Appendix D.1). **(c)** bias^2 by memory size (solid: surface, dashed: semantic). At fixed training budget a larger memory leaves the timescales unchanged but lowers the final surface residual. M=1 does not complete locking within the budget.

details. Only after further training the memory enters a *locking phase*, in which recall becomes *verbatim*, i.e., the target image is fully reproduced. No labels, captions, or semantic criteria are provided during memory formation. Abstraction arises from the preferential acquisition of semantic structure already represented by the prior. Memory recall becomes a probe of relevance: what the model stores, what it leaves to prior knowledge, and in what order. Our contributions are:

- We identify and quantify an *abstraction regime* in single-image memorization, Fig. 1, in which semantic content is acquired substantially before instance-specific appearance. A similar phenomenology can also be traced in different settings and data modalities, Fig. 4.
- We formulate memorization as prior-mediated control and provide a minimal theoretical framework in which learning depends jointly on the residual information in the target and the susceptibility of the generative prior. We then retrace two qualitative predictions on the abstraction regime in our experimental setting, Fig. 2.
- We show that abstraction is a distinct functional regime, Fig. 3, by evaluating three capabilities of the learned memory: *editability* (integrating new information at recall), *one-shot recognition* (recognizing unseen realizations of the same subject), and *text–memory retrieval* (accessing memories through textual cues). These capabilities emerge during abstraction and deteriorate as reconstruction approaches the verbatim target, revealing a trade-off between recall precision and memory reusability.

2 Memorizing through a diffusion prior

Memory as a conditioning signal. Consider a pretrained diffusion model with data distribution p_0 and score $\boldsymbol{s}_0(\boldsymbol{x}, t)$. We study how such a model can be made to recall a single target image $\boldsymbol{x}^\star$. Importantly, we aim to achieve this goal not by fine-tuning the parameters of the generative model, which are kept frozen at their pretraining values, but by learning a compact conditioning signal $\boldsymbol{C}$.

In the following, we focus on the training trajectory of $\boldsymbol{C}$, which we regard as the *memorization* process, and we refer to $\boldsymbol{C}$ as the *memory*. During generation, the memory acts as a *control*: it steers the reverse diffusion toward $\boldsymbol{x}^\star$, while relying on the structure encoded by the pretrained prior.

Residual to a prior. The target has negative log-likelihood $-\log p_0(\boldsymbol{x}^\star)$ under the pretrained generative model, which can be interpreted as its ideal codelength under the prior (Rissanen, 1978; Grünwald & Roos, 2019).[1] In this view, the memory $\boldsymbol{C}$ need not encode the datum from scratch: it should only supply what is not already explained by the model. Learning can therefore be viewed as progressively reducing a residual relative to the prior. The central question of this work is how this residual is resolved along the training trajectory: which properties of the datum are *relevant* under the prior and stored first, and which are initially left for the model to guess.

Training objective. We learn the control signal $\boldsymbol{C}$ using the diffusion model's denoising objective on a single target image $\boldsymbol{x}^\star$, keeping all pretrained parameters frozen. Under Gaussian noising, $\boldsymbol{x}_t^\star = \alpha_t \boldsymbol{x}^\star + \sigma_t \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the noised target's exact score is

$$\boldsymbol{s}^\star(\boldsymbol{x}_t^\star, t) = -\sigma_t^{-2}(\boldsymbol{x}_t^\star - \alpha_t \boldsymbol{x}^\star). \tag{1}$$

Let $\boldsymbol{s}_{\boldsymbol{C}}(\boldsymbol{x}, t)$ denote the frozen model's score under memory conditioning $\boldsymbol{C}$. We model *memorization* by minimizing, with respect to $\boldsymbol{C}$ only,

$$\mathcal{L}(\boldsymbol{C}) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{s}_{\boldsymbol{C}}(\boldsymbol{x}_t^\star, t) - \boldsymbol{s}^\star(\boldsymbol{x}_t^\star, t)\|^2\right]. \tag{2}$$

Instantiation. We instantiate this construction primarily in Stable Diffusion 2.1 (SD2.1) (Rombach et al., 2022). This is a latent diffusion model: a pretrained VAE maps an image to a latent representation, and a UNet learns to denoise noisy versions of that latent. Thus, we apply the denoising objective (equation 2) to the target's latent $\boldsymbol{z}^\star = \mathrm{VAE}(\boldsymbol{x}^\star)$, using the model's native v-prediction parameterization and uniform time sampling.

SD2.1 is text-conditioned. A text encoder maps a caption to a sequence of token embeddings, which condition the UNet through cross-attention: image features provide queries, while the text embeddings provide keys and values. When no caption is supplied, the model instead receives the embedding of the empty caption $\boldsymbol{c}_\varnothing$. We take this empty-caption model as our generative prior p_0.

We represent the memory by M learnable vectors $\boldsymbol{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\}$, each of dimension $D = 1024$. These vectors enter a second cross-attention pathway parallel to the text pathway, using the same frozen projection parameters. Thus, at each UNet cross-attention layer,

$$\mathrm{Attn}_{\mathrm{tot}}(\boldsymbol{h}) = \mathrm{Attn}_{\mathrm{text}}(\boldsymbol{h}, \boldsymbol{c}_\varnothing) + \mathrm{Attn}_{\mathrm{mem}}(\boldsymbol{h}, \boldsymbol{C}). \tag{3}$$

Only $\boldsymbol{C}$ is optimized; all parameters of the VAE, UNet, text encoder, and attention projections remain frozen. We initialize $\boldsymbol{C}$ with small random values so that the memory contribution is initially negligible and the model starts close to the prior p_0.

Model-generated vs. external targets. Unless stated otherwise, our main experiments use targets generated by SD2.1 from a curated set of captions. We make this choice for two reasons. First, targets within the model's generative repertoire should have smaller initial residuals and permit lower reconstruction bias in the asymptotic training regime of $\boldsymbol{C}$. Second, each target comes with a known semantic reference, the caption it was generated from, which we use in subsequent analyses. Crucially, this caption is not provided during memorization: $\boldsymbol{C}$ is learned from the target image alone. In Appendix E.1, we repeat our experiments on external data.

3 From gist to verbatim

We first ask what a memory acquires, and in which order. As displayed in Fig. 1 (a), following a single-image memory throughout training (minimizing equation 2) reveals a consistent trajectory.

Abstraction then Locking. At initialization, samples reproduce the pretrained prior. After relatively few updates, generations become semantically aligned with the target while remaining diverse in pose, appearance, background, and other instance-specific details. With further training, and a

[1] Strictly, this coding interpretation assumes a discretized or finite-precision representation.

growing norm of the memory vectors (see Appendix E.2), this variability progressively disappears until different noise realizations converge toward the same image. We call the intermediate phase the *abstraction regime*: the memory has acquired the target's semantic content while its particular appearance remains substantially unconstrained. This semantic organization is not imposed by the memory-training objective, which provides neither labels nor semantic supervision. Abstraction is therefore not the emergence of semantics from scratch, but a resonance between datum features and semantic structure already represented by the prior. We call the later training stage the *locking phase*, in which realization-specific spatial and surface details are progressively fixed to the target.

Semantics vs Surface. We quantify this transition in Fig. 1 (b). At each checkpoint, we generate multiple samples from independent initial noises and compare them with the target in DINOv2 (Oquab et al., 2024) (semantic) and VAE (surface) representation spaces. We measure two quantities: the squared distance between the mean sample and the target representations (*bias*^{2}), and the dispersion of individual samples around their mean (*variance*). Both are normalized by the squared norm of the target representation; exact definitions are given in Appendix D.1.

The two types of information are acquired on markedly different timescales. The median ratio of surface to semantic half-decay steps, i.e., the step at which each bias^2 curve has completed half of its total decrease, is 21.4 (Fig. 1b). Semantic bias drops rapidly and is largely removed during the early phase of training, whereas surface bias remains high over the same interval and decreases only later, in the locking phase. In Appendix E.1, we show that the same phenomenology extends to external targets.

Impact of memory size. Memory size, M, primarily affects the locking phase. As shown in Fig. 1(c), increasing the number of learned vectors lowers the final surface error, but the early semantic timescale is largely unaffected. In particular, a single-vector memory ($M = 1$) acquires semantic content on approximately the same schedule as larger memories but does not approach exact reconstruction. Abstraction is therefore not simply a consequence of insufficient representational capacity: even memories that ultimately reproduce the target pass through a phase in which semantic content has been acquired while finer details remain unresolved.

4 A FRAMEWORK FOR ABSTRACTION AS PRIOR-MEDIATED CONTROL

To organize these observations, we introduce a minimal mathematical model that captures conditions for an abstraction regime to emerge. Let $\boldsymbol{\phi} : \mathbb{R}^d \to \{0,1\}^F$ be a collection of measurable binary features capturing properties represented by the generative model. These features may range from high-level semantic attributes (e.g., "a face") to spatially localized geometric details (e.g., "a particular pose at a particular image location"). We model the effect of a learned memory as an exponential tilt of the pretrained generative distribution,

$$p_{\boldsymbol{u}}(\boldsymbol{x}) = \frac{e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{x})}}{Z(\boldsymbol{u})} p_0(\boldsymbol{x}), \qquad Z(\boldsymbol{u}) = \mathbb{E}_{p_0}\left[e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{X})}\right]. \tag{4}$$

Here, $\boldsymbol{u} \in \mathbb{R}^F$ is the theoretical counterpart of the learned memory $\boldsymbol{C}$. While $\boldsymbol{C}$ is represented by learned conditioning vectors in SD2.1, $\boldsymbol{u}$ directly weights the abstract feature coordinates $\boldsymbol{\phi}$, determining which properties of the target are emphasized. The tilted score decomposes as $\boldsymbol{s}_{\boldsymbol{u}}(\boldsymbol{x}, t) = \boldsymbol{s}_0(\boldsymbol{x}, t) + \delta \boldsymbol{s}_{\boldsymbol{u}}(\boldsymbol{x}, t)$, with

$$\delta \boldsymbol{s}_{\boldsymbol{u}}(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \log \mathbb{E}_0\left[e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{X}_0)} \,\middle|\, \boldsymbol{X}_t = \boldsymbol{x}\right], \tag{5}$$

where $\mathbb{E}_0[\cdot \mid \boldsymbol{X}_t = \boldsymbol{x}]$ denotes expectation under the posterior induced by p_0 and the forward noising process. The correction $\delta \boldsymbol{s}_{\boldsymbol{u}}$ is a state- and time-dependent control field: at each noise level, it evaluates under the pretrained model's own posterior how far the demanded properties are from being realized, and pushes the dynamics accordingly.

To study how the memory is acquired, define the current feature residual

$$\boldsymbol{r}_{\boldsymbol{\phi}}(\boldsymbol{u}) = \boldsymbol{\phi}(\boldsymbol{x}^\star) - \mathbb{E}_{p_{\boldsymbol{u}}}\left[\boldsymbol{\phi}(\boldsymbol{X})\right]. \tag{6}$$

In the appropriate time scaling, the denoising objective becomes equivalent to minimizing the target negative log-likelihood, and the gradient-flow dynamics for the control $\boldsymbol{u}$ gives

$$\frac{d\boldsymbol{u}}{d\tau} = -\eta \nabla_{\boldsymbol{u}} \mathcal{L} = \eta\, \boldsymbol{r}_{\boldsymbol{\phi}}(\boldsymbol{u}), \qquad \frac{d\boldsymbol{r}_{\boldsymbol{\phi}}}{d\tau} = -\eta \boldsymbol{\Sigma}_{\boldsymbol{u}} \boldsymbol{r}_{\boldsymbol{\phi}}, \tag{7}$$

where the feature covariance $\mathbf{\Sigma}_{\boldsymbol{u}} = \mathrm{Cov}_{p_{\boldsymbol{u}}}[\boldsymbol{\phi}(\boldsymbol{X}_0)]$ under the tilted model is a linear *susceptibility*, since $\partial_{\boldsymbol{u}} \mathbb{E}_{p_{\boldsymbol{u}}}[\boldsymbol{\phi}] = \mathbf{\Sigma}_{\boldsymbol{u}}$: feature-space directions with larger eigenvalues respond more strongly to infinitesimal changes in the tilt parameters $\boldsymbol{u}$. The memorization dynamics in equation 7 are governed by two ingredients: the residual $\boldsymbol{r}_{\boldsymbol{\phi}}$, specifying which properties of the target are not yet reproduced, and the *susceptibility* $\mathbf{\Sigma}_{\boldsymbol{u}}$, determining which feature-space directions can change most readily.

Considering a simple binary feature helps the intuition. If a feature is present with probability p under the current tilted model, its susceptibility is $\mathrm{Var}(\phi) = p(1-p)$. A highly specific event ("a pose at a location") has $p \ll 1$, and hence low susceptibility. A broader property ("a face") can have substantially larger susceptibility and therefore be acquired more readily. In the multivariate setting, correlations add another source of susceptibility: semantic properties often participate in coherent combinations, while highly localized details are typically less strongly coupled to other features. This is consistent with Fig. 1, where identity and style emerge while pose and clothing still vary.

This framework yields two qualitative predictions. (i) Early in memorization, the tilt is weak, so the susceptibility structure remains close to that of the pretrained prior. Through pretraining, text conditioning aligns with high-susceptibility semantic directions encoded in the prior. Yet captions do not fully constrain generated images: different initial noises produce different realizations consistent with the same text. If an abstract memory acts as a similarly weak semantic constraint, it should preserve some of this noise-dependent prior variability. (ii) As the fast semantic residuals are removed, learning shifts toward slower, low-susceptibility features. Exact reconstruction must capture these finer constraints, including those that specify and control the spatial realization of the target. In our implementation, this should appear as increasingly position-dependent addressing.

5 Analysis of the abstraction regime

Section 4 suggests two signatures of the abstraction and locking phases. First, an abstract memory should preserve variation inherited from the prior, yielding caption-like behavior during this phase. Second, as the memory fixes spatial details during locking, its influence should become more position-specific. In the following, we test these predictions.

5.1 Generalization through the prior

To test whether memory preserves variation inherited from the prior, we generate a collection of images from a fixed caption using different initial noise realizations, then train a single memory jointly on N of these images. For $N > 1$, this becomes a form of prompt inversion. We evaluate generalization on held-out initial noises, comparing, for each noise, the images generated under the original caption and the learned memory. The same test remains meaningful at $N = 1$: systematic correspondence across held-out noises indicates that the memory recovers, through the prior, variation absent from its single training image.

Paired advantage. We measure this correspondence using DINOv2 distance between memory- and caption-conditioned samples. For each held-out initial noise, we compare the distance between the two samples generated from the same noise with the average distance obtained by pairing the memory sample with caption-conditioned samples generated from different noises. We call the unpaired-minus-paired distance the *paired advantage*, normalized by the mean pairwise distance. Further details on the metric formulation are provided in Appendix D.2, Equation 31.

At initialization, memory conditioning is negligible and the prior alone differs substantially from caption conditioning, so sharing the same initial noise provides little correspondence and the advantage is near zero. During abstraction, a positive advantage indicates that memory and caption conditioning steer the prior similarly enough for the same initial noise to produce corresponding realizations. As exact recall is approached, memory conditioning suppresses this noise dependence, causing the advantage to return toward zero. Details in Appendix D.2.

Generalization from a single image. Figure 2 (a-b) shows two examples and the aggregate result over twelve captions. At initialization, memory samples are unrelated to the caption-conditioned references. Near the paired advantage peak, samples tend to follow properties of their own paired reference, including pose, framing, and color, while remaining distinct from the training image.

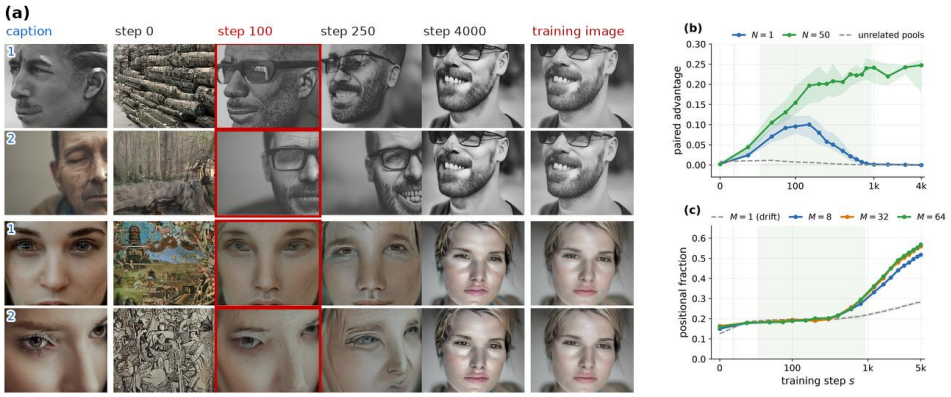

Figure 2: **Generalization through the prior. (a)** Each row corresponds to a caption–noise pair. For each of two captions, we show two input-noise realizations (1-2). Columns: the caption-conditioned sample; memory-conditioned samples (N=1) at initialization, near the peak of the paired advantage (step 100, red frame), later in training (step 250) and after locking (step 4000); the training image. The input noise is held fixed along each row. For each caption, the two realizations shown are those with the largest paired advantage at the peak. **(b)** Paired advantage over training for twelve captions describing faces, one memory per caption at N=1 (blue) and N=50 (green); median, bands: IQR. Dashed grey: the same samples scored against pools from unrelated captions at the same input noise. **(c)** Positional fraction f over training, mean over 30 memories per memory size (10 each of animals, vehicles and faces; bands: s.e.m.). Dashed: the M=1 control, which cannot directly learn position-dependent addressing.

Later in training, some of this variation remains, but the samples increasingly acquire the target's identity, palette, and background. This intermediate form of *biased generalization* (Garnier-Brun et al., 2026; Rodriguez et al., 2026) is consistent with instance-specific constraints strengthening after the semantic content has been acquired. With further training, dependence on the initial noise disappears as the outputs converge toward the memorized image, and the paired advantage returns to zero.

At the peak, the paired caption reference is the nearest neighbour for about 20% of memory-conditioned samples (chance level is 2% in this experiment). The effect is therefore substantial but not uniform across noise realizations, showing a restriction of the invariance supported by the caption-conditioned prior. The same samples show no advantage when compared with images from unrelated captions, while memories trained on $N = 50$ images from the same caption reach roughly twice the peak advantage and preserve the correspondence later in training.

5.2 From content dependence to positional specificity

The second prediction concerns how the memory acts within the network. Memory tokens influence generation through cross-attention, where they are projected into keys and values. Query–key interactions determine *where* each token contributes, while values determine *what* information it supplies. We therefore examine how memory's addressing changes during training.

Content and position in the attention logits. We want to distinguish whether the memory's addressing depends on the image being processed or instead becomes tied to fixed spatial locations. We examine the pre-softmax interaction between image queries and memory keys. We use the cosine similarity between image queries and memory keys, which factors out the large change in memory norm during training and remains informative even for $M = 1$, where the softmax attention weight is identically one.

Using a fixed set of held-out images, we decompose the variability of these scores by the law of total variance. The first component, V_{cont}, measures how much the score varies across images at fixed spatial positions. The second, V_{pos}, measures the variation across positions that remains after averaging over images. We report the *positional fraction*, f, the fraction of the total variance accounted for by V_{pos}. Small values therefore indicate predominantly content-dependent addressing, whereas large values indicate increasing positional specificity. Further details are given in Appendix D.3.

Addressing becomes position-specific during locking. Figure 2 (c) shows a clear transition. Throughout the abstraction regime, the positional fraction remains low and approximately constant. As locking sets in, it rises sharply for multi-token memories, indicating that a spatial pattern shared across images accounts for an increasing fraction of the addressing variability.

The $M = 1$ case provides a useful control. With a single memory token, the softmax attention weight is always one, so positional attention cannot be learned directly. The shared memory vector is still updated through the value pathway, which can indirectly change the query–key logits and produce a gradual positional drift. This can support good reconstruction through the frozen prior, but produces only the gradual change seen for $M = 1$. The sharp increase in positional specificity appears only for $M > 1$ as memorization approaches the verbatim regime.

6 THE FIDELITY–FUNCTIONALITY TRADE-OFF

Memory theories have long distinguished gist-like from verbatim representations, and behavioral work suggests that generalization or accessibility can come at the cost of precision (Brainerd & Reyna, 2002; Berens et al., 2020). We now ask whether a related trade-off emerges spontaneously during the formation of our memory. In particular, does increasing fidelity to the stored instance necessarily improve what the memory can do?

We probe three complementary forms of functionality. *Editability* asks whether the memory can be integrated with new information supplied at recall. *One-shot recognition* asks whether it retains useful invariance and can provide a good explanation across unseen realizations of the same subject. *Text–memory retrieval* asks whether it can be accessed through a textual cue. Across all three probes, functionality emerges during abstraction and subsequently deteriorates as reconstruction approaches the verbatim target (Fig. 3), even as recall fidelity continues to improve.

Editability. We follow a protocol analogous to diffusion-based editing, which starts from a partially noised input and lets the prior complete the denoising process (Meng et al., 2022). For each target x_{orig} and edited counterpart x_{edit} (generated with FLUX (Labs et al., 2025); see Appendix D.4), we partially noise the edit and complete denoising under the checkpointed memory, following a U-turn protocol. We then measure the resulting sample's DINOv2 cosine distance from the edit, D_{edit}, and from the stored image, D_{orig}. During abstraction, changes in pose, expression, accessories, and scene remain visible in the output. As memorization progresses, D_{edit} increases while D_{orig} decreases: the same memory that reconstructs its target more faithfully becomes progressively less compatible with the external edit (Fig. 3a). The same transition occurs when the new information to be integrated is an image of a different person: an abstract memory preserves substantial information from the input, whereas a verbatim memory increasingly overrides it with its stored subject. Sampling, noising, and averaging details are given in Appendix D.4.

One-shot recognition. Conditional denoising error can itself be used as a generative classifier, both in zero-shot (Li et al., 2023) and in few-shot (Chattopadhyay et al., 2024) settings. Here, we ask whether our trained memories transiently acquire enough invariance to recognize unseen realizations of their corresponding subject. We train one memory for each of a set of distinct identities (human faces), and construct a held-out query set showing the same identities in different poses and scenes. At each checkpoint, we evaluate the denoising loss in equation 2 for each query under every candidate memory, and assign the query to the identity whose memory gives the lowest loss. Recognition accuracy rises rapidly during abstraction: top-3 accuracy peaks above 0.9, with top-1 accuracy following the same non-monotonic trajectory, before both deteriorate as locking sets in (Fig. 3b). Thus, while reconstruction of the training image continues to improve, the memory progressively loses the invariance required to recognize the same subject under new appearances.

Text–memory retrieval. While text never enters memory formation, we can still ask whether a memory trained from an image alone becomes accessible through semantic cues, and how this accessibility changes as recall becomes more precise. For each identity, we generate a single training image from a detailed caption and train a corresponding memory. At evaluation time, we query the memory with three textual descriptions of increasing specificity: a category label plus one distinguishing attribute, a brief description, and a detailed description of the image (see Appendix D.6). At each checkpoint, we mean-pool the M conditioning vectors and compare the resulting representation with the frozen caption embeddings by cosine similarity. For each cue level, retrieval is

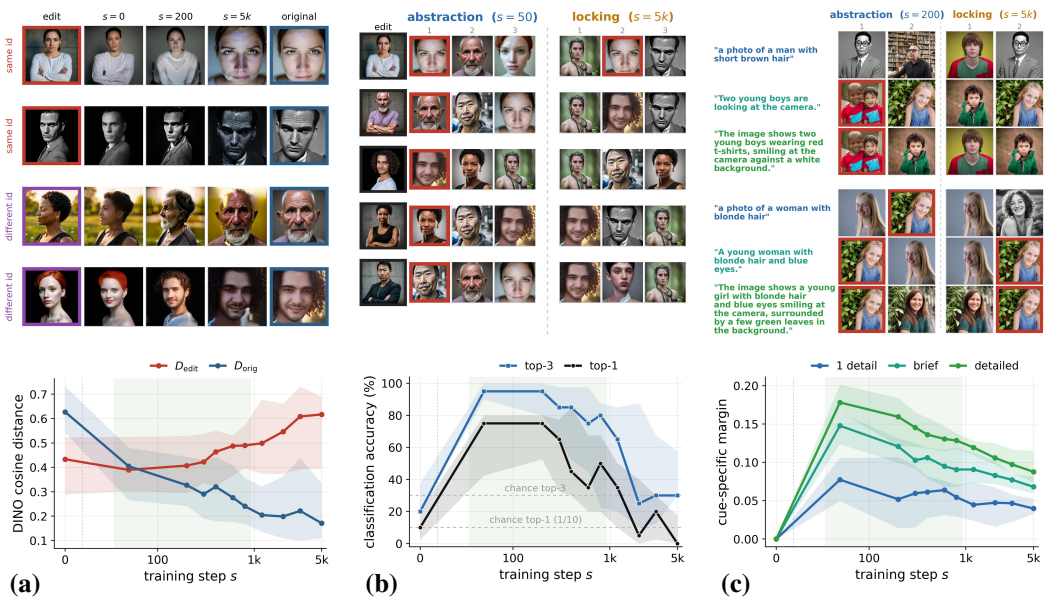

Figure 3: **Fidelity–functionality trade-off. (a) Editability.** Edited images are partially noised and denoised under the target's memory. Top: two same-identity edits and two cross-identity inputs. Columns show the edit, outputs at three checkpoints, and the target. Outputs retain edits during abstraction and revert toward the target during locking. Bottom: DINOv2 distance to the edit (D_{edit}) increases as distance to the target (D_{orig}) decreases, measured over 20 same-identity target–edit pairs with two samples per pair. **(b) One-shot recognition.** Each of 100 held-out images (10 per identity) is classified among the 10 identity memories, assigned to whichever yields the lowest denoising error over 40 shared timestep–noise pairs (chance: 1/10). Top: the three best-scoring identities (of 10) for five queries (edits) during abstraction and locking. Bottom: top-1 and top-3 accuracy, over 10 queries per identity. **(c) Text–memory retrieval.** Each of 15 memories is trained without text cues on a single image generated from a detailed caption. Cues at three specificity levels rank memories by cosine similarity between text-encoder EOS embeddings and mean-pooled memory vectors. Top: the two highest-ranked memories for two identities during abstraction and locking, represented by their training images. Bottom: similarity to the correct memory minus mean similarity to the other memories. In (b) and (c), red frames indicate correct matches. Curves: medians, bands: IQR across target–edit pairs in (a) and identities in (b–c).

given by the similarity between each description and the corresponding identity, measured relative to the other identities. Alignment with the semantic cues emerges rapidly and peaks in the abstraction regime, despite text never entering the memory-training objective. More detailed cues retrieve the memory more strongly, while all levels lose alignment as the memory becomes increasingly specialized to the particular training realization (Fig. 3c).

These experiments reveal a fidelity–functionality trade-off: as recall becomes more precise, memory becomes less editable, less invariant across realizations, and less accessible through semantic cues. Additional instance-specific information thus changes memory functionality.

7 Extensions to Other Models and Modalities

We now test whether the gist-to-verbatim ordering observed in SD 2.1 extends across architectures, generative objectives, memory parameterizations, and modalities. We repeat single-datum memorization with CogVideoX-2b (Yang et al., 2025) and the rectified-flow image model FLUX.2 (Black Forest Labs, 2024), adapting memory injection to each model's native attention architecture.

Abstraction across architectures and modalities. Figure 4 shows that both tested models pass through an intermediate regime of abstraction before verbatim reconstruction. Strikingly, in CogVideoX the memory trajectory reveals scene structure (street/beach) and characteristic motion (the direction of the car or clouds) among the first properties to be acquired, while instance-specific appearance still varies across initial noise realizations. Only later do the generated clips converge

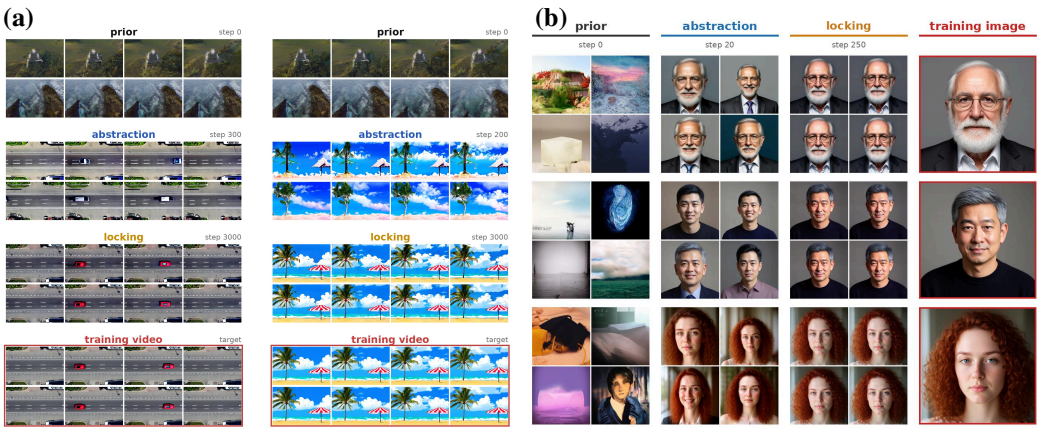

Figure 4: **Abstraction across models and modalities.** **(a)** Video memory on CogVideoX-2b, each block shows two seeds as filmstrips. During abstraction, characteristic motion, e.g., motion of the car and moving direction of the clouds, stabilize while frame content and layout vary. Finally, every seed reproduces the same target. **(b)** Image memories on FLUX.2 for three identities (rows), each sampled with four seeds. Semantic content emerges well before verbatim appearance.

toward the target. In FLUX.2, identity and pose likewise emerge before the exact appearance of the training image. In Appendix E.3, we further show that LoRA adaptation (Hu et al., 2021) of SD 2.1 to a single image produces a similar abstraction phase to that in Fig. 1. Together, these results suggest that abstraction before verbatim recall is not tied to the SD 2.1 cross-attention interface or to our particular memory parameterization, and can persist across changes in architecture, generative objective, and modality.

8 Limitations

We identify four limitations. (i) The observed semantic-first organization cannot be fully disentangled from the influence of text conditioning during pretraining. Although memory formation itself receives no textual supervision, data structure and language may both be required to shape the semantic directions along which abstraction emerges. (ii) Our experiments on strongly out-of-distribution targets are limited. In these extreme regimes, the memorization trajectory may change qualitatively, and the abstraction phase may narrow or disappear altogether. (iii) The theoretical framework serves primarily as an interpretative model rather than a complete mechanistic account of the observed dynamics. Extracting more detailed predictions would require additional assumptions on the model structure, reducing the generality of the framework. (iv) The fidelity–functionality trade-off is established for our memorization protocol and a limited set of functional probes. Different objectives or memory parameterizations may weaken, or completely avoid this trade-off.

9 Conclusions

Results suggest that gist-like and verbatim recall can emerge at different stages of the same memorization process, as a datum is learned relative to a generative prior. What is worth storing first is not intrinsic to the datum, but depends on what the prior can reconstruct and which residual directions can be acquired most efficiently. The results suggest a downside of exact recall: increasing fidelity enriches an abstract memory, but can remove the invariances that make that memory reusable.

Several questions follow from this picture. The alignment between abstraction and semantic structure may reflect both language supervision and regularities already present in the data. Disentangling these contributions will require priors learned without text or under controlled data statistics. More broadly, the fidelity–functionality trade-off raises the question of whether useful artificial memories should be optimized all the way to verbatim reconstruction, or instead stopped or regularized at an intermediate level of specificity. Finally, following what is acquired along a memorization trajectory may provide a way to expose the relevance structure of a generative prior itself.

AI USE STATEMENT

LLMs implemented code, executed experiments, and substantially assisted results analysis and manuscript preparation, including organization, prose revision, and clarity checks. The authors reviewed and approved all scientific claims, experimental design, analysis, and final text.

REFERENCES

Andrei Atanov, Jesse Allardice, Roman Bachmann, Oğuzhan Fatih Kar, R Devon Hjelm, David Griffiths, Peter Fu, Afshin Dehghan, and Amir Zamir. Videoflextok: Flexible-length coarse-to-fine video tokenization. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=KyVlaw4BxE.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DgdOkUUBzf.

Frederic C. Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, Cambridge, UK, 1932.

Christopher J Bates and Robert A Jacobs. Efficient data compression in perception and perceptual memory. *Psychological review*, 127(5):891, 2020.

Sam C Berens, Blake A Richards, and Aidan J Horner. Dissociating memory accessibility and precision in forgetting. *Nature Human Behaviour*, 4(8):866–877, 2020.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024. Accessed 2026.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237. IEEE, 2018.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.

Charles J Brainerd and Valerie F Reyna. Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in child development and behavior*, 28:41–100, 2002.

Tankut Can. Statistical mechanics of semantic compression. *arXiv preprint arXiv:2503.00612*, 2025.

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.

Sarthak Chandra, Sugandha Sharma, Rishidev Chaudhuri, and Ila Fiete. Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*, 638(8051):739–751, 2025.

Soumitri Chattopadhyay, Sanket Biswas, Emanuele Vivoli, and Josep Lladós. Towards generative class prompt learning for fine-grained visual recognition. In *British Machine Vision Conference (BMVC)*, 2024.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *International Conference on Learning Representations*, volume 2025, pp. 53791–53846, 2025.

Carles Domingo-Enrich, Yuanqi Du, and Michael S Albergo. A unified perspective on fine-tuning and sampling with diffusion and flow models. *arXiv preprint arXiv:2605.00229*, 2026.

Bradley Efron. Tweedie's formula and selection bias. *J. Am. Stat. Assoc.*, 106(496):1602–1614, 2011.

Zahra Fayyaz, Aya Altamimi, Carina Zoellner, Nicole Klein, Oliver T Wolf, Sen Cheng, and Laurenz Wiskott. A model of semantic completion in generative episodic memory. *Neural Computation*, 34(9):1841–1870, 2022.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Jerome Garnier-Brun, Luca Biggio, Davide Beltrame, Marc Mezard, and Luca Saglietti. Biased generalization in diffusion models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IqTvyp40O6.

Peter Grünwald and Teemu Roos. Minimum description length revisited. *International journal of mathematics for industry*, 11(01):1930001, 2019.

Pernille Hemmer and Mark Steyvers. A bayesian account of reconstructive memory. *Topics in cognitive science*, 1(1):189–202, 2009.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Hilbert J. Kappen. Linear theory for control of nonlinear stochastic systems. *Phys. Rev. Lett.*, 95: 200201, Nov 2005. doi: 10.1103/PhysRevLett.95.200201. URL https://link.aps.org/doi/10.1103/PhysRevLett.95.200201.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941. IEEE, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217. IEEE, 2023.

Qi Lin, Zifan Li, John Lafferty, and Ilker Yildirim. Images with harder-to-reconstruct visual representations leave stronger memory traces. *Nature human behaviour*, 8(7):1309–1320, 2024.

Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pp. 359–366, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047. IEEE, 2023.

David G Nagy, Balázs Török, and Gergő Orbán. Optimal forgetting: Semantic compression of episodic memories. *PLoS Computational Biology*, 16(10):e1008367, 2020.

David G Nagy, Gergő Orbán, and Charley M Wu. Adaptive compression as a unifying framework for episodic and semantic memory. *Nature Reviews Psychology*, 4(7):484–498, 2025.

Xueyan Niu, Bo Bai, Nian Guo, Weixi Zhang, and Wei Han. Rate–distortion–perception trade-off in information theory, generative models, and intelligent communications. *Entropy*, 27(4):373, 2025.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pp. 291–309. Springer, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Peter Potaptchik, Lee Cheuk Kit, and Michael Samuel Albergo. Tilt matching for scalable sampling and fine-tuning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=dQA4Gjt4KU.

Lucas Relic, Roberto Azevedo, Markus Gross, and Christopher Schroers. Lossy image compression with foundation diffusion models. In *European Conference on Computer Vision*, pp. 303–319. Springer, 2024.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Marta Aparicio Rodriguez, Anastasia Borovykh, Grigorios A. Pavliotis, and Daniel James Korchinski. Diffusion models preferentially memorize prototypical examples or: Why does my diffusion model love slop? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ZIFQJumCbK.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Daniel L Schacter and Donna Rose Addis. The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):773, 2007.

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

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021a. Curran Associates Inc. ISBN 9781713845393.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Eleanor Spens and Neil Burgess. A generative model of memory construction and consolidation. *Nature human behaviour*, 8(3):526–543, 2024.

Eleanor Spens and Neil Burgess. Hippocampo-neocortical interaction as compressive retrieval-augmented generation. *Nature Communications*, 2026.

Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. *Advances in neural information processing systems*, 31, 2018.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15897–15907. IEEE, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. URL https://arxiv.org/abs/2308.06721.

George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. *Advances in Neural Information Processing Systems*, 34: 11517–11529, 2021.

APPENDIX OVERVIEW

A Extended related works **14**

B Mathematical framework derivation **15**

C Implementation details **17**

D Experimental details **18**

D.1 From gist to verbatim (Fig. 1) 18

D.2 Generalization through the prior (Fig. 2a,b) 18

D.3 Positional specificity of addressing (Fig. 2c) 19

D.4 Editability (Fig. 3a) 19

D.5 One-shot recognition (Fig. 3b) 20

D.6 Text–memory retrieval (Fig. 3c) 21

D.7 Other models and modalities (Fig. 4) 21

E Additional experiments **22**

E.1 Out-of-distribution targets 22

E.2 Memory-norm growth 23

E.3 Memorizing through Low Rank Adaptation 23

E.4 Additional qualitative examples 24

A EXTENDED RELATED WORKS

Generative priors and lossy compression. Viewing memory as storing what cannot be supplied by a prior is closely related to generative approaches to lossy compression. Classical rate–distortion theory formalizes the trade-off between representation rate and reconstruction accuracy, while the perception–distortion and rate–distortion–perception frameworks distinguish sample-wise fidelity from the requirement that reconstructions remain perceptually consistent with the data distribution (Blau & Michaeli, 2018; 2019; Niu et al., 2025). Close to our setting, Tschannen et al. (2018) study distribution-preserving lossy compression, whose limiting regimes range from generation from the learned distribution at zero rate to exact reconstruction at sufficiently high rate. More recently, pretrained diffusion models have been used directly as strong priors for recovering information discarded by compression (Relic et al., 2024). Information-theoretic work also studies representations that can be progressively refined while varying distortion and perceptual constraints (Zhang et al., 2021). Bachmann et al. (2025) and Atanov et al. (2026) learn ordered representations in which short token prefixes capture abstract content and additional tokens refine reconstruction. These works vary an explicit representation budget, such as bitrate or token-prefix length. Here, instead, representation size is fixed and the ordering unfolds over optimization time.

Compression and reconstructive memory. A parallel literature in cognitive science and computational neuroscience treats memory as a compressed representation whose missing information can be reconstructed from prior knowledge. From a Bayesian viewpoint, recall can be formalized as an interaction between noisy episodic evidence and prior knowledge at different levels of abstraction (Hemmer & Steyvers, 2009), while rate–distortion approaches have argued that limits on perceptual and memory resources naturally produce selective, lossy representations (Bates & Jacobs, 2020). In the semantic-compression framework of Nagy et al. (2020), compression through a learned generative model produces systematic memory distortions and a continuum between detailed and increasingly gist-like recall. Fayyaz et al. (2022) similarly propose that episodic memory

stores an incomplete "episodic gist", with a generative semantic system completing information omitted from the trace. These works share with ours the idea that what must be explicitly stored depends on what a model of prior experience can reconstruct.

Generative memory and residual information. Recent work makes the relation between memory and generative prediction increasingly explicit. The hippocampo–neocortical model of Spens & Burgess (2026) describes hippocampal traces as compressed representations used to retrieve and condition a generative network, avoiding explicit storage of information that can instead be supplied by neocortical knowledge. The adaptive-compression framework of Nagy et al. (2025) extends this idea to a setting in which semantic knowledge itself changes over time: surprising experiences can update the model rather than merely being compressed relative to a fixed one. Our construction deliberately freezes the generative prior, isolating the complementary problem of forming a single new memory relative to an existing body of knowledge.

A complementary perspective comes from Chandra et al. (2025), who study how pre-existing hippocampal spatial structure changes what must be stored explicitly for episodic and associative recall. Their spatial scaffolds reduce sequence learning to structured transitions, with recall detail depending on available representational resources. More directly related to our residual formulation, Lin et al. (2024) show that reconstruction residuals of compressed visual representations predict both the strength of human memory traces and retrieval latency. This supports the broader view that information poorly accounted for by an existing representation can receive preferential memory resources.

Gist, meaning, and abstraction. Beyond the gist–verbatim distinction discussed in Sec. 6, a recent work in Statistical Mechanics formulates semantic compression as minimizing description length while preserving position in a semantic representation space, yielding distinct lossless/lossy and extractive/abstractive regimes (Can, 2025). Although this work concerns semantic compression rather than memory formation, it provides a complementary formalization of abstraction as preserving semantic structure while discarding finer information.

Tilting and control of generative priors. Our theoretical description of memory as a learned modification of a frozen generative distribution connects to recent control-theoretic views of diffusion and flow models. Uehara et al. (2024) formulate diffusion fine-tuning as entropy-regularized stochastic control relative to a pretrained process, while Domingo-Enrich et al. (2025) cast reward fine-tuning of diffusion and flow models as a stochastic optimal-control problem. More recent work directly considers target distributions obtained by exponentially tilting a base generative density (Potaptchik et al., 2026; Domingo-Enrich et al., 2026). As opposed to this approach, our exponential-tilt model is not to be intended as part of a new fine-tuning algorithm, and we instead use it to build a minimal description of memory formation.

Learned conditioning and inversion. Beyond the personalization methods discussed in the main text, ELITE and IP-Adapter learn mappings from reference images to conditioning representations (Wei et al., 2023; Ye et al., 2023), while Null-text Inversion optimizes unconditional embeddings around an image-specific inversion trajectory to support reconstruction and text-guided editing (Mokady et al., 2023). These methods are concerned primarily with obtaining useful conditioning or editable reconstructions, while we focus on how the properties of a single learned conditioning signal change over the course of its optimization.

B MATHEMATICAL FRAMEWORK DERIVATION

In this appendix we will provide a derivation for the equations presented in Section 4.
Let us start by defining a forward noising process in $\mathbb{R}^d$ which gradually mixes a data point $\boldsymbol{x}_0 \sim p_0$ with Gaussian noise:

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}\,, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})\,, \quad t \in [0, T]\,, \tag{8}$$

usually, one takes $\alpha_t \to 0$ and $\sigma_t \to 1$ as $t \to T$ so that the ending distribution p_T is a standard Gaussian. This formulation is pretty general, and includes both DDPMs (Ho et al., 2020) and SDE based diffusion (Song et al., 2021b). At each timestep, the marginal distribution of the noised points

x_t is given by

$$p_t(\boldsymbol{x}) = \int d\boldsymbol{x}_0\, p_0(\boldsymbol{x}_0)\, \mathcal{N}\big(\boldsymbol{x}; \alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}\big)\ . \tag{9}$$

Sampling from p_0 is done by inverting the forward process, i.e. by finding a process which, starting from p_T, brings us to p_0. The idea is that p_T is easy to sample from, so getting samples from p_0 becomes as simple as drawing from p_T and then following the backward process. This can be achieved in many different ways: we can for example train a model to infer the original $\boldsymbol{x}_0$ from $\boldsymbol{x}_t$, or to deduce the added noise $\boldsymbol{\epsilon}$, and then invert equation 8. These approaches are all related to the computation of the score function $\boldsymbol{s}(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ by Tweedie's formula (Efron, 2011), which can be retrieved through direct computation from equation 9:

$$\boldsymbol{s}_0(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \log \int d\boldsymbol{x}_0\, p_0(\boldsymbol{x}_0)\, \mathcal{N}\big(\boldsymbol{x}; \alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}\big) = \frac{\alpha_t \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] - \boldsymbol{x}_t}{\sigma_t^2} = -\frac{1}{\sigma_t}\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{x}_t]\,. \tag{10}$$

We now move on to the exponential tilt of the binary features $\boldsymbol{\phi}$ introduced in equation 4:

$$p_{\boldsymbol{u}}(\boldsymbol{x}_0) = \frac{e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{x}_0)}}{Z(\boldsymbol{u})} p_0(\boldsymbol{x}_0), \qquad Z(\boldsymbol{u}) = \mathbb{E}_{p_0}\Big[e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{X}_0)}\Big]\ .$$

When noising samples from this new target distribution, we get a slightly modified version of the marginals in equation 9:

$$p_t^{\boldsymbol{u}}(\boldsymbol{x}_t) = \frac{1}{Z(\boldsymbol{u})} \int d\boldsymbol{x}_0\, p_0(\boldsymbol{x}_0)\, \mathcal{N}\big(\boldsymbol{x}; \alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}\big)\ e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{X}_0)} = \tag{11}$$

$$= \frac{1}{Z(\boldsymbol{u})} p_t(\boldsymbol{x}_t) \mathbb{E}_0\Big[e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{X}_0)} \,\Big|\, \boldsymbol{X}_t = \boldsymbol{x}\Big] \tag{12}$$

where we have used Bayes' theorem to write

$$p_0(\boldsymbol{x}_0)\, \mathcal{N}\big(\boldsymbol{x}; \alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}\big) = p_0(\boldsymbol{x}_0) p(\boldsymbol{x}_t|\boldsymbol{x}_0) = p(\boldsymbol{x}_0|\boldsymbol{x}_t) p_t(\boldsymbol{x}_t)\,, \tag{13}$$

and we have denoted by $\mathbb{E}_0[\cdot]$ the expectation over $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$, i.e. the posterior induced by the original p_0 and the forward noising process. The expression of equation 12 allows us to directly derive the new score's decomposition $\boldsymbol{s}_{\boldsymbol{u}}(\boldsymbol{x}, t) = \boldsymbol{s}_0(\boldsymbol{x}, t) + \delta \boldsymbol{s}_{\boldsymbol{u}}(\boldsymbol{x}, t)$ given by equation 5 (Kappen, 2005):

$$\delta \boldsymbol{s}_{\boldsymbol{u}}(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \log \mathbb{E}_0\Big[e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{X}_0)} \,\Big|\, \boldsymbol{X}_t = \boldsymbol{x}\Big]\ . \tag{14}$$

The conditioning vector $\boldsymbol{u}$ is learned by choosing a single datum $\boldsymbol{x}^\star$ and making $p_u(\boldsymbol{x}_0)$ collapse on $p^\star(\boldsymbol{x}_0) = \delta(\boldsymbol{x}_0 - \boldsymbol{x}^\star)$, i.e. by forcing the conditioned model to only produce $\boldsymbol{x}^\star$. This can be done by minimizing the weighted objective in equation 2, which we can write more explicitly as:

$$\mathcal{L}(u) = \int_0^T dt\, w(t)\, \mathbb{E}_{p^\star}\ \big[||\boldsymbol{s}_{\boldsymbol{u}}(\boldsymbol{X}_t, t) - \boldsymbol{s}^\star(\boldsymbol{X}_t, t)||^2\big]\ , \tag{15}$$

where

$$\boldsymbol{p}^\star(\boldsymbol{x}_t, t) = \mathcal{N}\big(\boldsymbol{x}; \alpha_t \boldsymbol{x}^\star, \sigma_t^2 \boldsymbol{I}\big) \tag{16}$$

is the marginal of the forward process starting from $\boldsymbol{x}^\star$, while

$$\boldsymbol{s}^\star(\boldsymbol{x}_t, t) = -\sigma_t^{-2}(\boldsymbol{x}_t - \alpha_t \boldsymbol{x}^\star)$$

is the score associated to it, as found in equation 1.

To further study the evolution of $\boldsymbol{u}$ induced by the gradient flow, we put ourselves in the SDE-based generation framework. In particular, we use the Ornstein-Uhlenbeck process in the noising phase, which corresponds to the variance preserving (VP) scheme:

$$d\boldsymbol{x}_\theta = -\boldsymbol{x} d\theta + \sqrt{2} d\boldsymbol{W}_\theta \tag{17}$$

where $d\boldsymbol{W}_\theta$ is a Wiener process and $\theta \in [0, \Theta]$. It can be shown that the solution of this process is given by:

$$\boldsymbol{x}_t = e^{-\theta}\boldsymbol{x}_0 + \sqrt{1 - e^{-2\theta}}\boldsymbol{\epsilon} \tag{18}$$

which can be reduced to equation 8 by putting $\theta = -\log \alpha_t$, $\sigma_t = \sqrt{1 - \alpha_t^2}$ and taking $\Theta \to \infty$. If we now define a backward time $s = \Theta - \theta$ with negative increments ds, the backward process can be written as:

$$d\boldsymbol{x}_s = (-\boldsymbol{x}_s - 2\boldsymbol{s}(\boldsymbol{x}_s, \theta = \Theta - s))ds + \sqrt{2} d\tilde{\boldsymbol{W}}_s\,, \tag{19}$$

where $d\tilde{\boldsymbol{W}}_s$ is independent from the original one. In this framework, there is an elegant way of rewriting the objective $\mathcal{L}(\boldsymbol{u})$: first, similarly to (Lyu, 2009), we can show that the Kullback–Leibler divergence between two distributions p^1 and p^2 noised through equation 18 will evolve as

$$\frac{d}{d\theta} D_{KL}(p^1_\theta || p^2_\theta) = -\mathbb{E}_{p^1_\theta}\left[||\nabla_{\boldsymbol{x}} \log p^1_\theta(\boldsymbol{X}_\theta) - \nabla_{\boldsymbol{x}} \log p^2_\theta(\boldsymbol{X}_\theta)||^2\right] \tag{20}$$

$$= -\mathbb{E}_{p^1_\theta}\left[||\boldsymbol{s}^1(\boldsymbol{X}_\theta, \theta = \Theta - s) - \boldsymbol{s}^2(\boldsymbol{X}_\theta, \theta = \Theta - s)||^2\right] . \tag{21}$$

Where the minus sign is consistent with the fact that the two distributions are both converging toward a standard Gaussian. In the limit of large Θ, $p^1_\Theta = p^2_\Theta = \mathcal{N}(0, \boldsymbol{I})$: starting from this distribution and using the exact scores, the backward process (equation 19) will reproduce the exact marginals p^1_θ and p^2_θ at $s = \Theta - \theta$. As a result, we can integrate the previous equation (with a sign flip) starting from $D_{KL}(p^1_\Theta || p^2_\Theta) = 0$ to get

$$D_{KL}(p^1||p^2) = \int_0^\Theta ds\, \mathbb{E}_{p^1_s}\left[||\boldsymbol{s}_1(\boldsymbol{X}_s, \theta = \Theta - s) - \boldsymbol{s}_2(\boldsymbol{X}_s, \theta = \Theta - s)||^2\right] , \tag{22}$$

and then we can plug this expression, together with the likelihood time weighting $w(t)dt \propto ds$ (Song et al., 2021a) to get:

$$\mathcal{L}(u) = D_{KL}(p^\star||p_{\boldsymbol{u}}) = -\log p_{\boldsymbol{u}}(\boldsymbol{x}^\star) = -\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{x}^\star) + \log Z(\boldsymbol{u}) + \text{const.} \tag{23}$$

We stress the fact that while this choice of $w(t)$ is the most convenient in this theoretical setting, it is not the one we use in the experiments. The expression in equation 23 allows us to easily get the gradient flow dynamics of $\boldsymbol{u}$ with learning rate η found in equation 7:

$$\frac{d\boldsymbol{u}}{d\tau} = -\eta \nabla_{\boldsymbol{u}} \mathcal{L} = \eta\, \boldsymbol{r}_{\boldsymbol{\phi}}(\boldsymbol{u}), \tag{24}$$

where we have used the definition in equation 6

$$\boldsymbol{r}_{\boldsymbol{\phi}}(\boldsymbol{u}) = \boldsymbol{\phi}(\boldsymbol{x}^\star) - \mathbb{E}_{p_{\boldsymbol{u}}}\left[\boldsymbol{\phi}(\boldsymbol{X})\right] ,$$

and we have used the fact that

$$\nabla \log Z(\boldsymbol{u}) = \frac{1}{Z(\boldsymbol{u})} \int d\boldsymbol{x}_0 p(\boldsymbol{x}_0)\, \boldsymbol{\phi}(\boldsymbol{x}_0)\, e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{x}_0)} = \mathbb{E}_{p_{\boldsymbol{u}}}[\boldsymbol{\phi}(\boldsymbol{X}_0)] . \tag{25}$$

We are only left with finding the derivative of $\boldsymbol{r}_{\boldsymbol{\phi}}(\boldsymbol{u})$, which, by the chain rule, can be expressed as

$$\frac{d\boldsymbol{u}}{d\tau} = \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{u}} \frac{d\boldsymbol{u}}{d\tau} , \tag{26}$$

where $\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{u}}$ is the matrix of entries $\frac{\partial r_i}{\partial u_j}$. These elements can be easily expressed in the following way:

$$\frac{\partial r_i}{\partial u_j} = -\frac{\partial \mathbb{E}_{p_{\boldsymbol{u}}}[\boldsymbol{\phi}_i(\boldsymbol{X}_0)]}{\partial u_j} = -\frac{\partial}{\partial u_j} \frac{1}{Z(\boldsymbol{u})} \int d\boldsymbol{x}_0 p(\boldsymbol{x}_0)\, \boldsymbol{\phi}(\boldsymbol{X}_0)\, e^{\boldsymbol{u}^\top \boldsymbol{\phi}(\boldsymbol{X}_0)} = \tag{27}$$

$$= -\mathbb{E}_{p_{\boldsymbol{u}}}[\boldsymbol{\phi}_i(\boldsymbol{X}_0)\phi_j(\boldsymbol{X}_0)] + \mathbb{E}_{p_{\boldsymbol{u}}}[\boldsymbol{\phi}_i(\boldsymbol{X}_0)]\, \mathbb{E}_{p_{\boldsymbol{u}}}[\boldsymbol{\phi}_j(\boldsymbol{X}_0)] = -[\boldsymbol{\Sigma}_{\boldsymbol{u}}]_{ij} , \tag{28}$$

which finally gets us to the other expression in equation 7:

$$\frac{d\boldsymbol{r}_{\boldsymbol{\phi}}}{d\tau} = -\eta \boldsymbol{\Sigma}_{\boldsymbol{u}} \boldsymbol{r}_{\boldsymbol{\phi}}, \tag{29}$$

C Implementation details

Memory parameterization. The memory is a set of M learnable vectors $\boldsymbol{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\}$, $\boldsymbol{c}_m \in \mathbb{R}^{1024}$, living in the cross-attention context space of the frozen UNet (8,192 trainable parameters at M=8). At every cross-attention layer the memory is read through a second attention pass that reuses the layer's frozen query/key/value/output projections. The value and output projections are applied without their biases, so a zero memory's contribution is exactly equal to zero. As a consequence, at $\boldsymbol{C} = \boldsymbol{0}$ generation coincides with the empty-caption model exactly. Notice that this parallel construction allows us to vary freely the size of the memory M, instead of being tied to the text embedding space size (77 tokens). The text pathway is conditioned on the empty caption throughout the whole process.

Initialization. Each vector is drawn i.i.d. Gaussian with standard deviation 0.01 (per-vector norm ≈ 0.32). We use this random initialization instead of putting all vectors to zero because the latter is degenerate: all keys would coincide and receive identical gradients indefinitely.

Training. We minimize the frozen model's own denoising objective (equation 2) on the target latent (VAE posterior mean of $\boldsymbol{x}^\star$ at 768px), with the native v-prediction target and t sampled uniformly. Optimization uses AdamW with learning rate 10^{-2}, no weight decay, and a *constant* schedule with no warm-up. Each gradient step averages 8 independent $(t, \boldsymbol{\epsilon})$ draws against the same target. Runs last $5{,}000$ steps with $\sim$20 checkpoints on a fixed grid unless otherwise specified.

Sampling. All samples use DPM-Solver++ (multistep, order 3), 50 steps, 768px, classifier-free guidance $\omega = 3$; the unconditional branch is the same model with the memory detached. We use $\omega = 3$ for the caption-conditioned generated images used as reference as well, so that memory- and caption-conditioned samples are always compared under an identical protocol. Sampling is deterministic given the noise seed, which is what licenses the seed-paired comparisons of Section 5. Evaluation seeds are fixed per experiment and disjoint from any seed involved in creating a training target.

Compute. Each single-image run (training plus checkpoint sampling) takes a few hours on one A100/H100-class GPU; all experiments in the paper are single-GPU.

D Experimental details

D.1 From gist to verbatim (Fig. 1)

Targets. The bank contains 30 in-distribution targets (10 animals, 10 vehicles, 10 faces), each generated by the frozen model itself at $\omega = 7.5$, 50 steps, 768px, from a per-image caption with a recorded generating seed. All 30 are trained at $M \in \{1, 8, 32, 64\}$.

Evaluation samples. At each of $\sim$20 checkpoints we store $S = 20$ samples from fixed held-out seeds under the protocol of Appendix C. Because targets are model generated, a target's own generating seed can fall inside the evaluation set. In that case, the generating seed is excluded and the run is evaluated on the remaining 19 seeds, so no evaluation sample ever shares the initial noise that produced its run's target.

Bias and variance. For an embedding E and samples $\{\boldsymbol{g}_s\}_{s=1}^{S}$ at one checkpoint, let $\bar{\boldsymbol{e}} = S^{-1}\sum_s E(\boldsymbol{g}_s)$. We study

$$\text{bias}^2 = \frac{\|\bar{\boldsymbol{e}} - E(\boldsymbol{x}^\star)\|^2}{\|E(\boldsymbol{x}^\star)\|^2}, \qquad \text{variance} = \frac{1}{S}\sum_s \frac{\|E(\boldsymbol{g}_s) - \bar{\boldsymbol{e}}\|^2}{\|E(\boldsymbol{x}^\star)\|^2}, \tag{30}$$

which together form an exact decomposition of the normalized mean squared distance to the target embedding. The semantic space is the DINOv2 (Oquab et al., 2024)embedding (ViT-B/14), mean-pooled over patch tokens and unit-normalized, while the surface space is the raw VAE latent.

Half-decay steps. We define the half-decay step of a curve as the training step at which it has covered half of its total fall. We compute it through a linear interpolation of the checkpoints immediately before and after the crossing. This way, we assume no functional form for the curve. In Section 3, we have computed the per-image ratio between the half-decay steps of the surface and semantic bias^2 curves. The value quoted in the text is the median of the distribution computed over the 30 images.

D.2 Generalization through the prior (Fig. 2a,b)

Setup. Each caption defines a reference pool: 100 generated images at fixed noise seeds under the sampling protocol of Appendix C, split into 50 train and 50 held-out test seeds. A memory is trained from scratch on N of the training pictures (N=1 or N=50, M=8, 4,000 steps, 18 checkpoints with

held-out sampling at every checkpoint). Twelve face captions enter the pooled result, one memory per caption at each N, all under one protocol: target images, memory samples, and reference pools share the same sampler, resolution, and $\omega = 3$. Again, training seeds never intersect the test split, so no reference image is ever evaluated against what was trained on.

Paired advantage. At a checkpoint, let $\boldsymbol{g}_s$ be the memory's sample at held-out seed s and $\boldsymbol{y}_s$ the caption's own generated image from that same noise. Calling $\hat{E}$ the unit-normalized DINOv2 embedding, we compute the embedding distances $D_{ss'} = \|\hat{E}(\boldsymbol{g}_s) - \hat{E}(\boldsymbol{y}_{s'})\|$ over the $S = 50$ test seeds and then define

$$\text{adv} \;=\; \frac{1}{\bar{D}}\,\frac{1}{S}\sum_{s}\Big(\tfrac{1}{S-1}\textstyle\sum_{s'\neq s} D_{ss'} \;-\; D_{ss}\Big), \tag{31}$$

which corresponds to the mean excess distance of a sample to *other* seeds' outputs over its own seed's, normalized by the mean entry $\bar{D}$ so values are comparable across captions. The statistic is zero both for an untrained memory (samples unrelated to any seed) and for a fully memorized one (identical samples at every seed), and it can only be positive when the output varies with the seed the way the caption's generated images do.

Significance and controls. Holding D fixed, we randomly re-pair samples with references and recompute equation 31.As a negative control, the same samples are scored against pools generated from three unrelated captions at the same seeds, leading to a metric which stays close to 0 throughout the training process.

D.3 Positional specificity of addressing (Fig. 2c)

Probe. In the positional addressing experiment, we evaluate each memory using a fixed ensemble of P=12 held-out probe images generated from 12 different captions, for all three classes: animals, vehicles, and faces. The probes belong to the same class as the memory's training target, and no probe image is a training target in any run. To compute the metrics, we add noise to the probes at timestep t=800 using a fixed noise draw and pass them through the UNet with the memory being evaluated attached. This evaluation noise is separate from the noise used to generate the probe images.

Measurement. At every cross-attention layer we record the query–key cosine similarity $a_{p,\ell,m} = \cos(\boldsymbol{q}_{p,\ell}, \boldsymbol{k}_m)$ for probe image p, position ℓ, and memory token m, averaged over heads. Fixing a layer and a token and writing $\bar{a}_\ell = \frac{1}{P}\sum_p a_{p,\ell}$ for the image-averaged map, we define

$$V_{\text{pos}} = \text{Var}_\ell(\bar{a}_\ell)\,, \qquad V_{\text{cont}} = \mathbb{E}_\ell[\text{Var}_p(a_{p,\ell})]\,, \tag{32}$$

so that $\text{Var}_{p,\ell}(a) = V_{\text{pos}} + V_{\text{cont}}$ exactly, by the law of total variance: V_{pos} is the positional structure shared by all probe images, V_{cont} the residual across-image variability. Positional structure that differs per image is counted in V_{cont}, so the reported fraction $f = V_{\text{pos}}/(V_{\text{pos}} + V_{\text{cont}})$ is a conservative estimate of positional addressing. Components are averaged over tokens and the 16 layers before the ratio, so layers contribute in proportion to the variance they carry.

Runs and baseline. We measure 30 memories per memory size (10 per class) at $M \in \{1, 8, 32, 64\}$ at every stored checkpoint, and plot their mean $\pm$ s.e.m.. With a single token the softmax weight is identically one, so the M=1 addressing pathway receives no gradient; its slow drift reflects the token moving under the value-pathway loss, read through the frozen key projection, and serves as the baseline against which the $M \geq 8$ rise is read.

D.4 Editability (Fig. 3a)

Data. We use the ten model-generated face targets described in Appendix D.1 and their existing single-image memories. We denote each target by x_{orig} (corresponding to $\boldsymbol{x}^\star$ in Section 2) and its memory after s optimization steps by $\boldsymbol{C}^{(s)}$. For each target, we generate two edited counterparts x_{edit} using FLUX.2 Kontext (Labs et al., 2025), yielding 20 same-identity target–edit pairs. Edits vary pose, expression, accessories, and scene or background. The qualitative panel additionally

includes two cross-identity examples in which the edited input depicts a different person from the memory's stored subject.

For both editability and one-shot recognition (Appendix D.5), we evaluate the randomly initialized memories ($s = 0$) and checkpoints $s \in \{50, 200, 300, 400, 600, 800, 1200, 2000, 3000, 5000\}$. The initialization checkpoint retains the memory branch and is distinct from the memory-free prior, for which that branch is disabled.

U-turn protocol. Following diffusion-based editing (Meng et al., 2022), we partially noise the edited image and complete denoising under the target's memory. We encode the edit as $z_0^{\text{edit}} = \mathcal{E}(x_{\text{edit}})$ using the frozen VAE encoder and sample

$$z_{t_{\text{switch}}} = \alpha_{t_{\text{switch}}} z_0^{\text{edit}} + \sigma_{t_{\text{switch}}} \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{33}$$

Here, $\alpha_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$ are the signal and noise coefficients of the SD2.1 schedule, consistent with Section 2. We then denoise from t_{switch} to 0 with classifier-free guidance:

$$\hat{v}_t = v_\theta(z_t, t, \varnothing) + \omega\big[v_\theta(z_t, t, \boldsymbol{C}^{(s)}) - v_\theta(z_t, t, \varnothing)\big]. \tag{34}$$

The frozen v-prediction denoiser v_θ receives empty-caption text conditioning $\boldsymbol{c}_\varnothing$ in both branches; $\varnothing$ denotes the memory-disabled branch. Thus, the memory is the sole additional conditioning signal.

We use stochastic DDIM sampling with $\eta = 1$, 50 sampling steps, guidance scale $\omega = 3$, and resolution 768×768, with $t_{\text{switch}} = 800$ on the $T = 1000$ diffusion-time grid. We write the output as $\hat{x}_k = \mathcal{U}(x_{\text{edit}}, \boldsymbol{C}^{(s)}, t_{\text{switch}}, \xi_k)$, where ξ_k seeds both the forward noise and the stochastic denoising trajectory. We generate $K = 2$ samples per target–edit pair at each checkpoint.

Metric. Let E_{DINO} denote the unit-normalized DINOv2 `[CLS]` embedding (ViT-B/14, `dinov2-base`), and define the cosine distance $d_{\text{DINO}}(a, b) = 1 - E_{\text{DINO}}(a)^\top E_{\text{DINO}}(b)$. For each target–edit pair and checkpoint, we compute

$$D_{\text{edit}} = \frac{1}{K}\sum_{k=1}^{K} d_{\text{DINO}}(\hat{x}_k, x_{\text{edit}}), \qquad D_{\text{orig}} = \frac{1}{K}\sum_{k=1}^{K} d_{\text{DINO}}(\hat{x}_k, x_{\text{orig}}). \tag{35}$$

Lower D_{edit} indicates greater similarity to the edited input, whereas lower D_{orig} indicates greater similarity to the stored target. Increasing D_{edit} together with decreasing D_{orig} is consistent with the memory overwriting the edit as it approaches verbatim recall. At each checkpoint we report the mean of D_{edit} and D_{orig} over the $N = 20$ target–edit pairs.

D.5 One-shot recognition (Fig. 3b)

Data. We use the same ten face targets, corresponding to ten different identities, and memory checkpoints as in Appendix D.4. Using the same editing pipeline, we generate ten held-out images per identity, yielding 100 images with variations in pose, expression, or scene, all of which are excluded from memory training. The task tests whether a memory learned from a single target image recognizes other realizations of the same identity.

Scoring protocol. We encode each held-out image x as $z_0 = \mathcal{E}(x)$ and draw $K = 40$ uniformly sampled timestep–noise pairs (t_k, ε_k), shared across candidate memories for that image. Each pair defines a noised latent and its v-prediction target:

$$z_{t_k} = \alpha_{t_k} z_0 + \sigma_{t_k} \varepsilon_k, \qquad v_{t_k} = \alpha_{t_k} \varepsilon_k - \sigma_{t_k} z_0, \tag{36}$$

where $\varepsilon_k \sim \mathcal{N}(0, I)$ and the encoder, denoiser, and noise coefficients are as defined in Appendix D.4. We score each memory by its estimated expected denoising error:

$$\hat{L}\big(x, \boldsymbol{C}_j^{(s)}\big) = \frac{1}{K}\sum_{k=1}^{K} \big\| v_\theta\big(z_{t_k}, t_k, \boldsymbol{C}_j^{(s)}\big) - v_{t_k} \big\|_2^2, \tag{37}$$

where by $\boldsymbol{C}_j^{(s)}$ we denote the identity j's memory at optimization step s. The text pathway remains conditioned on the empty caption. We evaluate the denoiser directly, without a reverse diffusion trajectory, using images at resolution 768×768. All candidate memories are compared at the same optimization step.

Metric. We rank identities by increasing denoising error and assign each held-out image to

$$\hat{\jmath}^{(s)}(x) = \arg\min_{j \in \mathcal{J}} \hat{L}\big(x, \boldsymbol{C}_j^{(s)}\big)\,, \tag{38}$$

Where $\mathcal{J}$, with $|\mathcal{J}| = 10$, denotes the set of all identity indices. Top-1 and top-3 accuracy measure the fraction of images whose true identity appears among the one or three lowest-error candidates, respectively, with chance levels 0.1 and 0.3. At each checkpoint, we compute each identity's own accuracy over its 10 held-out queries, then report the median and interquartile range (25th–75th percentile) of these $N = 10$ per-identity accuracies.

D.6 Text–memory retrieval (Fig. 3c)

Data. We use a separate set of $N = 15$ model-generated identities, distinct from those used for editability and one-shot recognition. For each identity, we construct three text cues at different levels of specificity and generate one target image per cue level with SD2.1 (Fig. 3c, top). We train a single memory on the *detailed*-level image alone, following Section 2, without supplying any text cue during training. The other three generated images are not used for memory training. We denote identity i's memory after s optimization steps by $\boldsymbol{C}_i^{(s)}$. All three cue levels are evaluated against this same memory, using the initialization and evaluation checkpoints specified in Appendix D.4.

Text cues. *1 detail* adds one distinguishing attribute; and *brief* and *detailed* are automatically generated descriptions at two levels of detail. For one identity, the three cues are:

> *1 detail*: "a photo of a man with short light brown hair"
> *brief*: "A man standing in front of a colorful wall."
> *detailed*: "The image shows a man standing in front of a colorful wall with a smile on his face."

Text and memory representations. For cue level ℓ, let $\boldsymbol{t}_i^{(\ell)} \in \mathbb{R}^{1024}$ denote the EOS-token representation produced by the frozen SD2.1 text encoder. We represent each memory by the mean of its $M = 8$ conditioning vectors:

$$\bar{\boldsymbol{c}}_i^{(s)} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{c}_{i,m}^{(s)}.$$

Memory vectors and text-token representations share the same dimensionality and frozen UNet cross-attention key/value projections. We compare the pooled memory and text representations using cosine similarity.

Metric. We measure how well each identity's own text cue at level ℓ retrieves its own memory, relative to the other $N - 1$ identities' memories:

$$\mathrm{margin}_i^{(\ell)}(s) = \cos\big(\boldsymbol{t}_i^{(\ell)}, \bar{\boldsymbol{c}}_i^{(s)}\big) - \frac{1}{N-1} \sum_{j \neq i} \cos\big(\boldsymbol{t}_i^{(\ell)}, \bar{\boldsymbol{c}}_j^{(s)}\big). \tag{39}$$

A positive margin indicates greater similarity to the identity's own memory than to the other identities' memories on average. Cues shared by multiple identities remain in the comparison set.

D.7 Other models and modalities (Fig. 4)

Shared setup. We extend the memory construction of Section 2 to CogVideoX-2b and FLUX.2. In both experiments, we freeze the pretrained model and optimize only $M = 8$ conditioning vectors on a single target using the backbone's native training objective and empty-prompt text conditioning. The memory branch uses a separate softmax over memory keys and frozen projection weights with biases removed, ensuring zero contribution when the memory vectors are zero. We initialize the vectors with independent Gaussian entries of standard deviation 0.01, keeping the initial model close to the empty-prompt prior. At generation, we apply classifier-free guidance with scale $\omega = 3$, combining the memory-conditioned prediction with the zero-memory prediction, which is the empty-prompt prior.

CogVideoX-2b. CogVideoX-2b (Yang et al., 2025) is a latent video diffusion transformer with a causal 3D VAE and a native v-prediction objective. Each transformer block applies joint attention to text and spatiotemporal video tokens, using a hidden dimension of 1920 with 30 attention heads. Text conditioning is obtained from a T5 encoder and held at the empty caption throughout memory training and evaluation. We generate a single 49-frame target video at resolution 480×720 using the frozen model and a hand-written caption describing a fixed-camera scene. The VAE encodes this target into a latent with 13 temporal positions, 16 channels, and spatial dimensions 60×90. The generating caption is not supplied during memory training. We represent the memory by eight 1920-dimensional vectors and attach a memory branch to the joint attention in every transformer block. The joint text–video queries attend to memory keys and values obtained through the frozen key and value projections, with projection biases removed. We apply the base attention pathway's key normalization to the memory keys. The branch output passes through the frozen output projection without bias and is added to the joint attention output before it is separated into text and video streams. We train on the target latent using the native v-prediction parameterization of the denoising objective in equation 2, sampling timesteps uniformly and drawing fresh noise at each update. Optimization uses AdamW with learning rate 10^{-2} and no weight decay for 3000 steps. A 300-step linear warm-up is followed by cosine decay to one tenth of the peak learning rate. We save checkpoints throughout training, with denser coverage during the first 500 updates. Evaluation uses 50 sampling steps, held-out noise seeds, and tiled VAE decoding.

FLUX.2. We use the approximately four-billion-parameter base variant of FLUX.2 [klein] (Black Forest Labs, 2024), a rectified-flow transformer. We attach the memory branch to its double-stream blocks, which jointly process text and image tokens, and leave the subsequent single-stream blocks unchanged. The eight memory vectors lie in the text-feature space consumed by the frozen context embedder. In each double-stream block, image queries attend to memory keys and values obtained by passing these vectors through the context embedder and the text-side key/value projections, with biases removed. We apply the corresponding key normalization and assign every memory token the first text rotary position. The memory branch output passes through the frozen image output projection without bias and is added to the image stream. For this backbone, we use the native flow-matching objective in place of the diffusion denoising objective. Given a packed target latent $\mathbf{z}_0$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the velocity target is $\mathbf{v} = \boldsymbol{\epsilon} - \mathbf{z}_0$. We sample timesteps from the flow-matching schedule with its image-sequence-dependent shift and average eight independent timestep-noise pairs per optimization step. Training uses AdamW with learning rate 10^{-2} and no weight decay. Empty-prompt conditioning is fixed throughout. Sampling uses 50 flow-matching Euler steps at resolution 1024×1024. At each step, we combine memory-conditioned and zero-memory predictions using the shared guidance scale $\omega = 3$.

E Additional experiments

E.1 Out-of-distribution targets

Setup. Out-of-distribution targets are real photographs at 768px: faces from FFHQ (Karras et al., 2019), and animals, vehicles, and plants from DOCCI (Onoe et al., 2024), a corpus of photographs with detailed human-written descriptions. We train 80 single-image memories at $M{=}8$ (20 per class), 40 of which (10 per class) are additionally trained at every $M \in \{1, 8, 32, 64\}$ for the size comparison. These targets were never generated by the model, and no caption is involved at any stage; training, sampling, and evaluation follow Appendices C and D.1 unchanged, except that no target has a generating seed, so no seed exclusion applies.

The same two regimes, a narrower window. Figure 5 reproduces the trajectory of Fig. 1 on photographs: the semantic bias^2 drops before the surface bias^2 later, and the diversity of the generated images collapses only when memorization is completed. The memory size M is again important only for the final surface residual, rather than the for the timescale of early semantic acquisition. However, the separation between the two regimes contracts, and the per-image median ratio of surface to semantic half-decay steps is 10.6, against 21.4 for in-distribution targets. The contraction is visible in the grid itself, whose abstraction columns sit at steps 75–400 versus 75–600 for in-distribution targets. This phenomenon could be expected from the fact that a photograph's semantic content is less readily expressible through the prior, so less of the image can be delegated to it. As we



Figure 5: **Abstraction and locking on real photographs.** The external data counterpart of Fig. 1, same layout. **(a)** three memories (M=8), each trained on a single FFHQ photograph (red frame) and sampled at held-out noise seeds: prior draws at initialization, other people sharing the target's type, then the target's identity with expression, lighting and background still varying, then verbatim recall. **(b)** bias^2 and variance over all 80 single-image memories at M=8 (20 per class; median±IQR), semantic (DINOv2) and surface (VAE) spaces as in Fig. 1b. **(c)** the same distances by memory size over the 40 images trained at every M (10 per class).

discuss in the Limitations, the same mechanism could lead to the disappearance of the abstraction phase for strongly out-of-distribution targets.

Addressing. The positional-fraction transition of Appendix D.3 was nearly identical on out-of-distribution memories (20 runs per size, animals and vehicles): a flat content-driven plateau through abstraction, followed by a size-ordered rise as locking sets in, above the M=1 drift baseline. Thus, the mechanism of Section 5 extends beyond model-generated targets.

E.2 Memory-norm growth

Figure 6 tracks the mean per-token norm of the memory vectors, referenced from Section 3. We can notice three interesting phenomena: first of all, the norm grows by roughly two orders of magnitude, from 0.32 at initialization to a median of 29–38 at step 5,000, without saturating; in addition to that, the trajectories are essentially independent of memory size: M=1 through M=64 overlap throughout, so memory size acts through the *number* of tokens, not through single large tokens; finally, the M=1 memory grows its norm exactly like the larger ones while never completing locking (Fig. 1c), which implies that reconstruction quality is not norm-limited, consistently with the addressing account of Section 5.

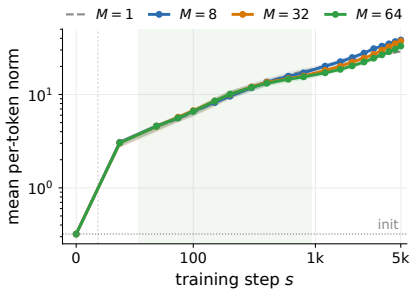

Figure 6: **Memory-norm growth.** Mean per-token norm over training (ID bank, 30 runs per memory size, median with IQR band, log scale). Dotted: the initialization norm $\|c\| \approx 0.32$.

E.3 Memorizing through Low Rank Adaptation

We study an alternative route to memorization in SD2.1, closer to the implementation in (Ruiz et al., 2023). Instead of adding a parallel memory cross-attention pathway, we keep the architecture of SD2.1 unaltered, but we learn a lightweight correction to the weights of SD2.1's UNet. The optimization objective is still a single image denoising objective analogous to equation 2, but now the optimization variables are the LoRA parameters.

In particular, here we consider a rank-1 correction to all weights of the UNet excluding the text-cross attention weights. This exclusion provides an ablation on the necessity of a direct interaction with

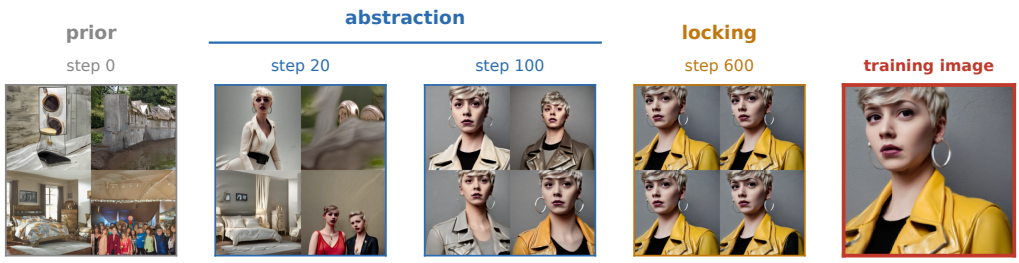

Figure 7: **Abstraction phase with LoRA**: A single SD2.1 native image, generated with a curated caption, is then memorized via a rank-1 correction of the UNet's weights (cross-attention excluded). *Abstraction* then *locking* emerge, in analogy to the experiments reported in the Main Text.

the text semantic space. In Fig. 7, we show that an abstraction phase still emerges, similar to that induced via the conditioning strategy.

E.4 Additional qualitative examples

The qualitative panels of the main text were selected for visual clarity. Here we show some additional examples.

Memorization trajectories, in-distribution. Figure 8 follows three memories from the bank of Section 3 across stored checkpoints at fixed evaluation seeds. The trajectory of Fig. 1 is reproduced in all three.

Memorization trajectories, external targets. Figure 9 repeats the layout on the external bank of Appendix E.1, including a plant, a class which we have not yet shown. The same ordering holds in every block, with general details emerging before specific ones: for example, the airport is fixed before the specific aircraft, and "roses in general" emerge before the specific target rose.

Generalization at random seeds. Figure 10 shows the generalization experiment of Section 5 at held-out seeds drawn at random rather than the top-ranked seeds of Fig. 2a, for three face captions and the animal caption. The results are consistent with Section 5: at step 100 most seeds follow their reference in pose and framing, some only weakly, and one has not yet left the prior at all.

Editability, additional examples. Figure 11 extends Fig. 3a with five more same-identity and five more different-identity U-turn examples. The same trend holds throughout: the edit survives at $s{=}0$ and $s{=}200$ and is increasingly overridden by $s{=}5000$. At $s{=}5000$ this override is not always a clean reversion to the stored identity: in several rows (e.g. the last one) the memory instead fails to reconstruct a coherent face at all, collapsing into a blurred, unrecognizable output.

Video memories, additional examples. Figure 12 extends the CogVideoX-2b result of Fig. 4 with four single-clip memories spanning distinct kinds of motion: breaking *waves*, an *octopus* waving its tentacles, a flying *butterfly*, and a *spinning top*. The same trend holds throughout: at initialization the memory returns unrelated draws from the prior, in the abstraction regime it fixes the scene and its characteristic motion (the wave breaking, the tentacles curling, the butterfly drifting, the top spinning), while the instance-specific appearance still varies across noise seeds (the umbrella's shape, the octopus' body, the wing pattern, the top's silhouette), and during locking every seed converges to the training clip.

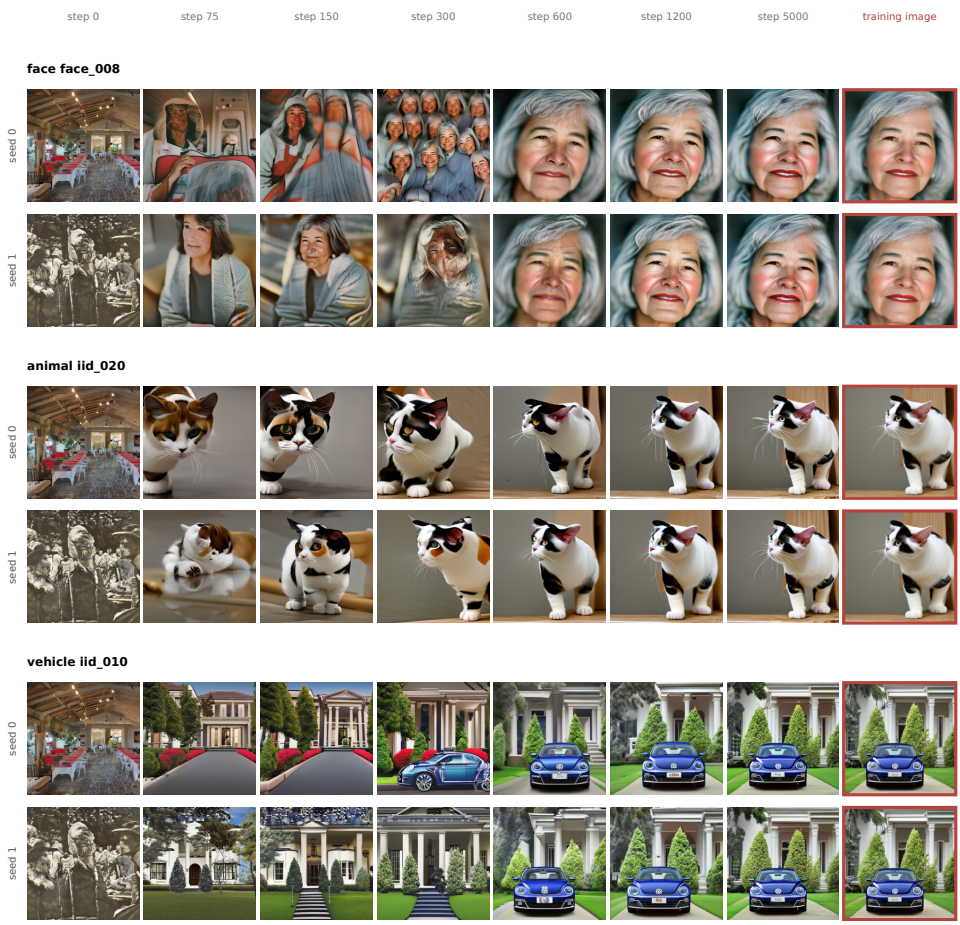

Figure 8: **Additional memorization trajectories, in-distribution.** Rows: the first two evaluation seeds of each run (a target's own generating seed is skipped). Columns: seven stored checkpoints and the training image (red frame). The face was drawn at random among those not shown in Fig. 1. The animal and the vehicle are the runs whose ratio of surface to semantic half-decay steps is the median of their class.

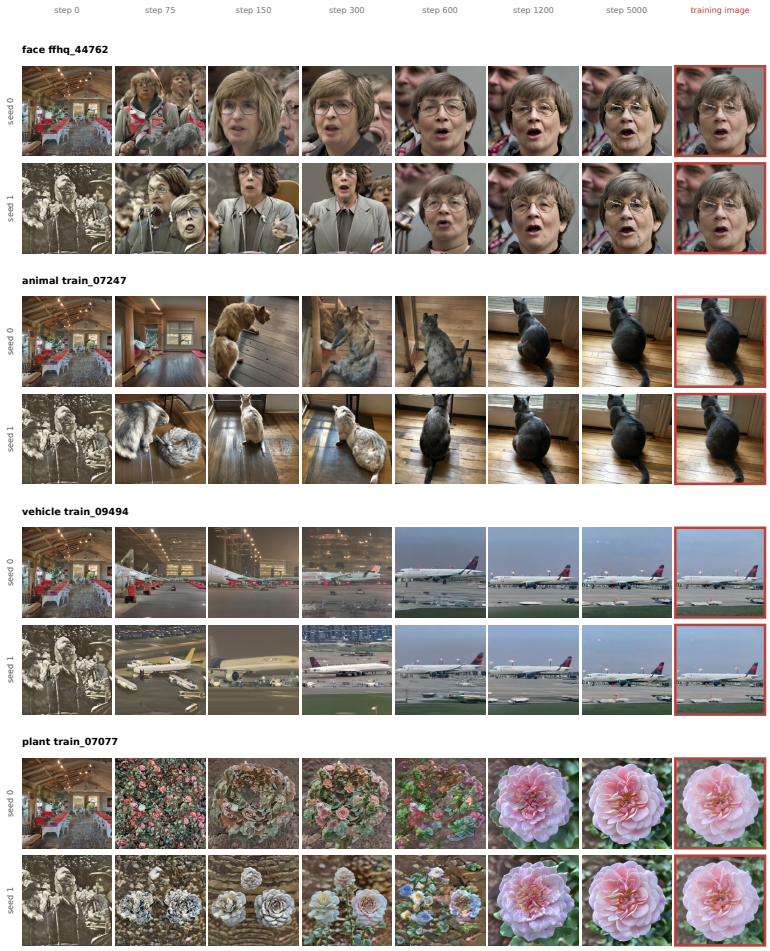

Figure 9: **Memorization trajectories, external targets.** Same layout as Fig. 8 on the bank of Appendix E.1. The animal and the vehicle are the runs whose ratio of surface to semantic half-decay steps is the median of their class.

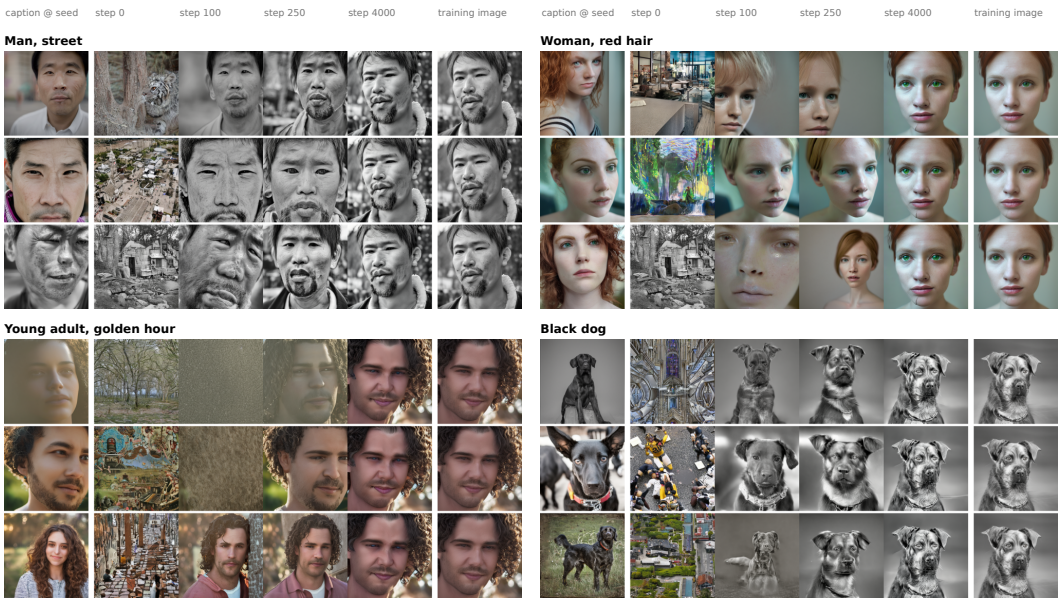

Figure 10: **Generalization at random test seeds.** Three face captions at the median of the twelve by peak advantage, and a black dog picture generated by SD2.1, each at three held-out seeds drawn at random. Columns as in Fig. 2a.

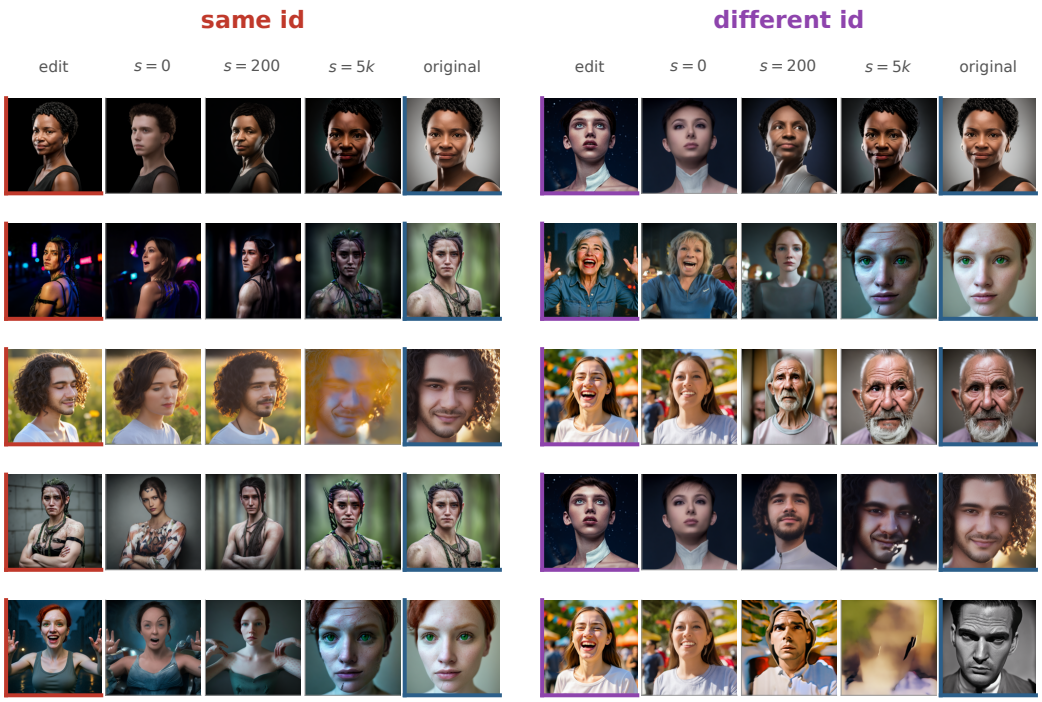

Figure 11: **Additional U-turn qualitative examples**, extending Fig. 3a with five more same-identity rows (left, red border) and five more different-identity rows (right, purple border). The protocol is the same as in Appendix D.4. Each row shows, left to right, the edited input x_{edit}, the U-turn output $\hat{x}$ at the untrained memory (s=0), an abstract checkpoint (s=200), a memorized checkpoint (s=5000), and the stored target x_{orig} (blue border). In same-identity rows the memory belongs to the same person shown in the edit; in different-identity rows it belongs to a different person, so the output overwrites the edit with a foreign identity as memorization progresses. Across both groups, the edit is largely preserved at s=0 and s=200 and is overridden by s=5000, which is consistent with what is observed in Fig. 3a.

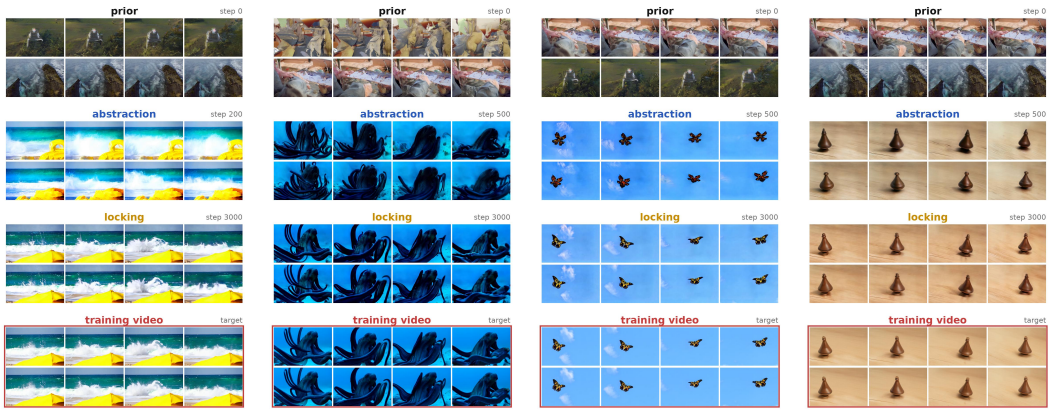

Figure 12: **Additional examples (CogVideoX-2b).** Each column trains one memory on a single target clip on top of the frozen video model and samples it at held-out noise seeds; columns are, left to right, breaking *waves*, an *octopus* waving its tentacles, a flying *butterfly*, and a *spinning top*. Rows follow the training trajectory top to bottom: draws from the prior at initialization, the abstraction regime, locking, and the training clip (red frame). Each block shows two seeds, each rendered as a filmstrip of four time-ordered frames. In the abstraction regime the memory fixes the scene and its characteristic *motion* (the wave breaking, the tentacles curling, the butterfly drifting across the sky, the top spinning) while the instance-specific appearance still varies across seeds (the umbrella's shape, the octopus' body, the butterfly's wing pattern, the top's silhouette). Only during the locking phase does every seed converge to the target clip.