LEARNING TO GENERATE CONTINUALLY IN LATENT SPACE VIA BURES-WASSERSTEIN OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

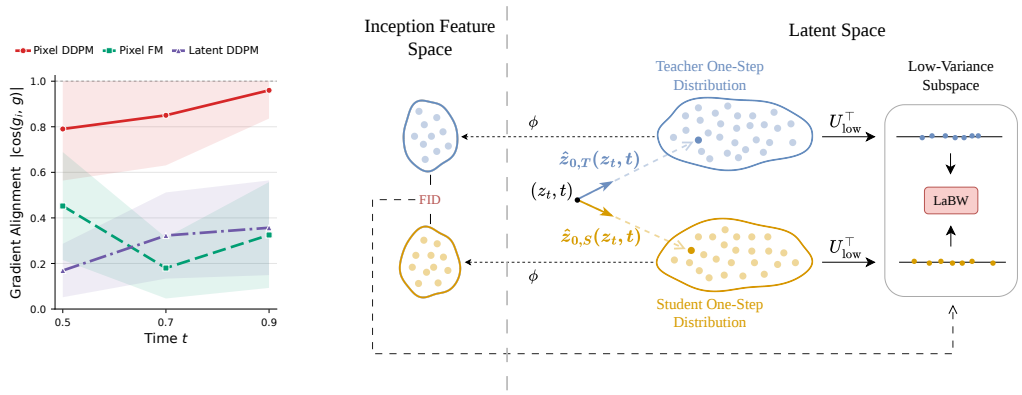
Continual learning for image generation aims to enable models to learn and synthesize images from new domains while preserving their ability to generate samples from previously learned ones, thereby avoiding the substantial computational cost of retraining modern generative models from scratch. The key challenge is to learn from new data without forgetting previously acquired knowledge. Existing approaches typically combine generative replay or distillation with regularization techniques, such as elastic weight consolidation (EWC), using generated samples to approximate previously learned distributions and constraining parameter updates to mitigate forgetting. However, while effective for pixel-space diffusion, we find that the assumptions underlying these regularization approaches do not generalize well to latent generative processes, such as flow matching. In this work, we propose **La**tent **B**ures-**W**asserstein (LaBW), a general framework for continual learning in latent image generative models with theoretical support. Instead of constraining model parameters, LaBW operates in the autoencoder latent space of generative models in a way that indirectly optimizes the generated image distribution in the Inception feature space. Specifically, our theoretical analysis establishes a local connection between perturbations of the latent endpoint distribution and changes in FID, revealing increased sensitivity to equal absolute covariance changes along low-variance latent directions. We further evaluate our approach across diffusion and flow matching models with both U-Net and DiT backbones on multiple benchmarks (e.g., CIFAR-10, LSUN Churches, and CelebA-HQ), demonstrating a substantial reduction in forgetting while maintaining or improving FID on both original and new domains compared to recent state-of-the-art approaches.

1 INTRODUCTION

Continual learning was initially studied in the context of discriminative tasks, such as image classification and recognition (Li & Hoiem, 2016; Kirkpatrick et al., 2017; Rebuffi et al., 2017; Lopez-Paz & Ranzato, 2017; Shin et al., 2017). Unlike vanilla fine-tuning, which focuses solely on adapting a model to a new task, continual learning additionally requires the model to retain knowledge acquired from previously learned tasks. With the growing popularity of generative models, continual learning has recently been extended to generative modeling (Wu et al., 2018; Zhai et al., 2019; Zajac et al., 2023), with the goal of enabling models to learn and generate samples from previously unseen domains while preserving their ability to synthesize images from previously learned ones. This setting is particularly important for modern generative models, such as diffusion and flow matching models (Ho et al., 2020; Lipman et al., 2023), whose training and adaptation require substantial computational resources, making retraining from scratch prohibitively expensive in many practical scenarios. In both discriminative and generative settings, however, catastrophic forgetting remains a fundamental challenge (McCloskey & Cohen, 1989; Goodfellow et al., 2015; Wang et al., 2026).

To mitigate catastrophic forgetting in generative models, existing approaches typically combine generative replay with additional regularization techniques to preserve previously acquired knowledge. Replay (Shin et al., 2017; Wu et al., 2018; Zajac et al., 2023) approximates previously learned distributions using either raw training data or synthesized samples, while regularization (Seff et al., 2017; Zhai et al., 2019) introduces explicit constraints on model parameters or intermediate and output predictions. Building on this framework, recent studies strengthen replay with generative distilla-

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.



(a) Gradient collinearity

(b) Our proposed LaBW framework

Figure 1: (a) Mean absolute cosine similarities at later timesteps t between per-sample gradients g_i and their expectation g for generative models trained on MNIST. Gradient alignment is high for the pixel-space DDPM, but substantially lower for the latent DDPM and pixel-space flow matching models. (b) Overview of the LaBW framework. Teacher and student predictions at a shared intermediate latent (z_t, t) are converted into one-step endpoint distributions. Their covariances are matched through the LaBW loss, after projection onto a low-variance subspace, the choice of which is motivated by the increased sensitivity of FID to latent perturbations along low-variance eigendirections.

tion losses (Masip et al., 2025) and constrain updates to parameters that are important to previously learned tasks using low-rank representations of the Fisher information matrix (e.g., rank-1 approximation) (Wang et al., 2026). However, we observe that *the key per-sample gradient collinearity assumption* underlying these regularization approaches does not generalize well beyond pixel-space diffusion models, e.g., to latent diffusion and flow matching models. We attribute this generalization challenge to two factors. First, the theoretical analysis is strongly tied to the structure of the standard variance-preserving diffusion process (i.e., DDPMs (Ho et al., 2020)) with noise prediction, and therefore does not directly extend to other generative dynamics. Second, in latent generative models, the autoencoder decoder introduces an additional nonlinear mapping between the latent and image spaces, which changes the underlying geometry and can further weaken the gradient collinearity observed in the original pixel-space formulation. As empirically demonstrated in Fig. 1a, collinearity deteriorates substantially when moving to flow matching and latent diffusion models.

Motivated by these limitations, we propose to move the regularization from the model parameter space to the *latent space of the underlying autoencoder*, and directly *regulate the latent endpoint distribution*, guided by its connection to FID in the Inception feature space. To the best of our knowledge, this is the first replay-based continual learning framework for image generative models that performs regularization directly in the autoencoder latent space. This formulation provides a scalable adaptation space for modern latent generative models and offers a more robust learning signal than per-sample gradients by operating on the endpoint distribution. However, connecting perturbations in the latent space to changes in the generated image distribution is non-trivial, as the latent perturbations propagate through the decoder and the Inception feature extractor before affecting the feature space distribution. To theoretically establish this connection, we analyze the local sensitivity of FID to latent perturbations, revealing that perturbations along low-variance latent directions can have a larger effect, as detailed in Sec. 3.

Building upon our theoretical insights, we propose **La**tent **B**ures-**W**asserstein (LaBW), a continual learning framework that performs regularization in the latent space through a low-variance Bures-Wasserstein objective, which we refer to as the *LaBW loss*. Specifically, LaBW matches the covariances of the teacher and student one-step endpoint predictions within a low-variance latent subspace. We obtain these predictions from shared noised replay samples and estimate their moments using exponential moving averages to stabilize the loss during minibatch training. We evaluate LaBW across multiple benchmarks (e.g., CIFAR-10 (Krizhevsky et al., 2009), LSUN Churches (Yu et al., 2016), and CelebA-HQ (Karras et al., 2018)), covering both diffusion and flow matching models with U-Net (Ronneberger et al., 2015) and DiT (Peebles & Xie, 2023) backbones. Across these

diverse settings, LaBW consistently reduces catastrophic forgetting while preserving adaptation to new domains, and achieves competitive or improved FID on both previously learned and new domains compared to recent state-of-the-art approaches. Relative to Rank-1 Fisher EWC, LaBW reduces forgetting from 3.34 to 2.17 on CIFAR-10 and from 1.89 to 1.25 on LSUN Churches, while maintaining OOD performance comparable to the strongest continual learning baselines.

Overall, our contributions can be summarized as follows:

- We theoretically characterize how perturbations in the latent distribution affect feature-space FID through a local expansion, and identify an increased sensitivity to equal absolute covariance changes along low-variance directions, which we further validate empirically.
- Motivated by this analysis, we propose **LaBW**, a latent-space continual learning framework that regularizes the endpoint distribution by matching the covariances of teacher and student one-step predictions within a low-variance latent subspace.
- We extensively evaluate LaBW across multiple datasets and base generative model backbones, demonstrating substantially reduced forgetting while maintaining or improving adaptation to new domains compared to recent SOTAs.

2 Related Work

Diffusion and Flow Matching in Latent Space. Diffusion and flow matching models (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023) define continuous-time generative processes that transform a reference distribution, typically Gaussian, into a data distribution through a time-dependent probability path. A broad class of such paths can be expressed as an affine interpolation $x_t = a_t x_0 + \sigma_t \epsilon$, where x_0 is a data sample and $\epsilon \sim \mathcal{N}(0, I)$. Diffusion models are commonly parameterized to predict the score $s_\theta(x_t, t)$, noise $\epsilon_\theta(x_t, t)$, or x_0-prediction $\hat{x}_{0,\theta}(x_t, t)$ (Ho et al., 2020; Song et al., 2021; Karras et al., 2022), while flow models learn to predict the velocity field $v_\theta(x_t, t)$ (Lipman et al., 2023; Liu et al., 2023).

Importantly, modern large-scale generative models increasingly operate in an autoencoder latent space rather than directly in pixel space for better scalability, as exemplified by the Stable Diffusion family (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Dao et al., 2023). Despite differences in their underlying formulations and prediction targets, these models share a common generative latent space between the autoencoder and the generative backbone. This shared latent space is also the space in which our proposed LaBW method performs continual adaptation.

Continual Learning for Generative Models. Continual learning refers to a model's ability to learn from a sequence of new tasks while retaining knowledge from previous ones. A central challenge in this paradigm is catastrophic forgetting (Goodfellow et al., 2015; McCloskey & Cohen, 1989), in which the model's performance on earlier tasks degrades when trained on subsequent ones. In generative models, catastrophic forgetting manifests as a degradation in the model's ability to generate samples from previously learned data distributions (Wu et al., 2018; Zhai et al., 2019). In recent years, several approaches have been explored to mitigate this phenomenon in diffusion models. Zajac et al. (2023) evaluated the effectiveness of established continual learning methods, namely L_2 regularization, experience replay, and generative replay, when applied to diffusion models. Masip et al. (2025) subsequently showed that generative replay can lead to a loss of the model's denoising capabilities, and introduced generative distillation, which distills the reverse process of the diffusion model. More recently, Wang et al. (2026) complemented generative distillation with an EWC penalty derived from the approximately rank-1 structure of the empirical Fisher, which arises from gradient collinearity in the low SNR regions of the diffusion process. Our work builds on this line of replay-based methods by introducing a distribution-level objective that aims to preserve the previously learned data distributions, and, to the best of our knowledge, is the first one to extend replay-based continual learning to latent generative models.

Fréchet Distance. The Fréchet distance (Dowson & Landau, 1982) between two multivariate normal distributions P and R with means μ_P and μ_R and covariances Σ_P and Σ_R is defined as

$$\mathrm{FD}(P, R) = \|\mu_P - \mu_R\|_2^2 + \underbrace{\mathrm{tr}\left(\Sigma_P + \Sigma_R - 2\left(\Sigma_P^{1/2}\Sigma_R\Sigma_P^{1/2}\right)^{1/2}\right)}_{d_{\mathrm{BW}}{}^2(\Sigma_P, \Sigma_R)}. \tag{1}$$

The term $d_{\mathrm{BW}}(\Sigma_P, \Sigma_R)$ is the Bures-Wasserstein distance (Bhatia et al., 2019), which defines a metric on the space of positive semidefinite matrices. When the distributions are defined in the feature space of Inception-v3 (Szegedy et al., 2016), the resulting distance is the Fréchet Inception Distance (FID) (Heusel et al., 2017), which has become one of the most widely adopted metrics for evaluating image generative models. More recently, Yang et al. (2026) showed that FID can also be effectively optimized as a distributional training objective. The Fréchet distance forms the basis of both our sensitivity analysis and proposed framework, introduced in Sections 3 and 4, respectively.

3 Connecting Latent Perturbations to FID Sensitivity

We first present a theoretical analysis of the connection between local perturbations in the latent space and the resulting FID sensitivity, as a principled basis for distributional continual adaptation of latent generative models. In these models, the decoder defines a nonlinear, generally non-isometric map, with local geometry that can vary across latent positions and directions (Arvanitidis et al., 2018; Rao et al., 2026). Consequently, the magnitude of a change in the latent space does not directly determine the magnitude of the corresponding change after decoding (Berrada et al., 2025) and its subsequent effect on FID. This disconnect motivates us to study the local sensitivity of feature space FID to controlled perturbations of the latent distribution P_Z.

3.1 Local FID Sensitivity to Latent Perturbations

We derive a local expansion of FID under a mean square differentiable perturbation of the latent distribution, which connects FID sensitivity to the induced first-order changes in the mean and covariance of the Inception feature distribution.

Theorem 1 *Let $(Z_\varepsilon)_\varepsilon$ be a family of square-integrable latent random variables with $Z = Z_0 \sim P_Z$, mean square differentiable at $\varepsilon = 0$ with mean square derivative $V \in L^2$, i.e.,*

$$\lim_{\varepsilon \to 0} \left\| \frac{Z_\varepsilon - Z}{\varepsilon} - V \right\|_{L^2} = 0. \tag{2}$$

Let $X_\varepsilon = \phi(Z_\varepsilon) \sim P_X^{(\varepsilon)}$, where $\phi := \mathrm{Inc} \circ \mathrm{Dec}$ is the composite map from the latent to the Inception feature space. At $\varepsilon = 0$, $X = X_0 \sim P_X$ has mean μ_X and covariance Σ_X. Assume that $\Sigma_X \succ 0$ and that X_ε admits the first-order Taylor expansion $X_\varepsilon = X + \varepsilon J_\phi(Z)V + o(\varepsilon)$ in L^2. Under this assumption, the feature space mean $\mu_X^{(\varepsilon)}$ and covariance $\Sigma_X^{(\varepsilon)}$ are differentiable at $\varepsilon = 0$, with derivatives $\dot{\mu}_X(V)$ and $\dot{\Sigma}_X(V)$, respectively. Then, as $\varepsilon \to 0$,

$$\mathrm{FID}\left(P_X, P_X^{(\varepsilon)}\right) = \varepsilon^2 \underbrace{\left(\|\dot{\mu}_X(V)\|_2^2 + \frac{1}{2}\mathrm{tr}\left(\mathcal{L}_{\Sigma_X}\left[\dot{\Sigma}_X(V)\right] \dot{\Sigma}_X(V) \right) \right)}_{Q(V)} + o(\varepsilon^2), \tag{3}$$

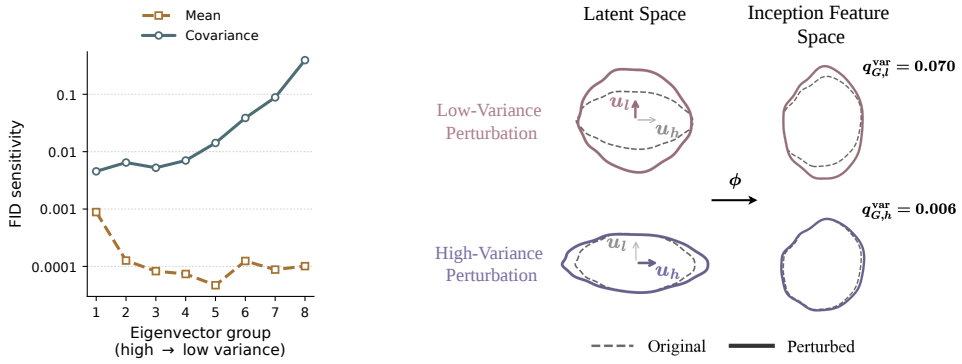
where $\mathcal{L}_{\Sigma_X}[H]$ denotes the unique symmetric solution A to the Lyapunov equation $A\Sigma_X + \Sigma_X A = H$.

Detailed proof of the theorem and sufficient conditions for the differentiability of the feature moments are provided in Appendix B. The term $Q(V)$ describes the local sensitivity of FID to the latent perturbation determined by its mean square derivative V.

3.2 Variants for Mean and Covariance Perturbations

To understand how this sensitivity depends on the latent covariance spectrum, we evaluate Q for controlled mean and covariance perturbations of P_Z along the latent covariance eigenvectors. Let μ_Z and Σ_Z denote the mean and covariance of P_Z, and let $\Sigma_Z = U\Lambda U^\top$, where $U = [u_1, \ldots, u_d]$ contains the orthonormal eigenvectors of Σ_Z and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ their corresponding eigenvalues. We study controlled perturbations along the eigenvectors u_i in order to characterize how the FID sensitivity varies along the latent covariance spectrum.

Latent Mean Perturbations. We first consider perturbations that preserve the latent covariance and induce the same absolute mean change along a latent covariance eigenvector.



(a) FID sensitivity across the spectrum (b) Feature space effects of equal-norm latent covariance perturbations

Figure 2: **FID sensitivity to latent mean and covariance perturbations on LSUN Churches**. (a) Sensitivity across eight groups of 512 latent eigenvectors, ordered from high to low variance. Covariance sensitivity increases toward the low-variance region, while mean sensitivity initially decreases and then remains approximately stable. (b) Equal-magnitude covariance perturbations in low- and high-variance groups and their effects under the decoder–Inception map ϕ. The low-variance perturbation has higher sensitivity (q_G^{var}) and induces a larger feature distribution change.

Proposition 1 *For an eigenvector u_i, let $Z_\varepsilon = Z + \varepsilon u_i$. This family is mean square differentiable at $\varepsilon = 0$, with mean square derivative $V_i^{\text{mean}} = u_i$. Its mean and covariance satisfy $\mu_Z^{(\varepsilon)} = \mu_Z + \varepsilon u_i$ and $\Sigma_Z^{(\varepsilon)} = \Sigma_Z$, so $\dot{\mu}_Z = u_i$ and $\dot{\Sigma}_Z = 0$. Its local FID sensitivity is $q_i^{\text{mean}} := Q\left(V_i^{\text{mean}}\right) = Q(u_i)$.*

Importantly, the perturbation does not introduce any explicit dependence on the latent eigenvalue λ_i.

Latent Covariance Perturbations. We next consider perturbations that preserve the latent mean and induce the same absolute variance change along a latent covariance eigenvector. Because a covariance change does not uniquely determine the corresponding distribution, we select, among the distributions with the desired moments, the one minimizing the 2-Wasserstein distance from P_Z.

Proposition 2 *Let $\Sigma_Z \succ 0$, and let u_i be an eigenvector of Σ_Z with associated eigenvalue λ_i. For sufficiently small ε, among the distributions with mean μ_Z and covariance $\Sigma_Z^{(\varepsilon)} = \Sigma_Z + \varepsilon u_i u_i^\top$ the minimum 2-Wasserstein distance from P_Z is attained by the positive semidefinite affine pushforward of P_Z (Nguyen et al., 2025). The corresponding random variable is*

$$Z_\varepsilon = Z + \left(\sqrt{1 + \frac{\varepsilon}{\lambda_i}} - 1\right) u_i u_i^\top (Z - \mu_Z), \tag{4}$$

whose mean square derivative at $\varepsilon = 0$ is

$$V_i^{\text{var}} = v_i^{\text{var}}(Z) = \frac{1}{2\lambda_i} u_i u_i^\top (Z - \mu_Z) = \frac{1}{2\sqrt{\lambda_i}} \eta_i u_i, \qquad \eta_i := \frac{u_i^\top (Z - \mu_Z)}{\sqrt{\lambda_i}}, \tag{5}$$

where η_i denotes the standardized coordinate along u_i. Because Q is quadratic in V, its local FID sensitivity is $q_i^{\text{var}} := Q\left(V_i^{\text{var}}\right) = \frac{1}{4\lambda_i} Q\left(\eta_i u_i\right)$.

The derivation is provided in Appendix C.2. Thus, for equal absolute variance changes, the local FID sensitivity contains an explicit factor inversely proportional to λ_i, while the remaining term $Q\left(\eta_i u_i\right)$ depends on the latent distribution, the decoder-Inception map, and the resulting FID geometry. To examine whether low-variance latent directions exhibit greater FID sensitivity empirically, we measure sensitivity to covariance perturbations over groups of eigenvectors ordered by their corresponding eigenvalues (see Appendix C.3 for details). As shown in Fig. 2, the sensitivity strongly increases toward the low-variance region of the spectrum. These results suggest that preserving the covariance along low-variance latent directions can help limit FID degradation.

4 LaBW for Continual Learning in Latent Space

Building on our sensitivity analysis, we introduce LaBW (**La**tent **B**ures-**W**asserstein), a framework for continual learning in latent generative models that aims to preserve the teacher's marginal endpoint distribution through covariance matching. By operating on a common endpoint representation, LaBW is agnostic to the specific diffusion or flow matching parameterization.

4.1 Problem Formulation

We consider continual learning for class-conditional generative models that operate in the learned latent space of an autoencoder (Rombach et al., 2022; Dao et al., 2023; Vahdat et al., 2021). A model $q_T(z \mid y)$ is pretrained on an in-domain (ID) distribution over a set of classes $\mathcal{Y}_{\mathrm{ID}}$. During continual adaptation, q_T is kept frozen as the teacher model, and a student model $q_\theta(z|y)$, initialized from q_T, is subsequently trained on data from an out-of-domain (OOD) distribution over classes $\mathcal{Y}_{\mathrm{OOD}}$, under the assumption that the ID data are no longer available. We further assume the imbalanced setting in which $|D_{\mathrm{ID}}| \gg |D_{\mathrm{OOD}}|$. This setup reflects a *realistic continual adaptation scenario* for large generative models, where the original training data are no longer accessible and newly available OOD data are typically much more limited. Following prior work (Zajac et al., 2023; Masip et al., 2025; Wang et al., 2026), we quantify forgetting ($\mathcal{F}$) as the model's performance degradation on the ID distribution, as measured by FID, $\mathcal{F} = \mathrm{FID}_{\mathrm{ID}}(q_\theta) - \mathrm{FID}_{\mathrm{ID}}(q_T)$.

4.2 Key Components of Proposed LaBW

We now describe the main components of the LaBW framework in the continual learning setting described above.

One-Step Endpoint Distributions. LaBW aims to prevent the student's endpoint marginal distribution from deviating from that of the teacher during adaptation. Directly estimating these distributions, however, would require repeatedly running the full generative process. Instead, we construct one-step endpoint estimates from intermediate latent points and use their distributions as tractable proxies. Let $\mathcal{B}_T$ denote the replay buffer generated by the teacher model using the sampler selected for each setting. At each training iteration, we sample a minibatch $\{(z_0^{(b)}, y^{(b)})\}_{b=1}^B$ from $\mathcal{B}_T$. For each sample b, we independently sample $\epsilon \sim \mathcal{N}(0, I)$ and $t \sim \mathrm{LogitNormal}(0, 1)$, following Esser et al. (2024), and construct the intermediate latent point $z_t^{(b)}$ according to the generative process. The teacher and student are then evaluated on the same points $\{(z_{t^{(b)}}^{(b)}, t^{(b)}, y^{(b)})\}_{b=1}^B$, and their outputs are converted to one-step endpoint estimates $\hat{z}_{0,T}$ and $\hat{z}_{0,S}$, respectively (see Appendix A for further details). These estimates define the teacher and student one-step endpoint distributions.

EMA Moment Estimation. To obtain stable estimates of the covariances of these distributions during minibatch training, we follow Yang et al. (2026) and maintain exponential moving averages (EMA) of their first and uncentered second moments, thereby decoupling the effective population size used to estimate the distributional loss from the batch size used for the gradient computation. For $a \in \{T, S\}$ and EMA decay rate $\beta \in (0, 1)$, the batch mean and uncentered second moments at training iteration k are given by:

$$\mu_{\mathrm{batch},a}^{(k)} = \frac{1}{B}\sum_{b=1}^{B} \hat{z}_{0,a}^{(b)}, \qquad M_{\mathrm{batch},a}^{(k)} = \frac{1}{B}\sum_{b=1}^{B} \hat{z}_{0,a}^{(b)} \hat{z}_{0,a}^{(b)\top}. \tag{6}$$

Their running estimates are then updated as:

$$\begin{gathered} \bar{\mu}_a^{(k)} = \beta \bar{\mu}_a^{(k-1)} + (1-\beta)\mu_{\mathrm{batch},a}^{(k)}, \qquad \bar{M}_a^{(k)} = \beta \bar{M}_a^{(k-1)} + (1-\beta) M_{\mathrm{batch},a}^{(k)}, \\ \bar{\Sigma}_a^{(k)} = \bar{M}_a^{(k)} - \bar{\mu}_a^{(k)} \bar{\mu}_a^{(k)\top}. \end{gathered} \tag{7}$$

Low-Variance Bures-Wasserstein Loss. Having obtained the moment estimates, we now define the distributional objective of LaBW, which matches the teacher and student one-step endpoint covariances using the Bures-Wasserstein distance. Guided by our sensitivity analysis in Sec. 3, we

restrict this matching to a low-variance subspace of the teacher replay covariance. Let Σ_T denote the teacher covariance estimated from the replay buffer $\mathcal{B}_T$. We eigendecompose

$$\Sigma_T = U\Lambda U^\top, \qquad U = [u_1, \dots, u_d] \qquad \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d), \qquad \lambda_1 \geq \cdots \geq \lambda_d. \tag{8}$$

For a subspace dimension r, we define $U_{\mathrm{low}}^{(r)} \coloneqq [u_{d-r+1}, \dots, u_d]$, which spans the r lowest-variance directions of Σ_T. At each training step, the EMA covariance estimates of the teacher and student one-step endpoint distributions are projected onto this subspace as $\widetilde{\Sigma}_a = U_{\mathrm{low}}^{(r)\top} \bar{\Sigma}_a U_{\mathrm{low}}^{(r)}$, $a \in \{T, S\}$. The resulting LaBW loss is:

$$\mathcal{L}_{\mathrm{BW}} = d_{\mathrm{BW}}^2 \left(\widetilde{\Sigma}_T, \widetilde{\Sigma}_S \right). \tag{9}$$

This loss corresponds to the covariance term of the Fréchet distance, whose corresponding mean term we omit (see Appendix D.2 for further discussion). Our full training objective then becomes:

$$\mathcal{L} = \mathcal{L}_{\mathrm{OOD}} + \lambda_{\mathrm{GD}} \mathcal{L}_{\mathrm{GD}} + \lambda_{\mathrm{BW}} \mathcal{L}_{\mathrm{BW}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{OOD}}$ denotes the standard generative training objective on the OOD data and $\mathcal{L}_{\mathrm{GD}}$ the pointwise Generative Distillation loss on the replay samples.

5 Experiments

5.1 Experimental Setup

Continual Learning Settings and Datasets. We evaluate our method under two continual learning settings. First, in the ID/OOD setting described in Sec. 4.1, we consider four dataset-model configurations. At 32×32 resolution, we use classes 0-7 of CIFAR-10 (Krizhevsky et al., 2009) as the ID distribution and classes 8-9 as the OOD distribution, and train an LDM. At 256×256 resolution, we adapt from LSUN Churches (Yu et al., 2016) to AFHQ (Choi et al., 2020) using both an LDM and an LFM, and from CelebA-HQ (Karras et al., 2018) to Flowers102 (Nilsback & Zisserman, 2008) using an LFM. Subsequently, we evaluate on the class-incremental CIFAR-10 setting adopted by prior works (Masip et al., 2025; Wang et al., 2026), where the ten classes are divided into five sequential two-class tasks. Additional implementation details about the dataset sizes, model architectures, autoencoders, and latent dimensions are provided in Appendix D.1.

Evaluation Metrics. For the ID/OOD setting, we report the FID on both the ID and OOD distributions. Forgetting is quantified as the increase in ID FID relative to the pretrained teacher, as defined in Sec. 4.1. For the class-incremental CIFAR-10 setting, let m_k denote the model after learning task k. Following prior work, we report the average FID of all K tasks and the final average forgetting:

$$\mathrm{AFID@}k = \frac{1}{k} \sum_{i=1}^{k} \mathrm{FID}_i(m_k), \qquad \mathcal{F}_{\mathrm{avg}} = \frac{1}{K} \sum_{i=1}^{K} \left(\mathrm{FID}_i(m_K) - \mathrm{FID}_i(m_i) \right). \tag{11}$$

All results in the main experiments (Tables 1-3) are reported as the mean and standard deviation over three training seeds, after averaging over three evaluation seeds for each training seed.

Base Generative Models and Baselines. We compare our method against vanilla fine-tuning, Generative Replay (Zajac et al., 2023), Generative Distillation (GD) (Masip et al., 2025), and Rank-1 Fisher EWC (Wang et al., 2026). Fine-tuning trains the pretrained model using only the data from the new task, without any mechanism for preserving the previous ones.

5.2 Main Experiments and Ablation Analysis

Different ID/OOD Settings. We present the results on the four ID/OOD settings in Tables 1 and 2. LaBW consistently has the lowest forgetting across all four settings, while maintaining OOD performance comparable to GD, which generally achieves the strongest OOD results among the continual baselines. In contrast, naive fine-tuning obtains slightly better OOD performance, but, as expected, suffers from catastrophic forgetting on the ID distribution. Additionally, Fig. 3 presents qualitative examples of generated samples from the ID and OOD distributions. In the ID examples, LaBW preserves fine-grained details of the teacher samples that are sometimes diminished under the other continual learning baselines, while the OOD examples are visually similar across methods, consistent with the relatively small differences in OOD FID.

Table 1: **Quantitative results for the ID/OOD settings using LDMs.** Our method achieves the lowest forgetting while maintaining OOD performance comparable to GD and Rank-1 Fisher EWC.

	CIFAR-10			LSUN Churches → AFHQ		
Method	ID FID ↓	OOD FID ↓	$\mathcal{F}$ ↓	ID FID ↓	OOD FID ↓	$\mathcal{F}$ ↓
Fine-Tuning	92.42 ± 0.10	**10.05** ± 0.02	80.99 ± 0.10	178.37 ± 0.01	**16.09** ± 0.03	174.12 ± 0.01
Generative Replay	18.79 ± 0.22	10.30 ± 0.04	7.36 ± 0.22	31.89 ± 0.02	16.87 ± 0.05	27.65 ± 0.02
GD	15.15 ± 0.16	10.71 ± 0.22	3.72 ± 0.16	6.37 ± 0.10	16.46 ± 0.12	2.12 ± 0.10
Rank-1 Fisher EWC	14.77 ± 0.10	11.03 ± 0.13	3.34 ± 0.10	6.17 ± 0.06	16.65 ± 0.15	1.92 ± 0.06
LaBW (ours)	**13.60** ± 0.04	10.69 ± 0.04	**2.17** ± 0.04	**5.70** ± 0.03	16.62 ± 0.24	**1.46** ± 0.03

Table 2: **Quantitative results for the ID/OOD settings using LFMs.** LaBW reduces forgetting in both settings while maintaining comparable OOD FID to the strongest continual learning baselines.

	CelebA-HQ → Flowers102			LSUN Churches → AFHQ		
Method	ID FID ↓	OOD FID ↓	$\mathcal{F}$ ↓	ID FID ↓	OOD FID ↓	$\mathcal{F}$ ↓
Fine-Tuning	297.51 ± 0.26	**22.89** ± 0.21	290.19 ± 0.26	177.88 ± 0.05	**10.15** ± 0.10	171.55 ± 0.05
Generative Replay	21.59 ± 0.15	24.51 ± 0.24	14.27 ± 0.15	24.67 ± 0.08	11.37 ± 0.17	18.34 ± 0.08
GD	8.15 ± 0.03	23.39 ± 0.12	0.82 ± 0.03	9.37 ± 0.01	10.71 ± 0.25	3.04 ± 0.01
Rank-1 Fisher EWC	8.11 ± 0.04	23.71 ± 0.36	0.77 ± 0.02	8.22 ± 0.07	11.79 ± 0.23	1.89 ± 0.07
LaBW (ours)	**7.89** ± 0.03	23.44 ± 0.33	**0.56** ± 0.03	**7.58** ± 0.02	11.03 ± 0.22	**1.25** ± 0.02



Figure 3: **Qualitative results on the** 256×256 **ID/OOD settings.** Each column uses the same initial Gaussian noise sample across all methods. In some examples, LaBW better preserves fine-grained details from the teacher ID samples, while the OOD samples remain visually similar across methods.

Class-Incremental Setting. To provide a direct comparison with prior baselines in their adopted setting, we additionally evaluate our method on the class-incremental CIFAR-10 setting described in Sec. 5.1. As shown in Table 3, our method improves both final average FID and average forgetting compared to the baselines, with the final model achieving the lowest per-task FID across all 5 tasks.

Ablation on the Covariance Subspace. We ablate the choice of the latent covariance subspace used by LaBW. In addition to the low-variance subspace, we consider the corresponding high-variance subspace of the same rank and the full latent space. Because these variants can differ substantially in the scale of the resulting loss, we tune the LaBW weight to obtain comparable gradient norms. The results are reported in Table 4. On CIFAR-10, training on the low-variance subspace leads to better OOD adaptation and substantially lower forgetting, supporting our theoretical motivation. On CelebA-HQ → Flowers102, the three variants perform similarly. Further analysis shows that they achieve comparable covariance preservation across the latent spectrum, regardless of the subspace on which the loss operates. Additional discussion is provided in Appendix D.2.

Ablation on the LaBW Loss Weight. Lastly, we study how varying the LaBW loss weight affects forgetting and OOD adaptation. The results are presented in Table 5. Across all three settings, increasing the weight consistently reduces forgetting. For the CelebA-HQ → Flowers102 and

Table 3: **Quantitative results for the five-task class-incremental CIFAR-10 setting.** LaBW improves final average FID and average forgetting, and achieves the lowest final FID on every task.

	Average Metrics		Per-Task FID				
Method	AFID@5 ↓	Avg. Forgetting ↓	Task 0 ↓	Task 1 ↓	Task 2 ↓	Task 3 ↓	Task 4 ↓
GD	22.74 ± 0.26	6.39 ± 0.23	23.61 ± 0.31	31.62 ± 0.29	26.84 ± 0.39	19.42 ± 0.17	12.22 ± 0.18
Rank-1 Fisher EWC	22.97 ± 0.25	6.25 ± 0.17	23.46 ± 0.20	32.02 ± 0.27	27.56 ± 0.44	19.66 ± 0.35	12.19 ± 0.12
LaBW (ours)	**21.41** ± 0.25	**5.01** ± 0.19	**20.76** ± 0.41	**30.04** ± 0.39	**25.71** ± 0.35	**18.64** ± 0.17	**11.91** ± 0.11

Table 4: **Quantitative results for the latent covariance subspace ablation in LaBW.** The low-variance subspace substantially improves performance on CIFAR-10. All three variants perform similarly on CelebA-HQ → Flowers102.

	CIFAR-10			CelebA-HQ → Flowers102		
Subspace	ID FID ↓	OOD FID ↓	$\mathcal{F}$ ↓	ID FID ↓	OOD FID ↓	$\mathcal{F}$ ↓
Low-variance	**13.60** ± 0.04	**10.69** ± 0.04	**2.17** ± 0.04	7.89 ± 0.03	23.44 ± 0.33	0.56 ± 0.03
High-variance	16.26 ± 0.08	11.14 ± 0.18	4.83 ± 0.08	7.93 ± 0.02	**23.41** ± 0.31	0.60 ± 0.02
Full latent space	15.17 ± 0.11	11.37 ± 0.07	3.74 ± 0.11	**7.81** ± 0.09	23.64 ± 0.27	**0.48** ± 0.09

Table 5: **Ablation of the LaBW loss weight.** Results are obtained from one training seed and reported as the mean and standard deviation across three evaluation seeds.

Setting	LaBW Weight	ID FID ↓	OOD FID ↓	Forgetting ↓
CIFAR-10	2.0	14.32 ± 0.13	**10.42** ± 0.06	2.89 ± 0.13
	4.0	13.63 ± 0.06	10.69 ± 0.12	2.20 ± 0.08
	6.0	**13.16** ± 0.10	10.87 ± 0.09	**1.73** ± 0.05
CelebA-HQ → Flowers102	0.25	7.95 ± 0.07	**23.96** ± 0.13	0.62 ± 0.02
	0.50	7.85 ± 0.07	24.13 ± 0.14	0.52 ± 0.01
	1.00	**7.72** ± 0.08	24.40 ± 0.16	**0.39** ± 0.01
LDM LSUN Churches → AFHQ	0.10	5.79 ± 0.06	16.97 ± 0.16	1.54 ± 0.05
	0.15	5.63 ± 0.02	**16.66** ± 0.14	1.39 ± 0.01
	0.25	**5.48** ± 0.03	16.73 ± 0.13	**1.23** ± 0.02

CIFAR-10 settings, OOD FID also increases gradually with the weight, highlighting the expected retention-adaptation trade-off. For LSUN Churches → AFHQ, increasing the weight initially improves both retention and adaptation, while a further increase continues to reduce forgetting at a small cost in OOD performance.

6 Conclusion and Discussions

In this work, we study continual learning for latent generative models from a distributional perspective and introduce LaBW, a framework for preserving previously learned distributions through covariance matching in the latent space. Our theoretical analysis connects local perturbations of a latent distribution to FID sensitivity and motivates prioritizing low-variance latent directions. In experiments across latent diffusion and flow matching models, LaBW consistently reduces forgetting while maintaining performance on the new task comparable to existing continual learning baselines, in both ID/OOD and class-incremental settings.

Limitations and Future Work. The theoretical analysis characterizes FID sensitivity across the spectrum, but cannot identify which direction will be more affected during continual learning. This motivates exploring more flexible methods for selecting the latent subspace that account for both sensitivity and the distributional changes induced by adaptation. Moreover, since LaBW relies on one-step endpoint estimates as tractable proxies for the teacher and student marginal distributions, more accurate ways of approximating them could be developed. Lastly, future work can explore alternative distributional objectives that capture more informative aspects of the changes induced by continual learning, such as differences beyond second-order statistics.

AI USE STATEMENT

In this work, we used AI tools for language refinement during the writing (e.g., suggestions on phrasing and clarity), to support literature search and discovery, and to assist with research execution, namely the implementation and debugging of experiments. AI writing suggestions were used as a reference. Papers identified through AI-assisted literature search were subsequently read in full and assessed by the authors, and all citations and bibliographic entries were added manually. AI-assisted code was reviewed and tested for correctness. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims, experimental results or artifacts.

ETHICS STATEMENT

Our work aims to improve continual learning for generative models. It does not involve the collection of new data or human-subject experiments and uses established, publicly available datasets and pretrained generative models. As with generative modeling more broadly, this work may have societal implications related to the use of generative artificial intelligence, and models adapted using our method may inherit biases or other undesirable properties from their training data or pretrained models. We do not foresee, however, distinct ethical risks beyond those associated with current generative models.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our theoretical and empirical results throughout the paper and the appendices. The assumptions and derivations related to our FID sensitivity analysis are stated in Sec. 3, with proofs and additional technical details provided in Appendices B and C. The LaBW framework is described in Sec. 4, experimental settings and evaluation metrics are outlined in Sec. 5, and additional details on model architectures, training hyperparameters, evaluation settings, and ablation studies are provided in Appendix D. An anonymous implementation of our method and the configurations used in our experiments are provided in the supplementary material.

REFERENCES

J. Aitchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

Tariq Berrada, Pietro Astolfi, Melissa Hall, Marton Havasi, Yohann Benchetrit, Adriana Romero-Soriano, Karteek Alahari, Michal Drozdzal, and Jakob Verbeek. Boosting latent diffusion with perceptual objectives. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 12128–12141, 2025.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023.

D.C Dowson and B.V Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

Nicolas Dufour, Alexei A. Efros, and Patrick Pérez. The fid lottery: Quantifying hidden randomness in generative model evaluation. *arXiv preprint arXiv:2606.20536*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations*, 2014.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629. Springer, 2016.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein riemannian geometry of positive definite matrices, 2018.

Sergi Masip, Pau Rodriguez, Tinne Tuytelaars, and Gido M van de Ven. Continual learning of diffusion models with generative distillation. In Vincenzo Lomonaco, Stefano Melacci, Tinne Tuytelaars, Sarath Chandar, and Razvan Pascanu (eds.), *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 431–456. PMLR, 2025.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pp. 109–165. Academic Press, 1989.

Viet Anh Nguyen, Soroosh Shafiee, Damir Filipović, and Daniel Kuhn. Mean-covariance robust risk measurement, 2025.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, ICVGIP '08, pp. 722–729. IEEE Computer Society, 2008.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Mingxing Rao, Bowen Qu, and Daniel Moyer. Latent diffusion inversion requires understanding the latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34971–34980, 2026.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental Classifier and Representation Learning . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542. IEEE Computer Society, 2017.

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 (CVPR)*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer International Publishing, 2015.

Ari Seff, Alex Beatson, Daniel Suo, and Han Liu. Continual learning in generative adversarial nets, 2017.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Christian Soize. *The Fokker-Planck Equation for Stochastic Dynamical Systems and Its Explicit Steady State Solutions*. 1994.

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*, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11287–11302. Curran Associates, Inc., 2021.

Zekun Wang, Anant Gupta, Zihan Dong, and Christopher J. MacLellan. Avoid catastrophic forgetting with rank-1 fisher from diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Memory replay gans: Learning to generate new categories without forgetting. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation. *arXiv:2604.28190*, 2026.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2016.

Michał Zajac, Kamil Deja, Anna Kuzina, Jakub M Tomczak, Tomasz Trzciński, Florian Shkurti, and Piotr Miłoś. Exploring continual learning of diffusion models. *arXiv preprint arXiv:2303.15342*, 2023.

Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2759–2768, 2019.

APPENDICES

The appendix is organized as follows. Sec. A describes how the one-step endpoint estimates used by our method are constructed from model predictions at (z_t, t) under different diffusion and flow matching parameterizations. Sec. B provides the technical details for the local FID sensitivity analysis. Sec. C derives the latent mean and covariance perturbations considered in our analysis, as well as extensions of the covariance perturbations to groups of covariance eigenvectors. Finally, Sec. D presents additional experimental results and ablation studies.

A DIFFUSION AND FLOW MATCHING PARAMETERIZATIONS

Our method operates on distributions of one-step endpoint estimates $\hat{z}_0$ obtained from model predictions at intermediate latents (z_t, t). Because diffusion and flow matching models can be trained with different prediction targets, the conversion from the model output to an endpoint estimate depends on the underlying parameterization. The conversions used in our experiments are summarized below. For notational simplicity, we omit the class condition y. We consider the affine interpolation

$$z_t = a_t z_0 + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{12}$$

where a_t and σ_t are the signal and noise coefficients at time t.

Diffusion Models. Diffusion models are commonly parameterized to predict the noise $\epsilon_\theta(z_t, t)$, the score $s_\theta(z_t, t)$, or the endpoint $\hat{z}_{0,\theta}(z_t, t)$. For noise prediction, from Eq. 12 we have

$$\hat{z}_{0,\theta} = \frac{z_t - \sigma_t \epsilon_\theta(z_t, t)}{a_t}. \tag{13}$$

For score prediction, using $\epsilon_\theta(z_t, t) = -\sigma_t s_\theta(z_t, t)$ gives

$$\hat{z}_{0,\theta} = \frac{z_t + \sigma_t^2 s_\theta(z_t, t)}{a_t} \tag{14}$$

For endpoint prediction, the model output is used as the endpoint estimate:

$$\hat{z}_{0,\theta} = \hat{z}_{0,\theta}(z_t, t) \tag{15}$$

Flow Matching. For the affine interpolation in Eq. 12, differentiating with respect to t gives

$$\dot{z}_t = \dot{a}_t z_0 + \dot{\sigma}_t \epsilon. \tag{16}$$

If the model is trained to predict the velocity $v_\theta(z_t, t)$, by solving Equations 12 and 16 for z_0, we have

$$\hat{z}_{0,\theta} = \frac{\dot{\sigma}_t z_t - \sigma_t v_\theta(z_t, t)}{a_t \dot{\sigma}_t - \dot{a}_t \sigma_t}. \tag{17}$$

In our flow matching experiments, we use the linear interpolation $z_t = (1-t)z_0 + t\epsilon$, where $a_t = 1 - t$ and $\sigma_t = t$, so Equation 17 simplifies to

$$\hat{z}_{0,\theta} = z_t - t v_\theta(z_t, t). \tag{18}$$

B TECHNICAL DETAILS FOR THE LOCAL FID SENSITIVITY ANALYSIS

B.1 FEATURE MOMENT EXPANSIONS

We consider $(Z_\varepsilon)_\varepsilon$ as a second-order random process in $\mathbb{R}^d$, indexed by ε. Following the standard definition of mean square differentiability (Soize, 1994), $(Z_\varepsilon)_\varepsilon$ is mean square differentiable at $\varepsilon = 0$ with mean square derivative V if

$$\lim_{\varepsilon \to 0} \left\| \frac{Z_\varepsilon - Z}{\varepsilon} - V \right\|_{L^2} = 0, \tag{19}$$

where for a square-integrable random vector Y,

$$\|Y\|_{L^2} \coloneqq (\mathbb{E}\left[\|Y\|_2^2\right])^{1/2}. \tag{20}$$

Let $X_\varepsilon = \phi(Z_\varepsilon)$ and $X = X_0$, where $\phi : \mathbb{R}^d \to \mathbb{R}^D$ is the composite map from the latent to the Inception feature space. We assume that ϕ is differentiable at Z almost surely and that $X, J_\phi(Z)V \in L^2$, where $J_\phi(Z)$ is the Jacobian of ϕ at Z. We further assume that the following first-order Taylor expansion holds in L^2:

$$X_\varepsilon = X + \varepsilon J_\phi(Z)V + o(\varepsilon). \tag{21}$$

Equivalently,

$$\|X_\varepsilon - X - \varepsilon J_\phi(Z)V\|_{L^2} = o(\varepsilon). \tag{22}$$

Under this assumption, $(X_\varepsilon)_\varepsilon$ is mean square differentiable at $\varepsilon = 0$, with mean square derivative:

$$\dot{X}(V) = J_\phi(Z)V. \tag{23}$$

Let $\mu_X^{(\varepsilon)} \coloneqq \mathbb{E}[X_\varepsilon]$ and $\Sigma_X^{(\varepsilon)} \coloneqq \mathrm{Cov}(X_\varepsilon)$ denote the mean and covariance of the perturbed feature distribution. Taking expectations in Eq. 21 gives

$$\mu_X^{(\varepsilon)} = \mu_X + \varepsilon\mathbb{E}[\dot{X}(V)] + o(\varepsilon) \implies \dot{\mu}_X(V) \coloneqq \left.\frac{d\mu_X^{(\varepsilon)}}{d\varepsilon}\right|_{\varepsilon=0} = \mathbb{E}[\dot{X}(V)]. \tag{24}$$

The covariance can be written as:

$$\Sigma_X^{(\varepsilon)} = \mathbb{E}\left[X_\varepsilon X_\varepsilon^\top\right] - \mu_X^{(\varepsilon)}\mu_X^{(\varepsilon)\top} \tag{25}$$

Using the first-order expansions above,

$$\begin{aligned} \mathbb{E}\left[X_\varepsilon X_\varepsilon^\top\right] &= \mathbb{E}[XX^\top] + \varepsilon\mathbb{E}\left[X\dot{X}(V)^\top + \dot{X}(V)X^\top\right] + o(\varepsilon), \\ \mu_X^{(\varepsilon)}\mu_X^{(\varepsilon)\top} &= \mu_X\mu_X^\top + \varepsilon\left[\mu_X\dot{\mu}_X(V)^\top + \dot{\mu}_X(V)\mu_X^\top\right] + o(\varepsilon). \end{aligned} \tag{26}$$

The omitted terms are $o(\varepsilon)$ by the Cauchy–Schwarz inequality and because $O(\varepsilon^2) \subset o(\varepsilon)$. Thus,

$$\Sigma_X^{(\varepsilon)} = \Sigma_X + \varepsilon\dot{\Sigma}_X(V) + o(\varepsilon), \tag{27}$$

where

$$\begin{aligned} \dot{\Sigma}_X(V) &\coloneqq \left.\frac{d\Sigma_X^{(\varepsilon)}}{d\varepsilon}\right|_{\varepsilon=0} \\ &= \mathbb{E}\left[X\dot{X}(V)^\top + \dot{X}(V)X^\top\right] - \mu_X\dot{\mu}_X(V)^\top - \dot{\mu}_X(V)\mu_X^\top \\ &= \mathbb{E}\left[(X-\mu_X)\dot{X}(V)^\top + \dot{X}(V)(X-\mu_X)^\top\right]. \end{aligned} \tag{28}$$

Since $\dot{X}(V) = J_\phi(Z)V$ is linear in V, both $\dot{\mu}_X(V)$ and $\dot{\Sigma}_X(V)$ are also linear in V.

B.2 Local Bures-Wasserstein Expansion

Let $\mathrm{Sym}(D)$ denote the space of $D \times D$ symmetric matrices, and $\mathrm{Sym}^{++}(D)$ denote the set of $D \times D$ symmetric positive definite matrices. Following Malagò et al. (2018), for $\Sigma \in \mathrm{Sym}^{++}(D)$ and $H \in \mathrm{Sym}(D)$, we denote by $\mathcal{L}_\Sigma[H]$ the unique symmetric solution A to the Lyapunov equation $A\Sigma + \Sigma A = H$. The Bures-Wasserstein distance induces the Riemannian metric

$$W_\Sigma(H,K) = \mathrm{tr}\left(\mathcal{L}_\Sigma[H]\Sigma\mathcal{L}_\Sigma[K]\right) = \frac{1}{2}\mathrm{tr}\left(\mathcal{L}_\Sigma[H]K\right), \tag{29}$$

for $H, K \in \mathrm{Sym}(D)$. Then, for a covariance perturbation $\Sigma^{(\varepsilon)} = \Sigma + \varepsilon H + o(\varepsilon)$, the local second-order expansion of the squared Bures-Wasserstein distance is:

$$d_{\mathrm{BW}}^2\left(\Sigma, \Sigma^{(\varepsilon)}\right) = \varepsilon^2 W_\Sigma(H,H) + o(\varepsilon^2) = \frac{\varepsilon^2}{2}\mathrm{tr}\left(\mathcal{L}_\Sigma[H]H\right) + o(\varepsilon^2). \tag{30}$$

Applying this result to the feature covariance expansion derived in Sec. B.1, with $H = \dot{\Sigma}_X(V)$, gives:

$$d_{\mathrm{BW}}^2\left(\Sigma_X, \Sigma_X^{(\varepsilon)}\right) = \frac{\varepsilon^2}{2}\mathrm{tr}\left(\mathcal{L}_{\Sigma_X}\left[\dot{\Sigma}_X(V)\right]\dot{\Sigma}_X(V)\right) + o(\varepsilon^2). \tag{31}$$

B.3 Local FID Expansion

From the feature mean expansion derived in Sec. B.1, the mean term of FID can be expressed as

$$\left\|\mu_X^{(\varepsilon)} - \mu_X\right\|_2^2 = \varepsilon^2 \left\|\dot{\mu}_X(V)\right\|_2^2 + o(\varepsilon^2). \tag{32}$$

Combining this with the local Bures-Wasserstein expansion from Sec. B.2, we have:

$$\mathrm{FID}\left(P_X, P_X^{(\varepsilon)}\right) = \varepsilon^2 Q(V) + o(\varepsilon^2), \tag{33}$$

where

$$Q(V) = \left\|\dot{\mu}_X(V)\right\|_2^2 + \frac{1}{2}\,\mathrm{tr}\left[\mathcal{L}_{\Sigma_X}\left[\dot{\Sigma}_X(V)\right]\dot{\Sigma}_X(V)\right]. \tag{34}$$

Since $\dot{\mu}_X(V)$ and $\dot{\Sigma}_X(V)$ are linear in V and $\mathcal{L}_{\Sigma_X}$ is linear, Q is quadratic in V.

C Latent Perturbation Derivations

For each perturbation family considered below, we assume that the feature space regularity assumptions of the theorem hold.

C.1 Latent Mean Perturbations

Let u_i be a unit eigenvector of the latent covariance $\Sigma_Z = U\Lambda U^\top$, with eigenvalue λ_i. For $Z_\varepsilon = Z + \varepsilon u_i$, the family is mean square differentiable at $\varepsilon = 0$ with derivative $V_i^{\mathrm{mean}} = u_i$. The mean and covariance of Z_ε satisfy

$$\mu_Z^{(\varepsilon)} = \mu_Z + \varepsilon u_i, \qquad \Sigma_Z^{(\varepsilon)} = \Sigma_Z, \tag{35}$$

and therefore

$$\dot{\mu}_Z = u_i, \qquad \dot{\Sigma}_Z = 0. \tag{36}$$

The corresponding local FID sensitivity is

$$q_i^{\mathrm{mean}} = Q\left(V_i^{\mathrm{mean}}\right) = Q(u_i). \tag{37}$$

C.2 Latent Covariance Perturbations

Let $\Sigma_Z \succ 0$, and let u_i be a unit eigenvector of Σ_Z with eigenvalue λ_i. For sufficiently small ε, we consider a covariance perturbation that preserves the latent mean and changes the variance along u_i at unit rate:

$$\mu_Z^{(\varepsilon)} = \mu_Z, \qquad \Sigma_Z^{(\varepsilon)} = \Sigma_Z + \varepsilon u_i u_i^\top, \qquad \left.\frac{d\Sigma_Z^{(\varepsilon)}}{d\varepsilon}\right|_{\varepsilon=0} = u_i u_i^\top. \tag{38}$$

Specifying the mean and covariance does not uniquely determine the corresponding distribution. We therefore select, among all the distributions with these moments, the distribution that minimizes the 2-Wasserstein distance from the original latent distribution.

For any two distributions P and R with means μ_P, μ_R and covariances Σ_P, Σ_R, the Gelbrich bound (Gelbrich, 1990) gives:

$$W_2^2(P, R) \geq \|\mu_P - \mu_R\|_2^2 + d_{\mathrm{BW}}^2(\Sigma_P, \Sigma_R). \tag{39}$$

When $\Sigma_P \succ 0$, Nguyen et al. (2025) show that this bound is attained when R is a positive semidefinite affine pushforward of P:

$$X' = \mu_R + A(X - \mu_P), \qquad X \sim P, \tag{40}$$

where

$$A = \Sigma_P^{-1/2}\left(\Sigma_P^{1/2}\Sigma_R\Sigma_P^{1/2}\right)^{1/2}\Sigma_P^{-1/2}. \tag{41}$$

For the covariance perturbation above, we set:

$$\mu_P = \mu_R = \mu_Z, \qquad \Sigma_P = \Sigma_Z, \qquad \Sigma_R = \Sigma_Z + \varepsilon u_i u_i^\top. \tag{42}$$

Applying the eigendecomposition $\Sigma_Z = U\Lambda U^\top$, Σ_R can be written as

$$\Sigma_R = U(\Lambda + \varepsilon e_i e_i^\top)U^\top, \tag{43}$$

where e_i is the i-th standard basis vector. Therefore,

$$\begin{aligned}\Sigma_Z^{1/2}\Sigma_R\Sigma_Z^{1/2} &= U\Lambda^{1/2}(\Lambda + \varepsilon e_i e_i^\top)\Lambda^{1/2}U^\top \\ &= U\mathrm{diag}\left(\lambda_1^2, \ldots, \lambda_i(\lambda_i+\varepsilon), \ldots, \lambda_d^2\right)U^\top.\end{aligned} \tag{44}$$

Substituting this expression into the definition of A, we get

$$A_\varepsilon = U\mathrm{diag}\left(1, \ldots, \sqrt{1+\frac{\varepsilon}{\lambda_i}}, \ldots, 1\right)U^\top = I + \left(\sqrt{1+\frac{\varepsilon}{\lambda_i}} - 1\right)u_i u_i^\top. \tag{45}$$

The corresponding random variable is

$$Z_\varepsilon = \mu_Z + A_\varepsilon(Z - \mu_Z) \implies Z_\varepsilon - Z = \left(\sqrt{1+\frac{\varepsilon}{\lambda_i}} - 1\right)u_i u_i^\top (Z - \mu_Z). \tag{46}$$

Because

$$\frac{\sqrt{1+\varepsilon/\lambda_i}-1}{\varepsilon} = \frac{1/\lambda_i}{\sqrt{1+\varepsilon/\lambda_i}+1} \longrightarrow \frac{1}{2\lambda_i} \qquad \text{as } \varepsilon \to 0, \tag{47}$$

the family $(Z_\varepsilon)_\varepsilon$ is mean square differentiable at $\varepsilon = 0$, with derivative

$$V_i^{\mathrm{var}} = v_i^{\mathrm{var}}(Z) = \frac{1}{2\lambda_i}u_i u_i^\top (Z - \mu_Z). \tag{48}$$

Indeed, because u_i is a unit eigenvector of Σ_Z,

$$\left\|u_i u_i^\top (Z - \mu_Z)\right\|_{L^2}^2 = u_i^\top \Sigma u_i = \lambda_i, \tag{49}$$

and therefore,

$$\left\|\frac{Z_\varepsilon - Z}{\varepsilon} - V_i^{\mathrm{var}}\right\|_{L^2} = \left|\frac{\sqrt{1+\varepsilon/\lambda_i}-1}{\varepsilon} - \frac{1}{2\lambda_i}\right|\sqrt{\lambda_i} \to 0. \tag{50}$$

The corresponding mean and covariance derivatives are $\dot{\mu}_Z = 0$ and $\dot{\Sigma}_Z = u_i u_i^\top$. Defining the standardized latent coordinate

$$\eta_i \coloneqq \frac{u_i^\top (Z - \mu_Z)}{\sqrt{\lambda_i}}, \tag{51}$$

we can write V_i^{var} as

$$V_i^{\mathrm{var}} = \frac{1}{2\sqrt{\lambda_i}}\eta_i u_i. \tag{52}$$

Since Q is quadratic in V, the corresponding local FID sensitivity is:

$$q_i^{\mathrm{var}} = Q\left(V_i^{\mathrm{var}}\right) = \frac{1}{4\lambda_i}Q(\eta_i u_i). \tag{53}$$

C.3 Perturbations over Eigenvector Groups

We further consider perturbations over groups of consecutive covariance eigenvectors, ordered by their eigenvalues, to examine how FID sensitivity varies across different regions of the spectrum. For each group G containing m eigenvectors, we consider mean or variance changes of equal magnitude along all eigenvector directions.

Mean perturbations. For mean perturbations, there is no unique direction that induces a mean shift with unit norm within the m-dimensional subspace. Therefore, we use normalized Rademacher directions:

$$V_G^{\mathrm{mean}}(s) = \frac{1}{\sqrt{m}}\sum_{i\in G} s_i u_i, \qquad s_i \in \{-1, +1\}. \tag{54}$$

We draw independent Rademacher signs s_i, with $\mathbb{P}(s_i = 1) = \mathbb{P}(s_i = -1) = 1/2$. For each fixed sign vector s, we consider the translation $Z_\varepsilon^{(s)} = Z + \varepsilon V_G^{\text{mean}}(s)$. Because the eigenvectors are orthonormal, $\|V_G^{\text{mean}}\|_2 = 1$ and the corresponding FID sensitivity is:

$$q_G^{\text{mean}}(s) = Q(V_G^{\text{mean}}(s)) = \frac{1}{m} \sum_{i,j \in G} s_i s_j \left(\dot{\mu}_X(u_i)^\top \dot{\mu}_X(u_j) + \frac{1}{2} \text{tr} \left(\mathcal{L}_{\Sigma_X}[\dot{\Sigma}_X(u_i)] \dot{\Sigma}_X(u_j) \right) \right), \tag{55}$$

where the last equation holds from the linearity of the feature moment velocities and of $\mathcal{L}_{\Sigma_X}$.

For independent Rademacher signs, $\mathbb{E}[s_i s_j] = 0$ for $i \neq j$ and $\mathbb{E}[s_i^2] = 1$, so averaging over Rademacher directions gives:

$$\mathbb{E}_s \left[q_G^{\text{mean}}(s) \right] = \mathbb{E}_s \left[Q(V_G^{\text{mean}}(s)) \right] = \frac{1}{m} \sum_{i \in G} q_i^{\text{mean}}. \tag{56}$$

Covariance perturbations. We define

$$H_G = \frac{1}{\sqrt{m}} \sum_{i \in G} u_i u_i^\top = \frac{1}{\sqrt{m}} P_G, \qquad \|H_G\|_F = 1, \tag{57}$$

where P_G is the orthogonal projection matrix onto the subspace spanned by the eigenvectors in G. The corresponding covariance perturbation is

$$\Sigma_Z^{(\varepsilon)} = \Sigma_Z + \varepsilon H_G. \tag{58}$$

Following Appendix C.2, the corresponding random variable is

$$Z_\varepsilon = \mu_Z + A_\varepsilon^{(G)}(Z - \mu_Z), \qquad A_\varepsilon^{(G)} = I + \sum_{i \in G} \left(\sqrt{1 + \frac{\varepsilon}{\sqrt{m}\lambda_i}} - 1 \right) u_i u_i^\top, \tag{59}$$

with mean square derivative at $\varepsilon = 0$

$$V_G^{\text{var}} = v_G^{\text{var}}(Z) = \frac{1}{\sqrt{m}} \sum_{i \in G} V_i^{\text{var}}. \tag{60}$$

The corresponding local FID sensitivity is

$$q_G^{\text{var}} = Q\left(V_G^{\text{var}}\right). \tag{61}$$

D Additional Experimental Details and Results

D.1 Implementation Details

Model Architectures. We provide additional information about the model and latent space configurations used in the continual learning settings outlined in Sec. 5.1. For both CIFAR-10 settings, we use a class-conditional DDPM++ U-Net (Song et al., 2021), with a Variational Autoencoder (VAE) (Kingma & Welling, 2014) of latent shape $4 \times 16 \times 16$ trained on all CIFAR-10 classes. For the LDM experiments on the LSUN Churches $\rightarrow$ AFHQ setting, we use the publicly released U-Net pretrained on LSUN Churches, along with its associated VAE (Rombach et al., 2022) with latent shape $4 \times 32 \times 32$. For the LFM experiments on CelebA-HQ $\rightarrow$ Flowers102 and LSUN Churches $\rightarrow$ AFHQ, we use the publicly released DiT-L/2 models (Dao et al., 2023) pretrained on their corresponding ID distributions, together with the Stable Diffusion VAE (Rombach et al., 2022) with latent shape $4 \times 32 \times 32$. The latent scaling factors are 0.6564 for the CIFAR-10 settings, 0.2458 for the LSUN Churches $\rightarrow$ AFHQ setting, and 0.1822 for both LFM settings. Details about the model architectures are provided in Table 6.

Dataset Statistics. Table 7 summarizes the dataset sizes used in the ID/OOD experiments.

Table 6: Architectures of the generative models used in our experiments.

Setting	Backbone	Architecture	Prediction	Parameters
CIFAR-10	U-Net	4 input/output channels, base width 128, multipliers $(1, 2, 2)$, 2 residual blocks, attention at 8×8 and 4×4	Noise	29.1M
LDM LSUN Churches $\rightarrow$ AFHQ	U-Net	4 input/output channels, base width 192, multipliers $(1, 2, 2, 4, 4)$, 2 residual blocks, attention at 32×32, 16×16, 8×8, and 4×4	Noise	295.0M
LFM settings	DiT-L/2	4 input/output channels, width 1024, 24 blocks, 16 heads, patch size 2, MLP ratio 4	Velocity	457.1M

Table 7: Dataset statistics for the ID/OOD experiments.

Setting	ID Dataset	ID Images	OOD Dataset	OOD Train
LDM CIFAR-10 8-to-2	CIFAR-10 0–7	40,000	CIFAR-10 8–9	10,000
LDM LSUN Churches $\rightarrow$ AFHQ	LSUN Churches	119,915	AFHQ	16,130
LFM CelebA-HQ $\rightarrow$ Flowers102	CelebA-HQ	30,000	Flowers102	6,149
LFM LSUN Churches $\rightarrow$ AFHQ	LSUN Churches	119,915	AFHQ	16,130

Training Hyperparameters. We next provide an overview of hyperparameters used in our experiments. Table 8 summarizes the training hyperparameters shared across methods within each setting. Table 9 reports method-specific loss weights and the hyperparameters of LaBW for each setting. The EWC weight applies only to the Rank-1 Fisher EWC baseline. For the one-step endpoint estimates used by LaBW, we sample t from a logit-normal distribution (Aitchison & Shen, 1980; Esser et al., 2024) with underlying Gaussian parameters $\mu = 0$ and $\sigma^2 = 1$. For the CIFAR-10 experiments, we additionally restrict the diffusion timesteps we sample from to $t \in \{1, \dots, 700\}$ for the one-step endpoint predictions to avoid instabilities at later timesteps. This restriction is not applied to the LSUN Churches $\rightarrow$ AFHQ setting with an LDM, which uses the beta schedule of the pretrained model. Furthermore, for the CIFAR-10 ID/OOD setting, we initialize the LaBW weight at 8 and linearly decay it to 4 over the first 50 epochs. This schedule was motivated by the quadratic nature of the Bures-Wasserstein objective. For the class-incremental CIFAR-10 setting, following Wang et al. (2026), the replay buffer contains $10,000 \times k$ samples before training on task $k + 1$, and is regenerated after each task.

Table 8: Training hyperparameters used across all methods for each setting.

Setting	Optimizer	Learning Rate	Batch Size	Replay Buffer	Training Epochs
CIFAR-10 (ID/OOD)	AdamW	2×10^{-4}	128	2,500 per class	800
CIFAR-10 (class-incremental)	AdamW	2×10^{-4}	128	$10,000 \times k$	1,000 per task
LSUN Churches $\rightarrow$ AFHQ (LDM)	AdamW	5×10^{-5}	64	12,623	225
CelebA-HQ $\rightarrow$ Flowers102	AdamW	5×10^{-5}	64	6,149	200
LSUN Churches $\rightarrow$ AFHQ (LFM)	AdamW	5×10^{-5}	64	12,623	400

FID Evaluation. To account for variability in FID across training and sampling seeds (Dufour et al., 2026), all reported FID and forgetting results in the main experiments are computed across three training seeds, with three evaluation seeds per training seed. For each training seed, we average the results across the three evaluation seeds, and then report the mean and standard deviation of these averages across the three training seeds. For the CIFAR-10 runs, we generate the same number of samples per class as in the corresponding reference distribution. For the other ID/OOD settings, we generate 30,000 samples for evaluating the ID distributions, and the same number of OOD samples as in the corresponding OOD dataset. For each training run, we train for the number of epochs reported in Table 8 and, among the saved checkpoints, select the one with the lowest OOD FID.

Computational Efficiency. We additionally report in Table 10 the time per epoch on an A100 GPU for each method, across all settings.

Table 9: Method-specific loss weights and LaBW hyperparameters for each setting.

	Weights				
Setting	GD	Rank-1 Fisher EWC	LaBW	Subspace Rank	EMA Decay Rate
CIFAR-10 (ID/OOD)	0.5	1,000	4	512	0.98
CIFAR-10 (class-incremental)	1	1,000	0.5	512	0.98
LSUN Churches → AFHQ (LDM)	1	15,000	0.15	2,048	0.98
CelebA-HQ → Flowers102	1	650	0.5	2,048	0.98
LSUN Churches → AFHQ (LFM)	1	5,000	2.5	2,048	0.98

Table 10: Average time per epoch in seconds on an A100 GPU for each method across the ID/OOD settings.

Setting	GD	Rank-1 Fisher EWC	LaBW
CIFAR-10	13.13	19.66	22.64
LSUN Churches → AFHQ (LDM)	212.34	224.53	497.82
CelebA-HQ → Flowers102	136.43	140.11	271.76
LSUN Churches → AFHQ (LFM)	324.45	325.38	606.81

D.2 Additional Ablation Studies

We provide further analysis of the covariance subspace ablation presented in Sec. 5.2 and perform an additional ablation on the added mean term of the Fréchet distance in LaBW.

Covariance Subspace. We provide additional analysis of the covariance subspace ablation from Sec. 5.2. In addition to the CIFAR-10 and CelebA-HQ → Flowers102 results reported in the main paper, we evaluate the same low-variance, high-variance and full space variants on the LSUN Churches → AFHQ setting with an LDM. As shown in Table 11, the three variants perform similarly, with only small differences in both ID and OOD FID.

Table 11: Additional results for the latent covariance subspace ablation on LSUN Churches → AFHQ with an LDM. The mean and standard deviation are computed over three evaluation seeds for one training seed.

Subspace	ID FID ↓	OOD FID ↓	$\mathcal{F}$ ↓
Low-variance	$\mathbf{5.68} \pm 0.01$	16.44 ± 0.17	$\mathbf{1.43} \pm 0.03$
High-variance	5.72 ± 0.05	16.40 ± 0.06	1.48 ± 0.03
Full latent space	5.74 ± 0.04	$\mathbf{16.27} \pm 0.24$	1.49 ± 0.02

To further analyze these results, we measure how well each variant preserves the teacher covariance across different parts of the latent spectrum. For a teacher covariance eigenspace U_B, where B denotes the low-variance subspace, high-variance subspace, or full latent space, we compute the Bures-Wasserstein distance

$$D_B = d^2_{\mathrm{BW}}\big(U_B^\top \Sigma_S U_B,\, U_B^\top \Sigma_T U_B\big)\,. \tag{62}$$

All checkpoints within a setting are evaluated using the same replay samples, timesteps, and noise used to construct the one-step estimates. The results are reported in Table 12. On CIFAR-10, the choice of the training subspace has a clear effect on covariance preservation. Matching the covariances in the low-variance subspace also reduces the BW distance in the high-variance subspace and, therefore, across the full latent space, whereas matching in the high-variance subspace does not reduce the distance in the low-variance subspace. This is consistent with the substantially lower forgetting observed for the low-variance variant. In contrast, for CelebA-HQ → Flowers102 and LSUN Churches → AFHQ, the three LaBW variants lead to very similar covariance discrepancies across the latent spectrum. This helps explain why the choice of subspace has little effect on FID in these settings, as the resulting covariance preservation extends beyond the subspace on which the loss is applied. Lastly, the full space variant is substantially slower than the low- and high-variance ones, as expected due to the larger covariance matrices required.

Table 12: Bures–Wasserstein covariance distances between the teacher and student endpoint distributions, evaluated in different latent covariance subspaces.

Setting	Variant	Low-variance	High-variance	Full Latent Space
CIFAR-10	GD	0.0929	0.2546	0.5502
	Low-variance LaBW	**0.0520**	**0.1510**	**0.3311**
	High-variance LaBW	0.0949	0.3279	0.6328
	Full-space LaBW	0.0666	0.2171	0.4348
CelebA-HQ → Flowers102	GD	3.6695	12.8683	26.1095
	Low-variance LaBW	3.6309	12.8170	25.9636
	High-variance LaBW	3.6483	12.8156	26.0037
	Full-space LaBW	**3.5998**	**12.7377**	**25.7872**
LSUN Churches → AFHQ (LDM)	GD	0.5488	1.3672	3.0697
	Low-variance LaBW	**0.3831**	0.9852	**2.1901**
	High-variance LaBW	0.4155	0.9733	2.2541
	Full-space LaBW	0.3996	**0.9707**	2.2083

Mean Term. Lastly, our final objective is the Bures-Wasserstein distance between the projected teacher and student covariances, which coincides with the covariance term of the Fréchet distance, and does not include the corresponding mean term. To explain this decision, we conduct ablations with the added Euclidean mean loss, which we apply in the full latent space, while keeping the remaining training configuration fixed. Because the mean loss operates in the full latent space, its weight is tuned separately to ensure that the two terms have comparable gradient norm magnitudes. We present the results in Table 13, where the added mean term provides no consistent benefit. Across all settings, the differences in ID and OOD FID are small relative to the reported variability.

Table 13: Ablation of the latent-mean term. CIFAR-10 results are the mean and standard deviation across three training seeds. CelebA-HQ → Flowers102 and LDM LSUN-Churches → AFHQ use one training seed and report variation across three evaluation seeds.

Setting	Objective	ID FID ↓	OOD FID ↓	Forgetting ↓
CIFAR-10	Covariance only	**13.60** $\pm$ 0.04	**10.69** $\pm$ 0.04	**2.17** $\pm$ 0.04
	Covariance + mean	13.62 $\pm$ 0.08	10.79 $\pm$ 0.10	2.19 $\pm$ 0.08
CelebA-HQ → Flowers102	Covariance only	**7.90** $\pm$ 0.09	23.40 $\pm$ 0.22	**0.58** $\pm$ 0.01
	Covariance + mean	8.02 $\pm$ 0.08	**23.36** $\pm$ 0.25	0.70 $\pm$ 0.02
LDM LSUN-Churches → AFHQ	Covariance only	**5.68** $\pm$ 0.01	16.44 $\pm$ 0.17	**1.43** $\pm$ 0.03
	Covariance + mean	5.72 $\pm$ 0.02	**16.35** $\pm$ 0.11	1.48 $\pm$ 0.02

We additionally isolate the contribution of the latent mean shift between teacher and student at the selected checkpoints of the main experiments. Given endpoint distributions generated by a teacher-student pair, we define $\delta_\mu = \mu_S - \mu_T$ and translate every student sample according to $z'^{(i)}_S = z^{(i)}_S - \delta_\mu$. This translation matches the teacher and student latent means while leaving the student latent covariance unchanged. We then decode the translated endpoints and recompute ID FID. The results are reported in Table 14. Correcting the latent mean shift leaves FID effectively unchanged, suggesting that the latent mean difference between teacher and student is not a meaningful source of forgetting in these settings.

Table 14: Effect of removing the teacher-relative latent mean shift. Negative ΔFID values indicate an improvement. FID is computed using 10,000 generated samples.

Setting	Checkpoint	$\|\mu_S - \mu_T\|_2$	Original FID	Corrected FID	ΔFID
CIFAR-10	GD	0.512	14.970	14.931	−0.039
	EWC	0.689	14.745	14.717	−0.028
	LaBW	0.766	13.629	13.649	+0.020
CelebA-HQ → Flowers102	GD	0.115	8.951	8.943	−0.008
	EWC	0.110	8.853	8.854	+0.001
	LaBW	0.108	8.672	8.672	+0.001
LDM LSUN-Churches → AFHQ	GD	0.242	6.382	6.394	+0.012
	EWC	0.212	6.275	6.272	−0.003
	LaBW	0.204	5.695	5.691	−0.005