SEQLORA: BILEVEL ORTHOGONAL ADAPTATION FOR CONTINUAL MULTI-CONCEPT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient fine-tuning enables fast personalization of text-to-image diffusion models to user-provided concepts (objects, people, or styles), but composing multiple such concepts remains challenging due to representation interference. Existing modular methods, usually built on low-rank adaptation (LoRA), either rely on expensive post-hoc fusion or freeze the LoRA adaptation subspaces, which limit expressiveness and concept fidelity. To address this trade-off, we propose Sequential regularized LoRA (SeqLoRA), a constrained continual learning framework that jointly optimizes both LoRA factors via bilevel optimization while keeping each new basis orthogonal to all previously learned ones. Theoretically, we establish monotone descent and convergence to a critical point of the constrained problem, and model the residual layer activations as a matrix sub-Gaussian process to derive a high-probability bound on catastrophic forgetting in multi-layer nonlinear networks. This bound depends on the basis only through a residual interference energy, and within the feasible subspace we prove that a data-adapted basis minimizes it, whereas a random frozen basis is suboptimal in expectation. Experiments on Stable Diffusion with up to 101 concepts show that SeqLoRA improves identity preservation over fusion-based methods, attains the lowest cross-concept leakage in multi-concept compositions, requires no fusion step, and scales to concept counts at which fusion runs out of memory.

1 INTRODUCTION

Text-to-image diffusion models generate high-quality, diverse images from natural-language prompts (Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022) and now support a wide range of creative and scientific applications (Po et al., 2024b; Luo et al., 2025). Parameter-efficient fine-tuning, especially Low-Rank Adaptation (LoRA) (Hu et al., 2022), has made it feasible for individual users to adapt these foundation models to new concepts, such as a specific object, person, or style, from a small set of reference images (Ruiz et al., 2023; Gal et al., 2023).

While single-concept customization has become highly mature (Ruiz et al., 2023; Gal et al., 2023; Hu et al., 2022), a key challenge arises when users want to generate images that include multiple personalized concepts at once, for instance, two specific pets interacting in the same scene, or a particular artistic style combined with a custom object (Kumari et al., 2023; Gu et al., 2023). The central challenge is *representation interference*: because concepts are usually learned separately, their corresponding adaptations may occupy overlapping regions in the model's parameter space (Po et al., 2024a; Liang and Li, 2024). When these adaptations are combined at inference time, the shared directions can lead to *attribute entanglement*, where identity-related features from one concept bleed into another, resulting in mixed attributes, weakened identities, or even incoherent generations (Po et al., 2024a; Gu et al., 2023). This problem is especially pronounced for semantically similar concepts, such as different human faces, whose learned weight residuals tend to be highly correlated and lie in nearly the same subspaces (Po et al., 2024a). As the number of concepts increases, the combinatorial growth in possible concept subsets makes it infeasible to train a separate model for every combination, motivating modular methods in which independently fine-tuned components can be composed at test time without retraining (Gu et al., 2023; Po et al., 2024a).

Existing remedies for concept interference come with distinct trade-offs. *Joint training* methods such as Custom Diffusion (Kumari et al., 2023) and Break-A-Scene (Avrahami et al., 2023) optimize all concepts simultaneously within a shared model, which can reduce interference but requires concurrent access to all concept data, a requirement that is unscalable to large concept libraries and raises data privacy concerns in multi-user settings. Gradient fusion approaches, exemplified by Mix-of-Show (Gu et al., 2023), allow independent fine-tuning of each concept and then merge the resulting

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.

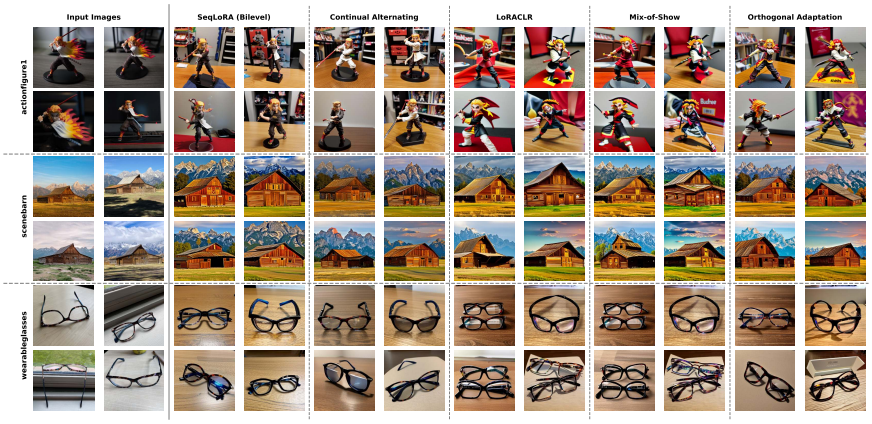


Figure 1: Qualitative comparison after learning 32 concepts into one model. We show the original concept images (Input) and 4 generated samples (seeds) for all methods.

models through an optimization-based fusion step to preserve identity; however, the fusion must be re-executed for every new combination of concepts, preventing truly modular composition where individual adapters can be added or removed on the fly. Orthogonal Adaptation (Po et al., 2024a) takes a different approach by sampling each concept's LoRA basis $\mathbf{B}_i$ from a shared orthogonal matrix and freezing it during training, optimizing only the coefficient matrix $\mathbf{A}_i$. This guarantees approximate orthogonality between concepts and enables instant merging via simple summation, but freezing $\mathbf{B}_i$ restricts the feasible space of the adaptation and limits per-concept fidelity. In the continual learning literature, InfLoRA (Liang and Li, 2024) designs task-specific LoRA subspaces for image classification that are orthogonal to the gradient subspaces of previously learned tasks, achieving strong stability–plasticity trade-offs for sequential classification; however, it similarly freezes the basis matrix for each task, inheriting the same expressiveness limitation.

These approaches share a tension between orthogonality, needed to reduce interference, and expressiveness, needed for high-fidelity adaptation: no existing method optimizes both LoRA factors freely while enforcing the subspace orthogonality required for low-interference composition. We address this tension by formulating multi-concept generation as a constrained continual learning problem. Figure 1 previews the effect after 32 concepts: on challenging concepts such as `wearableglasses` the baselines entangle attributes, whereas SeqLoRA preserves the identity.

Contributions.

- **Method.** SeqLoRA jointly optimizes both LoRA factors under exact subspace orthogonality via a bilevel update with a cross-Hessian correction and a closed-form projection (§3), yielding a single model that composes any subset of concepts by summation, with no fusion step and no stored data of past concepts.
- **Theory.** Monotone descent and convergence to a critical point of the constrained problem (Theorem 1); a high-probability forgetting bound for the *nonlinear* network (Theorem 2) identifying residual activation energy as the driver of interference, with an interpretable rate and a dominance result, within the feasible subspace, for covariance-adapted over random frozen bases.
- **Experiments.** On Stable Diffusion v1.5 with 8 to 101 concepts, SeqLoRA improves identity preservation over the baselines and scales past the point at which they run out of memory. In multi-concept compositions it attains the lowest cross-concept leakage of all methods; we also quantify robustness to concept order, rank and concept similarity, compute cost, and report a preliminary transfer to SDXL (Podell et al., 2024).

2 Problem Formulation

2.1 Continual Learning over Concepts

We consider a sequential concept stream $\tau_1, \ldots, \tau_T$, where each task τ_i provides a small reference dataset $\mathcal{D}_i = \{(\mathbf{x}_{i,m}, c_i)\}_{m=1}^{M_i}$ pairing M_i reference images with a concept-specific text token c_i. Only $\mathcal{D}_i$ is accessible when task τ_i arrives; prior datasets are discarded. The goal is a compact, composable representation per concept such that any subset of the T concepts can be jointly rendered

at inference time without retraining, balancing *plasticity* (adapting to new concepts) and *stability* (preserving previous ones).

2.2 Single-Concept Fine-Tuning with LoRA

Rather than updating the full weight matrix $\mathbf{W}_i \in \mathbb{R}^{n\times m}$ of an attention layer, LoRA (Hu et al., 2022) learns a low-rank residual for each concept i:

$$\mathbf{W}_i = \mathbf{W}_0 + \mathbf{A}_i\mathbf{B}_i^\top, \tag{1}$$

where $\mathbf{W}_0$ contains the pretrained parameters for that layer, $\mathbf{A}_i \in \mathbb{R}^{n\times r}$ and $\mathbf{B}_i \in \mathbb{R}^{m\times r}$ are the trainable low-rank factors with rank $r \ll \min\{m, n\}$. For a single concept i, the standard training objective minimizes the denoising loss

$$\mathcal{L}_i(\mathbf{A}_i, \mathbf{B}_i) = \mathbb{E}_{\mathbf{x}_i,\boldsymbol{\epsilon},t}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}(\mathbf{W}_i)}(\mathbf{x}_i^{(t)}, t, c_i)\right\|_2^2\right], \tag{2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the noise target, t is a uniformly sampled diffusion timestep, $\mathbf{x}_i^{(t)}$ is the noised version of the reference image, $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is the noise-prediction network, and $\boldsymbol{\theta}(\mathbf{W}_i)$ denotes the full network parameters with the adapted layer weight set to $\mathbf{W}_i$. Adaptation is applied to L layers; we omit the layer index whenever the argument is the same for every layer.

2.3 Multi-Concept Aggregation and the Interference Problem

In a practical deployment scenario, typically each concept is fine-tuned independently, potentially by different users in parallel, producing a separate LoRA module $(\mathbf{A}_i, \mathbf{B}_i)$ per concept (Gu et al., 2023; Po et al., 2024a). At inference time, the goal is to compose an arbitrary subset of these modules into a single model for multi-concept generation. The simplest aggregation strategy is to sum the individual LoRA residuals per layer:

$$\mathbf{W}_T = \mathbf{W}_0 + \sum_{i=1}^{T} \mathbf{A}_i\mathbf{B}_i^\top. \tag{3}$$

To explain why naive aggregation fails, consider a linear layer adapted for concept i with input $\mathbf{X}_i$. The single-concept output is $\mathbf{O}_i(\mathbf{X}_i) = (\mathbf{W}_0 + \mathbf{A}_i\mathbf{B}_i^\top)\mathbf{X}_i$, while the merged output becomes $\hat{\mathbf{O}}_i(\mathbf{X}_i) = (\mathbf{W}_0 + \mathbf{A}_i\mathbf{B}_i^\top + \mathbf{A}_j\mathbf{B}_j^\top)\mathbf{X}_i$. The discrepancy $\hat{\mathbf{O}}_i - \mathbf{O}_i = \mathbf{A}_j\mathbf{B}_j^\top\mathbf{X}_i$ is the crosstalk term (Po et al., 2024a). Concept identity is preserved when this term vanishes, but since $\mathbf{X}_i$ typically has full column rank, exact cancellation is impossible. A natural relaxation (Po et al., 2024a) projects $\mathbf{X}_i$ onto the orthogonal complement of $\mathrm{col}(\mathbf{B}_j)$. Let $\bar{\mathbf{B}}_j \in \mathbb{R}^{m\times(m-r)}$ span this complement; then $\mathbf{A}_j\mathbf{B}_j^\top\bar{\mathbf{B}}_j\bar{\mathbf{B}}_j^\top\mathbf{X}_i = \mathbf{0}$ since $\mathbf{B}_j^\top\bar{\mathbf{B}}_j = \mathbf{0}$. Because $r \ll \min\{m, n\}$, the complement covers most of $\mathbb{R}^m$, so $\bar{\mathbf{B}}_j\bar{\mathbf{B}}_j^\top\mathbf{X}_i \approx \mathbf{X}_i$. At the same time, for concept i's residual to remain effective, the columns of $\mathbf{B}_i$ must not lie in $\mathrm{col}(\mathbf{B}_j)$, which requires $\mathbf{B}_i^\top\mathbf{B}_j = \mathbf{0}$. This establishes pairwise orthogonality of the LoRA bases as the key condition for low interference aggregation.

2.4 Desiderata

The preceding analysis motivates three requirements: (i) pairwise orthogonality $\mathbf{B}_i^\top\mathbf{B}_j = \mathbf{0}$ for $i \neq j$ to reduce inference-time interference; (ii) joint optimization of $\mathbf{A}_i$ and $\mathbf{B}_i$ to preserve expressiveness; and (iii) continual learning, accommodating new concepts sequentially without access to past data or joint retraining. Prior works (Po et al., 2024a; Liang and Li, 2024) sacrifice (ii) by freezing $\mathbf{B}_i$. The following warm-up motivates why learning the basis is preferable to freezing it.

Example 1 (Warm-Up: why the basis matters)**.** *Consider a single-layer linear map with input $\mathbf{x} \in \mathbb{R}^m$ and output $\mathbf{y} \in \mathbb{R}^n$. For concept j with distribution $\mathcal{D}_j$, define the population risk $\mathcal{L}_j(\mathbf{W}) \triangleq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_j}\{\|\mathbf{y} - \mathbf{W}\mathbf{x}\|_2^2\}$ and the input covariance $\boldsymbol{\Sigma}_j \triangleq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_j}\{\mathbf{x}\mathbf{x}^\top\}$. Under orthogonality constraints, the interference for concept j after learning T concepts is proportional to its uncaptured residual energy: $Tr(\mathcal{P}_{B_j}^{\perp}\boldsymbol{\Sigma}_j)$, where $\mathcal{P}_{B_j}^{\perp}$ is the orthogonal projector onto the complement of $\mathrm{col}(\mathbf{B}_j)$ (Theorem 8 in Appendix E). To minimize this interference, the rank-r basis $\mathbf{B}_j$ must align with the top-r principal components of $\boldsymbol{\Sigma}_j$. Consequently, randomly frozen bases (e.g., Orthogonal Adaptation (Po et al., 2024a)) cannot exploit this spectral geometry and leave more residual interference in expectation (Proposition 2).*

The example does not model the nonlinear denoising network, and the covariance-optimal basis *is not necessarily* optimal for the end-to-end loss in (2). Rather, its role is to isolate, within a stylized linear model, the quantity that a learned basis can reduce but a frozen basis cannot.

3 PROPOSED METHOD

SeqLoRA treats the previously learned bases $\{\mathbf{B}_j\}_{j<i}$ as a compact memory of past concepts and constrains each new basis $\mathbf{B}_i$ to the orthogonal complement of their column span; the full procedure is summarized in Algorithm 1.

3.1 CONSTRAINED OPTIMIZATION PROBLEM

When concept i arrives, we seek LoRA factors $(\mathbf{A}_i, \mathbf{B}_i)$ that minimize the denoising loss while keeping $\mathbf{B}_i$ orthogonal to all previously learned bases:

$$\min_{\mathbf{A}_i,\, \mathbf{B}_i \in \mathcal{B}_i} \; \mathcal{L}_i(\mathbf{A}_i, \mathbf{B}_i), \tag{4}$$

where

$$\mathcal{B}_i = \{\mathbf{B}_i \in \mathbb{R}^{m\times r} \mid \mathbf{B}_j^\top \mathbf{B}_i = \mathbf{0}, \quad \forall\, j < i, \mathbf{0} \in \mathbb{R}^{r\times r}\}. \tag{5}$$

Problem (4) is jointly non-convex in $(\mathbf{A}_i, \mathbf{B}_i)$, but it has favorable structure: $\mathcal{L}_i$ is differentiable with respect to each factor individually, and the constraints are linear in $\mathbf{B}_i$.

3.2 BILEVEL OPTIMIZATION ALGORITHM

Block-coordinate schemes such as BSUM (Razaviyayn et al., 2013) or prox-linear methods (Drusvyatskiy et al., 2019) solve (4) by alternating between $\mathbf{A}_i$ for fixed $\mathbf{B}_i$ and $\mathbf{B}_i$ for fixed $\mathbf{A}_i$, ignoring the coupling between the two blocks. We instead rewrite (4) as a bilevel problem (Dempe, 2002):

$$\begin{aligned} &\min_{\mathbf{B}_i \in \mathcal{B}_i} \; \mathcal{L}_i(\mathbf{A}_i^*(\mathbf{B}_i), \mathbf{B}_i) \\ &\text{subject to} \quad \mathbf{A}_i^*(\mathbf{B}_i) = \operatorname*{argmin}_{\mathbf{A}_i} \mathcal{L}_i(\mathbf{A}_i, \mathbf{B}_i). \end{aligned} \tag{6}$$

To solve the lower-level optimization problem, we use a quadratic approximation $\hat{\mathcal{L}}_i(\mathbf{A}_i, \mathbf{B}_i)$ of $\mathcal{L}_i(\mathbf{A}_i, \mathbf{B}_i)$

$$\hat{\mathcal{L}}_i(\mathbf{A}_i, \mathbf{B}_i; \mathbf{A}_i^{(k)}) = \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i) + \left\langle \nabla_{\mathbf{A}_i} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i),\; \mathbf{A}_i - \mathbf{A}_i^{(k)} \right\rangle + \frac{1}{2\alpha} \left\| \mathbf{A}_i - \mathbf{A}_i^{(k)} \right\|_F^2. \tag{7}$$

Substituting this quadratic approximation into the lower-level optimization problem in (6) leads to:

$$\mathbf{A}_i^{(k+1)}(\mathbf{B}_i) = \arg\min_{\mathbf{A}_i} \; \hat{\mathcal{L}}_i(\mathbf{A}_i, \mathbf{B}_i; \mathbf{A}_i^{(k)}) = \mathbf{A}_i^{(k)} - \alpha\, \nabla_{\mathbf{A}_i} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i), \tag{8}$$

where α is a step size. Note that the model update depends on the $\mathbf{B}_i$. This dependence is accounted for in the upper-level optimization in (6), whose cost function becomes

$$\Phi_i(\mathbf{B}_i) \triangleq \mathcal{L}_i\big(\mathbf{A}_i^{(k+1)}(\mathbf{B}_i),\, \mathbf{B}_i\big) = \mathcal{L}_i\Big(\mathbf{A}_i^{(k)} - \alpha\, \nabla_{\mathbf{A}_i} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i),\; \mathbf{B}_i\Big), \tag{9}$$

which is now a function of $\mathbf{B}_i$ alone. The constrained upper-level problem is

$$\min_{\mathbf{B}_i \in \mathcal{B}_i} \; \Phi_i(\mathbf{B}_i). \tag{10}$$

By (9), $\mathbf{B}_i$ enters Φ_i twice: directly through $\nabla_{\mathbf{B}_i} \mathcal{L}_i$, and indirectly through the dependence of the updated factor $\mathbf{A}_i^{(k+1)}(\mathbf{B}_i)$ on $\mathbf{B}_i$, which is captured by the mixed derivative $\nabla^2_{\mathbf{B}_i \mathbf{A}_i} \mathcal{L}_i$. Solving (10) therefore seeks a basis that both lowers the loss at the current iterate and steers the lower-level update of $\mathbf{A}_i$ toward directions that lower it further, while the projection onto $\mathcal{B}_i$ maintains orthogonality across concepts.

Since Φ_i is non-convex and $\mathcal{B}_i$ is a constraint set, we again use a quadratic model $\hat{\Phi}_i(\mathbf{B}_i)$ around $\mathbf{B}_i^{(k)}$:

$$\hat{\Phi}_i(\mathbf{B}_i; \mathbf{B}_i^{(k)}) \;=\; \Phi_i(\mathbf{B}_i^{(k)}) + \left\langle \nabla_{\mathbf{B}_i} \Phi_i(\mathbf{B}_i^{(k)}),\; \mathbf{B}_i - \mathbf{B}_i^{(k)} \right\rangle + \frac{1}{2\beta} \left\| \mathbf{B}_i - \mathbf{B}_i^{(k)} \right\|_F^2, \tag{11}$$

where $\beta > 0$ is the upper-level step-size. Note that evaluating (8) at $\mathbf{B}_i = \mathbf{B}_i^{(k)}$ yields $\mathcal{L}_i(\mathbf{A}_i^{(k+1)}(\mathbf{B}_i^{(k)}), \mathbf{B}_i)$, which recovers the standard alternating-minimization pattern: (i) fix $\mathbf{B}_i = \mathbf{B}_i^{(k)}$ and update $\mathbf{A}_i$, then (ii) fix $\mathbf{A}_i = \mathbf{A}_i^{(k+1)}$ and update $\mathbf{B}_i$. However, unlike standard alternating minimization, which descends only the partial objective $\mathcal{L}_i(\mathbf{A}_i^{(k+1)}, \cdot)$ with frozen $\mathbf{A}_i^{(k+1)}$, the bilevel gradient $\nabla_{\mathbf{B}_i} \Phi_i$ additionally captures the coupling between factors through a cross-Hessian correction $\nabla^2_{\mathbf{B}_i \mathbf{A}_i} \mathcal{L}_i$, yielding a richer descent direction on the true reduced objective. We include

this standard alternating minimization variant (without the cross-Hessian correction) as a baseline in our experiments, denoted `Continual Alternating` in Section 4.

Minimizing this quadratic model (11) subject to the orthogonality constraints replaces (10) by

$$\mathbf{B}_i^{(k+1)} = \underset{\mathbf{B}_i \in \mathcal{B}_i}{\operatorname{argmin}} \hat{\Phi}_i(\mathbf{B}_i; \mathbf{B}_i^{(k)}). \tag{12}$$

By taking the gradient of the main cost function and dropping constant terms, the above update is equivalent to

$$\min_{\mathbf{B}_i \in \mathcal{B}_i} \left\| \mathbf{B}_i - \widetilde{\mathbf{B}}_i \right\|_F^2, \tag{13}$$

where

$$\widetilde{\mathbf{B}}_i = \mathbf{B}_i^{(k)} - \beta \, \nabla_{\mathbf{B}_i} \Phi_i(\mathbf{B}_i^{(k)}). \tag{14}$$

We use the chain rule to compute the gradient $\nabla_{\mathbf{B}_i} \Phi_i(\mathbf{B}_i^{(k)})$. So, it is straightforward to show that

$$\nabla_{\mathbf{B}_i} \Phi_i(\mathbf{B}_i^{(k)}) = \nabla_{\mathbf{B}_i} \mathcal{L}_i(\mathbf{A}_i^{(k+1)}(\mathbf{B}_i^{(k)}), \mathbf{B}_i)|_{\mathbf{B}_i = \mathbf{B}_i^{(k)}} - \alpha \, \mathcal{H}_{\mathbf{AB}}^{(k)} \left[\nabla_{\mathbf{A}_i} \mathcal{L}_i\big(\mathbf{A}_i^{(k+1)}(\mathbf{B}_i^{(k)}), \mathbf{B}_i^{(k)}\big) \right]. \tag{15}$$

The term $\mathcal{H}_{\mathbf{AB}}^{(k)}[\cdot] : \mathbb{R}^{n \times r} \to \mathbb{R}^{m \times r}$ is the cross-Hessian contraction operator. Directly instantiating this 4th-order tensor might be computationally prohibitive. Instead, we compute the required Hessian-matrix product efficiently using a nested automatic differentiation strategy (a double grad or Hessian-vector product trick). For a given inner residual $\mathbf{g}_{\mathbf{A}} = \nabla_{\mathbf{A}_i} \mathcal{L}_i$, this operation is mathematically equivalent to:

$$\mathcal{H}_{\mathbf{AB}}^{(k)}[\mathbf{g}_{\mathbf{A}}] = \nabla_{\mathbf{B}_i} \mathrm{Tr} \left({\mathbf{g}_{\mathbf{A}}}^{\top} \nabla_{\mathbf{A}_i} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i) \right) \Big|_{\mathbf{B}_i = \mathbf{B}_i^{(k)}}. \tag{16}$$

The optimization problem (13) is a convex quadratic program with linear equality constraints. Let $\mathbf{B}_{\text{int}} = [\mathbf{B}_1, \ldots, \mathbf{B}_{i-1}] \in \mathbb{R}^{m \times (i-1)r}$ denote the column-wise concatenation of all previously learned bases, allowing us to express the orthogonality constraint compactly as $\mathbf{B}_{\text{int}}^{\top} \mathbf{B}_i = \mathbf{0}$.

The minimizer of this problem is simply the exact orthogonal projection of the unconstrained step $\widetilde{\mathbf{B}}_i$ onto the null space of $\mathbf{B}_{\text{int}}^{\top}$. By solving the KKT conditions, this yields the closed-form geometric update:

$$\mathbf{B}_i^{(k+1)} = \underbrace{\left[\mathbf{I} - \mathbf{B}_{\text{int}} \big(\mathbf{B}_{\text{int}}^{\top} \mathbf{B}_{\text{int}} \big)^{-1} \mathbf{B}_{\text{int}}^{\top} \right]}_{\triangleq \mathcal{P}_{B_i}^{\perp}} \widetilde{\mathbf{B}}_i = \mathrm{Proj}_{\mathrm{span}\{\mathbf{B}_j\}_{j<i}^{\perp}} \left(\widetilde{\mathbf{B}}_i \right). \tag{17}$$

The projector $\mathcal{P}_{B_i}^{\perp}$ strips gradient components overlapping with past concepts. Given $\mathbf{B}_i^{(k+1)}$, the lower-level factor is updated by (17):

$$\mathbf{A}_i^{(k+1)} = \mathbf{A}_i^{(k)} - \alpha \nabla_{\mathbf{A}_i} \mathcal{L}_i \left(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k+1)} \right). \tag{18}$$

After K bilevel iterations we freeze $(\mathbf{A}_i, \mathbf{B}_i)$, append $\mathbf{B}_i$ to $\mathbf{B}_{\text{int}}$, and update $\mathbf{W}_i = \mathbf{W}_0 + \sum_{j \le i} \mathbf{A}_j \mathbf{B}_j^{\top}$. The same procedure is applied to all adapted layers, and processing the concept stream $i = 1, \ldots, T$ in sequence yields the full method (Algorithm 1; in practice each bilevel iteration uses S_A and S_B local gradient steps for the two factors). The cross-Hessian–vector product costs one additional backward pass per $\mathbf{B}$-step; Section 4.3 quantifies the resulting overhead.

3.3 Theoretical Results

SeqLoRA uses quadratic surrogate models $\hat{\mathcal{L}}_i$ and $\hat{\Phi}_i$ to simplify bilevel optimization. A key question is whether minimizing these local approximations guarantees convergence to a critical point of the main loss function in (4), which is addressed in the following theorem.

Theorem 1. *Under Assumption 1 (see Appendix A), let $\mathcal{L}_i(\mathbf{A}, \mathbf{B}) + \delta_{\mathcal{B}_i}(\mathbf{B})$ be the composite objective, $\delta_{\mathcal{B}_i}(\mathbf{B})$ is an indicator function, and let $\{(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})\}_{k=0}^{K}$ be generated by Algorithm 1. If the step sizes satisfy $0 < \alpha < \frac{1}{L}$ and $0 < \beta < \frac{1}{L_\Phi}$, where $L_\Phi \le L(1 + \alpha L)^2 + \alpha \rho M$, then:*

1. ***Monotonic Descent:** The composite objective decreases monotonically at each iteration k. In particular, the step sizes $\alpha^* = \frac{1}{2L}$ and $\beta^* = \frac{1}{2L_\Phi}$ maximize the descent coefficients, yielding the sharpest guaranteed decrease.*

2. ***Asymptotic Stationarity:*** *The sequence of iterates* $\{(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})\}$ *asymptotically converges to a critical point of the constrained optimization problem* (4).

Proof. See Appendix B. □

We next bound the *forgetting* of concept j, defined as the increase $\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_j)$ of its loss between the model θ_j obtained right after learning it and the final model θ_T. Building on Example 1, we model the part of each layer's activations *not* captured by the learned basis as a matrix sub-Gaussian process and propagate the resulting per-layer crosstalk through the network.

Theorem 2 (Hanson-Wright High-Probability Bound on Catastrophic Forgetting)**.** *Under Assumption 2 (see Appendix A), let* $\Gamma_\ell \triangleq \prod_{\ell'=\ell+1}^{L} \gamma_{\ell'} \|\mathbf{W}_T^{(\ell')}\|_2$ *be the downstream amplification factor at layer* ℓ*, let* $\mathbf{C}_j^{(\ell)} = \sum_{k=j+1}^{T} \mathbf{A}_k^{(\ell)} (\mathbf{B}_k^{(\ell)})^\top$ *be the crosstalk operator for concept* j *with* $\kappa_\ell \triangleq \|\mathbf{C}_j^{(\ell)}\|_2$*, and let* $\epsilon_\ell \triangleq Tr(\mathbf{\Psi}_j^{(\ell)}) Tr(\mathbf{\Sigma}_j^{(\ell),\perp})$ *be the residual activation energy at layer* ℓ*. For any* $\xi \in (0,1)$*, with probability at least* $1-\xi$*, the end-to-end catastrophic forgetting satisfies*

$$\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_j) = \mathcal{O}\left(\sum_{\ell=1}^{L} \Gamma_\ell \kappa_\ell \sqrt{\epsilon_\ell} \right), \tag{19}$$

where $\theta_T = \{\mathbf{W}_T^{(\ell)}\}_{\ell=1}^L$ *and* $\theta_j = \{\mathbf{W}_j^{(\ell)}\}_{\ell=1}^L$ *are the full model's parameters after learning concept* T *and* j*, respectively. Additionally, minimizing the expected residual energy* $Tr(\mathbf{\Psi}_j^{(\ell)}) Tr(\mathbf{\Sigma}_j^{(\ell),\perp})$ *requires the rank-*r *basis* $\mathbf{B}_j^{(\ell)}$ *to span the top-*r *principal components of* $\mathbf{\Sigma}_j^{(\ell)}$*, the activation covariance of concept* j *at layer* ℓ*, within the available orthogonal complement. Thus, randomly frozen bases are generically suboptimal, yielding higher residual energy unless the remaining feature space is isotropic.*

Proof. See Appendix C. □

The explicit form of the bound, with all constants, is (83) in Appendix C. Theorem 2 shows that forgetting at each layer is governed by the residual covariance $\mathbf{\Sigma}_j^{(\ell),\perp}$, which measures the activation energy that *leaks* outside concept j's learned subspace $\mathrm{col}(\mathbf{B}_j^{(\ell)})$. Orthogonality ensures this leaked energy is the only source of interference. The downstream factors Γ_ℓ is typical of norm-based bounds for deep networks (Bartlett et al., 2017; Neyshabur et al., 2018). The bound is minimized when each $\mathbf{B}_j^{(\ell)}$ aligns with the top principal components of its local activations. SeqLoRA does not necessarily compute this eigenspace; it learns $\mathbf{B}_j^{(\ell)}$ by minimizing the end-to-end denoising loss (Remark 1). The theorem therefore explains *why* a learned basis can reduce interference where a frozen one cannot. Appendix D interprets the constants, Appendix F provides a numerical verification in a small nonlinear network, while Section 4 empirically evaluates the resulting prediction that learned orthogonal bases exhibit less forgetting and interference than randomly frozen bases.

4 Experiments

Baselines and data. We evaluate SeqLoRA against state-of-the-art baselines: Mix-of-Show (Gu et al., 2023), Orthogonal Adaptation (Po et al., 2024a), LoRACLR (Simsar et al., 2025), and `Continual Alternating` (see Subsection 3.2) to isolate the contribution of the cross-Hessian coupling. To assess scalability, we vary the number of concepts from 8 to 101. All methods are evaluated using prompts from the CustomConcept101 dataset (Kumari et al., 2023), generating 16 images per prompt for metric computation.

Metrics and protocols. We measure identity preservation with DINO/DINOv2/DINOv3, holistic similarity with DreamSim, image and text alignment with CLIP-I/CLIP-T, and visual quality with HPSv2/HPSv3 (Appendix G). Three protocols are used. *Single-concept retention* queries the final T-concept model for each concept in isolation and tests whether learning T concepts into one model degrades any of them. *Multi-concept composition* renders three or four concepts in one image under the regional sketch/keypose protocol of Gu et al. (2023) and measures, from each concept's region, its DINOv2 similarity to its own references (*identity*) and to the co-composed concepts' references

Table 1: **Single-concept retention after learning 32 concepts.** Mean ± standard error over the 32 concepts (16 images each); best in bold, second best underlined.

	Reference similarity					Prompt following & aesthetics		
Method	DINO↑	DINOv2↑	DINOv3↑	DreamSim↑	CLIP-I↑	CLIP-T↑	HPSv2↑	HPSv3↑
Mix-of-Show	0.436 ± 0.023	0.377 ± 0.020	0.378 ± 0.017	0.448 ± 0.017	0.677 ± 0.010	0.276 ± 0.003	0.271 ± 0.001	8.039 ± 0.457
Orthogonal Adaptation	0.428 ± 0.023	0.361 ± 0.020	0.359 ± 0.017	0.442 ± 0.018	0.671 ± 0.009	0.277 ± 0.003	0.271 ± 0.001	8.336 ± 0.432
LoRACLR	0.434 ± 0.023	0.372 ± 0.020	0.371 ± 0.017	0.445 ± 0.018	0.674 ± 0.010	0.277 ± 0.003	**0.271** ± 0.001	8.382 ± 0.431
Continual Alternating	0.463 ± 0.020	0.412 ± 0.017	0.406 ± 0.015	0.470 ± 0.015	0.688 ± 0.008	**0.287** ± 0.004	0.271 ± 0.001	**9.242** ± 0.338
SeqLoRA (ours)	**0.468** ± 0.021	**0.418** ± 0.019	**0.418** ± 0.016	**0.477** ± 0.016	**0.693** ± 0.009	0.286 ± 0.003	0.270 ± 0.001	9.043 ± 0.399

Table 2: **Quantitative multi-concept regional composition at 32 concepts** (DINOv2 features, 5 seeds per composition). Identity is similarity of each concept's region to its own references (↑); leakage is similarity of each concept's region to *co-composed* concepts' references (↓); CLIP-T measures text alignment (↑). SeqLoRA attains the lowest cross-concept leakage; the reduction over Continual Alternating is statistically significant (paired Wilcoxon $p = 0.014$); margin is identity minus leakage.

Method	Identity↑	Leakage↓	Margin↑	CLIP-T↑
Mix-of-Show	**0.374**	0.234	0.140	**0.262**
Orthogonal Adaptation	0.344	0.231	0.113	0.259
LoRACLR	0.349	0.230	0.119	0.258
Continual Alternating	0.366	0.223	0.143	0.259
SeqLoRA (ours)	0.347	**0.203**	**0.144**	0.257

(*leakage*); it tests interference directly. *Forgetting* compares each concept's score right after it is learned with its score in the final model. Tables report mean ± standard error across concepts; figures report means (Appendix H). SeqLoRA uses $K = 3$ bilevel iterations, $S_A = S_B = 2$ local steps, $r = 4$, and $\epsilon = 10^{-8}$ unless stated otherwise; all methods use the same rank and data per concept, on 2× NVIDIA A100 GPUs.

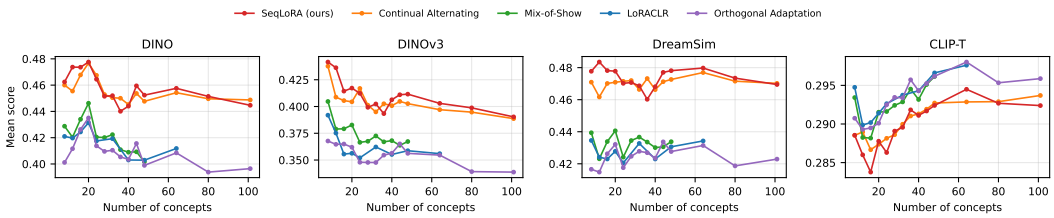
4.1 Main results

Single-concept retention. Table 1 reports the 32-concept experiment. The two sequential methods, SeqLoRA and Continual Alternating, outperform all three fusion-based baselines on every identity and similarity metric (DINO, DINOv2, DINOv3, DreamSim, CLIP-I) by a wide margin, and SeqLoRA is the better of the two on all five; CLIP-T and HPSv2/HPSv3 are within one standard error across the sequential methods, with Continual Alternating slightly ahead on HPSv3. The dominant gain thus comes from the constrained continual formulation, and the bilevel update adds a consistent further improvement in identity preservation. Figure 1 and Appendix J show the images for all 32 concepts.

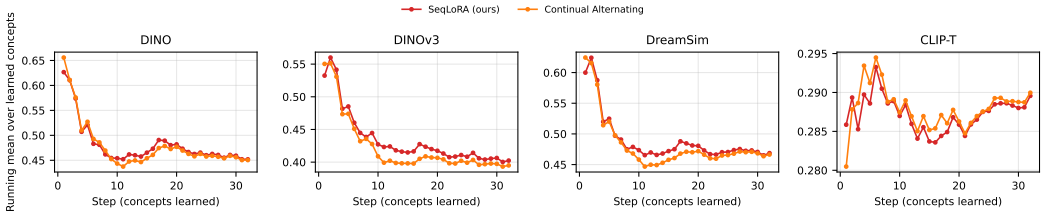
Multi-concept composition. Table 2 quantifies interference on multi-concept regional composition at 32 concepts. SeqLoRA attains the lowest cross-concept leakage of all methods, and its improvement over Continual Alternating, the closest method, is statistically significant (paired Wilcoxon test over compositions, $p = 0.014$). Its in-region identity (0.347) lies within the range spanned by the baselines: Mix-of-Show and Continual Alternating obtain higher absolute identity but at higher leakage, so that the identity–leakage margin of SeqLoRA (0.144) is on par with Continual Alternating (0.143) and above the fusion-based methods, while text alignment is essentially unchanged across methods. Mix-of-Show, whose fusion step is optimized for exactly this composition, remains the strongest on absolute in-region identity. Leakage is the quantity the orthogonality constraint targets, and it is where learned and frozen orthogonal bases differ most: Orthogonal Adaptation, which also composes by summation, has leakage 0.231 against 0.203. Although SeqLoRA is not the best on the identity score, the lower leakage is visible in Figure 7 (Appendix K): at the same seed and control inputs, SeqLoRA keeps each concept's appearance inside its own region, whereas the baselines mix attributes across regions.

4.2 Scalability and forgetting

Scalability. Figure 2a reports final-model performance for $T \in \{8, \ldots, 101\}$ (all metrics in Table 14). The identity metrics of the two sequential methods stay well above those of the fusion-based methods at every T, and the gap does not close as T grows. Mix-of-Show and LoRACLR fail



(a) Final-model performance versus number of concepts ($T = 8$ to 101), averaged over all T concepts. Mix-of-Show and LoRACLR run out of memory beyond 48 and 64 concepts.



(b) Running mean over all concepts learned so far, after each step of the 32-concept stream. The evaluated pool grows with the step, so the curve mixes forgetting with the difficulty of newly added concepts; Table 3 isolates forgetting per concept.

Figure 2: Scalability (a) and stability during sequential training (b) on four representative metrics; all eight metrics are in Figures 11 and 12.

Table 3: **Forgetting for SeqLoRA across stream lengths.** Each concept's score right after it is learned minus its score in the final T-concept model, mean $\pm$ standard error over concepts; initial and final scores in Table 12.

T	CLIP-I	CLIP-T	DINO	DINOv2	DINOv3	DreamSim	HPSv2	HPSv3
8	0.014 ± 0.006	-0.002 ± 0.001	0.022 ± 0.009	0.025 ± 0.009	0.026 ± 0.010	0.017 ± 0.008	-0.001 ± 0.000	-0.025 ± 0.020
16	0.009 ± 0.004	-0.002 ± 0.001	0.018 ± 0.007	0.020 ± 0.007	0.022 ± 0.008	0.014 ± 0.005	-0.000 ± 0.000	0.019 ± 0.029
32	0.008 ± 0.003	-0.001 ± 0.000	0.013 ± 0.004	0.016 ± 0.005	0.017 ± 0.005	0.012 ± 0.004	-0.000 ± 0.000	0.001 ± 0.025

with out-of-memory errors at 64 and 80 concepts, because their fusion step must hold all per-concept adapters (and, for optimization-based fusion, the optimizer state) in memory at once; SeqLoRA and Orthogonal Adaptation, which compose by summation, run to 101 concepts (Section 4.3).

Forgetting. Figure 2b tracks the running mean over all concepts learned so far along the 32-concept stream: SeqLoRA is at or above Continual Alternating on the visual-fidelity metrics at almost every step. Because the evaluated pool grows, this curve measures system-level stability; Table 3 therefore compares each concept's score right after it is learned with its score in the final model. The drop is small (e.g., 0.013 DINO and 0.008 CLIP-I at 32 concepts) and does not grow with the stream length, while CLIP-T and HPSv2 are unchanged: SeqLoRA mitigates forgetting substantially but does not eliminate it, consistent with Theorem 2. Appendix L gives per-step numbers and images of every concept across training.

4.3 Efficiency

Table 4 compares end-to-end cost at 32 concepts on the same hardware. Parallel methods (LoRACLR, Mix-of-Show, Orthogonal Adaptation) train per-concept adapters independently ($\sim$78s each) and fuse them post-hoc, while SeqLoRA and Continual Alternating learn sequentially in a single model with no fusion step. The accumulating orthogonality constraints raise the per-concept cost to $\sim$454s for Continual Alternating and $\sim$660s for SeqLoRA (the latter due to cross-Hessian evaluations). However, while parallel methods can leverage distributed computing to reduce wall-clock training time (at the cost of higher total compute), they require a complex and often time-consuming optimization step to fuse the independent adapters into a single multi-concept model. At 32 concepts, post-hoc fusion itself takes $\sim$33 min for Mix-of-Show and $\sim$8 min for LoRACLR (Orthogonal Adaptation fuses in 15s via simple summation), an overhead SeqLoRA avoids entirely while also supporting truly sequential lifelong learning without data accumulation.

Table 4: **End-to-end cost at 32 concepts on the same hardware.** Training time per concept, and fusion time and peak memory for one composition. The fusion step is repeated for every new subset of concepts; SeqLoRA and Continual Alternating compose by summation (memory versus T in Table 9).

Method	Train / concept	Fusion time	Peak fusion mem.
Orthogonal Adaptation	∼78 s	15 s	—
LoRACLR	∼78 s	8 min	44.7 GB
Mix-of-Show	∼78 s	33 min	18.4 GB
Continual Alternating	∼454 s	none	—
SeqLoRA (ours)	∼660 s	none	—

4.4 Robustness and ablations

Full tables are in Appendix I.

Concept order. Over four random orderings of a 16-concept stream (five seeds each), composition identity varies by ± 0.007 and leakage by ± 0.006 (coefficient of variation $\leq 3\%$; Table 6), so the sequential projection does not make later concepts systematically worse.

LoRA rank. Raising the rank from 4 to 8 or 16 increases leakage (0.200 to 0.221/0.219) without improving the identity–leakage margin (Table 7); since the constraint admits at most $\lfloor m/r \rfloor$ concepts per layer, a small rank is preferable for capacity and interference alike.

Concept similarity. The projection matters most where interference is largest: for visually similar concepts it reduces composition leakage from 0.434 to 0.386, whereas for a diverse group leakage is already low and unchanged (0.070 vs. 0.078; Table 8).

Hyperparameters. Varying local steps, bilevel iterations and ϵ changes identity metrics by less than one standard error (Table 10).

Backbone. The projection is backbone-agnostic; a preliminary port to SDXL (Podell et al., 2024) (four concepts) trains unchanged and reaches a mean DINO identity of 0.67 with a $+0.47$ margin over cross-concept similarity (Table 11).

5 Limitations and Discussion

Capacity. The orthogonality constraint admits at most $\lfloor m/r \rfloor$ concepts per layer (192 for m=768, r=4), which covers the scales we evaluate; later concepts are learned in a shrinking complement, and raising r increases leakage (Table 7).

Compute. The bilevel formulation adds a cross-Hessian–vector product per iteration, increasing per-concept training cost (Table 4), but avoids the costly post-hoc fusion required by parallel methods (e.g., ∼33 minutes for Mix-of-Show at 32 concepts).

Theory and scope. Our theoretical results rely on Assumptions 1 and 2, which are mild standard conditions discussed in Appendix A. Current experiments use Stable Diffusion v1.5 for consistency with prior work; a preliminary extension to SDXL is provided in Table 11.

6 Conclusion

SeqLoRA casts multi-concept customization as constrained continual learning: both LoRA factors of each new concept are learned, and the basis is kept orthogonal to all previous bases by a closed-form projection inside a bilevel update. We prove monotone descent and convergence to critical points of the constrained problem, derive a high-probability forgetting bound for nonlinear multi-layer networks whose basis-dependent term is the residual activation energy, and show that a data-adapted basis reduces this term where a random frozen basis cannot. On Stable Diffusion with up to 101 concepts, SeqLoRA improves identity preservation over fusion-based methods, attains the lowest cross-concept leakage in compositions, needs no fusion step, and scales past the point at which fusion runs out of memory; the bilevel correction gives a consistent further gain over alternating minimization.

Ethics Statement

SeqLoRA advances multi-concept personalization of text-to-image diffusion models, which has broad positive applications in creative design, media production, and accessibility. As with all generative modeling research, improvements in image fidelity could potentially be misused for generating misleading visual content. However, SeqLoRA does not introduce new generative capabilities beyond what existing diffusion models already provide; it improves the composability of personalized adaptations within established frameworks. We encourage the responsible use of such technology in accordance with applicable guidelines and regulations.

AI Use Statement

LLMs were used only to edit the text for readability and to review the manuscript for consistency of notation. They were not used to develop the method, to formulate or prove the mathematical claims, or to design experiments. All AI-assisted edits were reviewed and verified by the authors.

References

Hedy Attouch, Jérôme Bolte, and Benar Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. *Mathematical Programming*, 137, 01 2011. doi: 10.1007/s10107-011-0484-9.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6240–6249, 2017.

Jerome Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014. doi: 10.1007/s10107-013-0701-9. URL `https://inria.hal.science/hal-00916090`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

A Philip Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274, 1981.

Stephan Dempe. *Foundations of Bilevel Programming*, volume 61 of *Nonconvex Optimization and Its Applications*. Springer Science & Business Media, 2002. doi: 10.1007/978-1-4615-1503-8.

Dmitriy Drusvyatskiy, Adrian S Lewis, and Catherine A Paquette. Efficiency of the prox-linear algorithm for composite optimization. *Mathematical Programming*, 178:503–558, 2019.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International conference on artificial intelligence and statistics*, pages 1082–1092. PMLR, 2020.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

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

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36:15890–15902, 2023.

A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman & Hall/CRC, 2000.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

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. In *International Conference on Learning Representations (ICLR)*, 2022.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23638–23647, 2024.

Ziwei Luo, Fredrik Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas Schön. Taming diffusion models for image restoration: a review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 383(2299), 2025.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *International Conference on Learning Representations (ICLR)*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit Bermano, Eric Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. In *Computer graphics forum*, volume 43, page e15063. Wiley Online Library, 2024b.

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 *International Conference on Learning Representations*, volume 2024, pages 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23 (2):1126–1153, 2013.

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*, pages 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*, pages 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Enis Simsar, Thomas Hofmann, Federico Tombari, and Pinar Yanardag. Loraclr: Contrastive adaptation for customization of diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13189–13198, 2025.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

CONTENTS

1 Introduction **1**

2 Problem Formulation **2**

2.1 Continual Learning over concepts 2

2.2 Single-Concept Fine-Tuning with LoRA 3

2.3 Multi-Concept Aggregation and the Interference Problem 3

2.4 Desiderata 3

3 Proposed Method **4**

3.1 Constrained Optimization Problem 4

3.2 Bilevel Optimization Algorithm 4

3.3 Theoretical Results 5

4 Experiments **6**

4.1 Main results 7

4.2 Scalability and forgetting 7

4.3 Efficiency 8

4.4 Robustness and ablations 9

5 Limitations and Discussion **9**

6 Conclusion **9**

A Mathematical Preliminaries **15**

A.1 Notations 15

A.2 Definitions 16

A.3 Assumptions 16

B Proof of Theorem 1 **18**

C Proof of Theorem 2 **21**

C.1 Analysis of Catastrophic Forgetting 21

C.2 Optimal basis selection inside the free complement 25

D Decomposition of Catastrophic Forgetting: Mean Interference vs. Stochastic Deviation **27**

E Population Risk Analysis in a Stylized Linear Setting **29**

F Numerical Illustration of Theorem 2 in a Small Nonlinear Network **31**

G Evaluation Metrics **34**

H Evaluation Protocols **34**

H.1 Single-concept retention (Table 1, Figure 2a, Table 14) 34

H.2 Forgetting (Figure 2b, Table 3, Table 13) 34

H.3 Multi-concept composition (Table 2, Figure 7) 34

I Additional Experiments **35**

I.1 Sensitivity to concept order 35

I.2 Sensitivity to LoRA rank 35

I.3 Effect of concept similarity 35

I.4 Compute and memory 36

I.5 Hyperparameter ablation 36

I.6 Preliminary transfer to SDXL 36

J Qualitative Results **36**

K Multi-Concept Regional Generation **40**

L Additional Forgetting Results **41**

M Full Scalability and Stability Results **45**

Algorithm 1 Sequential Regularized LoRA (SeqLoRA)

Require: Pretrained weights $\{\mathbf{W}_0^{(\ell)}\}_{\ell=1}^{L}$; concept datasets $\mathcal{D}_1, \ldots, \mathcal{D}_T$; step sizes α, β; bilevel iterations K; local iterations S_B, S'_A; regularization ϵ

Ensure: LoRA factors $\{(\mathbf{A}_i^{(\ell)}, \mathbf{B}_i^{(\ell)})\}_{i,\ell}$

for concept $i = 1$ to T **do**
  Initialize $\{\mathbf{A}_i^{(\ell,0)}, \mathbf{B}_i^{(\ell,0)}\}_{\ell=1}^{L}$ randomly
  $\mathbf{B}_{\text{int}}^{(\ell)} \leftarrow [\mathbf{B}_1^{(\ell)}, \ldots, \mathbf{B}_{i-1}^{(\ell)}]$ **for all** $\ell \in \{1, \ldots, L\}$ ▷ Previous bases at each layer
  for $k = 0$ to $K - 1$ **do** ▷ Bilevel iterations

    // Step 1: Tentative lower-level update of $\{\mathbf{A}_i^{(\ell)}\}_{\ell=1}^{L}$
    $\widetilde{\mathbf{A}}^{(\ell)} \leftarrow \mathbf{A}_i^{(\ell,k)} - \alpha \nabla_{\mathbf{A}^{(\ell)}} \mathcal{L}_i\Big(\{\mathbf{A}_i^{(\ell,k)}, \mathbf{B}_i^{(\ell,k)}\}_{\ell=1}^{L}\Big)$ **for all** ℓ

    // Steps 2–3: Reduced gradient + projected $\{\mathbf{B}_i^{(\ell)}\}_{\ell=1}^{L}$ *update*
    $\mathbf{B}^{(\ell)} \leftarrow \mathbf{B}_i^{(\ell,k)}$ **for all** ℓ
    for $s = 1$ to S_B **do** ▷ Local B-iterations
      $\mathbf{g}_{\mathbf{B}}^{(\ell)} \leftarrow \nabla_{\mathbf{B}_i^{(\ell)}} \mathcal{L}_i(\{\widetilde{\mathbf{A}}^{(\ell)}, \mathbf{B}^{(\ell)}\}_{\ell=1}^{L})$ **for all** ℓ ▷ Direct partial
      $\mathbf{g}_{\mathbf{A}}^{(\ell)} \leftarrow \nabla_{\mathbf{A}_i^{(\ell)}} \mathcal{L}_i(\{\widetilde{\mathbf{A}}^{(\ell)}, \mathbf{B}^{(\ell)}\}_{\ell=1}^{L})$ **for all** ℓ ▷ Inner residual
      // Cross-Hessian via autodiff at current iterate
      $\mathcal{H}_{\mathbf{AB}}^{(k,\ell)}[\mathbf{g}_{\mathbf{A}}^{(\ell)}] \leftarrow \nabla_{\mathbf{B}_i^{(\ell)}} \text{Tr}\Big((\mathbf{g}_{\mathbf{A}}^{(\ell)})^\top \nabla_{\mathbf{A}_i^{(\ell)}} \mathcal{L}_i(\mathbf{A}_i^{(\ell,k)}, \mathbf{B}_i^{(\ell)})\Big)\Big|_{\mathbf{B}_i^{(\ell)}=\mathbf{B}_i^{(\ell,k)}}$ **for all** ℓ
      $\nabla_{\mathbf{B}^{(\ell)}} \Phi_i \leftarrow \mathbf{g}_{\mathbf{B}}^{(\ell)} - \alpha\, \mathcal{H}_{\mathbf{AB}}^{(k,\ell)}[\mathbf{g}_{\mathbf{A}}^{(\ell)}]$ **for all** ℓ ▷ Reduced gradient
      $\widetilde{\mathbf{B}}^{(\ell)} \leftarrow \mathbf{B}^{(\ell)} - \beta \nabla_{\mathbf{B}^{(\ell)}} \Phi_i$ **for all** ℓ
      $\mathbf{B}^{(\ell)} \leftarrow \left[\mathbf{I} - \mathbf{B}_{\text{int}}^{(\ell)}\Big((\mathbf{B}_{\text{int}}^{(\ell)})^\top \mathbf{B}_{\text{int}}^{(\ell)} + \epsilon \mathbf{I}\Big)^{-1} (\mathbf{B}_{\text{int}}^{(\ell)})^\top\right] \widetilde{\mathbf{B}}^{(\ell)}$ **for all** ℓ ▷ Regularized projection
    end for

    $\mathbf{B}_i^{(\ell,k+1)} \leftarrow \mathbf{B}^{(\ell)}$ **for all** ℓ

    // Step 4: Final update of $\{\mathbf{A}_i^{(\ell)}\}_{\ell=1}^{L}$ *using* $\{\mathbf{B}_i^{(\ell,k+1)}\}_{\ell=1}^{L}$
    $\mathbf{A}^{(\ell)} \leftarrow \mathbf{A}_i^{(\ell,k)}$ **for all** ℓ
    for $s = 1$ to S'_A **do** ▷ Local A-iterations at new $\{\mathbf{B}^{(\ell)}\}$
      $\mathbf{A}^{(\ell)} \leftarrow \mathbf{A}^{(\ell)} - \alpha \nabla_{\mathbf{A}^{(\ell)}} \mathcal{L}_i\Big(\{\mathbf{A}^{(\ell)}, \mathbf{B}_i^{(\ell,k+1)}\}_{\ell=1}^{L}\Big)$ **for all** ℓ
    end for

    $\mathbf{A}_i^{(\ell,k+1)} \leftarrow \mathbf{A}^{(\ell)}$ **for all** ℓ

  end for
  Store $(\mathbf{A}_i^{(\ell)}, \mathbf{B}_i^{(\ell)}) \leftarrow (\mathbf{A}_i^{(\ell,K)}, \mathbf{B}_i^{(\ell,K)})$ **for all** ℓ
end for

A Mathematical Preliminaries

This section gathers the notation, definitions, and assumptions used in the paper.

A.1 Notations

We use bold lowercase letters for vectors and bold uppercase letters for matrices. For a matrix $\mathbf{A}$, $\mathbf{A}^\top$, $\text{Tr}(\mathbf{A})$, and $\text{vec}(\mathbf{A})$ denote its transpose, trace, and vectorization, respectively. The Frobenius norm and the operator (spectral) norm are denoted by $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_2$. The Kronecker product is denoted by $\mathbf{A} \otimes \mathbf{B}$. For a matrix $\mathbf{B}$, $\text{col}(\mathbf{B})$ denotes its column space. The sub-Gaussian norm of a random variable Z is denoted by $\|Z\|_{\psi_2}$. We use superscripts in parentheses (e.g., $\mathbf{W}^{(\ell)}$) to index

network layers; subscripts denote either the LoRA factors associated with concept i (e.g., $\mathbf{A}_i, \mathbf{B}_i$) or the full parameter state after learning concept i (e.g., $\mathbf{W}_i$), with the meaning clear from context.

Lower-level iterates. In the bilevel formulation the lower-level solution depends on the upper-level variable, and we write $\mathbf{A}_i^{(k+1)}(\mathbf{B}) = \mathbf{A}_i^{(k)} - \alpha \nabla_{\mathbf{A}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B})$ to make this dependence explicit (Eq. (8)). Two instances recur: the *tentative* step at the current basis, $\widetilde{\mathbf{A}}_i^{(k)} \triangleq \mathbf{A}_i^{(k+1)}(\mathbf{B}_i^{(k)})$, and the *accepted* iterate at the projected basis, $\mathbf{A}_i^{(k+1)} \triangleq \mathbf{A}_i^{(k+1)}(\mathbf{B}_i^{(k+1)})$. With this convention the reduced objective of Eq. (9) satisfies $\Phi_i(\mathbf{B}_i^{(k)}) = \mathcal{L}_i(\widetilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)})$ and $\Phi_i(\mathbf{B}_i^{(k+1)}) = \mathcal{L}_i(\mathbf{A}_i^{(k+1)}, \mathbf{B}_i^{(k+1)})$, identities that are used in Appendix B. When concept and iteration indices are irrelevant we drop them and write $\mathbf{A}^{(k+1)}(\mathbf{B})$ with base point $\mathbf{A}_0 = \mathbf{A}_i^{(k)}$.

A.2 Definitions

Definition 1. *(Attouch et al., 2011) The subdifferential of a PLSC function g at $\mathbf{x} \in \mathbb{R}^n$ is defined as*

$$\partial g(\mathbf{x}) \stackrel{\triangle}{=} \{\zeta \in \mathbb{R}^n \mid \exists \mathbf{x}_k \to \mathbf{x},\ g(\mathbf{x}_k) \to g(\mathbf{x}),\ \zeta_k \to \zeta \in \partial \hat{g}(\mathbf{x}_k)\}$$

where $\partial \hat{g}(\mathbf{x}_k)$ is the Fréchet subdifferential of g at $\mathbf{x}_k \in \mathbb{R}^n$, defined as

$$\partial \hat{g}(\mathbf{x}_k) = \left\{ \zeta \in \mathbb{R}^n \mid \liminf_{\mathbf{v} \neq \mathbf{x}, \mathbf{v} \to \mathbf{x}} \frac{1}{\|\mathbf{v} - \mathbf{x}\|_2^2} \left[g(\mathbf{v}) - g(\mathbf{x}) - \langle \mathbf{v} - \mathbf{x}, \zeta \rangle \right] \geq 0 \right\}. \tag{20}$$

Definition 2. *(Attouch et al., 2011) A point $\mathbf{x}^*$ is called a critical point of a PLSC function $f(\mathbf{x})$ if $0 \in \partial f(\mathbf{x}^*)$.*

Lemma 3 (Descent lemma (Bolte et al., 2014))**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function whose gradient ∇f is L-Lipschitz continuous. Then, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2. \tag{21}$$

Definition 3 (Operator norm of the cross-Hessian (Horn and Johnson, 2012))**.** *For a linear map $\mathcal{H} : \mathbb{R}^{n \times r} \to \mathbb{R}^{m \times r}$, the operator norm is defined as:*

$$\|\mathcal{H}\|_2 = \sup_{\mathbf{V} \neq \mathbf{0}} \frac{\|\mathcal{H}[\mathbf{V}]\|_F}{\|\mathbf{V}\|_F}. \tag{22}$$

A.3 Assumptions

To establish the convergence of the SeqLoRA algorithm (Theorem 1), we require the following standard regularity conditions on the optimization landscape:

Assumption 1. *The loss function $\mathcal{L}_i(\mathbf{A}, \mathbf{B})$ satisfies the following properties:*

- *$\mathcal{L}_i$ is jointly L-smooth with respect to $\mathbf{A}$ and $\mathbf{B}$ in the Frobenius norm. This inherently bounds the operator norm of the cross-Hessian: $\|\nabla^2_{\mathbf{AB}} \mathcal{L}_i(\mathbf{A}, \mathbf{B})\|_2 \leq L$.*
- *The cross-Hessian is ρ-Lipschitz continuous with respect to $\mathbf{B}$:*

$$\|\nabla^2_{\mathbf{AB}} \mathcal{L}_i(\mathbf{A}_0, \mathbf{B}_1) - \nabla^2_{\mathbf{AB}} \mathcal{L}_i(\mathbf{A}_0, \mathbf{B}_2)\|_2 \leq \rho \|\mathbf{B}_1 - \mathbf{B}_2\|_F. \tag{23}$$

- *The partial gradient of the loss with respect to $\mathbf{A}$ is bounded by a constant $M > 0$:*

$$\|\nabla_{\mathbf{A}} \mathcal{L}_i(\mathbf{A}, \mathbf{B})\|_F \leq M. \tag{24}$$

Discussion of Assumption 1. These are standard regularity conditions in the analysis of non-convex and bilevel optimization (Fallah et al., 2020). Joint smoothness holds locally in any bounded region where training is stable. The bounded gradient assumption for the inner variable is a mild condition, further justified by the common deep learning practice of gradient clipping, which explicitly enforces such a bound to prevent training instability. The Lipschitz continuity of the cross-Hessian is a second-order smoothness condition necessary for analyzing algorithms that leverage parameter coupling.

Using the operator norm is the rigorous choice, as the cross-Hessian is a linear operator (a 4th-order tensor) mapping one matrix space to another.

To bound the catastrophic forgetting across the continuous learning stream (Theorem 2), we employ the following structural assumptions on the network mappings and the statistical distribution of the residual activations:

Assumption 2 (Multi-Layer Structure and Sub-Gaussian Residuals)**.** *The network applies LoRA adaptation at L layers. For a concept j at layer ℓ, let $\mathbf{X}_j^{(\ell)} \in \mathbb{R}^{m_\ell \times p}$ be the input activations and let $\mathcal{P}_{B_j}^{(\ell),\perp}$ denote the orthogonal projector onto the orthogonal complement of the column space of $\mathbf{B}_j^{(\ell)}$. We assume:*

1. ***Lipschitz Network Mappings:*** *Between consecutive LoRA-adapted layers ℓ and $\ell+1$, the intermediate network operations compose a mapping $\phi_{\ell+1} : \mathbb{R}^{n_\ell \times p} \to \mathbb{R}^{m_{\ell+1} \times p}$ that is γ_ℓ-Lipschitz continuous with respect to the Frobenius norm. Furthermore, the task loss applied to the final layer output is L_o-Lipschitz.*

2. ***Sub-Gaussian Residuals:*** *At each LoRA-adapted layer ℓ, the residual input activation $\mathbf{X}_j^{(\ell),\perp} = \mathcal{P}_{B_j}^{(\ell),\perp} \mathbf{X}_j^{(\ell)}$ follows a matrix sub-Gaussian process:*

$$\mathbf{X}_j^{(\ell),\perp} = (\mathbf{\Sigma}_j^{(\ell),\perp})^{1/2} \mathbf{Z}^{(\ell)} (\mathbf{\Psi}_j^{(\ell)})^{1/2}, \tag{25}$$

where $\mathbf{\Sigma}_j^{(\ell),\perp} \in \mathbb{R}^{m_\ell \times m_\ell}$ captures the feature-dimension covariance, $\mathbf{\Psi}_j^{(\ell)} \in \mathbb{R}^{p \times p}$ captures the token-dimension correlation (sequence dependencies), and $\mathbf{Z}^{(\ell)} \in \mathbb{R}^{m_\ell \times p}$ consists of independent, mean-zero, unit-variance sub-Gaussian entries bounded by $\|Z_{a,b}^{(\ell)}\|_{\psi_2} \leq K$.

Discussion of Assumption 2. These assumptions are based on the standard architecture of modern diffusion models. Item 1 (Lipschitz Mappings) is well-justified because deep networks are compositions of Lipschitz-continuous operations, a property required for stable training. Crucially, Item 2 (Sub-Gaussian Residuals) is an assumption applied only to the *residual* activations. The learned LoRA basis $\mathbf{B}_j^{(\ell)}$ is optimized to capture the dominant, highly-structured components of the concept's features. The residual, therefore, represents the uncaptured "noise floor," which can be modeled as a sub-Gaussian process. This is a mild assumption, as the sub-Gaussian class is broad and includes not only perfect Gaussians but also any bounded distribution (such as those produced by LayerNorm or Tanh) and mixtures thereof. The separable Kronecker covariance structure in (25) is a standard modeling choice in high-dimensional matrix-variate statistics (Dawid, 1981; Gupta and Nagar, 2000), adopted here for analytical tractability.

B PROOF OF THEOREM 1

We analyze the convergence of the SeqLoRA algorithm for the concept i.

Proof. To prove the theorem, we first establish the following lemma.

Lemma 4 (Gradient Lipschitz Continuity of Φ_i)**.** *Under Assumptions 1, the reduced objective function $\Phi_i(\mathbf{B}) \triangleq \mathcal{L}_i(\mathbf{A}^*(\mathbf{B}), \mathbf{B})$ where $\mathbf{A}^*(\mathbf{B}) = \mathbf{A}_0 - \alpha\nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}_0, \mathbf{B})$ and it has Lipschitz continuous gradients (i.e., is L_Φ-smooth) with the Lipschitz constant:*

$$L_\Phi \leq L(1+\alpha L)^2 + \alpha\rho M. \tag{26}$$

That is, for any $\mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{m\times r}$, we have:

$$\|\nabla\Phi_i(\mathbf{B}_1) - \nabla\Phi_i(\mathbf{B}_2)\|_F \leq L_\Phi\|\mathbf{B}_1 - \mathbf{B}_2\|_F. \tag{27}$$

Proof. Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be two arbitrary matrices.

Using the definition of the lower-level update (8) for $\mathbf{A}_0 = \mathbf{A}_i^{(k)}$ and the L-smoothness from Assumption 1:

$$\begin{aligned}\|\mathbf{A}^*(\mathbf{B}_1) - \mathbf{A}^*(\mathbf{B}_2)\|_F &= \|(\mathbf{A}_0 - \alpha\nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}_0, \mathbf{B}_1)) - (\mathbf{A}_0 - \alpha\nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}_0, \mathbf{B}_2))\|_F \\ &= \alpha\|\nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}_0, \mathbf{B}_1) - \nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}_0, \mathbf{B}_2)\|_F \leq \alpha L\|\mathbf{B}_1 - \mathbf{B}_2\|_F. \end{aligned} \tag{28}$$

Using (15), the total derivative of Φ_i with respect to $\mathbf{B}$ is given by:

$$\nabla\Phi_i(\mathbf{B}) = \underbrace{\nabla_{\mathbf{B}}\mathcal{L}_i(\mathbf{A}^*(\mathbf{B}), \mathbf{B})}_{T_1} - \alpha\underbrace{\nabla^2_{\mathbf{AB}}\mathcal{L}_i(\mathbf{A}_0, \mathbf{B})\big[\nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}^*(\mathbf{B}), \mathbf{B})\big]}_{T_2}. \tag{29}$$

To bound the $\|\nabla\Phi_i(\mathbf{B}_1) - \nabla\Phi_i(\mathbf{B}_2)\|_F$, we use the triangle inequality. We first bound the term T_1. Using the Assumption 1 and substituting the bound from (28):

$$\begin{aligned}&\|\nabla_{\mathbf{B}}\mathcal{L}_i(\mathbf{A}^*(\mathbf{B}_1), \mathbf{B}_1) - \nabla_{\mathbf{B}}\mathcal{L}_i(\mathbf{A}^*(\mathbf{B}_2), \mathbf{B}_2)\|_F \leq L\Big(\|\mathbf{A}^*(\mathbf{B}_1) - \mathbf{A}^*(\mathbf{B}_2)\|_F + \|\mathbf{B}_1 - \mathbf{B}_2\|_F\Big) \\ &\leq L\Big(\alpha L\|\mathbf{B}_1 - \mathbf{B}_2\|_F + \|\mathbf{B}_1 - \mathbf{B}_2\|_F\Big) = L(1+\alpha L)\|\mathbf{B}_1 - \mathbf{B}_2\|_F. \end{aligned} \tag{30}$$

To bound the term T_2 in (29), let $\mathcal{H}_j = \nabla^2_{\mathbf{AB}}\mathcal{L}_i(\mathbf{A}_0, \mathbf{B}_j)$ and $G_j = \nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}^*(\mathbf{B}_j), \mathbf{B}_j)$ for $j \in \{1, 2\}$. We need to bound $\|\mathcal{H}_1[G_1] - \mathcal{H}_2[G_2]\|_F$. Adding and subtracting the cross term $\mathcal{H}_1[G_2]$, and applying the triangle inequality and sub-multiplicativity of operator norms:

$$\begin{aligned}&\|\mathcal{H}_1[G_1] - \mathcal{H}_1[G_2] + \mathcal{H}_1[G_2] - \mathcal{H}_2[G_2]\|_F \leq \|\mathcal{H}_1[G_1 - G_2]\|_F + \|(\mathcal{H}_1 - \mathcal{H}_2)[G_2]\|_F \\ &\leq \|\mathcal{H}_1\|_2\|G_1 - G_2\|_F + \|\mathcal{H}_1 - \mathcal{H}_2\|_2\|G_2\|_F. \end{aligned} \tag{31}$$

We now bound each of the four components in the above inequality: by Assumption 1, we have $\|\mathcal{H}_1\|_2 \leq L$ and $\|G_2\|_F \leq M$. Using the ρ-Lipschitz property in Assumption 1 yields

$$\|\mathcal{H}_1 - \mathcal{H}_2\|_2 \leq \rho\|\mathbf{B}_1 - \mathbf{B}_2\|_F$$

Now, we use joint smoothness and (28), resulting in

$$\begin{aligned}&\|G_1 - G_2\|_F = \|\nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}^*(\mathbf{B}_1), \mathbf{B}_1) - \nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}^*(\mathbf{B}_2), \mathbf{B}_2)\|_F \leq \\ &L\Big(\|\mathbf{A}^*(\mathbf{B}_1) - \mathbf{A}^*(\mathbf{B}_2)\|_F + \|\mathbf{B}_1 - \mathbf{B}_2\|_F\Big) \leq L(1+\alpha L)\|\mathbf{B}_1 - \mathbf{B}_2\|_F. \end{aligned} \tag{32}$$

Substituting these four bounds back into (31) yields:

$$\begin{aligned}&\|\mathcal{H}_1[G_1] - \mathcal{H}_1[G_2] + \mathcal{H}_1[G_2] - \mathcal{H}_2[G_2]\|_F \leq L\Big[L(1+\alpha L)\|\mathbf{B}_1 - \mathbf{B}_2\|_F\Big] + \Big[\rho\|\mathbf{B}_1 - \mathbf{B}_2\|_F\Big]M \\ &= \Big(L^2(1+\alpha L) + \rho M\Big)\|\mathbf{B}_1 - \mathbf{B}_2\|_F. \end{aligned} \tag{33}$$

Combining the bounds (30) and (33) gives us:

$$\|\nabla\Phi_i(\mathbf{B}_1) - \nabla\Phi_i(\mathbf{B}_2)\|_F \leq \Big[L(1+\alpha L) + \alpha\big(L^2(1+\alpha L) + \rho M\big)\Big]\|\mathbf{B}_1 - \mathbf{B}_2\|_F$$
$$= \Big[L(1+\alpha L) + \alpha L^2(1+\alpha L) + \alpha\rho M\Big]\|\mathbf{B}_1 - \mathbf{B}_2\|_F = \Big[L(1+\alpha L)(1+\alpha L) + \alpha\rho M\Big]\|\mathbf{B}_1 - \mathbf{B}_2\|_F$$
$$= \Big(L(1+\alpha L)^2 + \alpha\rho M\Big)\|\mathbf{B}_1 - \mathbf{B}_2\|_F. \tag{34}$$

Thus, $\nabla\Phi_i(\mathbf{B})$ is Lipschitz continuous with constant $L_\Phi \leq L(1+\alpha L)^2 + \alpha\rho M$. This completes the proof of Lemma 4. □

Next, we show the rest of the results of Theorem 1.

The tentative lower-level update is $\tilde{\mathbf{A}}_i^{(k)} = \mathbf{A}_i^{(k)} - \alpha\nabla_{\mathbf{A}}\mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})$. Applying the Descent Lemma on the L-smooth loss $\mathcal{L}_i$ with respect to $\mathbf{A}_i$:

$$\mathcal{L}_i(\tilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)}) \leq \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}) + \Big\langle \nabla_{\mathbf{A}_i}\mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}), \tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)} \Big\rangle + \frac{L}{2}\|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\|_F^2. \tag{35}$$

Since $\widetilde{\mathbf{A}}_i^{(k)}$ is the minimizer of (8) when $\mathbf{B}_i = \mathbf{B}_i^{(k)}$, we have

$$\mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}) + \big\langle \nabla_{\mathbf{A}_i}\mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}),\ \tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)} \big\rangle + \frac{1}{2\alpha}\big\|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\big\|_F^2 \leq \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}), \tag{36}$$

Adding (35) into the above inequality gives us

$$\mathcal{L}_i(\tilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)}) \leq \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}) + (\frac{L}{2} - \frac{1}{2\alpha})\|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\|_F^2. \tag{37}$$

Algorithm 1 projects the gradient step of Φ_i onto $\mathcal{B}_i$. This is exactly the minimizer of the proximal quadratic model in which $\delta_{\mathcal{B}_i}(\mathbf{B})$ is an indicator function

$$\mathbf{B}_i^{(k+1)} = \arg\min_{\mathbf{B}} \left\{ \langle \nabla_{\mathbf{B}_i}\Phi_i(\mathbf{B}_i^{(k)}), \mathbf{B} - \mathbf{B}_i^{(k)} \rangle + \frac{1}{2\beta}\|\mathbf{B} - \mathbf{B}_i^{(k)}\|_F^2 + \delta_{\mathcal{B}_i}(\mathbf{B}) \right\}. \tag{38}$$

Since $\mathbf{B}_i^{(k+1)}$ minimizes this model compared to the feasible point $\mathbf{B}_i^{(k)}$, we have:

$$\langle \nabla_{\mathbf{B}}\Phi_i(\mathbf{B}_i^{(k)}; \mathbf{A}_i^{(k)}), \mathbf{B}_i^{(k+1)} - \mathbf{B}_i^{(k)} \rangle \leq \delta_{\mathcal{B}_i}(\mathbf{B}_i^{(k)}) - \delta_{\mathcal{B}_i}(\mathbf{B}_i^{(k+1)}) - \frac{1}{2\beta}\|\mathbf{B}_i^{(k+1)} - \mathbf{B}_i^{(k)}\|_F^2. \tag{39}$$

Applying the descent lemma to Φ_i (which is L_Φ-smooth by Lemma 4) between iterates k and $k+1$ results in

$$\Phi_i(\mathbf{B}_i^{(k+1)}) \leq \Phi_i(\mathbf{B}_i^{(k)}) + \langle \nabla_{\mathbf{B}_i}\Phi_i(\mathbf{B}_i^{(k)}), \mathbf{B}_i^{(k+1)} - \mathbf{B}_i^{(k)} \rangle + \frac{L_\Phi}{2}\|\mathbf{B}_i^{(k+1)} - \mathbf{B}_i^{(k)}\|_F^2 \tag{40}$$

Adding the above inequality to (39) yields:

$$\Phi_i(\mathbf{B}_i^{(k+1)}) + \delta_{\mathcal{B}_i}(\mathbf{B}_i^{(k+1)}) \leq \Phi_i(\mathbf{B}_i^{(k)}) + \delta_{\mathcal{B}_i}(\mathbf{B}_i^{(k)}) - \left(\frac{1}{2\beta} - \frac{L_\Phi}{2}\right)\|\mathbf{B}_i^{(k+1)} - \mathbf{B}_i^{(k)}\|_F^2. \tag{41}$$

By definition (9), $\Phi_i(\mathbf{B}_i^{(k+1)}) = \mathcal{L}_i(\mathbf{A}_i^{(k+1)}, \mathbf{B}_i^{(k+1)})$, and $\Phi_i(\mathbf{B}_i^{(k)}) = \mathcal{L}_i(\tilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)})$. Substituting these and adding (37) into (41) yields:

$$F_i(\mathbf{A}_i^{(k+1)}, \mathbf{B}_i^{(k+1)}) \leq F_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}) - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\|_F^2$$
$$- \left(\frac{1}{2\beta} - \frac{L_\Phi}{2}\right)\|\mathbf{B}_i^{(k+1)} - \mathbf{B}_i^{(k)}\|_F^2. \tag{42}$$

where $F_i(\mathbf{A}_i, \mathbf{B}_i) = \mathcal{L}_i(\mathbf{A}_i, \mathbf{B}_i) + \delta_{\mathcal{B}_i}(\mathbf{B}_i)$. Because $\alpha < \frac{1}{L}$ and $\beta < \frac{1}{L_\Phi}$, both coefficients are non-negative, resulting in a decreasing loss function. Substituting the tentative lower-level

update $\tilde{\mathbf{A}}_i^{(k)} = \mathbf{A}_i^{(k)} - \alpha \nabla_{\mathbf{A}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})$, using (14) and (17) upper-level update $\mathbf{B}_i^{(k+1)} = \mathcal{P}_{B_i}^{\perp}(\mathbf{B}_i^{(k)} - \beta \nabla_{\mathbf{B}_i} \Phi_i(\mathbf{B}_i^{(k)}))$ in the above inequality gives us

$$F_i(\mathbf{A}_i^{(k+1)}, \mathbf{B}_i^{(k+1)}) \leq F_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}) - \left(\frac{\alpha}{2} - \frac{\alpha^2 L}{2}\right) \|\nabla_{\mathbf{A}_i} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})\|_F^2 - \left(\frac{\beta}{2} - \frac{L_\Phi \beta^2}{2}\right) \|\mathcal{P}_{B_i}^{\perp} \nabla_{\mathbf{B}_i} \Phi_i(\mathbf{B}_i^{(k)})\|_F^2. \tag{43}$$

Maximizing $\left(\frac{\alpha}{2} - \frac{\alpha^2 L}{2}\right)$ and $\left(\frac{\beta}{2} - \frac{L_\Phi \beta^2}{2}\right)$ over $\alpha \in (0, \frac{1}{L})$ and $\beta \in (0, \frac{1}{L_\Phi})$, respectively, gives $\alpha^* = \frac{1}{2L}$ and $\beta^* = \frac{1}{2L_\Phi} \geq \frac{1}{2L(1+\alpha^* L)^2 + 2\alpha^* \rho M} = \frac{1}{\frac{9}{2}L + \frac{\rho M}{L}}$.

Summing (42) over $k = 0$ to ∞ reveals that the sum of the squared differences is bounded by the initial suboptimality $F_i^{(0)} - F_i^{\infty}$. Therefore, we obtain the limits:

$$\lim_{k \to \infty} \|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\|_F = 0 \quad \text{and} \quad \lim_{k \to \infty} \|\mathbf{B}_i^{(k+1)} - \mathbf{B}_i^{(k)}\|_F = 0. \tag{44}$$

We now prove that the convergence of the algorithm implies the stationary conditions. First, for $\mathbf{A}_i$, recall that $\mathbf{A}_i^{(k)} - \tilde{\mathbf{A}}_i^{(k)} = \alpha \nabla_{\mathbf{A}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})$. From (44), it trivially follows that:

$$\lim_{k \to \infty} \|\nabla_{\mathbf{A}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})\|_F = 0, \tag{45}$$

satisfying the critical point condition.

Second, for $\mathbf{B}_i$, using (38) the projected gradient mapping of the reduced objective vanishes, meaning:

$$-\nabla_{\mathbf{B}} \Phi_i(\mathbf{B}_i^{(k)}; \mathbf{A}_i^{(k)}) \in \partial \delta_{\mathcal{B}_i}(\mathbf{B}_i^{(k)}) \quad \text{as } k \to \infty. \tag{46}$$

where $\partial \delta_{\mathcal{B}_i}(\mathbf{B}_i^{(k)})$ is the sub-gradient of the indicator $\delta_{\mathcal{B}_i}(\mathbf{B})$ at $\mathbf{B}_i^{(k)}$. We want to show that $\nabla_{\mathbf{B}} \Phi_i$ converges to the partial gradient of the true loss $\nabla_{\mathbf{B}} \mathcal{L}_i$. Using the chain rule from Lemma 4:

$$\nabla_{\mathbf{B}} \Phi_i(\mathbf{B}_i^{(k)}) = \nabla_{\mathbf{B}} \mathcal{L}_i(\tilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)}) - \alpha \nabla^2_{\mathbf{AB}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}) \big[\nabla_{\mathbf{A}} \mathcal{L}_i(\tilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)})\big]. \tag{47}$$

We analyze the discrepancy $\mathbf{D}^{(k)} = \nabla_{\mathbf{B}} \Phi_i(\mathbf{B}_i^{(k)}; \mathbf{A}_i^{(k)}) - \nabla_{\mathbf{B}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})$:

$$\|\mathbf{D}^{(k)}\|_F \leq \|\nabla_{\mathbf{B}} \mathcal{L}_i(\tilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)}) - \nabla_{\mathbf{B}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})\|_F + \alpha \|\nabla^2_{\mathbf{AB}} \mathcal{L}_i\|_2 \|\nabla_{\mathbf{A}} \mathcal{L}_i(\tilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)})\|_F. \tag{48}$$

By the joint L-smoothness of $\mathcal{L}_i$, the first term is bounded by $L\|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\|_F$. The cross-Hessian norm is bounded by L. Furthermore, the gradient evaluates to:

$$\|\nabla_{\mathbf{A}} \mathcal{L}_i(\tilde{\mathbf{A}}_i^{(k)}, \mathbf{B}_i^{(k)})\|_F \leq \|\nabla_{\mathbf{A}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)})\|_F + L\|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\|_F. \tag{49}$$

Since $\alpha \nabla_{\mathbf{A}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}) = \mathbf{A}_i^{(k)} - \tilde{\mathbf{A}}_i^{(k)}$, every component of the discrepancy is bounded by a constant multiple of $\|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\|_F$. Because $\|\tilde{\mathbf{A}}_i^{(k)} - \mathbf{A}_i^{(k)}\|_F \to 0$ as established in (44), we conclude that:

$$\lim_{k \to \infty} \|\mathbf{D}^{(k)}\|_F = 0 \implies \lim_{k \to \infty} \nabla_{\mathbf{B}} \Phi_i(\mathbf{B}_i^{(k)}; \mathbf{A}_i^{(k)}) = \lim_{k \to \infty} \nabla_{\mathbf{B}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}). \tag{50}$$

Therefore, substituting this equivalence into (46), the algorithm perfectly satisfies the original composite stationarity condition:

$$-\nabla_{\mathbf{B}} \mathcal{L}_i(\mathbf{A}_i^{(k)}, \mathbf{B}_i^{(k)}) \in \partial \delta_{\mathcal{B}_i}(\mathbf{B}_i^{(k)}) \quad \text{as } k \to \infty, \tag{51}$$

Concluding the proof that the sequence converges to a joint critical point of the main composite loss function F_i. □

C Proof of Theorem 2

The proof of Theorem 2 has two parts. *Part 1 (Subsection C.1)* establishes the high-probability bound (19): Lemma 5 bounds the per-layer crosstalk $\|\mathbf{\Delta}_j^{(\ell)}\|_F$ via Hanson–Wright, after Proposition 1 reduces it to the residual subspace using orthogonality; Lemma 6 propagates these local bounds through the L Lipschitz layers via a linear recursion; a union bound over $\ell \in [L]$ yields the network-wide statement. *Part 2 (Subsection C.2)* establishes basis optimality: Theorem 7 shows the top-r eigenspace of the compressed covariance $\widetilde{\mathbf{\Sigma}}_j^{(\ell)}$ minimizes residual energy among feasible rank-r projectors, and Proposition 2 shows random frozen bases are strictly suboptimal except in the isotropic case.

C.1 Analysis of Catastrophic Forgetting

To establish the end-to-end bound on catastrophic forgetting, we systematically decouple the local statistical interference from its global deterministic propagation through the deep network. We structure the proof into two complementary lemmas. First, we leverage the Hanson-Wright inequality to provide a high-probability statistical bound on the isolated activation crosstalk at any single layer. Second, the network-wide error propagation is analyzed, unrolling the Lipschitz-continuous inter-layer mappings to bound the total catastrophic forgetting as a weighted sum of these local perturbations. Finally, we synthesize these two results using a union bound across all L layers to complete the theorem.

Lemma 5 (Hanson-Wright High-Probability Bound on Crosstalk)**.** *Under Assumption 2 item 2, let* $\mathbf{C}_j^{(\ell)} = \sum_{k=j+1}^{T} \mathbf{A}_k^{(\ell)} (\mathbf{B}_k^{(\ell)})^\top$ *be the crosstalk operator for concept* j *and let the crosstalk be* $\mathbf{\Delta}_j^{(\ell)} = \mathbf{C}_j^{(\ell)} \mathbf{X}_j^{(\ell)}$*. There exists an absolute universal constant* $c > 0$ *such that for any* $\xi \in (0,1)$*, the crosstalk with probability at least* $1 - \xi$ *is bounded by*

$$\|\mathbf{\Delta}_j^{(\ell)}\|_F \leq \sqrt{Tr(\mathbf{\Psi}_j^{(\ell)})\, \|\mathbf{Q}_j^{(\ell)}\|_F^2 + t^{(\ell)}}, \tag{52}$$

where

$$t^{(\ell)} = \max\left(\sqrt{C_1} K^2 \|\mathbf{\Psi}_j^{(\ell)}\|_F \|(\mathbf{Q}_j^{(\ell)})^\top \mathbf{Q}_j^{(\ell)}\|_F \sqrt{\log \frac{2}{\xi}},\ C_1 K^2 \|\mathbf{\Psi}_j^{(\ell)}\|_2 \|\mathbf{Q}_j^{(\ell)}\|_2^2 \log \frac{2}{\xi} \right), \tag{53}$$

$C_1 = 1/c$*, and* $\mathbf{Q}_j^{(\ell)} \triangleq \mathbf{C}_j^{(\ell)} (\mathbf{\Sigma}_j^{(\ell),\perp})^{1/2} \in \mathbb{R}^{n_\ell \times m_\ell}$.

Proof. Consider a model with L layers adapted for T concepts via SeqLoRA. At layer ℓ, the pre-trained weight is $\mathbf{W}_0^{(\ell)} \in \mathbb{R}^{n_\ell \times m_\ell}$ and the LoRA factors for concept k are $\mathbf{A}_k^{(\ell)} \in \mathbb{R}^{n_\ell \times r}$ and $\mathbf{B}_k^{(\ell)} \in \mathbb{R}^{m_\ell \times r}$.

In the sequential continual learning setting, the accumulated weight matrix after learning concept j at layer ℓ is:

$$\mathbf{W}_j^{(\ell)} = \mathbf{W}_0^{(\ell)} + \sum_{k=1}^{j} \mathbf{A}_k^{(\ell)} (\mathbf{B}_k^{(\ell)})^\top. \tag{54}$$

The composed output for concept j is:

$$O_j(\mathbf{X}_j^{(\ell)}) = \mathbf{W}_j^{(\ell)} \mathbf{X}_j^{(\ell)} \tag{55}$$

while the composed output (with all T concepts) is:

$$\tilde{O}_T(\mathbf{X}_j^{(\ell)}) = \mathbf{W}_T^{(\ell)} \mathbf{X}_j^{(\ell)} = O_j(\mathbf{X}_j^{(\ell)}) + \sum_{k=j+1}^{T} \mathbf{A}_k^{(\ell)} (\mathbf{B}_k^{(\ell)})^\top \mathbf{X}_j^{(\ell)}. \tag{56}$$

The total crosstalk at layer ℓ for concept j is:

$$\mathbf{\Delta}_j^{(\ell)} \triangleq \tilde{O}_T(\mathbf{X}_j^{(\ell)}) - O_j(\mathbf{X}_j^{(\ell)}) = \mathbf{C}_j^{(\ell)} \mathbf{X}_j^{(\ell)} \in \mathbb{R}^{n_\ell \times p}, \tag{57}$$

where the crosstalk operator is:

$$\mathbf{C}_j^{(\ell)} \triangleq \sum_{k=j+1}^{T} \mathbf{A}_k^{(\ell)} (\mathbf{B}_k^{(\ell)})^\top \in \mathbb{R}^{n_\ell \times m_\ell}. \tag{58}$$

This captures the interference of the parameters added after concept j acting on concept j's activations. We decompose the activation into its component within concept j's basis and its residual:

$$\mathbf{X}_j^{(\ell)} = \mathcal{P}_{B_j}^{(\ell)} \mathbf{X}_j^{(\ell)} + \mathcal{P}_{B_j}^{(\ell),\perp} \mathbf{X}_j^{(\ell)}. \tag{59}$$

where $\mathcal{P}_{B_j}^{(\ell)} = \mathbf{B}_j^{(\ell)} \left((\mathbf{B}_j^{(\ell)})^\top \mathbf{B}_j^{(\ell)} \right)^{-1} (\mathbf{B}_j^{(\ell)})^\top$ is the projector onto $\mathrm{col}(\mathbf{B}_j^{(\ell)})$ and its complement is $\mathcal{P}_{B_j}^{(\ell),\perp} = \mathbf{I}_{m^{(\ell)}} - \mathcal{P}_{B_j}^{(\ell)}$. Since $\mathbf{B}_j^{(\ell)}$ is the learned LoRA basis for concept j, it captures the relevant features of $\mathbf{X}_j^{(\ell)}$. The first term carries most of the activation energy; the second term is the leaked energy outside the learned subspace.

Proposition 1. *The crosstalk depends only on the residual component:*

$$\mathbf{\Delta}_j^{(\ell)} = \mathbf{C}_j^{(\ell)} \mathbf{X}_j^{(\ell)} = \mathbf{C}_j^{(\ell)} \mathcal{P}_{B_j}^{(\ell),\perp} \mathbf{X}_j^{(\ell)} = \mathbf{C}_j^{(\ell)} \mathbf{X}_j^{(\ell),\perp}. \tag{60}$$

That is, the signal component $\mathcal{P}_{B_j}^{(\ell)} \mathbf{X}_j^{(\ell)}$ is invisible to the crosstalk.

Proof. Substituting the decomposition (59):

$$\mathbf{C}_j^{(\ell)} \mathbf{X}_j^{(\ell)} = \mathbf{C}_j^{(\ell)} \mathcal{P}_{B_j}^{(\ell)} \mathbf{X}_j^{(\ell)} + \mathbf{C}_j^{(\ell)} \mathcal{P}_{B_j}^{(\ell),\perp} \mathbf{X}_j^{(\ell)}. \tag{61}$$

We show the first term vanishes. Expanding:

$$\mathbf{C}_j^{(\ell)} \mathcal{P}_{B_j}^{(\ell)} = \sum_{k=j+1}^{T} \mathbf{A}_k^{(\ell)} (\mathbf{B}_k^{(\ell)})^\top \mathcal{P}_{B_j}^{(\ell)}. \tag{62}$$

Now $(\mathbf{B}_k^{(\ell)})^\top \mathcal{P}_{B_j}^{(\ell)} = (\mathbf{B}_k^{(\ell)})^\top \mathbf{B}_j^{(\ell)} (\mathbf{G}_j^{(\ell)})^{-1} (\mathbf{B}_j^{(\ell)})^\top$, where $\mathbf{G}_k^{(\ell)} = (\mathbf{B}_k^{(\ell)})^\top \mathbf{B}_k^{(\ell)}$. Since $(\mathbf{B}_k^{(\ell)})^\top \mathbf{B}_j^{(\ell)} = 0$ for $k \neq j$:

$$\mathbf{C}_j^{(\ell)} \mathcal{P}_{B_j}^{(\ell)} = 0. \tag{63}$$

Therefore $\mathbf{\Delta}_j^{(\ell)} = \mathbf{C}_j^{(\ell)} \mathcal{P}_{B_j}^{(\ell),\perp} \mathbf{X}_j^{(\ell)} = \mathbf{C}_j^{(\ell)} \mathbf{X}_j^{(\ell),\perp}$. □

This result implies that the entire activation energy within $\mathrm{col}(\mathbf{B}_j^{(\ell)})$, regardless of its magnitude, does not contribute to the crosstalk, and the crosstalk sees only the residual

The residual activation covariance $\mathbf{x}_j^{(\ell),\perp} = \mathrm{vec}\left(\mathbf{X}_j^{(\ell),\perp} \right)$ is defined as follows

$$\mathrm{Cov}(\mathbf{x}_j^{(\ell),\perp}) \triangleq \mathbb{E}\left\{ \mathbf{x}_j^{(\ell),\perp} (\mathbf{x}_j^{(\ell),\perp})^\top \right\} = \mathbf{\Psi}_j^{(\ell)} \otimes \mathbf{\Sigma}_j^{(\ell),\perp} \in \mathbb{R}^{m_\ell p}. \tag{64}$$

It is important to note that since $\mathbf{B}_j^{(\ell)}$ captures most of the activation energy of concept j, the residual covariance $\mathbf{\Sigma}_j^{(\ell),\perp}$ may have a small energy (small eigenvalues).

Substituting the matrix sub-Gaussian process into the crosstalk (60), we obtain:

$$\mathbf{\Delta}_j^{(\ell)} = \mathbf{C}_j^{(\ell)} \mathbf{X}_j^{(\ell),\perp} = \mathbf{C}_j^{(\ell)} (\mathbf{\Sigma}_j^{(\ell),\perp})^{1/2} \mathbf{Z}^{(\ell)} (\mathbf{\Psi}_j^{(\ell)})^{1/2} = \mathbf{Q}_j^{(\ell)} \mathbf{Z}^{(\ell)} (\mathbf{\Psi}_j)^{1/2}, \tag{65}$$

where $\mathbf{Q}_j^{(\ell)} \triangleq \mathbf{C}_j^{(\ell)} (\mathbf{\Sigma}_j^{(\ell),\perp})^{1/2} \in \mathbb{R}^{n_\ell \times m_\ell}$. The squared Frobenius norm of the crosstalk is a quadratic form:

$$\begin{aligned} z \triangleq \|\mathbf{\Delta}_j^{(\ell)}\|_F^2 &= \mathrm{Tr}\left((\mathbf{Q}_j^{(\ell)} \mathbf{Z}^{(\ell)} (\mathbf{\Psi}_j^{(\ell)})^{1/2})^\top (\mathbf{Q}_j^{(\ell)} \mathbf{Z}^{(\ell)} (\mathbf{\Psi}_j^{(\ell)})^{1/2}) \right) \\ &= \mathrm{Tr}\left((\mathbf{Z}^{(\ell)})^\top \left((\mathbf{Q}_j^{(\ell)})^\top \mathbf{Q}_j^{(\ell)} \right) \mathbf{Z}^{(\ell)} \mathbf{\Psi}_j^{(\ell)} \right). \end{aligned} \tag{66}$$

By using the standard matrix trace identity $\text{Tr}(\mathbf{A}^\top\mathbf{BAC}) = \text{vec}(\mathbf{A})^\top(\mathbf{C}\otimes\mathbf{B})\,\text{vec}(\mathbf{A})$, we vectorize the independent sub-Gaussian matrix $\mathbf{Z}^{(\ell)}$ into $\mathbf{z}^{(\ell)} \in \mathbb{R}^{m_\ell p}$. The trace perfectly rewrites as a standard vector quadratic form:

$$z = (\mathbf{z}^{(\ell)})^\top \mathbf{H}\mathbf{z}^{(\ell)}, \tag{67}$$

where $\mathbf{H} = \mathbf{\Psi}_j^{(\ell)} \otimes \left((\mathbf{Q}_j^{(\ell)})^\top \mathbf{Q}_j^{(\ell)}\right)$. Taking the expectation of the variable z and using $\text{Tr}(\mathbf{A}\otimes\mathbf{B}) = \text{Tr}(\mathbf{A})\text{Tr}(\mathbf{B})$ (Horn and Johnson, 1991)

$$\begin{aligned}\mu_z = \mathbb{E}\{(\mathbf{z}^{(\ell)})^\top \mathbf{H}\mathbf{z}^{(\ell)}\} = \text{Tr}(\mathbb{E}\{\mathbf{z}^{(\ell)}(\mathbf{z}^{(\ell)})^\top\}\mathbf{H}) = \text{Tr}(\mathbf{H}) = \\ \text{Tr}(\mathbf{\Psi}_j^{(\ell)})\text{Tr}\left(\left((\mathbf{Q}_j)^{(\ell)}\right)^\top \mathbf{Q}_j^{(\ell)}\right) = \text{Tr}(\mathbf{\Psi}_j^{(\ell)})\|\mathbf{Q}_j^{(\ell)}\|_F^2,\end{aligned} \tag{68}$$

where, based on the unit variance sub-Gaussian assumption, we have $\mathbb{E}\{\mathbf{z}^{(\ell)}(\mathbf{z}^{(\ell)})^\top\} = \mathbf{I}$. By the Hanson-Wright inequality for sub-Gaussian random vectors (Vershynin, 2026, Theorem 6.2.2), for any $t > 0$:

$$\mathbb{P}(|z - \mu_z| > t) \le 2\exp\left[-c\min\left(\frac{t^2}{K^4\|\mathbf{H}\|_F^2}, \frac{t}{K^2\|\mathbf{H}\|_2}\right)\right], \tag{69}$$

where $c > 0$ is a universal constant. We now bound the norms of $\mathbf{H}$.

For the ℓ_2 norm and using $\|\mathbf{A}\otimes\mathbf{B}\|_2 = \|\mathbf{A}\|_2\|\mathbf{B}\|_2$ (Horn and Johnson, 1991), we have:

$$\|\mathbf{H}\|_2 = \|\mathbf{\Psi}_j^{(\ell)} \otimes \left((\mathbf{Q}_j^{(\ell)})^\top \mathbf{Q}_j^{(\ell)}\right)\|_2 = \|\mathbf{\Psi}_j^{(\ell)}\|_2\|(\mathbf{Q}_j^{(\ell)})^\top \mathbf{Q}_j^{(\ell)}\|_2 = \|\mathbf{\Psi}_j^{(\ell)}\|_2\|\mathbf{Q}_j^{(\ell)}\|_2^2. \tag{70}$$

To compute the squared Frobenius norm $\|\mathbf{H}\|_F^2$, we use the property $\|\mathbf{A}\otimes\mathbf{B}\|_F = \|\mathbf{A}\|_F\|\mathbf{B}\|_F$ (Horn and Johnson, 1991). Consequently,

$$\|\mathbf{H}\|_F = \|\mathbf{\Psi}_j^{(\ell)} \otimes \left((\mathbf{Q}_j^{(\ell)})^\top \mathbf{Q}_j^{(\ell)}\right)\|_F = \|\mathbf{\Psi}_j^{(\ell)}\|_F\|(\mathbf{Q}_j^{(\ell)})^\top \mathbf{Q}_j^{(\ell)}\|_F. \tag{71}$$

To derive the exact minimal t that guarantees a failure probability of at most ξ, we set the right-hand side equal to ξ:

$$\min\left(\frac{t^2}{K^4\|\mathbf{H}\|_F^2}, \frac{t}{K^2\|\mathbf{H}\|_2}\right) = \frac{1}{c}\log\frac{2}{\xi}. \tag{72}$$

For the minimum of two non-negative terms to equal a value x, both terms must be at least x. This yields two lower bounds for t:

$$\frac{t^2}{K^4\|\mathbf{H}\|_F^2} \ge \frac{1}{c}\log\frac{2}{\xi} \quad\Longrightarrow\quad t \ge \frac{K^2\|\mathbf{H}\|_F}{\sqrt{c}}\sqrt{\log\frac{2}{\xi}}, \tag{73}$$

$$\frac{t}{K^2\|\mathbf{H}\|_2} \ge \frac{1}{c}\log\frac{2}{\xi} \quad\Longrightarrow\quad t \ge \frac{K^2\|\mathbf{H}\|_2}{c}\log\frac{2}{\xi}. \tag{74}$$

To tightly satisfy both conditions, t must be exactly the maximum of the two bounds. Using (70) and (71) in the above inequalities, we have

$$t^{(\ell)} = \max\left(\sqrt{C_1}K^2\|\mathbf{\Psi}_j^{(\ell)}\|_F\|(\mathbf{Q}_j^{(\ell)})^\top \mathbf{Q}_j^{(\ell)}\|_F\sqrt{\log\frac{2}{\xi}},\ C_1K^2\|\mathbf{\Psi}_j^{(\ell)}\|_2\|\mathbf{Q}_j^{(\ell)}\|_2^2\log\frac{2}{\xi}\right), \tag{75}$$

where $C_1 = 1/c$. Using (68) and the above equation in the Hanson-Wright inequality results in the following crosstalk upper bound with probability $1-\xi$

$$\|\mathbf{\Delta}_j^{(\ell)}\|_F \le \sqrt{\text{Tr}(\mathbf{\Psi}_j^{(\ell)})\,\|\mathbf{Q}_j^{(\ell)}\|_F^2 + t^{(\ell)}}. \tag{76}$$

This concludes the proof of Lemma 5. □

Lemma 6 (End-to-End Catastrophic Forgetting)**.** *Under Assumptions 2 item 1, let* $\Gamma_\ell \triangleq \prod_{\ell'=\ell+1}^{L}\gamma_{\ell'}\|\mathbf{W}_T^{(\ell')}\|_2$ *be the downstream amplification factor at layer* ℓ*. For any* $\xi \in (0,1)$*, the end-to-end catastrophic forgetting satisfies, with probability at least* $1-\xi$*:*

$$\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_j) \le L_o\sum_{\ell=1}^{L}\Gamma_\ell\|\mathbf{\Delta}_j^{(\ell)}\|_F, \tag{77}$$

where $\theta_T = \{\mathbf{W}_T^{(\ell)}\}_{\ell=1}^L$ *and* $\theta_j = \{\mathbf{W}_j^{(\ell)}\}_{\ell=1}^L$ *are the full model after learning concept* j *and* T*, respectively.*

Proof. Let $O_j(\mathbf{X}_j^{(\ell)}) = \mathbf{W}_j^{(\ell)}\mathbf{X}_j^{(\ell)}$ and $\tilde{O}_T(\tilde{\mathbf{X}}_j^{(\ell)}) = \mathbf{W}_T^{(\ell)}\tilde{\mathbf{X}}_j^{(\ell)}$ denote the layer-ℓ inputs under model's parameters $\mathbf{W}_j^{(\ell)}$ and the composed model's parameters T concepts ($\mathbf{W}_T^{(\ell)}$), respectively. Defining the layer-wise perturbation as $e_\ell \triangleq \|O_j(\mathbf{X}_j^{(\ell)}) - \tilde{O}_T(\tilde{\mathbf{X}}_j^{(\ell)})\|_F$, we have $e_0 = 0$ at the network input.

For $\ell \geq 1$, the inputs to the LoRA layer are $\mathbf{X}_j^{(\ell+1)} = \phi_\ell(O_j(\mathbf{X}_j^{(\ell)}))$ and $\hat{\mathbf{X}}_j^{(\ell+1)} = \phi_\ell(\tilde{O}_T(\tilde{\mathbf{X}}_j^{(\ell)}))$. The output difference expands as:

$$\tilde{O}_T(\tilde{\mathbf{X}}_j^{(\ell)}) - O_j(\mathbf{X}_j^{(\ell)}) = \mathbf{W}_T^{(\ell)}\tilde{\mathbf{X}}_j^{(\ell)} - \mathbf{W}_j^{(\ell)}\mathbf{X}_j^{(\ell)} = \mathbf{W}_T^{(\ell)}\big(\tilde{\mathbf{X}}_j^{(\ell)} - \mathbf{X}_j^{(\ell)}\big) + \mathbf{C}_j^{(\ell)}\mathbf{X}_j^{(\ell)}. \tag{78}$$

Taking the Frobenius norm, applying the triangle inequality, and utilizing the γ_ℓ-Lipschitz property of the inter-layer mapping ϕ_ℓ yields the linear recursion:

$$e_\ell \leq \|\mathbf{W}_T^{(\ell)}\|_2\|\tilde{\mathbf{X}}_j^{(\ell)} - \mathbf{X}_j^{(\ell)}\|_F + \|\mathbf{C}_j^{(\ell)}\mathbf{X}_j^{(\ell)}\|_F \leq \alpha_\ell e_{\ell-1} + \|\mathbf{\Delta}_j^{(\ell)}\|_F, \tag{79}$$

where we define $\alpha_\ell \triangleq \gamma_\ell\|\mathbf{W}_T^{(\ell)}\|_2$ as the effective per-layer amplification and $\mathbf{\Delta}_j^{(\ell)} \triangleq \mathbf{C}_j^{(\ell)}\mathbf{X}_j^{(\ell)}$ as the injected crosstalk.

Unrolling the linear recursion (79) entirely from $\ell = 1$ to L, and recognizing that the cumulative amplification from layer ℓ to L is exactly $\Gamma_\ell = \prod_{\ell'=\ell+1}^{L}\alpha_{\ell'}$, we have the following bound

$$e_L \leq \sum_{\ell=1}^{L}\Gamma_\ell\|\mathbf{\Delta}_j^{(\ell)}\|_F. \tag{80}$$

Finally, by the L_o-Lipschitz continuity of the downstream mapping f_j, the end-to-end catastrophic forgetting is bounded directly by the final-layer perturbation:

$$\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_j) = f_j(\tilde{O}_T(\hat{\mathbf{X}}_j^{(L)})) - f_j(O_j(\hat{\mathbf{X}}_j^{(L)})) \leq L_o\|\tilde{O}_T(\hat{\mathbf{X}}_j^{(L)}) - O_j(\hat{\mathbf{X}}_j^{(L)})\|_F = L_o e_L. \tag{81}$$

Using (80) in the above inequality results in

$$\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_j) \leq L_o\sum_{\ell=1}^{L}\Gamma_\ell\|\mathbf{\Delta}_j^{(\ell)}\|_F. \tag{82}$$

□

We now synthesize the results of Lemma 5 and Lemma 6 to complete the proof of the bound (19) in Theorem 2. To ensure a global probabilistic guarantee across the entire network, we apply the statistical bound from Lemma 5 with an adjusted layer-wise failure probability of ξ/L. Substituting this high-probability upper bound (52) for the crosstalk norm $\|\mathbf{\Delta}_j^{(\ell)}\|_F$ into the end-to-end deterministic bound (77), we apply Boole's inequality (the union bound) over all $\ell \in [L]$. This yields to the inequality (19). This concludes the proof of catastrophic forgetting bound (19) in Theorem 2.

In Theorem 2, due to the residual model $\mathbf{X}_j^{(\ell),\perp} = (\mathbf{\Sigma}_j^{(\ell),\perp})^{1/2}\mathbf{Z}^{(\ell)}(\mathbf{\Psi}_j^{(\ell)})^{1/2}$ in Assumption 2, the trace of the covariance matrix of residual is given by

$$\begin{aligned}
\mathrm{Tr}\Big(\mathbb{E}\{(\mathbf{X}_j^{(\ell),\perp})^\top\mathbf{X}_j^{(\ell),\perp}\}\Big) &= \mathrm{Tr}\Big(\mathbb{E}\{(\mathbf{\Psi}_j^{(\ell)})^{1/2}(\mathbf{Z}^{(\ell)})^\top\mathbf{\Sigma}_j^{(\ell),\perp}\mathbf{Z}^{(\ell)}(\mathbf{\Psi}_j^{(\ell)})^{1/2}\}\Big) \\
&= \mathrm{Tr}\left(\mathbf{\Psi}_j^{(\ell)}\mathbb{E}\{(\mathbf{Z}^{(\ell)})^\top\mathbf{\Sigma}_j^{(\ell),\perp}\mathbf{Z}^{(\ell)}\}\right) \\
&= \mathrm{Tr}\left(\mathbf{\Psi}_j^{(\ell)} \cdot \mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp})\mathbf{I}_p\right) \\
&= \mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp})\mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)}).
\end{aligned}$$

Also, we know $\mathbf{X}_j^{(\ell),\perp} = \mathcal{P}_{B_j}^{(\ell),\perp}\mathbf{X}_j^{(\ell)}$, $\mathbf{\Sigma}_j^{(\ell)} = \mathbb{E}\{\mathbf{X}_j^{(\ell)}(\mathbf{X}_j^{(\ell)})^\top\}$ and using the above equation results in

$$\mathrm{Tr}(\mathbb{E}\{\mathbf{X}_j^{(\ell),\perp}(\mathbf{X}_j^{(\ell),\perp})^\top\}) = \mathrm{Tr}(\mathcal{P}_{B_j}^{(\ell),\perp}\mathbf{\Sigma}_j^{(\ell)}) = \mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp})\mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)}).$$

The term $\mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp})\mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})$ governs the expected crosstalk energy within the end-to-end forgetting bound because

$$\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_j) \le L_o \sum_{\ell=1}^{L} \Gamma_\ell \sqrt{\mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})\, \|\mathbf{Q}_j^{(\ell)}\|_F^2 + t^{(\ell)}}, \tag{83}$$

identifying the residual covariance mass as the primary driver of catastrophic forgetting. To prove the rate (19), we write $\epsilon_\ell = \mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})\mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp})$, and $\kappa_\ell = \|\mathbf{C}_j^{(\ell)}\|_2$. First,

$$\|\mathbf{Q}_j^{(\ell)}\|_F^2 = \mathrm{Tr}\big(\mathbf{C}_j^{(\ell)}\mathbf{\Sigma}_j^{(\ell),\perp}(\mathbf{C}_j^{(\ell)})^\top\big) \le \|\mathbf{C}_j^{(\ell)}\|_2^2\,\mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp}) = \kappa_\ell^2\,\mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp}), \tag{84}$$

using $\mathrm{Tr}(\mathbf{C}\mathbf{M}\mathbf{C}^\top) \le \|\mathbf{C}\|_2^2\mathrm{Tr}(\mathbf{M})$ for positive semidefinite $\mathbf{M}$. Hence the expected crosstalk energy satisfies $\mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})\|\mathbf{Q}_j^{(\ell)}\|_F^2 \le \kappa_\ell^2\epsilon_\ell$. Second, for positive semidefinite matrices, the Frobenius and operator norms are bounded by the trace, so $\|\mathbf{\Psi}_j^{(\ell)}\|_F \le \mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})$, $\|\mathbf{\Psi}_j^{(\ell)}\|_2 \le \mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})$, $\|(\mathbf{Q}_j^{(\ell)})^\top\mathbf{Q}_j^{(\ell)}\|_F \le \mathrm{Tr}((\mathbf{Q}_j^{(\ell)})^\top\mathbf{Q}_j^{(\ell)}) = \|\mathbf{Q}\|_F^2 \le \kappa_\ell^2\,\mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp})$ and $\|\mathbf{Q}_j^{(\ell)}\|_2^2 \le \|\mathbf{Q}_j^{(\ell)}\|_F^2 \le \kappa_\ell^2\,\mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp})$. Substituting into (53),

$$t^{(\ell)} \le \kappa_\ell^2\epsilon_\ell \cdot \max\left(\sqrt{C_1}K^2\sqrt{\log\tfrac{2L}{\xi}},\; C_1K^2\log\tfrac{2L}{\xi}\right) = \kappa_\ell^2\epsilon_\ell\, c_{K,\xi}. \tag{85}$$

Therefore $\sqrt{\mathrm{Tr}(\mathbf{\Psi})\|\mathbf{Q}\|_F^2 + t^{(\ell)}} \le \kappa_\ell\sqrt{\epsilon_\ell}\sqrt{1+c_{K,\xi}}$, and summing over layers gives (19).

This leads to a fundamental optimization question: among all feasible orthogonal rank-r subspaces, which one minimizes this residual energy? In the next subsection, we address this question, which proves the second part of Theorem 2 regarding minimizing the energy of interference.

C.2 Optimal basis selection inside the free complement

Let

$$\mathbf{B}_{\mathrm{int}}^{(\ell)} \triangleq [\mathbf{B}_1^{(\ell)}, \dots, \mathbf{B}_{j-1}^{(\ell)}], \qquad \mathcal{P}_{<j}^{(\ell)} \triangleq \mathbf{B}_{\mathrm{int}}^{(\ell)}\left((\mathbf{B}_{\mathrm{int}}^{(\ell)})^\top\mathbf{B}_{\mathrm{int}}^{(\ell)}\right)^{-1}(\mathbf{B}_{\mathrm{int}}^{(\ell)})^\top, \tag{86}$$

and define the currently available orthogonal complement

$$\mathcal{P}_{\mathrm{free},j}^{(\ell)} \triangleq \mathbf{I} - \mathcal{P}_{<j}^{(\ell)}. \tag{87}$$

A rank-r projector $\mathcal{P}^{(\ell)}$ is *feasible* for concept j if

$$\mathcal{P}^{(\ell)} = (\mathcal{P}^{(\ell)})^\top = (\mathcal{P}^{(\ell)})^2, \qquad \mathrm{rank}(\mathcal{P}^{(\ell)}) = r, \qquad \mathrm{range}(\mathcal{P}^{(\ell)}) \subseteq \mathrm{range}(\mathcal{P}_{\mathrm{free},j}^{(\ell)}). \tag{88}$$

Equivalently, $\mathcal{P}^{(\ell)}$ is any rank-r subspace that obeys the orthogonality constraints with all previously learned concepts. For any projector $\mathcal{P}^{(\ell)}$, write $\mathcal{P}^{(\ell),\perp} \triangleq \mathbf{I} - \mathcal{P}^{(\ell)}$. Define the compressed covariance inside the currently free subspace by

$$\widetilde{\mathbf{\Sigma}}_j^{(\ell)} \triangleq \mathcal{P}_{\mathrm{free},j}^{(\ell)}\mathbf{\Sigma}_j^{(\ell)}\mathcal{P}_{\mathrm{free},j}^{(\ell)}, \tag{89}$$

and let $d_{\mathrm{free},j} \triangleq \mathrm{rank}(\mathcal{P}_{\mathrm{free},j}^{(\ell)})$. Denote the eigenvalues of $\widetilde{\mathbf{\Sigma}}_j^{(\ell)}$ by

$$\lambda_1(\widetilde{\mathbf{\Sigma}}_j^{(\ell)}) \ge \lambda_2(\widetilde{\mathbf{\Sigma}}_j^{(\ell)}) \ge \cdots \ge \lambda_{d_{\mathrm{free},j}}(\widetilde{\mathbf{\Sigma}}_j^{(\ell)}) \ge 0. \tag{90}$$

Theorem 7 (Optimal feasible basis captures maximal task energy)**.** *Assume $r \le d_{\mathrm{free},j}$. Then for any feasible rank-r projector $\mathcal{P}^{(\ell)}$,*

$$\mathit{Tr}(\mathcal{P}^{(\ell),\perp}\mathbf{\Sigma}_j^{(\ell)}) = \mathit{Tr}\Big((\mathbf{I} - \mathcal{P}_{\mathrm{free},j}^{(\ell)})\mathbf{\Sigma}_j^{(\ell)}\Big) + \mathit{Tr}\Big((\mathcal{P}_{\mathrm{free},j}^{(\ell)} - \mathcal{P}^{(\ell)})\widetilde{\mathbf{\Sigma}}_j^{(\ell)}\Big). \tag{91}$$

Hence minimizing the residual energy $\mathit{Tr}(\mathcal{P}^{(\ell),\perp}\mathbf{\Sigma}_j^{(\ell)})$ over all feasible rank-r projectors is equivalent to maximizing the captured energy $\mathit{Tr}(\mathcal{P}^{(\ell)}\widetilde{\mathbf{\Sigma}}_j^{(\ell)})$. The minimizer is the projector $\mathcal{P}_j^{(\ell)\star}$ onto the top-r eigenspace of $\widetilde{\mathbf{\Sigma}}_j^{(\ell)}$, and the minimum residual energy is

$$\mathit{Tr}\Big((\mathcal{P}_j^{(\ell)\star})^\perp\mathbf{\Sigma}_j^{(\ell)}\Big) = \mathit{Tr}\Big((\mathbf{I} - \mathcal{P}_{\mathrm{free},j}^{(\ell)})\mathbf{\Sigma}_j^{(\ell)}\Big) + \sum_{q=r+1}^{d_{\mathrm{free},j}} \lambda_q(\widetilde{\mathbf{\Sigma}}_j^{(\ell)}). \tag{92}$$

Proof. Because $\mathrm{range}(\mathcal{P}^{(\ell)}) \subseteq \mathrm{range}(\mathcal{P}^{(\ell)}_{\mathrm{free},j})$, we have

$$\mathcal{P}^{(\ell)}_{\mathrm{free},j}\mathcal{P}^{(\ell)} = \mathcal{P}^{(\ell)}\mathcal{P}^{(\ell)}_{\mathrm{free},j} = \mathcal{P}^{(\ell)}. \tag{93}$$

Therefore

$$\mathbf{I} - \mathcal{P}^{(\ell)} = (\mathbf{I} - \mathcal{P}^{(\ell)}_{\mathrm{free},j}) + (\mathcal{P}^{(\ell)}_{\mathrm{free},j} - \mathcal{P}^{(\ell)}), \tag{94}$$

and thus

$$\mathrm{Tr}(\mathcal{P}^{(\ell),\perp}\mathbf{\Sigma}^{(\ell)}_j) = \mathrm{Tr}\Big((\mathbf{I} - \mathcal{P}^{(\ell)}_{\mathrm{free},j})\mathbf{\Sigma}^{(\ell)}_j\Big) + \mathrm{Tr}\Big((\mathcal{P}^{(\ell)}_{\mathrm{free},j} - \mathcal{P})\mathbf{\Sigma}^{(\ell)}_j\Big). \tag{95}$$

Since $\mathcal{P}^{(\ell)}_{\mathrm{free},j} - \mathcal{P}$ acts entirely inside the free complement,

$$\mathrm{Tr}\Big((\mathcal{P}^{(\ell)}_{\mathrm{free},j} - \mathcal{P})\mathbf{\Sigma}^{(\ell)}_j\Big) = \mathrm{Tr}\Big((\mathcal{P}^{(\ell)}_{\mathrm{free},j} - \mathcal{P}^{(\ell)})\mathcal{P}^{(\ell)}_{\mathrm{free},j}\mathbf{\Sigma}^{(\ell)}_j\mathcal{P}^{(\ell)}_{\mathrm{free},j}\Big) = \mathrm{Tr}\left((\mathcal{P}^{(\ell)}_{\mathrm{free},j} - \mathcal{P})\widetilde{\mathbf{\Sigma}}^{(\ell)}_j\right). \tag{96}$$

This proves the decomposition.

Next, observe that

$$\mathrm{Tr}\left((\mathcal{P}^{(\ell)}_{\mathrm{free},j} - \mathcal{P}^{(\ell)})\widetilde{\mathbf{\Sigma}}^{(\ell)}_j\right) = \mathrm{Tr}(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) - \mathrm{Tr}(\mathcal{P}^{(\ell)}\widetilde{\mathbf{\Sigma}}^{(\ell)}_j). \tag{97}$$

Therefore minimizing $\mathrm{Tr}(\mathcal{P}^{(\ell),\perp}\mathbf{\Sigma}^{(\ell)}_j)$ is equivalent to maximizing $\mathrm{Tr}(\mathcal{P}^{(\ell)}\widetilde{\mathbf{\Sigma}}^{(\ell)}_j)$ over all feasible rank-r projectors. By the variational characterization of eigenvalues, the maximum is

$$\max_{\mathcal{P}^{(\ell)}} \mathrm{Tr}(\mathcal{P}^{(\ell)}\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) = \sum_{q=1}^{r} \lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j), \tag{98}$$

attained when $\mathcal{P}^{(\ell)} = \mathcal{P}^{(\ell)^\star}_j$ is the projector onto the top-r eigenspace of $\widetilde{\mathbf{\Sigma}}^{(\ell)}_j$. Substituting back yields

$$\mathrm{Tr}\Big((\mathcal{P}^{(\ell)^\star}_j)^{\perp}\mathbf{\Sigma}^{(\ell)}_j\Big) = \mathrm{Tr}\Big((\mathbf{I} - \mathcal{P}^{(\ell)}_{\mathrm{free},j})\mathbf{\Sigma}^{(\ell)}_j\Big) + \mathrm{Tr}(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) - \sum_{q=1}^{r} \lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) \tag{99}$$

$$= \mathrm{Tr}\Big((\mathbf{I} - \mathcal{P}^{(\ell)}_{\mathrm{free},j})\mathbf{\Sigma}^{(\ell)}_j\Big) + \sum_{q=r+1}^{d_{\mathrm{free},j}} \lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j). \tag{100}$$

□

Theorem 7 identifies the best possible basis for continual-learning: among all subspaces compatible with the orthogonality constraints, the optimal one is the dominant task covariance subspace inside the currently available complement. This immediately explains why learning $\mathbf{B}^{(\ell)}_j$ is beneficial: a frozen basis cannot adapt to the spectral structure of $\widetilde{\mathbf{\Sigma}}^{(\ell)}_j$.

The next result makes this comparison explicit for random frozen bases, which are a natural abstraction of methods that choose an orthogonal basis a priori and then keep it fixed.

Proposition 2 (Random frozen bases are optimal only in isotropic free complements)**.** *Let $\mathcal{P}_{\mathrm{rand}}$ be a Haar-random rank-r projector inside* $\mathrm{range}(\mathcal{P}^{(\ell)}_{\mathrm{free},j})$*. Then*

$$\mathbb{E}[\mathcal{P}_{\mathrm{rand}}] = \frac{r}{d_{\mathrm{free},j}}\mathcal{P}^{(\ell)}_{\mathrm{free},j}, \qquad \mathbb{E}\{Tr(\mathcal{P}_{\mathrm{rand}}\widetilde{\mathbf{\Sigma}}^{(\ell)}_j)\} = \frac{r}{d_{\mathrm{free},j}}Tr(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j), \tag{101}$$

and therefore

$$\mathbb{E}\{Tr(\mathcal{P}^{\perp}_{\mathrm{rand}}\mathbf{\Sigma}^{(\ell)}_j)\} = Tr\Big((\mathbf{I} - \mathcal{P}^{(\ell)}_{\mathrm{free},j})\mathbf{\Sigma}^{(\ell)}_j\Big) + \left(1 - \frac{r}{d_{\mathrm{free},j}}\right)Tr(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j). \tag{102}$$

Moreover,

$$\mathbb{E}\{Tr(\mathcal{P}^{\perp}_{\mathrm{rand}}\mathbf{\Sigma}^{(\ell)}_j)\} - Tr\Big((\mathcal{P}^{(\ell)^\star}_j)^{\perp}\mathbf{\Sigma}^{(\ell)}_j\Big) = \sum_{q=1}^{r} \lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) - \frac{r}{d_{\mathrm{free},j}}\sum_{q=1}^{d_{\mathrm{free},j}} \lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) \geq 0. \tag{103}$$

Thus, except in degenerate isotropic or capacity-saturated cases, a learned feasible basis strictly improves upon a frozen random feasible basis in expectation.

Proof. By rotational invariance of the Haar distribution on the $d_{\mathrm{free},j}$-dimensional free subspace, the expectation $\mathbb{E}[\mathcal{P}_{\mathrm{rand}}]$ must commute with every orthogonal transformation acting on $\mathrm{range}(\mathcal{P}^{(\ell)}_{\mathrm{free},j})$. Hence it must be a scalar multiple of $\mathcal{P}^{(\ell)}_{\mathrm{free},j}$, say

$$\mathbb{E}[\mathcal{P}_{\mathrm{rand}}] = \alpha \mathcal{P}^{(\ell)}_{\mathrm{free},j}. \tag{104}$$

Taking traces gives

$$r = \mathrm{Tr}(\mathbb{E}[\mathcal{P}_{\mathrm{rand}}]) = \alpha\, d_{\mathrm{free},j}, \tag{105}$$

so $\alpha = r/d_{\mathrm{free},j}$, proving the first identity.

Using $\widetilde{\mathbf{\Sigma}}^{(\ell)}_j = \mathcal{P}^{(\ell)}_{\mathrm{free},j}\widetilde{\mathbf{\Sigma}}^{(\ell)}_j$,

$$\mathbb{E}\{\mathrm{Tr}(\mathcal{P}_{\mathrm{rand}}\widetilde{\mathbf{\Sigma}}^{(\ell)}_j)\} = \mathrm{Tr}\left(\mathbb{E}[\mathcal{P}_{\mathrm{rand}}]\widetilde{\mathbf{\Sigma}}^{(\ell)}_j\right) = \frac{r}{d_{\mathrm{free},j}}\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j). \tag{106}$$

The formula for $\mathbb{E}[\mathrm{Tr}(\mathcal{P}^{\perp}_{\mathrm{rand}}\mathbf{\Sigma}^{(\ell)}_j)]$ then follows directly from Theorem 7.

Finally, since the average of the top r eigenvalues is at least the average over all $d_{\mathrm{free},j}$ eigenvalues,

$$\frac{1}{r}\sum_{q=1}^{r}\lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) \geq \frac{1}{d_{\mathrm{free},j}}\sum_{q=1}^{d_{\mathrm{free},j}}\lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j), \tag{107}$$

which implies

$$\sum_{q=1}^{r}\lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) - \frac{r}{d_{\mathrm{free},j}}\sum_{q=1}^{d_{\mathrm{free},j}}\lambda_q(\widetilde{\mathbf{\Sigma}}^{(\ell)}_j) \geq 0. \tag{108}$$

This proves the claim. □

Combining the results in Section C.1 and Section C.2 proves Theorem 2.

Remark 1 (Relation to SeqLoRA)**.** *Theorem 7 should not be interpreted as saying that SeqLoRA explicitly computes the top eigenspace of $\widetilde{\mathbf{\Sigma}}^{(\ell)}_j$. Rather, it identifies the population quantity that matters for continual-learning stability under the orthogonality constraint. In the full nonlinear diffusion model, SeqLoRA seeks a favorable subspace implicitly by optimizing the end-to-end denoising loss over both $\mathbf{A}^{(\ell)}_j$ and $\mathbf{B}^{(\ell)}_j$, instead of freezing $\mathbf{B}^{(\ell)}_j$ a priori; the loss-optimal basis does not necessarily coincide with the covariance-optimal one, and the theory makes no claim about the critical points the algorithm reaches.*

Theorem 7 provides a simple learning-theoretic justification for SeqLoRA. Orthogonality ensures that future concepts can only interfere with concept j through the residual covariance mass outside $\mathrm{col}(\mathbf{B}_j)$, while learning $\mathbf{B}^{(\ell)}_j$ from data minimizes exactly this exposed residual energy within the feasible complement. Proposition 2 further shows that frozen random bases are statistically optimal only when the task covariance is essentially isotropic inside the free complement. In this precise sense, optimizing both LoRA factors yields a strictly better continual-learning bias than fixing the basis in advance.

D Decomposition of Catastrophic Forgetting: Mean Interference vs. Stochastic Deviation

The end-to-end catastrophic forgetting bound derived in Theorem 2 reveals a multi-scale structure. At each of the L adapted layers, the locally injected crosstalk $\|\mathbf{\Delta}^{(\ell)}_j\|^2_F$ decomposes into two distinct sources: the expected systematic interference $\mu^{(\ell)}_z$ and the stochastic tail deviation $t^{(\ell)}$. These local statistical fluctuations are then propagated and aggregated through the downstream network layers, controlled by the amplification factors Γ_ℓ, to produce the final global forgetting. We now analyze the physical meaning of each component and its role in the total network-wide interference.

Systematic Interference ($\mu_z^{(\ell)}$). The mean crosstalk $\mu_z = \mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})\|\mathbf{Q}_j^{(\ell)}\|_F^2$ factorizes into two physical quantities. The first, $\|\mathbf{Q}_j^{(\ell)}\|_F^2 = \mathrm{Tr}(\mathbf{C}_j^{(\ell)}\mathbf{\Sigma}_j^{(\ell),\perp})$, is the feature-level interference energy. Crucially, it depends exclusively on the residual covariance $\mathbf{\Sigma}_j^{(\ell),\perp}$, not the full covariance $\mathbf{\Sigma}_j^{(\ell)}$. This is a direct consequence of SeqLoRA's orthogonality constraint, which perfectly annihilates the primary signal component ($\mathbf{C}_j^{(\ell)}\mathcal{P}_{B_j}^{(\ell)} = \mathbf{0}$). The better the LoRA basis $\mathbf{B}_j^{(\ell)}$ captures concept j's features, the smaller the residual $\mathbf{\Sigma}_j^{(\ell),\perp}$ becomes, driving feature interference toward small values. The second term, $\mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})$, represents the total token sequence energy. Together, orthogonality guarantees that the expected forgetting is confined entirely to the small, uncaptured residual subspace. Additionally, because the activation $\mathbf{X}_j^{(\ell)}$ is itself the output of a network optimized for concept j, the LoRA bases across all layers work in concert to capture the concept's energy; any features not absorbed by $\mathbf{B}_j^{(\ell-1)}$ can be captured by $\mathbf{B}_j^{(\ell)}$, ensuring the per-layer residual covariance $\mathbf{\Sigma}_j^{(\ell),\perp}$ remains uniformly small throughout the network depth. This is a heuristic description of the mechanism rather than a proven property.

Stochastic Deviation ($t^{(\ell)}$). The deviation $t^{(\ell)}$ captures the random fluctuation of the crosstalk around its mean due to the realization of the sub-Gaussian noise. It is bounded by $t^{(\ell)} = \max(T_1, T_2)$, where:

$$T_1 = K^2\sqrt{C_1}\|\mathbf{\Psi}_j^{(\ell)}\|_F\|(\mathbf{Q}_j^{(\ell)})^\top\mathbf{Q}_j^{(\ell)}\|_F\sqrt{\log\frac{2}{\xi}} \qquad \text{(Sub-Gaussian Term)}$$

$$T_2 = K^2 C_1\|\mathbf{\Psi}_j^{(\ell)}\|_2\|(\mathbf{Q}_j^{(\ell)})^\top\mathbf{Q}_j^{(\ell)}\|_2\log\frac{2}{\xi} \qquad \text{(Sub-Exponential Term).}$$

Defining the effective rank as $\mathrm{r}_{\mathrm{eff}}(\mathbf{A}) \triangleq \|\mathbf{A}\|_F/\|\mathbf{A}\|_2$, we determine the dominant regime by setting $T_1 \geq T_2$. Solving this inequality yields the threshold condition for the sub-Gaussian term to dominate:

$$\mathrm{r}_{\mathrm{eff}}(\mathbf{\Psi}_j^{(\ell)}) \cdot \mathrm{r}_{\mathrm{eff}}\big((\mathbf{Q}_j^{(\ell)})^\top\mathbf{Q}_j^{(\ell)}\big) \geq \sqrt{C_1\log\frac{2}{\xi}}. \tag{109}$$

The heavy-tailed sub-exponential regime only triggers if this strict inequality is reversed.

Regime Analysis: Diffuse vs. Spiky Interference. The threshold condition in Equation (109) reveals that catastrophic forgetting behaves differently depending on the magnitude of the joint effective rank. This transition is governed by whether the residual geometry is diffuse or spiky:

- **The Sub-Gaussian Regime (Diffuse Interference):** When the joint effective rank is large ($T_1 \geq T_2$), the sub-Gaussian term dominates. Because the learned LoRA basis $\mathbf{B}_j^{(\ell)}$ successfully absorbs the dominant principal components of the activations, the residual covariance $\mathbf{\Sigma}_j^{(\ell),\perp}$ retains only a flat, unstructured spectral tail. The interference is *diffuse*, spread evenly across many dimensions, acting as benign noise that scales gracefully as $\sqrt{\log(2/\xi)}$.
- **The Sub-Exponential Regime (Spiky Interference):** Conversely, the heavier-tailed sub-exponential regime ($T_2 > T_1$) only triggers if the joint effective rank collapses (≈ 1). This occurs if the residual covariance still contains a massive, uncaptured principal direction, making the interference highly *spiky* and concentrated. In this state, forgetting is dictated by worst-case fluctuations along that single dominant axis, scaling poorly as $\log(2/\xi)$.

The Dual Role of Orthogonality. SeqLoRA's inter-concept orthogonality ($\mathbf{B}_i^\top\mathbf{B}_j = \mathbf{0}$) serves a dual mathematical purpose to protect the network. First, it shrinks the systematic mean μ_z by restricting interference exclusively to the residual subspace. Second, by projecting out the dominant principal components captured by $\mathbf{B}_j^{(\ell)}$, it exposes the naturally diffuse structure of the residual spectrum, ensuring a large joint effective rank that keeps the crosstalk in the well-concentrated sub-Gaussian regime.

E Population Risk Analysis in a Stylized Linear Setting

Theorem 1 establishes the optimization dynamics of SeqLoRA, and Theorem 2 provides a high-probability bound on end-to-end catastrophic forgetting in deep, non-linear networks. To complement this result, we now derive an *exact*, closed-form decomposition of catastrophic forgetting in a stylized linear setting with a single adapted layer. This is not intended to model the full diffusion network; rather, by removing non-linearities, it isolates the precise mechanism through which orthogonality and learned-basis adaptation interact, and makes explicit the residual-energy quantity that controls forgetting. The two analyses agree in their conclusion: the high-probability bound in Theorem 2 and the exact decomposition below both identify $\mathrm{Tr}(\mathcal{P}_{B_j}^{\perp}\mathbf{\Sigma}_j)$ as the driver of forgetting, providing convergent evidence from two distinct analytical viewpoints.

Stylized population model. Consider a single linear map with fixed input features $\mathbf{x} \in \mathbb{R}^m$ and output $\mathbf{y} \in \mathbb{R}^n$. For concept j, let $\mathcal{D}_j$ be a population distribution over $(\mathbf{x}, \mathbf{y})$, and define the squared population risk

$$\mathcal{L}_j(\mathbf{W}) \triangleq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_j}\{\|\mathbf{y} - \mathbf{W}\mathbf{x}\|_2^2\}, \qquad \mathbf{\Sigma}_j \triangleq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_j}\{\mathbf{x}\mathbf{x}^\top\}. \tag{110}$$

As in LoRA, after learning t concepts the adapted matrix is

$$\mathbf{W}_t = \mathbf{W}_0 + \sum_{s=1}^{t} \mathbf{A}_s\mathbf{B}_s^\top. \tag{111}$$

Fix a concept j. After learning concept j, the model is

$$\mathbf{W}_j = \mathbf{W}_0 + \sum_{s=1}^{j} \mathbf{A}_s\mathbf{B}_s^\top, \tag{112}$$

while after all later concepts have been learned,

$$\mathbf{W}_T = \mathbf{W}_j + \mathbf{C}_j, \qquad \mathbf{C}_j \triangleq \sum_{s=j+1}^{T} \mathbf{A}_s\mathbf{B}_s^\top. \tag{113}$$

Let

$$\mathcal{P}_{B_j} \triangleq \mathbf{B}_j(\mathbf{B}_j^\top\mathbf{B}_j)^{-1}\mathbf{B}_j^\top, \qquad \mathcal{P}_{B_j}^{\perp} \triangleq \mathbf{I} - \mathcal{P}_{B_j}, \tag{114}$$

denote the orthogonal projector onto the learned right-subspace of concept j and its orthogonal complement. In this simplified view, the role of the right LoRA factor $\mathbf{B}_j$ is summarized by $\mathcal{P}_{B_j}$, while the cumulative effect of all future concepts is summarized by the crosstalk operator $\mathbf{C}_j$.

The first result shows that, under orthogonality, forgetting is driven only by the part of the concept covariance that lies outside the learned basis.

Theorem 8 (Residual-subspace decomposition of forgetting)**.** *Assume that the later bases satisfy*

$$\mathbf{B}_s^\top\mathbf{B}_j = \mathbf{0}, \qquad \forall s > j. \tag{115}$$

Then the population forgetting on concept j admits the exact decomposition

$$\mathcal{L}_j(\mathbf{W}_T) - \mathcal{L}_j(\mathbf{W}_j) = Tr\Big(\mathbf{C}_j\mathcal{P}_{B_j}^{\perp}\mathbf{\Sigma}_j\mathcal{P}_{B_j}^{\perp}\mathbf{C}_j^\top\Big) + \langle\nabla\mathcal{L}_j(\mathbf{W}_j),\, \mathbf{C}_j\rangle\,. \tag{116}$$

Consequently, if $\|\nabla\mathcal{L}_j(\mathbf{W}_j)\|_F \le \varepsilon_j$, then

$$\mathcal{L}_j(\mathbf{W}_T) - \mathcal{L}_j(\mathbf{W}_j) \le \|\mathbf{C}_j\|_2^2 Tr\Big(\mathcal{P}_{B_j}^{\perp}\mathbf{\Sigma}_j\Big) + \varepsilon_j\|\mathbf{C}_j\|_F. \tag{117}$$

In particular, if $\mathbf{W}_j$ is a stationary point of $\mathcal{L}_j$, then

$$\mathcal{L}_j(\mathbf{W}_T) - \mathcal{L}_j(\mathbf{W}_j) = Tr\Big(\mathbf{C}_j\mathcal{P}_{B_j}^{\perp}\mathbf{\Sigma}_j\mathcal{P}_{B_j}^{\perp}\mathbf{C}_j^\top\Big) \le \|\mathbf{C}_j\|_2^2 Tr\Big(\mathcal{P}_{B_j}^{\perp}\mathbf{\Sigma}_j\Big)\,. \tag{118}$$

Proof. Because $\mathbf{B}_s^\top\mathbf{B}_j = \mathbf{0}$ for every $s > j$, we also have

$$\mathbf{B}_s^\top\mathcal{P}_{B_j} = \mathbf{0}, \qquad \forall s > j, \tag{119}$$

since $\mathrm{range}(\mathcal{P}_{B_j}) = \mathrm{col}(\mathbf{B}_j)$. Therefore

$$\mathbf{C}_j \mathcal{P}_{B_j} = \sum_{s=j+1}^{T} \mathbf{A}_s \mathbf{B}_s^\top \mathcal{P}_{B_j} = \mathbf{0}, \tag{120}$$

which implies $\mathbf{C}_j = \mathbf{C}_j \mathcal{P}_{B_j}^\perp$.

Now expand the population risk difference:

$$\mathcal{L}_j(\mathbf{W}_T) - \mathcal{L}_j(\mathbf{W}_j) = \mathbb{E}\{\|\mathbf{y} - (\mathbf{W}_j + \mathbf{C}_j)\mathbf{x}\|_2^2 - \|\mathbf{y} - \mathbf{W}_j\mathbf{x}\|_2^2\} \tag{121}$$

$$= \mathbb{E}\{\|\mathbf{C}_j\mathbf{x}\|_2^2\} + 2\,\mathbb{E}\{(\mathbf{W}_j\mathbf{x} - \mathbf{y})^\top \mathbf{C}_j\mathbf{x}\}. \tag{122}$$

Since

$$\nabla \mathcal{L}_j(\mathbf{W}) = 2\,\mathbb{E}\{(\mathbf{W}\mathbf{x} - \mathbf{y})\mathbf{x}^\top\}, \tag{123}$$

the second term is exactly $\langle \nabla \mathcal{L}_j(\mathbf{W}_j), \mathbf{C}_j \rangle$. For the first term, using $\mathbf{C}_j = \mathbf{C}_j \mathcal{P}_{B_j}^\perp$,

$$\mathbb{E}\{\|\mathbf{C}_j\mathbf{x}\|_2^2\} = \mathbb{E}\{\mathbf{x}^\top \mathbf{C}_j^\top \mathbf{C}_j \mathbf{x}\} \tag{124}$$

$$= \mathrm{Tr}\big(\mathbf{C}_j \mathbf{\Sigma}_j \mathbf{C}_j^\top\big) \tag{125}$$

$$= \mathrm{Tr}\Big(\mathbf{C}_j \mathcal{P}_{B_j}^\perp \mathbf{\Sigma}_j \mathcal{P}_{B_j}^\perp \mathbf{C}_j^\top\Big). \tag{126}$$

This proves the identity.

For the upper bound, note that for any positive semidefinite matrix $\mathbf{M}$,

$$\mathrm{Tr}(\mathbf{C}_j \mathbf{M} \mathbf{C}_j^\top) \le \|\mathbf{C}_j\|_2^2 \mathrm{Tr}(\mathbf{M}). \tag{127}$$

Applying this with $\mathbf{M} = \mathcal{P}_{B_j}^\perp \mathbf{\Sigma}_j \mathcal{P}_{B_j}^\perp$ gives

$$\mathrm{Tr}\Big(\mathbf{C}_j \mathcal{P}_{B_j}^\perp \mathbf{\Sigma}_j \mathcal{P}_{B_j}^\perp \mathbf{C}_j^\top\Big) \le \|\mathbf{C}_j\|_2^2 \mathrm{Tr}\Big(\mathcal{P}_{B_j}^\perp \mathbf{\Sigma}_j\Big). \tag{128}$$

Finally, by Cauchy–Schwarz,

$$|\langle \nabla \mathcal{L}_j(\mathbf{W}_j),\, \mathbf{C}_j \rangle| \le \|\nabla \mathcal{L}_j(\mathbf{W}_j)\|_F \|\mathbf{C}_j\|_F \le \varepsilon_j \|\mathbf{C}_j\|_F. \tag{129}$$

The stationary case follows by setting $\varepsilon_j = 0$. □

This result is perfectly consistent with Theorem 2, which derives the catastrophic forgetting by $\mathrm{Tr}(\mathcal{P}_{B_j}^{(\ell),\perp} \mathbf{\Sigma}_j^{(\ell)}) = \mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp}) \mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)})$. The alignment between our deep-network concentration bounds and this population-risk analysis underscores that minimizing this residual energy is the key to sequential stability.

F Numerical Illustration of Theorem 2 in a Small Nonlinear Network

Evaluating Theorem 2 directly on a diffusion model is challenging because the quantities it involves, such as residual covariances, crosstalk matrices, and products of layerwise Lipschitz constants, are not readily available in that setting. Its structure can be checked, however, and the theorem makes three predictions that are concrete enough to test. First, under the orthogonality constraint the crosstalk injected at a layer acts only on the residual part of the activations, so its energy should equal $\mathrm{Tr}(\mathbf{C}_j^{(\ell)}\mathbf{\Sigma}_j^{(\ell),\perp}(\mathbf{C}_j^{(\ell)})^\top)$ (Proposition 1 and Lemma 5). Second, the per-concept quantity $\sum_\ell \Gamma_\ell \kappa_\ell \sqrt{\epsilon_\ell}$ that appears in the bound should order the concepts by how much they are forgotten. Third, a basis learned inside the feasible complement should leave less residual energy, and therefore forget less, than a random basis frozen in the same complement (Theorem 7 and Proposition 2). We test these three predictions in a small nonlinear network.

Setup. The network has $L = 3$ LoRA-adapted linear layers of sizes 48×48, 48×48 and 16×48, separated by $\tanh$ ($\gamma_\ell = 1$), and every layer carries a rank-$r = 4$ residual $\mathbf{A}_j^{(\ell)}(\mathbf{B}_j^{(\ell)})^\top$. A stream of $T = 8$ concepts arrives sequentially. Concept j has its own input distribution $\mathbf{x} = \mathbf{\Sigma}_j^{1/2}\mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{48})$; the covariance $\mathbf{\Sigma}_j$ has six dominant eigen-directions, with eigenvalues decreasing from 3.5 to 1.3, and an isotropic floor of 0.0225 on the remaining 42 directions. Each concept also has its own teacher, obtained by adding a concept-specific rank-4 perturbation to every layer of the pretrained network. As in the theory, a batch is a matrix $\mathbf{X}_j^{(\ell)} \in \mathbb{R}^{m\times p}$ whose p columns are tokens; here $p = 256$ and the tokens are i.i.d., so $\mathbf{\Psi}_j^{(\ell)} = \mathbf{I}_p$ and the factor $\mathrm{Tr}(\mathbf{\Psi}_j^{(\ell)}) = p$ is common to all quantities and is dropped. The LoRA factors of concept j are trained for 400 steps of gradient descent (step size 0.05) on the squared loss to the teacher, with $\mathbf{A}_j^{(\ell)}$ initialised at zero and $\mathbf{B}_j^{(\ell)}$ at random. Three ways of handling the basis are compared:

- *learned, orthogonal*: both factors are trained, and $\mathbf{B}_j^{(\ell)}$ is projected onto the complement of the previous bases after every step, as in (17);
- *frozen random, orthogonal*: $\mathbf{B}_j^{(\ell)}$ is drawn at random inside the feasible complement and kept fixed, and only $\mathbf{A}_j^{(\ell)}$ is trained; this is the abstraction of frozen-basis methods analysed in Proposition 2;
- *learned, unconstrained*: both factors are trained and no orthogonality is imposed.

Once the stream is complete we take $4{,}096$ fresh samples of each concept $j < T$ (the last concept has no successors and hence no crosstalk) and compute, at every layer, the residual energy $\epsilon_\ell = \mathrm{Tr}(\mathbf{\Sigma}_j^{(\ell),\perp})$ of the layer inputs under θ_j, the crosstalk magnitude $\kappa_\ell = \|\mathbf{C}_j^{(\ell)}\|_2$, the amplification $\Gamma_\ell = \prod_{\ell'>\ell} \|\mathbf{W}_T^{(\ell')}\|_2$, the injected crosstalk energy $\|\mathbf{C}_j^{(\ell)}\mathbf{X}_j^{(\ell)}\|_F^2/p$ together with its residual-only counterpart $\mathrm{Tr}(\mathbf{C}_j^{(\ell)}\mathbf{\Sigma}_j^{(\ell),\perp}(\mathbf{C}_j^{(\ell)})^\top)$, and finally the forgetting $\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_j)$, i.e. the measured increase of the loss between the model right after concept j and the final model. With five seeds this gives 35 (concept, seed) pairs per strategy. Figure 3 shows all of them and Table 5 reports the means.

Table 5: Summary of the synthetic study (mean $\pm$ standard error over 35 (concept, seed) pairs; energies summed over the three layers). "Forgetting" is the measured loss increase $\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_j)$, not the bound. "Residual-only" is $\sum_\ell \mathrm{Tr}(\mathbf{C}_j^{(\ell)}\mathbf{\Sigma}_j^{(\ell),\perp}(\mathbf{C}_j^{(\ell)})^\top)$ and "measured" is $\sum_\ell \|\mathbf{C}_j^{(\ell)}\mathbf{X}_j^{(\ell)}\|_F^2/p$; their difference is the crosstalk that passes through the captured subspace, which orthogonality removes. The last column is the Spearman correlation, within the strategy, between the predictor of Theorem 2 and the measured forgetting.

Basis strategy	Residual energy $\sum_\ell \epsilon_\ell$	Loss after learning	Forgetting	Crosstalk energy		Spearman
				residual-only	measured	
Frozen random, orthogonal	34.3 ± 0.7	1.32 ± 0.11	14.2 ± 0.8	23.8	23.8	0.70
Learned, orthogonal	27.9 ± 0.2	0.34 ± 0.02	3.1 ± 0.2	6.67	6.67	0.77
Learned, unconstrained	26.9 ± 0.2	0.28 ± 0.01	3.5 ± 0.2	4.95	7.02	0.69

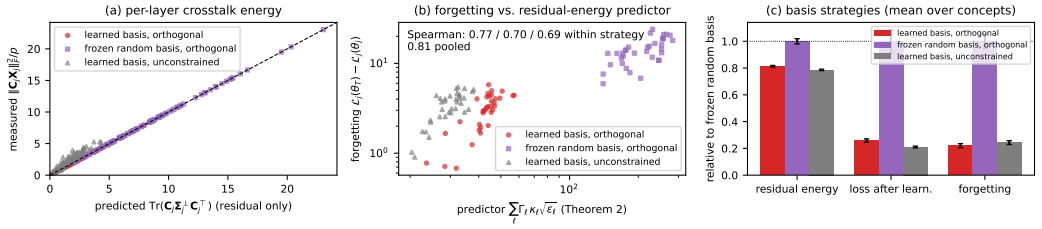


Figure 3: Synthetic study of Theorem 2. (a) Measured per-layer crosstalk energy $\|\mathbf{C}_j^{(\ell)}\mathbf{X}_j^{(\ell)}\|_F^2/p$ against the residual-only value $\mathrm{Tr}(\mathbf{C}_j^{(\ell)}\mathbf{\Sigma}_j^{(\ell),\perp}(\mathbf{C}_j^{(\ell)})^\top)$ of Lemma 5, one point per (concept, layer, seed). Under orthogonality the two coincide (Proposition 1); without the constraint the measured value is larger because the crosstalk also acts on the captured subspace. (b) Measured forgetting of each concept against the predictor $\sum_\ell \Gamma_\ell \kappa_\ell \sqrt{\epsilon_\ell}$ of Theorem 2 (log–log axes, one point per concept and seed). (c) Mean residual energy, loss right after learning, and forgetting of the three strategies, relative to the frozen random basis (error bars: standard error over the 35 pairs).

The residual covariance as the only interference channel. For the two orthogonal strategies the measured crosstalk energy coincides with the residual-only value at every layer, seed and concept: all their points in panel (a) lie on the diagonal, and the summed values in Table 5 agree to all digits. This is not surprising in itself, since under exact orthogonality the two quantities are algebraically equal ($\mathbf{C}_j^{(\ell)}\mathcal{P}_{B_j}^{(\ell)} = \mathbf{0}$), so for these strategies the panel mainly confirms that the projection is exact at every layer of the trained network. The informative case is the unconstrained one. There the measured crosstalk is larger than the residual-only value, by 30% on average (7.02 against 4.95 summed over layers). The excess is interference that enters through the captured subspace $\mathrm{col}(\mathbf{B}_j^{(\ell)})$ once later bases overlap with it, and it is the channel that the orthogonality constraint removes.

Ordering of concepts by the predictor. Within each strategy, the Spearman correlation between $\sum_\ell \Gamma_\ell \kappa_\ell \sqrt{\epsilon_\ell}$ and the measured forgetting is 0.77 for the learned orthogonal basis, 0.70 for the frozen orthogonal basis and 0.69 for the unconstrained basis; pooled over the three strategies it is 0.81, with a Pearson correlation of 0.93. The relation is monotone but not linear. Panel (b) also shows that the predictor separates the three strategies by close to an order of magnitude, in the same direction as the measured forgetting. Its absolute value overestimates the forgetting by a roughly constant factor, which is what one expects from the conservative prefactors in the bound; the predictor should therefore be used to rank concepts and strategies, not as a numerical estimate of the loss increase.

Learned versus frozen basis, and what the constraint costs. Compared with a random basis frozen in the feasible complement, learning the basis lowers the summed residual energy from 34.3 to 27.9, the loss right after learning from 1.32 to 0.34, and the forgetting from 14.2 to 3.1 (panel (c)). The reduction in forgetting is considerably larger than the reduction in residual energy. The reason is that a frozen basis affects every factor of the bound at once: the concept is fitted less well, the factor $\mathbf{A}_j^{(\ell)}$ grows to compensate and thereby inflates the crosstalk magnitudes (κ_ℓ averages 3.1 against 1.6), and the larger weights also increase the downstream amplification ($\Gamma_1 = 15.0$ against 4.8). This is the mechanism behind Proposition 2 and the second part of Theorem 2.

The unconstrained learned basis reaches a slightly lower residual energy than the orthogonal one (26.9 against 27.9) and a slightly lower loss after learning (0.28 against 0.34). This is the expressiveness cost of the constraint that Theorem 7 makes explicit: a feasible basis has to lie in the free complement, so whatever part of concept j's activation energy falls inside the subspaces of earlier concepts, $\mathrm{Tr}((\mathbf{I} - \mathcal{P}_{\mathrm{free},j})\mathbf{\Sigma}_j)$, necessarily stays in the residual. Consistently with this, the residual-energy gap between the two strategies is 0.0 for the first concept and grows along the stream (0.0, -0.1, 0.1, 0.9, 1.7, 2.3, 1.6 for $j = 1, \ldots, 7$). In spite of its smaller residual, the unconstrained basis forgets slightly more (3.5 against 3.1): the 2.1 units of crosstalk that pass through its captured subspace outweigh the 1.0 unit of residual energy it saves. The constraint thus exchanges a somewhat larger residual for a closed leakage channel. With $m = 48$ and eight rank-4 concepts, bases learned independently

overlap little, so the net effect on forgetting is small in this setting; the concept-similarity experiment in Appendix I.3 (Table 8) suggests that it becomes larger as the concepts become more similar.

G Evaluation Metrics

We use the following metrics to evaluate different aspects of the generated images:

- **CLIP-I / CLIP-T** (Radford et al., 2021; Ruiz et al., 2023): image-to-image and text-to-image alignment in CLIP space, measuring whether generated images match the reference concept and the prompt.
- **DINO / DINOv2 / DINOv3** (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025): cosine similarity of self-supervised features to the reference images, which is sensitive to identity rather than category.
- **DreamSim** (Fu et al., 2023): holistic visual similarity (layout, pose, and semantic content) to the reference images.
- **HPSv2 / HPSv3** (Wu et al., 2023; Ma et al., 2025): human preference scores for overall visual quality.

H Evaluation Protocols

This section details the three protocols of Section 4 and the aggregation rules. In all protocols each per-concept score is first averaged over 16 generated images obtained with different random seeds.

H.1 Single-concept retention (Table 1, Figure 2a, Table 14)

For a stream of T concepts we evaluate the model only after it has been trained on all T concepts. The final composed model generates 16 images for every one of the T concepts using the CustomConcept101 prompts, yielding $16 \times T$ images; each metric is computed for every image against the concept's reference images. Figure 2a plots the mean of the $16 \times T$ scores (at $T = 101$ this is the mean of 1,616 generations). Tables report mean $\pm$ standard error of the mean (SEM) over the T per-concept scores, $\mathrm{SEM} = \sigma/\sqrt{T}$ with σ the unbiased standard deviation across concepts; SEMs therefore capture concept-level variability and decrease naturally with T.

H.2 Forgetting (Figure 2b, Table 3, Table 13)

Figure 2b tracks the 32-concept sequential run: a point at step k is the state of the model right after learning the k-th concept, evaluated on all concepts $1, \ldots, k$ seen so far ($16 \times k$ images), and the plotted value is the mean over these images. Because the evaluated pool grows with k and later concepts may be intrinsically harder, a downward slope of this curve is not, by itself, forgetting; the curve measures the aggregate health of the model along the stream and is best read as a comparison between methods trained on the same stream. Table 3 isolates forgetting: for each concept j the *Initial* score is measured right after concept j is learned and the *Final* score in the final T-concept model, and *Forgetting* is their difference, averaged over concepts.

H.3 Multi-concept composition (Table 2, Figure 7)

We follow the regional generation protocol of Gu et al. (2023): three or four concepts learned in the same 32-concept stream are rendered in one image, each in a region specified by a bounding box, with sketch and keypose conditioning shared across methods and the same random seed for every method. For each composition and seed we crop each concept's region and compute DINOv2 cosine similarities: *identity* is the similarity of the crop to the concept's own reference images, and *leakage* is its similarity to the reference images of the other concepts in the same composition (averaged over those concepts). CLIP-T is computed on the full image against the composition prompt. Scores are averaged over five seeds per composition and over compositions; the significance test in Table 2 is a paired Wilcoxon signed-rank test over compositions between SeqLoRA and Continual Alternating.

I Additional Experiments

I.1 Sensitivity to concept order

Because SeqLoRA projects each new basis onto the complement of the bases learned before it, later concepts are learned in a smaller subspace, and one may ask whether the order of the stream matters. Table 6 reports composition identity and leakage (protocol of Appendix H) for four random orderings of a 16-concept stream, five seeds each. The variation across orderings is small (standard deviation ≤ 0.007, coefficient of variation $\leq 3\%$), indicating limited sensitivity to order at this stream length; with $m = 768$ and $r = 4$, sixteen concepts occupy 64 of 768 dimensions per layer, so the complement available to the last concept remains large.

Table 6: Sensitivity to concept order: SeqLoRA on four random orderings of a 16-concept stream (5 seeds each, DINOv2 features, composition protocol).

Ordering	Identity↑	Leakage↓
Order 1	0.316	0.206
Order 2	0.311	0.201
Order 3	0.324	0.216
Order 4	0.324	0.207
Mean $\pm$ std	0.319 ± 0.007	0.208 ± 0.006

I.2 Sensitivity to LoRA rank

The orthogonality constraint admits at most $\lfloor m/r \rfloor$ concepts per layer, so the rank trades capacity against per-concept expressiveness. Table 7 varies the rank on a 16-concept stream. Identity is not monotone in the rank, and a larger rank increases leakage without improving the identity–leakage margin; we therefore use $r = 4$ throughout.

Table 7: LoRA-rank sensitivity of SeqLoRA (16 concepts, DINOv2 features, composition protocol). Margin is identity minus leakage.

LoRA rank	Identity↑	Leakage↓	Margin↑
$r = 4$	0.322	0.200	0.122
$r = 8$	0.316	0.221	0.095
$r = 16$	0.338	0.219	0.119

I.3 Effect of concept similarity

Interference is expected to be largest between visually similar concepts, whose activations share directions. To test where the orthogonal projection matters, we compare a group of visually similar concepts with a group of diverse concepts under the composition protocol, with and without the projection step of (17). Table 8 shows that the projection reduces leakage substantially for the similar group and is essentially a no-op for the diverse group, whose leakage is already low. The mechanism therefore helps precisely in the regime where a dimension-counting capacity argument is least informative.

Table 8: Cross-concept leakage (↓, DINOv2 features) for a group of visually similar concepts and a group of diverse concepts, with the orthogonal projection of (17) disabled and enabled.

Concept group	Without projection	With projection
Visually similar	0.434	0.386
Diverse	0.070	0.078

I.4 COMPUTE AND MEMORY

Table 4 in the main text reports end-to-end cost at 32 concepts. Table 9 reports the peak memory of the composition step as a function of the number of concepts. Optimization-based fusion must instantiate all per-concept adapters (and the optimizer state) simultaneously, so its peak memory grows with the number of concepts; LoRACLR's grows most steeply and reaches 44.7 GB at 32 concepts, which is consistent with the out-of-memory failures of Mix-of-Show and LoRACLR beyond 48 and 64 concepts in Table 14. Training-step peak memory is independent of the number of concepts for all methods, since each concept's adapter is trained with the previously learned bases held fixed (7.3 GB for SeqLoRA vs. 6.9 GB for the parallel baselines, per step).

Table 9: Peak memory (GB) of the composition step versus number of concepts. For SeqLoRA the composition step is the summation of adapters into the base weights.

# Concepts	Mix-of-Show	LoRACLR	SeqLoRA
4	8.9	12.9	9.2
16	12.8	23.1	12.9
32	18.4	44.7	17.9

I.5 HYPERPARAMETER ABLATION

Table 10 varies the number of local steps $S_A = S_B$, the regularization ϵ of the projection, and the number of bilevel iterations K for the 32-concept stream. Reducing the local steps or the bilevel iterations to one leaves the identity metrics (DINO, DINOv2, DINOv3, DreamSim) essentially unchanged but lowers HPSv3; five local steps raise HPSv3 with a marginal change in identity; larger ϵ (10^{-4}) slightly reduces identity scores. The default configuration keeps near-optimal identity preservation with a competitive preference score. Since a single bilevel iteration matches the default identity metrics, the per-concept training cost of Section 4.3 can be reduced roughly threefold when preference scores are not the priority.

Table 10: Ablation study on hyperparameters for SeqLoRA with 32 concepts.

Factor	Value	DINO↑	DINOv2↑	DINOv3↑	DreamSim↑	CLIP-I↑	CLIP-T↑	HPSv2↑	HPSv3↑
Default	-	0.468 ± 0.021	0.418 ± 0.019	0.418 ± 0.016	0.477 ± 0.016	0.693 ± 0.009	0.286 ± 0.003	0.270 ± 0.001	9.043 ± 0.399
Local Steps	1 Step	0.466 ± 0.021	0.418 ± 0.017	0.414 ± 0.015	0.479 ± 0.016	0.694 ± 0.009	0.287 ± 0.003	0.270 ± 0.001	8.814 ± 0.422
	5 Steps	0.469 ± 0.021	0.417 ± 0.020	0.415 ± 0.017	0.475 ± 0.016	0.690 ± 0.009	0.286 ± 0.003	0.271 ± 0.001	9.378 ± 0.346
Epsilon	10^{-4}	0.459 ± 0.021	0.406 ± 0.018	0.405 ± 0.016	0.470 ± 0.017	0.686 ± 0.009	0.288 ± 0.003	0.271 ± 0.001	8.838 ± 0.402
	10^{-5}	0.466 ± 0.021	0.422 ± 0.019	0.417 ± 0.017	0.476 ± 0.016	0.692 ± 0.009	0.287 ± 0.004	0.271 ± 0.001	8.914 ± 0.402
Bilevel Iters	1 Iter	0.466 ± 0.021	0.418 ± 0.019	0.417 ± 0.016	0.477 ± 0.017	0.695 ± 0.009	0.288 ± 0.003	0.270 ± 0.001	8.784 ± 0.377
	4 Iters	0.464 ± 0.020	0.408 ± 0.018	0.410 ± 0.016	0.474 ± 0.016	0.691 ± 0.009	0.288 ± 0.003	0.271 ± 0.001	8.855 ± 0.390

I.6 PRELIMINARY TRANSFER TO SDXL

The projection (17) is a closed-form operation on the basis factors that depends only on the LoRA rank and the layer width, with no assumption specific to Stable Diffusion v1.5 or to a single text encoder. To test this, we ported SeqLoRA to SDXL-base (Podell et al., 2024), whose architecture differs substantially (two text encoders, 2048-dimensional cross-attention, 70 cross-attention layers, and pooled/time micro-conditioning); only the architecture-specific wiring (dual-encoder conditioning and the per-layer embedding count) was adapted. Table 11 reports a four-concept stream with rank 4 and single-concept evaluation: DINO (ViT-S/16) similarity of generated images to the concept's own references (identity) and to the other concepts' references (cross-concept, a genericness baseline). Identity is in the usual personalization range with a large margin over cross-concept similarity, indicating faithful, concept-specific transfer.

J QUALITATIVE RESULTS

We present qualitative comparisons for all 32 concepts of the main experiment, ordered by their position in the training sequence. Figures 4, 5, and 6 show concepts 1–12, 13–24, and 25–32. As in

Table 11: SeqLoRA on SDXL-base (4 concepts, rank 4): per-concept DINO identity versus cross-concept similarity.

Concept	DINO identity↑	Cross-concept	Margin↑
things_cup2	0.747	0.193	+0.554
plushie_cow	0.620	0.207	+0.413
wearable_glasses	0.591	0.186	+0.405
actionfigure_1	0.711	0.218	+0.493
Mean	0.667	0.201	+0.466

Figure 1, each row shows the reference images and samples from SeqLoRA, Continual Alternating, LoRACLR, Mix-of-Show, and Orthogonal Adaptation, all generated with the same seeds.

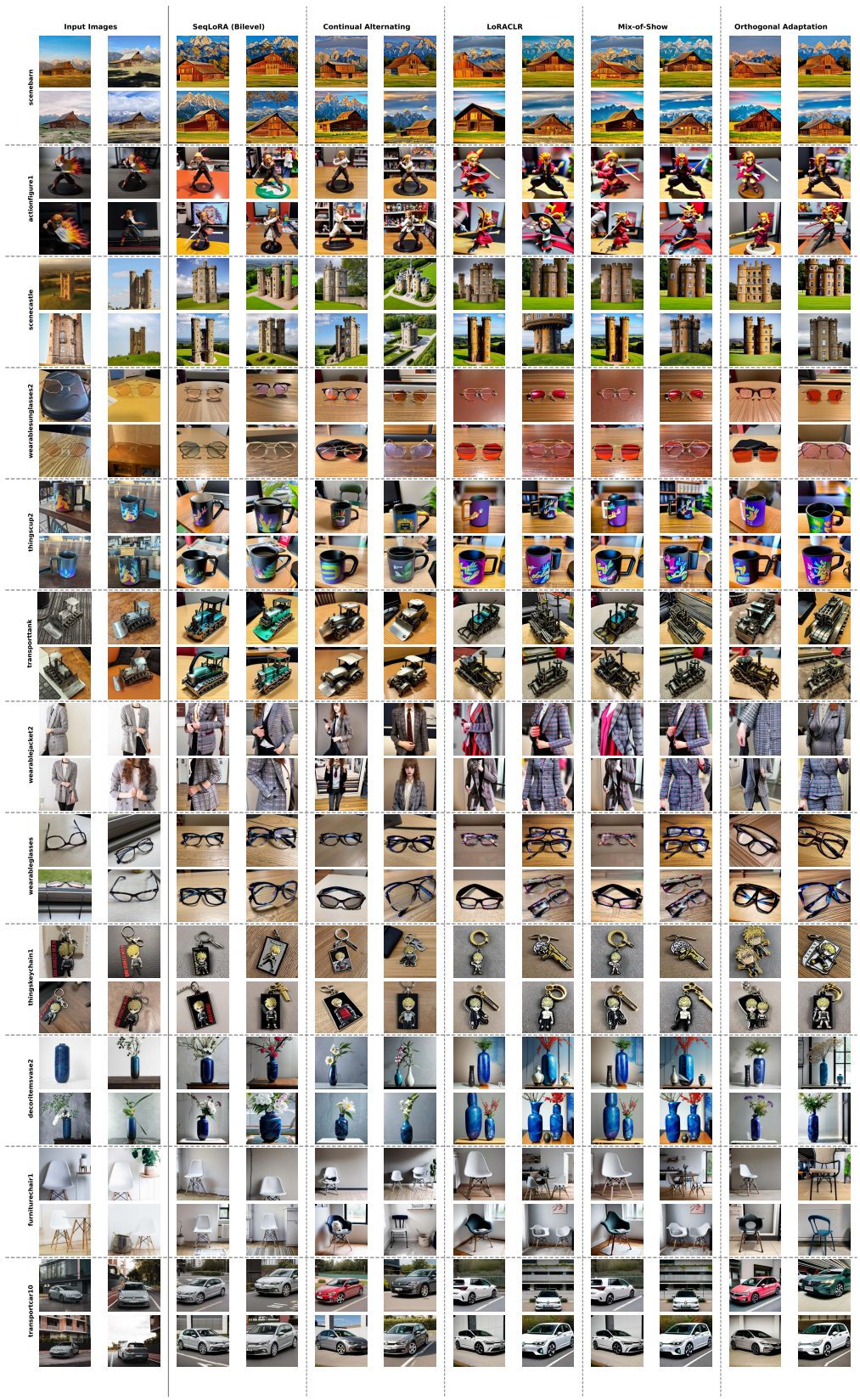


Figure 4: Qualitative comparison for concepts 1–12 (ordered by training sequence); same seeds for all methods.

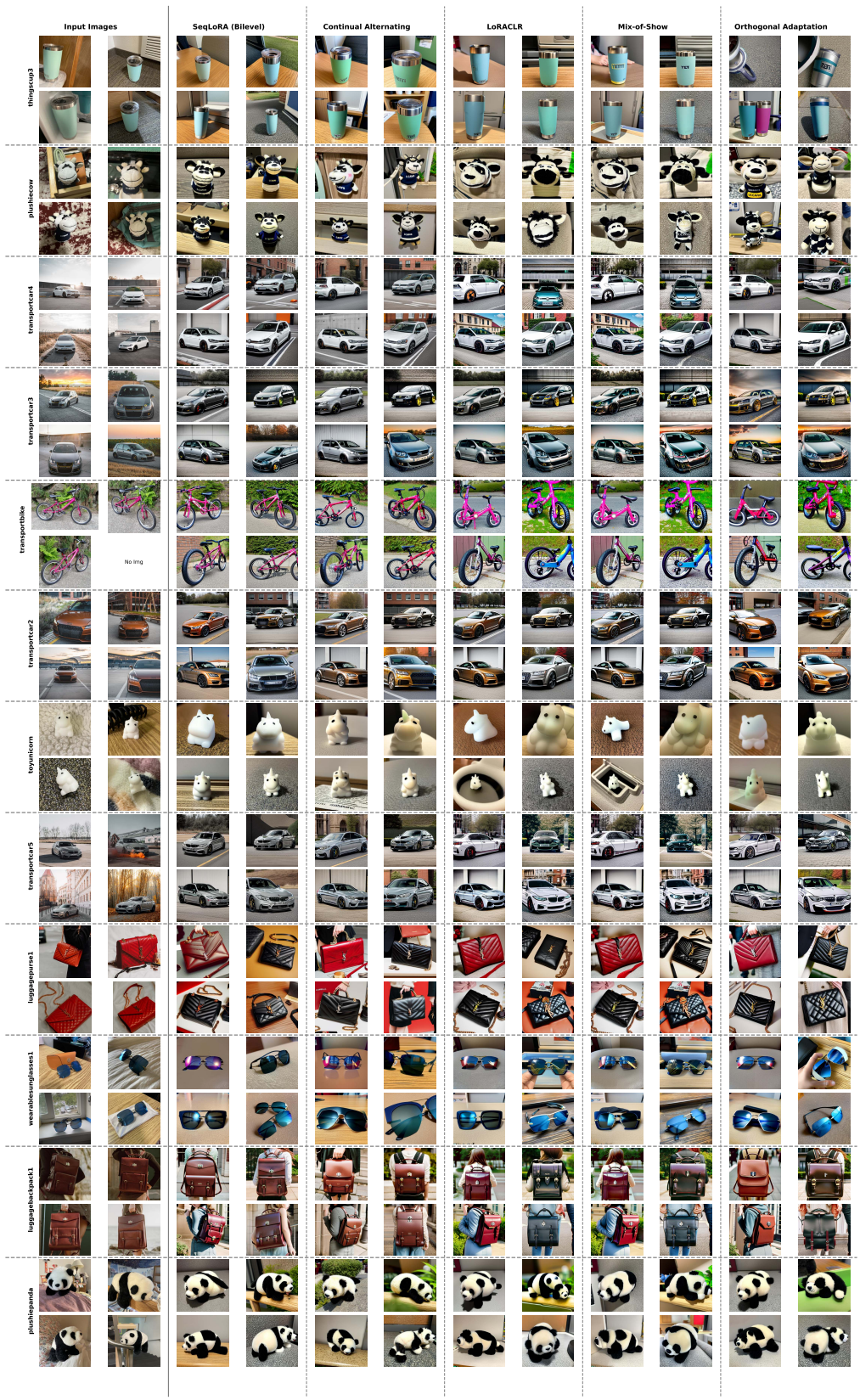


Figure 5: Qualitative comparison for concepts 13–24 (ordered by training sequence); same seeds for all methods.

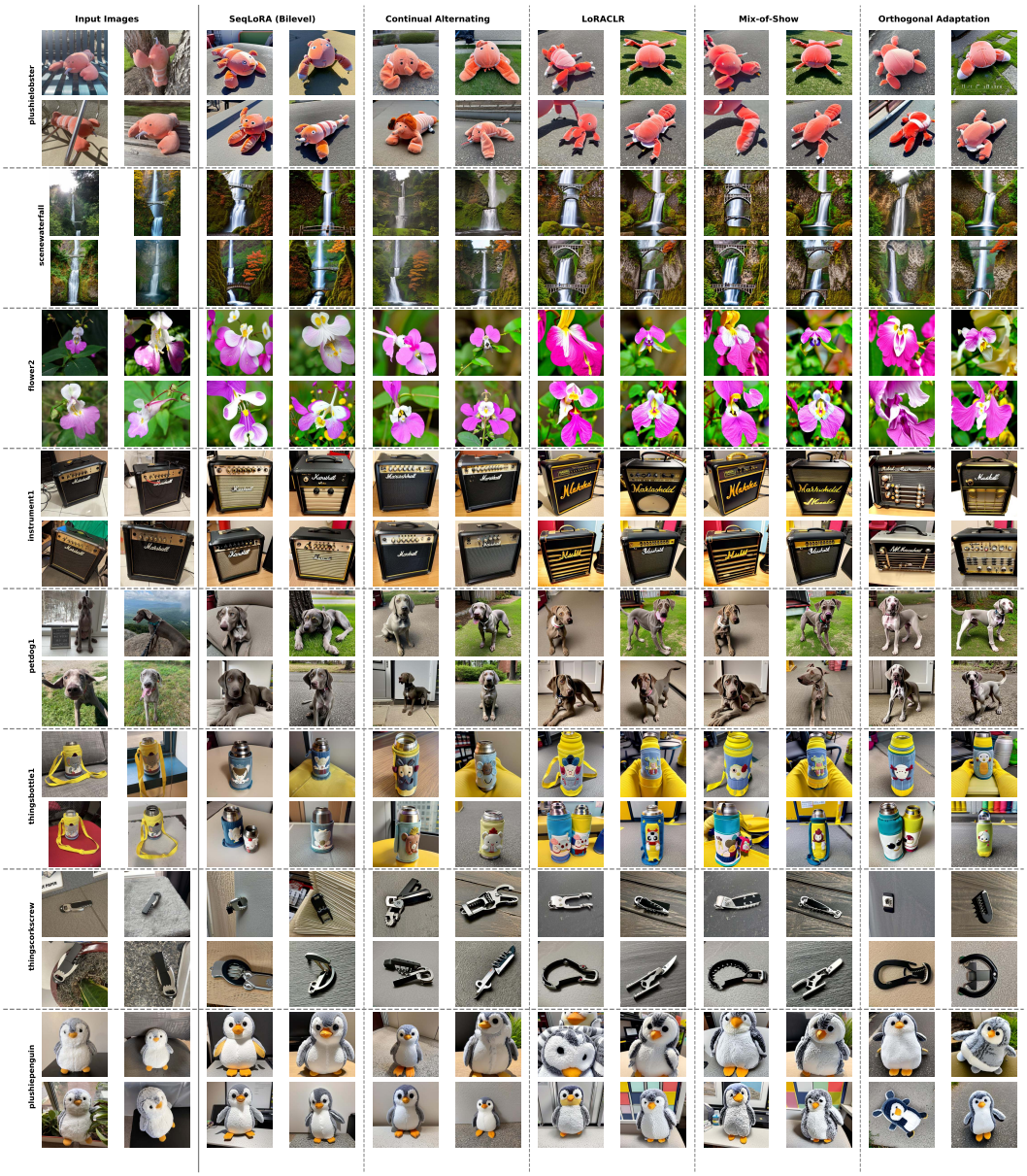


Figure 6: Qualitative comparison for concepts 25–32 (ordered by training sequence); same seeds for all methods.

K Multi-Concept Regional Generation

Figure 7 shows the compositions that underlie Table 2: a three-concept scene combining two chairs, a table and a vase, and a four-concept scene that adds a dog. Every method uses the same random seed and the same regional control inputs (sketch, keypose, and per-region bounding boxes), so visual differences are attributable to the multi-concept mechanism rather than to randomness or spatial conditioning. SeqLoRA preserves per-region identity (chair upholstery, vase silhouette, dog appearance), whereas the fusion-based baselines and Continual Alternating exhibit visible attribute leakage across regions.



(a) **Three concepts.** Prompt: *"two chairs, a table, and a vase, in a living room"*.



(b) **Four concepts.** Prompt: *"a chair, a vase, a dog and a table, in a living room"*.

Figure 7: **Multi-concept regional generation: qualitative comparison.** Each subfigure shows, in the top-left cell, the input concept reference images, followed by five generated outputs (one per method) arranged in a 3×2 grid. All methods use the *same* random seed and the same regional sketch/keypose conditioning, so visual differences reflect the underlying multi-concept fusion mechanism rather than randomness or spatial conditioning. SeqLoRA preserves per-region concept identity (e.g., the chair upholstery, vase silhouette, and dog appearance) more faithfully than the baselines, which exhibit attribute leakage and identity collapse across regions.

L Additional Forgetting Results

Table 12 expands Table 3 with the initial and final scores from which the forgetting values are computed. Table 13 lists the running-mean CLIP-I, CLIP-T, DINO, and DreamSim scores over all concepts learned so far (the quantity plotted in Figure 2b) at steps 1, 8, 16, 24, and 32 of the

Table 12: **Forgetting for SeqLoRA across stream lengths.** *Initial*: score of each concept measured right after it is learned; *Final*: score of the same concept in the final T-concept model; *Forgetting*: Initial − Final (lower is better; negative means improvement). Mean ± standard error over the T concepts.

T	Measure	CLIP-I	CLIP-T	DINO	DINOv2	DINOv3	DreamSim	HPSv2	HPSv3
8	Initial	0.709 ± 0.018	0.286 ± 0.005	0.485 ± 0.047	0.491 ± 0.049	0.468 ± 0.039	0.495 ± 0.040	0.270 ± 0.003	9.513 ± 0.579
	Final	0.695 ± 0.014	0.289 ± 0.005	0.462 ± 0.041	0.465 ± 0.041	0.442 ± 0.030	0.478 ± 0.035	0.271 ± 0.003	9.539 ± 0.588
	Forgetting	0.014 ± 0.006	−0.002 ± 0.001	0.022 ± 0.009	0.025 ± 0.009	0.026 ± 0.010	0.017 ± 0.008	−0.001 ± 0.000	−0.025 ± 0.020
16	Initial	0.703 ± 0.013	0.281 ± 0.004	0.492 ± 0.030	0.436 ± 0.032	0.437 ± 0.027	0.493 ± 0.026	0.268 ± 0.002	9.494 ± 0.416
	Final	0.694 ± 0.012	0.284 ± 0.004	0.474 ± 0.029	0.416 ± 0.028	0.414 ± 0.022	0.478 ± 0.024	0.269 ± 0.002	9.476 ± 0.424
	Forgetting	0.009 ± 0.004	−0.002 ± 0.001	0.018 ± 0.007	0.020 ± 0.007	0.022 ± 0.008	0.014 ± 0.005	−0.000 ± 0.000	0.019 ± 0.029
32	Initial	0.696 ± 0.009	0.288 ± 0.003	0.465 ± 0.022	0.413 ± 0.021	0.420 ± 0.018	0.481 ± 0.017	0.270 ± 0.001	9.044 ± 0.396
	Final	0.687 ± 0.009	0.290 ± 0.003	0.452 ± 0.022	0.398 ± 0.019	0.402 ± 0.017	0.469 ± 0.016	0.270 ± 0.001	9.043 ± 0.399
	Forgetting	0.008 ± 0.003	−0.001 ± 0.000	0.013 ± 0.004	0.016 ± 0.005	0.017 ± 0.005	0.012 ± 0.004	−0.000 ± 0.000	0.001 ± 0.025

32-concept stream, for SeqLoRA and Continual Alternating, with standard errors over the concepts evaluated at each step. Figures 8, 9 and 10 show, for every concept, the generated image at steps 1, 4, 8, 12, 16, 20, 24, 28, and 32, using the same random seed and the basic prompt template ("photo of a <concept>") at every step so that changes reflect the model rather than the sampling.

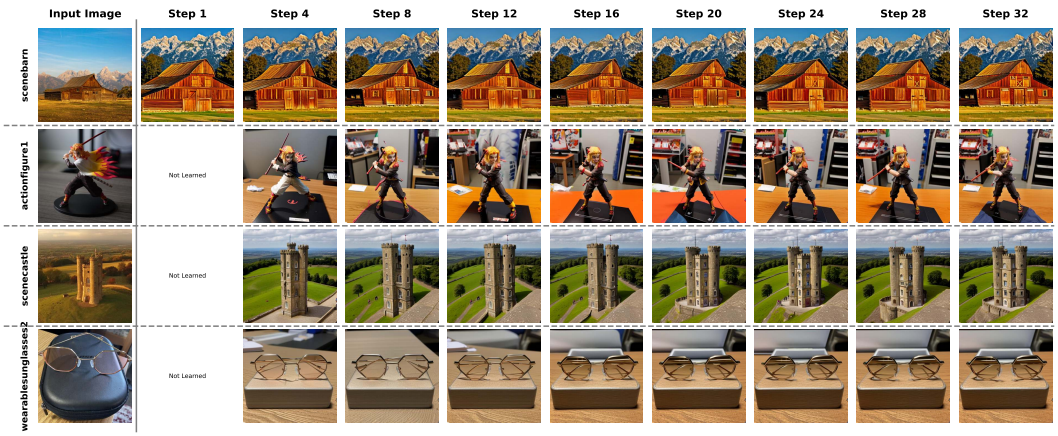


Figure 8: Evolution of generated images for selected concepts across training steps of the 32-concept SeqLoRA run (input image followed by samples at steps 1, 4, 8, 12, 16, 20, 24, 28, and 32).



Figure 9: Evolution of generated images for concepts 1–16 across training steps (same seed and prompt template at every step).

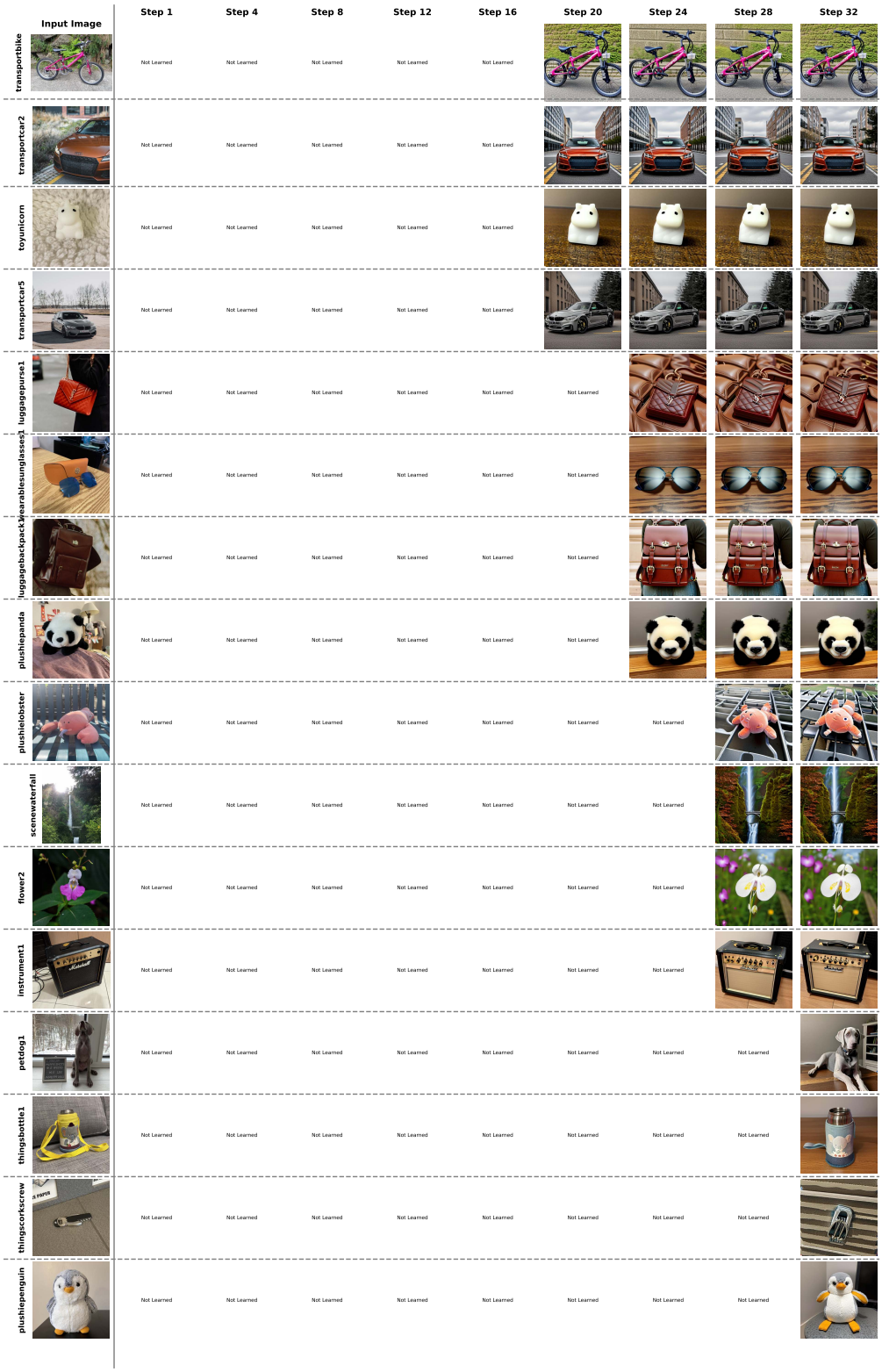


Figure 10: Evolution of generated images for concepts 17–32 across training steps (same seed and prompt template at every step).

Table 13: Summary of forgetting results for 32 concepts across selected steps. We show the average performance across all concepts for SeqLoRA (SeqL) and Continual Alternating (Alt).

Metric	Step 1		Step 8		Step 16		Step 24		Step 32	
	SeqL	Alt	SeqL	Alt	SeqL	Alt	SeqL	Alt	SeqL	Alt
CLIP-I	0.739	0.754	0.694 ± 0.014	0.693 ± 0.012	0.692 ± 0.011	0.682 ± 0.012	0.691 ± 0.009	0.688 ± 0.009	0.687 ± 0.009	0.683 ± 0.008
CLIP-T	0.286	0.280	0.289 ± 0.005	0.289 ± 0.006	0.284 ± 0.004	0.285 ± 0.004	0.288 ± 0.004	0.288 ± 0.004	0.290 ± 0.003	0.290 ± 0.004
DINO	0.627	0.656	0.462 ± 0.043	0.469 ± 0.041	0.473 ± 0.026	0.461 ± 0.028	0.465 ± 0.025	0.463 ± 0.025	0.452 ± 0.022	0.450 ± 0.021
DreamSim	0.600	0.624	0.475 ± 0.035	0.473 ± 0.034	0.475 ± 0.020	0.461 ± 0.021	0.470 ± 0.019	0.465 ± 0.018	0.469 ± 0.016	0.466 ± 0.015

M Full Scalability and Stability Results

Figure 11 shows the scalability curves of Figure 2a for all eight metrics, and Figure 12 the corresponding running-mean stability curves along the 32-concept stream. Table 14 gives the numbers for all metrics and all stream lengths from 8 to 101 concepts; entries marked "–" correspond to out-of-memory failures of the fusion step.

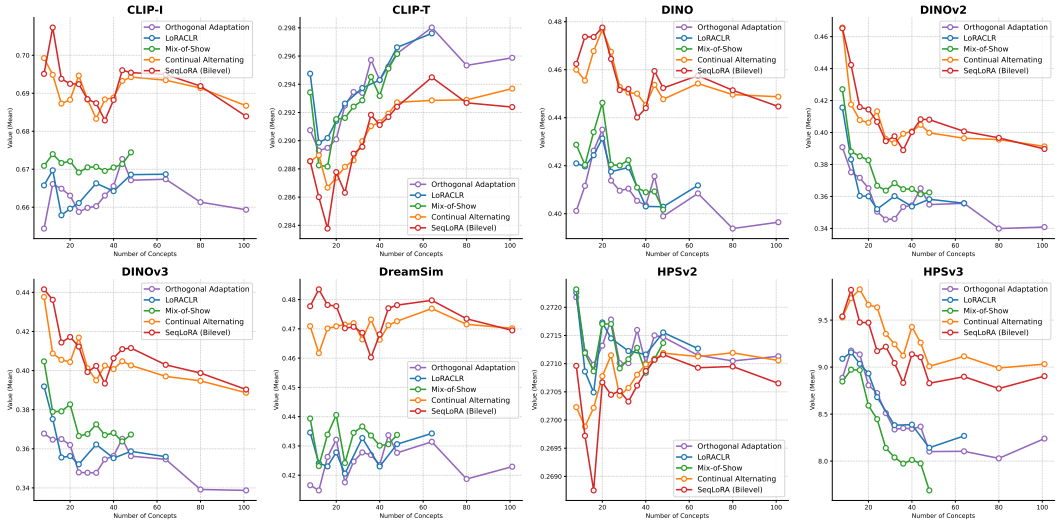


Figure 11: Final-model performance versus number of concepts for all eight metrics (averaged over all T concepts).

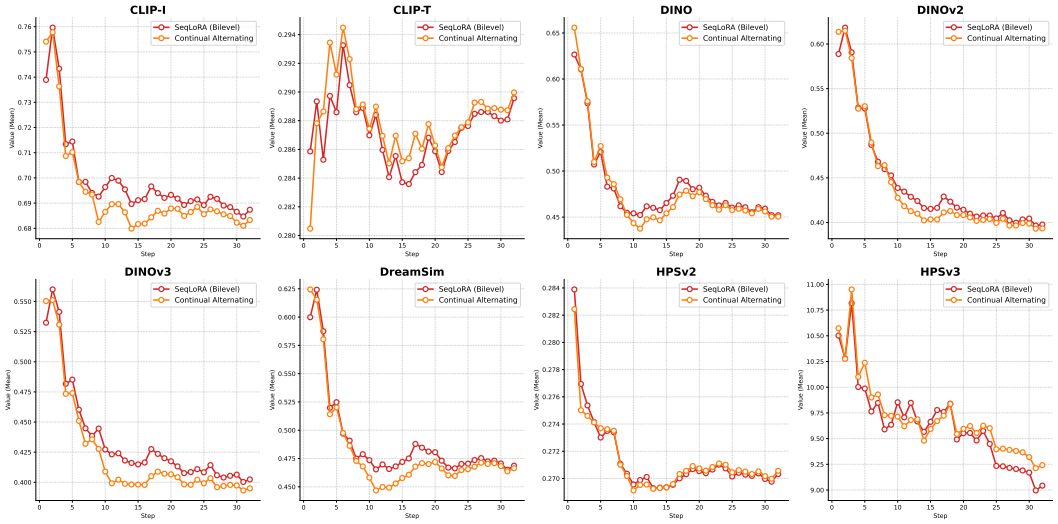


Figure 12: Running mean over all concepts learned so far after each step of the 32-concept stream, for all eight metrics.

Table 14: Supplementary overall comparison results for all metrics across different number of concepts.

Metric	Method	8	16	24	32	40	48	64	80	101
CLIP-I	SeqLoRA (Bilevel)	0.710 ± 0.015	0.704 ± 0.012	0.695 ± 0.010	0.693 ± 0.009	0.707 ± 0.007	0.714 ± 0.007	0.715 ± 0.006	0.711 ± 0.006	0.705 ± 0.005
	Continual Alternating	0.717 ± 0.015	0.696 ± 0.012	0.698 ± 0.009	0.688 ± 0.008	0.705 ± 0.007	0.711 ± 0.007	0.711 ± 0.006	0.708 ± 0.006	0.706 ± 0.005
	Mix-of-Show	0.677 ± 0.018	0.677 ± 0.011	0.677 ± 0.009	0.677 ± 0.010	0.678 ± 0.008	0.681 ± 0.007	-	-	-
	Orthogonal Adaptation	0.664 ± 0.016	0.673 ± 0.011	0.668 ± 0.009	0.671 ± 0.009	0.675 ± 0.008	0.677 ± 0.008	0.677 ± 0.007	0.670 ± 0.007	0.669 ± 0.006
	LoRACLR	0.674 ± 0.017	0.664 ± 0.012	0.669 ± 0.009	0.674 ± 0.010	0.674 ± 0.008	0.678 ± 0.008	0.677 ± 0.007	-	-
CLIP-T	SeqLoRA (Bilevel)	0.287 ± 0.005	0.282 ± 0.004	0.284 ± 0.004	0.286 ± 0.003	0.278 ± 0.003	0.279 ± 0.003	0.280 ± 0.002	0.279 ± 0.002	0.277 ± 0.002
	Continual Alternating	0.287 ± 0.006	0.288 ± 0.004	0.286 ± 0.004	0.287 ± 0.004	0.277 ± 0.003	0.279 ± 0.003	0.278 ± 0.002	0.279 ± 0.002	0.278 ± 0.002
	Mix-of-Show	0.278 ± 0.004	0.273 ± 0.003	0.274 ± 0.003	0.276 ± 0.003	0.277 ± 0.003	0.279 ± 0.002	-	-	-
	Orthogonal Adaptation	0.274 ± 0.004	0.274 ± 0.003	0.275 ± 0.003	0.277 ± 0.003	0.278 ± 0.003	0.279 ± 0.002	0.280 ± 0.002	0.277 ± 0.002	0.278 ± 0.002
	LoRACLR	0.278 ± 0.004	0.273 ± 0.003	0.275 ± 0.003	0.277 ± 0.003	0.278 ± 0.003	0.280 ± 0.002	0.280 ± 0.002	-	-
DINO	SeqLoRA (Bilevel)	0.502 ± 0.041	0.484 ± 0.029	0.474 ± 0.026	0.468 ± 0.021	0.478 ± 0.017	0.488 ± 0.017	0.494 ± 0.014	0.487 ± 0.013	0.484 ± 0.011
	Continual Alternating	0.504 ± 0.040	0.472 ± 0.026	0.477 ± 0.025	0.463 ± 0.020	0.475 ± 0.017	0.482 ± 0.016	0.486 ± 0.015	0.482 ± 0.013	0.485 ± 0.011
	Mix-of-Show	0.440 ± 0.053	0.446 ± 0.028	0.435 ± 0.026	0.436 ± 0.023	0.424 ± 0.018	0.416 ± 0.017	-	-	-
	Orthogonal Adaptation	0.416 ± 0.048	0.440 ± 0.029	0.429 ± 0.026	0.428 ± 0.023	0.420 ± 0.018	0.417 ± 0.016	0.427 ± 0.014	0.410 ± 0.013	0.416 ± 0.011
	LoRACLR	0.435 ± 0.052	0.434 ± 0.032	0.431 ± 0.028	0.434 ± 0.023	0.419 ± 0.018	0.420 ± 0.017	0.429 ± 0.015	-	-
DINOv2	SeqLoRA (Bilevel)	0.509 ± 0.041	0.452 ± 0.028	0.427 ± 0.021	0.418 ± 0.019	0.437 ± 0.015	0.445 ± 0.014	0.439 ± 0.013	0.435 ± 0.012	0.431 ± 0.012
	Continual Alternating	0.515 ± 0.045	0.439 ± 0.022	0.433 ± 0.020	0.412 ± 0.017	0.432 ± 0.014	0.434 ± 0.014	0.429 ± 0.013	0.429 ± 0.012	0.429 ± 0.011
	Mix-of-Show	0.439 ± 0.049	0.394 ± 0.026	0.375 ± 0.019	0.377 ± 0.020	0.375 ± 0.016	0.372 ± 0.015	-	-	-
	Orthogonal Adaptation	0.405 ± 0.040	0.384 ± 0.025	0.361 ± 0.020	0.361 ± 0.020	0.368 ± 0.016	0.369 ± 0.015	0.371 ± 0.013	0.354 ± 0.013	0.358 ± 0.012
	LoRACLR	0.431 ± 0.046	0.367 ± 0.028	0.361 ± 0.022	0.372 ± 0.020	0.367 ± 0.016	0.372 ± 0.015	0.370 ± 0.014	-	-
DINOv3	SeqLoRA (Bilevel)	0.475 ± 0.031	0.439 ± 0.023	0.425 ± 0.018	0.418 ± 0.016	0.440 ± 0.013	0.446 ± 0.013	0.439 ± 0.011	0.434 ± 0.010	0.428 ± 0.011
	Continual Alternating	0.475 ± 0.030	0.427 ± 0.016	0.429 ± 0.018	0.406 ± 0.015	0.431 ± 0.012	0.435 ± 0.013	0.428 ± 0.012	0.426 ± 0.010	0.424 ± 0.010
	Mix-of-Show	0.411 ± 0.035	0.385 ± 0.020	0.371 ± 0.016	0.378 ± 0.017	0.374 ± 0.013	0.372 ± 0.013	-	-	-
	Orthogonal Adaptation	0.379 ± 0.027	0.375 ± 0.020	0.356 ± 0.017	0.359 ± 0.017	0.365 ± 0.013	0.367 ± 0.014	0.366 ± 0.011	0.350 ± 0.011	0.351 ± 0.010
	LoRACLR	0.403 ± 0.032	0.361 ± 0.021	0.359 ± 0.019	0.371 ± 0.017	0.364 ± 0.013	0.369 ± 0.014	0.367 ± 0.012	-	-
DreamSim	SeqLoRA (Bilevel)	0.513 ± 0.035	0.494 ± 0.024	0.478 ± 0.020	0.477 ± 0.016	0.501 ± 0.014	0.512 ± 0.012	0.515 ± 0.010	0.508 ± 0.010	0.507 ± 0.008
	Continual Alternating	0.510 ± 0.033	0.481 ± 0.020	0.480 ± 0.020	0.470 ± 0.015	0.495 ± 0.013	0.504 ± 0.012	0.508 ± 0.011	0.503 ± 0.009	0.504 ± 0.008
	Mix-of-Show	0.447 ± 0.041	0.443 ± 0.022	0.436 ± 0.020	0.448 ± 0.017	0.442 ± 0.014	0.445 ± 0.013	-	-	-
	Orthogonal Adaptation	0.429 ± 0.037	0.438 ± 0.021	0.430 ± 0.019	0.442 ± 0.018	0.437 ± 0.014	0.442 ± 0.013	0.446 ± 0.010	0.432 ± 0.010	0.438 ± 0.008
	LoRACLR	0.445 ± 0.041	0.431 ± 0.024	0.431 ± 0.021	0.445 ± 0.018	0.437 ± 0.014	0.444 ± 0.013	0.447 ± 0.011	-	-
HPSv2	SeqLoRA (Bilevel)	0.271 ± 0.003	0.269 ± 0.002	0.270 ± 0.001	0.270 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001
	Continual Alternating	0.270 ± 0.003	0.270 ± 0.002	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001
	Mix-of-Show	0.272 ± 0.002	0.271 ± 0.002	0.272 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	-	-	-
	Orthogonal Adaptation	0.272 ± 0.003	0.271 ± 0.001	0.272 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001
	LoRACLR	0.272 ± 0.002	0.270 ± 0.002	0.271 ± 0.001	0.271 ± 0.001	0.271 ± 0.001	0.272 ± 0.001	0.271 ± 0.001	-	-
HPSv3	SeqLoRA (Bilevel)	9.539 ± 0.588	9.476 ± 0.424	9.171 ± 0.509	9.043 ± 0.399	9.136 ± 0.283	8.828 ± 0.326	8.898 ± 0.252	8.772 ± 0.231	8.904 ± 0.195
	Continual Alternating	9.529 ± 0.552	9.827 ± 0.328	9.636 ± 0.407	9.242 ± 0.338	9.427 ± 0.303	9.008 ± 0.347	9.115 ± 0.261	8.991 ± 0.228	9.032 ± 0.200
	Mix-of-Show	8.847 ± 0.735	8.968 ± 0.444	8.446 ± 0.503	8.039 ± 0.457	8.013 ± 0.337	7.688 ± 0.314	-	-	-
	Orthogonal Adaptation	8.886 ± 0.788	9.135 ± 0.431	8.723 ± 0.413	8.336 ± 0.432	8.343 ± 0.312	8.101 ± 0.314	8.104 ± 0.290	8.029 ± 0.244	8.240 ± 0.205
	LoRACLR	9.089 ± 0.722	9.035 ± 0.400	8.681 ± 0.432	8.382 ± 0.431	8.388 ± 0.304	8.142 ± 0.308	8.268 ± 0.270	-	-