# Low-Rank Embedding Adaptation for Models with Expanding Vocabularies

**Anonymous Authors**[1]

## Abstract

Pretrained models are routinely extended with new vocabulary entries, for example new items in recommenders and new tokens in LLMs. We show that jointly training old and new embeddings has a hidden failure mode: old-entry quality degrades while new entries are still learning. On sequential recommendation, old items overfit while new items improve, forcing premature early-stopping; on LLMs, base-token perplexity rises within 1–2 epochs. We propose *population-specific low-rank subspaces* that extend the low-rank adaptation principle to embedding tables: each entry is parameterized as $e_i = \bar{e}_i + u_i V_k^\top$, where $V_k$ is shared within a population and separate across populations with different data regimes. The shared $V_k$ provides implicit regularization for sparse new entries while the separation prevents gradient interference between populations. Three instantiations (Freeze-SV, Freeze1-SV, Dual-SV) cover different regimes. On sequential recommendation (SASRec and GRU4Rec architectures; Taobao e-commerce and MerRec marketplace datasets), our methods Pareto-dominate joint fine-tuning and continual-learning baselines. On LLM vocabulary expansion (GPT-2, Pythia-410M, Pythia-1.4B on BillSum legal and PubMed medical domains), our low-rank methods improve overall perplexity over joint training in 8 out of 9 model-scale cells, outperforming full-rank frozen baselines in 8/9 cells at $2\times$ compression ($r/d$=50%) and 6/9 at $4\times$ compression ($r/d$=25%).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

## 1. Introduction

Low-rank adaptation (LoRA) (Hu et al., 2022) has become the standard for parameter-efficient fine-tuning: decompose weight updates as $\Delta W = BA$ with $r \ll \min(d, k)$, and train only the low-rank factors. This works because transformer weight matrices concentrate most model parameters in LLMs.

But the parameter distribution inverts in other architectures. In sequential recommenders like SASRec (Kang & McAuley, 2018) serving millions of items, **99.9% of parameters are in the embedding table** and only 0.1% in the transformer. LoRA adapts the 0.1% and leaves the dominant block frozen. Even in LLMs, vocabulary expansion (adding domain-specific tokens) creates new embedding rows that LoRA cannot reach.

We propose *population-specific low-rank embedding subspaces* – extending the low-rank adaptation principle from weight matrices to embedding tables. The factorization mirrors LoRA's structure but operates per-entry:

$$e_i = \bar{e}_i + u_i \cdot V_k^\top, \quad u_i \in \mathbb{R}^r,\ V_k \in \mathbb{R}^{d \times r} \tag{1}$$

where $V_k$ is shared within a population (analogous to LoRA's shared $B, A$) and $u_i$ provides per-entry coordinates (analogous to per-input activations). Combined with LoRA on the transformer, this gives a **complete low-rank adaptation stack** covering both weight matrices and embedding tables.

A key challenge in embedding adaptation is that vocabularies expand incrementally: new entries arrive with sparse data alongside well-established entries. Naïve low-rank adaptation of the full table fails because established-entry gradients dominate the shared subspace. We show that population-specific $V_k$ matrices – separate projections for base and new entries (where "populations" are groups of entries with different data regimes) – are essential, resolving an anti-correlation where base quality degrades while new entries improve (Figure 1).

Three instantiations differ in base-population handling:

1. **Freeze-SV**: base frozen throughout; only new-entry parameters learned. The simplest variant; most competitive

on well-pretrained LLMs where base representations are already strong.

2. **Freeze1-SV**: one epoch with all parameters trainable, then base frozen. Epoch-robust (trains 20 epochs without base degradation). Wins on seqrec and smaller LLMs.
3. **Dual-SV**: separate subspaces for base and new at different learning rates. Best new-entry quality under early stopping on seqrec; competitive with Freeze-SV on LLMs.

**Contributions.** (1) We extend low-rank adaptation from weight matrices to embedding tables, providing parameter efficiency in embedding-heavy models and preventing base overfitting during LLM vocabulary expansion. (2) We show population-specific structure is necessary when vocabularies expand incrementally, with a convergence analysis ($O(d/r)$ improvement). (3) We demonstrate Pareto-dominance on two seqrec datasets and improvement in 8/9 LLM cells over joint training.

## 2. Related Work

**Low-rank adaptation of weight matrices.** LoRA (Hu et al., 2022) decomposes weight updates as $\Delta W = BA$, training only low-rank factors. GaLore (Zhao et al., 2024) projects gradients onto a low-rank subspace; QLoRA (Dettmers et al., 2023) combines quantization with LoRA. All target weight matrices; none address the embedding table. SD-LoRA (Wu et al., 2025) uses separate LoRA branches for old/new tasks – the closest structural precedent, but on weight matrices for task-level separation rather than embedding tables for population-level separation.

**Incremental recommendation.** EWC (Kirkpatrick et al., 2017) adds Fisher-weighted regularization; ADER (Mi et al., 2020) uses replay with distillation. Both focus on preventing forgetting but do not examine the base-vs-new interaction during joint training. Fatkulin et al. (Fatkulin et al., 2025) propose bounded-delta embeddings but do not prevent base degradation (our reimplementation: $-5.8$pp base NDCG over 20 epochs).

**LLM vocabulary expansion.** Cui et al. (Cui et al., 2023b) and Yamaguchi et al. (Yamaguchi et al., 2024) study token initialization but do not identify that joint training degrades base-token quality during vocabulary expansion.

**Embedding compression.** Mixed Dimension Embeddings (Ginart et al., 2021) assign different dimensions by frequency; FIITED (Lian et al., 2021) prunes per-embedding dimensions. Both are static techniques, not designed for incremental settings where new entries arrive post-deployment.

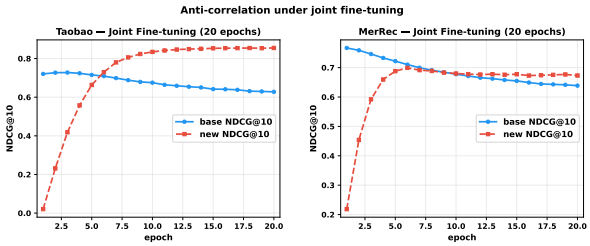


*Figure 1.* Anti-correlation under joint fine-tuning (W0, 20 epochs without early stopping). Blue: base NDCG@10 (drops). Red: new NDCG@10 (rises). Left: Taobao. Right: MerRec.

## 3. The Anti-Correlation Problem

We formalize the setting: a pretrained model with embedding table $\bar{E}$ is extended with new entries and fine-tuned on adaptation data containing both populations. The joint objective $\mathcal{L} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{new}}$ shares encoder parameters $\theta$.

**Sequential recommendation.** On SASRec (Kang & McAuley, 2018) (Taobao, MerRec), base-only NDCG@10 drops 13% (Taobao) and 16% (MerRec) over 20 epochs while new NDCG rises continuously (Figure 1). Our methods hold base within $\pm 1\%$.

*Table 1.* Training dynamics (W0, 20 epochs, no early stopping). Base-only $\Delta$ = base-only NDCG change ep1$\rightarrow$20. FT degrades base; our methods preserve it.

| Dataset | Method | base $\Delta$ | new $\Delta$ | peak ovrl | ep |
|---|---|---|---|---|---|
| Taobao | FT | $-0.093$ | $+0.834$ | 0.729 | 3 |
| | Freeze1-SV | $+0.001$ | $+0.537$ | 0.730 | 19 |
| | Dual-SV | $-0.013$ | $+0.538$ | 0.733 | 7 |
| MerRec | FT | $-0.128$ | $+0.455$ | 0.751 | 3 |
| | Freeze1-SV | $-0.009$ | $+0.080$ | 0.768 | 18 |
| | Dual-SV | $-0.021$ | $+0.142$ | 0.767 | 2 |

Table 1 quantifies the tradeoff: FT peaks on overall NDCG at epoch 3 (base dominates the metric) while new items are still at 0.02–0.42 NDCG. Our methods achieve higher peak overall NDCG at later epochs because base quality is preserved, allowing the overall metric to benefit from new-item improvement.

**LLM vocabulary expansion.** On GPT-2 and Pythia (410M, 1.4B) expanded with 2000 domain tokens (BillSum, PubMed), base perplexity rises within 1–2 epochs under joint training. Base-freezing wins on overall perplexity in 8 of 9 cells (Figure 4).

## 4. Method

### 4.1. Framework

The general formulation (Eq. 1) has three properties: (1) gradient decoupling via separate $V_k$; (2) implicit regularization via shared $V_k$ within each population; (3) parameter efficiency ($r$ params per new entry, $4\times$ compression at $r/d = 25\%$). Figure 2 illustrates the end-to-end pipelines for both sequential recommendation and LLM vocabulary expansion.

The structural analogy to LoRA is precise: LoRA decomposes $\Delta W = BA$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are shared across all inputs. Our factorization decomposes $\Delta E = UV$ where $V \in \mathbb{R}^{d\times r}$ is shared across entries (analogous to LoRA's shared structure) and $U \in \mathbb{R}^{n\times r}$ provides per-entry coordinates (analogous to per-input activations flowing through LoRA). The key difference: LoRA's $\Delta W$ is the same for every input, while our $\Delta e_i = u_i V^\top$ is different per entry.

### 4.2. Freeze-SV (No Warmup)

The simplest variant: base embeddings are frozen from epoch 0 and never updated. New entries are parameterized as $e_j = \bar{e}_j + u_j V^\top$ with shared $V$. Most effective on LLM vocabulary expansion where pretrained base representations are already strong and even one epoch of joint training starts overfitting base tokens.

### 4.3. Freeze1-SV (Warmup + Freeze)

Trains all parameters for one warmup epoch (encoder observes joint distribution), then freezes base embeddings from epoch 2 onward. New entries: $e_j = \bar{e}_j + u_j V^\top$. Epoch-robust: trains 20 epochs without base degradation (<0.01 NDCG drift). Dominates on sequential recommendation where the encoder is small relative to the embedding table.

**Algorithm 1** Freeze1-SV Training (per window)

**Input:** Pretrained $\bar{E}$, new entries $\{j\}$, rank $r$
Initialize anchors: $\bar{e}_j \leftarrow$ task-dependent (mean of base embeddings, content embedding, or multivariate normal)
Init: $u_j \leftarrow \mathbf{0}$, $V \leftarrow \text{Xavier}(r, d)$; embedding: $e_j = \bar{e}_j + u_j V^\top$
**Epoch 1 (warmup):** Train all ($\bar{E}$, $\{u_j\}$, $V$, encoder + LoRA)
**Epoch 2+:** Freeze $\bar{E}$ and anchors $\{\bar{e}_j\}$; train $\{u_j\}$, $V$, encoder
**Return:** Updated model for next window

### 4.4. Dual-SV (Separate Subspaces)

Separate subspaces for base and new at different learning rates:

$$e_i^{\text{base}} = \bar{e}_i + u_i^{\text{base}} (V^{\text{base}})^\top,\ \text{lr} = 0.1 \times \text{lr}_{\text{new}} \quad (2)$$

$$e_j^{\text{new}} = \bar{e}_j + u_j^{\text{new}} (V^{\text{new}})^\top,\ \text{lr} = \text{lr}_{\text{new}} \quad (3)$$

Base entries receive a controlled low-rank delta at $10\times$ lower learning rate, allowing mild co-adaptation without catastrophic drift. This achieves the best new-entry quality at the best-overall epoch under standard early stopping, because the encoder can co-adapt to serve both populations simultaneously.

**Algorithm 2** Dual-SV Training (per window)

**Input:** Pretrained $\bar{E}$, new entries $\{j\}$, rank $r$
Initialize anchors: $\bar{e}_j \leftarrow$ task-dependent (same as Freeze1-SV)
Base anchors: $\bar{e}_i \leftarrow$ pretrained embeddings from $\bar{E}$
Init: $u_j^{\text{new}} \leftarrow \mathbf{0}$, $V^{\text{new}} \leftarrow \text{Xavier}(r, d)$
Init: $u_i^{\text{base}} \leftarrow \mathbf{0}$, $V^{\text{base}} \leftarrow \text{Xavier}(r, d)$
Set $\text{lr}_{\text{base}} = 0.1 \times \text{lr}_{\text{new}}$
**for** each epoch **do**
  Train $\{u_j^{\text{new}}\}$, $V^{\text{new}}$ at $\text{lr}_{\text{new}}$
  Train $\{u_i^{\text{base}}\}$, $V^{\text{base}}$ at $\text{lr}_{\text{base}}$
  Train encoder (+ LoRA)
  Stop if val-overall metric degrades
**end for**
**Return:** Updated model for next window

### 4.5. Trainable Parameters

For a window with $n_b$ active base items and $n_n$ new items at rank $r$, Dual-SV trains:

$$(n_b + n_n) \cdot r + 2rd + \text{LoRA params} \quad (4)$$

On MerRec ($n_b$=1.5M, $n_n$=1.5M, $r$=32, $d$=150): 106M embedding params vs. 678M for full FT ($6.4\times$ reduction). Each item learns only $r$ parameters regardless of $d$. Freeze1-SV is even cheaper: only new items have trainable $u_j$ ($n_n \cdot r + rd$).

### 4.6. Theoretical Motivation

**Proposition 1** (Shared-$V$ error bound)**.** *Consider $n$ entries, each observed $k_j$ times with noise variance $\sigma^2$. Under approximate low-rank structure ($\|e_j^* - u_j^* V^*\| \le \tau$), the shared-$V$ estimator achieves per-entry error $O(r\sigma^2/k_j)$ vs. $O(d\sigma^2/k_j)$ for independent estimation – an $O(d/r)$ improvement.*

*Proof sketch.* The shared $V$ is estimated from all $N = \sum_j k_j$ observations jointly. Once $V$ is known, each entry estimates only its $r$-dimensional coordinate $u_j$ from $k_j$

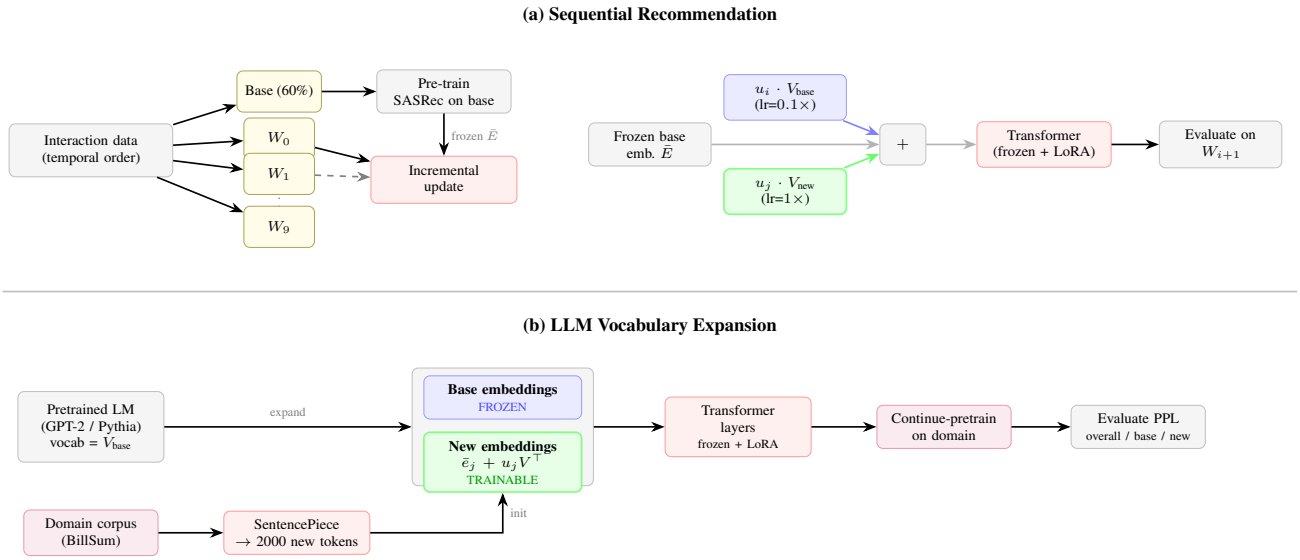


*Figure 2.* End-to-end pipelines. **(a)** Sequential recommendation: frozen base embeddings with separate low-rank corrections ($V_{\text{base}}$, $V_{\text{new}}$), LoRA on attention. **(b)** LLM vocabulary expansion: new rows parameterized as $\bar{e}_j + u_j V^\top$, base frozen, LoRA on all layers.

samples, reducing variance by factor $d/r$ compared to estimating the full $d$-dimensional embedding independently. The condition for this to help is $N \gg rd/(d-r)$, which is satisfied by orders of magnitude on both datasets (Taobao: $N \approx 10^7$, threshold $\approx 873$).

The bound predicts: (P1) advantage concentrated in sparse entries; (P2) diminishing returns as per-entry data grows; (P3) a data-scale crossover where full-rank becomes preferable. P1 follows from the bound structure ($r/k_j$ dominates when $k_j$ is small) and is consistent with our new-item results (median $k_j = 1$). P2 is confirmed by rank ablations. P3 is confirmed by the GPT-2 15K exception.

## 5. Experiments

### 5.1. Sequential Recommendation

**Setup.** SASRec ($d$=150, 2 layers). Taobao (Zhu et al., 2018) (91M interactions, 1.2M items, 2h windows) and MerRec (Sato et al., 2025) (280M interactions, 4.5M items, 1d windows). 10-window continual evaluation, 20 epochs/window, sampled-100 protocol (Figure 2a). These datasets represent two turnover regimes: Taobao has 10% new items per window (base-dominated validation), MerRec has 61% (new-dominated).

**Baselines.** (1) Stale: no incremental training. (2) FT: all parameters trainable. (3) EWC (Kirkpatrick et al., 2017): Fisher-weighted regularization. (4) ADER (Mi et al., 2020): experience replay with distillation.

**Our methods.** Freeze1-SV and Dual-SV ($r$=128) with LoRA (rank 8) on Q/K/V attention projections. Combined: embedding low-rank + transformer low-rank = full-model low-rank adaptation.

*Table 2.* 10-window continual: mean NDCG@10 (val-overall ES).

| | Taobao | | MerRec | |
|---|---|---|---|---|
| Method | new | base | new | base |
| FT | 0.067 | 0.717 | 0.661 | 0.742 |
| EWC | 0.063 | 0.717 | 0.671 | 0.748 |
| ADER | 0.147 | 0.712 | 0.660 | 0.753 |
| Freeze1-SV | 0.071 | 0.717 | 0.659 | 0.754 |
| Dual-SV | **0.624** | 0.709 | **0.711** | 0.754 |

Table 2: Under val-overall ES, Dual-SV achieves 0.624 new NDCG on Taobao (4× ADER) because its decoupled optimization allows the overall metric to keep improving. Freeze1-SV's advantage is epoch-robustness: without ES it reaches 0.770 new NDCG with stable base (0.728).

**Pareto dominance** (Figure 3): On W0 without ES, Freeze1-SV and Dual-SV occupy the top-right quadrant (high base AND high new), dominating every point on FT's trajectory. Without ES, FT reaches 0.855 new NDCG but base collapses to 0.632 (−12%).

**GRU4Rec (architecture generality).** We verify on GRU4Rec (Hidasi et al., 2016) (RNN-based, no attention). The same pattern holds: Freeze1-SV trains longer with higher overall NDCG (Taobao: 0.833 vs. FT 0.820; MerRec: 0.920 vs. 0.890). The phenomenon is not specific to transformers.

**Rank ablation** (Table 3): At $r$=64 (43%), Dual-SV matches full FT on new quality (0.863 vs. 0.867) while

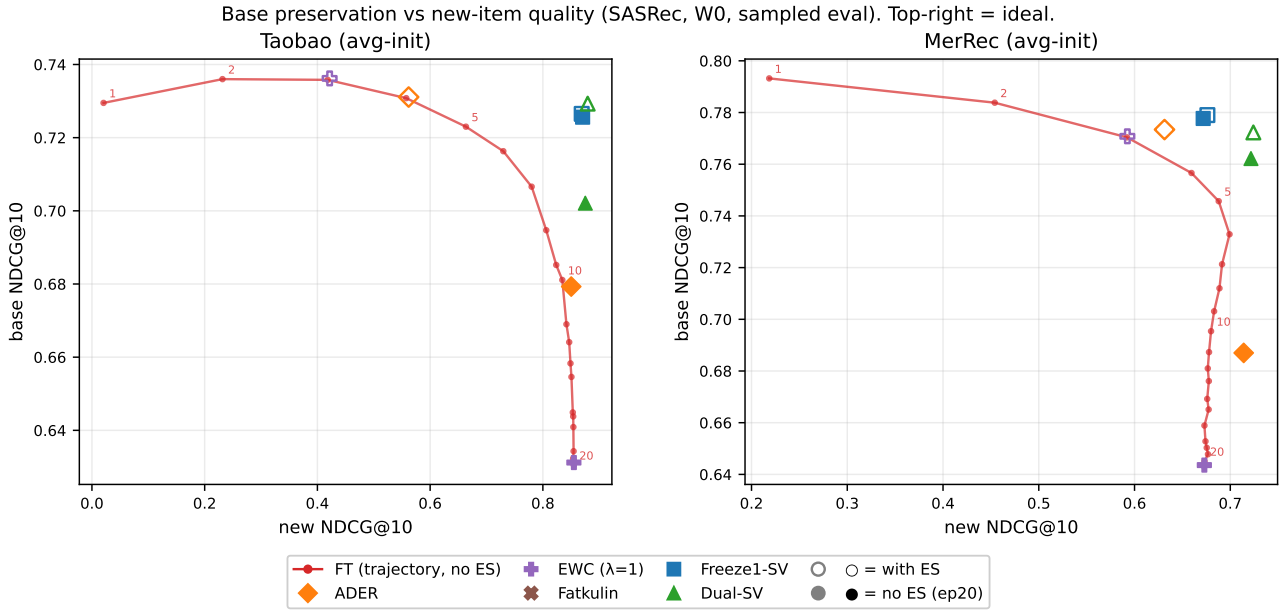


*Figure 3.* Pareto comparison (W0, 20 epochs without early stopping). Gray trajectory: FT epochs 1–20. Our methods (Freeze1-SV = blue, Dual-SV = green) Pareto-dominate every point on FT's trajectory. Left: Taobao 2h. Right: MerRec 1d.

preserving base. At $r$=128 (85%), it exceeds FT (0.872) – the rank constraint provides regularization even at high capacity.

*Table 3.* Rank ablation (Taobao W0, no ES). Higher rank improves new quality with diminishing returns; base is stable.

| Variant | $r$ | $r/d$ | new | base |
|---|---|---|---|---|
| Dual-SV | 16 | 11% | 0.604 | 0.721 |
| Dual-SV | 32 | 21% | 0.794 | 0.718 |
| Dual-SV | 64 | 43% | 0.863 | 0.718 |
| Dual-SV | 128 | 85% | 0.872 | 0.722 |
| Full FT | $d$ | 100% | 0.867 | 0.690 |

**Content-initialized embeddings.** When new items have content features (E5 (Wang et al., 2024) text embeddings), the anchor $\bar{e}_j$ can be set to the content embedding rather than the base mean. On MerRec with content initialization (Table 4), lower rank suffices: Dual-SV $r$=32 achieves 0.849 new NDCG, outperforming $r$=128 (0.832), because the content anchor already provides a good starting point. The Fatkulin (Fatkulin et al., 2025) bounded-delta baseline does not prevent base degradation ($-5.8$pp).

*Table 4.* Content-init embeddings on MerRec (W0, best-overall-NDCG epoch). Anchor = E5 content embedding. Δbo = base NDCG change ep1→20.

| Method | rank | ep | overall | base | new | Δbo |
|---|---|---|---|---|---|---|
| FT (content) | full | 1 | 0.804 | 0.807 | 0.784 | −0.074 |
| Fatkulin ($\delta$) | full | 2 | 0.803 | 0.801 | 0.813 | −0.058 |
| Freeze1-SV | 128 | 9 | 0.817 | 0.819 | 0.800 | +0.016 |
| Freeze1-SV | 32 | 6 | 0.810 | 0.804 | 0.861 | −0.002 |
| Dual-SV | 128 | 6 | **0.822** | 0.820 | 0.832 | +0.006 |
| Dual-SV | 32 | 7 | 0.820 | 0.816 | **0.849** | −0.000 |

### 5.2. LLM Vocabulary Expansion

**Setup.** GPT-2 (Radford et al., 2019) (124M, $d$=768), Pythia-410M and Pythia-1.4B (Biderman et al., 2023) ($d$=1024, $d$=2048). 2000 new tokens via SentencePiece on BillSum (Kornilova & Eidelman, 2019) legal text. Continue-pretrain on 1K/5K/15K documents (10–20 epochs, Figure 2b). LoRA rank 8 on all linear layers. New-token embeddings initialized via multivariate normal matching the base embedding distribution.

**Baselines.** (1) Basetrn: base embeddings trainable, full-dim new rows (Cui et al. (2023b) Stage-1 recipe). (2) Full-rank frozen: base frozen, full-dim new rows.

**Our methods.** Freeze-SV, Freeze1-SV, and Dual-SV at $r/d \approx 50\%$.

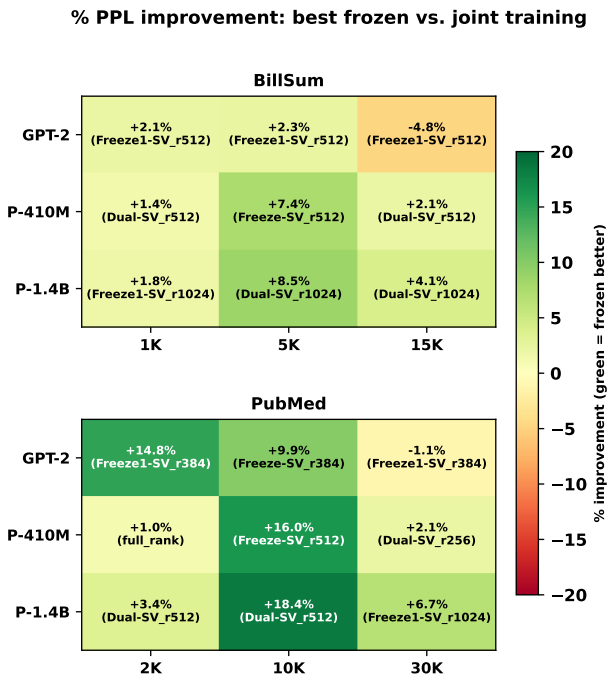


*Figure 4.* Best frozen method vs. joint training on BillSum (top) and PubMed (bottom). Green = frozen better. Base-freezing wins in 8/9 cells on both domains.

*Table 5.* BillSum: best-epoch overall PPL (↓). Bold = best in row.

| Model | Scale | basetrn full_ft | frozen_base full_ft | freeze_sv | freeze1_sv | dual_v |
|---|---|---|---|---|---|---|
| GPT-2 | 1K | 6.40 | 6.42 | 6.39 / 6.49 | **6.30** / 6.40 | 6.55 / 6.65 |
| | 5K | 5.95 | 5.98 | 5.96 / 6.04 | **5.84** / 5.90 | 5.94 / 6.02 |
| | 15K | **4.79** | 5.27 | 5.26 / 5.29 | 5.02 / 5.05 | 5.11 / 5.18 |
| P-410M | 1K | 4.85 | 4.80 | 4.80 / 4.81 | 4.79 / 4.81 | **4.78** / 4.80 |
| | 5K | 5.06 | **4.69** | **4.69** / 4.71 | 4.83 / 4.84 | **4.69** / 4.70 |
| | 15K | 4.13 | 4.06 | 4.06 / 4.06 | 4.08 / 4.09 | **4.04** / 4.05 |
| P-1.4B | 1K | 4.14 | 4.09 | 4.09 / 4.10 | **4.07** / **4.07** | **4.07** / 4.09 |
| | 5K | 4.40 | 4.04 | **4.03** / 4.04 | 4.09 / 4.11 | **4.03** / 4.04 |
| | 15K | 3.67 | 3.54 | 3.54 / 3.54 | 3.56 / 3.56 | **3.52** / 3.53 |

$r/d$: 50% / 25%. GPT-2 $r$=384/192; P-410M $r$=512/256; P-1.4B $r$=1024/512.

Table 5: Base-freezing wins in 8/9 cells (Figure 4). The exception (GPT-2 15K) occurs when the small model has sufficient data to tolerate joint training. At both $r/d = 50\%$ and 25%, our low-rank methods match full-rank frozen within 0.01–0.02 PPL, confirming the rank constraint does not hurt while providing 2–4× parameter efficiency. (All LLM results are single-run; differences <0.05 PPL between frozen variants may not be significant.)

At $r/d = 50\%$, our low-rank methods match or slightly beat full-rank frozen (e.g., Dual-SV 4.03 vs. 4.04 on Pythia-1.4B 5K). At $r/d = 25\%$, quality remains within 0.01–0.02 PPL at 4× fewer parameters. The improvement over joint training comes from base preservation: basetrn degrades base-token PPL by 7–18% while our frozen methods prevent this, winning on overall PPL (Figure 5).

**PubMed replication.** We replicate on PubMed medical abstracts (U.S. National Library of Medicine, 2024) (token-matched scales: 2K/10K/30K documents) and observe the same 8/9 pattern (Table 6).

*Table 6.* PubMed: best-epoch overall PPL (↓). Bold = best in row.

| Model | Scale | basetrn full_ft | frozen_base full_ft | freeze_sv | freeze1_sv | dual_v |
|---|---|---|---|---|---|---|
| GPT-2 | 2K | 51.6 | 45.6 | **44.0** / 44.2 | **44.0** / 44.1 | 48.0 / 46.1 |
| | 10K | 41.4 | 37.7 | **37.3** / 37.6 | 37.5 / 37.7 | 41.7 / 40.1 |
| | 30K | **29.8** | 31.3 | 31.1 / 31.5 | 30.1 / 30.6 | 32.0 / 32.3 |
| P-410M | 2K | 23.8 | **23.5** | 23.6 / 23.8 | 23.6 / 23.8 | 23.7 / 23.7 |
| | 10K | 26.1 | **21.9** | **21.9** / 22.1 | 22.1 / 22.4 | 22.3 / 22.0 |
| | 30K | 14.0 | **13.8** | 14.1 / 13.9 | 14.0 / 13.8 | 13.9 / **13.7** |
| P-1.4B | 2K | 17.5 | 17.4 | **17.0** / 17.1 | **17.0** / 17.1 | 17.1 / **16.9** |
| | 10K | 22.6 | 19.3 | 19.2 / 18.8 | 19.4 / 19.2 | **19.0** / **18.4** |
| | 30K | 10.7 | 10.1 | 10.1 / 10.1 | **10.0** / 10.1 | 10.3 / 10.1 |

$r/d$: 50% / 25%. GPT-2 $r$=384/192; P-410M $r$=512/256; P-1.4B $r$=1024/512.

## 6. Analysis

**Why separate V? (SASRec, Taobao).** A single shared $V$ for both populations achieves only 0.293 new NDCG vs. 0.872 for Freeze1-SV (Table 7). Base gradients dominate the shared subspace, preventing new items from learning useful directions. The population-specific structure is essential, not merely beneficial.

*Table 7.* Ablations (Taobao W0, no ES, best epoch). All use LoRA on attention.

| Variant | new NDCG | base NDCG |
|---|---|---|
| LoRA only (frozen emb.) | 0.000 | 0.710 |
| LoRA + full emb. update | 0.867 | 0.690 |
| Shared-V (single $V$, all items) | 0.293 | 0.721 |
| Freeze1-SV ($r$=128) | 0.872 | 0.722 |
| Dual-SV ($r$=128) | 0.863 | 0.718 |

**LoRA alone fails (SASRec, Taobao).** Frozen embeddings + LoRA on attention gives zero new-item quality: the embedding delta is the critical component. LoRA provides complementary encoder adaptation but cannot substitute for embedding-level learning.

**Architecture generality (SASRec + GRU4Rec).** On GRU4Rec (Hidasi et al., 2016) (RNN-based, no attention), the same pattern holds: Freeze1-SV achieves 0.833 overall NDCG vs. FT's 0.820 on Taobao; 0.920 vs. 0.890 on MerRec.

**V-initialization (SASRec, Taobao/MerRec).** The shared projection $V$ can be initialized from PCA of the base embedding table (providing a meaningful starting subspace) or Xavier random. On Taobao, PCA outperforms Xavier

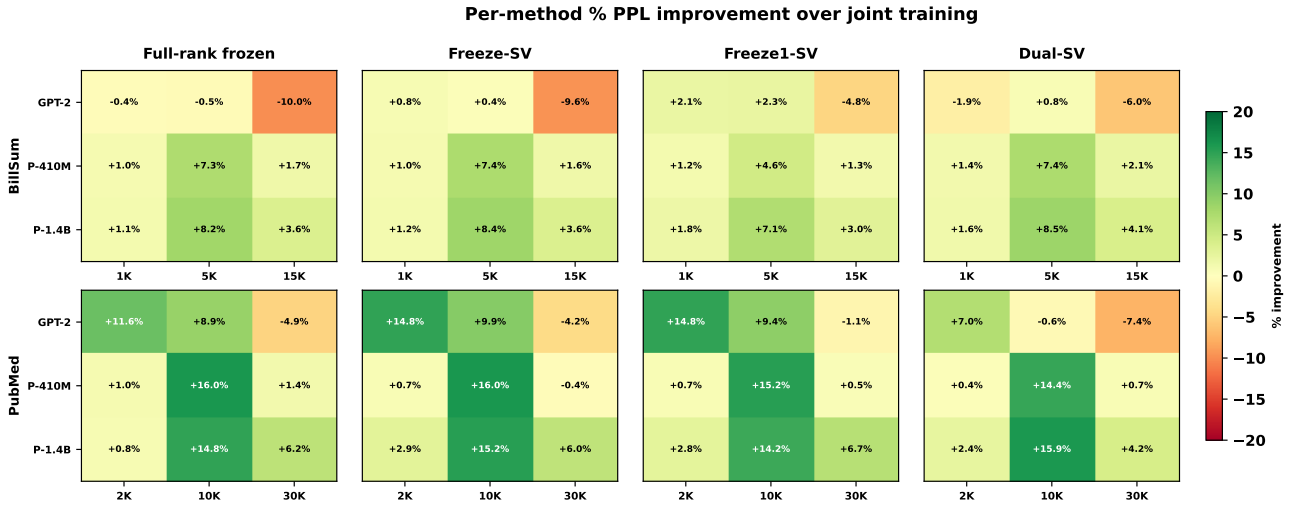


*Figure 5.* Per-method PPL improvement over joint training at $r/d \approx 50\%$. Top row: BillSum. Bottom row: PubMed. Ranks: GPT-2 $r$=384; P-410M $r$=512; P-1.4B $r$=1024.

by $+0.8$pp NDCG; on MerRec, Xavier outperforms PCA by $+1.9$pp. Zero initialization fails completely (model cannot learn useful directions in few epochs). The sensitivity is mild: both non-zero initializations converge to similar quality by epoch 10, suggesting $V$ learns task-appropriate directions regardless of starting point when given sufficient training signal.

**Rank selection (LLM vocabulary expansion).** $r/d \geq 50\%$ matches or beats full-rank (regularization helps); $\approx 25\%$ competitive with multivariate anchors (within 0.01 PPL); $< 15\%$ underpowered. The ratio $r/d$ is the relevant quantity, not absolute $r$.

**The parameter inversion (cross-architecture).** In LLMs (GPT-2: 124M params, 31% in embeddings), most parameters are in transformer layers – LoRA's target. In recommendation (SASRec with 4.5M items: 99.9% in embeddings), the ratio inverts; here our method both adapts the dominant block and provides parameter efficiency ($4\times$ compression at $r/d = 25\%$). In LLM vocabulary expansion, new embedding rows are a small fraction of total parameters, so parameter efficiency is a side benefit; the primary value is preventing base-token overfitting that joint training causes within 1–2 epochs.

**Limitations.** Our LLM experiments reach 1.4B parameters; verification at 7B+ is future work. We evaluate on two LLM domains (BillSum legal text, PubMed medical abstracts) and two seqrec datasets (Taobao 91M interactions, MerRec 280M interactions). The anti-correlation persists across LLM scales (124M to 1.4B) and seqrec architectures (SASRec, GRU4Rec), but our seqrec encoder is small (2 layers, $d$=150); whether larger seqrec encoders reduce the effect is untested. On GPT-2 at 15K documents, joint training wins because small models with abundant data tolerate the anti-correlation. On knowledge graph completion (TransE (Bordes et al., 2013) on ENTITY (Cui et al., 2023a)), the anti-correlation does not manifest and low-rank hurts because TransE requires precise pairwise distances without a shared encoder. Our seqrec evaluation uses sampled-100 protocol; base-only evaluation confirms method ordering. Results are single-run for LLMs (differences <0.05 PPL between frozen variants may not be significant).

## 7. Discussion

**LoRA for embeddings vs. LoRA for weights.** The structural parallel between our method and LoRA is precise but the design considerations differ. In LoRA, the shared $\Delta W = BA$ applies identically to every input token. In our factorization, the shared $V$ defines the subspace but each entry has its own coordinate $u_i$ – making it closer to matrix completion than to adapter-based PEFT. This per-entry structure is necessary because embedding tables store per-entity information, not shared transformations.

**When does population-specific structure help?** The separation into population-specific $V_k$ is essential when: (1) populations have different data densities (sparse new vs. abundant base), causing gradient imbalance in a shared subspace; (2) the encoder creates gradient coupling between populations through shared computation layers. On knowledge graph completion (TransE (Bordes et al., 2013) on the ENTITY benchmark (Cui et al., 2023a)), where each embedding is trained independently without a shared encoder, the anti-correlation does not manifest.

**Connection to continual learning.** EWC (Kirkpatrick et al., 2017) implicitly implements a similar base/new separation: new entries have zero Fisher importance (unconstrained) while base entries have high importance (penalized). Our method achieves the same effect structurally: the rank constraint limits base drift while separate $V_{\text{new}}$ allows unconstrained new-entry learning. The structural approach avoids EWC's computational cost (Fisher computation scales poorly to million-item catalogs) and its sensitivity to the regularization strength $\lambda$.

**Practical deployment considerations.** Freeze1-SV is the most deployment-friendly variant: it requires no early stopping criterion, no population-specific validation set, and trains for a fixed number of epochs with guaranteed base stability. Dual-SV achieves higher new-entry quality but benefits from monitoring overall validation quality to select the best epoch. In production systems where the validation set composition changes daily, Freeze1-SV's epoch-robustness is a significant practical advantage.

**Multi-population extensions.** The framework generalizes to $K > 2$ populations. For example, items could be stratified by interaction frequency into cold ($k < 5$), warm ($5 \leq k < 50$), and hot ($k \geq 50$) cohorts, each with its own $V_k$ and learning rate. The theoretical bound suggests that sparser cohorts benefit more from the shared subspace, so rank could also be adapted per cohort ($r_{\text{cold}} < r_{\text{warm}} < r_{\text{hot}}$). We leave this exploration to future work.

## 8. Conclusion and Future Work

We extend low-rank adaptation from weight matrices to embedding tables for models with expanding vocabularies. The key insight is that the parameter distribution inverts across architectures: in LLMs, LoRA adapts the dominant transformer weights; in recommendation, our method adapts the dominant embedding table. Population-specific subspaces resolve the anti-correlation between base and new entries by decoupling their optimization while providing implicit regularization for sparse entries.

The framework Pareto-dominates baselines on sequential recommendation and wins in 8/9 LLM cells, with $4\times$ parameter efficiency at matched quality. Future directions include: (1) extending to $K > 2$ populations (e.g., frequency-based cohorts), (2) scaling to 7B+ LLMs where the anti-correlation effect may grow with model size, and (3) applying to DLRM-style models where embedding tables dominate even more than in sequential recommendation.

Code available at `https://anonymous.4open.science/r/base-overfit-2026-E184`.

## References


Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *NeurIPS*, 2013.

Cui, Y., Wang, Y., Sun, Z., Liu, W., Jiang, Y., Han, K., and Hu, W. Lifelong embedding learning and transfer for growing knowledge graphs. In *AAAI*, 2023a. ENTITY benchmark dataset derived from Freebase (CC BY 2.5).

Cui, Y., Yang, Z., and Yao, X. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023b.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized language models. In *NeurIPS*, 2023.

Fatkulin, A., Zhukova, A., and Makarov, I. Let it go? not quite: Addressing item cold start in sequential recommendations with content-based initialization. In *Proceedings of the Web Conference*, 2025.

Ginart, A., Neel, S., Roth, A., and Waldon, B. Mixed dimension embeddings with application to memory-efficient recommendation systems. In *IEEE International Symposium on Information Theory*, 2021.

Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Kang, W.-C. and McAuley, J. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining*, 2018.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Vinyals, O., Mohamed, S., et al. Overcoming catastrophic forgetting in neural networks. In *Proceedings of the National Academy of Sciences*, 2017.

Kornilova, A. and Eidelman, V. Billsum: A corpus for automatic summarization of us legislation. In *Workshop on New Frontiers in Summarization*, 2019.

Lian, D., Wang, H., Liu, Z., Lian, J., Chen, E., and Xie, X. Personalized embedding size search under memory constraint. In *SIGIR*, 2021.

Mi, F., Lin, B., and Luo, X. Ader: Adaptively distilled exemplar replay towards continual learning for session-based recommendation. In *RecSys*, 2020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Sato, L. et al. Merrec: A large-scale multipurpose mercari dataset for consumer-to-consumer recommendation systems. In *KDD*, 2025.

U.S. National Library of Medicine. Pubmed/medline baseline, 2024. Courtesy of the U.S. National Library of Medicine.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2024.

Wu, J. et al. SD-LoRA: Scalable decoupled low-rank adaptation for class incremental learning. *arXiv preprint arXiv:2501.13198*, 2025.

Yamaguchi, A., Chrysostomou, G., and Aletras, N. An empirical study on vocabulary expansion for language models. *arXiv preprint arXiv:2404.03416*, 2024.

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. GaLore: Memory-efficient LLM training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

Zhu, H., Li, X., Zhang, P., Li, G., He, J., Li, H., and Gai, K. Learning tree-based deep model for recommender systems. *KDD*, 2018. Dataset: https://tianchi.aliyun.com/dataset/649, CC BY-NC-SA 4.0.

## A. Multi-Seed Stability

*Table 8.* Multi-seed stability (W0, sampled eval). Mean ± std across 3 seeds (42, 123, 456). Method ordering is stable.

| Dataset | Method | new NDCG | base NDCG |
|---|---|---|---|
| Taobao | FT | $0.806 \pm 0.000$ | $0.696 \pm 0.000$ |
| | FT + freeze@ep1 | $0.865 \pm 0.001$ | $0.727 \pm 0.000$ |
| | Freeze1-SV ($r$=64) | $0.871 \pm 0.001$ | $0.723 \pm 0.000$ |
| | Freeze1-SV ($r$=128) | $0.874 \pm 0.001$ | $0.725 \pm 0.001$ |
| MerRec | FT | $0.674 \pm 0.014$ | $0.751 \pm 0.006$ |
| | FT + freeze@ep1 | $0.698 \pm 0.000$ | $0.767 \pm 0.000$ |
| | Freeze1-SV ($r$=64) | $0.721 \pm 0.003$ | $0.762 \pm 0.003$ |
| | Freeze1-SV ($r$=128) | $0.734 \pm 0.001$ | $0.766 \pm 0.001$ |

Standard deviation is below 0.003 NDCG on Taobao and below 0.014 on MerRec, confirming that method ordering is stable.

## B. Empirical Validation of Theoretical Bound

We estimate the low-rank approximation quality from trained models on Taobao (W0). Computing the SVD of centered new-item embeddings from a full-rank frozen model, the top-128 singular vectors capture 95% of variance ($\tau^2 = 0.006$ residual per item). Even at $r$=32, 54% of variance is captured ($\tau^2 = 0.056$), validating the approximate low-rank structure assumed by the proposition.

The bound predicts shared-$V$ outperforms independent estimation when $k_j$ is small relative to $d/r$. On both datasets, new items have median $k_j = 1$. At our primary rank $r$=128 ($d/r \approx 1.2$), the estimation-error improvement is modest; the empirical gains at high rank come primarily from the regularization effect of the shared subspace (preventing overfitting on sparse items) rather than pure variance reduction. At lower ranks ($r$=32, $d/r \approx 5$), the variance reduction is larger but approximation error increases. Prediction (P3) – data-scale crossover – is confirmed by the LLM experiments: at 15K documents on GPT-2, full-rank becomes preferable as per-token data grows large relative to model capacity. Prediction (P2) – diminishing returns – is confirmed by the rank ablation (Table 3): gains from $r$=64 to $r$=128 are smaller than from $r$=32 to $r$=64. Prediction (P1) – advantage concentrated in sparse entries – follows from the bound structure and is consistent with our new-item evaluation (all new items have $k_j \leq 5$, well below the crossover threshold).