UNDERSTANDING COLLABORATION MECHANISM IN VAE RECOMMENDER SYSTEMS

Anonymous authors

000

001

002 003 004

010 011

012

013

014

016

017

018

019

021

025

026

027

028

029

031

033

035

037

040

041

042

043

044

045

046

047

048

051

052

Paper under double-blind review

ABSTRACT

Variational Autoencoders (VAEs) are a powerful alternative to matrix factorization for recommendation. A common technique in VAE-based collaborative filtering (CF) consists in applying binary input masking to user interaction vectors, which improves performance but remains underexplored theoretically. In this work, we analyze how collaboration arises in VAE-based CF and show it is governed by latent proximity: we derive a latent sharing radius that informs when an SGD update on one user strictly reduces the loss on another user, with influence decaying as the latent Wasserstein distance increases. We further study the induced geometry: with clean inputs, VAE-based CF primarily exploits *local* collaboration between input-similar users and under-utilizes *global* collaboration between far-but-related users. We compare two mechanisms that encourage global mixing and characterize their trade-offs: \bigcirc β -KL regularization directly tightens the information bottleneck, promoting posterior overlap but risking representational collapse if too large; ② input masking induces stochastic geometric contractions and expansions, which can bring distant users onto the same latent neighborhood but also introduce neighborhood drift. To preserve user identity while enabling global consistency, we propose an anchor regularizer that aligns user posteriors with item embeddings, stabilizing users under masking and facilitating signal sharing across related items. Our analyses are validated on the Netflix, MovieLens-20M, and Million Song datasets. We also successfully deployed our proposed algorithm on an Amazon streaming platform following a successful online experiment.

1 Introduction

Recommender systems are essential for delivering personalized user experiences. Having benefited from three decades of research (Resnick et al., 1994), Collaborative filtering (CF) remains an essential approach in today's recommender systems (Zhu et al., 2025). CF predictive process consists in predicting user preferences by identifying and leveraging similarity patterns between users and items (Su & Khoshgoftaar, 2009; Ricci et al., 2010), which naturally aligns with the framework of latent variable models (LVMs) (Bishop & Nasrabadi, 2006), where latent representations are used to capture the shared structure of user-item interactions. Due to their simplicity and effectiveness, LVMs have historically played a central role in CF research. However, these models are inherently linear, which limits their capacity to model the complex and non-linear nature of real-world user behavior (Paterek, 2007; Mnih & Salakhutdinov, 2007). To overcome these limitations, researchers have increasingly explored the integration of neural networks (NNs) into CF frameworks, enabling more expressive modeling and yielding notable improvements in recommendation accuracy (He et al., 2017; Wu et al., 2016; Liang et al., 2018; Truong et al., 2021; Li et al., 2021). A particularly successful line of work is VAE-based collaborative filtering (Liang et al., 2018), which extends the variational autoencoder (VAE) framework (Kingma & Welling, 2013; Rezende et al., 2014) to collaborative filtering tasks. Unlike traditional latent factor models (Hu et al., 2008; Paterek, 2007; Mnih & Salakhutdinov, 2007), which require learning a separate latent vector for each user, VAE-based models offer a userindependent parameterization, where the number of trainable parameters remains fixed regardless of the number of users Lobel et al. (2019), leading to remarkable scalability. Additionally, empirically, VAE-based CF models consistently outperform many existing LVM-based alternatives (Liang et al., 2018; Kim & Suh, 2019; Walker et al., 2022; Ma et al., 2019; Guo et al., 2022; Wang et al., 2023; 2022; Guo et al., 2024; Husain & Monteil, 2024; Li et al., 2021; Tran & Lauw, 2025).

A key driver of the strong performance of VAE-based CF is the use of a binary mask that corrupts the user interaction vector, producing a partial history from which the model is trained to reconstruct the full interaction vector (Liang et al., 2018). This masking strategy has been shown to significantly enhance recommendation accuracy (see Section 4.2 and Figure 2, comparing settings with and without masking). Although input noise has become a standard component in VAE-based CF models due to its empirical effectiveness, its underlying mechanisms and potential side effects remain largely unexplored. Existing works treat masking as a simple performance-enhancing heuristic, without thoroughly examining how it influences the learning process or affects the latent representations. Our work aims to fill this gap with a comprehensive study the effect of input noise in VAE-based CF. The contributions of this paper are summarized as follows:

- We conduct an in-depth analysis of the collaboration mechanism in VAE-based collaborative filtering models and reveal that ① collaboration in VAE-CF is fundamentally governed by latent proximity; ② VAE-CF with clean inputs primarily leverages local collaboration and fails to utilize global collaborative signals when input distances are large; ③ Both β -KL regularization and input masking can encourage global collaborative signals, but they operate through distinct mechanisms with different trade-offs i.e., β -KL regularization promotes posterior mixing by directly constraining the information bottleneck, but suffers the risk of representational collapse when applied too aggressively while *Input masking* achieves mixing through *geometric* and stochastic means such that it can bring distant users into the same latent neighborhood, and latent space expansions can introduce neighborhood drift effects.
- Guided by our theoretical analysis, we propose a regularization scheme that addresses the issues induced by input masking, mitigating the loss of local collaboration while preserving its benefits for global alignment. Specifically, we model items as learnable anchors in latent space, and during training, the masked encoder outputs are pulled toward the user's anchor centroid. This acts as a training-only auxiliary condition that helps stabilize user representations under input corruption without tightening the information bottleneck, promoting consistent, semantically grounded latent proximity. To our knowledge, ours is the first work to systematically analyze the collaboration mechanisms in VAE-based CF, showing that both β -KL regularization and input masking can promote global collaboration. In contrast to prior works that focus on addressing the problems of β -KL regularization (ref. Appendix B for a detailed discussion), we address the issues induced by input masking.
- Our experimental results demonstrate the strong benefits of the proposed PIA approach compared to vanilla VAE-CF on benchmark datasets and especially the success on the A/B testing at Amazon streaming platform. We conducted ablation studies across user groups segmented by interaction count to validate the effectiveness of global collaborative signals. Additionally, we provide visualizations of the learned latent space that support our theoretical analysis.

2 COLLABORATION MECHANISM IN VAE-BASED CF

Notations. We index users by $u \in \{1, 2, \dots, U\}$ and items by $i \in \{1, 2, \dots, I\}$, where U, I are the number of users and items, respectively. $\mathbf{X} \in \{0, 1\}^{U \times I}$ represents the user-item interaction matrix (e.g., click, watch, check-in, etc.) and $\mathbf{x}_u = [\mathbf{x}_{u1}, \mathbf{x}_{u2}, \dots, \mathbf{x}_{uI}]$ is an I-dimensional binary vector (the u-th row of \mathbf{X}) whereby $\mathbf{x}_{ui} = 1$ implies that user u has interacted with item i and $\mathbf{x}_{ui} = 0$ indicates otherwise. Note that $\mathbf{x}_{ui} = 0$ does not necessarily mean user n dislikes item i; the item may never be shown to the user. For simplicity, we use \mathbf{x} to denote a general user interaction vector and retain \mathbf{x}_u when specifically referring to user u. Additionally, we measure distances between input vectors with the ℓ_1 norm $\|\cdot\|_1$ (Hamming) and between latent distributions with the 1-Wasserstein distance $W_1(\cdot,\cdot)$. We refer to the Appendix Table A for the notation summary.

2.1 VAE-BASED COLLABORATIVE FILTERING

Given a user's interaction history $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I]^{\top} \in \{0, 1\}^I$, the goal is to predict the full interaction behavior of this user with all remaining items. To simulate this process during training, as well as to avoid overfitting to non-informative patterns (Steck, 2020), VAE-based CF (Liang et al., 2018; Kim & Suh, 2019; Truong et al., 2021; Vančura & Kordík, 2021) introduces a random binary mask $\mathbf{b} \in \{0, 1\}^I$ is , with the entry 1 as *un-masked*, and 0 as *masked*. Thus, $\mathbf{x}_h = \mathbf{x} \odot \mathbf{b}$ is the user's partial interaction history, and the goal is to recover the full \mathbf{x} given \mathbf{x}_h .

Training Objective. The parameters ϕ , θ of the VAE-based collaborative filtering model are learnt by minimizing the negative β -regularized Evidence Lower Bound (ELBO):

$$\mathcal{L}_{VAE}(\mathbf{x}; \theta, \phi) = -\mathbb{E}_{q_{\phi}(\mathbf{z} \mid \mathbf{x}_{h})} \left[\log p_{\theta}(\mathbf{x} \mid \mathbf{z}) \right] + \beta \operatorname{KL} \left(q_{\phi}(\mathbf{z} \mid \mathbf{x}_{h}) \parallel p(\mathbf{z}) \right), \tag{1}$$

with standard reparameterization

$$\mathbf{b} \sim \text{Bernoulli}(\rho)^I, \quad \mathbf{x}_h = \mathbf{x} \odot \mathbf{b}, \quad (\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = q_{\phi}(\mathbf{x}_h), \ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \ \mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \boldsymbol{\sigma}$$
 (2)

where ρ is the hyperparameter of a Bernoulli distribution, q_{ϕ} is a ϕ -parameterized neural network, which outputs the mean μ and variance σ^2 of the Gaussian distribution.

2.2 THEORETICAL ANALYSIS

Before presenting our theoretical analysis, we first recall the CF target: predict a user's preferences from other users' interaction patterns, without relying on item or user side content. Figure 1 illustrates two typical situations. First, users \mathbf{x}_1 and \mathbf{x}_4 are close in the input space (e.g., under ℓ_1), so each serves as a natural reference for the other by the definition of the collaborative filtering. Second, let $S_{\mathbf{x}} = \{i : \mathbf{x}_i = 1\}$ contain the positive items for user \mathbf{x} , if $S_{\mathbf{x}_2} \subset S_{\mathbf{x}_1}$ i.e., \mathbf{x}_1 is a more active user with similar interests, then \mathbf{x}_1 should inform recommendations for \mathbf{x}_2 (e.g., items $i_1, i_2, i_3 \in S_{\mathbf{x}_1} \setminus S_{\mathbf{x}_2}$). We refer to this latter, "far-but-related" influence as a global collaborative signal. To make these notions precise, for $\delta > 0$, we define the input-space neighborhood for user u $N_{\delta}(u) = \int_{-\infty}^{\infty} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{2\pi$

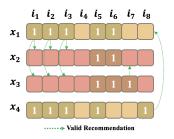


Figure 1: *Local* and *global* collaborative signal example.

$$\{v: \|\mathbf{x}_u - \mathbf{x}_v\|_1 \le \delta\}$$
 and the (nonzero) overlap indicator: $\operatorname{ov}(\mathbf{x}_u, \mathbf{x}_v) = \mathbf{1}_{\{|S_{\mathbf{x}_u} \cap S_{\mathbf{x}_v}| > 0\}}$.

Definition 2.1 (Local collaborative signal). The prediction for user u depends on other users within the neighborhood $N_{\delta}(u)$.

Definition 2.2 (Global collaborative signal). A model exhibits a *global* collaborative signal at scale δ if there exist users u and v which are "far-but-related" i.e., $\|\mathbf{x}_u - \mathbf{x}_v\|_1 > \delta$ and $\operatorname{ov}(\mathbf{x}_u, \mathbf{x}_v) = 1$, and their predictions strictly influence each other.

We now examine how VAE-based CF methods leverage local and global collaborative signals.

Theorem 2.3 (Latent- W_1 sharing radius). Assume decoder gradient is Lipschitz in z such that for any given $\mathbf{z} \sim q_{\phi}(\cdot \mid \mathbf{x}_u)$, uniformly in \mathbf{x} and $\forall \mathbf{z}'$, $\|\nabla_{\theta}\ell_{\theta}(\mathbf{x}, \mathbf{z}) - \nabla_{\theta}\ell_{\theta}(\mathbf{x}, \mathbf{z}')\| \leq L_{\theta z} \|\mathbf{z} - \mathbf{z}'\|$. Let

$$q_u := q_{\phi}(\cdot \mid \mathbf{x}_u), \quad q_v := q_{\phi}(\cdot \mid \mathbf{x}_v), \quad \mathcal{L}_u(\theta) := \mathbb{E}_{q_u}[\ell_{\theta}(\mathbf{x}_u, \mathbf{z})], \quad g_u(\theta) := \nabla_{\theta} \mathcal{L}_u(\theta).$$

Define the content-mismatch term

$$\Delta_x(u,v) := \left\| \mathbb{E}_{q_v} \left[\nabla_{\theta} \ell_{\theta}(\mathbf{x}_u, \mathbf{z}) - \nabla_{\theta} \ell_{\theta}(\mathbf{x}_v, \mathbf{z}) \right] \right\|, \quad D_{u,v} := L_{\theta z} W_1(q_u, q_v) + \Delta_x(u, v).$$

For one SGD step on user v, $\theta^+ = \theta - \eta q_v(\theta)$, the first-order change satisfies

$$\mathcal{L}_{u}(\theta^{+}) - \mathcal{L}_{u}(\theta) \leq -\eta \|g_{u}(\theta)\| (\|g_{u}(\theta)\| - D_{u,v}) + O(\eta^{2}).$$

Consequently, the step on v strictly decreases \mathcal{L}_u to first order whenever $D_{u,v} < \|g_u(\theta)\|$, i.e.

$$W_1(q_u, q_v) < r_{\text{share}}(u, v; \theta) := \frac{[\|g_u(\theta)\| - \Delta_x(u, v)]_+}{L_{\theta z}}$$

Proof. See Appendix E.1.

Result-1 (Collaboration is governed by *latent* **proximity).** The quantity $D_{u,v} = L_{\theta z} W_1(q_u,q_v) + \Delta_x(u,v)$ is the *transfer penalty*: it upper-bounds the per-step regret at user u by updating the model using user v. When $D_{u,v} < \|g_u\|$, the step computed on v still impacts u; thus collaboration during training is localized to the latent neighborhood $\{v: W_1(q_u,q_v) < r_{\text{share}}(u,v;\theta)\}$ and collaborative signal reduces with increased latent distance.

Consequently, to understand how VAE-CF exploits both local and global collaborative signals, we examine the correspondence between the *geometry of the input space* and the *geometry of the latent space* in VAE-based collaborative filtering, in the following section.

2.3 CORRESPONDENCE BETWEEN GEOMETRY OF THE INPUT AND LATENT SPACE

We summarize the key theoretical results and present the discussion subsequently.

Lemma 2.4. Assume the encoder is L_{ϕ} -Lipschitz, i.e., $\|q_{\phi}(\mathbf{x}_u) - q_{\phi}(\mathbf{x}_v)\| \le L_{\phi} \|\mathbf{x}_u - \mathbf{x}_v\| \ \forall \mathbf{x}_u, \mathbf{x}_v \in \{0,1\}^I$ then $W_1(q_{\phi}(\cdot \mid \mathbf{x}_u), q_{\phi}(\cdot \mid \mathbf{x}_v)) \le L_{\phi} \|\mathbf{x}_v - \mathbf{x}_u\|_1, \ \forall \mathbf{x}_u, \mathbf{x}_v \in \{0,1\}^I$.

Theorem 2.5. Assume $p_{\theta}(\mathbf{x} \mid \mathbf{z})$ is a regular exponential family with sufficient statistics $T(\mathbf{x})$, natural parameter $\eta(\mathbf{z})$, and log-partition A. Let $q_u = q_{\phi}(\cdot \mid \mathbf{x}_u)$ and $q_v = q_{\phi}(\cdot \mid \mathbf{x}_v)$, and define $\alpha(\mathbf{z}) = \frac{q_u(\mathbf{z})}{q_u(\mathbf{z}) + q_v(\mathbf{z})}$. Then the β -regularized pairwise objective satisfies

$$\min_{\eta(\cdot)} \left\{ \sum_{i \in \{u,v\}} \mathbb{E}_{q_i} \left[-\log p_{\theta}(\mathbf{x}_i \mid \mathbf{z}) \right] + \beta \sum_{i \in \{u,v\}} \text{KL} \left(q_{\phi}(\mathbf{z} \mid \mathbf{x}_i) \parallel p(\mathbf{z}) \right) \right\} \\
= C + \int \left(q_u(\mathbf{z}) + q_v(\mathbf{z}) \right) \Delta_{A^*} \left(\mathbf{x}_u, \mathbf{x}_v; \alpha(\mathbf{z}) \right) d\mathbf{z} + \beta \sum_{i \in \{u,v\}} \text{KL} \left(q_{\phi}(\mathbf{z} \mid \mathbf{x}_i) \parallel p(\mathbf{z}) \right), \tag{3}$$

where C is a constant, A^* is the convex conjugate of A, and for $\alpha \in [0,1]$

$$\Delta_{A^*}(\mathbf{x}_u, \mathbf{x}_v; \alpha) = \alpha A^*(T(\mathbf{x}_u)) + (1 - \alpha) A^*(T(\mathbf{x}_v)) - A^*(\alpha T(\mathbf{x}_u) + (1 - \alpha)T(\mathbf{x}_v)) \ge 0,$$
with equality iff $T(\mathbf{x}_u) = T(\mathbf{x}_v)$.

Additionally, assume the prior p satisfies the Bobkov–Götze/Talagrand $T_1(C)$ inequality (Bobkov & Götze, 1999) with a constant C > 0 (e.g., Normal prior $p = \mathcal{N}(0, \sigma^2 I)$):

$$W_1(q_u, q_v) \leq \left(\sqrt{2C \operatorname{KL}(q_u \| p)} + \sqrt{2C \operatorname{KL}(q_v \| p)}\right). \tag{4}$$

Result-2 (Clean inputs favor local over global collaboration).

- Input locality induces latent locality. We show in Lemma 2.4 that nearby inputs (in ℓ_1) induce nearby posteriors in $\mathcal Z$ i.e., input neighborhoods embed into latent neighborhoods, preserving local collaboration.
- Reconstruction favors posterior separation. When $T(\mathbf{x}_u) \neq T(\mathbf{x}_v)$, any overlap of q_u and q_v incurs the strictly positive gap $\int (q_u + q_v) \Delta_{A^*} d\mathbf{z}$ in Eq. (3), discouraging posterior overlap and preventing input-distant, content-mismatched users from becoming latent-near, discouraging global collaboration.
- β -KL encourages posterior overlap. Increasing β reduces the optimal KL terms, tightening the upper bound of $W_1(q_u,q_v)$, thereby increasing latent overlap i.e., potentially bringing input-distant users closer in latent space and enabling global collaboration. However, an excessively large β can lead to posterior collapse and weaken predictive performance. In practice, β is typically kept small in ranking tasks.

Consequently, for moderate β , similar users remain latent-near while content-mismatched users are pushed apart, yielding a clustered latent geometry aligned with input similarity (Setting-1, Figure 2). By Theorem 2.3, SGD updates are therefore shared predominantly within these latent neighborhoods (i.e., across pairs with $W_1(q_u,q_v) < r_{\rm share}(u,v;\theta)$). Hence, with clean inputs, VAE-CF primarily exploits local collaborative signals and suppresses global influence from input-distant users.

Theorem 2.6. Let $\mathbf{x}_u, \mathbf{x}_v \in \{0,1\}^I$ be binary inputs. Let $b_{\mathbf{x}_u}, b_{\mathbf{x}_v} \sim \mathrm{Bern}(\rho)^I$ be independent masks and set $\mathbf{x}_u' = \mathbf{x}_u \odot b_{\mathbf{x}_u}$ and $\mathbf{x}_v' = \mathbf{x}_v \odot b_{\mathbf{x}_v}$. Denote the number of non-overlapped items $h = \|\mathbf{x}_u - \mathbf{x}_v\|_1$ and the number of overlapped items $s = \langle \mathbf{x}_u, \mathbf{x}_v \rangle$. For any $\delta > 0$, define $T_{\delta} = \lceil \delta \rceil - 1$ and $U_{\delta} = \lceil \delta \rceil$. Then:

Contraction.
$$\Pr[\|\mathbf{x}'_u - \mathbf{x}'_v\|_1 < \delta] \geq (\rho^2 + (1-\rho)^2)^s \sum_{k=0}^{\min\{h, T_\delta\}} {h \choose k} \rho^k (1-\rho)^{h-k}.$$

Expansion.
$$\Pr[\|\mathbf{x}'_u - \mathbf{x}'_v\|_1 \ge \delta] \ge \sum_{k=U_\delta}^s {s \choose k} (2\rho(1-\rho))^k (1-2\rho(1-\rho))^{s-k}.$$

Result-3 (Masking induces stochastic neighborhood mixing). Random masks perturb pairwise distances stochastically i.e., each draw can contract or expand $\|\mathbf{x}'_u - \mathbf{x}'_v\|_1$. Under Lemma 2.4, contractions propagate to latent space; if the realized δ satisfies $L_{\phi} \delta < r_{\text{share}}(u, v; \theta)$, an SGD step on v strictly reduces \mathcal{L}_u (cf. Theorem 2.3). On the other hand, expansions can push pairs outside the sharing radius, weakening local collaboration. This creates a pattern where latent representations of distant groups become intermittently entangled via contractions, while within-group cohesion diffuses due to occasional expansions (Setting-2, Figure 2).

3 Proposed Method

The previous section established that collaboration is governed by *latent* proximity via the sharing radius (Theorem 2.3). With clean (unmasked) inputs, ℓ_1 locality makes latent locality and reconstruction favors posterior *separation*, so updates are shared primarily within local neighborhoods; increasing β can counteract this by uniformly contracting pairwise latent distances and enabling long-range sharing, but overly large β weakens reconstruction gradients and can shrink the sharing radius. In contrast, input masking stochastically contracts/expands neighborhoods across batches, producing *intermittent* long-range sharing alongside potential neighborhood drift.

To convert these geometric properties into *actionable insights*, we connect latent distances to the KL terms that we actually optimize in the ELBO.

3.1 From Latent Geometry to an Information Lens: β -KL vs. Masking

Clean input. Theorem 2.5 yields for any users u, v:

$$W_1(q_{\phi}(\cdot \mid \mathbf{x}_u), q_{\phi}(\cdot \mid \mathbf{x}_v)) \leq \sqrt{2C \operatorname{KL}(q_{\phi}(\cdot \mid \mathbf{x}_u) \parallel p)} + \sqrt{2C \operatorname{KL}(q_{\phi}(\cdot \mid \mathbf{x}_v) \parallel p)}.$$
 (5)

Averaging over user pairs and using Jensen's inequality (concavity of the square root),

$$\mathbb{E}_{u,v} W_1(q_u, q_v) \leq 2\sqrt{2C} \sqrt{\mathbb{E}_{\mathbf{x}} \operatorname{KL}(q_\phi(\cdot \mid \mathbf{x}) \parallel p)}.$$
 (6)

By rearranging the ELBO terms with $q(Z) := \mathbb{E}_X q_{\phi}(Z \mid X)^{-1}$,

$$\mathbb{E}_{\mathbf{x}}\Big[\mathrm{KL}\big(q_{\phi}(Z\mid \mathbf{x}) \parallel p(Z)\big)\Big] = I_{q_{\phi}}(X;Z) + \mathrm{KL}\big(q(Z) \parallel p(Z)\big),$$

so lowering $I_q(X; Z)$ and/or the aggregated divergence $\mathrm{KL}(q \| p)$ shrinks expected pairwise latent distances, increasing the chance that users fall within the sharing radius (Theorem 2.3).

Input masking. Let $\mathbf{x}_h = \mathbf{x} \odot \mathbf{b}$ with independent $\mathbf{b} \sim \mathrm{Bern}(\rho)^I$ and define $q_h(\mathbf{z}) := \mathbb{E}_{\mathbf{x},\mathbf{b}} q_\phi(\mathbf{z} \mid \mathbf{x} \odot \mathbf{b})$. Applying Eq. (5) to masked posteriors and averaging over masks,

$$\mathbb{E}_{u,v,\mathbf{b}} W_1(q_{\phi}(\cdot \mid \mathbf{x}_u \odot \mathbf{b}_u), q_{\phi}(\cdot \mid \mathbf{x}_v \odot \mathbf{b}_v)) \leq 2\sqrt{2C} \sqrt{\mathbb{E}_{\mathbf{x},\mathbf{b}} \operatorname{KL}(q_{\phi}(\cdot \mid \mathbf{x} \odot \mathbf{b}) \parallel p)}.$$
(7)

Rearranging ELBO terms for the masked variable X_h gives

$$\mathbb{E}_{\mathbf{x},\mathbf{b}} \operatorname{KL} (q_{\phi}(Z \mid \mathbf{x} \odot \mathbf{b}) \parallel p(Z)) = I_{q_{\phi}}(X_h; Z) + \operatorname{KL} (q_h(Z) \parallel p(Z)).$$

Hence Eq. (7) ties the *mask-averaged* latent distances directly to $I_q(X_h; Z)$ and $\mathrm{KL}(q_h \| p)$. In words: masking acts by reducing the usable instance information (data-processing: $I_q(X_h; Z) \leq I_q(X; Z)$), which *on average* pulls masked posteriors closer together and to the prior, thereby increasing the frequency of global sharing events, while still allowing per-draw contractions/expansions (Result-3).

Distinct geometries at two levels. Both mechanisms increase posterior overlap but reshape geometry differently:

1. β -KL (objective-level). Scaling the KL by β directly reduces $I_q(X;Z)$ and $\mathrm{KL}(q\|p)$, shrinking the RHS of Eq. (6). Geometry: uniform shrinkage of inter-user distances, hence uniform global mixing. Trade-off: too large β weakens reconstruction gradients ($\|g_u\|$) and risks posterior collapse, blurring user semantics.

¹We use uppercase letters for random variables e.g., X and bold lowercase for their realizations x.

2. Input Masking (data-level). Masking reduces $I_q(X_h; Z)$ and $\mathrm{KL}(q_h \| p)$, shrinking the RHS of Eq. (7). Geometry: stochastic contractions/expansions across batches (Result-3), which can intermittently bring far-but-related users within the sharing radius, but also introduce neighborhood drift that may weaken local reliability.

Remark. Eq. (3) in Theorem 2.5 shows that reconstruction favors posterior *separation* when $T(\mathbf{x}_u) \neq T(\mathbf{x}_v)$, while Eq. (5)–Eq. (7) show how the β -KL or masking can increase *overlap*. Prior work mostly tunes the β -KL pathway i.e., scaling the KL by β or redesigning the prior p (ref. Section B for detail discussion). In contrast, we operate on the *masking* pathway: preserve instance information (avoid uniform shrinkage) but stabilize its stochastic geometry so that beneficial long-range sharing occurs more often and with less drift. We formalize this next.

3.2 Personalized Item Alignment (PIA)

During training the encoder takes masked inputs $\mathbf{x}_h = \mathbf{x} \odot \mathbf{b}$ (cf. Eq. 2). Rather than conditioning $q_{\phi}(\mathbf{z} \mid \cdot)$ on auxiliary features at inference, we introduce a *regularizer* that, *during training*, pushes the posterior given \mathbf{x}_h towards a user-specific target derived from the user's positive items.

We define the overall objective as:

$$\mathcal{L}_{\text{PIA-VAE}}(\mathbf{x}; \theta, \phi, E) = \mathbb{E}_{\mathbf{b}} \Big[\mathcal{L}_{\text{VAE}}(\mathbf{x}; \theta, \phi; \mathbf{x}_h) \Big] + \lambda_{\text{A}} \mathbb{E}_{\mathbf{b}} \Big[\mathcal{L}_{\text{A}}(\mathbf{x}_h, \mathbf{x}; \phi, E) \Big], \tag{8}$$

where $E = \{\mathbf{e}_i \in \mathbb{R}^d\}_{i=1}^I$ are learnable *item anchors* in latent space (same dimension as \mathbf{z}), and $\lambda_A > 0$ is small. In particular, let $S_{\mathbf{x}} = \{i : \mathbf{x}_i = 1\}$ be the positives for user \mathbf{x} , we have:

$$\mathcal{L}_{A}(\mathbf{x}_{h}, \mathbf{x}; \phi, E) = \frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S_{\mathbf{x}}} \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x}_{h})} [\|\mathbf{z} - \mathbf{e}_{i}\|_{2}^{2}].$$
(9)

Intuitively, this pulls the masked posterior toward the anchors of the items the user liked, making users with shared items close in the *latent* space.

Proposition 3.1. Assume the encoder posterior is diagonal-Gaussian, $q_{\phi}(\mathbf{z} \mid \mathbf{x}_h) = \mathcal{N}(\boldsymbol{\mu}_{\phi}(\mathbf{x}_h), \operatorname{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_h)))$. Let the item centroid be $\bar{\mathbf{e}}_{\mathbf{x}} := \frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S_{\mathbf{x}}} \mathbf{e}_i$, then

$$\mathcal{L}_{A}(\mathbf{x}_{h}, \mathbf{x}; \phi, E) = \underbrace{\left\| \boldsymbol{\mu}_{\phi}(\mathbf{x}_{h}) - \bar{\mathbf{e}}_{\mathbf{x}} \right\|_{2}^{2}}_{align \ mean \ to \ item \ centroid} + \underbrace{\operatorname{tr} \Sigma_{\phi}(\mathbf{x}_{h})}_{variance \ shrinkage} + const(\mathbf{x}, E), \tag{10}$$

where $\Sigma_{\phi}(\mathbf{x}_h) = \operatorname{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_h))$ and $\operatorname{const}(\mathbf{x}, E) = \frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S_{\mathbf{x}}} \|\mathbf{e}_i\|_2^2 - \|\bar{\mathbf{e}}_{\mathbf{x}}\|_2^2$.

Proposition 3.1 indicates that PIA 1 centers masked latents near the user's *item barycenter* $\overrightarrow{\mathbf{e}}_{\mathbf{x}}$ and 2 modestly reduces posterior spread. Two users u,v with similar positive-item sets have close centroids, so their masked posteriors become *latently near* more frequently, increasing the chance they fall within the sharing radius and benefit from each other's updates.

Proposition 3.2. Fix \mathbf{x} and its neighborhood in which the ELBO objective $\mathcal{L}_{VAE}(\mathbf{x}; \theta, \phi; \mathbf{x}_h)$, defined in Eq. (1), written as a function of the encoder mean $\boldsymbol{\mu}_{\phi}(\mathbf{x}_h)$, admits a quadratic approximation with Hessian $H \succeq mI$ and $\|H\| \le L$. Adding $\lambda_A \|\boldsymbol{\mu}_{\phi}(\mathbf{x}_h) - \bar{\mathbf{e}}_{\mathbf{x}}\|_2^2$ to this objective yields an effective Hessian $H_{eff} = H + 2\lambda_A I$. Let $\boldsymbol{\mu}^{(0)}(\mathbf{x}_h)$ be the unregularized minimizer over masks and $\boldsymbol{\mu}^{(A)}(\mathbf{x}_h)$ the minimizer with alignment. Then with $\tau = \left(\frac{L}{L+2\lambda_A}\right)$, we obtain the following inequalities

$$\operatorname{Var}_{\mathbf{b}} \big[\boldsymbol{\mu}^{(A)}(\mathbf{x}_h) \big] \ \preceq \ \tau^2 \operatorname{Var}_{\mathbf{b}} \big[\boldsymbol{\mu}^{(0)}(\mathbf{x}_h) \big], \qquad \mathbb{E}_{\mathbf{b}} \big[\| \boldsymbol{\mu}^{(A)}(\mathbf{x}_h) - \bar{\mathbf{e}}_{\mathbf{x}} \|^2 \big] \ \leq \ \tau \mathbb{E}_{\mathbf{b}} \big[\| \boldsymbol{\mu}^{(0)}(\mathbf{x}_h) - \bar{\mathbf{e}}_{\mathbf{x}} \|^2 \big].$$

Proof. See Appendix E.5.
$$\Box$$

Proposition 3.2 indicates that adding the PIA term makes the masked encoder *locally better conditioned* and pulls its mean $\mu_{\phi}(\mathbf{x}_h)$ toward a per-user item centroid. Quantitatively, it shrinks ① the variance of $\mu_{\phi}(\mathbf{x}_h)$ across different masks and ② the average drift of $\mu_{\phi}(\mathbf{x}_h)$ from the centroid by a

multiplicative factor $\tau \in (0,1)$. Hence, **masked views of the same user are more alike and less noisy**, so the neighborhoods we train on are closer to the neighborhoods we infer on at test time.

In summary, PIA ① preserves instance information: PIA does not add a penalty on $I_q(X; Z)$; it acts on geometry during training. With full- ${\bf x}$ inference, the encoder can still exploit all available signal; ② stabilizes the geometric pathway: aligning $q_{\phi}({\bf z} \mid {\bf x}_h)$ to a fixed per-user $\bar{\bf e}_{\bf x}$ reduces masked-vs-clean drift and gradient variance; expansions are less likely to eject genuine neighbors from the sharing radius; ③ promotes meaningful global mixing: shared items pull users toward nearby centroids, creating consistent, semantically grounded latent proximity instead of relying purely on stochastic contractions ((Setting-3, Figure 2)); ④ introduces no test-time burden: E and the regularizer are estimated during training-only; inference uses the standard $q_{\phi}({\bf z} \mid {\bf x})$.

4 EXPERIMENTS

We validate our analysis using three real-world recommendation datasets: MovieLens-20M, Netflix, and Million Song Dataset and the A/B testing on an Amazon streaming platform. Specifically:

- First, we assess benefits of the proposed personalized item alignment approach compared to vanilla VAE-based CF (Multi-VAE (Liang et al., 2018)) on these benchmark datasets.
- Second, we provide visualizations of the learned latent space under three conditions: VAE without masking, VAE with masking, and VAE with PIA, to support our theoretical analysis.
- Finally, we conducted ablation studies across user groups segmented by interaction count to validate the effectiveness of global collaborative signals.

4.1 EFFECTIVENESS OF PERSONALIZED ITEM ALIGNMENT

Public dataset. Table 1 presents the performance of our framework, which adds personalized item alignment to Multi-VAE, and RecVAE (Shenbin et al., 2020) on the MovieLens-20M, Netflix and Million Song datasets respectively. We follow the preprocessing procedure from (Liang et al., 2018). The detailed data preprocessing steps and train/validation/test split methodology are presented in Section D.1. Our code for reproducibility is available anonymously at https://anonymous.4open.science/r/PIAVAE-E082/.

Table 1: Our method (with PIA) achieves the best performance for MovieLens and Netflix Prize datasets while having the 3rd rank for Million Song. The best results are highlighted in bold.

	Mo	vieLens-2	20M	Netflix Prize			Million Song		
Model	Recall	Recall	NDCG	Recall	Recall	NDCG	Recall	Recall	NDCG
	@20	@50	@100	@20	@50	@100	@20	@50	@100
Popularity	0.162	0.235	0.191	0.116	0.175	0.159	0.043	0.068	0.058
EASE	0.391	0.521	0.420	0.362	0.445	0.393	0.333	0.428	0.389
MF	0.367	0.498	0.399	0.335	0.422	0.369	0.258	0.353	0.314
WMF	0.362	0.495	0.389	0.321	0.402	0.349	0.211	0.312	0.257
GRALS	0.376	0.505	0.401	0.335	0.416	0.365	0.201	0.275	0.245
PLRec	0.394	0.527	0.426	0.357	0.441	0.390	0.286	0.383	0.344
WARP	0.310	0.448	0.348	0.273	0.360	0.312	0.162	0.253	0.210
LambdaNet	0.395	0.534	0.427	0.352	0.441	0.386	0.259	0.355	0.308
	Nonlinear autoencoders: MLP for encoder								
CDAE	0.391	0.523	0.418	0.343	0.428	0.376	0.188	0.283	0.237
RaCT	0.403	0.543	0.434	0.357	0.450	0.392	0.268	0.364	0.319
Multi-VAE	0.395	0.537	0.426	0.351	0.444	0.386	0.266	0.364	0.316
Multi-VAE + PIA	0.408	0.546	0.437	0.360	0.448	0.392	0.275	0.372	0.326
Uplift (%)	3.29	1.68	2.58	2.56	0.90	1.55	3.38	2.20	3.16
Nonlinear autoencoders: densely connected layers for encoder									
RecVAE	0.414	0.553	0.442	0.361	0.452	0.394	0.276	0.374	0.326
RecVAE + PIA	0.417	0.556	0.446	0.365	0.454	0.396	0.278	0.376	0.329
Uplift (%)	0.72	0.54	0.90	1.01	0.44	0.51	0.72	0.54	0.92

The results demonstrate that PIA consistently improves the performance over the base VAE recommenders, in terms of nDCG and Recall. It also exhibits competitive performance across the three

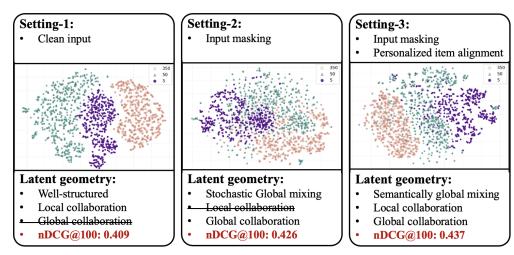


Figure 2: t-SNE visualization of the latent representations for three user groups differentiated by the number of interactions from ML-20M dataset. Purple, Teal, and Orange denote users with 5, 50, and 350 interactions, respectively, correspond to different VAE model configurations. We select group of users with 5, 50, and 350 interactions as they are clearly separated (i.e., L1) on the input space.

datasets considered, with RecVAE+PIA being the top performing approach on MovieLens-20M and Netflix datasets, and the 3rd performing approach on Million Song dataset.

A/B testing on an Amazon streaming platform. On the basis of offline results, we run one week of A/B testing in September 2025 for the Multi-VAE+PIA algorithm on one streaming platform of Amazon. The approach was implemented as an offline system, with weekly training considering a 3-month window for collecting streaming behavior, and daily inference for active customers. The personalized scores computed daily include about 25 millions of users and 4000 movies. The performance of our system was compared with a statistical baseline (control group) on 2 movie cards present on the Home page and on the Movie page. As shown in Table 2, our approach outperforms the control group significantly, with improved performance on the card click rate by 117% - 267% (per daily view) and 123% - 283% (per daily user view).

4.2 LATENT SPACE VISUALIZATION

We present t-SNE visualizations (Maaten & Hinton, 2008) of the latent spaces learned under three settings: 1 clean input, 2 input masking, and 3 input masking with personalized item alignment, to examine the correspondence between the geometry of the input space and the latent space.

Setup: We focus on user cohorts with 5, 50, and 350 interactions to clearly contrast local versus global collaboration. Note that standard set-based distances (Hamming) inflate cross-cohort dissimilarity: even two 5-interaction users with disjoint histories are closer to each other than any 5–50 pair, even when the 50-interaction user subsumes the 5-interaction user's items; the same pattern holds for the 50–350 cohorts.

Setting-1: As illustrated in Figure 2, when masking is disabled and interaction counts differ substantially, the learned representations are cleanly segregated by cohort. This indicates that the model in this setting primarily leverages local collaboration and has limited ability to capture global collaborative signals. Moreover, the latent geometry mis-aligns with the global structure of the input space: the 350-interaction cluster lies closer to the 5-interaction cluster than to the 50-interaction cluster which contrary to the expected ordering, where distance (350, 50) < distance(350, 5).

Setting-2: the representations from different cohorts become stochastically entangled, which encourages global sharing. However, within-cohort structure is more diffuse, weakening local collaboration. Despite this trade-off, **Setting-2** substantially outperforms **Setting-1** (nDCG@100= 0.426 vs. 0.409), demonstrating the benefit of encouraging global collaboration.

Setting-3 augments masking with PIA, which ① helps the VAE remain discriminative under input corruption, yielding a more structured latent space and ② promotes globally consistent user repre-

sentations. The resulting latent manifold is both well organized and globally aligned, exhibiting smooth transitions from the 5- to 50- to 350-interaction cohorts. This balance of local and global collaboration yields the best performance (nDCG@100 = 0.437).

4.3 Performance of user groups with different number of interactions

Table 2: Offline and online results on an Amazon streaming Table 3: Results across user groups for platform.

MovieLens20M.

nDCG

@100

0.386

	Multi-VAE+PIA	0.609	0.302	0.405			
	Uplift (%)	2.87	4.88	5.13			
	Model	Playtime (sec) per user view	Click Rate (%) per view	Click Rate (%) per user view			
	Home Card						
Online	Control Group	27.7	4.4	5.3			
	Multi-VAE+PIA	74.6	9.5	12.0			
	Uplift (%)	169	117	123			
	Movie Card						
	Control Group	16.8	3.4	4.2			
	Multi-VAE+PIA	102.6	12.5	16.2			
	Unlift (%)	500	267	283			

Recall

@50

0.288

@20

0.592

Group	Model	Re	nDCG	
r		@20	@50	@100
	Multi-VAE	0.461	0.625	0.317
[5–10]	Multi-VAE + PIA	0.473	0.629	0.323
	Uplift (%)	2.72	0.55	1.63
[11–50]	Multi-VAE	0.421	0.595	0.429
	Multi-VAE + PIA	0.424	0.598	0.434
	Uplift (%)	0.86	0.49	0.13
[51–100]	Multi-VAE	0.313	0.478	0.497
	Multi-VAE + PIA	0.314	0.479	0.502
	Uplift (%)	0.26	0.09	0.85
[100+]	Multi-VAE	0.418	0.386	0.474
	Multi-VAE + PIA	0.435	0.393	0.486
	Uplift (%)	4.09	0.72	2.57

Our proposed framework provides both a well-structured latent space and the capacity to capture global collaborative signals. As a result, we expect it to benefit users across groups, including cold-start, neutral, and warm-start users. To assess this, we partition the test-set users based on their number of interactions and evaluate the performance of our method within each group.

As shown in Table 3, our framework improves performance for all user groups. Notably, the cold-start group (within 5 to 10 interactions) and the warm-start group (more than 100 interactions) benefit the most. This can be attributed to the inherent challenges each group faces: cold-start users have limited historical data, making recommendation difficult, while warm-start users —often found in the long tail of the user distribution—typically lack sufficient collaborative overlap. Our framework addresses both issues by enhancing access to global collaborative signals.

5 CONCLUSION

Model

Offline

In this work, we analyzed how collaboration emerges in VAE–CF and showed that it is fundamentally governed by *latent* proximity: SGD updates are shared within a data-dependent *sharing radius*, clean inputs bias the model toward *local* collaboration, and global signals can be induced by either the β –KL/prior pathway (near-uniform contraction of latent distances, with collapse risk if over-used) or by input masking (stochastic neighborhood mixing with potential drift). Guided by these insights, we introduced *Personalized Item Alignment* (PIA), a training-only regularizer that attaches learnable item anchors and softly pulls masked encodings toward each user's anchor centroid. PIA preserves instance information, stabilizes the geometry under masking, and promotes *semantically grounded* global mixing without adding test-time overhead. Empirically, PIA improves over vanilla VAE–CF on standard benchmarks and in an A/B test on a large-scale streaming platform, with ablations across user-activity strata and latent-space visualizations corroborating the theory.

Limitations. The benefits of capturing global collaborative signals still depend heavily on how well the input masking is designed (most current research relies on Bernoulli masking). If the masking is too noisy, even with alignment mechanisms, the model may struggle to learn meaningful representations. Therefore, a promising direction for future research is to explore more effective masking strategies that better support global collaboration.

REFERENCES

- Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
 - Sergej G Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.
 - Monroe D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28:1–47, 1975.
 - Werner Fenchel. On conjugate convex functions. Canadian Journal of Mathematics, 1:73-77, 1949.
 - Zhiqiang Guo, Guohui Li, Jianjun Li, and Huaicong Chen. Topicvae: Topic-aware disentanglement representation learning for enhanced recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 511–520, 2022.
 - Zhiqiang Guo, Guohui Li, Jianjun Li, Chaoyang Wang, and Si Shi. Dualvae: Dual disentangled variational autoencoder for recommendation. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pp. 571–579. SIAM, 2024.
 - Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.
 - Nicholas J. Higham and Lijing Lin. Matrix functions: A short course. MIMS EPrint 2013.73, Manchester Institute for Mathematical Sciences, The University of Manchester, 2013. Theorem 3.4(c): if X commutes with A, then X commutes with f(A).
 - Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In 2008 Eighth IEEE international conference on data mining, pp. 263–272. Ieee, 2008.
 - Hisham Husain and Julien Monteil. Geometric collaborative filtering with convergence. *Proceedings* of The 28th International Conference on Artificial Intelligence and Statistics, 2024.
 - Giannis Karamanolakis, Kevin Raji Cherian, Ananth Ravi Narayan, Jie Yuan, Da Tang, and Tony Jebara. Item recommendation with variational autoencoders and heterogeneous priors. In *Proceedings of the 3rd workshop on deep learning for recommender systems*, pp. 10–14, 2018.
 - Daeryong Kim and Bongwon Suh. Enhancing vaes for collaborative filtering: flexible priors & gating mechanisms. In *Proceedings of the 13th ACM conference on recommender systems*, pp. 403–407, 2019.
 - Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.
 - Xiucheng Li, Jin Yao Chin, Yile Chen, and Gao Cong. Sinkhorn collaborative filtering. In *Proceedings of the web conference 2021*, pp. 582–592, 2021.
 - Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pp. 689–698, 2018.
 - Sam Lobel, Chunyuan Li, Jianfeng Gao, and Lawrence Carin. Towards amortized ranking-critical training for collaborative filtering. *ICLR* 2020, 2019.
 - Teng Long, Yanshuai Cao, and Jackie Chi Kit Cheung. Preventing posterior collapse in sequence vaes with pooling. *ICLR*, 2019.
- Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. Learning disentangled representations for recommendation. *Advances in neural information processing systems*, 32, 2019.
 - Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

- Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20, 2007.
- Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In Proceedings of KDD cup and workshop, volume 2007, pp. 5–8, 2007.
 - Nikhil Rao, Hsiang-Fu Yu, Pradeep K Ravikumar, and Inderjit S Dhillon. Collaborative filtering with graph information: Consistency and scalable methods. *Advances in neural information processing systems*, 28, 2015.
 - Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pp. 175–186, New York, NY, USA, 1994. Association for Computing Machinery. doi: 10.1145/192844.192905.
 - Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.
 - Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pp. 1–35. Springer, 2010.
 - R. Tyrrell Rockafellar. Convex Analysis. Princeton University Press, 1970.
 - R. Sh. Rubinstein. On the problem of optimal transportation of masses. *Doklady Akademii Nauk SSSR*, 191:1240–1243, 1970.
 - Suvash Sedhain, Hung Bui, Jaya Kawale, Nikos Vlassis, Branislav Kveton, Aditya Krishna Menon, Trung Bui, and Scott Sanner. Practical linear models for large-scale one-class collaborative filtering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 3854–3860, 2016.
 - Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I Nikolenko. Recvae: A new variational autoencoder for top-n recommendations with implicit feedback. In *Proceedings of the 13th international conference on web search and data mining*, pp. 528–536, 2020.
 - Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*, pp. 3251–3257, 2019.
 - Harald Steck. Autoencoders that don't overfit towards the identity. *Advances in Neural Information Processing Systems*, 33:19598–19608, 2020.
 - Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009(1):421425, 2009.
 - Jakub Tomczak and Max Welling. Vae with a vampprior. In *International conference on artificial intelligence and statistics*, pp. 1214–1223. PMLR, 2018.
 - Nhu-Thuat Tran and Hady W Lauw. Varium: Variational autoencoder for multi-interest representation with inter-user memory. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 156–164, 2025.
 - Quoc-Tuan Truong, Aghiles Salah, and Hady W Lauw. Bilateral variational autoencoder for collaborative filtering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 292–300, 2021.
 - Vojtěch Vančura and Pavel Kordík. Deep variational autoencoder with shallow parallel path for top-n recommendation (vasp). In *International Conference on Artificial Neural Networks*, pp. 138–149. Springer, 2021.
 - Joojo Walker, Fan Zhou, Edward Y Baagyere, Emmanuel Ahene, and Fengli Zhang. Implicit optimal variational collaborative filtering. *Complex & Intelligent Systems*, 8(5):4369–4384, 2022.

- Siyu Wang, Xiaocong Chen, Quan Z Sheng, Yihong Zhang, and Lina Yao. Causal disentangled variational auto-encoder for preference understanding in recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 1874–1878, 2023.
- Xin Wang, Hong Chen, Yuwei Zhou, Jianxin Ma, and Wenwu Zhu. Disentangled representation learning for recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):408–424, 2022.
- Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. Collaborative denoising autoencoders for top-n recommender systems. In *Proceedings of the ninth ACM international conference on web search and data mining*, pp. 153–162, 2016.
- W. H. Young. Über die Form der Additiven Funktionen. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, pp. 367–377, 1912.
- Guijuan Zhang, Yang Liu, and Xiaoning Jin. Adversarial variational autoencoder for top-n recommender systems. In 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), pp. 853–856. IEEE, 2018.
- Yaochen Zhu, Chao Wan, Harald Steck, Dawen Liang, Yesu Feng, Nathan Kallus, and Jundong Li. Collaborative retrieval for large language model-based conversational recommender systems. In *Proceedings of the ACM on Web Conference* 2025, pp. 3323–3334, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3696410.3714908.

THE USE OF LARGE LANGUAGE MODELS

We used a large language model (ChatGPT) to help with editing this paper. It was only used for simple tasks such as fixing typos, rephrasing sentences for clarity, and improving word choice. All ideas, experiments, and analyses were done by the authors, and the use of LLMs does not affect the reproducibility of our work.

We also used ChatGPT to assist with proof verification and theorem refinement. Our workflow involved providing initial drafts to ChatGPT, which would then suggest improvements to the mathematical presentation and formatting. We subsequently edited and refined these suggestions.

APPENDIX

This supplementary material provides a summary of common notations, detailed experimental settings, and proofs for the theoretical results stated in the main paper. It is organized as follows:

- We summarize common notation in Section A.
- We present the Related Work in Section B.
- The pseudo-code of the algorithm is provided in Section C.
- Detailed experimental settings and implementation details are described in Section D.
- We present all proofs relevant to the theory developed in our paper in Section E.

A NOTATION SUMMARY

Table 4: Table of Notations

Symbol	Description			
Users and Items Input Data				
U, I $\mathbf{x}_u = [\mathbf{x}_{u1}, \mathbf{x}_{u2}, \dots, \mathbf{x}_{uI}]$	Number of users and number of items I -dimensional binary vector (the u -th row of \mathbf{X}); $\mathbf{x}_{ui} = 1$ implies that user u has a positive interaction with item i ; $\mathbf{x}_{ui} = 0$ indicates otherwise			
$\mathbf{b} \in \{0, 1\}^{I}$ $\mathbf{x}_{h} = \mathbf{x} \odot \mathbf{b}$ $S_{\mathbf{x}} = \{ \forall i \leq I : \mathbf{x}_{i} = 1 \}$	a binary mask, i.e., $\mathbf{b} \sim \mathrm{Bern}(\rho)^I$ user's partial interaction history a set of positive items from user \mathbf{x}			
VAE Models				
$egin{aligned} \phi, heta \ q_\phi, p_ heta \ \mathbf{z} \end{aligned}$ $\mathbf{e}_i \in \mathbb{R}^d$	VAE encoder p_{θ} and decoder parameters q_{ϕ} ϕ -parameterized and θ -parameterized neural networks Latent space of the VAE, e.g., $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = q_{\phi}(\mathbf{x}_h), \ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \ \mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \boldsymbol{\sigma}$ learnable item embedding in latent space (same dimension as \mathbf{z}) and $E = \{\mathbf{e}_i \in \mathbb{R}^d\}_{i=1}^I$			
Theoretical Constants and Bounds				
$N_{\delta}(u) = \{ v : \ \mathbf{x}_u - \mathbf{x}_v\ _1 \le \delta \}$ $\ \cdot\ _1$ $W_1(\cdot, \cdot)$	input-space neighborhood L_1 norm (sum of absolute values) 1-Wasserstein distance			

B RELATED WORKS

Motivated by our theoretical analysis, both β -weighted KL regularization and input masking can promote *global* collaboration, albeit with different trade-offs. Prior work has largely focused on

controlling the KL term, primarily along two lines: (i) scheduling the β factor in the regularizer and (ii) adopting more flexible priors.

 β -scheduling. Liang et al. (Liang et al., 2018) introduce a β -scaling factor to modulate the strength of the regularization $\mathrm{KL}(q_{\phi}(\mathbf{z}\mid\mathbf{x})\parallel p(\mathbf{z}))$, while Long et al. (Long et al., 2019) propose gradually increasing this weight over training to mitigate posterior collapse.

Flexible priors and architectures. Several works replace the standard normal prior with richer alternatives to better match the data. Examples include VampPrior and its hierarchical variants (e.g., HVamp) (Tomczak & Welling, 2018; Kim & Suh, 2019), as well as implicit or learned priors (Walker et al., 2022). RecVAE (Shenbin et al., 2020) combines a redesigned encoder–decoder, a composite prior, input-dependent $\beta(\mathbf{x})$ rescaling, alternating training, and a non-denoising decoder. Other lines incorporate user-dependent priors (Karamanolakis et al., 2018) or impose an arbitrary target prior via adversarial training (Zhang et al., 2018).

To our knowledge, ours is the first work to systematically analyze the collaboration mechanisms in VAE-based CF, showing that both β -weighted KL regularization and input masking can promote *global* collaboration. **In contrast to prior works**, guided by our theoretical analysis, we propose a regularization scheme that addresses the issues induced by input masking, mitigating the loss of local collaboration while preserving its benefits for global alignment.

C ALGORITHM

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719 720 721

722 723

724

725

726

727

728

729 730 731 The model is trained by optimizing the objective defined in Eq. (8). However, for the hyperparameter $\lambda_{\text{alignment}}$, which controls the strength of the personalized item alignment regularization, instead of consider it as fixed hyper-parameter, we gradually increase its value during training .Specifically, we use the validation set to monitor whether the latent space is getting trapped in a local optimum i.e., when the validation performance does not improve after ρ consecutive epochs. If such a case is detected, we increase $\lambda_{\text{alignment}}$ by a scaling factor λ_{scale} , as detailed in Algorithm 1.

Algorithm 1 Personalized Item Alignment VAE

```
732
            1: Initialize:
733
               Models: encoder q_{\phi}, sequence of invertible transformations f_1, ..., f_T, decoder p_{\theta} and item-
734
               embeddings E.
               Hyper-parameters: \lambda_A, \lambda_{\text{scale}} and \rho.
735
               Variables: best\_val\_epoch := 0, best\_val\_ndcg := 0
736
           2: for epoch in n_epochs do
737
                  // Training
           3:
738
           4:
                  for iter in iterations do
739
           5:
                     Sample a mini-batch x
740
                     \mathbf{z}_0 \sim q_{\phi}(\cdot \mid \mathbf{x}) // using reparametrization trick
           6:
741
           7:
                     \mathbf{z}_T = (f_T \circ \dots \circ f_1) (\mathbf{z}_0)
742
           8:
                     Update q_{\phi}, f_1, ... f_T, p_{\theta} and E based on \mathcal{L}_{PIA-VAE}(\mathbf{x}; \theta, \phi, f_1, ..., f_T) in Equation (8)
743
           9:
                  end for
744
           10:
                  //Validation
745
          11:
                  Compute nDCG@K on validation set: epoch_ndcg
746
          12:
                  if epoch_ndcg > best_val_ndcg then
          13:
                     best\_val\_ndcg := epoch\_ndcg
747
                     best\_val\_epoch := epoch
          14:
748
          15:
                  end if
749
                  // Increase \lambda_A if training is stuck in a local optimum, i.e., when the validation performance
          16:
750
                  does not improve after \rho consecutive epochs.
751
          17:
                  if best\_val\_epoch < epoch + \rho then
752
          18:
                     \lambda_{A} := \lambda_{scale} \times \lambda_{A}
753
          19:
                  end if
754
          20: end for
755
          21: Return: the optimal encoder q_{\phi} and decoder p_{\theta} at best_val_epoch.
```

D EXPERIMENTAL SETTINGS

D.1 DATASET

We validate our analysis using three real-world recommendation datasets: MovieLens-20M², Netflix³ and Million Song (MSD) ⁴, where each record consists of a user-item pair along with a rating that the user has given to the item. We follow the preprocessing procedure from MultVAE (Liang et al., 2018). For MovieLens-20M and Netflix, we retain users who have rated at least five movies and treat ratings of four or higher as positive interactions. For MSD, we keep only users with at least 20 songs in their listening history and songs that have been listened to by at least 200 users.

The user data is split into training, validation, and test sets as presented in Table 5.

Table 5: Dataset statistics

	ML-20M	Netflix	MSD
# of users	136,677	463,435	571,355
# of items	20,108	17,769	41,140
# of interactions	10.0M	56.9M	33.6M
% of interactions	0.36%	0.69%	0.14%
# of held-out users	10,000	40,000	50,000

For every user in the training set, we utilize all interaction history, whereas for users in the validation or test set, a fraction of the history (80%) is used to predict the remaining interaction.

D.2 IMPLEMENTATION DETAILS

Hyperparameters. We use a batch size of 500 and train the model for 200 epochs using the Adam optimizer with a learning rate of 1×10^{-3} across all experiments. The specific hyperparameters $\lambda_{\rm A}$, ρ , and $\lambda_{\rm scale}$ are selected based on validation performance. We set $\rho=5$, $\lambda_{\rm scale}=2$ and $\lambda_{\rm A}=8$ consistently across all experiments, as the model is not particularly sensitive to these values.

Implementation details and instructions for reproducing all experimental results and visualizations are provided in the accompanying source code.

D.3 BASELINES

We have selected following models as baselines:

- Matrix factorization (MF); we consider MF trained with ALS with uniform weights (Hu et al., 2008), which is a simple and computationally efficient baseline, and also weighted matrix factorization (wMF) (Hu et al., 2008);
- Regularization based on item-item interactions; here we selected GRALS (Rao et al., 2015) that employs graph regularization;
- *Linear models*; we have chosen full-rank models EASE Steck (2019) and a low-rank model PLRec (Sedhain et al., 2016);
- *Nonlinear autoencoders*; here we consider the shallow autoencoder CDAE (Wu et al., 2016), variational autoencoder MultVAE (Liang et al., 2018), and its successors: RaCT (Lobel et al., 2019) and RecVAE (Shenbin et al., 2020).

E THEORETICAL DEVELOPMENT

In this Section, we present all proofs relevant to theory developed in our paper.

²https://grouplens.org/datasets/movielens/20m/

³https://www.kaggle.com/netflix-inc/netflix-prize-data

⁴http://millionsongdataset.com

E.1 PROOF OF THEOREM 2.3

Theorem E.1 (Latent- W_1 Sharing Radius). Assume decoder gradient is Lipschitz in z such that for any given $\mathbf{z} \sim q_{\phi}(\cdot \mid \mathbf{x}_u)$, uniformly in \mathbf{x} and for all \mathbf{z}' , $\|\nabla_{\theta}\ell_{\theta}(\mathbf{x}, \mathbf{z}) - \nabla_{\theta}\ell_{\theta}(\mathbf{x}, \mathbf{z}')\| \leq L_{\theta z} \|\mathbf{z} - \mathbf{z}'\|$. Let

$$\mathcal{L}_u(\theta) := \mathbb{E}_{q_{\phi}(\cdot|\mathbf{x}_u)}[\ell_{\theta}(\mathbf{x}_u, \mathbf{z})], \qquad g_u(\theta) := \nabla_{\theta} \mathcal{L}_u(\theta),$$

and define the content-mismatch term

$$\Delta_x(u,v) := \left\| \mathbb{E}_{q_{\phi}(\cdot|\mathbf{x}_v)} \left[\nabla_{\theta} \ell_{\theta}(\mathbf{x}_u, \mathbf{z}) - \nabla_{\theta} \ell_{\theta}(\mathbf{x}_v, \mathbf{z}) \right] \right\|.$$

For one SGD step on user v with step size $\eta > 0$, setting $\theta^+ = \theta - \eta \, q_v(\theta)$, we have:

$$\mathcal{L}_{u}(\theta^{+}) - \mathcal{L}_{u}(\theta) \leq -\eta \|g_{u}(\theta)\|^{2} + \eta \|g_{u}(\theta)\| \Big(L_{\theta z} W_{1}(q_{\phi}(\cdot \mid \mathbf{x}_{u}), q_{\phi}(\cdot \mid \mathbf{x}_{v})) + \Delta_{x}(u, v) \Big) + O(\eta^{2}).$$

In particular, the step on user v strictly decreases \mathcal{L}_u to first order whenever

$$W_1(q_{\phi}(\cdot \mid \mathbf{x}_u), q_{\phi}(\cdot \mid \mathbf{x}_v)) < r_{\text{share}}(u, v; \theta) := \frac{\|g_u(\theta)\| - \Delta_x(u, v)}{L_{\theta z}}.$$

Proof. For shorthand, write $q_u := q_{\phi}(\cdot \mid \mathbf{x}_u)$ and $q_v := q_{\phi}(\cdot \mid \mathbf{x}_v)$.

A first-order Taylor expansion gives

$$\mathcal{L}_u(\theta - \eta g_v) = \mathcal{L}_u(\theta) - \eta \langle g_u(\theta), g_v(\theta) \rangle + O(\eta^2). \tag{11}$$

Then, we decompose the inner product:

$$-\langle g_u, g_v \rangle = -\|g_u\|^2 - \langle g_u, g_v - g_u \rangle \le -\|g_u\|^2 + \|g_u\| \|g_v - g_u\|.$$

Next, we bound

$$||g_{v} - g_{u}|| = \left\| \mathbb{E}_{q_{v}} \nabla_{\theta} \ell_{\theta}(\mathbf{x}_{v}, \mathbf{z}) - \mathbb{E}_{q_{u}} \nabla_{\theta} \ell_{\theta}(\mathbf{x}_{u}, \mathbf{z}) \right\|$$

$$\leq \underbrace{\left\| \mathbb{E}_{q_{v}} \left[\nabla_{\theta} \ell_{\theta}(\mathbf{x}_{u}, \mathbf{z}) - \nabla_{\theta} \ell_{\theta}(\mathbf{x}_{v}, \mathbf{z}) \right] \right\|}_{=\Delta_{x}(u, v)} + \underbrace{\left\| \mathbb{E}_{q_{v}} \nabla_{\theta} \ell_{\theta}(\mathbf{x}_{u}, \mathbf{z}) - \mathbb{E}_{q_{u}} \nabla_{\theta} \ell_{\theta}(\mathbf{x}_{u}, \mathbf{z}) \right\|}_{\leq L_{\theta z} W_{1}(q_{v}, q_{u})}.$$

For the last inequality, let $\psi(\mathbf{z}) := \nabla_{\theta} \ell_{\theta}(\mathbf{x}_u, \mathbf{z})$. By assumption decoder gradient is Lipschitz, ψ is $L_{\theta z}$ -Lipschitz in \mathbf{z} ; for any coupling π of (q_u, q_v) ,

$$\left\| \mathbb{E}_{q_u}[\psi] - \mathbb{E}_{q_v}[\psi] \right\| = \left\| \int (\psi(\mathbf{z}) - \psi(\mathbf{z}')) \, d\pi(\mathbf{z}, \mathbf{z}') \right\| \le L_{\theta z} \int \|\mathbf{z} - \mathbf{z}'\| \, d\pi \le L_{\theta z} \, W_1(q_u, q_v).$$

Combining the bounds and substituting into Eq. (11) yields the stated result, and the strict-decrease condition follows by inspecting the coefficient of η .

E.2 Proof of Theorem 2.5

Before proving the theorem, we recall some inequalities:

• The Kantorovich-Rubinstein dual form (Rubinstein, 1970):

$$W_1(\mu,\nu) = \sup_{\substack{f: \mathcal{X} \to \mathbb{R} \\ \text{Lip}(f) \le 1}} \int_{\mathcal{X}} f \, d(\mu - \nu), \tag{12}$$

• The Donsker-Varadhan variational formula (Donsker & Varadhan, 1975): for any measurable g with $\int e^g dp < \infty$,

$$KL(\nu||p) \ge \int g \, d\nu - \log \int e^g \, dp. \tag{13}$$

• The Bobkov–Götze/Talagrand $T_1(C)$ inequality (Bobkov & Götze, 1999) is known to be equivalent to the following *sub-Gaussian moment generating function (mgf) bound* for Lipschitz functions: p satisfies the $T_1(C)$ inequality with constant C > 0, then for every 1-Lipschitz f and every $\lambda \in \mathbb{R}$,

$$\log \int \exp(\lambda(f - \mathbb{E}_p f)) dp \le \frac{C\lambda^2}{2}. \tag{14}$$

Lemma E.2. Assume the encoder is L_{ϕ} -Lipschitz, i.e., $\|q_{\phi}(\mathbf{x}_u) - q_{\phi}(\mathbf{x}_v)\| \le L_{\phi} \|\mathbf{x}_u - \mathbf{x}_v\|$ for all $\mathbf{x}_u, \mathbf{x}_v \in \{0, 1\}^I$ then $W_1(q_{\phi}(\cdot \mid \mathbf{x}_u), q_{\phi}(\cdot \mid \mathbf{x}_v)) \le L_{\phi} \|\mathbf{x}_v - \mathbf{x}_u\|_1$ for all $\mathbf{x}_u, \mathbf{x}_v \in \{0, 1\}^I$.

Proof. First, since $q_{\phi}(\mathbf{z}|\mathbf{x}_i) = \mathcal{N}\Big(\mu_{\phi}(\mathbf{x}_i), \operatorname{diag}(\sigma_{\phi}^2(\mathbf{x}_i))\Big)$, the Wasserstein-2 distance $W_2(q_{\phi}(\mathbf{z}|\mathbf{x}_1), q_{\phi}(\mathbf{z}|\mathbf{x}_2))$ has the following closed form:

$$W_2(q_{\phi}(z|\mathbf{x}_u), q_{\phi}(z|\mathbf{x}_v))^2 = \|\mu_{\phi}(\mathbf{x}_u) - \mu_{\phi}(\mathbf{x}_v)\|^2 + \|\sigma_{\phi}(\mathbf{x}_u) - \sigma_{\phi}(\mathbf{x}_v)\|^2, \tag{15}$$

which, combined with the definition $Q_{\phi}(x) = \begin{bmatrix} \mu_{\phi}(\mathbf{x}) \\ \sigma_{\phi}(\mathbf{x}) \end{bmatrix}$, yields

$$||Q_{\phi}(\mathbf{x}_u) - Q_{\phi}(\mathbf{x}_v)||^2 = W_2(q_{\phi}(\mathbf{z}|\mathbf{x}_1), q_{\phi}(\mathbf{z}|\mathbf{x}_2))^2.$$
(16)

Since Q_{ϕ} is L_{ϕ} -Lipschitz continuous, we have $\|Q_{\phi}(\mathbf{x}_u) - Q_{\phi}(\mathbf{x}_v)\| \le L_{\phi} \|\mathbf{x}_u - \mathbf{x}_v\|$, and

$$W_2(q_{\phi}(\mathbf{z} \mid \mathbf{x}_u), q_{\phi}(\mathbf{z} \mid \mathbf{x}_v)) \le L_{\phi} \|\mathbf{x}_u - \mathbf{x}_v\|. \tag{17}$$

Since
$$W_1 \leq W_2$$
, we have $W_1(q_{\phi}(\mathbf{z} \mid \mathbf{x}_u), q_{\phi}(\mathbf{z} \mid \mathbf{x}_v)) \leq L_{\phi} ||\mathbf{x}_u - \mathbf{x}_v||$.

Theorem E.3. Assume $p_{\theta}(\mathbf{x} \mid \mathbf{z})$ is a regular exponential family with sufficient statistics $T(\mathbf{x})$, natural parameter $\eta(\mathbf{z})$ and log-partition A. Let $\alpha(\mathbf{z}) = \frac{q_{\phi}(\mathbf{z}|\mathbf{x}_1)}{q_{\phi}(\mathbf{z}|\mathbf{x}_1) + q_{\phi}(\mathbf{z}|\mathbf{x}_2)}$ on $\{q_1 + q_2 > 0\}$. Then

$$\min_{\eta(\cdot)} \left\{ \sum_{i=1}^{2} \mathbb{E}_{q_{\phi}(\cdot \mid \mathbf{x}_{i})} \left[-\log p_{\theta}(\mathbf{x}_{i} \mid \mathbf{z}) \right] + \beta \sum_{i=1}^{2} \text{KL} \left(q_{\phi}(\mathbf{z} \mid \mathbf{x}_{i}) \parallel p(\mathbf{z}) \right) \right\}$$
(18)

$$= C + \int (q_{\phi}(\mathbf{z} \mid \mathbf{x}_{1}) + q_{\phi}(\mathbf{z} \mid \mathbf{x}_{2})) \Delta_{A^{*}}(\mathbf{x}_{1}, \mathbf{x}_{2}; \alpha(\mathbf{z})) d\mathbf{z} + \beta \sum_{i=1}^{2} KL(q_{\phi}(\mathbf{z} \mid \mathbf{x}_{i}) \parallel p(\mathbf{z})),$$

where A^* is the convex conjugate of A, C is independent of $\eta(\cdot)$, and

$$\Delta_{A^*}(\mathbf{x}_1, \mathbf{x}_2; \alpha) = \alpha A^*(T(\mathbf{x}_1)) + (1 - \alpha) A^*(T(\mathbf{x}_2)) - A^*(\alpha T(\mathbf{x}_1) + (1 - \alpha)T(\mathbf{x}_2)) \ge 0,$$

with equality iff either $(q_{\phi}(\cdot \mid \mathbf{x}_1), q_{\phi}(\cdot \mid \mathbf{x}_2) = 0$ almost everywhere or $T(\mathbf{x}_1) = T(\mathbf{x}_2)$.

Additionally, assume the prior p satisfies the Bobkov–Götze/Talagrand $T_1(C)$ inequality (Bobkov & Götze, 1999) with constant C > 0 (e.g., Normal prior $p = \mathcal{N}(0, \sigma^2 I)$):

$$W_1(q_{\phi}(\mathbf{z} \mid \mathbf{x}_1), q_{\phi}(\mathbf{z} \mid \mathbf{x}_2)) \leq \sqrt{2C \operatorname{KL}(q_{\phi}(\mathbf{z} \mid \mathbf{x}_1) \parallel p(\mathbf{z}))} + \sqrt{2C \operatorname{KL}(q_{\phi}(\mathbf{z} \mid \mathbf{x}_2) \parallel p(\mathbf{z}))}.$$
(19)

Proof of Eq. (18). We write the conditional likelihood in its exponential-family form

$$p_{\theta}(\mathbf{x} \mid \mathbf{z}) = h(\mathbf{x}) \exp(\langle T(\mathbf{x}), \eta(\mathbf{z}) \rangle - A(\eta(\mathbf{z}))),$$

so that

$$-\log p_{\theta}(\mathbf{x}_i \mid \mathbf{z}) = A(\eta(\mathbf{z})) - \langle T(\mathbf{x}_i), \eta(\mathbf{z}) \rangle - \log h(\mathbf{x}_i).$$

Let $q_i(\mathbf{z}) := q_{\phi}(\mathbf{z} \mid \mathbf{x}_i), i = 1, 2$, and define

$$Q(\mathbf{z}) := q_1(\mathbf{z}) + q_2(\mathbf{z}), \qquad \alpha(\mathbf{z}) := \begin{cases} \frac{q_1(\mathbf{z})}{Q(\mathbf{z})}, & Q(\mathbf{z}) > 0, \\ \text{arbitrary in } [0, 1], & Q(\mathbf{z}) = 0, \end{cases}$$

and the α -mixture of sufficient statistics

$$\overline{T}_{\alpha}(\mathbf{z}) := \alpha(\mathbf{z}) T(\mathbf{x}_1) + (1 - \alpha(\mathbf{z})) T(\mathbf{x}_2).$$

(Any choice of α on $\{Q=0\}$ is immaterial since all integrands below are multiplied by Q.)

Summing the two reconstruction terms and using $\int q_i = 1$,

$$\begin{split} \sum_{i=1}^2 \mathbb{E}_{q_i} \big[-\log p_{\theta}(\mathbf{x}_i \mid \mathbf{z}) \big] &= \int \Big(\sum_{i=1}^2 q_i(\mathbf{z}) \Big) \, A(\eta(\mathbf{z})) \, d\mathbf{z} - \int \big\langle T(\mathbf{x}_1) q_1(\mathbf{z}) + T(\mathbf{x}_2) q_2(\mathbf{z}), \, \eta(\mathbf{z}) \big\rangle d\mathbf{z} - \sum_{i=1}^2 \log h(\mathbf{x}_i) \\ &= \int Q(\mathbf{z}) \Big(A(\eta(\mathbf{z})) - \big\langle \overline{T}_{\alpha}(\mathbf{z}), \eta(\mathbf{z}) \big\rangle \Big) \, d\mathbf{z} - \sum_{i=1}^2 \log h(\mathbf{x}_i). \end{split}$$

Recall the convex conjugate A^* and Fenchel–Young inequality (Fenchel, 1949; Young, 1912; Rockafellar, 1970):

$$A^*(y) := \sup_{\eta} \{ \langle y, \eta \rangle - A(\eta) \}, \qquad A(\eta) - \langle y, \eta \rangle \ge -A^*(y),$$

with equality when $y \in \partial A(\eta)$; in the regular (Legendre) case (Rockafellar, 1970), A is essentially smooth and strictly convex, so $y = \nabla A(\eta)$ is the unique equality condition and A^* is strictly convex on its (convex) effective domain.

Because the integrand is separable in \mathbf{z} and $Q(\mathbf{z}) \geq 0$, minimizing the integral over all measurable $\eta(\cdot)$ reduces to pointwise minimization:

$$\min_{\eta(\cdot)} \int Q(\mathbf{z}) \Big(A(\eta(\mathbf{z})) - \langle \overline{T}_{\alpha}(\mathbf{z}), \eta(\mathbf{z}) \rangle \Big) d\mathbf{z} = \int Q(\mathbf{z}) \min_{\eta} \Big\{ A(\eta) - \langle \overline{T}_{\alpha}(\mathbf{z}), \eta \rangle \Big\} d\mathbf{z}
= - \int Q(\mathbf{z}) A^* \Big(\overline{T}_{\alpha}(\mathbf{z}) \Big) d\mathbf{z}.$$

Hence

$$\min_{\eta(\cdot)} \sum_{i=1}^{2} \mathbb{E}_{q_i} \left[-\log p_{\theta}(\mathbf{x}_i \mid \mathbf{z}) \right] = -\int Q(\mathbf{z}) A^* \left(\overline{T}_{\alpha}(\mathbf{z}) \right) d\mathbf{z} - \sum_{i=1}^{2} \log h(\mathbf{x}_i).$$
 (20)

(When $\overline{T}_{\alpha}(\mathbf{z})$ lies in the interior of $\operatorname{dom}(A^*)$, the minimizer is $\eta^*(\mathbf{z}) = \nabla A^*(\overline{T}_{\alpha}(\mathbf{z}))$; equivalently, $\nabla A(\eta^*(\mathbf{z})) = \overline{T}_{\alpha}(\mathbf{z})$.)

Add and subtract the quantity

$$\int Q(\mathbf{z}) \Big(\alpha(\mathbf{z}) A^* \big(T(\mathbf{x}_1) \big) + (1 - \alpha(\mathbf{z})) A^* \big(T(\mathbf{x}_2) \big) \Big) d\mathbf{z} = \sum_{i=1}^2 A^* \big(T(\mathbf{x}_i) \big) \underbrace{\int q_i(\mathbf{z}) d\mathbf{z}}_{i,i},$$

and collect the terms independent of $\eta(\cdot)$ into the constant

$$C := -\sum_{i=1}^{2} \left(\log h(\mathbf{x}_i) + A^* (T(\mathbf{x}_i)) \right).$$

Using Eq. (20), we obtain

$$\min_{\eta(\cdot)} \sum_{i=1}^{2} \mathbb{E}_{q_{i}} \left[-\log p_{\theta}(\mathbf{x}_{i} \mid \mathbf{z}) \right] = C + \int Q(\mathbf{z}) \left(\alpha(\mathbf{z}) A^{*}(T(\mathbf{x}_{1})) + (1 - \alpha(\mathbf{z})) A^{*}(T(\mathbf{x}_{2})) - A^{*}(\overline{T}_{\alpha}(\mathbf{z})) \right) d\mathbf{z}$$

$$= C + \int Q(\mathbf{z}) \Delta_{A^{*}}(\mathbf{x}_{1}, \mathbf{x}_{2}; \alpha(\mathbf{z})) d\mathbf{z}, \tag{21}$$

where

$$\Delta_{A^*}(\mathbf{x}_1, \mathbf{x}_2; \alpha) := \alpha A^*(T(\mathbf{x}_1)) + (1 - \alpha) A^*(T(\mathbf{x}_2)) - A^*(\alpha T(\mathbf{x}_1) + (1 - \alpha)T(\mathbf{x}_2)).$$

By convexity of A^* , $\Delta_{A^*}(\mathbf{x}_1, \mathbf{x}_2; \alpha) \geq 0$ for all $\alpha \in [0, 1]$ (Jensen gap). In the regular (Legendre) case, A^* is strictly convex on its effective domain, so $\Delta_{A^*}(\mathbf{x}_1, \mathbf{x}_2; \alpha) = 0$ iff either

- $T(\mathbf{x}_1) = T(\mathbf{x}_2)$, in which case the three arguments of A^* coincide, or
- $\alpha \in \{0, 1\}$, i.e. $Q(\mathbf{z})\alpha(\mathbf{z})(1 \alpha(\mathbf{z})) = 0$ for Q-a.e. \mathbf{z} . Equivalently, the posteriors have disjoint supports w.r.t. the measure $Q(\mathbf{z}) d\mathbf{z}$ (on each point with Q > 0 exactly one of q_1, q_2 is zero).

(If A^* were affine on the segment $[T(\mathbf{x}_1), T(\mathbf{x}_2)]$, equality could also occur with $0 < \alpha < 1$, but strict convexity rules this out unless $T(\mathbf{x}_1) = T(\mathbf{x}_2)$.)

Finally, the regularizer $\beta \sum_{i=1}^{2} \mathrm{KL}(q_{\phi}(\mathbf{z} \mid \mathbf{x}_{i}) \parallel p(\mathbf{z}))$ does not depend on $\eta(\cdot)$, hence it carries through unchanged. Combining with Eq. (21) yields Eq. (18).

Proof of Eq. (19). By the triangle inequality for W_1 ,

$$W_1(q_{\phi}(\cdot \mid \mathbf{x}_1), q_{\phi}(\cdot \mid \mathbf{x}_2)) \leq W_1(q_{\phi}(\cdot \mid \mathbf{x}_1), p) + W_1(q_{\phi}(\cdot \mid \mathbf{x}_2), p). \tag{22}$$

Thus it suffices to show that for any probability ν with $\mathrm{KL}(\nu || p) < \infty$,

$$W_1(\nu, p) \le \sqrt{2C \operatorname{KL}(\nu \| p)}. \tag{23}$$

Fix a 1-Lipschitz f and $\lambda > 0$. Apply Eq. (13) with $g = \lambda (f - \mathbb{E}_p f)$ to obtain

$$\mathrm{KL}(\nu \| p) \geq \lambda \int (f - \mathbb{E}_p f) d\nu - \log \int \exp(\lambda (f - \mathbb{E}_p f)) dp.$$

Using the mgf bound Eq. (14) gives

$$\int f \, d(\nu - p) \, \leq \, \frac{1}{\lambda} \operatorname{KL}(\nu \| p) \, + \, \frac{C\lambda}{2}.$$

Optimizing the right-hand side over $\lambda > 0$ yields the minimizer $\lambda^* = \sqrt{2 \operatorname{KL}(\nu || p)/C}$, and the minimum value

$$\frac{1}{\lambda^{\star}} \operatorname{KL}(\nu \| p) + \frac{C\lambda^{\star}}{2} = \sqrt{2C \operatorname{KL}(\nu \| p)}.$$

Therefore, for every 1-Lipschitz f,

$$\int f d(\nu - p) \le \sqrt{2C \operatorname{KL}(\nu || p)}.$$

Taking the supremum over all 1-Lipschitz f and invoking Eq. (12) gives exactly Eq. (23).

Applying Eq. (23) to $\nu = q_{\phi}(\cdot \mid \mathbf{x}_1)$ and to $\nu = q_{\phi}(\cdot \mid \mathbf{x}_2)$ and combining with the triangle inequality in Eq. (22) yields Eq. (19).

E.3 Proof of Theorem 2.6

Theorem E.4 (Masked input: contraction and expansion). Let $\mathbf{x}_1, \mathbf{x}_2 \in \{0, 1\}^I$ be binary inputs. Let $b_{\mathbf{x}_1}, b_{\mathbf{x}_2} \sim \mathrm{Bern}(\rho)^I$ be independent masks and set $\mathbf{x}_1' = \mathbf{x}_1 \odot b_{\mathbf{x}_1}$ and $\mathbf{x}_2' = \mathbf{x}_2 \odot b_{\mathbf{x}_2}$. Write $h = \|\mathbf{x}_1 - \mathbf{x}_2\|_1$ and $s = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle$ (so h is the number of disagreeing coordinates and s the count of shared 1's). For any $\delta > 0$, define $T_{\delta} = \lceil \delta \rceil - 1$ and $U_{\delta} = \lceil \delta \rceil$. Let $D' := \|\mathbf{x}_1' - \mathbf{x}_2'\|_1$. Then:

Contraction.

$$\Pr[D' < \delta] \ge (\rho^2 + (1 - \rho)^2)^s \sum_{k=0}^{\min\{h, T_\delta\}} \binom{h}{k} \rho^k (1 - \rho)^{h-k}.$$
 (24)

Expansion.

$$\Pr[D' \ge \delta] \ge \sum_{k=U_s}^{s} {s \choose k} (2\rho(1-\rho))^k (1-2\rho(1-\rho))^{s-k}.$$
 (25)

Proof. Partition coordinates into

$$H := \{j : x_{1j} \neq x_{2j}\}, \quad |H| = h, \qquad S := \{j : x_{1j} = x_{2j} = 1\}, \quad |S| = s.$$

For $j \in H$, exactly one of (x_{1j}, x_{2j}) equals 1. After masking, the post-mask difference at j equals 1 iff the unique 1 is kept, with probability ρ . Hence

$$Y := \sum_{i \in H} \mathbf{1}\{\text{post-mask difference at } j = 1\} \sim \text{Binomial}(h, \rho).$$

For $j \in S$, both entries are 1. The post-mask difference equals $|b_{1j} - b_{2j}|$, which is 1 iff the masks disagree; this happens with probability $\Pr(b_{1j} \neq b_{2j}) = \rho(1-\rho) + (1-\rho)\rho = 2\rho(1-\rho)$. Thus

$$Z := \sum_{j \in S} |b_{1j} - b_{2j}| \sim \text{Binomial}(s, 2\rho(1-\rho)).$$

Independence of masks across coordinates implies $Y \perp Z$, and the masked distance decomposes as

$$D' = \|\mathbf{x}_1' - \mathbf{x}_2'\|_1 = Y + Z.$$

Contraction. Because D' is integer-valued, $D' < \delta$ is equivalent to $D' \leq T_{\delta}$. Consider the event $\mathcal{E} := \{Z = 0\} \cap \{Y \leq T_{\delta}\}$. On \mathcal{E} we have $D' = Y + Z \leq T_{\delta}$, hence

$$\Pr[D' < \delta] \ge \Pr[\mathcal{E}] = \Pr[Z = 0] \Pr[Y \le T_{\delta}]$$

by independence of Y and Z. Now $\Pr[Z=0] = (\Pr[b_{1j}=b_{2j}])^s = (\rho^2 + (1-\rho)^2)^s$, and

$$\Pr[Y \le T_{\delta}] = \sum_{k=0}^{\min\{h, T_{\delta}\}} \binom{h}{k} \rho^{k} (1 - \rho)^{h-k}.$$

Multiplying the two factors yields Eq. (24).

Expansion. Using D' = Y + Z with $Y \perp Z$ and $U_{\delta} = \lceil \delta \rceil$,

$$\Pr[D' \ge \delta] = \Pr[Y + Z \ge U_{\delta}] = \sum_{m=0}^{h} \Pr[Y = m] \Pr[Z \ge U_{\delta} - m]$$

$$= \sum_{m=0}^{h} \binom{h}{m} \rho^m (1-\rho)^{h-m} \sum_{k=\max\{U_{\delta}-m,0\}}^{s} \binom{s}{k} (2\rho(1-\rho))^k (1-2\rho(1-\rho))^{s-k},$$

Finally, since $D' = Y + Z \ge Z$,

$$\Pr[D' \ge \delta] \ge \Pr[Z \ge U_{\delta}] = \sum_{k=U_{\delta}}^{s} {s \choose k} (2\rho(1-\rho))^{k} (1 - 2\rho(1-\rho))^{s-k},$$

which gives Eq. (25). This completes the proof.

E.4 PROOF OF PROPOSITION 3.1

Proposition E.5. Assume the encoder posterior is diagonal-Gaussian, $q_{\phi}(\mathbf{z} \mid \mathbf{x}_h) = \mathcal{N}(\boldsymbol{\mu}_{\phi}(\mathbf{x}_h), \operatorname{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_h)))$. Let the item centroid be $\bar{\mathbf{e}}_{\mathbf{x}} := \frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S_{\mathbf{x}}} \mathbf{e}_i$, then

$$\mathcal{L}_{A}(\mathbf{x}_{h}, \mathbf{x}; \phi, E) = \underbrace{\|\boldsymbol{\mu}_{\phi}(\mathbf{x}_{h}) - \bar{\mathbf{e}}_{\mathbf{x}}\|^{2}}_{\text{align mean to item centroid}} + \underbrace{\operatorname{tr} \Sigma_{\phi}(\mathbf{x}_{h})}_{\text{variance shrinkage}} + \operatorname{const}(\mathbf{x}, E), \tag{26}$$

where $\Sigma_{\phi}(\mathbf{x}_h) = \operatorname{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_h))$ and $\operatorname{const}(\mathbf{x}, E) = \frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S_{\mathbf{x}}} \|\mathbf{e}_i\|_2^2 - \|\bar{\mathbf{e}}_{\mathbf{x}}\|_2^2$.

Proof. Write $q_{\phi}(\mathbf{z} \mid \mathbf{x}_h) = \mathcal{N}(\boldsymbol{\mu}_{\phi}(\mathbf{x}_h), \Sigma_{\phi}(\mathbf{x}_h))$ with $\Sigma_{\phi}(\mathbf{x}_h) = \operatorname{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_h))$, we have:

$$\mathbb{E}[\|\mathbf{z} - \mathbf{e}_i\|_2^2] = \|\boldsymbol{\mu}_{\phi}(\mathbf{x}_h) - \mathbf{e}_i\|_2^2 + \operatorname{tr} \Sigma_{\phi}(\mathbf{x}_h).$$

Averaging over $i \in S_{\mathbf{x}}$ yields

$$\mathcal{L}_{A}(\mathbf{x}_{h}, \mathbf{x}; \phi, E) = \frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S_{\mathbf{x}}} \|\boldsymbol{\mu}_{\phi}(\mathbf{x}_{h}) - \mathbf{e}_{i}\|_{2}^{2} + \operatorname{tr} \Sigma_{\phi}(\mathbf{x}_{h}).$$

Given the item centroid be $\bar{\mathbf{e}}_{\mathbf{x}} := \frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S_{\mathbf{x}}} \mathbf{e}_i$, then, we obtain

$$\frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S} \|\boldsymbol{\mu}_{\phi}(\mathbf{x}_h) - \mathbf{e}_i\|_2^2 = \|\boldsymbol{\mu}_{\phi}(\mathbf{x}_h) - \bar{\mathbf{e}}_{\mathbf{x}}\|_2^2 + \frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S} \|\mathbf{e}_i\|_2^2 - \|\bar{\mathbf{e}}_{\mathbf{x}}\|_2^2.$$

Combining the last two displays gives

$$\mathcal{L}_{A}(\mathbf{x}_{h}, \mathbf{x}; \phi, E) = \left\| \boldsymbol{\mu}_{\phi}(\mathbf{x}_{h}) - \bar{\mathbf{e}}_{\mathbf{x}} \right\|_{2}^{2} + \operatorname{tr} \Sigma_{\phi}(\mathbf{x}_{h}) + \underbrace{\left(\frac{1}{|S_{\mathbf{x}}|} \sum_{i \in S_{\mathbf{x}}} \|\mathbf{e}_{i}\|_{2}^{2} - \|\bar{\mathbf{e}}_{\mathbf{x}}\|_{2}^{2}\right)}_{\text{const}(\mathbf{x}, E)},$$

which is exactly Eq. (26).

E.5 PROOF OF PROPOSITION 3.2

Proposition E.6. Fix \mathbf{x} and a neighborhood in which the mask-conditioned denoising ELBO objective, viewed as a function of the encoder mean $\boldsymbol{\mu} \in \mathbb{R}^d$, admits a quadratic approximation with mask-independent curvature

$$H \succeq mI$$
, $||H||_2 \leq L$ $(0 < m \leq L < \infty)$.

Augmenting the objective by $\lambda_A \| \boldsymbol{\mu}_{\phi}(\mathbf{x}_h) - \bar{\mathbf{e}}_{\mathbf{x}} \|_2^2$ yields an effective Hessian $H_{eff} = H + 2\lambda_A I$ for the $\boldsymbol{\mu}$ -subproblem. Let $\boldsymbol{\mu}^{(0)}(\mathbf{x}_h)$ be the (local) minimizer without alignment and $\boldsymbol{\mu}^{(A)}(\mathbf{x}_h)$ the minimizer with alignment. Define the masking variance $\operatorname{Var}_{\mathbf{b}}[v] := \mathbb{E}_{\mathbf{b}}[(v - \mathbb{E}_{\mathbf{b}}v)(v - \mathbb{E}_{\mathbf{b}}v)^{\top}]$ and set

$$\tau := \frac{L}{L + 2\lambda_{\mathcal{A}}} \in (0, 1).$$

Then

$$\|\operatorname{Var}_{\mathbf{b}}[\boldsymbol{\mu}^{(A)}(\mathbf{x}_h)]\|_{2} \leq \tau^{2} \|\operatorname{Var}_{\mathbf{b}}[\boldsymbol{\mu}^{(0)}(\mathbf{x}_h)]\|_{2}, \tag{27}$$

$$\operatorname{tr} \operatorname{Var}_{\mathbf{b}} \left[\boldsymbol{\mu}^{(A)}(\mathbf{x}_h) \right] \leq \tau^2 \operatorname{tr} \operatorname{Var}_{\mathbf{b}} \left[\boldsymbol{\mu}^{(0)}(\mathbf{x}_h) \right], \tag{28}$$

$$\mathbb{E}_{\mathbf{b}}\left[\|\boldsymbol{\mu}^{(\mathbf{A})}(\mathbf{x}_h) - \bar{\mathbf{e}}_{\mathbf{x}}\|_{2}^{2}\right] \leq \tau \,\mathbb{E}_{\mathbf{b}}\left[\|\boldsymbol{\mu}^{(0)}(\mathbf{x}_h) - \bar{\mathbf{e}}_{\mathbf{x}}\|_{2}^{2}\right]. \tag{29}$$

Moreover, if $Var_{\mathbf{b}}[\boldsymbol{\mu}^{(0)}]$ commutes with H (e.g., they are simultaneously diagonalizable), then the Löwner-order contraction

$$\operatorname{Var}_{\mathbf{b}}\left[\boldsymbol{\mu}^{(\mathrm{A})}(\mathbf{x}_h)\right] \leq \tau^2 \operatorname{Var}_{\mathbf{b}}\left[\boldsymbol{\mu}^{(0)}(\mathbf{x}_h)\right]$$

holds.

Proof. Fix \mathbf{x} and a mask \mathbf{b} , and let $F_{\mathbf{b}}(\boldsymbol{\mu})$ denote the (unregularized) mask-conditioned denoising objective as a function of the encoder mean $\boldsymbol{\mu} \in \mathbb{R}^d$ (all other quantities \mathbf{x} , the masked input \mathbf{x}_h , decoder parameters, are held fixed).

Choose a reference point $\tilde{\mu}$ in a neighborhood where $F_{\mathbf{b}}$ admits a quadratic approximation with mask-independent curvature matrix H, and assume

$$H \succ mI$$
, $||H||_2 < L$, $0 < m < L < \infty$.

Equivalently, we approximate the Hessian uniformly across masks by $\nabla^2 F_{\mathbf{b}}(\tilde{\mu}) \approx H$.

Quadratic surrogate. Define the quadratic model of $F_{\mathbf{b}}$ around $\tilde{\mu}$ by

$$J_{\mathbf{b}}(\boldsymbol{\mu}) := F_{\mathbf{b}}(\tilde{\boldsymbol{\mu}}) + \nabla F_{\mathbf{b}}(\tilde{\boldsymbol{\mu}})^{\top} (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) + \frac{1}{2} (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^{\top} H (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}).$$
(30)

Let

$$\mathbf{g_b} := \nabla F_{\mathbf{b}}(\tilde{\boldsymbol{\mu}}), \quad \mathbf{a_b} := \tilde{\boldsymbol{\mu}} - H^{-1}\mathbf{g_b}.$$

Completing the square yields an equivalent form

$$J_{\mathbf{b}}(\boldsymbol{\mu}) = c_{\mathbf{b}} + \frac{1}{2} (\boldsymbol{\mu} - \mathbf{a}_{\mathbf{b}})^{\top} H (\boldsymbol{\mu} - \mathbf{a}_{\mathbf{b}}), \tag{31}$$

where the (mask-dependent) constant

$$c_{\mathbf{b}} := F_{\mathbf{b}}(\tilde{\boldsymbol{\mu}}) - \frac{1}{2} \mathbf{g}_{\mathbf{b}}^{\top} H^{-1} \mathbf{g}_{\mathbf{b}}$$

is independent of μ .

In particular, the unique minimizer of J_b is a_b :

$$\arg\min_{\boldsymbol{\mu}} J_{\mathbf{b}}(\boldsymbol{\mu}) = \mathbf{a}_{\mathbf{b}}.$$

If $F_{\mathbf{b}}$ is exactly quadratic with curvature H in this neighborhood, then $\mu^{(0)}(\mathbf{x}_h) = \mathbf{a}_{\mathbf{b}}$; otherwise, $\mathbf{a}_{\mathbf{b}}$ is the minimizer of the local quadratic approximation to $F_{\mathbf{b}}$.

Adding the alignment penalty $\lambda_A \|\mu - \bar{\mathbf{e}}_{\mathbf{x}}\|_2^2$ gives the first-order condition

$$(H + 2\lambda_{\mathbf{A}}I)\,\boldsymbol{\mu} = H\,\mathbf{a_b} + 2\lambda_{\mathbf{A}}\bar{\mathbf{e_x}}.$$

Define

$$M := (H + 2\lambda_{\mathbf{A}}I)^{-1}H, \qquad \tau := \frac{L}{L + 2\lambda_{\mathbf{A}}} \in (0, 1).$$

Then the aligned minimizer is the affine *shrinkage* of a_b toward \bar{e}_x :

$$\boldsymbol{\mu}^{(A)} = M \, \mathbf{a_b} + (I - M) \, \bar{\mathbf{e}_x}. \tag{32}$$

Spectral bounds on M. Diagonalize $H = Q\Lambda Q^{\top}$ with $\Lambda = \operatorname{diag}(\lambda_i), m \leq \lambda_i \leq L$. Then

1136
$$M = Q \operatorname{diag}\!\left(\frac{\lambda_i}{\lambda_i + 2\lambda_{\mathrm{A}}}\right) Q^\top.$$

1138 Let $\alpha_i := \lambda_i/(\lambda_i + 2\lambda_{\rm A}) \in (0,1)$. It follows that

$$0 \le M \le I, \qquad \|M\|_2 = \max_i \alpha_i = \frac{\lambda_{\max}(H)}{\lambda_{\max}(H) + 2\lambda_A} \le \tau, \qquad \|M^2\|_2 = \|M\|_2^2 \le \tau^2, \quad (33)$$

and, eigenwise, $\alpha_i^2 \le \tau \alpha_i$, hence

$$M^2 \, \le \, \tau \, M \, \le \, \tau \, I. \tag{34}$$

Variance contraction (operator norm). Since \bar{e}_x is mask-independent, Eq. (32) gives

$$\operatorname{Var}_{\mathbf{b}}[\boldsymbol{\mu}^{(A)}] = M \operatorname{Var}_{\mathbf{b}}[\mathbf{a}_{\mathbf{b}}] M. \tag{35}$$

Taking spectral norms and using submultiplicativity,

$$\|\operatorname{Var}_{\mathbf{b}}[\boldsymbol{\mu}^{(A)}]\|_{2} \leq \|M\|_{2}^{2} \|\operatorname{Var}_{\mathbf{b}}[\mathbf{a}_{\mathbf{b}}]\|_{2} \leq \tau^{2} \|\operatorname{Var}_{\mathbf{b}}[\mathbf{a}_{\mathbf{b}}]\|_{2},$$

which is Eq. (27) upon noting $Var_{\mathbf{b}}[\mathbf{a}_{\mathbf{b}}] = Var_{\mathbf{b}}[\boldsymbol{\mu}^{(0)}].$

Variance contraction (trace). From Eq. (35),

$$\operatorname{tr} \operatorname{Var}_{\mathbf{b}} \left[\boldsymbol{\mu}^{(\mathbf{A})} \right] = \operatorname{tr} \left(\operatorname{Var}_{\mathbf{b}} [\mathbf{a}_{\mathbf{b}}] \, M^2 \right) \, \leq \, \| M^2 \|_2 \, \operatorname{tr} \operatorname{Var}_{\mathbf{b}} [\mathbf{a}_{\mathbf{b}}] \, \leq \, \tau^2 \, \operatorname{tr} \, \operatorname{Var}_{\mathbf{b}} [\mathbf{a}_{\mathbf{b}}],$$

where the inequality uses $M^2 \leq \|M^2\|_2 I$ and the fact that for $A, B \succeq 0$, $\operatorname{tr}(AB) \leq \|B\|_2 \operatorname{tr}(A)$. This yields Eq. (28).

Löwner-order contraction under commutation (Higham & Lin, 2013). If $\operatorname{Var}_{\mathbf{b}}[\mathbf{a}_{\mathbf{b}}]$ commutes with H, then it commutes with M. In the common eigenbasis, write $\operatorname{Var}_{\mathbf{b}}[\mathbf{a}_{\mathbf{b}}] = Q \operatorname{diag}(v_i) Q^{\top}$ with $v_i \geq 0$. Then

$$M \operatorname{Var}_{\mathbf{b}}[\mathbf{a}_{\mathbf{b}}] M = Q \operatorname{diag}(\alpha_i^2 v_i) Q^{\top} \leq Q \operatorname{diag}(\tau^2 v_i) Q^{\top} = \tau^2 \operatorname{Var}_{\mathbf{b}}[\mathbf{a}_{\mathbf{b}}],$$

proving the Löwner-order bound.

Mean–drift contraction. From Eq. (32),

$$\mu^{(A)} - \bar{\mathbf{e}}_{\mathbf{x}} = M(\mathbf{a}_{\mathbf{b}} - \bar{\mathbf{e}}_{\mathbf{x}}),$$

hence

$$\|\boldsymbol{\mu}^{(\mathrm{A})} - \bar{\mathbf{e}}_{\mathbf{x}}\|_{2}^{2} = (\mathbf{a}_{\mathbf{b}} - \bar{\mathbf{e}}_{\mathbf{x}})^{\top} M^{2} (\mathbf{a}_{\mathbf{b}} - \bar{\mathbf{e}}_{\mathbf{x}})$$

$$\leq (\mathbf{a}_{\mathbf{b}} - \bar{\mathbf{e}}_{\mathbf{x}})^{\top} (\tau I) (\mathbf{a}_{\mathbf{b}} - \bar{\mathbf{e}}_{\mathbf{x}}) \qquad \text{(by Eq. (34))}$$

$$= \tau \|\mathbf{a}_{\mathbf{b}} - \bar{\mathbf{e}}_{\mathbf{x}}\|_{2}^{2}.$$

Taking $\mathbb{E}_{\mathbf{b}}$ gives

$$\mathbb{E}_{\mathbf{b}} \big[\| \boldsymbol{\mu}^{(A)} - \bar{\mathbf{e}}_{\mathbf{x}} \|_2^2 \big] \ \leq \ \tau \, \mathbb{E}_{\mathbf{b}} \big[\| \boldsymbol{\mu}^{(0)} - \bar{\mathbf{e}}_{\mathbf{x}} \|_2^2 \big],$$

which is Eq. (29).