On the Mechanisms of Collaborative Learning in VAE Recommenders

Tung–Long Vuong[1], Julien Monteil[1], Hien Dang[2], Volodymyr Vaskovych[1], Trung Le[3], Vu Nguyen[1]
[1]Amazon, [2]University of Texas at Austin, [3]Monash University

Problem setup

VAE-CF trains from a masked interaction history

Objective

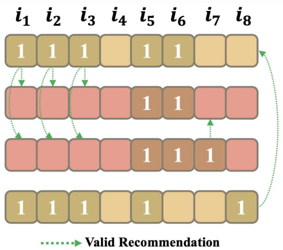
$$\mathcal{L}_{VAE}(x;\theta,\phi) = -\mathbb{E}_{q_\phi(z|x_h)}[\log p_\theta(x \mid z)] + \beta KL(q_\phi(z \mid x_h)||p(z))$$

$\rho = 1$ gives clean input. Smaller ρ means stronger masking and less observed information.

Collaborative signals

Local signal: nearby users should make similar predictions.

Global signal: far-but-related users can still help when they share positive items.



Masking changes which user pairs become neighbors.

Result-1: Collaboration is governed by latent proximity.

An SGD step on user v helps user u only if their posteriors are close enough in latent space.

$$D_{u,v} = L_{\theta z} W_1(q_u, q_v) + \Delta_x(u,v)$$

$W1(q_u, q_v)$: latent mismatch

$\Delta_x(u,v)$: content mismatch

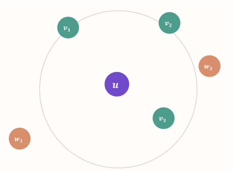
$$W_1(q_u,q_v) < r_{\text{share}}(u,v;\theta) = \frac{[\|g_u(\theta)\| - \Delta_x(u,v)]_+}{L_{\theta z}}$$

→ helpful gradient transfer is restricted to a latent neighborhood. Influence decays as users move farther apart in W_1.

- Raw input similarity is not enough by itself.
- What matters is whether users fall inside each other's sharing radius after encoding.
- Global collaboration requires some far-but-related pairs to become latent-close.

Latent neighborhood view

Inside the radius: useful transfer.
Outside the radius: it fades quickly.



sharing radius around u

green = inside | orange = outside | purple = u

Result-2: Clean inputs favor local over global collaboration.

$$W_1(q_u,q_v) \le L_\phi \|x_u - x_v\|_1$$

Lipschitz encoder: L_1-near users stay W_1-near after encoding.

VAE-CF without masking objective satisfies:

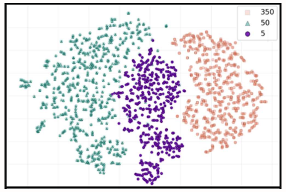
$$C + \int (q_u(z) + q_v(z))\Delta_{A^*} dz + \beta \sum_{i\in\{u,v\}} \mathrm{KL}(q_\phi(z \mid x_i)||p(z))$$

where Δ_{A^*} is the **compromise gap** from making one latent region explain two different users.

Reconstruction creates a compromise gap for mismatched users, so input-distant users are pushed apart instead of sharing posteriors.

$$W_1(q_u,q_v) \le \left(\sqrt{2C\ \mathrm{LK}(q_u||p)} + \sqrt{2C\ \mathrm{LK}(q_v||p)}\right)$$

Increasing β tightens the $W_1(q_u,q_v)$ bound, and enables global collaboration, but excessive values cause posterior collapse; thus, β is kept small in ranking-focused CF.



clean input only | nDCG@100 | 0.409

Clusters are cleanly separated by interaction level:

- **good for local structure**
- **poor for global alignment.**

Result-3: Masking can create useful long-range sharing, but it also destabilizes local neighborhoods.

Let $x_u x_v \in \{0,1\}^I$, $b_u, b_v \sim Bern(\rho)^I$ be independent masks and set $x_u' = x_u \odot b_u$ and $x_u' = x_u \odot b_u$

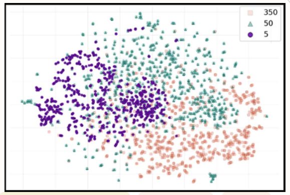
Contraction. $\Pr[\| x_u' - x_v'\|_1 \le \delta] \ge 0$

Expansion. $\Pr[\| x_u' - x_v'\|_1 \ge \delta] \ge 0$

Random masking can contract distances for some far-apart user pairs.

Those contractions can pull far-but-related users inside the sharing radius, enabling intermittent global transfer.

Masking can also expand genuine neighbors apart, causing neighborhood drift and noisier gradients.



clean input only | nDCG@100 | 0.426

The cohorts become entangled:

- **enables more global sharing,**
- **makes local neighborhoods more diffuse.**

Beyond latent geometry — β-KL vs. masking

$$\mathbb{E}_{u,v,b} W_1\left(q_\phi(\cdot| x_u \odot b_u), q_\phi(\cdot| x_v \odot b_v)\right) \le 2\sqrt{2C}\sqrt{\mathbb{E}_{x,b} KL(q_\phi(\cdot| x \odot b)||p)}$$

with

$$\mathbb{E}_{x,b} KL(q_\phi(\cdot| x \odot b)||p) = I_{q_\phi}(X_h; Z) + KL(q_h||p)$$

This shows that reducing the KL budget shrinks **expected pairwise latent distance**, which increases the chance that users fall inside each other's sharing radius.

1) Address the prior / KL pathway

Increase beta or redesign $p(z)$:

- **Effect:** near-uniform contraction of user posteriors.
- **Gain:** more overlap and more global sharing.
- **Risk:** too much beta weakens reconstruction and can cause collapse.

2) Address the masking pathway

Reduce rho so masking is stronger:

- **Effect:** stochastic contractions and expansions of pairwise distances.
- **Gain:** far-but-related users can become latent neighbors.
- **Risk:** true neighbors can drift apart, creating noisy and unstable sharing.

Method: Personalized Item Alignment (PIA)

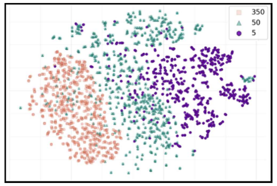
$$\mathcal{L}_{PIA-VAE} = \mathbb{E}_b[\mathcal{L}_{VAE}] - \lambda_A \mathcal{L}_A$$

$$\mathcal{L}_A(x_h, x; \phi, E) = \frac{1}{|S_x|}\sum_{i\in S_x} \mathbb{E}_{z\sim q_\phi(z|x_h)} \|z - e_i\|_2^2$$

Proposition 3.1: PIA aligns the posterior mean to the item centroid and also mildly shrinks posterior spread.

- Stabilizes masking-induced geometry: different masks of the same user map to a more consistent region.
- Reduces neighborhood drift and gradient variance.
- Promotes semantically grounded global mixing because users with overlapping items have closer centroids.
- No test-time cost: anchors and regularizer are used only during training.

Experiments



clean input only | nDCG@100 | 0.437

Keep global mixing from masking while restoring local reliability

One-week Amazon A/B test

	Model	Recall @20	Recall @50	nDCG @100
Offline	Multi-VAE	0.592	0.288	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
Online	**Home Card**			
	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
	Uplift (%)	509	267	283

Offline benchmarks

Model	MovieLens-20M Recall @20	MovieLens-20M Recall @50	MovieLens-20M NDCG @100	Netflix Prize Recall @20	Netflix Prize Recall @50	Netflix Prize NDCG @100	Million Song Recall @20	Million Song Recall @50	Million Song NDCG @100
Matrix factorization & Linear regression									
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

Results across user groups for MovieLens20M.

Group	Model	Recall @20	Recall @50	nDCG @100
[5–10]	Multi-VAE	0.461	0.625	0.317
	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

Code: https://github.com/amazon-science/PIAVAE