Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.086183

ARTICLE

α-FedVAE: Detached Posterior-Confidence Gating for Dimension-Wise KL Regularization in Personalized Federated Collaborative Filtering

Jincheng Cai[1], **Li Feng**[1,*] **and Ni Zhao**[2]

[1]The School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China
[2]The School of Finance and Economics, Shenzhen University of Information Technology, Shenzhen, China

*Corresponding Author: Li Feng. Email: lfeng@must.edu.mo

Received: 26 May 2026; Accepted: 10 August 2026; Published: 15 September 2026

ABSTRACT: Federated Variational Autoencoders (VAEs) keep interaction data local, but existing federated VAE recommenders typically apply uniform KL regularization and do not adapt dimension-wise penalties to unreliable posteriors in sparse interaction scenarios. We propose α-FedVAE, which uses a detached, clipped normalized signal-to-noise ratio as a local confidence gate for each KL dimension of a fused user posterior, without extra communication. Across MovieLens-100K, MovieLens-1M, and Amazon Video, α-FedVAE improves mean HR@20 by 7.8%–55.3% and NDCG@20 by 8.0%–65.8% over FedDAE. These results indicate that α-FedVAE improves personalized recommendation under sparse and decentralized settings while preserving the communication footprint.

KEYWORDS: Federated learning; self-adaptive training; variational autoencoder; collaborative filtering; regularization dynamics

1 Introduction

Collaborative Filtering (CF) models user preferences from interaction data [1–3], but centralizing such data raises privacy concerns under regulations such as the General Data Protection Regulation [4]. Federated Learning (FL) instead aggregates model updates while retaining raw data on clients [5,6], enabling Personalized Federated Collaborative Filtering (PFCF) [7,8]. Variational Autoencoders (VAEs) [9] are effective generative recommenders: an encoder maps interactions to a stochastic latent representation and a decoder reconstructs item preferences. Training maximizes an Evidence Lower Bound (ELBO) containing reconstruction and Kullback–Leibler (KL) regularization terms.

Dual-encoder methods such as FedDAE [10] separate shared and private representations, but retain one KL weight β for every latent dimension. They therefore regularize a reliable preference factor and an uncertain dimension equally. We study how this dimension-blind penalty affects sparse federated recommendation.

1.1 The Uniform Regularization Problem

For a sparse user, limited evidence leaves posterior dimensions with unequal reliability. Uniform KL regularization applies the same coefficient whether a posterior displacement is well determined or accompanied by high variance. Global β-annealing [11] and activity-based scaling [12] vary the penalty over time or users but remain unable to distinguish dimensions within one representation.

We use normalized SNR to calibrate the prior penalty: a well-determined displacement retains nearly the global KL strength, whereas a variance-dominated KL contribution is discounted. This rule distinguishes dimensions using encoder statistics without client-specific tuning; the gate study evaluates its empirical boundary rather than assuming unique optimality.

1.2 Our Contributions

Normalized SNR, stop-gradient, and KL decomposition are established components rather than individual contributions. Our contribution is their integration as a locally recomputed gate on each fused-user-posterior KL term, without a learned gate parameter, client-specific tuning, or additional payload. The raw confidence approaches one for a well-determined dimension and zero when uncertainty dominates; fixed global clipping to $[0.05, 0.95]$ prevents complete KL suppression or exact saturation.

Our contributions are summarized as follows:

- We formulate and empirically examine dimension-blind KL regularization as a bottleneck for sparse federated VAE recommendation.
- We introduce the integrated federated VAE mechanism: the clipped normalized-SNR coefficient is locally recomputed and detached on each KL dimension of a fused user posterior. Its placement and update rule across dimensions, users, and local steps constitute the method contribution, not the SNR formula.
- We establish a limited ARD-like posterior-mean gradient analogy, a bound on aggregate gate strength, and a conditional descent decomposition with explicit budgets for non-stationary gates; these results are neither full ARD inference nor an unconditional convergence theorem for the neural network.
- Three-dataset 100-round experiments show mean gains of 7.8%–55.3% in HR@20 and 8.0%–65.8% in NDCG@20 over FedDAE, with controlled analyses delimiting where those gains persist.

2 Related Work

2.1 Federated Collaborative Filtering

Federated CF spans matrix factorization [7,8], neural filtering (FedNCF [13]), personalization layers (FedPer [14]), dual or additive personalization (PFedRec [15]; FedRAP [16]), embedding-degradation mitigation (PLGC [17]), and common-representation learning (FedCR [18]). SCAFFOLD controls client drift [19]; FedVAE [20], EFVAE [21], FedDAE [10], and FedRKG [22] provide generative or knowledge-graph-based alternatives. These methods retain uniform within-user KL weighting; our focus is dimension-wise weighting of one fused client posterior, with FedDAE as the closest architectural baseline. LightGCN [23] is contextual because its architecture and centralized access differ.

2.2 KL Regularization in Variational Autoencoder

The β-VAE and KL annealing vary a global regularization weight but remain dimension-blind [11,24]. Variational Dropout uses the SNR of a variational *weight* posterior for multiplicative noise [25]. We compute the same normalized statistic $s_d = \mu_{u,d}^2/(\mu_{u,d}^2 + \sigma_{u,d}^2) = \mathrm{SNR}/(1+\mathrm{SNR})$ from a fused *user-latent* posterior and use $\alpha_d = \mathrm{clip}(s_d, 0.05, 0.95)$ to weight each VAE KL term. The statistic is shared, but its clipped optimization role, posterior object, and federated deployment differ.

Free-bits thresholds KL_d, Ladder VAE controls layers [26], Mult-VAE uses uniform KL, and RecVAE learns a centralized composite prior [12,27]. Annealing varies across rounds and activity scaling across users; our non-learned gate instead weights each local user–dimension posterior while retaining the standard Gaussian prior and adding no transmitted variable.

2.3 Automatic Relevance Determination

Automatic Relevance Determination (ARD) assigns independent prior precisions to dimensions [28,29]. With α_d detached, the gated KL has the same posterior-mean gradient scaling as a Gaussian prior with dimension-specific precision α_d. This is only an ARD-like local shrinkage analogy: α_d is not learned by evidence maximization, does not perform classical relevance selection, and does not define a full Bayesian ARD posterior. Section 5 states this scope.

3 Preliminaries

Let $\mathcal{U} = \{1, \ldots, N\}$ and $\mathcal{I} = \{1, \ldots, M\}$ denote users and items. Interactions are represented by binary matrix $\mathbf{R} \in \{0,1\}^{N \times M}$, where $r_{ui} = 1$ indicates user u interacted with item i. For each user u, the positive set is $P_u = \{i \in \mathcal{I} \mid r_{ui} = 1\}$ and interaction vector $\mathbf{r}_u \in \{0,1\}^M$ is stored locally.

Federated VAE. A K-dimensional latent variable $\mathbf{z}_u \in \mathbb{R}^K$ captures preferences. The generative process samples $\mathbf{z}_u \sim p(\mathbf{z}_u) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, transforms via decoder $f_\theta : \mathbb{R}^K \to \mathbb{R}^M$, and draws $\mathbf{r}_u \sim \text{Multinomial}(|P_u|, \pi(\mathbf{z}_u))$ where $\pi(\mathbf{z}_u) = \text{softmax}(f_\theta(\mathbf{z}_u))$. Training maximizes the β-weighted ELBO:

$$\mathcal{L}_u = \mathbb{E}_{q_\phi(\mathbf{z}_u|\mathbf{r}_u)} \left[\log p_\theta(\mathbf{r}_u|\mathbf{z}_u)\right] - \beta \cdot \text{KL}\left(q_\phi(\mathbf{z}_u|\mathbf{r}_u) \| p(\mathbf{z}_u)\right). \tag{1}$$

The encoder $q_\phi(\mathbf{z}_u|\mathbf{r}_u) = \mathcal{N}(\boldsymbol{\mu}_u, \text{diag}\{\boldsymbol{\sigma}_u^2\})$ outputs per-dimension mean $\mu_{u,d}$ and variance $\sigma_{u,d}^2$. Sampling uses reparameterization $\mathbf{z}_u = \boldsymbol{\mu}_u + \boldsymbol{\sigma}_u \odot \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Dual-Encoder Architecture. Following FedDAE [10], the encoder decomposes into global q_{ϕ_g} (shared) and local q_{ϕ_l} (private). A gating network h_{ψ_u} combines outputs:

$$\boldsymbol{\mu}_u = \omega_{u1} \cdot \boldsymbol{\mu}_g + \omega_{u2} \cdot \boldsymbol{\mu}_l, \quad \boldsymbol{\sigma}_u^2 = \omega_{u1}^2 \cdot \boldsymbol{\sigma}_g^2 + \omega_{u2}^2 \cdot \boldsymbol{\sigma}_l^2, \tag{2}$$

where $(\omega_{u1}, \omega_{u2}) = h_{\psi_u}(\mathbf{r}_u)$ with $\omega_{u1} + \omega_{u2} = 1$. Only ϕ_g and θ are communicated; ϕ_l and ψ_u remain on-device. The fused posterior $q_\phi(z_{u,d}|\mathbf{r}_u) = \mathcal{N}(\mu_{u,d}, \sigma_{u,d}^2)$ is obtained via Eq. (2). Bold symbols denote vectors (e.g., $\boldsymbol{\mu}_u \in \mathbb{R}^K$), non-bold denote scalars (e.g., $\mu_{u,d}$).

4 The α-FedVAE Framework

Fig. 1 shows the on-device confidence gate and shared-parameter aggregation. Each user is one client; raw interactions, local encoder parameters, and posterior-fusion parameters remain private, while only shared encoder and decoder updates are aggregated. The confidence gate itself has no learned parameter. The implementation provides data decentralization but not secure aggregation, homomorphic encryption, or differential privacy.

4.1 Diagnosing Uniform Regularization Waste

For a diagonal Gaussian posterior, $\text{KL}_d = \frac{1}{2}(\mu_{u,d}^2 + \sigma_{u,d}^2 - 1 - \log \sigma_{u,d}^2)$ and $\partial \text{KL}_d / \partial \mu_{u,d} = \mu_{u,d}$. Applying the same β to every dimension ignores whether a posterior displacement reflects a reliable preference or uncertainty, a mismatch that is amplified when user histories are sparse.

The issue is not that every uncertain dimension has a large gradient in isolation. Rather, a uniform objective has no mechanism to distinguish a small but reliable displacement from one accompanied by high posterior variance, and the same rule is repeated across all K dimensions. A confidence-aware coefficient supplies this missing dimension-level distinction while preserving the decomposable VAE objective.

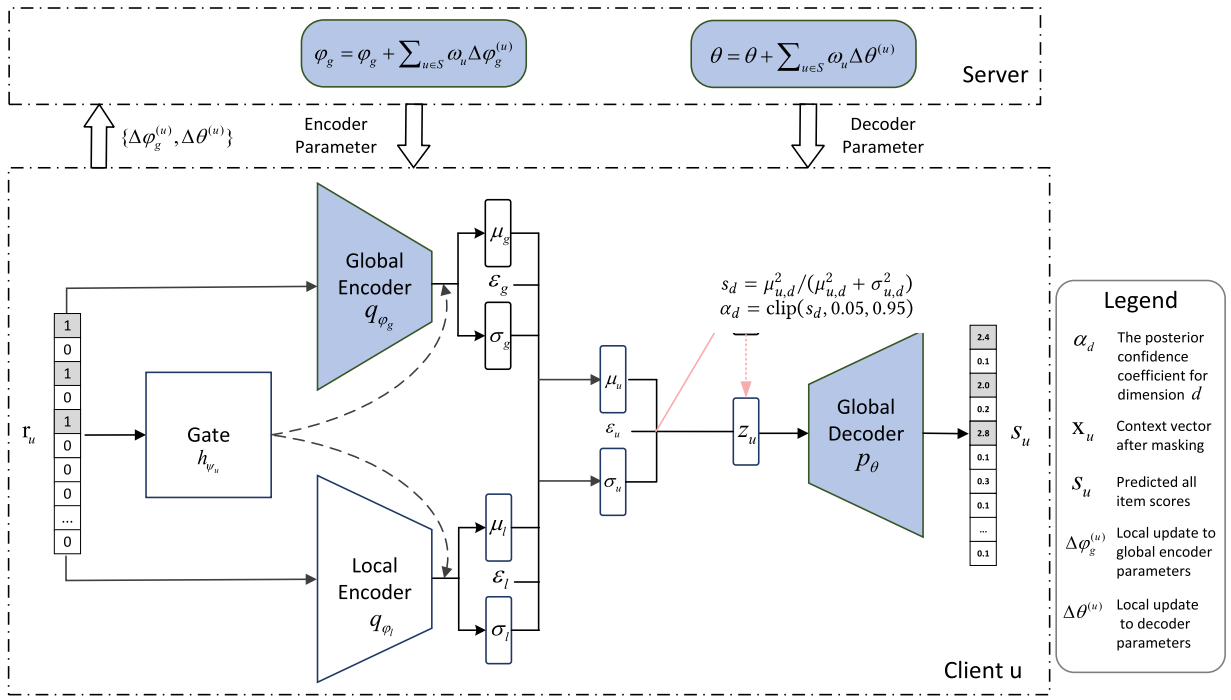


Figure 1: Architecture of α-FedVAE. Clients compute clipped confidence coefficients α_d from fused posterior statistics to modulate per-dimension KL penalties, then upload shared model updates only.

4.2 Posterior Confidence as a Dimension Gate

From the fused posterior in Eq. (2), define the raw confidence $s_d = \mu_{u,d}^2/(\mu_{u,d}^2 + \sigma_{u,d}^2)$, a bounded normalization of $\mu_{u,d}^2/\sigma_{u,d}^2$, and the implemented gate

$$\alpha_d = \text{clip}(s_d, a, b), \qquad a = 0.05, \quad b = 0.95.$$

The squared mean measures displacement from the prior and the variance measures uncertainty: $s_d \to 1$ for a well-determined displacement, $s_d \to 0$ when uncertainty dominates, and intermediate values yield proportionate regularization. Consequently, the implemented direction retains nearly the global KL strength for a high-SNR dimension and discounts a variance-dominated KL term. It does not relax a reliable factor to preserve its mean; rather, it uses posterior reliability to calibrate how strongly the prior penalty is applied. The common floor prevents any KL dimension from being removed completely, while the ceiling avoids exact saturation. Both bounds are fixed design constants used globally for all users, dimensions, datasets, and reported runs; they are not learned or client-specific, but they remain design choices. The gate is computed locally, detached during backpropagation, and never transmitted.

The coefficient has four useful properties:

(i) *Boundedness*: $\alpha_d \in [a, b]$ neither removes a KL term nor amplifies it beyond β.
(ii) *Continuity*: clipping preserves a continuous confidence transition.
(iii) *Non-learned adaptation*: the fixed global bounds introduce no learned, client-specific, or dimension-specific gate parameter.
(iv) *Interpretability*: s_d is the mean's fraction of the posterior second moment and α_d is its safe-guarded form.

The gated term is

$$\mathcal{L}_u^{\mathrm{KL}} = \beta \sum_{d=1}^{K} \alpha_d \cdot \frac{1}{2}\left(\mu_{u,d}^2 + \sigma_{u,d}^2 - 1 - \log \sigma_{u,d}^2\right). \tag{3}$$

Thus $\mathcal{J}_u = -\mathbb{E}_q[\log p_\theta(\mathbf{r}_u|\mathbf{z}_u)] + \mathcal{L}_u^{\mathrm{KL}}$. We summarize aggregate gate strength by the soft-count notation $K_{\mathrm{eff}} = \sum_d \alpha_d$; Section 5 bounds this quantity and Section 6 measures it.

K_{eff} is a soft summary with range $[aK, bK]$: variance-dominated posteriors approach the safeguarded floor and better-determined posteriors can approach the ceiling. It is not a count of nonzero or relevant latent variables. The theory also does not establish a monotone relationship between interaction count and K_{eff}; interaction count is used only to stratify descriptive recommendation results.

Detachment is central to this interpretation. If gradients were propagated through the gate, the KL derivative would contain both $\alpha_d \nabla \mathrm{KL}_d$ and $\mathrm{KL}_d \nabla \alpha_d$, allowing the encoder to change the weight while responding to the weighted term. Stop-gradient removes the second path within an update, so α_d acts as a measured coefficient. It is nevertheless recomputed from the latest fused posterior on the next batch, which preserves adaptation over training rather than freezing one client-level weight.

4.3 Training and Inference

We warm up β linearly [24] and detach α_d within each update; the 100-round benchmark uses 20 warm-up rounds and the reviewer controls use five. Detachment blocks within-step gate manipulation, but recomputation lets the gate evolve between updates, motivating Section 5. Inference uses the unchanged dual encoder and adds no operation or state.

Warm-up controls global KL strength across rounds, whereas confidence redistributes it across dimensions within each local step. At inference, the unchanged dual encoders and decoder rank items without computing s_d or α_d.

4.4 Federated Training Protocol

Algorithm 1 summarizes training. The server maintains $\Theta_g = \{\phi_g, \theta\}$; client u retains $\Theta_u = \{\phi_l^{(u)}, \psi_u\}$.

Relative to FedAvg [5], the only changes are the detached, clipped coefficient at line 8 and per-dimension KL weighting at line 10. Each participating client transfers $2|\Theta_g|$ parameters per round; α_d is local and adds no payload.

The server never observes α_d, posterior statistics, or private encoder parameters. Client sampling and aggregation therefore remain identical to the base federated VAE, which isolates the empirical effect of confidence gating from changes in communication or personalization architecture.

Operationally, a selected client combines the received shared model with its persistent private encoder and fusion parameters for E local epochs. Confidence affects the shared update only through the local objective, and its transient tensor is discarded after backpropagation. The server averages only returned shared parameters; private state remains local. Thus the mechanism changes local optimization but not the aggregation interface or payload size.

Algorithm 1: α-FedVAE

Input: local data $\{\mathbf{r}_u\}$; β, learning rate η, rounds T, local epochs E; fixed bounds $a = 0.05, b = 0.95$

Initialize shared (ϕ_g, θ) and private $\{\phi_l^{(u)}, \psi_u\}$
for $t = 1$ to T **do**
  Sample S_t and broadcast (ϕ_g, θ)
  for each $u \in S_t$ **in parallel do**
    $(\phi_g^{(u)}, \theta^{(u)}) \leftarrow (\phi_g, \theta)$
    for $e = 1$ to E **do**
      Obtain $(\boldsymbol{\mu}_u, \boldsymbol{\sigma}_u^2)$ from both encoders and Eq. (2)
      $\alpha_d \leftarrow \mathrm{stopgrad}\big(\mathrm{clip}\big(\mu_{u,d}^2/(\mu_{u,d}^2 + \sigma_{u,d}^2), a, b\big)\big), \forall d$
      Sample $\mathbf{z}_u = \boldsymbol{\mu}_u + \boldsymbol{\sigma}_u \odot \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
      $\mathcal{J}_u \leftarrow -\log p_\theta(\mathbf{r}_u|\mathbf{z}_u) + \beta \sum_d \alpha_d \mathrm{KL}_d$
      Update $(\phi_g^{(u)}, \theta^{(u)}, \phi_l^{(u)}, \psi_u)$ with Adam(η)
    end for
  Upload $(\phi_g^{(u)}, \theta^{(u)})$
  end for
  $(\phi_g, \theta) \leftarrow |S_t|^{-1} \sum_{u \in S_t} (\phi_g^{(u)}, \theta^{(u)})$ ▷ equal-client FedAvg
end for

5 Theoretical Analysis

We characterize the gate through an ARD-like mean-gradient analogy, an aggregate gate-strength bound, and a conditional descent decomposition under gate variation.

Proposition 1 *(ARD-Like Mean-Gradient Analogy): With α_d frozen by stop-gradient, the gated KL mean gradient equals that of a Gaussian prior $\tilde{p}_d = \mathcal{N}(0, 1/\alpha_d)$ with a dimension-specific precision.*

The proof of Proposition 1 is provided in Appendix A.

For the posterior mean, the gated gradient $\alpha_d \mu_{u,d}$ matches the precision-scaled mean gradient obtained by setting $\gamma_d = \alpha_d$ during one frozen update. Because α_d is recomputed afterward and is not evidence-optimized, this is not equality of complete objectives, variance gradients, relevance rankings, or Bayesian ARD posteriors [28,29].

Theorem 1 *(Bounded Aggregate Gate Strength): Let $s_d = \mu_{u,d}^2/(\mu_{u,d}^2 + \sigma_{u,d}^2)$ and $\alpha_d = \mathrm{clip}(s_d, a, b)$, where $0 \le a < b < 1$. For a diagonal Gaussian posterior with $\sigma_{u,d}^2 \ge \sigma_{\min}^2 > 0$,*

$$aK \le K_{\mathrm{eff}} = \sum_d \alpha_d \le \min\left\{bK,\ aK + \frac{\|\boldsymbol{\mu}_u\|_2^2}{\|\boldsymbol{\mu}_u\|_2^2/K + \sigma_{\min}^2}\right\}.$$

Hence $K_{\mathrm{eff}} \to aK$ as $\|\boldsymbol{\mu}_u\|_2^2 \to 0$; the un-clipped case $a = 0$ recovers a zero limit.

The proof of Theorem 1 is provided in Appendix B.

The variance floor makes the bound finite. Posterior mean norm is the signal quantity appearing in the bound: as the norm vanishes, $\alpha_d \to a$, while a larger norm only relaxes the upper bound and does not prove that K_{eff} increases. No theorem links interaction count monotonically to this quantity.

Proposition 2 *(Conditional Descent under Gate Variation): Let F_t be the stop-gradient surrogate formed by freezing the current gates during round t. Define objective variation $v_t = \sup_\Theta |F_{t+1}(\Theta) - F_t(\Theta)|$, gradient*

variation $\delta_t = \sup_\Theta \|\nabla F_{t+1}(\Theta) - \nabla F_t(\Theta)\|$, *and budgets* $V_T = \sum_{t<t} v_t$ *and* $D_T = \sum_{t<t} \delta_t^2$. *Assume the* F_t *are uniformly lower bounded and L-smooth, stochastic gradients and client heterogeneity are bounded, posterior variances have a positive floor, and parameters remain in a compact set. With* $\eta = \mathcal{O}(T^{-1/2})$, *the average surrogate stationarity measure satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F_t(\Theta^t)\|_2^2 \le \mathcal{O}(T^{-1/2}) + \mathcal{O}(E^2 T^{-1/2}) + \mathcal{O}(V_T T^{-1/2}) + \mathcal{O}(D_T/T).$$

The proof of Proposition 2 is provided in Appendix C.

The first terms are standard optimization and client drift [30]; V_T, D_T expose gate non-stationarity as in time-varying FL [31]. The usual order requires $V_T = \mathcal{O}(1)$ and $D_T = \mathcal{O}(\sqrt{T})$, while convergence requires $V_T = o(\sqrt{T})$ and $D_T = o(T)$. Stop-gradient alone proves neither, so this is a conditional descent decomposition, not convergence to one fixed objective.

6 Experiments

We evaluate recommendation accuracy, sparsity behavior, sensitivity, convergence, and overhead on three datasets.

6.1 Experimental Setup

6.1.1 Datasets

MovieLens-100K (ML-100K: 100,000 interactions, 943 users, 1682 items, 93.70% sparsity) and MovieLens-1M (ML-1M: 1,000,209 interactions, 6040 users, 3706 items, 95.53% sparsity) are movie benchmarks [32]; Amazon Video (Video: 23,181 interactions, 1372 users, 7957 items, 99.79% sparsity) is a sparse e-commerce dataset [33]. The logged files are treated as implicit feedback. Their minimum history lengths are 20 for both MovieLens datasets and 10 for Video. The logged leave-one-out protocol reserves one interaction per user for validation and one for testing.

6.1.2 Baselines

Federated baselines are FedVAE [20], PFedRec [15], FedRAP [16], and the dual-encoder FedDAE [10]. VAE-based methods share latent size, encoder/decoder capacity, optimizer, communication rounds, client sampling, and local epochs; α-FedVAE differs from FedDAE in its KL gate. RecVAE and LightGCN are excluded from that ranking because their architectures and data access differ.

6.1.3 Implementation Details

Each user is a naturally non-IID client. The server uses equal-client FedAvg, $(\phi_g, \theta) = |S_t|^{-1} \sum_{u \in S_t} (\phi_g^{(u)}, \theta^{(u)})$, for the shared encoder and decoder, rather than weighting clients by their numbers of local interactions; local encoder and posterior-fusion parameters remain private. The confidence gate has no learned parameter. The original benchmark in Table 1 uses 100 rounds, seeds $\{0, 1, 2\}$, $K = 256$, Adam at 10^{-3}, batch size 2048, $E = 5$, a 20-round KL warm-up, two-layer tanh MLPs of width 600, full participation on ML-100K and Video, and 10% participation (approximately 604 of 6040 users) on ML-1M.

Table 1: Overall recommendation performance (mean ± sample standard deviation over three runs). HR@20 equals Recall@20 under the one-positive protocol. Best federated results are in **bold**; second best are underlined.

Method	ML-100K (93.70%)		ML-1M (95.53%)		Video (99.79%)	
	HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20
FedVAE [20] (2021)	0.1200 ± 0.004	0.0501 ± 0.003	0.0639 ± 0.001	0.0250 ± 0.001	0.0634 ± 0.013	0.0227 ± 0.004
PFedRec [15] (2023)	0.0541 ± 0.005	0.0115 ± 0.003	0.0600 ± 0.000	0.0231 ± 0.000	0.0138 ± 0.006	0.0026 ± 0.001
FedRAP [16] (2024)	0.0626 ± 0.003	0.0122 ± 0.002	0.0625 ± 0.002	0.0240 ± 0.000	0.0142 ± 0.010	0.0028 ± 0.002
FedDAE [10] (2025)	0.1232 ± 0.002	0.0503 ± 0.001	0.0650 ± 0.004	0.0261 ± 0.001	0.0654 ± 0.018	0.0230 ± 0.003
α-FedVAE (Ours)	**0.1913** ± 0.003	**0.0834** ± 0.002	**0.0701** ± 0.002	**0.0282** ± 0.001	**0.0887** ± 0.005	**0.0358** ± 0.002
Improv. over FedDAE	+55.3%	+65.8%	+7.8%	+8.0%	+35.6%	+55.7%

The reviewer-requested controls in Tables 2–5 follow their logged 30-round configuration: seeds $\{1, 5, 9\}$, $K = 256$, Adam at 10^{-3}, batch size 2048, $E = 5$, three MLP layers, dropout 0.5, a five-round adaptive-KL warm-up, and full client participation on all three datasets. Thus their means are not mixed with the 100-round headline results. The data-loader records four training negatives per positive. Evaluation adds no sampled negatives: training interactions are masked and the held-out item is ranked against the remaining catalog.

Table 2: Thirty-round controlled sensitivity study under varying KL weight β with the normalized-SNR gate. The table consolidates all four recommendation metrics with the logged final-round mean confidence and aggregate gate strength $K_{\text{eff}} = \sum_d \alpha_d$. Metric values are mean ± sample standard deviation over seeds $\{1, 5, 9\}$; best means within each dataset are in **bold**.

Dataset	β	Recall@20	NDCG@20	Precision@20	MAP@20	Mean α	K_{eff}
ML-100K	0.001	0.1675 ± 0.0068	0.0680 ± 0.0025	0.0084 ± 0.0004	0.0407 ± 0.0029	0.7348	188.1
	0.01	**0.1732** ± 0.0018	0.0701 ± 0.0021	**0.0087** ± 0.0001	0.0420 ± 0.0027	0.5047	129.2
	0.1	0.1718 ± 0.0060	**0.0711** ± 0.0009	0.0086 ± 0.0003	**0.0437** ± 0.0009	0.2865	73.4
	1	0.1566 ± 0.0123	0.0649 ± 0.0050	0.0078 ± 0.0006	0.0396 ± 0.0031	0.1412	36.1
ML-1M	0.001	**0.0882** ± 0.0055	**0.0342** ± 0.0020	**0.0044** ± 0.0003	**0.0197** ± 0.0011	0.7613	194.9
	0.01	0.0832 ± 0.0055	0.0329 ± 0.0019	0.0041 ± 0.0003	0.0193 ± 0.0010	0.4926	126.1
	0.1	0.0709 ± 0.0008	0.0279 ± 0.0007	0.0036 ± 0.0001	0.0164 ± 0.0010	0.2149	55.0
	1	0.0713 ± 0.0006	0.0280 ± 0.0002	0.0036 ± 0.0001	0.0164 ± 0.0004	0.1116	28.6
Video	0.001	0.0487 ± 0.0157	0.0211 ± 0.0060	0.0024 ± 0.0008	0.0135 ± 0.0036	0.8182	209.5
	0.01	**0.0684** ± 0.0121	**0.0268** ± 0.0044	**0.0034** ± 0.0006	**0.0154** ± 0.0021	0.6551	167.7
	0.1	0.0646 ± 0.0177	0.0235 ± 0.0054	0.0032 ± 0.0009	0.0125 ± 0.0023	0.2941	75.3
	1	0.0611 ± 0.0223	0.0220 ± 0.0063	0.0030 ± 0.0011	0.0115 ± 0.0023	0.1002	25.6

Table 3: Posterior-source ablation on ML-100K. Recall equals the originally logged HR, and Precision is Recall/20. Only metrics available for every variant are shown. Bold values indicate the best value within each metric column.

Method	Recall@20	NDCG@20	Precision@20
FedDAE (baseline)	0.1232	0.0503	0.0062
α-FedVAE (global)	0.1801	0.0781	0.0090
α-FedVAE (local)	0.1756	0.0763	0.0088
α-FedVAE (fused)	**0.1913**	**0.0834**	**0.0096**

Table 4: Alternative bounded gates on ML-1M using the same dual-encoder federated VAE backbone (30 rounds, $\beta = 1$). All adaptive rows are detached, use the fused posterior, and share clipping bounds $[0.05, 0.95]$; Uniform KL uses $\alpha_d = 1$. Values are mean ± sample standard deviation over seeds $\{1, 5, 9\}$. The normalized-SNR statistic follows Variational Dropout; the α-FedVAE contribution is its application to fused-user-posterior KL terms. Bold values indicate the best mean within each metric column.

Gate	Recall@20	NDCG@20	Precision@20	MAP@20
Uniform KL [11]	0.0703 ± 0.0008	0.0274 ± 0.0002	0.0035 ± 0.0000	0.0158 ± 0.0002
Clipped $\lvert\mu_d\rvert$ [29]	0.0714 ± 0.0004	0.0278 ± 0.0002	0.0036 ± 0.0000	0.0161 ± 0.0004
Clipped inverse uncertainty $1/(1+\sigma_d^2)$ [28]	0.0713 ± 0.0005	0.0279 ± 0.0002	0.0036 ± 0.0001	0.0162 ± 0.0002
Clipped normalized KL share $\mathrm{KL}_d / \sum_j \mathrm{KL}_j$ [9]	**0.0719** ± 0.0007	**0.0280** ± 0.0007	**0.0036** ± 0.0000	0.0162 ± 0.0009
Clipped normalized SNR (α-FedVAE gate) [25]	0.0713 ± 0.0006	**0.0280** ± 0.0002	**0.0036** ± 0.0001	**0.0164** ± 0.0004

Table 5: Reviewer-requested centralized and global-KL references under the same 30-round protocol ($\beta = 1$, mean ± sample standard deviation over seeds $\{1, 5, 9\}$). Centralized Mult-VAE is a non-competing reference because it accesses all interactions centrally. Only metrics available for every method are shown. Bold values indicate the best mean within each metric column.

Dataset	Method	Recall@20	NDCG@20	Precision@20
ML-100K	Centralized Mult-VAE	0.1268 ± 0.0068	0.0548 ± 0.0029	0.0063 ± 0.0003
	FedVAE + KL annealing	0.1247 ± 0.0058	0.0546 ± 0.0012	0.0062 ± 0.0003
	FedDAE	0.1463 ± 0.0068	0.0574 ± 0.0027	0.0073 ± 0.0003
	α-FedVAE	**0.1566** ± 0.0123	**0.0649** ± 0.0050	**0.0078** ± 0.0006
ML-1M	Centralized Mult-VAE	**0.0754** ± 0.0014	**0.0305** ± 0.0004	**0.0038** ± 0.0001
	FedVAE + KL annealing	0.0688 ± 0.0004	0.0270 ± 0.0003	0.0034 ± 0.0000
	FedDAE	0.0685 ± 0.0013	0.0270 ± 0.0008	0.0034 ± 0.0001
	α-FedVAE	0.0713 ± 0.0006	0.0280 ± 0.0002	0.0036 ± 0.0001
Video	Centralized Mult-VAE	0.0719 ± 0.0062	0.0250 ± 0.0013	0.0036 ± 0.0003
	FedVAE + KL annealing	**0.0747** ± 0.0030	**0.0270** ± 0.0011	**0.0037** ± 0.0002
	FedDAE	0.0576 ± 0.0101	0.0209 ± 0.0016	0.0029 ± 0.0005
	α-FedVAE	0.0611 ± 0.0223	0.0220 ± 0.0063	0.0030 ± 0.0011

Tables report mean ± sample standard deviation. For the matched 30-round α-FedVAE/FedDAE comparison, two-sided paired t-tests use the three common seeds, and 95% confidence intervals use the Student-t critical value with two degrees of freedom. With only three pairs and multiple dataset–metric comparisons, these tests are exploratory and low-powered; no multiplicity-adjusted confirmatory claim is made. The exact results are reported with Table 5. No inferential claim is attached to the 100-round means because the consolidated revision logs do not retain their per-seed paired records.

6.1.4 Evaluation Metrics

At $K = 20$, Recall records retrieval, NDCG discounts a hit by rank, Precision measures the relevant fraction, and MAP uses precision at the relevant rank. The main table retains HR and NDCG for comparability; controlled tables add Precision and MAP when commonly available. Section 6.5 stratifies users by activity and test-item popularity.

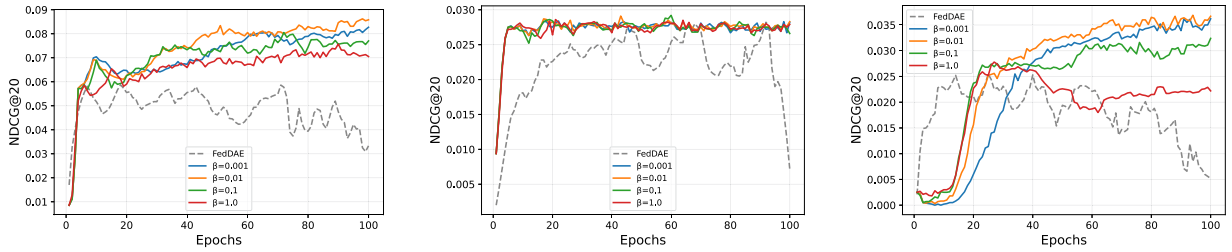
More precisely, let r_u be the one-based rank of user u's held-out item after masking training interactions, and let $\mathbb{I}_u(K) = \mathbb{I}[r_u \le K]$. The reported averages over the evaluated user set $\mathcal{U}_{\text{test}}$ are

$$\begin{aligned} \text{Recall@}K &= \frac{1}{|\mathcal{U}_{\text{test}}|}\sum_u \mathbb{I}_u(K), & \text{Precision@}K &= \frac{1}{K|\mathcal{U}_{\text{test}}|}\sum_u \mathbb{I}_u(K), \\ \text{NDCG@}K &= \frac{1}{|\mathcal{U}_{\text{test}}|}\sum_u \frac{\mathbb{I}_u(K)}{\log_2(r_u+1)}, & \text{MAP@}K &= \frac{1}{|\mathcal{U}_{\text{test}}|}\sum_u \frac{\mathbb{I}_u(K)}{r_u}. \end{aligned} \tag{4}$$

With one relevant test item, these are the leave-one-out forms of the standard metrics: Recall equals HR, Precision equals Recall/K, and NDCG and MAP retain different rank discounts. Comparisons omit MAP when it is not commonly available.

6.2 Overall Performance

Under the 100-round, validation-selected $\beta = 0.01$ protocol, α-FedVAE has the highest mean HR/NDCG on all datasets: gains over FedDAE are 55.3%/65.8% on ML-100K, 7.8%/8.0% on ML-1M, and 35.6%/55.7% on Video. The larger ML-100K and Video mean gains are consistent with the sparsity motivation, while the smaller ML-1M gain indicates less headroom over FedDAE. Fig. 2 shows stable training curves; the separate 30-round study below evaluates β under a matched budget.



(a) ML-100k (b) ML-1M (c) Video

Figure 2: NDCG@20 convergence over 100 rounds.

ML-100K combines moderate sparsity with a relatively small client population, where the fused global/local posterior can provide differentiated confidence across dimensions. Video is substantially sparser and has lower absolute scores, yet confidence gating still improves both hit occurrence and ranking quality.

This indicates that the method can remain useful under severe sparsity, although the later popularity analysis shows that the improvement should not be interpreted as success on tail items.

ML-1M yields smaller mean gains. Its larger user population and denser histories give FedDAE a stronger starting point, leaving less scope for changing the KL allocation alone. Together, the datasets suggest that the benefit depends on both sparsity and the quality of posterior estimates rather than increasing monotonically with catalog sparsity. The convergence curves support optimization stability but are not used to infer an unreported convergence threshold.

6.3 Sensitivity to KL Weight

Table 2 shows that gating remains β-sensitive: ML-100K favors 0.01 for Recall/Precision and 0.1 for NDCG/MAP, ML-1M favors 0.001, and Video favors 0.01 with larger variance. Mean α and K_{eff} decrease with β (ML-100K: 188.1 to 36.1), confirming systematic changes in KL strength. This 30-round diagnostic is separate from the 100-round, validation-selected main setting; β remains a validation-selected global hyperparameter.

6.4 Ablation and Confidence Analysis

Table 4 uses bounded counterparts of the suggested signal families to avoid singular raw precision and place every coefficient in the same allowed interval. This common interval does not equalize the empirical gate distributions: in particular, a normalized KL share can concentrate near the clipping floor when K is large. No Variational-Dropout row is duplicated because its normalized score is identical before clipping. Clipped normalized SNR improves Recall by 1.4% over uniform, ties the best NDCG, and leads MAP; normalized KL share leads Recall. The single-dataset results show that several non-uniform rules behave similarly, but do not establish unique SNR optimality or a scale-controlled ranking among gate families.

Under the shared 30-round budget, splits, and seeds, α-FedVAE leads on ML-100K, centralized Mult-VAE on ML-1M, and global annealing on Video. This controlled comparison isolates regularization more closely than the main experiment but provides context, not a universal ranking; the centralized row is non-competing and MAP is omitted when unavailable.

For the matched α-FedVAE/FedDAE seeds, the paired Recall/NDCG p-values are 0.4510/0.2035 (ML-100K), 0.0202/0.1914 (ML-1M), and 0.7978/0.7697 (Video). The corresponding 95% confidence intervals for the mean differences are $[-0.0373, 0.0578]/[-0.0099, 0.0249]$, $[0.0011, 0.0045]/[-0.0013, 0.0034]$, and $[-0.0476, 0.0546]/[-0.0134, 0.0157]$. ML-1M Recall is the only unadjusted p-value below 0.05. Because there are only three paired seeds and six reported comparisons, we treat all tests as exploratory and make no multiplicity-adjusted significance claim.

Table 3 isolates posterior source: fused confidence performs best (HR@20 0.1913), ahead of global (0.1801) and local (0.1756), supporting fusion within this architecture.

6.5 Per-User Performance Analysis

Table 6 stratifies ML-100K by user activity and test-item popularity; these descriptive results are separate from the significance tests.

The relative NDCG gain decreases from 49.4% for the least-active group to 19.7% for users with at least 100 interactions, consistent with the sparse-user motivation. Head-item Recall/NDCG rises from 0.1406/0.0980 to 0.1722/0.1268, but both methods have zero hits among only six tail cases; most evidence therefore concerns head items, not tail recommendation.

Table 6: ML-100K metrics by user activity and held-out-item popularity (group means across the available evaluation runs). Grouped Precision@20 is derived as Recall@20/20; only metrics available for every group are shown. Bold values indicate the best displayed method within each group and metric.

Grouping	Group	Method	Recall@20	NDCG@20	Precision@20
Activity	[10, 20)	FedDAE	0.2143	0.0774	0.0107
		α-FedVAE	**0.2786**	**0.1156**	**0.0139**
	[20, 50)	FedDAE	0.1655	0.0703	0.0083
		α-FedVAE	**0.2196**	**0.0907**	**0.0110**
	[50, 100)	FedDAE	0.1469	0.0553	0.0073
		α-FedVAE	**0.1781**	**0.0708**	**0.0089**
	≥100	FedDAE	0.1031	0.0398	0.0052
		α-FedVAE	**0.1115**	**0.0477**	**0.0056**
Popularity	Head (n = 530)	FedDAE	0.1406	0.0980	0.0070
		α-FedVAE	**0.1722**	**0.1268**	**0.0086**
	Mid (n = 405)	FedDAE	0.0004	0.0002	0.00002
		α-FedVAE	**0.0015**	**0.0009**	**0.00007**
	Tail (n = 6)	FedDAE	0.0000	0.0000	0.0000
		α-FedVAE	0.0000	0.0000	0.0000

Interaction count is only an observable proxy for data richness: two equally active users may still have posteriors with different uncertainty. Because gate strength was not logged by activity group, these accuracy results do not test or prove a monotone relationship between history length and K_{eff}. Likewise, the popularity groups are descriptive and too imbalanced for a tail-performance claim. A stronger evaluation would use larger tail groups, popularity-aware metrics, and repeated sampling across datasets.

6.6 Computational Overhead Analysis

α_d costs $\mathcal{O}(K)$ element-wise operations, one transient $B \times K$ tensor (about 2 MB for $B = 2048$, $K = 256$), no persistent state, and no payload. Uniform/SNR time per round is 32.92/34.59 s on ML-100K, 259.28/290.49 s on ML-1M, and 81.57/80.39 s on Video (−1.4% to +12.0%). Process-level peak memory is 12.85/12.88, 106.79/106.79, and 69.48/69.59 GB; these research-hardware values are not per-client requirements. Deployment cost remains hardware dependent, and wireless communication schemes are complementary.

7 Limitations

Extreme cold-start users may yield unreliable posterior means and variances, so side information or sequential context may be needed. Highly skewed popularity may concentrate confidence on head factors; the six-case ML-100K tail group provides no positive evidence, and most popularity evidence concerns head items.

The privacy scope is data decentralization, not secure aggregation, differential privacy, or homomorphic encryption; these mechanisms may alter optimization and require separate evaluation. The ML-1M gate study has different Recall and MAP leaders and no cross-dataset ranking. Explicit-feedback, sequential, cross-domain, rapidly changing, and substantially larger client populations remain untested.

The method also depends on the quality of posterior uncertainty. Misspecified variances can produce a bounded but poorly ranked confidence signal, and clipping prevents degeneracy without calibrating that signal. The fixed bounds a and b are global design choices rather than learned or client-specific parameters, and their sensitivity was not separately evaluated. Moreover, β remains a validation-selected global hyperparameter: dimension-wise gating redistributes its effect but does not make performance invariant to KL strength. The alternative-gate study covers only ML-1M and does not equalize the empirical distributions induced by different formulas. These boundaries motivate reporting the sensitivity, gate-function, and activity analyses together rather than treating any one result as universal.

Calibration and recommendation diversity were not measured, so no conclusion is drawn for either criterion.

8 Conclusion

We introduced a detached, clipped posterior-confidence gate that changes dimension-wise KL allocation without learned gate parameters, extra payload, or inference cost. Theory provides a scoped ARD-like mean-gradient analogy, a clipping-consistent aggregate gate-strength bound, and conditional descent with variation budgets. Across three datasets, descriptive 100-round mean HR@20 and NDCG@20 gains over FedDAE are 7.8%–55.3% and 8.0%–65.8%, with the strongest activity-group gain among the least-active evaluated users. Controlled results limit the claim: the three-seed tests are exploratory, annealing can lead, and no gate is uniquely optimal. The mechanism is applicable to data-decentralized recommendation where records cannot be centralized, subject to the reported timing, privacy, and evidence boundaries.

Local, communication-neutral gating suits decentralized media, e-commerce, mobile, and healthcare personalization. Future work should test stronger privacy, larger federations, sequential recommendation, popularity debiasing, and cold-start side information.

Acknowledgement: Not applicable.

Funding Statement: Not applicable.

Author Contributions: Conceptualization, Jincheng Cai and Li Feng; methodology, Jincheng Cai; software, Jincheng Cai; validation, Jincheng Cai and Ni Zhao; formal analysis, Jincheng Cai; investigation, Jincheng Cai; resources, Li Feng and Ni Zhao; data curation, Jincheng Cai; writing—original draft preparation, Jincheng Cai; writing—review and editing, Li Feng and Ni Zhao; supervision, Li Feng; project administration, Li Feng. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available: MovieLens datasets are available from GroupLens, and Amazon review data are available from the public Amazon review dataset collection. The processed data and implementation details are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of the ARD-Like Mean-Gradient Analogy

Proof: For $q_d = \mathcal{N}(\mu_{u,d}, \sigma^2_{u,d})$ and $p_d = \mathcal{N}(0,1)$, the gated mean gradient is $\alpha_d \, \partial \mathrm{KL}_d / \partial \mu_{u,d} = \alpha_d \mu_{u,d}$ when stop-gradient freezes α_d. An ARD prior $\tilde{p}_d = \mathcal{N}(0, 1/\gamma_d)$ gives mean gradient $\gamma_d \mu_{u,d}$; setting $\gamma_d = \alpha_d$ proves equality for that update. Recomputing α_d afterward makes the result local and gradient-level, not full ARD inference.

The equality concerns shrinkage of the posterior mean. The variance component of the Gaussian KL, the evolution of α_d across updates, and classical ARD relevance selection are not identified; these restrictions are why Proposition 1 is stated only as a frozen-step gradient analogy. □

Appendix B. Proof of Bounded Aggregate Gate Strength

Proof: Let $S = \sum_d s_d$ for the un-clipped normalized-SNR scores. Since $s_d = 1 - \sigma_{u,d}^2/(\mu_{u,d}^2 + \sigma_{u,d}^2)$, the variance floor and Cauchy–Schwarz give

$$S \le \frac{\|\boldsymbol{\mu}_u\|_2^2}{\|\boldsymbol{\mu}_u\|_2^2/K + \sigma_{\min}^2}.$$

For $\alpha_d = \mathrm{clip}(s_d, a, b)$, $a \le \alpha_d \le b$ and $\alpha_d \le a + s_d$. Summing these inequalities yields $aK \le K_{\mathrm{eff}} \le bK$ and $K_{\mathrm{eff}} \le aK + S$; combining the two upper bounds proves Theorem 1. As $\|\boldsymbol{\mu}_u\|_2^2 \to 0$ under the positive variance floor, every $s_d \to 0$, so every $\alpha_d \to a$ and $K_{\mathrm{eff}} \to aK$.

The result is an upper bound on aggregate gate strength, not an equality or monotonic relationship involving interaction count. It formalizes only the limiting case of weak posterior-mean signal under a nonzero variance floor; the activity analysis reports recommendation accuracy by history length and does not test gate-strength monotonicity. □

Appendix C. Derivation of Conditional Descent Bound

Proof: At round t, stop-gradient defines F_t conditional on $\alpha^{(t)}$. The variance floor, compact parameter set, bounded gates, and bounded stochastic-gradient variance provide uniform constants. Applying the smoothness descent inequality to the E local updates, taking conditional expectations, and bounding client drift as in FedAvg [30] gives the fixed-surrogate terms $C_0/(\eta T) + C_1\eta + C_2\eta E^2$ for the average $\|\nabla F_t(\Theta^t)\|^2$.

The descent terms do not telescope directly because $F_{t+1} \ne F_t$. Inserting $F_{t+1}(\Theta^{t+1})$ introduces at most v_t, while inserting and subtracting the adjacent gradients and applying Young's inequality introduces a constant multiple of δ_t^2. Summation therefore adds $C_3 V_T/(\eta T) + C_4 D_T/T$. Setting $\eta = \mathcal{O}(T^{-1/2})$ yields Proposition 2. This derivation deliberately exposes the required variation budgets: it does not infer them from stop-gradient and does not identify the changing surrogate sequence with one fixed objective. □

References

1. Ko H, Lee S, Park Y, Choi A. A survey of recommendation systems: recommendation models, techniques, and application fields. Electronics. 2022;11(1):141.
2. Zhang S, Yao L, Sun A, Tay Y. Deep learning based recommender system: a survey and new perspectives. ACM Comput Surv. 2019;52(1):5. doi:10.1145/3285029.
3. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer. 2009;42(8):30–7. doi:10.1109/mc.2009.263.
4. Voigt P, Von dem Bussche A. The EU general data protection regulation (GDPR). A practical guide. 1st ed. Berlin/Heidelberg, Germany: Springer; 2017.
5. McMahan B, Moore E, Ramage D, Hampson S, Aguera y Arcas B. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS); 2017 Apr 20–22; Fort Lauderdale, FL, USA. p. 1273–82.
6. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. Found Trends Mach Learn. 2021;14(1–2):1–210. doi:10.1561/2200000083.
7. Ammad-Ud-Din M, Ivannikova E, Khan SA, Oyomno W, Fu Q. Federated collaborative filtering for privacy-preserving personalized recommendation system. arXiv:1901.09888. 2019.

8. Chai D, Wang L, Chen K, Yang Q. Secure federated matrix factorization. IEEE Intell Syst. 2021;36(5):11–20. doi:10.1109/MIS.2020.3014880.
9. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114. 2013.
10. Li Z, Long G, Zhou T, Jiang J, Zhang C. Personalized federated collaborative filtering: a variational autoencoder approach. Proc AAAI Conf Artif Intell. 2025;39(17):18602–10. doi:10.1609/aaai.v39i17.34047.
11. Higgins I, Matthey B, Pal A, Burgess C, Glorot X, McWilliams M, et al. β-VAE: learning basic visual concepts with a constrained variational framework. In: Proceedings of the International Conference on Learning Representations (ICLR); 2017 Apr 24–26; Toulon, France.
12. Liang D, Krishnan RG, Hoffman MD, Jebara T. Variational autoencoders for collaborative filtering. In: Proceedings of the International World Wide Web Conference (WWW); 2018 Apr 23–27; Lyon, France. p. 689–98.
13. Jiang X, Liu B, Qin J, Zhang Y, FedNCF QJ. Federated neural collaborative filtering for privacy-preserving recommender system. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN); 2022 Jul 18–23; Padua, Italy. p. 1–8.
14. Arivazhagan MG, Aggarwal V, Singh AK, Choudhary S. Federated learning with personalization layers. arXiv:1912.00818. 2019.
15. Zhang C, Long G, Zhou T, Yan P, Zhang Z, Zhang C, et al. Dual personalization on federated recommendation. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI); 2023 Aug 19–25; Macau, China. p. 4558–66. doi:10.24963/ijcai.2023/507.
16. Li Z, Long G, Zhou T. Federated recommendation with additive personalization. In: Proceedings of the International Conference on Learning Representations (ICLR); 2024 May 7–11; Vienna, Austria.
17. Shen J, Mi Y, Zhao G, Shen J, Qian X. A model-agnostic strategy to mitigate embedding degradation in personalized federated recommendation. arXiv:2508.19591. 2025.
18. Zhang H, Li C, Dai W, Zou J, FedCR XH. Personalized federated learning based on across-client common representation with conditional mutual information regularization. In: Proceedings of the International Conference on Machine Learning (ICML); 2023 Jul 23–29; Honolulu, HI, USA. p. 41314–30.
19. Karimireddy SP, Kale S, Mohri M, Reddi S, Stich S, Suresh AT. SCAFFOLD: stochastic controlled averaging for federated learning. In: Proceedings of the International Conference on Machine Learning (ICML); 2020 Jul 12–18; Virtual. p. 5132–43.
20. Polato M. Federated variational autoencoder for collaborative filtering. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN); 2021 Jul 18–22; Shenzhen, China. p. 1–8.
21. Zhang L, Rong Q, Ding X, Li G, Yuan L. EFVAE: efficient federated variational autoencoder for collaborative filtering. In: Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM); 2024 Oct 21–25; Boise, ID, USA. p. 3176–85.
22. Yao D, Liu T, Cao Q, Jin H. FedRKG: a privacy-preserving federated recommendation framework via knowledge graph enhancement. In: Proceedings of the 18th International Conference on Green, Pervasive, and Cloud Computing (GPC); 2023 Sep 22–24; Harbin, China. p. 81–96. doi:10.1007/978-981-99-9896-8_6.
23. He X, Deng K, Wang X, Li Y, Zhang Y, LightGCN WM. Simplifying and powering graph convolution network for recommendation. In: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR); 2020 Jul 25–30; Xi'an, China. p. 639–48.
24. Bowman SR, Vilnis L, Vinyals O, Dai AM, Jozefowicz R, Bengio S. Generating sentences from a continuous space. In: Proceedings of the Conference on Computational Natural Language Learning (CoNLL); 2016 Aug 11–12; Berlin, Germany. p. 10–21.
25. Kingma DP, Salimans T, Welling M. Variational dropout and the local reparameterization trick. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS); 2015 Dec 7–12; Montreal, QC, Canada. p. 2575–83.
26. Sonderby CK, Raiko T, Maaloe L, Sonderby SK, Winther O. Ladder variational autoencoders. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS); 2016 Dec 5–10; Barcelona, Spain. p. 3738–46.

27. Shenbin I, Alekseev A, Tutubalina E, Malykh V, Nikolenko SI. RecVAE: a new variational autoencoder for top-N recommendations with implicit feedback. In: Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM); 2020 Feb 3–7; Houston, TX, USA. p. 528–36.
28. MacKay DJC. Bayesian interpolation. Neural Comput. 1992;4(3):415–47. doi:10.1162/neco.1992.4.3.415.
29. Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res. 2001;1:211–44.
30. Li T, Sanjabi M, Beirami A, Smith V. Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems (MLSys); 2020 Mar 2–4; Austin, TX, USA. p. 429–50.
31. Ganguly B, Aggarwal V. Online federated learning via non-stationary detection and adaptation amidst concept drift. arXiv:2211.12578. 2022. doi:10.48550/arXiv.2211.12578.
32. Harper FM, Konstan JA. The MovieLens datasets: history and context. ACM Trans Interact Intell Syst. 2015;5(4):19. doi:10.1145/2827872.
33. Ni J, Li J, McAuley J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); 2019 Nov 3–7; Hong Kong, China. p. 188–97.