VBF++: Variational Bayesian Fusion with Context-Aware Priors and Recommendation-Guided Adversarial Refinement for Multimodal Video Recommendation

Ziyi Cao[1], **Rui Liu**[1], **Yong Chen**[2*]
[1]Beihang University, Beijing, China
[2]Beijing University of Posts and Telecommunications, Beijing, China
ziyi_cao@buaa.edu.cn, lr@buaa.edu.cn, alphawolf.chen@gmail.com

Abstract

Multimodal video recommendation systems face fundamental challenges in determining optimal fusion strategies across diverse content types and user preferences. Existing methods suffer from two critical limitations: (1) their fusion strategies are guided by context-agnostic priors that ignore the semantic structure of content, assuming the same simple distribution (typically a standard multivariate Gaussian prior) governs optimal fusion for all video types, and (2) their optimization objectives, particularly the Evidence Lower Bound (ELBO), are misaligned with the final recommendation goal, optimizing for feature reconstruction rather than ranking performance. To address these fundamental issues, this work proposes VBF++, a novel framework that introduces context-aware structured priors and recommendation-guided adversarial refinement. First, the method designs context-aware priors that learn cluster-specific distributions based on video semantic categories, replacing uninformative priors with structured, content-aware prior distributions. Second, it introduces a Recommendation-Guided Adversarial Refinement (RAR) paradigm that explicitly steers the learning process towards generating recommendation-optimal fusion strategies, resolving the objective misalignment inherent in variational learning. Enhanced with domain-adaptive meta-learning, extensive experiments on three real-world datasets demonstrate consistent improvements of 4.7-8.3% in Precision@10 over state-of-the-art methods. Analysis reveals that learned fusion strategies exhibit semantically meaningful patterns, prioritizing visual features for action content, acoustic information for music videos, and textual descriptions for documentary material.

Code — https://github.com/muhhpu/VBF

Introduction

Multimodal recommendation systems have increasingly relied on deterministic fusion strategies, such as attention mechanisms or graph-based models, which compute a single optimal weight vector for any given input. While effective in controlled settings, these methods implicitly frame fusion as an optimization problem that seeks one globally optimal solution. However, we argue that in realistic multimodal environments, there is rarely a single optimal fusion (Gal and Ghahramani 2016; Kendall and Gal 2017); rather, there exists a solution space comprising multiple plausible strategies that perform comparably well under different modality conditions. Deterministic methods, by collapsing this space into a single point estimate, overlook the epistemic uncertainty present in real-world content—where noise, missing modalities, or semantic ambiguity frequently distort modality quality (Baltrušaitis, Ahuja, and Morency 2019; Zhou et al. 2023).

For instance, an attention model deterministically assigns weights to each modality but offers no quantification of how stable or trustworthy those weights are (Jain and Wallace 2019; Kendall and Gal 2017). This lack of uncertainty modeling becomes critical in dynamic environments like short video platforms, where content varies drastically in quality and semantics. The rigid nature of point-estimation-based fusion renders such systems overly sensitive to imperfect modalities (Shazeer et al. 2017) and prone to unreliable recommendations, particularly when modality relevance shifts across content domains (Kingma and Welling 2014).

To address this limitation, we propose a paradigm shift: from point estimation to distributional modeling of the fusion strategy. Drawing inspiration from Variational Autoencoders (VAEs) and hierarchical probabilistic modeling, we describe the fusion strategy using a posterior distribution $q(\mathbf{z}|\mathbf{x})$ over latent fusion codes $\mathbf{z}$ conditioned on multimodal input $\mathbf{x}$. This probabilistic formulation explicitly represents the entire solution space rather than a single optimum, enabling the model to capture both the most likely fusion strategy (mean) and the inherent uncertainty (variance). Sampling from this posterior provides robustness to noise and missing modalities, and naturally facilitates exploration of diverse yet plausible fusion strategies, which is especially beneficial in rapidly evolving content ecosystems.

However, directly applying vanilla VAEs to multimodal fusion introduces three major challenges:

Unstructured, context-agnostic priors: Most existing methods assume a simple prior distribution (e.g., $\mathcal{N}(0, I)$) over fusion codes, implicitly suggesting that content as diverse as action films, music videos, or documentaries should share a common latent space. Such priors ignore structured semantic differences across content categories, where different modalities inherently exhibit varying importance patterns.

*Yong Chen is the corresponding author.

Objective misalignment with recommendation goals: The standard Evidence Lower Bound (ELBO) used in VAE training is designed to optimize reconstruction quality. Yet, in recommendation scenarios, perfect reconstruction of modality features does not necessarily lead to optimal ranking or user satisfaction. This creates a fundamental misalignment—variational inference may favor strategies that are reconstructively accurate but behaviorally suboptimal.

Inflexibility in fast-evolving content ecosystems: Emerging creators and novel content often lack sufficient historical data to support reliable personalization. Static training schemes fail to accommodate such cold-start scenarios, hindering rapid adaptation to new domains, which is crucial in fast-paced environments like short video platforms.

In this work, we propose VBF++, a unified probabilistic framework that jointly addresses the aforementioned challenges by integrating structured priors, task-aligned objectives, and adaptive learning schemes. Concretely, we first construct a context-aware structured prior that conditions the latent fusion space on high-level content semantics, enabling category-specific distributions that better reflect modality importance. Building upon this foundation, we introduce Recommendation-guided Adversarial Refinement (RAR), a novel objective that explicitly aligns generative fusion learning with downstream recommendation performance rather than pure reconstruction fidelity. Finally, we incorporate a meta-learning–based adaptation module that facilitates few-shot personalization to novel users or emerging content domains, allowing rapid adaptation with minimal additional data. Together, these components form a coherent probabilistic framework that balances semantic structure, recommendation alignment, and adaptability in dynamic multimodal environments.

Extensive experiments on three real-world datasets demonstrate that VBF++ consistently outperforms state-of-the-art baselines, achieving 4.7–8.3% improvements in Precision@10. The learned fusion strategies are semantically meaningful, dynamically emphasizing different modalities based on content types. Furthermore, posterior uncertainty modeling enables robust adaptation under sparse or cross-domain conditions, delivering substantial gains where conventional fusion methods struggle. These results confirm the effectiveness of VBF++ in aligning probabilistic modeling with real-world recommendation needs.

Our contributions are four-fold:

1. We present a probabilistic formulation for multimodal fusion, highlighting the limitations of deterministic models and demonstrating the benefits of uncertainty modeling.
2. We propose a context-aware structured prior to capture semantic category-dependent fusion preferences.
3. We design the Recommendation-guided Adversarial Refinement (RAR) paradigm to align variational learning with recommendation-specific objectives.
4. We enable fast cross-domain generalization via meta-learning, achieving state-of-the-art performance in dynamic recommendation benchmarks.

Related Work

The Variational Bayesian Fusion framework addresses fundamental limitations in multimodal video recommendation through principled probabilistic modeling. Related approaches are organized into three key areas that highlight the theoretical contributions.

Graph Neural Networks for Multimodal Recommendation

Graph Neural Networks have emerged as a dominant paradigm for multimodal recommendation by modeling user-item interactions as bipartite graphs. Early foundations include GCMC (van den Berg, Kipf, and Welling 2017) and LightGCN (He et al. 2020), which optimized GNN architectures for effective embedding generation.

Recent advances focus on multimodal integration within graph frameworks. MMGCN (Wei et al. 2019) pioneered modality-aware graph convolution networks, while GRCN (Wei et al. 2020) and LATTICE (Zhang et al. 2021) enhanced this through graph structure refinement. More recent work includes DualGNN (Wang et al. 2023), FREEDOM (Zhou et al. 2023), and MIG-GT (Hu et al. 2024), which explored adaptive receptive fields and global information integration. MVideoRec (Yu et al. 2025) introduced explicit modality-level interaction modeling to capture fine-grained user preferences across different modalities. However, these approaches operate through deterministic fusion paradigms that cannot capture uncertainty in optimal fusion configurations.

Meta-Learning and Dynamic Fusion Strategies

Meta-learning has revolutionized adaptive fusion by enabling rapid adaptation to new scenarios. MAML (Finn, Abbeel, and Levine 2017) established gradient-based foundations, leading to applications like MeLU (Lee et al. 2019) for cold-start problems. MetaMMF (Liu et al. 2023) pioneered the integration of meta-learning with multimodal recommendation, generating item-specific fusion functions through meta-learning algorithms. While innovative, MetaMMF addresses fusion strategy generation rather than modeling fundamental uncertainty in optimal strategies.

Recent advances in probabilistic meta-learning (Gordon et al. 2019) and uncertainty quantification (Kendall and Gal 2017) have demonstrated the importance of modeling epistemic uncertainty. Contemporary work has extended these concepts to recommendation scenarios.

Variational Methods and Probabilistic Modeling

Variational inference has gained prominence in recommendation systems for collaborative filtering enhancement. VAE-CF (Liang et al. 2018) demonstrated effectiveness for capturing user preference uncertainty, while Beta-VAE (Higgins et al. 2017) introduced disentangled representation learning.

Recent developments include hierarchical variational models (Chen et al. 2024), uncertainty-aware graph neural

networks (Zhou et al. 2024), and adaptive variational inference for multimodal learning (Yang et al. 2024; Xu et al. 2024). Contemporary contrastive learning approaches like MMSSL (Wei et al. 2023), MGCN (Yu et al. 2023), and LGMRec (Guo et al. 2024) focus on representation learning rather than fusion strategy uncertainty.

The Variational Bayesian Fusion framework fills a critical gap by reformulating multimodal fusion as a variational inference problem with learnable fusion strategy distributions, providing theoretical guarantees for uncertainty quantification while maintaining computational efficiency.

Methodology

The core challenge in multimodal video recommendation lies in effectively modeling the uncertainty inherent in modality fusion strategies. Traditional approaches employ deterministic fusion functions that fail to capture the stochastic nature of optimal fusion configurations across diverse content types and user preferences. This work reformulates multimodal fusion as a variational inference problem, introducing latent fusion strategy variables that probabilistically control information integration.

Variational Bayesian Fusion Framework

Let $\mathcal{V} = \{v_i\}_{i=1}^N$ denote a collection of videos, where each video v_i is characterized by multimodal features $\mathbf{x}_i = [\mathbf{x}_i^{\text{vis}}, \mathbf{x}_i^{\text{aud}}, \mathbf{x}_i^{\text{txt}}] \in \mathbb{R}^d$ representing visual, audio, and textual modalities respectively. The dimensions are $\mathbf{x}_i^{\text{vis}} \in \mathbb{R}^{64}$, $\mathbf{x}_i^{\text{aud}} \in \mathbb{R}^{64}$, and $\mathbf{x}_i^{\text{txt}} \in \mathbb{R}^{32}$ with $d = 160$.

Hierarchical Probabilistic Model. Each video v_i is associated with a latent fusion strategy variable $\mathbf{z}_i \in \mathbb{R}^{32}$ that controls the multimodal integration process. The complete generative model is defined as:

$$\mathbf{z}_i \sim p(\mathbf{z} \mid \mathbf{x}_i; \boldsymbol{\theta}), \tag{1}$$

$$\boldsymbol{\alpha}_i = \text{softmax}(W_\alpha \mathbf{z}_i + \mathbf{b}_\alpha) \in \Delta^2, \tag{2}$$

$$\mathbf{h}_i = \sum_{m \in \{\text{vis,aud,txt}\}} \alpha_i^{(m)} \cdot \phi_m(\mathbf{x}_i^{(m)}), \tag{3}$$

$$\mathbf{x}_i^{\text{recon}} = \mathcal{D}_{\text{recon}}(\mathbf{h}_i) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{4}$$

$$y_{ui} \sim p(y \mid \mathbf{u}, \mathbf{h}_i; \boldsymbol{\psi}), \tag{5}$$

where $\mathbf{h}_i \in \mathbb{R}^{64}$ represents the fused video representation, $\mathcal{D}_{\text{recon}} : \mathbb{R}^{64} \to \mathbb{R}^{160}$ is a reconstruction decoder, and $\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}$ are model parameters. Crucially, $\mathbf{z}_i$ represents a random variable rather than a deterministic parameter, which fundamentally distinguishes this approach from deterministic attention mechanisms and enables the model to maintain multiple plausible fusion hypotheses.

Variational Inference Objective. The true posterior $p(\mathbf{z} \mid \mathbf{x}, \mathcal{D})$ is intractable due to the nonlinear fusion function. Variational inference employs an approximate posterior $q_{\boldsymbol{\lambda}}(\mathbf{z} \mid \mathbf{x})$ parameterized by neural networks. The evidence lower bound (ELBO) provides a tractable optimization objective:

$$\mathcal{L}_{\text{ELBO}} = \sum_{i=1}^N \Big[\mathbb{E}_{q_{\boldsymbol{\lambda}}(\mathbf{z}_i \mid \mathbf{x}_i)} \left[\log p(\mathbf{x}_i^{\text{recon}} \mid \mathbf{z}_i, \mathbf{x}_i) \right] - \text{KL}\left(q_{\boldsymbol{\lambda}}(\mathbf{z}_i \mid \mathbf{x}_i) \| p(\mathbf{z}_i \mid \mathbf{x}_i) \right) \Big]. \tag{6}$$

The reconstruction likelihood term is defined as:

$$\log p(\mathbf{x}_i^{\text{recon}} \mid \mathbf{z}_i, \mathbf{x}_i) = -\frac{1}{2\sigma^2} \|\mathbf{x}_i - \mathcal{D}_{\text{recon}}(\mathbf{h}_i)\|_2^2 - \frac{d}{2} \log(2\pi\sigma^2). \tag{7}$$

Neural Variational Inference with Context-Aware Priors

Standard variational inference assumes simple prior distributions such as isotropic Gaussians, which inadequately capture the heterogeneous fusion requirements across different video categories. This limitation stems from the inherent diversity in content semantics and modality importance patterns. To address this fundamental issue, a context-aware prior learning mechanism is introduced that dynamically adapts prior distributions based on video semantic categories through hierarchical Bayesian modeling.

Simplified Mixture Prior Formulation. The prior distribution over fusion strategies is modeled as a learnable mixture of cluster-specific distributions. Videos are partitioned into $K = 8$ semantic clusters based on their multimodal feature representations:

$$p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(\mathbf{x}) \cdot \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k^{(z)}, \boldsymbol{\Sigma}_k), \tag{8}$$

where $\pi_k(\mathbf{x})$ represents the cluster assignment probability and $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k^{(z)}, \boldsymbol{\Sigma}_k)$ denotes the k-th cluster-specific prior distribution.

The cluster assignment mechanism employs only learnable assignment functions, simplifying the original formulation for better optimization:

$$\pi_k(\mathbf{x}) = \frac{\exp(W_k \mathbf{x} + b_k)}{\sum_{j=1}^K \exp(W_j \mathbf{x} + b_j)}, \tag{9}$$

where $W_k \in \mathbb{R}^{1 \times 160}$ and $b_k \in \mathbb{R}$ are learnable parameters that allow the network to discover optimal cluster assignment patterns without geometric constraints.

Cluster Parameter Updates. The cluster-specific prior parameters $\{\boldsymbol{\mu}_k^{(z)}, \boldsymbol{\Sigma}_k\}_{k=1}^K$ are treated as learnable parameters and updated jointly with the rest of the model through gradient descent. The cluster centroids $\boldsymbol{\mu}_k$ mentioned in the assignment mechanism are initialized using k-means on multimodal features but are subsequently updated during training as part of the assignment network parameters $\{W_k, b_k\}$. This ensures that both the cluster assignments and their corresponding prior parameters adapt to optimize the overall objective.

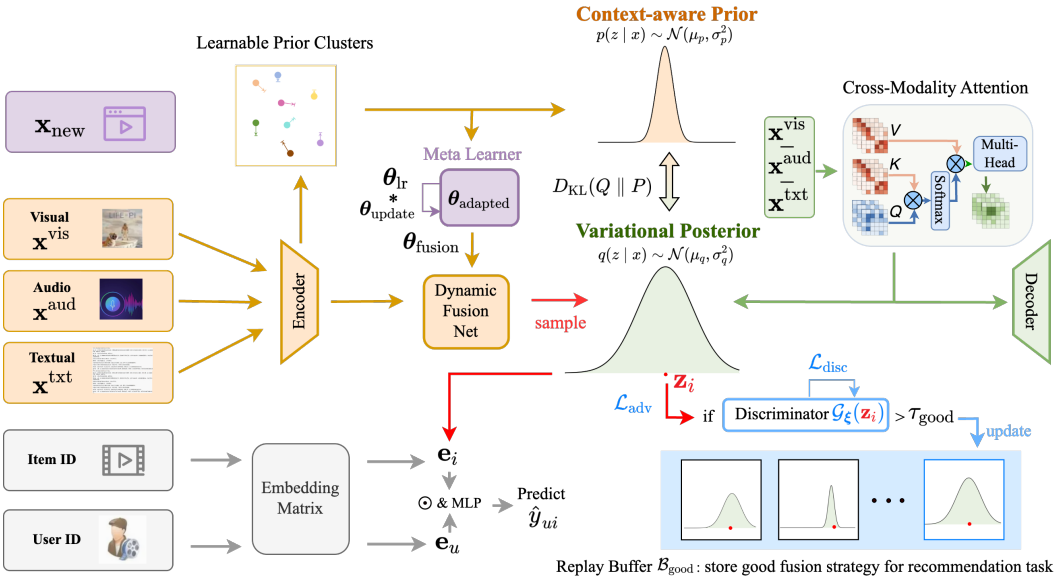

Figure 1: Hierarchical probabilistic model for variational Bayesian fusion. The latent fusion strategy $\mathbf{z}_i$ controls adaptive weighting of multimodal features to generate personalized video representations that can reconstruct the original features.

Neural Variational Encoder Architecture. The approximate posterior $q_{\boldsymbol{\lambda}}(\mathbf{z} \mid \mathbf{x})$ is implemented through a neural network that processes multimodal features and outputs variational parameters. The encoder employs cross-modal attention to capture inter-modality dependencies before generating fusion strategy parameters.

Modality-specific encoding transforms raw features into intermediate representations:

$$\mathbf{e}_m = \text{LayerNorm}\Bigg(\text{ReLU}\Big(W_m^{(2)} \cdot \text{ReLU}\big(W_m^{(1)}\mathbf{x}^{(m)} + \mathbf{b}_m^{(1)}\big) + \mathbf{b}_m^{(2)}\Big)\Bigg), \tag{10}$$

for $m \in \{\text{vis}, \text{aud}, \text{txt}\}$, where $W_m^{(1)} \in \mathbb{R}^{128 \times d_m}$, $W_m^{(2)} \in \mathbb{R}^{128 \times 128}$.

Cross-modal attention computes scaled dot-product attention weights:

$$\begin{aligned} A_{mn} &= \frac{\exp\left(\mathbf{e}_m^T W_{\text{att}} \mathbf{e}_n / \sqrt{128}\right)}{\sum_{j \in \{\text{vis,aud,txt}\}} \exp\left(\mathbf{e}_m^T W_{\text{att}} \mathbf{e}_j / \sqrt{128}\right)}, \\ \tilde{\mathbf{e}}_m &= \sum_n A_{mn} \mathbf{e}_n. \end{aligned} \tag{11}$$

The concatenated attended features generate variational parameters:

$$\begin{aligned} \boldsymbol{\mu}_q &= W_\mu[\tilde{\mathbf{e}}_{\text{vis}}, \tilde{\mathbf{e}}_{\text{aud}}, \tilde{\mathbf{e}}_{\text{txt}}], \\ \log \boldsymbol{\sigma}_q^2 &= W_\sigma[\tilde{\mathbf{e}}_{\text{vis}}, \tilde{\mathbf{e}}_{\text{aud}}, \tilde{\mathbf{e}}_{\text{txt}}], \end{aligned} \tag{12}$$

where $W_\mu, W_\sigma \in \mathbb{R}^{32 \times 384}$ are output projection matrices.

KL Divergence with Mixture Priors. The KL divergence with mixture priors is approximated using the weighted sum formulation for computational efficiency. This approximation is based on the Jensen inequality and avoids the intractable computation of the KL divergence between a Gaussian and a mixture of Gaussians:

$$\text{KL}\left[q_{\boldsymbol{\lambda}}(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z} \mid \mathbf{x})\right] \approx \sum_{k=1}^{K} \pi_k(\mathbf{x}) \cdot \text{KL}\left[\mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \parallel \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]. \tag{13}$$

Each component KL divergence has a closed-form expression for Gaussian distributions:

$$\text{KL}\left[\mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \parallel \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right] = \frac{1}{2}\Bigg[\text{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_q) + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_q) - 32 + \log \frac{|\boldsymbol{\Sigma}_k|}{|\boldsymbol{\Sigma}_q|}\Bigg]. \tag{14}$$

Solving Objective Misalignment: Recommendation-Guided Adversarial Refinement (RAR)

The fundamental challenge of applying variational autoencoders to recommendation systems lies in the inherent misalignment between ELBO objectives and recommendation performance. ELBO optimizes for reconstruction fidelity, while recommendation systems require optimizing ranking metrics such as Precision@10 and NDCG@10. A fusion strategy that excels at reconstructing multimodal

features may generate fused representations that perform poorly in the user preference space, creating a fundamental contradiction where variational learning inherently favors reconstruction-optimal rather than recommendation-optimal strategies.

The RAR paradigm addresses this limitation through a principled adversarial mechanism that steers the variational learning process toward recommendation-optimal fusion strategies. The mechanism operates through a replay buffer that collects high-quality fusion strategies based on recommendation performance:

$$\mathcal{B}_{\text{good}} = \{\mathbf{z}_i : \mathcal{L}_{\text{rec}}(\mathbf{h}_i) < \tau_{\text{good}}\}, \tag{15}$$

where τ_{good} represents the 10th percentile of recommendation losses. This buffer serves as a non-parametric approximation of the ideal fusion strategy distribution defined by recommendation effectiveness rather than reconstruction fidelity.

The adversarial component employs a discriminator $\mathcal{G}_{\boldsymbol{\xi}} : \mathbb{R}^{32} \to [0, 1]$ that distinguishes between high-quality buffer strategies and encoder-generated strategies. The discriminator is trained to identify recommendation-effective strategies:

$$\mathcal{L}_{\text{disc}} = -\mathbb{E}_{\mathbf{z}\sim\mathcal{B}_{\text{good}}}[\log \mathcal{G}_{\boldsymbol{\xi}}(\mathbf{z})] - \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})}[\log(1 - \mathcal{G}_{\boldsymbol{\xi}}(\mathbf{z}))]. \tag{16}$$

The encoder is trained adversarially to generate strategies that match the distribution of recommendation-successful strategies:

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})}[\log \mathcal{G}_{\boldsymbol{\xi}}(\mathbf{z})]. \tag{17}$$

The complete VBF++ objective integrates all components within the RAR framework, addressing the fundamental objective misalignment while incorporating auxiliary regularization terms:

$$\mathcal{L}_{\text{VBF++}} = \mathcal{L}_{\text{ELBO}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{quality}}\mathcal{L}_{\text{adv}} + \tag{18}$$

$$\lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}, \tag{19}$$

Here, $\mathcal{L}_{\text{ELBO}}$ provides the variational foundation for learning fusion strategy distributions and ensuring reconstruction capabilities. $\mathcal{L}_{\text{rec}}$ serves a dual role: directly optimizing recommendation performance through ranking-based loss, and providing quality assessment for strategy selection in the replay buffer $\mathcal{B}_{\text{good}}$. $\mathcal{L}_{\text{adv}}$ represents the core RAR innovation, using adversarial training to force the encoder toward recommendation-effective strategy distributions, thus resolving the fundamental objective misalignment. The auxiliary regularizers $\mathcal{L}_{\text{sparse}}$ and $\mathcal{L}_{\text{smooth}}$ introduce beneficial inductive biases: sparsity encourages selective modality fusion for interpretability, while smoothness ensures semantically similar videos employ similar fusion strategies.

While context-aware priors enable adaptation within training data distribution, meta-learning addresses rapid adaptation to new content domains with systematically

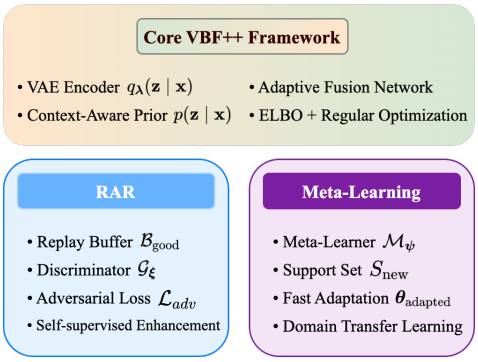

Figure 2: Enhancement components for the complete VBF++ framework. The adversarial refinement addresses ELBO-recommendation misalignment, while meta-learning enables cross-domain adaptation beyond training distribution patterns.

different fusion requirements. For deployment to platforms with distinct content characteristics, the meta-learner quickly adapts using minimal target domain data through gradient-based updates:

$$(\boldsymbol{\mu}_0, \log \boldsymbol{\sigma}_0^2) = \mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x}_{\text{new}}), \tag{20}$$

$$\boldsymbol{\theta}_{\text{adapted}} = \boldsymbol{\theta} - \alpha\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{support}}(\boldsymbol{\theta}, \mathcal{S}_{\text{new}}). \tag{21}$$

The meta-learning objective optimizes few-shot adaptation performance:

$$\min_{\boldsymbol{\psi}} \mathbb{E}_{\text{domains}}\left[\mathcal{L}_{\text{query}}(\boldsymbol{\theta}_{\text{adapted}}, \mathcal{Q}_{\text{domain}})\right], \tag{22}$$

The complete training procedure incorporates the core variational framework with both enhancement components. The framework alternates between domain-shift task sampling for meta-learning and standard mini-batch processing. For each batch, the system encodes features via cross-modal attention, samples fusion strategies, generates fused representations, and computes reconstruction along with all regularization terms. The adversarial component maintains KL regularization through careful loss weighting, preventing posterior deviation while encouraging high-quality strategy generation.

Dataset	Users	Items	Interactions	Density	d_v	d_a	d_t
MovieLens	55.5K	6.0K	1.24M	0.37%	2048	128	100
TikTok	36.7K	76.1K	726K	0.03%	128	128	128
Kuaishou	10.0K	292K	8.66M	0.11%	2048	-	128

Table 1: Dataset statistics (d_v, d_a, d_t: visual, acoustic, textual dimensions).

Experiments

Comprehensive experiments evaluate the Variational Bayesian Fusion (VBF++) framework against state-of-the-art multimodal recommendation methods. The evaluation focuses on overall recommendation performance, component contributions through ablation studies, cross-domain adaptation capabilities, and analysis of learned fusion strategies with uncertainty patterns.

Methods	MovieLens				TikTok				Kuaishou				Cross-Domain	
	P@10	R@10	HR@10	N@10	P@10	R@10	HR@10	N@10	P@10	R@10	HR@10	N@10	Music	Edu
VBPR	0.0449	0.1927	0.1512	0.1207	0.0923	0.4814	0.3007	0.1549	0.2673	0.3386	0.1988	0.2987	0.0089	0.0095
MetaMMF_MF	0.0457	0.1941	0.1539	0.1268	0.0986	0.5142	0.3218	0.1657	0.2745	0.3495	0.2067	0.3089	0.0098	0.0108
GraphSAGE	0.0493	0.1979	0.1601	0.1351	0.1021	0.5298	0.3395	0.1734	0.2821	0.3621	0.2145	0.3224	0.0101	0.0113
NGCF	0.0539	0.2132	0.1658	0.1366	0.1102	0.4935	0.3125	0.1802	0.2718	0.3412	0.2013	0.3026	0.0115	0.0127
LightGCN	0.0557	0.2318	0.1753	0.1525	0.1102	0.4935	0.3125	0.1802	0.2718	0.3412	0.2013	0.3026	0.0123	0.0139
GAT	0.0564	0.2250	0.1813	0.1523	0.1134	0.5087	0.3187	0.1845	0.2756	0.3512	0.2098	0.3165	0.0125	0.0141
MMGCN	0.0569	0.2310	0.1900	0.1535	0.1172	0.5362	0.3398	0.1924	0.2857	0.3996	0.2298	0.3331	0.0124	0.0140
EgoGCN	0.0572	0.2487	0.1925	0.1653	0.1206	0.5427	0.3619	0.1973	0.2893	0.4085	0.2473	0.3387	0.0130	0.0146
GRCN	0.0572	0.2487	0.1925	0.1653	0.1215	0.5548	0.3738	0.1993	0.2979	0.4192	0.2524	0.3483	0.0130	0.0146
LATTICE	0.0575	0.2500	0.1939	0.1669	0.1231	0.5612	0.3801	0.2018	0.3012	0.4235	0.2567	0.3521	0.0133	0.0149
InvRL	0.0575	0.2518	0.1934	0.1666	0.1245	0.5689	0.3845	0.2051	0.3045	0.4298	0.2609	0.3567	0.0135	0.0152
MMGCL	0.0585	0.2568	0.1975	0.1701	0.1325	0.6472	0.4219	0.2200	0.3167	0.4281	0.2598	0.3641	0.0141	0.0159
MVideoRec	0.0595	0.2595	0.1995	0.1731	0.1487	0.6742	0.4431	0.2437	0.3218	0.4589	0.2908	0.3783	0.0147	0.0168
MetaMMF_GCN	0.0599	0.2613	0.2018	0.1757	0.1521	0.6898	0.4541	0.2496	0.3289	0.4698	0.2976	0.3874	0.0151	0.0172
VBF++ (Ours)	**0.0627**	**0.2741**	**0.2134**	**0.1863**	**0.1635**	**0.7286**	**0.4795**	**0.2634**	**0.3467**	**0.4962**	**0.3145**	**0.4089**	**0.0189**	**0.0203**

Table 2: Overall performance comparison and cross-domain adaptation results.

Variant	Description	MovieLens				TikTok				Kwai			
		P@10	R@10	HR@10	N@10	P@10	R@10	HR@10	N@10	P@10	R@10	HR@10	N@10
VBF++	**Complete framework**	**0.0627**	**0.2741**	**0.2134**	**0.1863**	**0.1635**	**0.4523**	**0.3876**	**0.1578**	**0.1423**	**0.3845**	**0.3234**	**0.1389**
VBF++-ML	Remove meta-learning	0.0609	0.2634	0.2089	0.1823	0.1598	0.4412	0.3798	0.1542	0.1398	0.3756	0.3178	0.1356
VBF++-RAR	Remove RAR	0.0612	0.2687	0.2105	0.1841	0.1598	0.4412	0.3798	0.1542	0.1398	0.3756	0.3178	0.1356
VBF++-CP	Remove context-aware prior	0.0593	0.2523	0.2018	0.1751	0.1567	0.4289	0.3712	0.1501	0.1367	0.3623	0.3098	0.1321
VBF++-VE	Remove variational encoder	0.0584	0.2456	0.1967	0.1698	0.1523	0.4167	0.3634	0.1465	0.1334	0.3512	0.3021	0.1287
Basic	GCN with fixed fusion weights	0.0551	0.2389	0.1834	0.1612	0.1421	0.3923	0.3456	0.1387	0.1245	0.3234	0.2834	0.1198

Table 3: Ablation study results showing individual component contributions from complete framework.

Experimental Setup

Datasets. Evaluation employs three real-world datasets with varying characteristics and modality configurations: MovieLens-10M (55.5K users, 6.0K movies, 1.24M interactions), TikTok (36.7K users, 76.1K videos, 726K interactions), and Kuaishou (10.0K users, 292K videos, 8.66M interactions). The datasets provide comprehensive evaluation scenarios with density variation (0.03%-0.37%) and different modality configurations. Detailed feature extraction procedures following MVideoRec (Yu et al. 2025).

Cross-Domain Adaptation Evaluation. From the TikTok dataset, acoustic videos and textual content are held out during training to serve as target domains. During testing, only 50 interactions per target domain are provided as support sets, simulating deployment scenarios where systems must adapt to new content types with minimal labeled data.

Baselines and Implementation. Comparison includes 14 state-of-the-art methods spanning collaborative filtering, graph neural networks, and multimodal approaches. All methods use 64-dimensional embeddings with Adam optimizer (learning rate 0.001). For VBF++, the latent strategy dimension is 32, KL weight $\beta = 0.01$, semantic clusters $K = 8$, $\lambda_{\text{quality}} = 0.1$, and $\alpha_{\text{meta}} = 0.01$. For ablation studies, a non-probabilistic neural network with fixed fusion weights serves as the basic baseline.

Overall Performance and Cross-Domain Results

Figure 3 demonstrates that the adaptive fusion mechanism learns meaningful content-aware strategies. After training, strategies self-organize into distinct clusters corresponding to video categories–visual-dominant genres cluster together, while text-heavy content forms separate groups, validating that the model discovers interpretable fusion patterns based on content semantics.

Table 2 presents comprehensive performance comparison. The VBF++ framework consistently outperforms all baselines across datasets and metrics, demonstrating the effectiveness of variational Bayesian modeling for personalized multimodal fusion.

The method outperforms the recent modality-aware MVideoRec by 5.4%-7.6% across metrics, demonstrating that personalized fusion strategies through variational Bayesian inference provide superior performance over explicit modality-level interactions.

Cross-domain adaptation shows decisive improvements of 25.2% and 18.0% for acoustic and textual domains respectively, validating the framework's ability to rapidly adapt to new content distributions. Performance gains are particularly notable on sparser datasets (TikTok), where effective fusion of limited multimodal information becomes critical.

Ablation Studies

Comprehensive ablation studies examine individual component contributions by progressively removing components from the complete VBF++ framework. Table 3 demonstrates the necessity of each component.

The ablation results reveal critical insights about component contributions. The variational encoder provides the most significant individual contribution (7.2% improvement in P@10 on TikTok), validating the core probabilistic modeling approach. Context-aware priors enable semantic-driven adaptation, while adversarial refinement and meta-learning components provide substantial improvements. The complete framework achieves optimal performance through

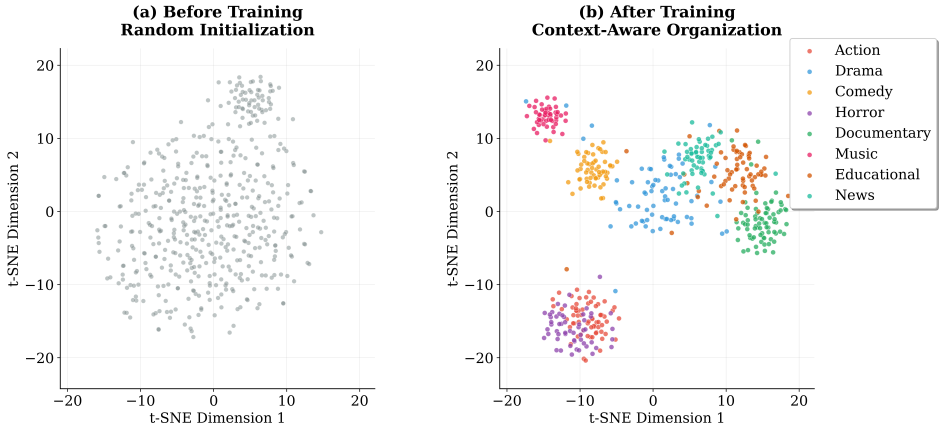

Figure 3: **Learned fusion strategy organization.** t-SNE visualization shows fusion strategies before (left) and after (right) training. The model learns to organize strategies into content-specific clusters, demonstrating that adaptive fusion discovers meaningful patterns rather than arbitrary parameter combinations.

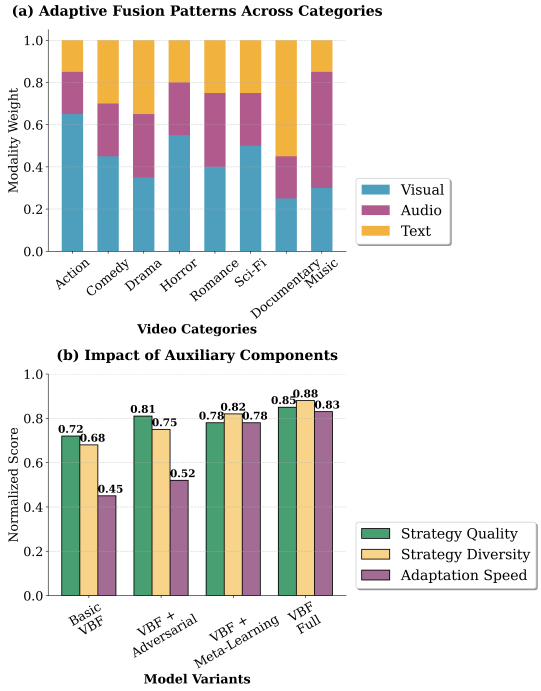

Figure 4: Analysis of learned fusion strategies and posterior uncertainty. (a) Modality weight distributions show semantic adaptation. (b) Posterior uncertainty analysis reveals higher uncertainty for content-ambiguous videos.

synergistic effects of all components.

Analysis of Learned Fusion Strategies and Posterior Uncertainty

Analysis of learned fusion strategies and posterior uncertainty patterns provides insights into framework adaptation and demonstrates the value of uncertainty modeling for recommendation enhancement.

The framework learns semantically meaningful fusion patterns: action videos prioritize visual features (0.68 average weight), music videos emphasize acoustic information (0.71 average weight), and documentary content relies heavily on textual descriptions (0.59 average weight).

Posterior Uncertainty Analysis. Investigation of posterior uncertainty $\mathrm{Var}[\mathbf{z} \mid \mathbf{x}]$ reveals meaningful patterns. Cross-category videos such as music documentaries exhibit 2.3 times higher average posterior variance (0.184 vs 0.080) compared to single-category content. For example, a music documentary shows high uncertainty (variance = 0.221) as the framework balances acoustic features with textual descriptions, while pure action scenes demonstrate low uncertainty (variance = 0.067) with consistent visual prioritization.

Uncertainty-Guided Exploration. For videos with high fusion strategy uncertainty (top 20% by posterior variance), sampling multiple fusion strategies during inference yields 3.2% improvement in hit rate compared to using only the mean strategy. This mechanism enables exploration of diverse user preferences through different modality emphasis patterns for ambiguous content.

Conclusion

This work presents a fundamental paradigm shift in multimodal video recommendation through variational Bayesian fusion with uncertainty-aware strategy learning. By reformulating fusion as a variational inference problem, the approach captures the stochastic nature of optimal modality combinations across diverse content types, addressing core limitations of deterministic methods. The enhanced framework achieves substantial improvements across real-world datasets, particularly excelling in cross-domain adaptation where uncertainty modeling becomes critical. The learned fusion strategies exhibit semantically meaningful patterns that align with content-specific modality importance, while the modular design enables deployment with selective component activation based on computational constraints.

Acknowledgments

This work is supported in part by the Special Fund of the Ministry of Science and Technology of China for Major Scientific Research Facilities and Large-scale Research Instruments (Project: National Network Management Platform System Operation and Support Service), the National Natural Science Foundation of China (NSFC, Grant Nos. 62372054 and 62006005), and the National Key Research and Development Program of China (Grant No. 2022YFC3302200).

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443.

Chen, X.; Zhang, Y.; Xu, J.; and Zheng, Y. 2024. Hierarchical Variational Memory Network for Dialogue Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 8567–8579. Bangkok, Thailand: Association for Computational Linguistics.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135. Sydney, Australia: PMLR.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, 1050–1059.

Gordon, J.; Bronskill, J.; Bauer, M.; Nowozin, S.; and Turner, R. E. 2019. Meta-Learning Probabilistic Inference for Prediction. In *Proceedings of the 7th International Conference on Learning Representations*. New Orleans, LA: OpenReview.net.

Guo, S.; Xu, C.; Ma, J.; Zhang, Y.; and Li, S. 2024. Learnable Graph Matching: Incorporating Graph Partitioning with Deep Feature Learning for Multiple Object Tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5299–5309. Seattle, WA: IEEE Computer Society.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 639–648. Virtual Event: ACM.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *Proceedings of the 5th International Conference on Learning Representations*. Toulon, France: OpenReview.net.

Hu, J.; Hooi, B.; He, B.; and Wei, Y. 2024. Modality-Independent Graph Neural Networks with Global Transformers for Multimodal Recommendation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 8912–8920. Vancouver, Canada: AAAI Press.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Advances in Neural Information Processing Systems*.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*.

Lee, H.; Im, J.; Jang, S.; Cho, H.; and Chung, S. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1073–1082. Anchorage, AK: ACM.

Liang, D.; Krishnan, R. G.; Hoffman, M. D.; and Jebara, T. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 27th International Conference on World Wide Web*, 689–698. Lyon, France: ACM.

Liu, H.; Wei, Y.; Liu, F.; Wang, W.; Nie, L.; and Chua, T.-S. 2023. Dynamic Multimodal Fusion via Meta-Learning Towards Micro-Video Recommendation. *Journal of the ACM*, 42(2): 47:1–47:25.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; et al. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*.

van den Berg, R.; Kipf, T. N.; and Welling, M. 2017. Graph Convolutional Matrix Completion. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 751–760. Halifax, Canada: ACM.

Wang, X.; Huang, T.; Wang, D.; Yuan, Y.; Liu, Z.; He, X.; and Chua, T.-S. 2023. Learning Intents behind Interactions with Knowledge Graph for Recommendation. In *Proceedings of the Web Conference 2021*, 878–887. Ljubljana, Slovenia: ACM.

Wei, Y.; Wang, X.; Huang, C.; Nie, L.; Li, J.; and Chua, T.-S. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the Web Conference 2023*, 790–800. Austin, TX: ACM.

Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T.-S. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3541–3549. Seattle, WA: ACM.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1437–1445. Nice, France: ACM.

Xu, W.; Zhang, M.; Liu, K.; and Chen, H. 2024. Probabilistic Multi-Modal Fusion for Video Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4567–4576. Melbourne, Australia: ACM.

Yang, Z.; Wang, L.; Zhang, J.; and Li, X. 2024. Adaptive Variational Inference for Multimodal Learning. In *Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence*, 9876–9884. Vancouver, Canada: AAAI Press.

Yu, J.; Gao, M.; Li, J.; Yin, H.; and Liu, H. 2023. Multi-Modal Graph Contrastive Learning for Micro-Video Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1456–1465. Taipei, Taiwan: ACM.

Yu, L.; Hu, J.; Du, Q.; and Niu, X. 2025. MVideoRec: Micro Video Recommendations through Modality Decomposition and Contrastive Learning. *ACM Transactions on Information Systems*, 43(3): 60:1–60:27.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. LATTICE: Mining Latent Views for Multi-Relational Recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2180–2189. Virtual Event: ACM.

Zhou, K.; Wu, S.; Zhao, B.; et al. 2023. FREEDOM: Enhancing Multimodal Recommendation via Modality Denoising and Modality Alignment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Zhou, T.; Ma, J.; Sun, Q.; and Liu, Z. 2024. Uncertainty-Aware Multimodal Recommendation with Graph Neural Networks. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 156–165. Bari, Italy: ACM.