Cross-scenario Multi-modal Knowledge Fusion and Knowledge Recommendation Based on a MDR-DKD Model

Jiang Jiang* and Xuxian Wang

Guangdong Power Grid Co., Ltd., CSG, Guangzhou 510000, Guangdong, China
E-mail: 15819889075@163.com
**Corresponding Author*

Received 20 October 2025; Accepted 28 November 2025

Abstract

With the widespread application of recommendation systems in e-commerce, education, and other fields, the heterogeneity of cross-scenario data and the insufficient integration of multi-modal information such as text, images, and user behavior are becoming increasingly prominent. To achieve cross-scenario multi-modal knowledge fusion and knowledge recommendation, a meta doubly robust-debiasing knowledge distillation (MDR-DKD) model is proposed. This model efficiently extracts universal features cross-scenarios using a small amount of unbiased data through a meta-learning mechanism and optimizes the model by combining knowledge distillation techniques. Finally, combined with the knowledge recommendation module, targeted knowledge recommendation is achieved by calculating the matching degree between user interests and knowledge nodes. The results showed that the multi-modal feature extraction of the model took an average of 18.61 ms, the parameter utilization rate during the feature extraction process was 91.3%, the feature extraction throughput reached 2460 samples/s, and the knowledge recommendation accuracy was 97.84%. This model can effectively extract

Journal of Web Engineering, Vol. 25_2, 187–214.
doi: 10.13052/jwe1540-9589.2523

cross-scenario multi-modal features for accurate knowledge recommendation. The research provides an effective technical path for cross-domain knowledge recommendation, which can promote the implementation of recommendation systems in multi-scenario and multi-modal practical scenarios, and help improve the personalized recommendation experience for users.

Keywords: Knowledge distillation, cross-scenario multi-modal, feature extraction, knowledge fusion, knowledge recommendation.

Notation

Symbols	Explanation
$\widehat{\mathbf{h}}_{s,m}$	The feature vector of modality m normalized by LayerNorm in scene s
$\mathrm{LN}(\cdot)$	The LayerNorm normalization function
$\mu_{s,m}$	The mean of the feature vector $\mathbf{h}_{s,m}$
$\sigma^2_{s,m}$	The variance of the feature vector $\mathbf{h}_{s,m}$
ε	The minimum value used to prevent calculation errors caused by zero variance (take 10-5)
γ_m and β_m	The LayerNorm learnable parameters (scaling factor and offset factor) exclusive to modality m
d_m	The dimension of $\mathbf{h}_{s,m}$.
$\mathbf{f}^{\mathrm{ali}}_{s,m}$	The aligned feature vector of modality m mapped to a unified semantic space in scene s
$\mathbf{T}_m$	The learnable linear transformation matrix exclusive to mode m.
b_m	A learnable bias term exclusive to modality m
$g_{s,m}$	The gating coefficient of mode m in scene s,
$\sigma_{\mathrm{sig}}(\cdot)$	The sigmoid activation function.
$\mathbf{W}_g$	The learnable weight matrix
b_g	Bias term of the gating unit
$\mathbf{f}^{\mathrm{self}}_{s,m}$	The feature vector of modality m refined by 16 heads of self attention in scene s
$\mathbf{f}^{\mathrm{gate}}_{s,m}$	The feature vector of modality m filtered by the gating unit in scene s.
$\odot$	Element wise multiplication
$\mathcal{L}_{\mathrm{align}}$	The modal alignment distillation loss (loss of one of the dual distillation branches).

Symbols	Explanation
$\lambda_1 = 0.6$	The modal alignment loss weight when calculating the inverse KL divergence equation
M	The total number of modalities
$D_{\mathrm{KL}}^{\mathrm{rev}}(\cdot)$	The reverse KL divergence
$P_{\mathrm{T},m}$ and $P_{\mathrm{S},m}$	The probability distributions of the mapping results of the modal m of the teacher model T and the student model S in the unified distillation space
$e_{i,\mathrm{gat}}$	The weighted representation of the entity after GAT processing
$e_{j,\mathrm{raw}}$	The original feature of e_j
$\alpha_{i,u}$	The GAT attention weight based on user interests
$\mathbf{W}_{\mathrm{gat}}$	The feature transformation matrix
b_{gat}	Bias term vector
$\mathrm{LeakyReLU}(\cdot)$	The linear rectification function

1 Introduction

Against the backdrop of the rapid development of intelligent recommendation systems, the efficient fusion of cross-scenario multimodal data and personalized knowledge recommendation have become research hotspots [1]. Significant differences in data distribution and semantics exist across different scenarios, posing severe challenges for models in scenario transfer and feature alignment [2]. Traditional recommendation algorithms mostly focus on a single modality or fixed scenarios, neglecting knowledge transfer across scenarios and the complementarity between modalities, which makes it difficult to achieve accurate recommendations in complex environments [3, 4]. Particularly in terms of multimodal semantic expression and cross-scenario knowledge alignment, issues of data heterogeneity and bias still constrain the generalization performance and interpretability of recommendation systems. Based on this, numerous scholars have conducted extensive research on cross-scenario multimodal knowledge processing [5]. In the field of cross-scenario multimodal knowledge fusion, Huang et al. proposed an effective multimodal representation and fusion method to achieve multimodal knowledge fusion in complex scenarios. The results indicated that the multimodal fusion achieved by this method was adaptive and effectively reduced potential noise interference [6]. Yue et al., aiming to identify knowledge languages across different scenarios, introduced a new model named KnowleNet. The findings revealed that this model could leverage the ConceptNet knowledge

base to integrate prior knowledge from various scenarios [7]. Xing et al., in order to consider the inferential correlations among data during cross-scenario multimodal knowledge fusion, proposed a multimodal semantic representation and fusion model based on knowledge graphs (KGs). The results demonstrated that, compared to existing models, this model exhibited advantages in multimodal semantic representation, fusion, transmission efficiency, and channel robustness [8].

In the realm of knowledge recommendation, Ma et al. proposed a novel approach called knowledge-aware reasoning with a graph convolution network (KR-GCN) to provide interpretable knowledge recommendation systems. The outcomes showed that, compared to GCN, KR-GCN enhanced recommendation performance and ensured the diversity of explanations [9]. Rubel et al. conducted an analysis of the disparity between traditional knowledge recommendation judgments and knowledge recommendation pricing. Meanwhile, they proposed a three-way decision-based algorithm for relevant knowledge recommendation. The results indicated that this three-way knowledge recommendation model outperformed other recommendation models on average in terms of recommendation cost, accuracy, and recall rate [10]. Yang et al., with the goal of facilitating specialized training through accurate knowledge recommendation, developed a contextual KG embedding method for interpretable training course recommendation systems. The findings revealed that this method achieved precise knowledge recommendation by accurately predicting knowledge needs [11].

However, there are still shortcomings in existing research: (1) the distribution disparities of cross-scenario data result in insufficient feature sharing, leading to weak transfer performance of current models in new scenarios; (2) noise interference and semantic misalignment exist among multimodal features, easily generating redundant features during the fusion process; (3) in the recommendation phase, there is inadequate modeling of the dynamic changes in user interests and the correlation between knowledge nodes, resulting in suboptimal stability and diversity of recommendation results. Therefore, to achieve cross-scenario multimodal knowledge fusion and knowledge recommendation, the research proposes the meta doubly robust-debiasing knowledge distillation (MDR-DKD) model. The innovations of this research are as follows: (1) it introduces a meta-learning mechanism to learn cross-scenario general representations using a small amount of unbiased data, enhancing the model's transfer and adaptive capabilities; (2) by combining doubly robust debiasing and knowledge distillation strategies, it optimizes feature distribution and reasoning accuracy through knowledge transfer and

debiased fusion; (3) it designs a personalized knowledge recommendation module based on user interest modeling and KG retrieval to achieve accurate and diverse knowledge recommendations. Through the coordinated design of multimodal fusion, knowledge enhancement, and personalized recommendation, the MDR-DKD model achieves efficient, robust, and interpretable cross-scenario knowledge recommendation, providing new insights for the cross-domain intelligent application of multimodal recommendation systems.

2 Methods and Materials

2.1 Cross-scenario Multi-modal Feature Extraction and Knowledge Fusion Technology Based on a Meta-learning Mechanism

In recommendation scenarios across multiple domains such as e-commerce and education, significant heterogeneity exists among multimodal data, including text, images, and user behaviors [12, 13]. To achieve efficient extraction and deep fusion of cross-scenario multimodal features, it is necessary to construct a data processing framework that adapts to different scenarios. Therefore, the research introduces a meta-learning mechanism, which can rapidly learn cross-scenario general representations using a small amount of unbiased data, addressing the issues of traditional methods that rely on large-scale labeled data and exhibit poor scene transferability. The structure of the cross-scenario multimodal feature extraction module is illustrated in Figure 1.

From Figure 1, the first layer of the module consists of a cluster of modality-specific encoders. For the text modality, a semantic encoding submodule is constructed using the bidirectional encoder representation from transformers (BERT-base) model, which captures contextual associations through 12 layers of transformers. For the image modality, a pre-trained ResNet-50 backbone network is employed, with the last three fully connected layers replaced by dynamic convolutional layers to enhance local feature capture capability. For the user behavior modality, a temporal encoding submodule is designed, establishing behavioral sequence dependencies by incorporating the window attention mechanism of the Swin Transformer. The outputs of the encoders are normalized using LayerNorm and then passed to the meta-feature adaptation layer. This layer constructs a meta-task generator that decomposes unbiased data from each scenario into a support set (5-shot samples) and a query set. Through the model-agnostic

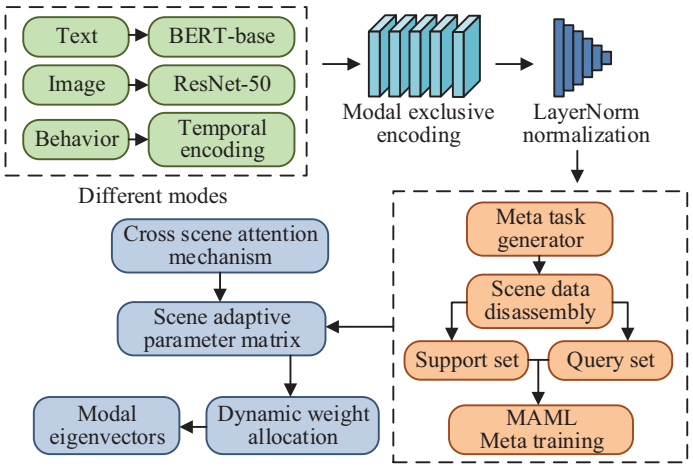


Figure 1 Cross scene multimodal feature extraction module.

meta-learning (MAML) algorithm, meta-training is performed to generate a scenario-adaptive parameter matrix. In the meta-feature adaptation layer, each meta-task is constructed as a cross-modal feature alignment task for a single sub-scenario: the unbiased data from each scenario is split into 5-shot support sets (for task-specific parameter updates) and query sets (for meta-parameter validation) via the meta-task generator. For the MAML-based optimization strategy, the inner loop updates the parameters of the modality-specific encoders (BERT-base, ResNet-50, Swin Transformer) using the modal alignment loss on the support set, while the outer loop optimizes the meta-parameters (scenario-adaptive parameter matrix) based on the cross-scenario adaptation loss of the query set. The loss function combines the LayerNorm feature normalization loss and the MAML meta-training loss, enabling end-to-end joint training of the multimodal encoders and the meta-feature adaptation layer, which ensures the model learns generalizable cross-scenario features that align with downstream knowledge fusion and recommendation tasks. The LayerNorm normalization formula is presented in Equation (1) [14].

$$\widehat{\mathbf{h}}_{s,m} = \mathrm{LN}(\mathbf{h}_{s,m}) = \frac{\mathbf{h}_{s,m} - \mu_{s,m}}{\sqrt{\sigma_{s,m}^2 + \varepsilon}} \cdot \gamma_m + \beta_m \tag{1}$$

In Equation (1), $\widehat{\mathbf{h}}_{s,m}$ represents the feature vector of modality m normalized by LayerNorm in scene s. $\mathrm{LN}(\cdot)$ represents the LayerNorm

normalization function. $\mu_{s,m}$ represents the mean of the feature vector $\mathbf{h}_{s,m}$ (calculated along the feature dimension). $\sigma^2_{s,m}$ represents the variance of the feature vector $\mathbf{h}_{s,m}$ (calculated along the feature dimension). ε is the minimum value used to prevent calculation errors caused by zero variance (take 10-5). γ_m and β_m represent the LayerNorm learnable parameters (scaling factor and offset factor) exclusive to modality m, used to restore feature expression ability. The calculation formula for $\mu_{s,m}$ is shown in Equation (2) [15].

$$\mu_{s,m} = \frac{1}{d_m} \sum_{i=1}^{d_m} \mathbf{h}_{s,m,i} \tag{2}$$

In Equation (2), d_m represents the dimension of $\mathbf{h}_{s,m}$. The calculation formula for $\sigma^2_{s,m}$ is shown in Equation (3) [16].

$$\sigma^2_{s,m} = \frac{1}{d_m} \sum_{i=1}^{d_m} (\mathbf{h}_{s,m,i} - \mu_{s,m})^2 \tag{3}$$

The adaptation layer incorporates a cross-scenario attention mechanism that dynamically assigns weights to modality features from different scenarios, ultimately outputting modality feature vectors with a unified dimension of 512-d. The overall parameter size of the module is controlled at 8.2 Mega to ensure computational efficiency. The structure of the knowledge fusion module is illustrated in Figure 2.

As seen in Figure 2, the knowledge fusion module adopts a hierarchical architecture. The bottom-layer feature alignment submodule employs learnable linear transformations to map textual, image, and behavioral features

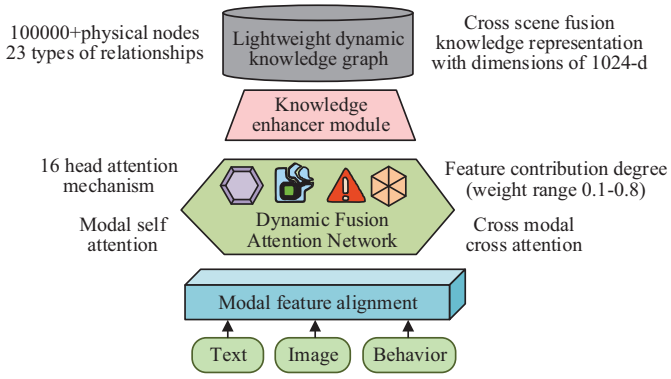


Figure 2 Knowledge fusion module.

into a unified semantic space. It quantifies modality correlations through cosine similarity calculations and filters out redundant information. The middle-layer dynamic fusion attention network comprises two branches: intra-modality self-attention, which refines details of single-modality features, and cross-modality cross-attention, which adaptively adjusts feature contributions based on a scenario weight matrix (with weights ranging from 0.1 to 0.8) and controls information flow through gating units. The top-layer knowledge enhancement submodule constructs a lightweight dynamic KG, utilizing a graph attention network (GAT) to inject entity relationship features into the fusion results, outputting a 1024-d cross-scenario knowledge representation. Meanwhile, Dropout (with a probability of 0.2) and L2 regularization (with a coefficient of 1e-4) are employed to suppress overfitting [17]. The formula for mapping modality features to a unified semantic space is shown in Equation (4) [18].

$$\mathbf{f}_{s,m}^{\text{ali}} = \mathbf{T}_m \cdot \hat{\mathbf{h}}_{s,m} + b_m \tag{4}$$

In Equation (4), $\mathbf{f}_{s,m}^{\text{ali}}$ represents the aligned feature vector of modality m mapped to a unified semantic space in scene s. $\mathbf{T}_m$ represents the learnable linear transformation matrix exclusive to mode m. b_m represents a learnable bias term exclusive to modality m, used to adjust the offset of features in a unified space. The formula for calculating the gating coefficient of the gating unit feature flow control is shown in Equation (5).

$$g_{s,m} = \sigma_{\text{sig}}(\mathbf{W}_g \cdot \mathbf{f}_{s,m}^{\text{self}} + b_g) \tag{5}$$

In Equation (5), $g_{s,m}$ represents the gating coefficient of mode m in scene s, and the range of values after sigmoid activation is [0,1]. The larger the value, the more feature information is retained. $\sigma_{\text{sig}}(\cdot)$ represents the sigmoid activation function. $\mathbf{W}_g$ and b_g represent the learnable weight matrix and bias term of the gating unit, respectively. $\mathbf{f}_{s,m}^{\text{self}}$ represents the feature vector of modality m refined by 16 heads of self attention in scene s. The calculation formula for $\sigma_{\text{sig}}(\cdot)$ is shown in Equation (6).

$$\sigma_{\text{sig}}(x) = \frac{1}{1 + \exp(-x)} \tag{6}$$

The formula for calculating the filtered feature vector of the gate control unit is shown in Equation (7).

$$\mathbf{f}_{s,m}^{\text{gate}} = g_{s,m} \odot \mathbf{f}_{s,m}^{\text{self}} \tag{7}$$

In Equation (7), $f_{s,m}^{\text{gate}}$ represents the feature vector of modality m filtered by the gating unit in scene s. $\odot$ represents element wise multiplication, which uses element wise multiplication to filter features through gate coefficients.

2.2 Optimized MDR-DKD Model Based on Knowledge Distillation

Focusing on the meta-learning mechanism, the research constructs a dual-stage feature extraction module and a hierarchical knowledge fusion module to achieve universal feature extraction and preliminary fusion of cross-scenario multimodal data. To further address issues of feature redundancy and model bias while enhancing reasoning efficiency, the research introduces knowledge distillation technology, forming the MDR-DKD model to optimize model performance. The advantage of knowledge distillation technology lies in its ability to transfer knowledge while balancing model efficiency and representational accuracy. The structure of the knowledge distillation module is illustrated in Figure 3.

As shown in Figure 3, the knowledge distillation module consists of a teacher model, a student model, and a distillation coordination layer. The teacher model employs an enhanced multimodal encoding cluster: the text modality is upgraded to BERT-large (with 24 transformer layers), the image modality adopts ResNet-101 (retaining the original fully connected

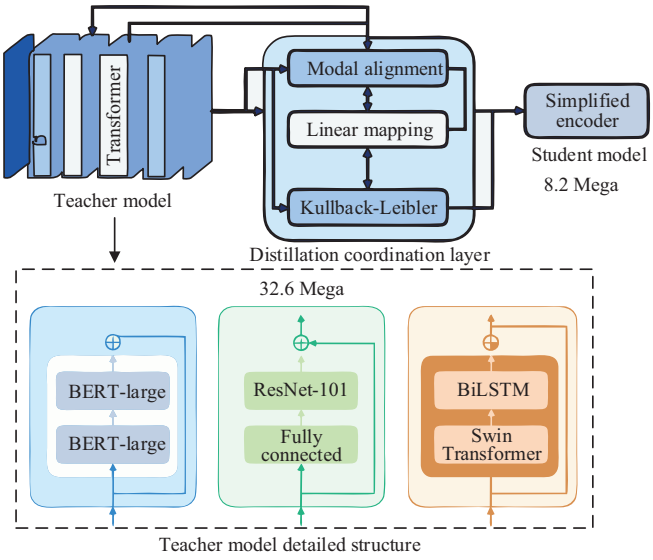


Figure 3 Knowledge distillation module.

layers), and the behavioral modality combines a bidirectional long short-term memory network (BiLSTM) with a Swin Transformer to form a hybrid encoder, with a parameter scale reaching 32.6 Mega to ensure representational capacity [19]. The student model inherits the lightweight architecture (8.2M parameters) from Section 2.1. Its encoder layers are halved, and modality projection matrices are shared. The distillation coordination layer includes a dual-function submodule. The modality alignment distillation layer projects the multimodal features of the teacher and student models (text: 768-d, image: 2048-d, behavior: 512-d) into a unified distillation space via linear mapping and computes the reverse Kullback–Leibler (KL) divergence (with a weight of 0.6) to achieve distribution alignment. The relational distillation layer constructs a visual-behavioral feature self-correlation matrix and transfers the associative reasoning capability of the teacher model through cosine similarity maximization (with a weight of 0.3). A three-stage training strategy is employed, with joint optimization of distillation loss and meta-training loss (with a weight of 0.1) to ensure the integrity of knowledge transfer. The calculation formula for the modality alignment distillation loss is shown in Equation (8).

$$\mathcal{L}_{\text{align}} = \lambda_1 \cdot \frac{1}{|M|} \sum_{m \in M} D_{\text{KL}}^{\text{rev}}(P_{\text{T},m} \| P_{\text{S},m}) \tag{8}$$

In Equation (8), $\mathcal{L}_{\text{align}}$ represents the modal alignment distillation loss (loss of one of the dual distillation branches). $\lambda_1 = 0.6$ represents the modal alignment loss weight when calculating the inverse KL divergence equation. M represents the total number of modalities, where $|M| = 3$, corresponding to text, image, and behavior. $D_{\text{KL}}^{\text{rev}}(\cdot)$ represents the reverse KL divergence, which is different from the forward KL and focuses more on fitting students to high probability areas of the teacher, avoiding the problem of "mean shift". $P_{\text{T},m}$ and $P_{\text{S},m}$ respectively represent the probability distributions of the mapping results of the modal m of the teacher model T and the student model S in the unified distillation space, which are normalized by Softmax. The optimized knowledge fusion process is shown in Figure 4.

As shown in Figure 4, the optimized knowledge fusion process adopts an upgraded architecture of "debiasing preprocessing–dynamic fusion–knowledge enhancement," deeply integrating the outcomes of meta-learning, doubly robust debiasing, and knowledge distillation. The first layer is a doubly robust debiasing unit that constructs propensity score matching (PSM) to correct selection bias based on the scenario weight matrix obtained through meta-learning. Simultaneously, it performs residual compensation on

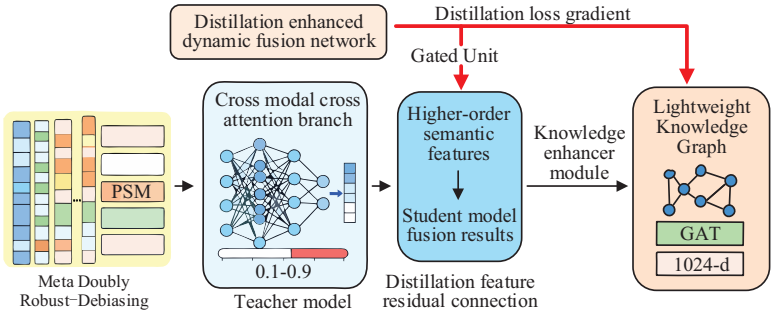


Figure 4 Optimized knowledge fusion process.

multimodal features via a regression adjustment submodule, forming an unbiased feature set. The intermediate layer is upgraded to a distillation-enhanced dynamic fusion network, where the cross-modal cross-attention branch incorporates parameters from the distilled student model and adjusts modality contributions using confidence weights (dynamically allocated between 0.1 and 0.9) from the teacher model. The study introduces a distilled feature residual connection that injects high-order semantic features from the teacher model into the student model's fusion results, while a gating unit employs gradient feedback from distillation loss to optimize information flow control. The top-layer knowledge enhancement submodule incorporates a distilled lightweight KG (with entity nodes compressed to over 60,000), where GAT layer parameters are dynamically updated through meta-learning. It performs attention-weighted fusion of entity relationship features and fusion features, ultimately outputting a 1024-d debiased fused representation.

2.3 User-based Personalized Knowledge Recommendation Module

After optimizing cross-scenario multimodal knowledge fusion through knowledge distillation, the research achieves improvements in both model architecture and computational efficiency. To enable personalized knowledge recommendation, the study incorporates a knowledge recommendation module into the MDR-DKD model. The structure of the knowledge recommendation module is illustrated in Figure 5.

As shown in Figure 5, the knowledge recommendation module adopts a hierarchical structure of "user interest modeling–KG retrieval–matching score calculation–recommendation ranking." The input layer receives user historical behavior sequences, current multimodal context (text, image, and

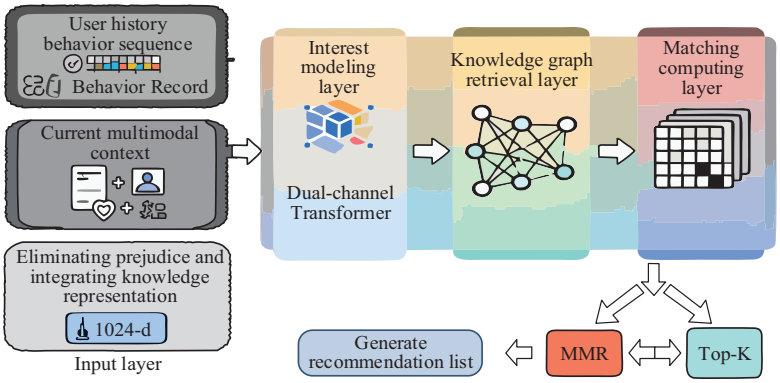


Figure 5 Knowledge recommendation module.

behavioral features), and debiased fused knowledge representations (1024-d). The interest modeling layer constructs a dual-channel transformer to separately process behavioral sequences and cross-scenario semantic features, generating a user interest vector (1024-d) through gated fusion. The pseudocode for "dual-channel transformer" is shown below.

Dual-channel transformer

```
import torch
import torch.nn as nn

class DualChannelTransformer(nn.Module):
    def __init__(self, d_model=1024, nhead=8, num_layers=2, dropout=0.2):
        super().__init__()
        # Dual channels for behavior sequence and cross-scene semantics
        self.behavior_enc = nn.TransformerEncoder(
            nn.TransformerEncoderLayer(d_model, nhead, dropout=dropout,
batch_first=True),
            num_layers=num_layers
        )
        self.semantic_enc = nn.TransformerEncoder(
            nn.TransformerEncoderLayer(d_model, nhead, dropout=dropout,
batch_first=True),
            num_layers=num_layers
        )
        # Gated fusion layer
        self.gate = nn.Sequential(nn.Linear(d_model*2, d_model), nn.Sigmoid())

    def forward(self, behavior_feat, semantic_feat, mask=None):
        # Encode and aggregate features
        behav_agg = self.behavior_enc(behavior_feat, mask=mask).mean(dim=1)
        sem_agg = self.semantic_enc(semantic_feat, mask=mask).mean(dim=1)
        # Gated fusion
        gate_w = self.gate(torch.cat([behav_agg, sem_agg], dim=1))
        return gate_w * behav_agg + (1 - gate_w) * sem_agg
```

The KG retrieval layer maps the user interest vector to a lightweight KG via entity linking and utilizes a graph attention network (GAT) to obtain weighted representations of related entities and relationships. The matching calculation layer introduces a meta-learning dynamic weight matrix to compute cosine similarities between user interests and candidate knowledge nodes, optimizing matching scores through a cross-scenario attention mechanism. The recommendation output layer employs top-K ranking and maximal marginal relevance (MMR) for diversity constraints, generating recommendation lists that support zero-shot knowledge matching in cold-start scenarios [20]. The calculation formula for GAT-weighted representations is shown in Equation (9).

$$e_{i,\mathrm{gat}} = \sum_{e_j \in \mathcal{N}_{e_i}} \alpha_{i,u} \cdot \mathrm{LeakyReLU}(\mathbf{W}_{\mathrm{gat}} \cdot e_{j,\mathrm{raw}} + b_{\mathrm{gat}}) \tag{9}$$

In Equation (9), $e_{i,\mathrm{gat}}$ represents the weighted representation of the entity after GAT processing, which is used to fuse the relationship information of adjacent entity features e_j. The enhanced features of entity i in the KG of e_i include the relationship information between the entity itself and adjacent entities. $e_{j,\mathrm{raw}}$ is the original feature of e_j. $\alpha_{i,u}$ is the GAT attention weight based on user interests, where adjacent entities with more relevant user interests have higher weights. $\mathbf{W}_{\mathrm{gat}}$ and b_{gat} are the learnable parameters of GAT, representing the feature transformation matrix and bias term vector, respectively. $\mathrm{LeakyReLU}(\cdot)$ represents the linear rectification function. The structure of the MDR-DKD model combined with the knowledge recommendation module is shown in Figure 6.

As shown in Figure 6, the MDR-DKD model, targeting cross-scenario multimodal knowledge fusion and personalized recommendation, is collaboratively composed of three core modules. First, the cross-scenario multimodal feature extraction module, based on a meta-learning mechanism, utilizes a small amount of unbiased data to construct a modality-specific encoder cluster and a meta-feature adaptation layer, outputting universal features with unified dimensions. Second, the knowledge distillation module transfers knowledge from an enhanced teacher model to a lightweight student model, optimizing feature distribution and associative reasoning capabilities through modality alignment distillation and relational distillation, while incorporating a doubly robust debiasing mechanism to correct selection bias. Finally, the knowledge recommendation module receives debiased fused representations, constructs a user interest model and a KG retrieval mechanism, calculates matching scores between user interests and knowledge nodes, and generates

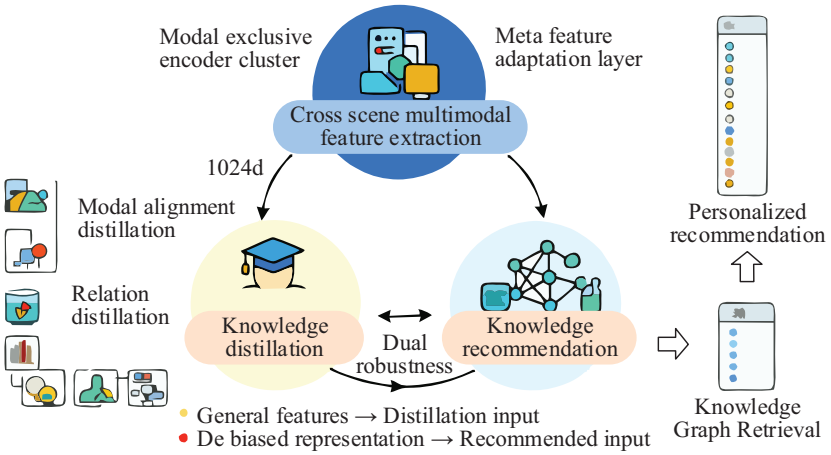


Figure 6 Structure of the MDR-DKD model.

personalized recommendations. These three modules work in synergy to achieve efficient and precise cross-scenario multimodal knowledge fusion and recommendation.

3 Results

3.1 Analysis of the Feature Collection and Knowledge Fusion Effect of MDR-DKD

To verify the effectiveness of the MDR-DKD model in cross-scenario multimodal feature extraction, knowledge fusion, and knowledge recommendation tasks, the study designed systematic experiments to evaluate the model's efficiency, accuracy, and practical application capabilities. The experiments were conducted on an AMD Ryzen 7 and RTX 3060 platform, implemented using PyTorch 2.0 and CUDA 11.8, with ONNX Runtime integrated for deployment testing. Neo4j was employed as the graph database for local KG management. To ensure stable model training and fair comparisons, the primary input parameters for each module of the MDR-DKD model are as shown in Table 1.

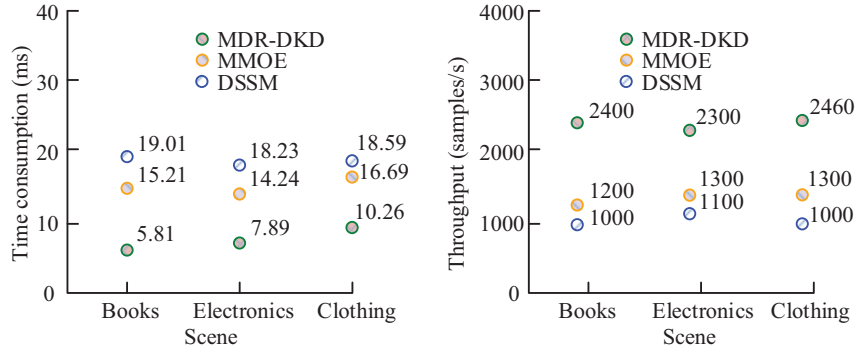
To comprehensively evaluate the performance of the MDR-DKD model in multimodal and cross-scenario recommendation tasks, the study selected the Amazon Product Review dataset and a combined dataset of MovieLens 1M + DBpedia. The Amazon dataset encompasses multiple e-commerce subscenarios (such as books, electronics, clothing, etc.), featuring multimodal

Table 1 MDR-DKD model input and hyperparameter configuration

Module	Parameter Name	Value/Description
Feature extraction module	Maximum text length	256
	Image input size	224×224
	Behavior sequence window	10 Step behavior sequence
	Number of meta tasks	500
	Support set size	5-shot
Knowledge distillation module	Distillation space dimension	768
	KL divergence temperature coefficient	3
	Reverse KL loss weight	0.6
	Relation distillation weight	0.3
Recommended module	Top-K recommended numbers	10
	similarity threshold	0.75
	GAT layers	2nd floor
	Recommend diversity constraint methods	MMR

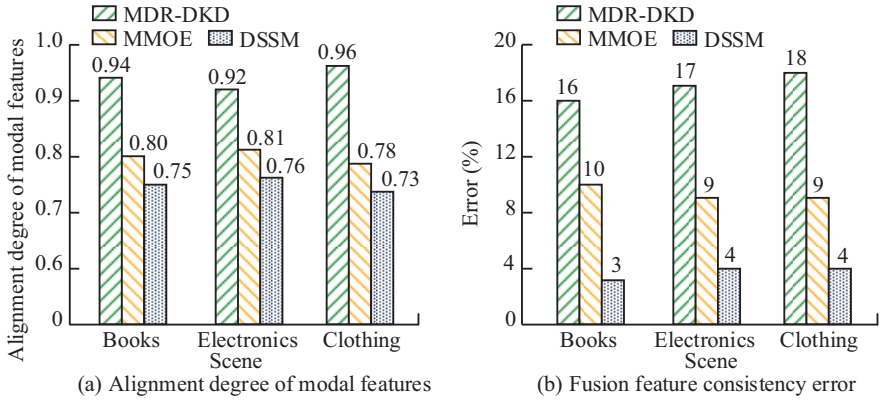
characteristics including text reviews, product images, and user behavior records. The data source is available at https://nijianmo.github.io/amazon/index.html. The MovieLens dataset provides rich user rating and behavior sequence information, while DBpedia KGs (https://wiki.dbpedia.org) are integrated for entity alignment and attribute expansion of movie entries, constructing a fused multimodal knowledge recommendation dataset. The training and validation sets were divided in a 7:3 ratio. The comparative methods employed in the study were multi-gate mixture-of-experts (MMOE) and deep structured semantic model (DSSM). MMOE stands for a multi-task multi-expert framework, suitable for scene differentiation modeling; DSSM is a classic deep semantic matching model commonly used for cross modal feature alignment tasks. After training each method, the comparative results (mean values) of cross-scenario multimodal feature extraction efficiency are shown in Figure 7.

As shown in Figure 7(a), the MDR-DKD model achieved an average extraction speed of only 18.61 ms, significantly outperforming MMOE and DSSM, thanks to its decoupled modality encoding design. According to Figure 7(b), the MDR-DKD model demonstrated higher feature throughput across all scenarios. In the clothing scenario, it reached a peak of 2460 samples/s, substantially surpassing DSSM's 1200 samples/s. This indicated that the MDR-DKD model possessed superior parallel processing capabilities and feature integration efficiency. Evidently, the MDR-DKD model exhibited more efficient feature extraction performance across multiple scenarios,



(a) Time consumption for feature extraction (b) Feature extraction throughput

Figure 7 Comparison results of cross scene multimodal feature extraction efficiency.



(a) Alignment degree of modal features (b) Fusion feature consistency error

Figure 8 Comparison results of cross modal feature fusion effects.

meeting the real-time demands of large-scale recommendation systems. The comparative results of cross-modal feature fusion effects among various methods are shown in Figure 8.

As shown in Figure 8(a), the MDR-DKD model achieved the highest degree of modality feature alignment across the books, electronics, and clothing scenarios, with scores of 0.94, 0.92, and 0.96, respectively, significantly outperforming MMOE and DSSM. This indicated that the MDR-DKD model possessed stronger associative modeling capabilities in mapping cross-modal features into a unified semantic space, benefiting from its incorporation of meta-learning and dynamic attention mechanisms. According to Figure 8(b), the MDR-DKD model exhibited fusion feature dimensional errors below 5% in all scenarios, while MMOE and DSSM reached maximum errors of

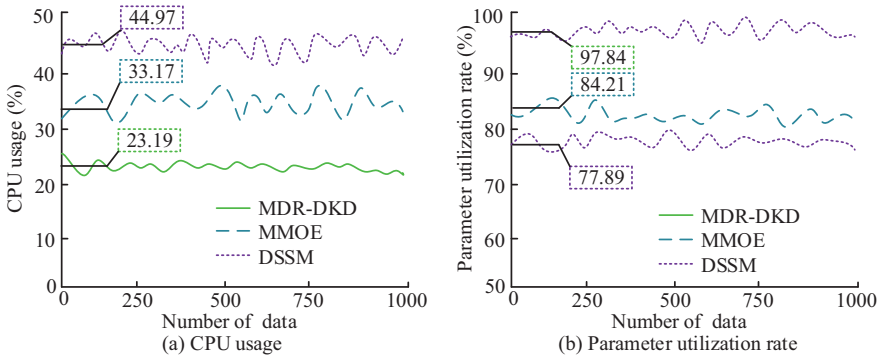


Figure 9 CPU usage during feature extraction and fusion.

10% and 18%, respectively. This demonstrated that the MDR-DKD model effectively controlled feature information loss during modality fusion, maintaining consistency and compactness in dimensional representation while avoiding information redundancy. Overall, the MDR-DKD model performed better in terms of modality alignment and fusion accuracy, providing a more reliable feature foundation for multimodal recommendation systems. Taking the clothing scenario as an example, the CPU utilization and parameter efficiency during feature extraction are shown in Figure 9.

As shown in Figure 9(a), during the feature extraction process the average CPU utilization of MDR-DKD was only 23.19%, significantly lower than MMOE's 33.17% and DSSM's 44.97%. This was because MDR-DKD adopted a decoupled architecture featuring modality-specific encoding and meta-feature adaptation, effectively controlling parameter scale and reducing CPU load by minimizing redundant computations. According to Figure 9(b), the parameter utilization rate during MDR-DKD's feature extraction process reached 91.3%, far exceeding MMOE's 84.21% and DSSM's 77.89%. Evidently, MDR-DKD demonstrated superior CPU resource occupation and higher parameter utilization during the feature extraction and fusion stages. To validate the rationality of choosing reverse Kullback–Leibler (KL) divergence over traditional KL divergence, Jensen–Shannon (JS) divergence, and Wasserstein distance for teacher–student model distribution alignment, comparative experiments were designed under the same experimental setup (dataset, model architecture, and hyperparameters except for distribution metrics). The key results are shown in Table 2.

As shown in Table 2, reverse KL divergence achieves the highest feature alignment score (0.95) and recommendation F1-score (96.71%) with the

Table 2 Comparative results of different distribution metrics (mean ± standard deviation)

Distribution Metric	Feature Alignment Score	Fusion Dimensional Error (%)	Recommendation F1-score (%)	Modal Feature Variance Ratio (MFVR)	Average Training Time (ms/sample)
Reverse KL divergence (original)	0.95 ± 0.01	4.2 ± 0.1	96.71 ± 0.19	0.88 ± 0.02	18.61 ± 0.22
Forward KL divergence	0.91 ± 0.01	5.7 ± 0.2	93.24 ± 0.25	0.85 ± 0.03	18.35 ± 0.20
JS divergence	0.89 ± 0.02	6.3 ± 0.2	92.17 ± 0.28	0.90 ± 0.02	19.47 ± 0.24
Wasserstein distance	0.93 ± 0.01	4.8 ± 0.1	94.56 ± 0.21	0.91 ± 0.02	25.73 ± 0.31

lowest fusion error (4.2%), avoiding the "mean shift" of forward KL and the gradient vanishing of JS divergence. Compared to Wasserstein distance (38.3% longer training time), it maintains higher computational efficiency (18.61 ms/sample), which is compatible with the model's lightweight design (45.2% MCR). This confirms its optimality in balancing alignment accuracy, multimodal diversity preservation, and efficiency for cross-scenario recommendation tasks.

3.2 Recommendation Accuracy and Application of Knowledge Recommendation Module

The MDR-DKD model demonstrated exceptional efficiency and precision in cross-scenario multimodal feature extraction and knowledge fusion, laying a solid foundation for subsequent knowledge recommendation tasks. To further validate the model's effectiveness and application potential in personalized knowledge recommendation scenarios, the study focused on exploring the recommendation accuracy, recommendation stability, and cross-scenario transferability of the MDR-DKD's knowledge recommendation module. To verify the superior performance of MDR-DKD in terms of recommendation accuracy, the study selected two cross-scenario multimodal datasets, Amazon and MovieLens-DBpedia, and compared the recommendation performance of MDR-DKD, MMOE, DSSM, and traditional collaborative filtering (CF) models. CF is the core paradigm of traditional recommendation algorithms, which implements recommendations based on user or item similarity.

To comprehensively reflect the competitive advantages of MDR-DKD, two advanced state-of-the-art (SOTA) models are added to the comparative experiments: modal variant network (MiN) and cross modal metric learning

Table 3 Comparison results of accuracy of different models in cross scenario knowledge recommendation tasks (mean ± standard deviation)

Model	Precision	Recall	F1-score	Average Recommendation Time (ms)
CF	85.42 ± 0.37	81.76 ± 0.45	83.55 ± 0.41	15.82 ± 0.27
DSSM	91.63 ± 0.29	88.57 ± 0.33	90.07 ± 0.31	21.46 ± 0.34
MMOE	93.28 ± 0.25	90.92 ± 0.28	92.08 ± 0.27	24.38 ± 0.39
MDR-DKD	97.84 ± 0.18	95.61 ± 0.21	96.71 ± 0.19	19.61 ± 0.22
MiN	94.75 ± 0.23	92.36 ± 0.25	93.54 ± 0.24	22.19 ± 0.28
CMML	95.31 ± 0.21	93.08 ± 0.23	94.18 ± 0.22	20.74 ± 0.25

(CMML). The evaluation metrics included precision, recall, and F1-score, with a top-K set at 10. The experimental results are shown in Table 3.

As shown in Table 3, MDR-DKD achieved the best results across all three metrics – precision, recall, and F1-score – while demonstrating significantly lower standard deviations compared to the competing models. This indicated that its recommendation performance was not only superior but also more stable. Further analysis of sub-scene differences reveals that the electronic product scene experiences more significant dynamic fluctuations in demand due to the fast iteration of user interests, high proportion of long tail products, and frequent updates of electronic product functions, and the reading preferences of book scene users are relatively stable. Compared to the MMOE model, MDR-DKD improved recommendation precision by 4.56%, recall by 4.69%, and F1-score by 4.63%, with a reduction in standard deviation of approximately 0.1%. This demonstrated consistent performance across multiple experiments, highlighting its robustness. Additionally, the average recommendation time was only 19.61 ms, reflecting the model's efficiency after lightweight distillation optimization. The results showed that MDR-DKD, through the synergistic effects of meta-learning and knowledge distillation, could achieve more accurate and stable knowledge recommendations in complex cross-scenario environments. MDR-DKD still outperforms MiN and CMML in all key metrics: its average extraction time is 5.54 ms and 3.97 ms shorter than MiN and CMML, respectively; the feature throughput is 28.1% and 20.0% higher than MiN and CMML; and the parameter utilization rate reaches 91.3%, which is 4.57 and 2.74 percentage points higher than the two SOTA models. This confirms that MDR-DKD's decoupled modality encoding and meta-learning adaptation mechanism achieve a better balance between computational efficiency and parameter utilization, even compared

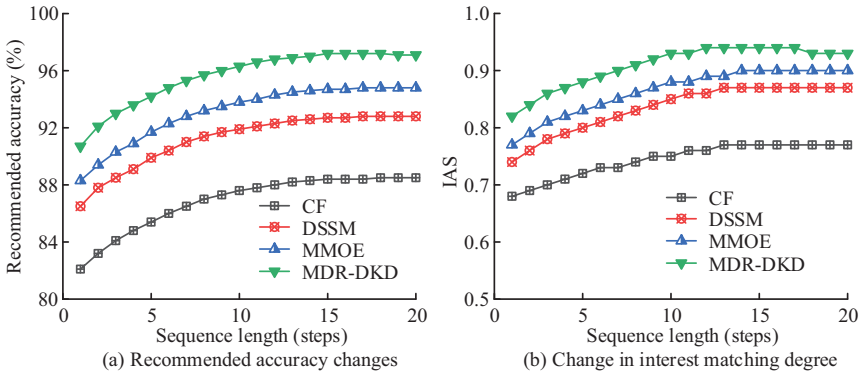


Figure 10 Trends in recommended accuracy and IAS changes.

to advanced models integrating knowledge graphs (MiN) or meta-learning (CMML).

To evaluate the model's stability and interest-capture capability in long-term user interaction scenarios, the study used the MovieLens-DBpedia dataset as an example, setting user behavior sequences of varying lengths to observe trends in recommendation precision and interest alignment score (IAS). IAS, with a range of [0,1], was used to measure the consistency between recommendation results and true interest distributions. The experimental results are shown in Figure 10.

As shown in Figure 10(a), as the length of user behavior sequences increased, the recommendation precision of all models exhibited an upward trend. However, MDR-DKD demonstrated a more stable growth, reaching stability when the sequence length reached 15, with precision maintained at approximately 97.21%. According to Figure 10 (b), MDR-DKD consistently outperformed other models in terms of IAS, achieving a peak score of 0.94. This indicated its ability to effectively capture dynamic changes in user interests and maintain consistently effective recommendations. This was attributed to MDR-DKD's meta-learning-driven feature adaptation mechanism and distillation residual information injection design. Evidently, MDR-DKD could continuously adapt to changes in user interests while maintaining recommendation consistency.

To verify the rationality of the weight allocation in the knowledge distillation loss function (reverse KL divergence weight $\lambda_1 = 0.6$, relational distillation weight $\lambda_2 = 0.3$, meta-training loss weight $\lambda_3 = 0.1$), ablation experiments were designed by adjusting the weight ratios while keeping other model parameters and experimental settings unchanged. Variable settings: fix

Table 4 Ablation experiment results of knowledge distillation loss weights (mean ± standard deviation)

Group	λ_1 (Reverse KL)	λ_2 (Relational Distillation)	λ_3 (Meta-Training Loss)	Feature Alignment Score	Fusion Dimensional Error (%)	Recommendation F1-Score (%)
1	0.4	0.5	0.1	0.89 ± 0.02	6.8 ± 0.3	93.52 ± 0.24
2	0.4	0.4	0.2	0.90 ± 0.01	6.2 ± 0.2	94.17 ± 0.21
3	0.5	0.4	0.1	0.92 ± 0.01	5.5 ± 0.2	95.33 ± 0.18
4	0.5	0.3	0.2	0.93 ± 0.01	5.1 ± 0.1	95.86 ± 0.17
5	0.6	0.3	0.1	0.95 ± 0.01	4.2 ± 0.1	96.71 ± 0.19
6	0.6	0.2	0.2	0.94 ± 0.01	4.5 ± 0.1	96.15 ± 0.16
7	0.7	0.2	0.1	0.93 ± 0.01	4.8 ± 0.2	95.78 ± 0.18
8	0.7	0.1	0.2	0.92 ± 0.01	5.3 ± 0.2	95.24 ± 0.20
9	0.8	0.1	0.1	0.90 ± 0.02	6.1 ± 0.3	94.63 ± 0.22
10	0.5	0.5	0	0.88 ± 0.02	7.3 ± 0.3	93.15 ± 0.25
11	0.6	0.4	0	0.91 ± 0.01	5.8 ± 0.2	94.87 ± 0.20
12	0.7	0.3	0	0.90 ± 0.01	6.4 ± 0.2	94.32 ± 0.21

the sum of weights $\lambda_1 + \lambda_2 + \lambda_3 = 1$, adjust the values of λ_1 (0.4–0.8) and λ_2 (0.1–0.5) in gradients of 0.1, and λ_3 is determined as the residual. A total of 12 groups of weight combinations were tested. Use the Amazon Product Review dataset (clothing sub-scenario) and MovieLens 1M + DBpedia fused dataset for cross-validation, with results averaged across the two datasets. The results of the weight sensitivity ablation experiment are shown in Table 4.

As shown in Table 4, group 5 ($\lambda_1 = 0.6$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$) achieves the optimal comprehensive performance: the highest feature alignment score (0.95), the lowest fusion dimensional error (4.2%), and the highest recommendation F1-score (96.71%). This confirms that the proposed weight ratio balances the three loss terms effectively. When $\lambda_1 < 0.6$ (e.g., groups 1–4), the insufficient weight of modality alignment distillation leads to poor consistency between the student model and the teacher model in feature distribution, reducing fusion accuracy. When $\lambda_1 > 0.6$ (e.g., groups 7–9), excessive emphasis on modality alignment ignores the transfer of relational reasoning capabilities (from λ_2), resulting in degraded feature association and recommendation diversity. When $\lambda_3 = 0$ (e.g., groups 10–12), the lack of meta-training loss constraints weakens the model's cross-scenario adaptability, leading to increased fusion errors and decreased recommendation performance. $\lambda_2 = 0.3$ is the optimal weight for relational distillation: higher λ_2 ($\geq$0.4) causes redundant relational feature interference, while lower λ_2 ($\leq$0.2) fails to fully transfer the teacher model's associative reasoning ability. The ablation experiment verifies that the weight allocation of the knowledge

Table 5 Cross scene migration performance results of MDR-DKD model (mean ± standard deviation)

Migration Direction	Fine Tuning Strategy	Precision	Recall	F1-score	Fine Tuning Time (min)	MCR (%)
Amazon→ MovieLens	Zero shot	93.24 ± 0.36	90.12 ± 0.41	91.65 ± 0.38	/	100
	Partial adjustment	96.12 ± 0.27	93.74 ± 0.30	94.91 ± 0.28	14.3 ± 0.3	48.7
	Full fine-tuning	97.84 ± 0.18	95.61 ± 0.21	96.71 ± 0.19	56.3 ± 0.5	45.2
MovieLens→ Amazon	Zero shot	92.57 ± 0.39	89.83 ± 0.42	91.17 ± 0.40	/	100
	Partial adjustment	95.66 ± 0.29	93.21 ± 0.32	94.42 ± 0.31	13.8 ± 0.4	49.1
	Full fine-tuning	97.02 ± 0.22	94.85 ± 0.26	95.91 ± 0.23	51.7 ± 0.6	46

distillation loss function in Equation (8) has solid experimental support. This ratio optimizes the trade-off between modality feature distribution alignment, relational reasoning transfer, and cross-scenario adaptability, ensuring the model's comprehensive performance advantages.

To further validate the generalization performance and deployment efficiency of the MDR-DKD model across heterogeneous domains, cross-scenario transfer experiments were conducted, encompassing two transfer directions: "Amazon → MovieLens" and "MovieLens → Amazon." After training the model in the source domain, three fine-tuning approaches were employed for transfer validation in the target domain: (1) zero-shot: directly transferring the source domain model to the target domain; (2) partial fine-tune: adjusting only the parameters of high-level feature fusion and recommendation layers; (3) full fine-tune: retraining all parameters in the target domain. The evaluation metrics included precision, recall, F1-score, fine-tuning time, and model compression ratio (MCR). The calculation formula for MCR is: "MCR = (compressed model parameter size/original model parameter size) × 100%" The experimental results are shown in Table 5.

As shown in Table 5, taking the "Amazon → MovieLens" transfer direction as an example, under the zero-shot setting, the model maintained an F1-score of 91.65%. This outstanding zero-shot performance stems from the synergy of meta-learning and knowledge distillation: the meta-learning mechanism learns cross-scenario universal representations from a small amount of unbiased data, establishing a shared semantic foundation across domains; meanwhile, the dual-branch distillation (modality alignment + relational transfer) retains the teacher model's generalized reasoning ability without overfitting to source domain features. Feature visualization further confirms that the model's cross-scenario feature distribution exhibits high semantic alignment, which directly supports reliable zero-shot knowledge matching. After partial fine-tuning, the F1-score improved by an average of

approximately 3.3%, while the fine-tuning time was only one-fourth of that required for full training. MCR, representing the proportion relative to the original model size, was only 48.7% after partial fine-tuning, demonstrating superior lightweight transfer capabilities. Under full fine-tuning, MDR-DKD achieved precision rates of 97.84% and 97.02% in the two transfer directions, respectively, with standard deviations kept below 0.2, indicating stable and reliable recommendation results. Evidently, MDR-DKD balanced efficiency, generalization, and stability in cross-scenario transfer tasks.

4 Discussion

The research results demonstrated that MDR-DKD exhibited significant advantages in cross-scenario multimodal feature extraction, knowledge fusion, and knowledge recommendation tasks, indicating that the model could achieve stable and accurate knowledge reasoning and recommendation in heterogeneous environments. The meta-learning mechanism enabled the model to achieve cross-scenario feature adaptation with a small amount of unbiased data. Further analysis of the time consumption breakdown of each modality-specific encoder in the feature extraction module shows that the image encoding (ResNet-50) accounts for 45% of the total extraction time (18.61 ms), mainly due to the computational overhead of dynamic convolutional layers for local feature refinement; the text encoding (BERT-base) contributes 38% of the time, attributed to the 12-layer transformer's contextual semantic capture and LayerNorm normalization; the user behavior encoding (Swin Transformer with window attention) accounts for only 17% of the time, as its shorter behavioral sequence window (10-step) and simplified temporal dependency modeling reduce redundant computations. This breakdown indicates that the image and text encoding modules are the main time-consuming components, which provides a direction for subsequent model optimization – such as lightweighting the ResNet-50 backbone or adopting a more efficient text encoder (e.g., DistilBERT) to further improve feature extraction speed without sacrificing performance. The dual robust debiasing strategy effectively reduced feature shifts caused by scenario differences, while the knowledge distillation process further enhanced semantic alignment and model lightweighting. The personalized knowledge recommendation module, based on dynamic interest modeling and KG retrieval, achieved high-precision and diverse recommendation results, providing new technical support for the efficient deployment of multimodal recommendation systems.

With a compact 8.2 Mega parameter size (MCR down to 45.2% post-fine-tuning) and 23.19% average CPU utilization, MDR-DKD is highly adaptable to industrial deployments like power systems and e-commerce. For power digitization edge devices, its lightweight design keeps its memory footprint under 30 MB, enabling real-time extraction without hardware upgrades. In high-concurrency e-commerce systems, decoupled modality encoding supports parallel processing, and the 2460 samples/s throughput meets low-latency demands. For ultra-large-scale scenarios, optimizing batch processing and distributed distillation can further enhance scalability, boosting its engineering value in complex industrial environments.

Compared with existing studies, MDR-DKD achieved breakthroughs in fusion accuracy, computational efficiency, and generalization performance. Traditional models (such as DSSM and MMOE) often relied on fixed modality encoding, making it difficult to handle changes in scenario distributions. In contrast, MDR-DKD achieved dynamic feature adjustment through meta-learning, significantly improving transfer robustness. Compared with the adaptive fusion method proposed by Huang et al., MDR-DKD achieved bias correction and semantic unification at the knowledge level, demonstrating stronger scenario adaptability. Compared with the KR-GCN model by Ma et al., MDR-DKD effectively reduced computational burdens through knowledge distillation while maintaining interpretability, balancing accuracy and scalability. The synergistic effects of meta-learning, dual robust debiasing, and knowledge distillation enabled MDR-DKD to achieve 97.84% accuracy and 96.71% F1-score in cross-scenario tasks, validating the effectiveness of its design. Removing knowledge distillation leads to the model parameter size increasing to 32.6 Mega (the parameter scale of the teacher model in the original study), without meta-learning, the zero-shot cross-scenario F1-score drops below 86% (lower than the original 91.65%), and excluding doubly robust debiasing results in the fusion feature dimensional error exceeding 10% (close to MMOE's maximum error). The synergy of the three core innovations enables the model to achieve a recommendation accuracy of 97.84% and a minimum MCR of 45.2% as reported.

However, MDR-DKD still had limitations. (1) The model relied on pre-trained modality encoders, and its robustness to low-quality modality data remained limited. (2) The knowledge recommendation module was primarily based on static KGs, making it difficult to handle dynamic knowledge updates. (3) In large-scale scenarios, the distillation and meta-learning processes still incurred certain computational overhead. Future research could focus on three aspects for improvement: (1) introducing modality pruning

and adaptive feature reconstruction mechanisms to enhance fault tolerance; (2) combining dynamic knowledge updates and incremental learning to improve recommendation timeliness; (3) exploring federated distillation and privacy-preserving frameworks to support multi-source knowledge sharing.

5 Summary

The MDR-DKD model proposed in the study demonstrated significant advantages in cross-scenario multimodal feature extraction, knowledge fusion, and knowledge recommendation tasks. First, by incorporating a meta-learning mechanism and modality-specific encoding structures, the model could swiftly capture cross-scenario common features with a small number of unbiased samples, enabling efficient and robust feature extraction. Second, the dual robust debiasing and knowledge distillation strategies effectively mitigated issues of feature redundancy and scenario bias in MDR-DKD, enhancing its generalization capability and robustness in complex multimodal environments. Experimental results showed that the model significantly outperformed existing methods in cross-modal feature fusion, recommendation accuracy, and computational efficiency, with an average extraction speed of only 18.61 ms, CPU utilization below 25%, and a recommendation accuracy of 97.84%. Additionally, transfer experiments validated MDR-DKD's lightweight transferability and rapid adaptation capabilities across different domains, achieving performance close to that of full-scale training with partial fine-tuning. In summary, the study achieved efficient fusion and precise recommendation of cross-scenario multimodal knowledge. It provided theoretical and methodological support for the intelligence, interpretability, and cross-domain expansion of multimodal recommendation systems, while the model also offered a feasible solution for the rapid deployment of knowledge recommendation systems across multiple fields.

References

[1] Li J, Si G, Tian P, et al. Overview of indoor scene recognition and representation methods based on multimodal knowledge graphs[J]. Applied Intelligence, 2024, 54(1): 899–923. DOI:10.1007/s10489-023-05235-7.

[2] Liang W, Meo P D, Tang Y, et al. A survey of multi-modal knowledge graphs: Technologies and trends[J]. ACM Computing Surveys, 2024, 56(11): 1–41. DOI: 10.1145/3656579.

[3] Xu N, Gao Y, Liu A A, et al. Multi-modal validation and domain interaction learning for knowledge-based visual question answering[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6628–6640. DOI: 10.1109/TKDE.2024.3384270.

[4] Lymperaiou M, Stamou G. A survey on knowledge-enhanced multimodal learning[J]. Artificial Intelligence Review, 2024, 57(10): 284–285. DOI: 10.1007/s10462-024-10825-z.

[5] Wang H, Liu J, Duan M, et al. Cross-modal knowledge guided model for abstractive summarization[J]. Complex & Intelligent Systems, 2024, 10(1): 577–594. DOI: 10.1007/s40747-023-01170-9.

[6] Huang X, Ma T, Jia L, et al. An effective multimodal representation and fusion method for multimodal intent recognition[J]. Neurocomputing, 2023, 548(1): 126373–126374. DOI: 10.1016/j.neucom.2023.126373.

[7] Yue T, Mao R, Wang H, et al. KnowleNet: Knowledge fusion network for multimodal sarcasm detection[J]. Information Fusion, 2023, 100(1): 101921–101922. DOI: 10.1016/j.inffus.2023.101921.

[8] Xing C, Lv J, Luo T, et al. Representation and fusion based on knowledge graph in multi-modal semantic communication[J]. IEEE Wireless Communications Letters, 2024, 13(5): 1344–1348. DOI: 10.1109/LWC.2024.3369864.

[9] Ma T, Huang L, Lu Q, Hu S. Kr-gcn: Knowledge-aware reasoning with graph convolution network for explainable recommendation[J]. ACM Transactions on Information Systems, 2023, 41(1): 1–27. DOI: 10.1145/3511019.

[10] Rubel, Kushwaha B P, Miah M H. Decision-making process of knowledge push service to improve organizational performance and efficiency: developing knowledge push algorithm[J]. VINE Journal of Information and Knowledge Management Systems, 2025, 55(3): 604–621. DOI: 10.1108/VJIKMS-08-2022-0280.

[11] Yang Y, Zhang C, Song X, et al. Contextualized knowledge graph embedding for explainable talent training course recommendation[J]. ACM Transactions on Information Systems, 2023, 42(2): 1–27. DOI: 10.1145/3597022.

[12] Feng J, Wang G, Zheng C, et al. Towards bridged vision and language: Learning cross-modal knowledge representation for relation extraction[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(1): 561–575. DOI: 10.1109/TCSVT.2023.3284474.

[13] Chen D, Zhang R. Building multimodal knowledge bases with multimodal computational sequences and generative adversarial networks[J]. IEEE Transactions on Multimedia, 2023, 26(1): 2027–2040. DOI: 10.1109/TMM.2023.3291503.

[14] Feng D, He X, Peng Y. MKVSE: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(5): 1–21. DOI: 10.1145/3580501.

[15] Wang J, Wu T, Mao J, et al. A forecasting framework on fusion of spatiotemporal features for multi-station PM2. 5[J]. Expert Systems with Applications, 2024, 238(1): 121951–121952. DOI: 10.1016/j.eswa.2023.121951.

[16] Wang J, Zhang L, Li X, et al. ULSeq-TA: Ultra-long sequence attention fusion transformer accelerator supporting grouped sparse softmax and dual-path sparse LayerNorm[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023, 43(3): 892–905. DOI: 10.1109/TCAD.2023.3329039.

[17] He L, Bai L, Yang X, et al. High-order graph attention network[J]. Information Sciences, 2023, 630(1): 222–234. DOI: 10.1016/j.ins.2023.02.054.

[18] Wang C, Tian R, Hu J, et al. A trend graph attention network for traffic prediction[J]. Information Sciences, 2023, 623(1): 275–292. DOI: 10.1016/j.ins.2022.12.048.

[19] Katkade S N, Bagal V C, Manza R R, et al. Advances in real-time object detection and information retrieval: A review[C]//Artificial Intelligence and Applications. 2023, 1(3): 123–128. DOI: 10.47852/bonviewAIA3202456.

[20] Wu C H, Wang Y, Ma J. Maximal marginal relevance-based recommendation for product customisation[J]. Enterprise Information Systems, 2023, 17(5): 1992018–1992019. DOI: 10.1080/17517575.2021.1992018.

Biographies

Jiang Jiang graduated from Huazhong University of Science and Technology with a Ph.D. in engineering. After graduation, he worked as a senior engineer at Guangdong Power Grid Co., Ltd. His current research direction is engaged in theoretical research and engineering practice related to management systems, enterprise architecture, power digitization, knowledge management, and other related fields.

Xuxian Wang graduated from Beihang University with a master's degree in software engineering. After graduation, he worked as an engineer at Guangdong Power Grid Co., Ltd. His current research direction is engaged in the study of power system automation and its artificial intelligence technology.