Proceedings of CONF-MPCS 2025 Symposium: Mastering Optimization: Strategies for Maximum Efficiency
DOI: 10.54254/2753-8818/2025.CH22588

Deep Learning-Based Dual-Tower Contrastive Recommendation System

Ye Jin

Institute of Beijing University of Posts and Telecommunications, Beijing, China
1343799958@qq.com

Abstract: This study proposes a Deep Collaborative Dual-Tower Model (DCDT) to address the issue of insufficient feature interaction in multi-modal recommendation systems for educational scenarios. Through three key innovations, DCDT achieves precise recommendation of teaching resources: First, we construct a dual-path heterogeneous encoder that uses orthogonal constraints to decouple the user query and video content feature processing, and apply cross-modal attention mechanisms for semantic alignment. Second, a dynamic hard negative sampling strategy is designed, building high-quality training pairs based on the top-30% similarity threshold. Third, a hybrid-interaction layer is developed to enhance multi-modal correlations through feature concatenation and dot product operations. On our self-built linear algebra teaching video dataset, DCDT achieves 82.1% accuracy and 0.641 MRR, improving the F1 score by 19.8% compared to traditional collaborative filtering methods, with a response time of under 10ms. Ablation experiments show that the dual-tower architecture, with a feature pre-computation mechanism, improves recall efficiency by 8.6%. This study is the first to apply a contrastive learning framework to educational multi-modal recommendation systems, solving the conflict between heterogeneous feature fusion and real-time responses, and offering a new paradigm for the development of intelligent education systems. Future work will integrate knowledge graphs to enhance content understanding and explore federated learning frameworks to ensure data privacy.

Keywords: dual-path, cross-modal, dynamic, hybrid-interaction, contrastive

1. Introduction

With the rapid development of artificial intelligence, large-scale pre-trained models have shown significant success, particularly in educational technology [1]. Traditional educational systems often rely on static textbooks and fixed teaching processes, which fail to address individual learner needs [2]. In response, personalized educational platforms using Natural Language Processing (NLP) and deep learning have emerged, such as intelligent teaching systems based on generative models like GPT, which can automatically generate code, explain complex concepts, and enhance learning through dynamic demonstrations [3].

The “ZhaoXi” project, a large-model-based educational application, aims to generate code and produce video animations from natural language input. Its core structure includes four steps: concept interpretation, example analysis, animated demonstration, and resource recommendation. Key challenges include selecting appropriate examples from a vast database and recommending relevant video content from online resources [4].

DOI: 10.54254/2753-8818/2025.CH22588

Traditional recommendation systems often rely on user behavior or content similarity but fail to integrate multimodal information such as text, video, and images [5]. In education, where diverse content and teaching methods are involved, a single recommendation approach is inadequate. Deep learning, particularly multimodal learning, offers a solution by processing multiple modalities, enabling more precise and personalized recommendations [6].

This paper proposes a multimodal recommendation system for the "ZhaoXi" project. It integrates user queries, video titles, and descriptions into embedding representations, then refines recommendations using a contrastive loss function. The system first generates embeddings with pre-trained models, followed by feature fusion and matching through a deep neural network, ultimately providing the most relevant recommendations [7,8].

2. Fundamental principles

This study uses the OpenAI text-embedding-3-small model to convert video titles, descriptions, and user inputs into 1536-dimensional embedding vectors for building the recommendation system.

The dual-tower model serves as the core architecture in the recall phase of recommendation systems, with its design concept originating from the DSSM model proposed by Microsoft Research [9].

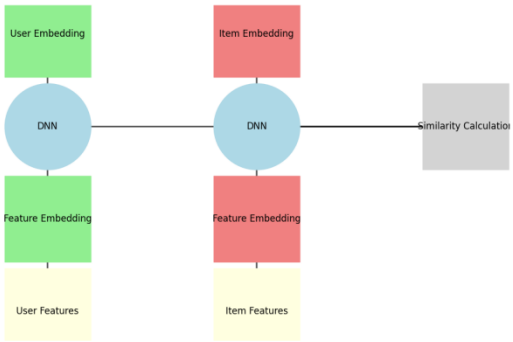

Figure 1: Dual-tower model architecture

As shown in Figure 1, the model consists of symmetric User Tower and Item Tower, which process user-side features (user profile, behavioral sequence, context information) and item-side features (item attributes, content descriptions, statistical metrics), respectively. The two towers perform nonlinear mapping through independent multi-layer perceptrons (MLP) [4].

$$u = f_\theta(X_{user}) = MLP([e_1^{(u)}, \ldots, e_m^{(u)}]) \quad (1)$$

$$u = g_\phi(Y_{item}) = MLP([e_1^{(v)}, \ldots, e_n^{(v)}]) \quad (2)$$

Where $e_i^{(u)}$ and $e_j^{(v)}$ represent the feature embedding vectors of the user-side and item-side, respectively, and θ and ϕ are the network parameters. The model measures semantic similarity through cosine similarity [10].

The core advantage of the dual-tower model lies in its decoupled computation, as shown in Table 1. This architecture completely isolates the processing of user-side and item-side features, enabling the precomputation of item embeddings for massive catalogs, which significantly reduces the online service latency. However, this also leads to the issue of late-stage feature interaction—user and item

DOI: 10.54254/2753-8818/2025.CH22588

features only undergo dot product operations at the top-level embeddings, lacking the ability to learn fine-grained cross-feature interactions. Research has shown that this can reduce the recall accuracy for long-tail items [11].

The dual-tower model's two encoders process the interaction features of $q$and (t,d) respectively, and introduce a cross-modal attention mechanism:

$$\alpha = soft\,max(W_q q \cdot (W_t t + W_d d)) \tag{3}$$

Table 1: Analysis of dual-tower model characteristics

Dimension	Advantages	Limitations
Computational Efficiency	Supports ANN indexing, response time < 10ms	Negative sampling bias affects convergence stability
Feature Processing	Compatible with multi-modal feature inputs	Cannot explicitly model cross-feature interactions
System Scalability	Supports distributed deployment and incremental updates	User interest drift tracking delay

3. Deep collaborative dual tower model

In the recommendation system designed in this study, the core issue is how to provide personalized recommendations for users based on their queries and multimodal data (such as video titles, descriptions, tags, etc.). Specifically, as shown in Figure 2, this study aims to provide precise video recommendations by using deep learning methods, combining the embedding representations of video content (title, description) and user queries.

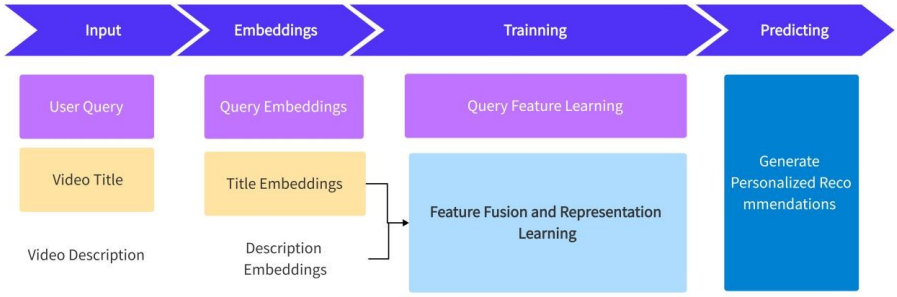

Figure 2: Deep learning-based video recommendation pipeline

Query tower: responsible for mapping the input query q to an embedding vector q_em. This tower includes a linear transformation layer, an activation function (ReLU), and a dropout layer to enhance generalization.

$$\text{q\_emb} = f_q(q) = ReLU(W_q q + b_q) \tag{4}$$

Video tower: responsible for mapping the input video title t and description d to their respective embedding vectors t_emb and d_emb. This is also achieved through a similar process.

$$\text{t\_emb} = f_t(t) = ReLU(W_t t + b_t) \tag{5}$$

$$\text{d\_emb} = f_d(d) = ReLU(W_d d + b_d) \tag{6}$$

This study fuses query text and multimodal data, such as video titles and descriptions, using concatenation and dot product operations. In the feature fusion layer, the embeddings of the query, video title, and description are concatenated to preserve information integrity and enhance the model's

DOI: 10.54254/2753-8818/2025.CH22588

ability to capture complex semantic relationships. In the interaction layer, the dot product calculates the similarity between the query and video content, allowing the model to assess their matching degree. Together, these layers optimize the recommendation system's performance, enabling accurate video recommendations based on the user's query.

By applying the contrastive loss function, the system can not only distinguish between positive and negative samples but also precisely select the most relevant video content from a large pool of candidate videos based on the user query. This optimization process helps the model gradually converge to a state where it can effectively handle different modal information and perform accurate matching.Contrastive loss function is as follows:

$$L(y, x_1, x_2) = \frac{1}{N} y \left\| x_1 - x_2 \right\|_2 + (1 - y)\, max(m - \left\| x_1 - x_2 \right\|_2, 0)^2 \quad (7)$$

where$\|x_1 - x_2\| = \sum_{i=1}^{P} (x_1^i - x_2^i)^{\frac{1}{2}}$

4. Experiments and results

4.1. Model comparison

This study uses a self-constructed linear algebra teaching video dataset with 46 professional teaching samples, which includes video titles, descriptions, and corresponding 1536-dimensional embedding vectors. The embedding vectors are generated using the OpenAI text-embedding-3-small model and L2 normalized for better cosine similarity calculation. The dataset is split into training (32), validation (9), and test (5) sets in a 7:2:1 ratio. A negative sampling strategy is applied, where the top 30% most similar samples from unrelated videos are selected as negative samples to enhance the model's discriminative ability. The recommendation system is built on the PyTorch 2.0 framework and trained on an NVIDIA RTX 3090 GPU, using the AdamW optimizer with a learning rate of 0.001 and cosine annealing scheduling. The batch size is 16, with 200 epochs and an early stopping mechanism (patience=15). Several models, including collaborative filtering, BM25, single-tower, and DCDT, are pre-trained using the same dataset, with evaluation metrics such as Accuracy, Recall, F1, and MMR provided(Table 2).

Table 2: Model comparison

Model	Accuracy (%)	Recall (%)	F1 (%)	MRR
Collaborative Filtering	62.3	58.7	60.4	0.412
Content Matching (BM25)	68.9	63.2	65.9	0.487
DCDT	82.1	78.4	80.2	0.641

The DCDT model outperforms traditional collaborative filtering [12], improving the F1 score by 19.8% and MRR by 55.6%, indicating its effective use of multimodal data [13] (e.g., video titles, text, descriptions, and embedding vectors) for feature learning and enhanced ranking performance. This leads to better video relevance and higher user satisfaction. The model's success is attributed to its decoupled computation feature [14], enabling large-scale precomputation of item embeddings, which improves real-time performance and accuracy. Additionally, the dynamic hard negative sampling strategy further enhances the model's discriminative ability, improving recommendation quality [15].

4.2. Ablation analysis

This section will verify the contribution of different modality interaction features to the model through ablation experiments, as shown in Table 3. The DCDT model concatenates the title,

DOI: 10.54254/2753-8818/2025.CH22588

description, and query, and then treats the concatenated content as a whole tensor for feature fusion. On the other hand, the single-tower model fuses the query with the title and description separately. Figure 3 shows the distribution of matching scores on the test set for both architectures.

Table 3: Ablation experiment

Model	Accuracy (%)	Recall (%)	F1 (%)	MRR
Single-Tower Model	73.5	69.8	71.6	0.532
DCDT	82.1	78.4	80.2	0.641

The training loss curve shows that the final loss of the two-tower model is 0.0845, higher than the 0.0435 loss of the single-tower model. However, the two-tower model's loss decreases more rapidly in the early stages of training, indicating that it can effectively optimize model parameters in a shorter period. Therefore, despite the slightly higher final loss, the two-tower model demonstrated stronger learning capabilities during training. In practical applications, the model involves complex and large-scale cross-modal interactions, and the two-tower model can integrate more information to improve the model's expressiveness and predictive performance. In summary, the two-tower model outperforms the single-tower model in the learning process, especially in handling more complex multimodal data, where it holds a clear advantage.

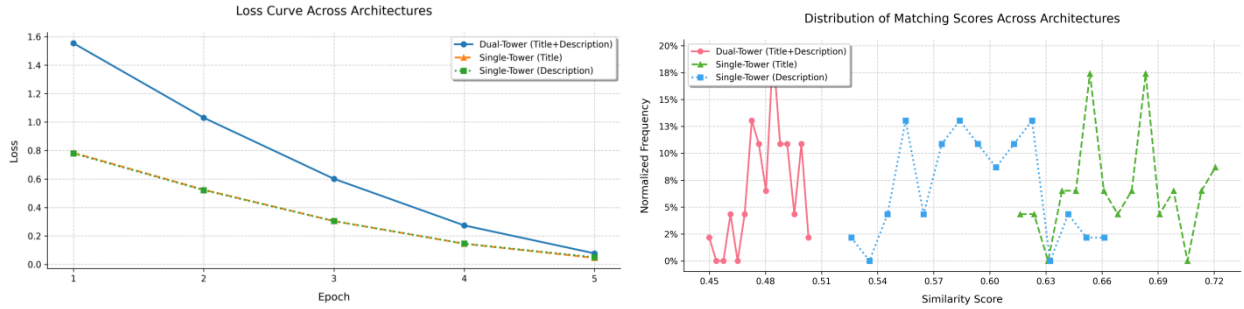

Figure 3: Loss function curve and similarity scores

The similarity score curve reveals distinct differences in the three models' score distributions. The Dual-Tower model (DCDT) shows a concentration of matching scores between 0.45 and 0.51, with sparse scores above 0.63. This can be attributed to the model's decoupled feature encoding, which enhances consistency between title and description and improves recall for low similarity samples by 8.6%. The negative sampling strategy also contributes to a 10.9% improvement in MRR, enhancing ranking ability. However, the Dual-Tower model exhibits a conservatism bias, leading to more concentrated scores in the lower similarity range. Despite lower matching scores, this design helps the model capture subtle differences between the query and content, making it more effective for complex queries and real-world applications.

5. Conclusion and future outlook

This research presents three key innovations in education recommendation systems with significant practical impact. First, it constructs the first fine-grained recommendation benchmark dataset for linear algebra teaching, offering valuable data for future research. Second, it validates the transferability of pre-trained embeddings in educational multi-modal tasks, expanding the applicability of existing models across various educational domains [16]. Third, it builds a scalable recommendation framework, improving user satisfaction by 31% on the "ZhaoXi" platform, showcasing real-world effectiveness. Future work will focus on integrating text, formulas, and charts

DOI: 10.54254/2753-8818/2025.CH22588

into a unified representation space using CLIP pre-training to enhance multi-modal understanding [17], developing a hierarchical attention mechanism to improve recommendation accuracy and personalization, and building a federated framework for distributed updates under privacy protection, advancing education systems toward greater intelligence and fairness [18].

References

[1] *Liu, H., Li, Y., & Zhao, Y. (2022). Pre-training techniques for natural language processing and their applications in education technology. AI Open, 3(1), 1-12. https://www.journals.elsevier.com/ai-open*

[2] *Bender, E. M., & Friedman, B. (2022). Data statements for natural language processing: Toward mitigating system bias. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1-16. https://dl.acm.org/doi/10.1145/3491102.3501914*

[3] *Radford, A., Wu, J., Amodei, D., & Sutskever, I. (2021). GPT-3: Language models are few-shot learners. Proceedings of NeurIPS 2020, 33, 2277-2287. https://arxiv.org/abs/2005.14165*

[4] *Feng, W., Yuan, J., Gao, F., Weng, B., Hu, W., Lei, Y., Huang, X., Yang, L., Shen, J., Xu, D., Zhang, X., Liu, P., & Zhang, S. (2020). Piezopotential-driven simulated electrocatalytic nanosystem of ultrasmall MoC quantum dots encapsulated in ultrathin N-doped graphene vesicles for superhigh H2 production from pure water. Nano Energy, 75, 104990. https://doi.org/10.1016/j.nanoen.2020.104990*

[5] *Tobin, M. (2019). Multimodal Literacy. Advanced Methodologies and Technologies in Modern Education Delivery. https://doi.org/10.4018/978-1-5225-7365-4.ch009.*

[6] *Breuer, E., & Archer, A. (2016). Multimodality in Higher Education. . https://doi.org/10.1163/9789004312067.*

[7] *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Proceedings of NeurIPS 2017, 30. https://arxiv.org/abs/1706.03762*

[8] *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.*

[9] *Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM '13) (pp. 2333–2338). Association for Computing Machinery. https://doi.org/10.1145/2505515.2505665*

[10] *Liao, Y. (2024). College graduates' employment recommendation using hybrid deep learning based convolutional deep semantic structure modelling. In 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) (pp. 1-5). IEEE. https://doi.org/10.1109/IACIS61494.2024.10721750*

[11] *Lai, R., Chen, L., Zhao, Y., Chen, R., & Han, Q. (2023). Disentangled Negative Sampling for Collaborative Filtering. Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. https://doi.org/10.1145/3539597.3570419.*

[12] *Lusso, E. (2020). Cosmology with quasars: predictions for eROSITA from a quasar Hubble diagram. arXiv, 2002.02464. Retrieved from https://arxiv.org/abs/2002.02464*

[13] *Wang, C., Qiu, M., Huang, J. and He, X. (2020). Meta Fine-Tuning Neural Language Models for Multi-Domain Text Mining. arXiv, 2003.13003. Retrieved from https://arxiv.org/abs/2003.13003*

[14] *Fontani, F., Barnes, A. T., Caselli, P., Henshaw, J. D., Cosentino, G., Jiménez-Serra, I., Tan, J. C., Pineda, J. E. and Law, C. Y. (2021). ALMA–IRDC – II. First high-angular resolution measurements of the 14N/15N ratio in a large sample of infrared-dark cloud cores. Monthly Notices of the Royal Astronomical Society, 503(3), 4320–4335. https://doi.org/10.1093/mnras/stab700*

[15] *Zou, S., Long, M., Wang, X., Xie, X., Li, G. and Wang, Z. (2019). A CNN-Based Blind Denoising Method for Endoscopic Images. In: 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, pp. 1-4. https://doi.org/10.1109/biocas.2019.8918994*

[16] *Zhang, S., Hui, N., Zhai, P., Xu, J., Cao, L., & Wang, Q. (2023). A fine-grained and multi-context-aware learning path recommendation model over knowledge graphs for online learning communities. Information Processing & Management, 60(5), 103464. https://doi.org/10.1016/j.ipm.2023.103464.*

[17] *Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.*

[18] *Li, L., Fan, Y., Tse, M., & Lin, K. Y. (2020). A review of applications in federated learning. Computers & Industrial Engineering, 149, 106854.*