

End-to-end training of Multimodal Model and ranking Model

Xiuqi Deng, Lu Xu, Xiyao Li, Jinkai Yu, Erpeng Xue, Zhongyuan Wang, Di Zhang,
Zhaojie Liu, Yang Song, Guorui Zhou, Na Mou, Shen Jiang

KuaiShou Inc.
Beijing, China

ABSTRACT

Traditional recommender systems heavily rely on ID features, which often encounter challenges related to cold-start and generalization. Modeling pre-extracted content features can mitigate these issues, but is still a suboptimal solution due to the discrepancies between training tasks and model parameters. End-to-end training presents a promising solution for these problems, yet most of the existing works mainly focus on retrieval models, leaving the multimodal techniques under-utilized. In this paper, we propose an industrial multimodal recommendation framework named EM3: End-to-end training of Multimodal Model and ranking Model, which sufficiently utilizes multimodal information and allows personalized ranking tasks to directly train the core modules in the multimodal model to obtain more task-oriented content features, without overburdening resource consumption. First, we propose Fusion-Q-Former, which consists of transformers and a set of trainable queries, to fuse different modalities and generate fixed-length and robust multimodal embeddings. Second, in our sequential modeling for user content interest, we utilize Low-Rank Adaptation technique to alleviate the conflict between huge resource consumption and long sequence length. Third, we propose a novel Content-ID-Contrastive learning task to complement the advantages of content and ID by aligning them with each other, obtaining more task-oriented content embeddings and more generalized ID embeddings. In experiments, we implement EM3 on different ranking models in two scenario, achieving significant improvements in both offline evaluation and online A/B test, verifying the generalizability of our method. Ablation studies and visualization are also performed. Furthermore, we also conduct experiments on two public datasets to show that our proposed method outperforms the state-of-the-art methods.

CCS CONCEPTS

- Information systems → Recommender systems; Multimedia information systems.

KEYWORDS

Multimodal recommendation, Recommender systems, Multimodal, Contrastive learning;

1 INTRODUCTION

For the past decades, recommender systems (RS) have achieved a fabulous performance[52]. Large-scale industrial RS usually use unique identities (ID) to represent users and items. Benefited from its strong abilities to remember and capture the user-item relationships, this ID-based paradigm dominates the RS fields until now [49, 54].

But there are still some shortcomings. On the one hand, ID embeddings have the cold-start problems because of data sparsity [14].

On the other hand, it may cause the estimation variance on items with similar materials [2], known as the generalization problems. One solution is to model content features so that inference can be made without interaction records [11], which shows better accuracy than general ID-based models [57]. This can be done in two ways: pre-extraction (PE) and end-to-end (E2E) [56].

The PE paradigm extracts frozen features from content models and feeds them into recommendation models (RM) as common features, also known as the two-stage paradigm [31, 49, 58]. However, the content-oriented pre-training tasks do not match well with the downstream personalized task [26, 47]. Besides, industrial RM require continuous training to follow the time-varying online distribution [25], while the content models remain frozen, giving rise to parametric mismatch. Both of the above reasons will result in suboptimal performance.

On the contrary, E2E refers to feeding low-level content features into RM and updating the content model together. Earlier studies have shown that E2E performs better than PE [25, 49]. However, most of them are based on two-tower [9, 13, 43, 48, 57] or session-based models [46, 51, 53], which are generally used as retrieval models. Few related studies on industrial ranking models only consider a single visual modality [3, 4, 10, 25]. Nowadays, online platforms usually have diverse content domains [6], utilizing multimodal information can capture useful information that is invisible in the single modality and can handle the modality missing problem to get a suitable content representation [56]. To summarize, the E2E training of multimodal model and ranking model is a valuable and promising direction that has not been fully explored.

In this paper, we propose an industrial multimodal recommendation framework named **EM3: End-to-end training of Multimodal Model and ranking Model**. As shown in fig. 1, EM3 sufficiently utilizes multimodal information and allows personalized ranking tasks to directly train the core modules in multimodal model, obtaining more task-oriented content representations without overburdening resource consumption. First, inspired from BLIP2 [22], we propose Fusion-Q-Former, which consists of transformers and a set of trainable queries, to fuse different modalities and generate fixed-length and robust multimodal embeddings. Second, in our sequential modeling for user content interest, we utilize Low-Rank Adaptation technique [16] to alleviate the conflict between huge resource consumption and long sequence length. Moreover, we propose a novel Content-ID Contrastive learning task to complement the advantages of content and ID by aligning them with each other, obtaining more task-oriented content embeddings and more generalized ID embeddings. In our comprehensive experiments, we verify the effect of EM3 on two different ranking models in our system, achieving significant improvements on the offline dataset with billions of records and online A/B test, contributing to millions of revenue. A series of ablation studies and a visualized analysis are also presented.

Furthermore, we conduct experiments on two public datasets to show that our proposed method outperforms the state-of-the-art methods in terms of recommendation accuracy, and the source code is available at <https://github.com/em3e2e-anonymous/em3>.

Our main contributions can be summarized as follows:

- i. To the best of our knowledge, this is the first work to propose an industrial framework for E2E training of multimodal model and ranking model, verifying the value and feasibility of this direction in both academia and industry.
- ii. We propose Fusion-Q-Former to fuse different modalities, which consists of transformers and a set of trainable queries, generating fixed-length and robust multimodal embeddings.
- iii. We utilize Low-Rank Adaptation technique to alleviate the conflict between the huge number of trainable parameters and the sequence length in sequential modeling.
- iv. We propose a novel Content-ID-Contrastive learning task to complement the advantages of content and ID by aligning them with each other, obtaining more task-oriented content embeddings and more generalized ID embeddings.

2 RELATED WORK

Our work is closely related to three research areas: recommendation system, multimodal model, and multimodal recommendation.

2.1 Recommender Systems

Industrial RS are generally divided into two stages: retrieval and ranking. Retrieval narrows down the selection to a small set of items, while ranking provides more precise estimation [54]. In this paper, we focus on ranking models, which are usually single-tower and sensitive to resource consumption.

In early ages, ranking model typically adopt collaborative filtering, logistic regression or matrix factorization [20, 30, 38] to capture user-item relationships. Later, the contextual features, user profiles and item attributes are integrated into the RS through more sophisticated models [11] such as FM [36], FFM [18], Wide&Deep [5] and DeepFM [12]. Recently, user interests are further excavated by sequential modeling such as DIN [55] and SIM [33]. Today, industrial ranking models are usually hybrid models that include the above various technologies.

2.2 Multimodal Model

The research on multimodal model was originally divided into two separate fields:

- Computer Vision (CV) includes tasks such as image classification and object detection. Initially, CNN were widely used [21, 39, 41]. Recently, Vision Transformer (ViT) has achieved remarkable results [8, 28, 45].
- Natural Language Processing (NLP) includes tasks such as machine translation and Q&A. RNN [1, 40] and BERT [7, 16, 27, 44] have successively dominated for many years. Currently, large language models (LLM) are ushering in a new era of generative models [32, 35].

Nowadays, multimodal learning unifies these two domains, aiming to extract and understand multi-media information better when various modalities are engaged [17]. Some researchers focus on

fusing different modalities into a single embedding, such as the single-flow paradigm in ViLT [19] and two-flow paradigm in ViLBert [29]. Others maintain two independent representations by aligning them across domains, such as CLIP [34]. ALBEF [24] and BLIP [23] integrate the two paradigms into a unified framework. BLIP2 [22] further merge visual features into LLM via Q-Former, becoming one of the most representative approaches of Multimodal LLM (MLLM).

2.3 Multimodal Recommendation

Most of the previous works on multimodal recommendation are in the PE paradigm [14, 31, 50, 56, 57], few E2E works are mainly based on two-tower models [13, 42, 43, 48, 50, 57], markedly different from the industrial ranking models, so we do not go into the PE or two-tower paradigm in detail in this section.

In the field of E2E training of multimodal model and ranking model, DeepCTR [3] incorporates a trainable CNN into the ranking model to capture visual features associated with advertising (ad). To accelerate the training speed, it groups samples with the same image into the same batch, thereby reducing the times of CNN forward propagation. CSCNN [25] makes use of the plug-in attention module to feed e-commerce (e-com) categories as side information. This approach helps CNN extract diverse visual features. DICM [10] not only adds a trainable image encoder on the item side but also implements it on user behavioral sequences, leading to a huge improvement. To enhance the training efficiency, it designs the AMS framework to separately deploy ID parts and multimodal parts on different devices. Besides, it selects the trainable-fixed hybrid paradigm by only updating the top fully connected (FC) layers of visual encoder. HCCM [4] combines the key benefits of the aforementioned methods: it utilizes ad categories as side information and extends its application to user behavioral sequences.

In summary, the pioneering works have preliminarily realized E2E training in industrial ranking scenario, but there is still room for improvement: First, they only utilize single visual modality, while multimodal information in the current online environment has not been utilized. Second, despite they indicate the improvement of sequential modeling, the huge resource consumption will limit the sequence length in practice. Third, the complementary advantages between content and ID can be further leveraged. Last but not least, they all conduct experiments on batch-training CTR models, we can generalize it to a wider range of scenarios such as online-learning models and CVR models.

3 METHOD

3.1 Ranking Model

In the ranking model, there are a lot of features such as ID, context and session features. We concatenate and feed them into DNN, and the softmaxed output of DNN is used to optimize the negative log-likelihood function:

$$L_{\text{ranking}} = -\frac{1}{B} \sum_{i=1}^B y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i), \quad (1)$$

where B is the batch size, $y_i \in \{0, 1\}$ is the class label denoting whether a click or conversion happens, \hat{y}_i is the softmaxed output of DNN.

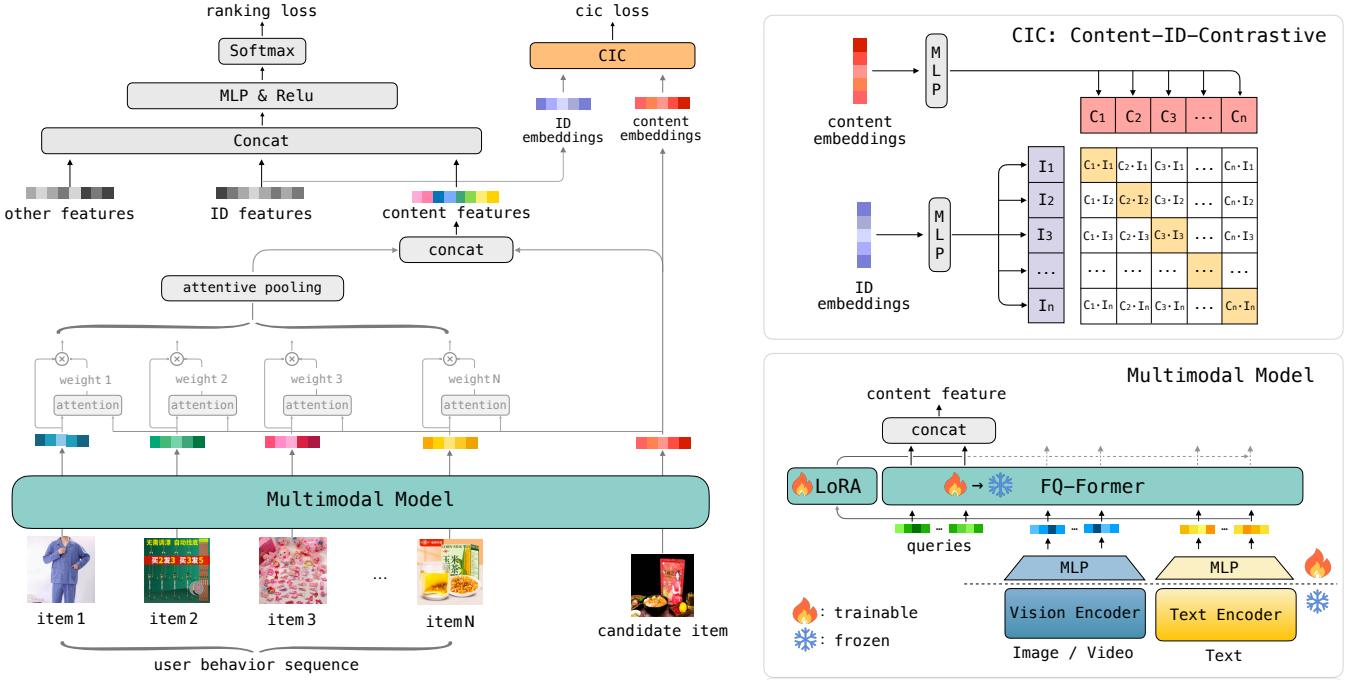


Figure 1: The overall framework of EM3. (a) Multimodal Model: aims to fuse modalities and shares parameters across different items. (b) CIC: a self-supervised task to align content and ID by maximizing the cosine similarities of the pairs on the diagonal.

Due to the confidentiality policy, we can't show much details of the ranking models, so we only use a simple model to represent our ranking model. Actually, it is a complicated and large-scale ID-based model. It can be replaced to any other models in practice.

3.2 Multimodal Model

3.2.1 Modalities. We select two modalities: visual modality from video frames, and text modality from titles, descriptions, ASR and OCR. Given the item A , we use $\text{img}_m(A)$ and $\text{txt}_k(A)$ to represent the raw content materials. Next, we extract single-modal features from vision encoder f_{ve} and text encoder f_{te} :

$$\bar{v}_m(A) = f_{\text{ve}}(\text{img}_m(A)), \quad m = 1, 2, \dots, M, \quad (2)$$

$$\bar{t}_k(A) = f_{\text{te}}(\text{txt}_k(A)), \quad k = 1, 2, \dots, K, \quad (3)$$

where M indicates the number of visual modalities, K indicates the number of text modalities, $\bar{v}_m(A)$ and $\bar{t}_k(A)$ represent the output of single-modal encoder.

Then, we downsize their dimensions through several FC layers:

$$v_m(A) = f_v(\bar{v}_m(A)), \quad (4)$$

$$t_k(A) = f_t(\bar{t}_k(A)), \quad (5)$$

in which f_v and f_t represent the FC layers, $v_m(A)$ and $t_k(A)$ are the final single-modal representations.

3.2.2 Multimodal Fusion. Motivated by [22], we propose Fusion-Q-Former (FQ-Former) to fuse different modalities, which is made up of transformers and a set of trainable queries.

Given the sets of single-modalities $v_A = \{v_1(A), \dots, v_M(A)\}$ and $t_A = \{t_1(A), \dots, t_K(A)\}$, we concatenate them with the globally

shared queries $q = \{q_1, \dots, q_Q\}$, where Q is the number of queries. Then we input them into the transformers, which are abbreviated as TRM in eq. (6), aiming to learn the relationships and importance between different modalities using self-attention (SA). Finally, we slice the first Q output tokens to generate the multimodal content features c_A :

$$c_A = f_{\text{FQ-Former}}(q, v_A, t_A) \\ = \text{TRM}(\text{concat}([q, v_A, t_A]))[:Q]. \quad (6)$$

FQ-Former has advantages over traditional fusion methods: (i) *Fixed-length*: the output size of FQ-Former is fixed and regardless of the number of modalities, making it more suitable for the industrial variable-length modalities. (ii) *Robust*: since the queries participate in SA, FQ-Former can relieve the potential negative impacts, which might be caused by the low-quality materials or already enough interaction data, by assigning more attention weight to queries.

3.2.3 Hybrid Training. It is widely acknowledged that recent open-source backbones have been sufficiently trained and very close to the ceiling. Therefore, freezing the weights in single-modal encoders and only training the top layers or additional modules are gradually adopted by recent works [4, 10, 16, 22]. We also choose this hybrid training paradigm by freezing f_{ve} and f_{te} ; and end-to-end training the f_v , f_t , q and $f_{\text{FQ-Former}}$, to balance the trade-off between efficiency and effectiveness.

3.3 Sequential Modeling

3.3.1 User Content Interest. Earlier works [33, 55] have shown that sequential modeling can improve the performance, guiding us to model user content interest in a similar manner.

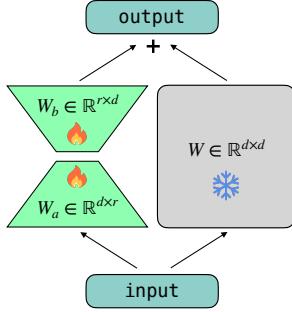


Figure 2: The details of LoRA module. The right $W^{d \times d}$ is frozen, while the left $W_a^{d \times r}$ and $W_b^{r \times d}$ are trainable.

Given the candidate item A and the user behavior sequence $\mathbf{u} = \{u_1, u_2, \dots, u_N\}$, we first utilize the multimodal model to generate their content embeddings c_A and $\{c_{u_1}, c_{u_2}, \dots, c_{u_N}\}$. Then we calculate the attention score between c_A and each c_{u_i} , and utilize the scores to weighted average the content sequence:

$$u_A = f(c_A, c_{u_1}, c_{u_2}, \dots, c_{u_N}) = \sum_{i=1}^N a(c_A, c_{u_i}) c_{u_i} = \sum_{i=1}^N w_i c_{u_i}, \quad (7)$$

$$w_i = \frac{\exp(a(c_A, c_{u_i}))}{\sum_{j=1}^N \exp(a(c_A, c_{u_j}))}, \quad (8)$$

where u_A represents the user content interest, $a(\cdot)$ represents the attention score function, N is the sequence length, w_i is the softmaxed attention score.

3.3.2 LoRA-based Long-term Content Interest. In practice, we encounter difficulties with GPU OOM when we intend to increase the sequence length. Given the hidden-size d , each FC layer in transformers requires $O(d^2)$ consumption. Taking the sequence length N into consideration, the consumption inflates to $O(N \cdot d^2)$, greatly limiting the sequence length.

We notice that the parameters in multimodal model will stabilize after a few days of training, which inspires us to utilize the Low-Rank Adaptation (LoRA) technique [16] by switching from full tuning to partial tuning. Let's consider the weights $W^{d \times d}$ that need to be optimized, we first train it on a short sequence length. After a period of sufficient training, we freeze W and add a trainable LoRA module to continuously follow the time-varying online distribution.

As shown in fig. 2, the LoRA module contains two bypassed FC layers $W_a^{d \times r}$ and $W_b^{r \times d}$, where $r \ll d$. Compared with the $W^{d \times d}$, they have the same output size but fewer trainable parameters. We add their outputs together:

$$f_{\text{LoRA}}(x) = \text{sg}(Wx) + W_b(W_ax), \quad (9)$$

where $\text{sg}(\cdot)$ is the stop gradient operator. By the way, we reduce the number of trainable parameters from $O(d^2)$ to $O(rd)$, allowing us to increase the sequence length N so that we can model user long-term content interest.

3.4 Content-ID-Contrastive Learning

As we know, ID embeddings have good memory and outperform on popular items, but have cold-start issues. Content embeddings are more generalizable, but they cannot benefit from interaction

data. We propose a Content-ID-Contrastive (CIC) learning task to complement their advantages.

Given the item i , we first select several important ID embeddings (e.g. ItemID, CategoryID) and concatenate them as id_i . Next, we linearly transform the content embedding c_i and ID embeddings id_i into the same vector space:

$$C_i = f_{\text{CIC}}(c_i), \quad (10)$$

$$I_i = f'_{\text{CIC}}(\text{id}_i), \quad (11)$$

where f_{CIC} and f'_{CIC} are the FC layers, C_i and I_i are the output embeddings with the same dimension. They are positive samples of each other in the following contrastive learning.

We randomly choose H negative samples for each C_i and I_i from the training batch, which are defined as $I_i^- = \{I_{i1}^-, I_{i2}^-, \dots, I_{iH}^-\}$ and $C_i^- = \{C_{i1}^-, C_{i2}^-, \dots, C_{iH}^-\}$. Then we utilize negative log-likelihood function to maximize the similarities of each positive pair and minimize the similarities of negative pairs:

$$L_{\text{C2I}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s(C_i, I_i)/\tau)}{\exp(s(C_i, I_i)/\tau) + \sum_{j=1}^H \exp(s(C_i, I_{ij}^-)/\tau)}, \quad (12)$$

$$L_{\text{I2C}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s(I_i, C_i)/\tau)}{\exp(s(I_i, C_i)/\tau) + \sum_{j=1}^H \exp(s(I_i, C_{ij}^-)/\tau)}, \quad (13)$$

where $s(\cdot)$ represents the cosine similarity, τ is the temperature parameter. Finally, we add their average to the ranking loss:

$$L = L_{\text{ranking}} + \alpha \cdot L_{\text{CIC}} = L_{\text{ranking}} + 0.5\alpha \cdot (L_{\text{C2I}} + L_{\text{I2C}}), \quad (14)$$

where α is a hyperparameter.

Considering the constraint of both ranking loss and CIC loss, CIC has different effects on different items:

- On *popular items*: ID will dominate the alignment and inject user behavioral information into the content embeddings, generating more task-oriented content embeddings and modeling user interest better.
- On *cold-start items*: content will dominate the alignment and guide the update of ID. As a result, the items with similar materials will also have similar ID embeddings, which promotes the generalization of RS.

3.5 Feature System

We design the feature system as shown in fig. 3 to optimize the efficiency in two aspects:

- *Training*: since the f_{ve} and f_{te} are frozen, we precalculate and cache the single-modal features $\bar{v}_m(A)$ and $\bar{t}_k(A)$ to accelerate the offline training.
- *Serving*: we infer and cache multimodal embeddings c_A for all items at regular intervals. When RS receiving an online request, the multimodal embeddings can be looked up directly without the forward propagation.

4 EXPERIMENT

4.1 Setup

Baselines. We conducted experiments on two state-of-the-art production models in our system:

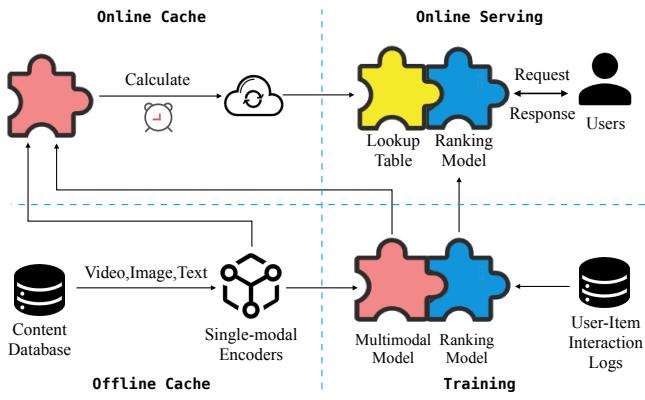


Figure 3: The feature system of EM3.

- An e-com online-learning CTR model that predicts whether a user will click on the item.
- An ad batch-training CVR model that estimates whether a user will stay in the app in the next day.

Datasets. Our experimental datasets come from our RS. In the e-com scenario, we use 2.1 billion records as the training set and 250 million as the test set. In the ad scenario, we use 82 million records as the training set and 2.2 million as the test set. Due to the confidentiality policy, we do not specify the beginning and ending time of dataset.

Metrics. In offline evaluation and ablation studies, we use AUC as the primary performance indicator. In online A/B test, we consider three key metrics for the e-com system: the Gross Merchandise Value (GMV), Order Volume, and CTR. For the ad system, we consider two important metrics: the Revenue Per Mille (RPM) and Income.

Backbones. We utilize the Swin-T-22K¹ as the vision encoder, and choose the RoBERTa-wwm-ext² as the text encoder. Both have been proven to be powerful single-modal models in practice.

Hyperparameters. The values of the hyperparameters that we select are as follows: the number of queries is 2, the number of FQ-Former layers is 1, the number of SA heads is 4, and the weight of CIC loss is 0.1, the CIC temperature is 0.1.

4.2 Offline Evaluation

During training, the e-com model needs extra 50% GPUs for keeping up with online-training data; the ad model does not require additional GPU. Both are acceptable to us.

As shown in table 1, the e-com AUC increases by **0.256%**, and the ad AUC increases by **0.242%**, which are both significant under the T-test.

4.3 Comparison & Ablation

Except for the LoRA experiment, which is conducted on the e-com model, our other studies are all conducted on the ad model due to its lower experimental cost.

¹<https://github.com/microsoft/Swin-Transformer>

²<https://github.com/ymcui/Chinese-BERT-wwm>

Table 1: Comparison of AUC.

scenario	Method	AUC	AUC gain
e-com	baseline	0.7803	-
	EM3	0.7823	0.256%
ad	baseline	0.7016	-
	EM3	0.7033	0.242%

4.3.1 Modalities. We combine the following modalities in various ways to evaluate how different modalities impact the performance:

- Text: a sentence composed of title, description, ASR and OCR.
- Image: the cover frame of a short video.
- Video: several frames from a short video.

The results are listed in table 2, demonstrating that multi-modalities can facilitate the modeling on recommendation, and increasing the number of modalities can further enhance performance.

Table 2: Comparison of modalities.

method	AUC	AUC gain
baseline	0.7016	-
only image	0.7024	0.114%
only text	0.7022	0.086%
image & text	0.7027	0.157%
video & text	0.7033	0.242%

4.3.2 Fusion Methods. We compare performance of the following fusion methods:

- 1flow [19]: concatenates modalities and fuses them using transformers. The variable-length sequential outputs with paddings are used as the content features.
- 2flow [29]: utilizes two independent transformers with cross-attention layers to allow the modalities to interact with each other. The output of text flow is used as the content feature.
- Masked FQ-Former: the queries interact with modalities via Q & K, but their V do not participate in pooling.
- FQ-Former: our proposed method.

As shown in table 3, FQ-Former outperforms traditional fusion methods, achieving the highest AUC gain, and taking queries into attentive pooling can further boost the improvements.

Table 3: Comparison of fusion methods.

method	AUC	AUC gain
baseline	0.7016	-
1flow	0.7030	0.199%
2flow	0.7029	0.190%
masked FQ-Former	0.7031	0.214%
FQ-Former	0.7033	0.242%

4.3.3 LoRA. We compare the AUC before and after using LoRA. As shown in table 4, when training all parameters with a sequence length 20, the AUC gain is 0.179%; when using LoRA with a sequence length 50, the AUC gain is 0.256%, which verifies that a

longer sequence with fewer trainable parameters can improve the performance.

Table 4: Ablation of LoRA.

method	AUC	AUC gain
baseline	0.7803	-
length=20 w/o LoRA	0.7817	0.179%
length=50 w/ LoRA	0.7823	0.256%

4.3.4 CIC. We conduct this ablation experiment to assess the improvement of CIC. As shown in table 5: when training without CIC, the AUC gain is 0.143%; when training with CIC, the AUC gain is 0.242%, demonstrating that CIC can benefit the performance.

Table 5: Ablation of CIC.

method	AUC	AUC gain
baseline	0.7016	-
w/o CIC	0.7026	0.143%
w/ CIC	0.7033	0.242%

4.3.5 Item & User. The final content features in our method are composed of two parts: the item-side c_A and the user-side u_A , this ablation study aims to understand the benefits of each. As shown in table 6, either item feature or user feature can improve the performance, combining them together achieves the best.

Table 6: Ablation of item & user features.

method	AUC	AUC gain
baseline	0.7016	-
only item	0.7021	0.071%
only user	0.7024	0.114%
item & user	0.7033	0.242%

4.3.6 Splitting the Gains. The gains of E2E may come from three aspects: (1) the injection of content information; (2) the guidance from recommendation task; (3) the parametric matching between the multimodal model and the ranking model. Accordingly, we set up three experimental groups to gradually split the benefits:

- PE: pre-extracts frozen features from a multimodal model, which has the similar structure with FQ-Former and has been fine-tuned in our system.
- Task-specific PE: we first fine-tune the content model through E2E method, and pre-extract content embeddings for all items. Finally, we feed them into the ranking model as frozen features.
- E2E: continuously trains the multimodal model with ranking model together.

As shown in table 7, the improvement of (1) is 0.071%; (2) adds another 0.114% increase; (3) provides an additional promotion of 0.057%. Combining them all together can result in a total promotion of 0.242%.

Table 7: Splitting the Gains of E2E.

method	AUC	AUC gain
baseline	0.7016	-
PE	0.7021	0.071%
task-specific PE	0.7029	0.185%
E2E	0.7033	0.242%

4.4 Online A/B Test

We set 7-day online A/B Test in both scenario. In the e-com scenario, EM3 contributes to a **3.22%** improvement on GMV, a **2.92%** increment on order volume, and an **1.75%** promotion on CTR. In the ad scenario, EM3 achieves a **2.64%** improvement on RPM and generates extra **3.17%** income. The results are all significant under the T-test.

It is worth mentioning that our method brings 2.07% more impressions for cold-start items, which will make the platform ecology healthier in the long run.

4.5 Impacts on Embeddings & Visualization

This section aims to quantitatively analyze and visualize how EM3 impacts the embeddings.

4.5.1 Cold-start Items. Cold-start IDs always lack generalization. In order to analyze the differences in ItemID embeddings between the baseline ranking model and EM3, we dump all their ItemID embeddings from the parameter server. Next, we randomly sample thousands of cold-start items that are created within 5 days, and search for similar items in their respective vector spaces using cosine similarity. Then we calculate the material similarities between each candidate item and its top3 similar items. This is done in an indifferent vector space from a frozen multimodal model, which has been fine-tuned in our system and has been utilized in section 4.3.6.

As the results shown in table 8 and the typical cases shown in fig. 4(a), the cold-start ItemIDs of EM3 achieve a great improvement in material similarity.

Table 8: Material similarity of ItemID.

model	material similarity	gain
baseline	0.3185	-
EM3	0.3363	5.588%

4.5.2 Popular Items. On popular items, we mainly focus on whether behavioral information has an impact on the content embeddings. For fairness, we choose the frozen multimodal model mentioned above as a comparison for EM3. First, we utilize the frozen model and EM3 to calculate content embeddings for all items. Next, we randomly sample thousands of popular items from top 30% GMV, and search for similar items in their respective vector spaces. Then we calculate the behavioral similarities between each candidate item and its top3 similar items. This is done in the vector space of baseline ItemID embeddings, which have only been trained by user-item interaction data and can be used to measure the behavioral similarity.



Figure 4: Visualization of the impacts on embeddings.

As the results shown in table 9 and the typical cases shown in fig. 4(b), we can conclude that the content embeddings of EM3 achieve a significant improvement in behavioral similarity.

Table 9: Behavioral similarity of content embeddings.

model	behavioral similarity	gain
frozen	0.2275	-
EM3	0.2514	10.505%

4.6 Public Evaluation

Because our proposed modules can be also combined with some two-tower models, we conduct experiments on two public datasets. Thanks to the MMRec³ framework contributed by [56], we can test at a very low cost.

Datasets. The Amazon dataset has been widely used in previous studies, providing both interaction records and multimodal information. We choose 2 categories: Baby and Sports. For each category, we randomly split 80% of historical records as a training set, 10% for validation and the remaining 10% for test.

³<https://github.com/enoche/MMRec>

Method. We integrate the fusion and CIC modules into the FREEDOM [57], which is one of the most effective works at the time of writing. In details, we first use FQ-Former to calculate the multimodal embeddings of all items. Then we utilize GCN to process them on both item-side and user-side. Next, we align the item-side multimodal embeddings with ItemID embeddings using CIC. Finally, we concatenate them for retrieving.

Same as the previous works, we fix the embedding size of both users and items to 64, and use the negative sampling strategy to pair each user-item interaction with one negative item.

Baselines. We compare our proposed model with the following methods: BPR [37], LightGCN [15], VBPR [14], SLMRec [42], LAT-TICE [50], BM3 [58], FREEDOM. For a fair comparison, we retrain FREEDOM and some other models in our environment.

Hyperparameters. We perform a grid search to find its optimal settings on different datasets: we search for the number of queries from {1, 2}, the dropout rate of transformer from {0.5, 0.8}, and the CIC temperature from {0.5, 0.1}.

Metrics. We select two widely-used evaluation metrics for top-K recommendation: Recall@K and NDCG@K, which are abbreviated as R@K and N@K. We use R@20 on the validation data as the

Table 10: Performance of different recommendation models on public datasets. The best results are marked in boldface and the second best results are underlined.

Datasets	Metrics	General Models				Multimodal Models			
		BPR	LightGCN	VBPR	SLMRec	LATTICE	BM3	FREEDOM	EM3
Baby	R@10	0.0357	0.0479	0.0423	0.0521	0.0551	0.0564	<u>0.0624</u>	0.0646
	R@20	0.0575	0.0754	0.0663	0.0772	0.0852	0.0883	<u>0.0980</u>	0.1032
	N@10	0.0192	0.0257	0.0223	0.0289	0.0292	0.0301	<u>0.0324</u>	0.0336
	N@20	0.0249	0.0328	0.0285	0.0354	0.0369	0.0383	<u>0.0416</u>	0.0435
Sports	R@10	0.0432	0.0569	0.0560	0.0663	0.0621	0.0656	<u>0.0713</u>	0.0726
	R@20	0.0653	0.0864	0.0854	0.0990	0.0957	0.0980	<u>0.1075</u>	0.1099
	N@10	0.0241	0.0311	0.0307	0.0365	0.0335	0.0355	<u>0.0384</u>	0.0391
	N@20	0.0298	0.0387	0.0383	0.0450	0.0422	0.0438	<u>0.0477</u>	0.0488

training stopping indicator, and report the average metrics of all users in the test sets for both K=10 and K=20.

Results. As shown in table 10, EM3 outperforms the baselines on both datasets. It shows that the method proposed by us is generalizable. The source code and evaluation logs are available at <https://github.com/em3e2e-anonymous/em3>.

Because the public evaluation is not the main focus of our research, we only simply integrate our method with one of previous works. Considering the flexibility of EM3 in decoupling from the original recommendation models, we believe that it can be combined with other models for a similar improvement.

5 CONCLUSION

In this paper, we propose an industrial multimodal recommendation framework named EM3 for end-to-end training of multimodal model and ranking model. EM3 sufficiently utilizes multimodalities and allows personalized ranking tasks to directly train the core modules in the multimodal model, obtaining more task-oriented content representations. In details, we propose FQ-Former to fuse different modalities and generate fixed-length and robust content embeddings. In user sequential modeling, we utilize LoRA technique to reduce the consumption of trainable parameters so that we can increase the length of behavioral sequence to model user content interest better. Besides, we propose a novel CIC learning task to complement the advantages of content and ID through alignment, which allows us to obtain more task-oriented content embeddings and more generalized ID embeddings. The experiments conducted in two different scenarios show that EM3 is achieves significant improvements and bring in millions of revenue, which also verify the generalizability of our method. The evaluation on public datasets also show that our proposed method outperforms the state-of-the-art methods.

In the future, we are going to focus on the direction of E2E training, to add more modalities such as audio, or to end-to-end train MLLM.

REFERENCES

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [2] Jin Chen, Tiezheng Ge, Gangwei Jiang, Zhiqiang Zhang, Defu Lian, and Kai Zheng. 2021. Efficient Optimal Selection for Composited Advertising Creatives with Tree Structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3967–3975.
- [3] Junxuan Chen, Baigui Sun, Hao Li, Hongtao Lu, and Xian-Sheng Hua. 2016. Deep ct prediction in display advertising. In *Proceedings of the 24th ACM international conference on Multimedia*. 811–820.
- [4] Xin Chen, Qingtao Tang, Ke Hu, Yue Xu, Shihang Qiu, Jia Cheng, and Jun Lei. 2022. Hybrid CNN Based Attention with Category Prior for User Image Behavior Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2336–2340.
- [5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.
- [6] Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-rec: Generative pretrained language models are open-ended recommender systems. *arXiv preprint arXiv:2205.08084* (2022).
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [9] Sherene Elsayed, Lukas Brinkmeyer, and Lars Schmidt-Thieme. 2022. End-to-end image-based fashion recommendation. In *Workshop on Recommender Systems in Fashion and Retail*. Springer, 109–119.
- [10] Tiezheng Ge, Liqin Zhao, Guorui Zhou, Keyu Chen, Shuying Liu, Huimin Yi, Zelin Hu, Bochao Liu, Peng Sun, Haoyu Liu, et al. 2018. Image matters: Visually modeling user behaviors using advanced model server. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2087–2095.
- [11] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.
- [12] Hufeng Guo, Ruiming Tang, Yuning Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).
- [13] Ruining He, Chunbin Lin, Jianguo Wang, and Julian McAuley. 2016. Sherlock: sparse hierarchical embeddings for visually-aware one-class collaborative filtering. *arXiv preprint arXiv:1604.05813* (2016).
- [14] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.
- [15] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgen: 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.
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- [17] Summaira Jabeen, Xi Li, Muhammad Shoib Amin, Omar Bourahla, Songyuan Li, and Abdul Jabbar. 2023. A review on methods and applications in multimodal

- deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 2s (2023), 1–41.
- [18] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM conference on recommender systems*. 43–50.
- [19] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.
- [20] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).
- [22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
- [23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.
- [24] Junnan Li, Ramprasad Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.
- [25] Hu Liu, Jing Lu, Hao Yang, Xiwei Zhao, Sulong Xu, Hao Peng, Zehua Zhang, Wenjie Niu, Xiaokun Zhu, Yongjun Bao, et al. 2020. Category-Specific CNN for Visual-aware CTR Prediction at JD. com. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2686–2696.
- [26] Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. 2023. Megef: Multimodal entity graph collaborative filtering for personalized recommendation. *ACM Transactions on Information Systems* 41, 2 (2023), 1–27.
- [27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.
- [29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).
- [30] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1222–1230.
- [31] Kaixiang Mo, Bo Liu, Lei Xiao, Yong Li, and Jie Jiang. 2015. Image feature learning for cold start problem in display advertising. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
- [32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- [33] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 2020. Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2685–2692.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [35] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [36] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.
- [37] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).
- [38] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.
- [39] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [40] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).
- [41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.
- [42] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* (2022).
- [43] Ivona Tautkute, Tomasz Trzcinski, Aleksander P Skorupa, Lukasz Brocki, and Krzysztof Marasek. 2019. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access* 7 (2019), 84613–84628.
- [44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [45] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 568–578.
- [46] Shitao Xiao, Zheng Liu, Yingxia Shao, Tao Di, Bhuvan Middha, Fangzhao Wu, and Xing Xie. 2022. Training large-scale news recommenders with pretrained language models in the loop. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4215–4225.
- [47] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. *arXiv preprint arXiv:2106.01804* (2021).
- [48] Yoonseok Yang, Kyu Seok Kim, Minsam Kim, and Junyeoung Park. 2022. GRAM: Fast Fine-tuning of Pre-trained Language Models for Content-based Collaborative Filtering. *arXiv preprint arXiv:2204.04179* (2022).
- [49] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. *arXiv preprint arXiv:2303.13835* (2023).
- [50] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.
- [51] Lingzi Zhang, Xin Zhou, and Zhiqi Shen. 2023. Multimodal Pre-training Framework for Sequential Recommendation via Contrastive Learning. *arXiv preprint arXiv:2303.11879* (2023).
- [52] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)* 52, 1 (2019), 1–38.
- [53] Zhipeng Zhang, Piao Tong, Yingwei Ma, Qiao Liu, Xujiang Liu, and Xu Luo. 2023. Language-Enhanced Session-Based Recommendation with Decoupled Contrastive Learning. *arXiv preprint arXiv:2307.10650* (2023).
- [54] Xiangyu Zhao, Maolin Wang, Xianjian Zhao, Jiansheng Li, Shucheng Zhou, Dawei Yin, Qing Li, Jiliang Tang, and Ruocheng Guo. 2023. Embedding in Recommender Systems: A Survey. *arXiv preprint arXiv:2310.18608* (2023).
- [55] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.
- [56] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions. *arXiv preprint arXiv:2302.04473* (2023).
- [57] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 935–943.
- [58] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.