TargetMR: Learning Modality Target for Multimodal Recommendation

Gu Tang
Shanghai Jiao Tong University
Shanghai, China
gutang@sjtu.edu.cn

Jinghe Wang
Shanghai Jiao Tong University
Shanghai, China
whalien-56@sjtu.edu.cn

Jiang Bo†
Shanghai Jiao Tong University
Shanghai, China
bjiang@sjtu.edu.cn

Ze Zhao
Shanghai Jiao Tong University
Shanghai, China
zhaoze@sjtu.edu.cn

Jianping Zhou
Shanghai Jiao Tong University
Shanghai, China
jianpingzhou@sjtu.edu.cn

Xiaoying Gan†
Shanghai Jiao Tong University
Shanghai, China
ganxiaoying@sjtu.edu.cn

Luoyi Fu
Shanghai Jiao Tong University
Shanghai, China
yiluofu@sjtu.edu.cn

Xinbing Wang
Shanghai Jiao Tong University
Shanghai, China
xwang8@sjtu.edu.cn

Chenghu Zhou
Chinese Academy of Sciences
Beijing, China
zhouch@lreis.ac.cn

Abstract

Rapid development of web services has led to an explosion of multimodal content, making multimodal recommender systems (MRSs) vital tools for mitigating information overload. Current MRSs have achieved remarkable progress by incorporating advanced technologies such as Graph Neural Networks (GNNs) and Large Language Models (LLMs). However, these studies still suffer from the *semantic shift problem*. Generally, item's multimodal content usually contain multiple objects, including target object (core content of item) and auxiliary objects (decorations of item). Existing MRSs overlooked this distinction, failing to prevent auxiliary objects from dominating the representation, leading to biased item representation. To address this issue, we propose a model-agnostic framework "**TargetMR**". Concretely, TargetMR comprises two core modules, including **Object Disentangler** and **Object Identifier**. The Object Disentangler decouples item text and image into multiple objects via text syntactic parsing and image segmentation. The Object Identifier performs knowledge distillation based on LLMs to efficiently identify the target text object. It then identifies the target image object through cross-modal semantic evaluation. Moreover, this module refines the representation of image target object by optimizing the semantic correlation. Owing to the model-agnostic design of TargetMR, it can be integrated into various backbone MRSs. Extensive experiments on three benchmark datasets show that TargetMR consistently improves the performance of five backbone MRSs, with an average improvement of 12.26%. Our codes are available at https://github.com/gutang-97/TargetMR/.

† Corresponding authors.

WWW '26, Dubai, United Arab Emirates.

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792466

CCS Concepts

• **Information systems** → **Recommender systems**; **Multimedia and multimodal retrieval**.

Keywords

Multimodal Recommendation; Semantic Shift; Knowledge Distillation

ACM Reference Format:
Gu Tang, Jinghe Wang, Jiang Bo†, Ze Zhao, Jianping Zhou, Xiaoying Gan†, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2026. TargetMR: Learning Modality Target for Multimodal Recommendation. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792466

1 Introduction

With the rapid growth of web services like e-commerce and social media, recommender systems (RSs) have become essential tools for providing personalized services. They infer user preferences based on user-item interactions to provide personalized recommendations. In recent years, the abundance of multimodal content—such as text, image, and audio—on web platforms has motivated numerous efforts [1–3], giving rise to the promising research field known as multimodal recommender systems (MRSs).

The development of MRSs has yielded many valuable research works. Early MRSs, such as VBPR [4] and DeepStyle [5], primarily integrated multimodal content with matrix factorization (MF) to derive modality-enhanced user and item representations. With the remarkable success of Graph Neural Networks (GNNs), a series of efforts (e.g., MGCN [1] and LATTICE [3]) have incorporated GNNs to model the high-order relationship between multimodal content and user behavior, resulting in impressive progress. Building upon these, subsequent studies [6, 7] further incorporate the rich knowledge of Large Language Models (LLMs) [8, 9] to enhance user/item profiles [6] or perform knowledge distillation [7], thus enhancing recommendation performance.

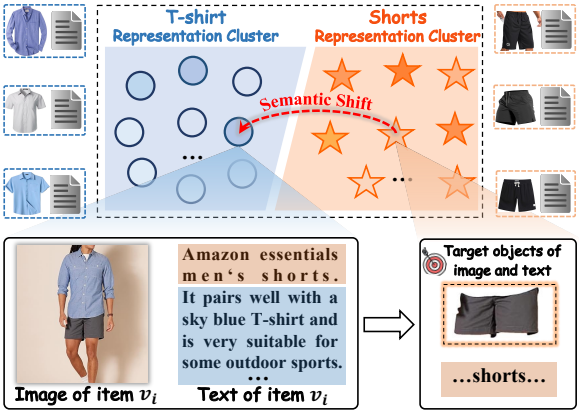

Figure 1: The illustration of semantic shift problem.

Despite the considerable progress, existing MRSs still suffer from the *semantic shift problem*. For a specific item, its multimodal content typically comprises multiple objects, including a target object (core content of item) and multiple auxiliary objects (decorations of item). However, current MRSs ignore the coexistence of these objects, which causes auxiliary objects to dominate the feature learning, resulting in biased item representation. As shown in Figure 1, the target object for the item v_i is "Shorts", and its auxiliary objects include a "T-shirt" (used to decorate the "Shorts"). Yet, the auxiliary object occupies a larger portion of image patches and textual descriptions. Hereby, "T-shirt" dominates the representation of item v_i, resulting in a biased representation.

To address the semantic shift problem, the key lies in overcoming the following two challenges: (i) Multiple objects within a modality (e.g., text or image) often lack clear discriminative criteria, which makes it difficult to disentangle a modality into distinct objects. (ii) The target object is hidden among multiple objects and lacks explicit annotation, posing a significant challenge for its accurate identification.

Consequently, we propose a model-agnostic framework termed "TargetMR: Learning Modality Target for Multimodal Recommendation" to tackle the aforementioned challenges. To be specific, TargetMR comprises two core modules, including **Object Disentangler** and **Object Identifier**. The Object Disentangler is designed to decouple modalities (e.g., text and image) into their multiple objects. It first decouples item text into multiple objects by parsing the part-of-speech (POS) and syntactic structure of the text. Thereafter, multiple image objects are obtained by performing image segmentation. Given the above-mentioned multiple text and image objects, the Object Identifier aims to identify target objects from them. It optimizes a discriminative architecture-based text target selector by distilling knowledge from LLMs, thereby enabling the identification of target text object and avoiding the issue of LLM instruction non-compliance. In addition, this module further develops an image target selector, which identifies target image object and refines its representations by evaluating and optimizing the semantic correlations. Owing to the model-agnostic design of TargetMR, it can be incorporated into various backbone MRSs, thereby enhancing their performance.

To summarize, the key contributions of this work are as follows:

- We propose TargetMR, a novel and model-agnostic framework, to address the semantic shift problem through modality decoupling and target identification. It can be integrated into various backbone MRSs, thereby enhancing their performance.
- We propose an Object Disentangler to decouple item text and images into multiple objects through performing syntactic parsing and image segmentation.
- The proposed Object Identifier determines target objects through knowledge distillation and semantic evaluation, followed by representation refinement of the image object via optimizing the semantic correlation.

Extensive experiments on three benchmark datasets show that TargetMR consistently enhances five backbone MRS, achieving state-of-the-art results and an average improvement of 12.26% across them.

2 Preliminary

In this section, we first formally define the concepts and notations used throughout this paper. Subsequently, we introduce the task definition of multimodal recommendation.

Concepts and Notations. Following previous works [1, 3, 10, 11], we denote the user-item interactions as a bipartite graph $\mathcal{G} = \{(u_i, y_{u_i,v_j}, v_j)|u_i \in \mathcal{U}, v_j \in \mathcal{V}\}$, where $\mathcal{U}$ and $\mathcal{V}$ correspond to the user set and item set, respectively. The $y_{u_i,v_j} \in \{0, 1\}$ is a binary indicator, where $y_{u_i,v_j} = 1$ signifies user u_i interacted with item v_j, and $y_{u_i,v_j} = 0$ otherwise. In MRS, each item is associated with multimodal content. In this paper, we focus on the two most prevalent modalities: text and image. Formally, given item v_i, its text and image are defined as sequence $\mathcal{T}_{v_i} = [t_1, t_2, \cdots, t_n]$ and $\mathcal{P}_{v_i} = [p_1, p_2, \cdots, p_m]$, respectively. t_i and p_i denote the i-th token and patch in $\mathcal{T}_{v_i}$ (including n tokens) and $\mathcal{P}_{v_i}$ (with m patches), respectively. For additional notations, we will introduce them in the corresponding section.

Task Formulation. Formally, the task of multimodal recommendation is formulated as follows: Given the user-item interaction graph $\mathcal{G}$ and multimodal content, we aim to optimize a learnable function $\mathcal{F}$ to predict the probability of a user adopting an item.

3 Methodology

The overall framework of TargetMR is shown in Figure 2, which consists of three components, including (i) Object Disentangler, (ii) Object Identifier and (iii) Downstream Task. In the following subsections, we will provide a detailed description for each module.

3.1 Object Disentangler

The goal of this module is to decouple item text and image into multiple objects, respectively. Structurally, this module consists of two parts: *Text Decouple* and *Image Decouple*.

3.1.1 ***Text Decouple.*** The key to decoupling item text into multiple text objects lies in identifying their corresponding tokens. Inspired by the part-of-speech (POS) tags and syntactic structures of text, we identify the tokens of different text objects by parsing the two features.

For the POS tags, given the text sequence $\mathcal{T}_{v_i} = [t_1, t_2, \cdots, t_n]$ of item v_i, we employ a POS tagging model $PosTag(\cdot)$ to derive the

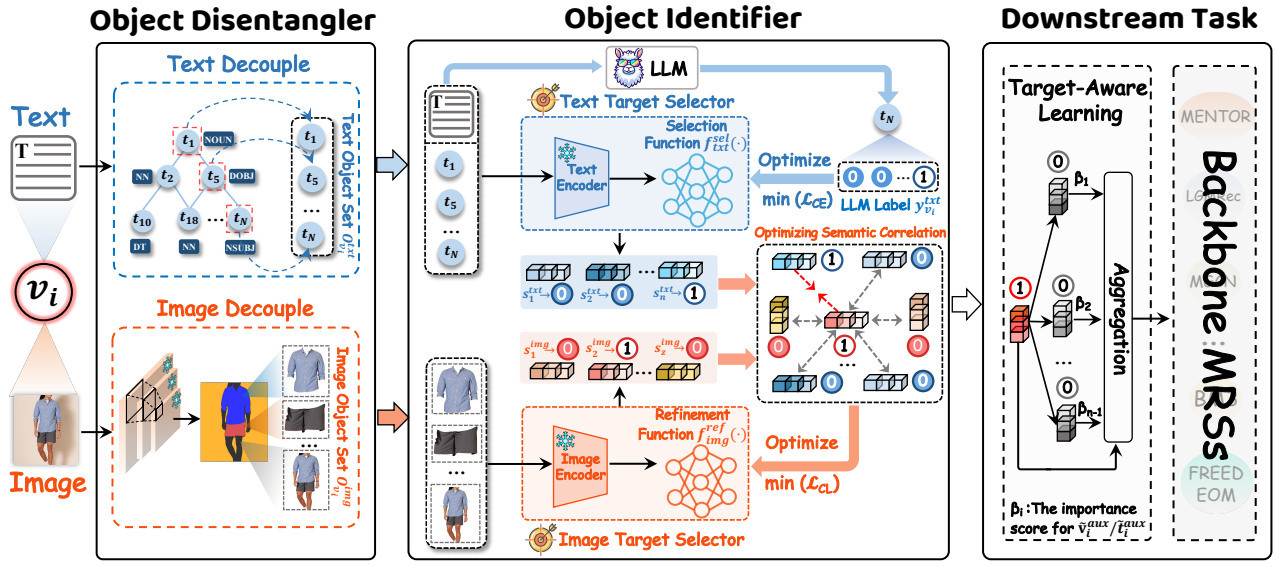

Figure 2: The overall framework of our proposed TargetMR. It consists of three components, i.e., Object Disentangler, Object Identifier and Downstream Task.

corresponding POS sequence $\mathcal{T}_{v_i}^p = [t_1^p, t_2^p, \cdots, t_n^p]$:

$$\mathcal{T}_{v_i}^p = PosTag(\mathcal{T}_{v_i}) = [t_1^p, t_2^p, \cdots, t_n^p], \tag{1}$$

where t_i^p denotes the POS tag of token t_i and the POS tagging model $PosTag(\cdot)$ is implemented using spaCy [12]. In MRS, the POS tags of item objects are predominantly covered by noun ($NOUN$). We therefore construct the MRS POS set $\mathcal{S}^p = \{NOUN\}$.

Regarding syntactic structure parsing, we adopt a parsing model $DepParse(\cdot)$ [12] to obtain the syntactic dependency tree $\mathcal{T}ree_{v_i} = \{(t_i, t_j, r_j)\}$ corresponding to $\mathcal{T}_{v_i}$, which is defined as follow:

$$\mathcal{T}ree_{v_i} = DepParse(\mathcal{T}_{v_i}) = \{(t_i, t_j, r_j) \mid t_i, t_j \in \mathcal{T}_{v_i}\}, \tag{2}$$

where t_i denotes head token, t_j is the dependent token, and r_j signifies the dependency relation of t_j. Each token t_j in $\mathcal{T}_{v_i}$ can correspond to a dependency relation r_j. Hence, we can obtain a syntactic sequence $\mathcal{T}_{v_i}^s = [r_1, r_2, \cdots, r_n]$. Within MRS, different objects of item text are primarily covered by nominal subject ($NSUBJ$), direct objects ($DOBJ$) or prepositional objects ($POBJ$). Therefore, we define the MRS relation set as $\mathcal{S}^r = \{NSUBJ, DOBJ, POBJ\}$.

Relying solely on either POS tags or syntactic structure parsing result in inaccurate localization of text object tokens, as evidenced in our ablation study (Sec. 4.3). Therefore, we perform a dual verification of object token t_i (i.e., $t_i^p \in \mathcal{S}^p \cap r_i \in \mathcal{S}^r$) from POS and syntactic perspectives, thus constructing the text object set $O_{v_i}^{txt}$:

$$O_{v_i}^{txt} = \left\{t_i \mid t_i^p \in \mathcal{S}^p \cap r_i \in \mathcal{S}^r\right\}. \tag{3}$$

3.1.2 ***Image Decouple.*** This part aims to decouple item image into multiple objects by identifying the image patches associated with each object. We leverage image segmentation techniques [13, 14] to achieve this goal.

Specifically, given the image patch sequence $\mathcal{P}_{v_i} = [p_1, p_2, \cdots, p_m]$, we employ a well-trained Segment Anything Model (SAM) [14] to generate a mask matrices $\mathbf{M}$ for each object, forming the mask set $\mathcal{M}_{v_i}$. These mask matrices are then applied to the original image to obtain different objects, thus forming the image object set $O_{v_i}^{img}$:

$$O_{v_i}^{img} = \{\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_z\} = \{\mathbf{M}_i \odot \mathcal{P}_{v_i} \mid \mathbf{M}_i \in \mathcal{M}_{v_i}\}, \tag{4}$$

$$\mathcal{M}_{v_i} = \{\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_z\} = SAM(\mathcal{P}_{v_i}), \tag{5}$$

where $\mathbf{V}_i \in O_{v_i}^{img}$ denotes the i-th image object of item v_i. Since the total number of image objects is not constant across different items, for modeling convenience, we fix z objects (prioritizing those with more non-zero values in mask matrice for retention).

3.2 Object Identifier

In this subsection, we propose an Object Identifier to determine target objects from the text and image object set ($O_{v_i}^{txt}$ and $O_{v_i}^{img}$). This module consists of two components: (i) *Text Target Selector* and (ii) *Image Target Selector*.

3.2.1 ***Text Target Selector.*** To identify the target object from the text object set $O_{v_i}^{txt}$, an intuitive solution is to leverage large language models (LLMs) [8, 15, 16]. However, this approach encounters some identification failures in our experiments, as evidenced in Sec. 4.4.2. Our analysis reveals that these failures are due to the instruction non-compliance that exists in the LLM's outputs.

The target object identification is fundamentally a discriminative task, requiring a model to assign a deterministic probability to each candidate object. In contrast, LLMs follow a generative paradigm, and their outputs may exhibit unpredictable or inconsistent forms.

In addition, e-commerce and content platforms generate a massive volume of new items daily. The large parameter size of LLMs poses significant challenges in efficiency and cost when processing such vast datasets. To address these issues, we propose the Text Target Selector, a task-specific discriminative model. It applies knowledge distillation based on LLMs to achieve stable output,

lightweight architecture, and competitive performance compared to LLMs.

Concretely, we first drive LLM (LLaMA2) [15] to label text objects based on the text object set $O_{v_i}^{txt}$ and the text sequence $\mathcal{T}_{v_i}$:

$$y_{v_i}^{txt} = LLM([O_{v_i}^{txt} || \mathcal{T}_{v_i}]; P), \tag{6}$$

where P is a LLM prompt, which is detailed in Appendix A.3. $||$ denotes concatenate operation. $y_{(v_i,j)}^{txt} \in \{0, 1\}$ signifies the label of j-th text object $t_j \in O_{v_i}^{txt}$. $y_{(v_i,j)}^{txt} = 1$ indicates that t_j is labeled as the target object, and 0 otherwise (i.e., an auxiliary object). Notably, samples where annotation failed due to LLM instruction non-compliance are discarded.

Thereafter, we develop a selection model to distill the knowledge from LLMs. This model first utilizes a frozen CLIP [17] to obtain the sentence representation $\mathbf{t}_{v_i} \in \mathbb{R}^d$ by encoding $\mathcal{T}_{v_i}$. Given a text object $t_j \in O_{v_i}^{txt}$, we adopt the text embedding layer of CLIP to capture its representation $\mathbf{t}_j \in \mathbb{R}^d$:

$$\mathbf{t}_{v_i} = CLIP_{txt}(\mathcal{T}_{v_i}), \ \mathbf{t}_j = CLIP_{txt}^{emb}(t_j), \tag{7}$$

where $CLIP_{txt}(\cdot)$ and $CLIP_{txt}^{emb}(\cdot)$ denote CLIP's text encoder and text embedding layer, respectively. Additionally, we further incorporate the position information of text object t_j within the text sequence $\mathcal{T}_{v_i}$ to generate the position embedding $\mathbf{p}_j \in \mathbb{R}^d$, thereby capturing sequential information.

Building on these, a trainable selection function $f_{txt}^{sel}(\cdot)$, structured primarily as MLPs (see Appendix A.3), is applied to estimate the probability that each text object in $O_{v_i}^{txt}$ is the target object. These steps are defined as follows:

$$s_j^{txt} = \frac{\exp(p_j^{txt})}{\sum_{t_i \in O_{vi}^{txt}} \exp(p_i^{txt})}, \tag{8}$$

$$p_j^{txt} = f_{txt}^{sel}([\mathbf{t}_{v_i} || \mathbf{t}_j || \mathbf{p}_j]; \theta_{sel}) = MLPs([\mathbf{t}_{v_i} || \mathbf{t}_j || \mathbf{p}_j]), \tag{9}$$

where θ_{sel} is the trainable parameter of $f_{txt}^{sel}(\cdot)$. $s_j^{txt} \in (0, 1)$ is the text target score, which indicates the probability that the text object t_j is the target text object of item v_i.

We optimize the learnable parameters of selection function $f_{txt}^{sel}(\cdot)$ by minimizing the cross-entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -\frac{1}{|\mathcal{V}|} \sum_{v_i \in |\mathcal{V}|} \sum_{j=1}^{|O_{v_i}^{txt}|} y_{(v_i,j)}^{txt} \log(s_j^{txt}), \tag{10}$$

where $y_{(v_i,j)}^{txt}$ denotes the label for the j-th text object of item v_i.

Based on the optimized Text Target Selector, we can perform rapid inference on unannotated item text, such as validation items (see Sec. 4.4.2 of Experiments) and cross-dataset scenarios (see Sec. 4.5 of Experiments), without relying on LLMs.

Next, we generate the representations for text objects, thereby supporting the downstream recommendation task. Specifically, supposing that text object t_j is identified as the target object by Text Target Selector (according to the text target score s_j^{txt}), t_j and its representation $\mathbf{t}_j \in \mathbb{R}^d$ (derived from Eq. (7)) are defined as $t_{v_i}^{tgt}$ and $\mathbf{t}_{v_i}^{tgt}$, respectively. The remaining objects in $O_{v_i}^{txt}$ are defined as auxiliary objects, where the k-th auxiliary object t_k is denoted as t_k^{aux}. Its representation $\mathbf{t}_k \in \mathbb{R}^d$ (driven from Eq. (7)) is renamed as $\mathbf{t}_k^{aux}$. All auxiliary text objects constitute the auxiliary text object set $\mathcal{A}_{v_i}^{txt} = \{t_1^{aux}, t_2^{aux}, \cdots, t_{n-1}^{aux}\}$.

In addition, to enrich the representation of different text objects, we combine a fixed context window with size k and $CLIP_{txt}$ to capture their contextual representation from $\mathcal{T}_{v_i}$:

$$\mathbf{c}_{v_i}^{tgt} = CLIP_{txt}\big(\textsc{Context}(t_{v_i}^{tgt}, \mathcal{T}_{v_i})\big), \tag{11}$$

$$\mathbf{c}_k^{aux} = CLIP_{txt}\big(\textsc{Context}(t_k^{aux}, \mathcal{T}_{v_i})\big), \tag{12}$$

$$\textsc{Context}(t_i, \mathcal{T}_{v_i}) = \Big[t_{i-\lfloor \frac{k-1}{2} \rfloor}, \cdots, t_i, \cdots, t_{i+\lceil \frac{k-1}{2} \rceil}\Big], \tag{13}$$

where $\mathbf{c}_{v_i}^{tgt} \in \mathbb{R}^d$ and $\mathbf{c}_k^{aux} \in \mathbb{R}^d$ are the context representations of target text object $t_{v_i}^{tgt}$ and the k-th auxiliary text object t_k^{aux}.

We then combine the representations of text objects with their context representations to form the final text object representations:

$$\widetilde{\mathbf{t}}_{v_i}^{tgt} = \mathbf{t}_{v_i}^{tgt} + \mathbf{c}_{v_i}^{tgt}, \ \widetilde{\mathbf{t}}_k^{aux} = \mathbf{t}_k^{aux} + \mathbf{c}_k^{aux}, \tag{14}$$

where $\widetilde{\mathbf{t}}_{v_i}^{tgt} \in \mathbb{R}^d$ and $\widetilde{\mathbf{t}}_k^{aux} \in \mathbb{R}^d$ are the final representations of the target text object $t_{v_i}^{tgt}$ and the k-th auxiliary text object t_k^{aux}. For clarity, the representations of all auxiliary text objects form the set $\mathbf{A}_{v_i}^{txt} = \{\widetilde{\mathbf{t}}_1^{aux}, \widetilde{\mathbf{t}}_2^{aux}, \cdots, \widetilde{\mathbf{t}}_{n-1}^{aux}\}$.

3.2.2 ***Image Target Selector.*** In this part, we identify the target object from the image object set $O_{v_i}^{img} = \{\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_z\}$ and refine the representations of different image objects.

Specifically, we first apply the image encoder of CLIP ($CLIP_{img}(\cdot)$) to embed the image objects in $O_{v_i}^{img}$:

$$\mathbf{O}_{v_i}^{img} = \{\mathbf{v}_i \mid \mathbf{v}_i = CLIP_{img}(\mathbf{V}_i), \mathbf{V}_i \in O_{v_i}^{img}\}, \tag{15}$$

where $\mathbf{v}_i \in \mathbb{R}^d$ represents the representation of the i-th object $\mathbf{V}_i \in O_{v_i}^{img}$ and $\mathbf{O}_{v_i}^{img} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_z\}$.

The target text object $\widetilde{\mathbf{t}}_{v_i}^{tgt}$ embodies the unique core content of item v_i, ensure a unique semantic correspondence with its image counterpart. Consequently, we develop an image target score $s_i^{img} \in (-1, 1)$ to identify the target image object by computing the cosine similarity between image object $\mathbf{v}_i$ and target text object $\widetilde{\mathbf{t}}_{v_i}^{tgt}$:

$$\mathbf{S}_{v_i}^{img} = \{s_i^{img} \mid s_i^{img} = Cos(\mathbf{v}_i, \widetilde{\mathbf{t}}_{v_i}^{tgt}), \mathbf{v}_i \in \mathbf{O}_{v_i}^{img}\}, \tag{16}$$

where $\mathbf{S}_{v_i}^{img} = \{s_1^{img}, s_2^{img}, \cdots, s_z^{img}\}$ is the image target score set.

Based on the $\mathbf{S}_{v_i}^{img}$, the representation in $\mathbf{O}_{v_i}^{img}$ with the highest image target score is designated as the target image object $\mathbf{v}_{v_i}^{tgt} \in \mathbb{R}^d$, and the remainder as auxiliary objects. The representation of the i-th auxiliary image object is defined as $\mathbf{v}_i^{aux} \in \mathbb{R}^d$, and all auxiliary objects form the set $\mathbf{A}_{v_i}^{img} = \{\mathbf{v}_1^{aux}, \mathbf{v}_2^{aux}, \cdots, \mathbf{v}_{z-1}^{aux}\}$.

Although we currently distinguish between target and auxiliary image objects based on the target image scores, this approach lacks differentiated learning at the representation level. Image object representations typically encompass multiple attributes—such as color and size—and when an auxiliary object shares certain basic attributes with the target, the distinctive features of the target (e.g., contour and texture) may not be sufficiently emphasized. To better model the representational differences between target and auxiliary image objects, we propose a representation refinement function $f_{img}^{ref}(\cdot)$ to refine their representations:

$$\widetilde{\mathbf{A}}_{v_i}^{aux} = \{\widetilde{\mathbf{v}}_i^{aux} \mid \widetilde{\mathbf{v}}_i^{aux} = f_{img}^{ref}(\mathbf{v}_i^{aux}, \theta_{ref}), \mathbf{v}_i^{aux} \in \mathbf{A}_{v_i}^{aux}\}, \tag{17}$$

$$\widetilde{\mathbf{v}}_{v_i}^{tgt} = f_{img}^{ref}(\mathbf{v}_{v_i}^{tgt}, \theta_{ref}), \ f_{img}^{ref}(\mathbf{x}, \theta_{ref}) = MLPs(\mathbf{x}), \tag{18}$$

where $\widetilde{\mathbf{v}}_{v_i}^{tgt}, \widetilde{\mathbf{v}}_i^{aux} \in \mathbb{R}^d$ are the refined representations of the target image object and the i-th auxiliary image object. θ_{ref} is the trainable parameter of $f_{img}^{ref}(\cdot)$, which is detailed in Appendix A.3.

For the optimization of $f_{img}^{ref}(\cdot)$, we use the representation of target text object $\widetilde{\mathbf{t}}_{v_i}^{tgt}$ as an anchor. The representation of target image object $\widetilde{\mathbf{v}}_i^{tgt}$ is encouraged to be close to the anchor, while distinct from the representations of text/image auxiliary objects. This is achieved by minimizing a contrastive loss $\mathcal{L}_{CL}$:

$$\mathcal{L}_{CL} = -\log\Big[\frac{\exp(cos(\widetilde{\mathbf{v}}_{v_i}^{tgt}, \widetilde{\mathbf{t}}_{v_i}^{tgt}))}{\sum_{\widetilde{\mathbf{v}}_j^{aux} \in \widetilde{\mathbf{A}}_{v_i}^{img}} \exp(cos(\widetilde{\mathbf{v}}_{v_i}^{tgt}, \widetilde{\mathbf{v}}_j^{aux}))} + \frac{\exp(cos(\widetilde{\mathbf{v}}_{v_i}^{tgt}, \widetilde{\mathbf{t}}_{v_i}^{tgt}))}{\sum_{\widetilde{\mathbf{t}}_j^{aux} \in \mathbf{A}_{v_i}^{txt}} \exp(cos(\widetilde{\mathbf{v}}_{v_i}^{tgt}, \widetilde{\mathbf{t}}_j^{aux}))}\Big]. \tag{19}$$

3.3 Downstream Task

This subsection aims to prevent the semantic shift problem by utilizing the identified target and auxiliary objects of modalities, thereby learning unbiased mutimodal representations of items. Owing to the plug-and-play architecture of TargetMR, it can be seamlessly integrated into any backbone MRS to enhance its performance.

3.3.1 ***Target-Aware Learning.*** In this subsection, we propose a Target-Aware Learning mechanism to mitigate the semantic shift problem and learn unbiased multimodal representations of items.

Specifically, given the target object representations ($\widetilde{\mathbf{t}}_{v_i}^{tgt}$ and $\widetilde{\mathbf{v}}_{v_i}^{tgt}$). We adopt them as guidance and incorporate attention mechanism to aggregate auxiliary representation in auxiliary object sets $\mathbf{A}_{v_i}^{txt} = \{\widetilde{\mathbf{t}}_1^{aux}, \widetilde{\mathbf{t}}_2^{aux}, \cdots, \widetilde{\mathbf{t}}_{n-1}^{aux}\}$ and $\widetilde{\mathbf{A}}_{v_i}^{img} = \{\widetilde{\mathbf{v}}_1^{aux}, \widetilde{\mathbf{v}}_2^{aux}, \cdots, \widetilde{\mathbf{v}}_{z-1}^{aux}\}$). The target and aggregated auxiliary representations are combined to obtain the final text and image representations:

$$\mathbf{x}_{v_i}^{txt} = \Big[\widetilde{\mathbf{t}}_{v_i}^{tgt} \,||\, \sum_{i=1}^{n-1} \beta_i^t \widetilde{\mathbf{t}}_i^{aux}\Big], \quad \mathbf{x}_{v_i}^{img} = \Big[\widetilde{\mathbf{v}}_{v_i}^{tgt} \,||\, \sum_{i=1}^{z-1} \beta_i^g \widetilde{\mathbf{v}}_i^{aux}\Big], \tag{20}$$

$$\beta_i^t = \frac{\exp\left(\mathbf{W}_t[\widetilde{\mathbf{t}}_{v_i}^{tgt} || \widetilde{\mathbf{t}}_i^{aux}\right)}{\sum_{i=1}^{n-1} \exp\left(\mathbf{W}_t[\widetilde{\mathbf{t}}_{v_i}^{tgt} || \widetilde{\mathbf{t}}_i^{aux}\right)}, \beta_i^g = \frac{\exp\left(\mathbf{W}_g[\widetilde{\mathbf{v}}_{v_i}^{tgt} || \widetilde{\mathbf{v}}_i^{aux}\right)}{\sum_{i=1}^{z-1} \exp\left(\mathbf{W}_g[\widetilde{\mathbf{v}}_{v_i}^{tgt} || \widetilde{\mathbf{v}}_i^{aux}\right)}, \tag{21}$$

where $\mathbf{W}_t, \mathbf{W}_g \in \mathbb{R}^{2d \times 1}$ are trainable parameters.

Similar to [3], we apply the sum of $\mathbf{x}_{v_i}^{txt}$ and $\mathbf{x}_{v_i}^{img}$, denoted as $\mathbf{x}_{v_i}^{com} \in \mathbb{R}^d$, to supplement latent positive interactions for users:

$$\widetilde{\mathcal{N}}_{u_i} = \{\mathcal{N}_{u_i}, \overline{\mathcal{N}}_{u_i}\}, \ \overline{\mathcal{N}}_{u_i} = \{v_k \mid v_j \in \mathcal{N}_{u_i}, Cos(\mathbf{x}_{v_j}^{com}, \mathbf{x}_{v_k}^{com}) > \eta\}, \tag{22}$$

where $\mathcal{N}_{u_i}$ and $\widetilde{\mathcal{N}}_{u_i}$ denote the neighbor set and the extended neighbor set of user u_i, respectively. Correspondingly, the extended user-item bipartite graph is denoted as $\widetilde{\mathcal{G}}$.

3.3.2 ***Backbone MRS.*** The framework of TargetMR is model-agnostic, enabling seamless compatibility with any MRS. Concretely, given the text feature set $\mathcal{X}^{txt} = \{\mathbf{x}_{v_i}^{txt} \mid v_i \in \mathcal{V}\}$, image feature set $\mathcal{X}^{img} = \{\mathbf{x}_{v_i}^{img} \mid v_i \in \mathcal{V}\}$, and the extended user-item bipartite graph $\widetilde{\mathcal{G}}$, we minimize the loss $\mathcal{L}_{BCK}$ derived from backbone model Backbone$(\cdot)$ to optimize the model's parameters:

$$\mathcal{L}_{BCK} = \textsc{Backbone}(\mathcal{X}^{txt}, \mathcal{X}^{img}, \widetilde{\mathcal{G}}; \theta_B), \tag{23}$$

where θ_B is the trainable parameters of the backbone model.

Overall, the Algorithm 1 in Appendix. A.3 outlines the overall algorithm of TargetMR.

4 Experiments

4.1 Experimental Settings

Datasets. Following [1, 10, 18], we select three types benchmark datasets from Amazon, including (i) Clothing, (ii) Sports and (iii) Electronics. A detailed introduction is provided in the Appendix A.1.

Evaluation Protocols. Following [3, 10, 11, 19], we employ two widely-adopted evaluation metrics, Recall@N and NDCG@N (where $N \in \{10, 20\}$ in this paper), to assess the performance of all models. The reported results represent the average performance across all users in the testing set.

Backbone Models. We select several prominent MRSs of recent years as backbone models, including **BM3 (2023)** [20], **FREEDOM (2023)** [20], **MGCN (2023)** [1], **LGMRec (2024)** [10], and **MENTOR (2025)** [18]. Detailed descriptions of these models can be found in Appendix A.2.

4.2 Overall Performance

To validate the effectiveness of TargetMR, we integrate it with five backbone MRSs and conduct performance comparison on three benchmark datasets. Table 1 presents the performance of these backbone MRSs both without and with TargetMR, respectively. The term *Imp* denotes the percentage of performance improvement brought by TargetMR to these backbone MRSs. Based on the experimental results, the following conclusions can be drawn.

After integrating TargetMR, these backbone MRSs achieved significant improvements in both Recall and NDCG. This demonstrates the effectiveness of TargetMR and its model-agnostic design. Specifically, TargetMR brings average performance gains of 17.45%, 8.59% and 10.74% across the Clothing, Sports and Electronics datasets for multiple backbone MRSs. These improvements can be attributed to two key components of TargetMR: (i) The Object Disentangler effectively decouples item text and image into multiple objects, thereby providing effective support for target object identification. (ii) The Object Identifier precisely identifies the target objects and auxiliary objects within multiple modality objects, facilitating the learning of unbiased modality representations.

By analyzing the performance of TargetMR's on different backbone MRSs, we observe the highest Recall and NDCG are achieved by MENTOR equipped with TargetMR. This superior performance stems from MENTOR's incorporation of self-supervised learning to align features across different modalities, establishing a solid baseline performance. The integration of TargetMR further enhanced this model, resulting in state-of-the-art performance.

4.3 Ablation Study

To investigate the impact of different components in TargetMR, we design multiple variant models through systematic component ablation. Notably, to ensure consistent evaluation standards, MENTOR is employed as the backbone MRS of TargetMR in this subsection. These variant models are defined as follows:

Table 1: Overall performance of all backbone models and their enhanced versions with TargetMR on three benchmark datasets. The highest metric within each block (corresponding to each backbone model) is indicated in bold, and the percentage of performance improvement is highlighted with a gray background. The † indicates statistically significant performance improvements with p-value < 0.01.

MODELS	Clothing				Sports				Electronics			
	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
BM3 (2023) [21]	0.0450	0.0669	0.0243	0.0295	0.0656	0.0980	0.0355	0.0438	0.0437	0.0648	0.0247	0.0302
+TargetMR	**0.0580**†	**0.0835**†	**0.0328**†	**0.0396**†	**0.0737**†	**0.1078**†	**0.0409**†	**0.0498**†	**0.0461**†	**0.0682**†	**0.0260**†	**0.0317**†
Imp(%)	28.89%	24.81%	34.98%	34.24%	12.35%	10.00%	15.21%	13.70%	5.49%	5.25%	5.26%	4.97%
FREEDOM (2023) [20]	0.0629	0.0941	0.0341	0.0420	0.0717	0.1089	0.0385	0.0481	0.0398	0.0603	0.0218	0.0271
+TargetMR	**0.0696**†	**0.1025**†	**0.0387**†	**0.0465**†	**0.0761**†	**0.1137**†	**0.0415**†	**0.0519**†	**0.0450**†	**0.0669**†	**0.0254**†	**0.0312**†
Imp(%)	10.65%	8.93%	13.49%	10.71%	6.14%	4.41%	7.79%	7.9%	13.07%	10.95%	16.51%	15.13%
MGCN (2023) [1]	0.0641	0.0945	0.0347	0.0428	0.0729	0.1106	0.0397	0.0496	0.0442	0.0650	0.0246	0.0302
+TargetMR	**0.0711**†	**0.1040**†	**0.0388**†	**0.0473**†	**0.0771**†	**0.1140**†	**0.0424**†	**0.0517**†	**0.0468**†	**0.0690**†	**0.0264**†	**0.0329**†
Imp(%)	10.92%	10.05%	11.82%	10.51%	5.76%	3.07%	6.80%	4.23%	5.88%	6.15%	7.32%	8.94%
LGMRec (2024) [10]	0.0555	0.0828	0.0302	0.0371	0.0720	0.1068	0.0390	0.0480	0.0408	0.0604	0.0226	0.0277
+TargetMR	**0.0688**†	**0.1003**†	**0.0376**†	**0.0453**†	**0.0781**†	**0.1148**†	**0.0435**†	**0.0526**†	**0.0452**†	**0.0671**†	**0.0253**†	**0.0311**†
Imp(%)	23.96%	21.35%	24.50%	22.10%	8.47%	7.49%	11.54%	9.58%	10.78%	11.09%	11.95%	12.27%
MENTOR (2025) [18]	0.0668	0.0989	0.0360	0.0443	0.0763	0.1139	0.0409	0.0511	0.0439	0.0655	0.0244	0.0300
+TargetMR	**0.0757**†	**0.1078**†	**0.0410**†	**0.0491**†	**0.0832**†	**0.1221**†	**0.0458**†	**0.0558**†	**0.0509**†	**0.0751**†	**0.0286**†	**0.0348**†
Imp(%)	13.32%	9.00%	13.89%	10.84%	9.04%	7.20%	11.98%	9.20%	15.95%	14.66%	17.21%	16.00%

Table 2: Ablation Study on TargetMR.

Variant Models	Clothing		Sports		Electronics	
	R@20	N@20	R@20	N@20	R@20	N@20
w/o POS	0.1038	0.0470	0.1176	0.0526	0.0716	0.0333
w/o SYP	0.1045	0.0474	0.1188	0.0533	0.0724	0.0337
w/o ITS	0.1025	0.0463	0.1168	0.0516	0.0713	0.0331
w/o RRF	0.1034	0.0466	0.1174	0.0522	0.0720	0.0334
w/o TAL	0.1007	0.0453	0.1152	0.0512	0.0684	0.0315
TargetMR	**0.1078**	**0.0491**	**0.1221**	**0.0558**	**0.0751**	**0.0348**

- **w/o POS:** We remove the POS tagging in text segmentation, relying solely on syntactic parsing to obtain the tokens corresponding to text objects.
- **w/o SYP:** We remove the syntactic parsing in text segmentation, relying solely on POS tagging to obtain the tokens corresponding to text objects.
- **w/o ITS:** We remove the Image Target Selector and directly feed the original image features into backbone MRS.
- **w/o RRF:** We remove the representation refinement function $f_{img}^{ref}(\cdot)$ in Image Target Selector.
- **w/o TAL:** We remove the Target-Aware Learning and directly feed the representations of target text and image objects into backbone MRS.

Table 2 shows the performance of TargetMR and multiple variant models. Based on these results, we draw the following conclusions: (i) TargetMR shows performance degradation when ablating both POS tagging and syntactic parsing in text decoupling, i.e., *w/o POS* and *w/o SYP*, underscoring their combined contribution to accurately decoupling item text. (ii) Removing the Image Target Selector (*w/o ITS*) or its representation refinement function $f_{img}^{ref}(\cdot)$ (*w/o RRF*) leads to performance degradation. These demonstrate the importance of identifying the target image object as well as refining its representation. (iii) The removal of Target-Aware Learning, i.e., *w/o TAL*, leads to performance degradation in TargetMR. The reason is that the target-guided auxiliary representation serves as an effective complement to the target object representation.

4.4 Results Visualization and Multi-LLM Distillation of TargetMR

4.4.1 ***The results visualization of TargetMR.*** In this subsection, we perform a visual analysis to intuitively investigate the effectiveness of TargetMR. Specifically, using the Clothing dataset as an example, we first extract the sum of text and image embeddings for all items as their representations via the optimized MENTOR and MENTOR+TargetMR, respectively. We then randomly select two items with target objects "T-shirt" and "Shorts", respectively. Based on their representations, we retrieve the top-200 most similar items for each of them using cosine similarity, forming two item groups. Finally, we apply t-SNE to visualize the representations of all items within these two groups. The experimental results are presented in Figure 3.

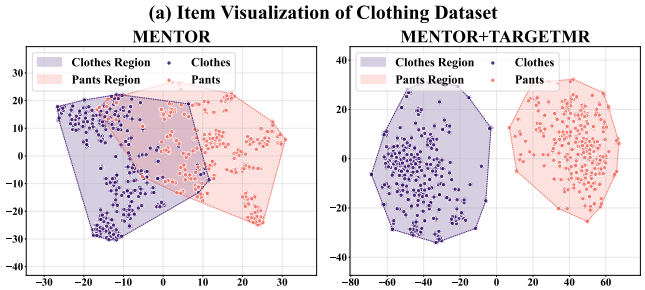

Figure 3: The t-SNE visualization for item representations of two item groups generated by MENTOR and MENTOR+TargetMR on Clothing Datasets.

Figure 3 reveals that the item representations generated by MENTOR exhibit significant overlap between the two item groups in

the Clothing dataset. This occurs because the multimodal representations of "T-shirt" and "Shorts" often include each other as auxiliary objects. However, MENTOR fails to distinguish the target object and auxiliary objects, resulting in a misleading overlap of representations between the two item groups. In contrast, MENTOR+TargetMR significantly reduces the representation overlap between the two item groups. These results demonstrate the effectiveness of TargetMR in identifying target objects, thereby learning unbiased item representations. Further experiments on the Sports and Electronics datasets support this conclusion, as demonstrated in Appendix A.4.

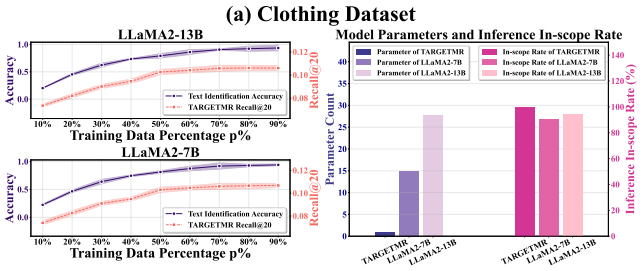

Figure 4: The identification accuracy and Recall@20 of TargetMR based on two LLMs (line chart). The efficiency comparison between TargetMR and two LLMs (bar chart).

4.4.2 ***The Effect of TargetMR in Distilling Diverse LLMs.*** This subsection evaluates the performance of TargetMR distilled from different LLMs. Specifically, we create two training datasets by annotating all item texts in the Clothing dataset with LLaMA2-7B and LLaMA2-13B, respectively. Samples with failed annotations due to the LLM's failure to follow instructions are removed from each dataset. For each dataset, the p% of the data is used for training TargetMR, and the remaining (100-p)% is held out for validation (metric is accuracy). Concurrently, we record the recommendation performance (Recall@20) of TargetMR+MENTOR in the corresponding state. As shown in Figure 4 (line chart), the accuracy of TargetMR and Recall@20 of TargetMR + MENTOR increase monotonically with the percentage of training data ratio (p%). At p=80%, accuracy surpasses 90%, indicating that TargetMR achieves performance comparable to the LLMs.

In addition, Figure 4 (bar chart) compares the parameter counts of all models (expressed as relative multiples to TargetMR) and their inference in-scope rates based on validation data (20% of training data). The inference in-scope rate is defined as the proportion of the inferred target text object for v_i that belongs to the text object set $O_{v_i}^{txt}$. It is evident that TargetMR has significantly fewer parameters than the LLMs while maintaining a 100% inference in-scope rate, highlighting its lightweight and stability. Appendix A.4 provides the experiments on more datasets.

4.5 The Cross-dataset Transferability

The optimization of TargetMR relies on item's multimodal content and LLM. Moreover, in recommendation scenarios, item's multimodal content generally exhibit cross-dataset relevance, endowing TargetMR with cross-dataset transferability. To validate this, we select MENTOR as backbone MRS and optimize TargetMR on source dataset and directly performing inference on target dataset, i.e., SD → TD. These inference results, i.e., the identified target objects and auxiliary objects, are leveraged to optimize the backbone MRS and evaluate its performance. Tabel 3 demonstrates the experimental results. Obviously, Clothing → Sports and Sports → Clothing yields the satisfied performance, whereas Sports/Clothing → Electronics yield poor performance. The reason is that items of Clothing and Sports share overlapping categories (e.g., clothing, shoes, and pants), while Electronics contains items from a distinct domain. Additional experiments involving more diverse settings are provided in Appendix A.4.

Table 3: The cross-dataset transferability experiments for TargetMR, where SD and TD denote the source and target dataset, respectively. The overall highest metric values are shown in bold.

SD → TD	R@10	R@20	N@10	N@20
Sports → Sports	**0.0832**	**0.1221**	**0.0458**	**0.0558**
Clothing → Sports	0.0815	0.1200	0.0445	0.0543
Electronics → Sports	0.0779	0.1168	0.0424	0.0526
Clothing → Clothing	**0.0757**	**0.1078**	**0.0410**	**0.0491**
Sports → Clothing	0.0722	0.1057	0.0394	0.0478
Electronics → Clothing	0.0684	0.1021	0.0376	0.0458
Electronics → Electronics	**0.0509**	**0.0751**	**0.0286**	**0.0348**
Sports → Electronics	0.0460	0.0683	0.0258	0.0315
Clothing → Electronics	0.0458	0.0681	0.0260	0.0316

4.6 Hyperparameter Analysis and Case Study

4.6.1 ***Hyperparameter Analysis.*** In this subsection, we conduct sensitivity analyses of the context window size k (Eq. (13)) in Text Target Selector and the cosine similarity threshold η (Eq. (22)) in Target-Aware Learning, as shown in Figure 5. We observe that values of k in the set $\{8, 16\}$ and values of η in the set $\{0.5, 0.7\}$ yield superior performance. This is because an overly large k introduces noise into the textual context information of text objects, while an overly small k provides insufficient context information; conversely, an overly large η fails to capture sufficient latent positive interactions, whereas an overly small η admits noisy signals. Extensive experiments on more datasets are demonstrated in Appendix A.4.

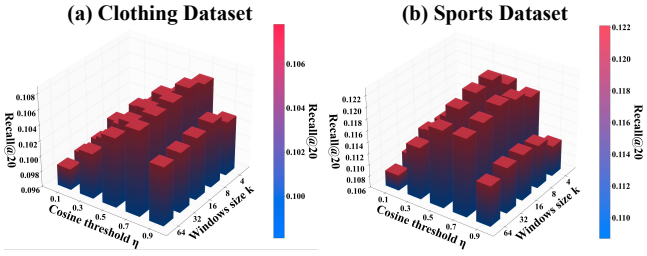

Figure 5: Hyperparameter analysis on cosine threshold η and windows size k in terms of Clothing and Sports datasets.

4.6.2 ***Case Study.*** To investigate the interpretability of TargetMR, we conducte a case study as illustrated in Figure 6. We randomly select an item (v_{6096}) from the Clothing dataset and identify its

target users with v_{6096} in their top-20 recommendation lists. We then perform recommendation for v_{6096} using the optimized model MENTOR and MENTOR+TargetMR, respectively. Additionally, we record the target objects identified by TargetMR in both the image and text modalities. Meanwhile, we record the change in cosine similarity between each image object and the target text object, before and after they are processed by the representation refinement function $f_{img}^{ref}(\cdot)$ in Eq. (18). For instance, $cos_{(4,1)}^{(img,txt)}$ denotes the similarity between the fourth image object and the target text object (i.e., the first text object). Experimental results show that TargetMR+MENTOR successfully recommended v_{6096} to three target users, whereas MENTOR only achieved one. Analysis shows that the superior performance of MENTOR+TargetMR stems from its ability to recognize the target objects—"sleeves"—in both modalities and refine the representations of image objects (the cosine similarity between the target image object and target text objects is improved). This case demonstrates the effectiveness of TargetMR in addressing the semantic shift problem.

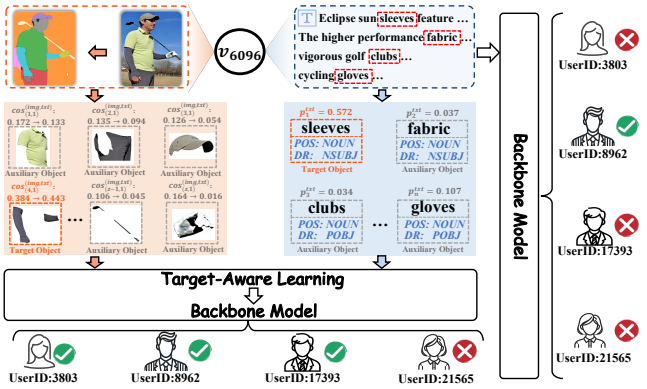

Figure 6: Case study of TargetMR on a random item v_{6096} from Clothing dataset, where the POS and DR denote par-of-speech and dependency relation, respectively.

5 Related Work

• **Multimodal Recommendation.** The primary motivation behind multimodal recommender systems (MRSs) is to enhance item representation by integrating their multimodal content, thereby alleviating the cold-start problem [22, 23]. Early approaches [4, 24] predominantly relied on linear combination or attention mechanisms [25] to fuse multimodal representations with item ID embeddings. For example, VBPR [4] and MRG [24] apply concatenate operation and attention mechanism to integrate item's multimodal features, respectively. Following the widespread adoption of graph neural networks (GNNs), numerous studies [1, 3, 10, 11, 26–30] leverage them to capture high-order user-item interactions. For instance, MGCN [1] develops various user-item bipartite graph based on item's multimodal content. LATTICE [3] proposes to model the relationships between items by building the item-item graph, thus enriching item representations. DRAGON [26] learns dual user-item representations via homogeneous graph construction to strengthen dyadic relations. Building on these efforts, self-supervised learning (SSL) has been widely integrated into subsequent studies [18, 21, 31–33] to improve recommendation performance. BM3 [21] employs a latent contrastive learning framework to achieve graph reconstruction from multi-view data and modality alignment. MMSSL [31] employs a modality-aware adversarial architecture to disentangle common and specific user preferences via cross-modal contrastive learning. MENTOR [18] achieves alignment across different modalities, preserves historical interaction information, and augments feature representations.R^{2}R propose to review and rewrite multimodal features based on information consensus theory and a latent mapping model, thus improving modality quality.

• **Knowledge Distillation for Recommendation.** Knowledge distillation (KD) techniques [34–36] have been widely applied in recommender systems. Early efforts [37–39] primarily employ knowledge distillation to maintain model performance while reducing the parameters, thereby enabling efficient online inference. For instance, RD [37] proposes a small student ranking model that integrates training data with supervisory signals from a teacher ranking model to achieve efficient online item ranking. DE-RRD [38] jointly learns from the latent knowledge encoded in the teacher model and its predictions, achieving superior performance and faster inference. The rapid advancement of Large Language Models (LLMs) has spurred the adoption of knowledge distillation in recommender systems [6, 7, 40–43], enabling them to leverage the rich knowledge embedded within LLMs for enhanced performance. For example, SLMRec [40] leverages a simple knowledge distillation strategy combined with post-training efficiency techniques to achieve robust recommendation with a lightweight architecture. RLMRec [7] introduces model-agnostic framework that amalgamates representation learning with LLMs, aimed at deciphering the nuanced semantic dimensions of user behaviors and preferences. ALKDRec [43] efficiently distills knowledge from LLMs using a small subset of predictions and an active learning strategy for theoretically-grounded effectiveness.

6 Conclusion

This paper presents TargetMR, a model-agnostic framework designed to address the semantic shift problem in MRSs. The framework employs an Object Disentangler to decouple multimodal content into multiple objects and an Object Identifier to figure out target objects. Extensive experiments demonstrate that TargetMR consistently enhances the performance of various backbone MRSs. In our future work, we intent to explore the high-order correlations among modality objects by constructing a topology between target and auxiliary objects.

Acknowledgments

This work was partially supported by NSF China (No.T2421002, 62272301, 92579211, 62525209).

References

[1] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6576–6585, 2023.

[2] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1807–1811, 2022.

[3] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3872–3880, 2021.

[4] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[5] Qiang Liu, Shu Wu, and Liang Wang. Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pages 841–844, 2017.

[6] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, pages 806–815, 2024.

[7] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Representation learning with large language models for recommendation. In *Proceedings of the ACM web conference 2024*, pages 3464–3475, 2024.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[10] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. Lgmrec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8454–8462, 2024.

[11] Gu Tang, Jinghe Wang, Xiaoying Gan, Bin Lu, Ze Zhao, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. R2mr: Review and rewrite modality for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, page 1337–1348, 2025.

[12] Duygu Altinok. *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*. Packt Publishing Ltd, 2021.

[13] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[15] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[16] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[18] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C-H Ngai. Mentor: Multi-level self-supervised learning for multimodal recommendation. *arXiv preprint arXiv:2402.19407*, 2024.

[19] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

[20] Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 935–943, 2023.

[21] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, page 845–854, 2023.

[22] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, 2002.

[23] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1073–1082, 2019.

[24] Quoc-Tuan Truong and Hady Lauw. Multimodal review generation for recommender systems. In *The World Wide Web Conference*, page 1864–1874, 2019.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[26] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In *ECAI 2023*, pages 3123–3130. IOS Press, 2023.

[27] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.

[28] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2021.

[29] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. Mgat: Multimodal graph attention network for recommendation. *Information Processing Management*, 57(5):102277, 2020.

[30] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Mind individual information! principal graph learning for multimedia recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13096–13105, 2025.

[31] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 790–800, 2023.

[32] Chunyu Wei, Jian Liang, Di Liu, and Fei Wang. Contrastive graph structure learning via information bottleneck for recommendation. *Advances in Neural Information Processing Systems*, pages 20407–20420, 2022.

[33] Junliang Yu, Hongzhi Yin, Jundong Li, Qinyong Wang, Nguyen Quoc Viet Hung, and Xiangliang Zhang. Self-supervised multi-channel hypergraph convolutional network for social recommendation. In *Proceedings of the web conference 2021*, pages 413–424, 2021.

[34] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[35] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019.

[36] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

[37] Jiaxi Tang and Ke Wang. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2289–2298, 2018.

[38] SeongKu Kang, Junyoung Hwang, Wonbin Kweon, and Hwanjo Yu. De-rrd: A knowledge distillation framework for recommender system. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 605–614, 2020.

[39] Jae-woong Lee, Minjin Choi, Jongwuk Lee, and Hyunjung Shim. Collaborative distillation for top-n recommendation. In *2019 IEEE international conference on data mining (ICDM)*, pages 369–378. IEEE, 2019.

[40] Wujiang Xu, Qitian Wu, Zujie Liang, Jiaojiao Han, Xuying Ning, Yunxiao Shi, Wenfang Lin, and Yongfeng Zhang. Slmrec: Distilling large language models into small for sequential recommendation. *arXiv preprint arXiv:2405.17890*, 2024.

[41] Zhixuan Chu, Hongyan Hao, Xin Ouyang, Simeng Wang, Yan Wang, Yue Shen, Jinjie Gu, Qing Cui, Longfei Li, Siqiao Xue, et al. Leveraging large language models for pre-trained recommender systems. *arXiv preprint arXiv:2308.10837*, 2023.

[42] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM conference on recommender systems*, pages 1007–1014, 2023.

[43] Yingpeng Du, Zhu Sun, Ziyan Wang, Haoyan Chua, Jie Zhang, and Yew-Soon Ong. Active large language model-based knowledge distillation for session-based recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11607–11615, 2025.

A Supplementary Materials

A.1 Datasets

To validate the effectiveness of TargetMR across diverse recommendation scenarios, we select three types benchmark datasets from Amazon, including (i) Clothing, Shoes and Jewelry, (ii) Sports and Outdoors, and (iii) Electronics Products. For ease of reference, these datasets are denoted as *Clothing*, *Sports*, and *Electronics*, respectively. Detailed statistical information of these datasets are presented in Table 4.

Table 4: Statistics of the three experimental datasets.

Dataset	#User	#Item	#Interaction	#Sparsity
Clothing	39,387	23,033	278,677	99.97%
Sports	35,598	18,357	296,337	99.95%
Electronics	192,403	63,001	1,689,188	99.06%

A.2 Backbone Models.

To validate the effectiveness and generalizability of TargetMR, we integrate it with several prominent backbone MRSs. These models are introduced as follows:

- **BM3 (2023)** [21] constructs multi-view item representations through dropout and incorporates self-supervised learning to enhance recommendation performance.
- **FREEDOM (2023)** [20] freezes item-item structure before training and introduces a degree-sensitive edge pruning technique to remove the noise in the user-item interactions.
- **MGCN (2023)** [1] propose a multi-view information encoder to separately learning collaborative and semantical signals. In addition, it further develops a self-supervised auxiliary task to enhance user representations.
- **LGMRec (2024)** [10] enhances recommendation performance by constructing both local graph and global hypergraph graph to supplement item and user representations from different perspectives.
- **MENTOR (2025)** [18] develops a cross-modal alignment task and a general feature enhancement task to alleviate data sparsity problem.

A.3 Parameters Settings and Algorithm

TargetMR is implemented using PyTorch. In TargetMR, the CLIP within the Object Identifier employs the ViT-B/32 version [17]. Meanwhile, the specific structures of the text selection function $f_{txt}^{sel}(\cdot)$ and the image refinement function $f_{img}^{ref}(\cdot)$ in the Object Identifier are detailed in Table 5. Based on experimental results on the validation set, the learning rate and batch size for the aforementioned two functions are set to $1e-3$ and 512, respectively. All learnable parameters in TargetMR are initialized using the Xavier Normal method and optimized with the Adam optimizer. In addition, the configuration of the LLM prompt in Eq. (6) is described in Figure 8.

The algorithm of TargetMR mainly comprises four steps: (i) Decoupling item text and images via the Object Disentangler; (ii) Optimizing the Text Target Selector via knowledge distillation to identify target and auxiliary text objects; (iii) Identifying target image object via Image Target Selector and optimizing its representation refinement function $f_{img}^{ref}(\cdot)$ to learn better image object representations; (iv) Applying Target-Aware Learning mechanism to learn items' multimodal features and optimizing Backbone MRS. The complete procedure is described as Algorithm 1.

Table 5: The detailed structures of text selection function $f_{txt}^{sel}(\cdot)$ and image refinement function $f_{img}^{ref}(\cdot)$.

Structure of text selection function $f_{txt}^{sel}(\cdot)$			Structure of image refinement function $f_{img}^{ref}(\cdot)$		
Num.	Layer	Weight Size	Num.	Layer	Weight Size
1	MLP	(1536, 512)	1	MLP	(512, 512)
×6	MLP	(512,512)	×8	MLP	(512,512)
	LayerNorm	-		LayerNorm	-
	ResidualAdd	-		ResidualAdd	-
1	MLP	(512,1)			
1	Softmax	-			

Algorithm 1 Implementation of TargetMR

Input:
The text $\mathcal{T}_{v_i}$ and image $\mathcal{P}_{v_i}$ of item v_i; the user-item bipartite graph $\mathcal{G}$; the text object selection function $f_{txt}^{sel}(\cdot)$ and representation refinement function $f_{img}^{ref}(\cdot)$; the Target-Aware Learning function $f_{tal}(\cdot)$; and the backbone MRS $\mathcal{F}(\cdot)$;
Output: The optimized parameters θ_{sel}, θ_{ref}, θ_{tal}, and θ_B for $f_{txt}^{sel}(\cdot)$, $f_{img}^{ref}(\cdot)$, $f_{tal}(\cdot)$, and $\mathcal{F}(\cdot)$, respectively.
Decoupling text and image via Object Disentangler (Eq.(1-5)) and labeling text target object via LLM (Eq. (6)).
for every epoch **do**
  Optimizing θ_{sel} for $f_{txt}^{sel}(\cdot)$ via minimizing $\mathcal{L}_{CE}$ through Eq. (10);
  Forming the representation of target text object $\widetilde{\mathbf{t}}_{v_i}^{tgt}$ and the representation set of auxiliary text objects $\mathbf{A}_{vi}^{txt}$ via Eq. (11-14);
  Early stopping strategy;
end
for every epoch **do**
  Identifying target image object via cosine similarity (Eq. (16));
  Refining target and auxiliary image objects via $f_{img}^{ref}(\cdot)$ (Eq.(17-18));
  Optimizing θ_{ref} for $f_{img}^{ref}(\cdot)$ via minimizing $\mathcal{L}_{CL}$ in Eq. (19);
  Early stopping strategy;
end
for every epoch **do**
  Learning the final text and image representations, i.e.,$\mathbf{x}_{v_i}^{txt}$ and $\mathbf{x}_{v_i}^{img}$, through the Target-Aware Learning (Eq. (20-21));
  Supplementing latent positive interactions for user-item bipartite graph $\mathcal{G}$ via Eq. (22);
  Optimizing θ_{tal} and θ_B for the Target-Aware Learning function and backbone MRS $\mathcal{F}(\cdot)$ via minimizing $\mathcal{L}_{BCK}$ (Eq. (23));
  Early stopping strategy;
end

A.4 Supplementary Experiments

- **The Results Visualization of TargetMR.** To further investigate the effectiveness of TargetMR in identifying target objects within multimodal features, we extend the visual analysis from Sec. 4.4.1 to the Sports and Electronics datasets. For this experiment,

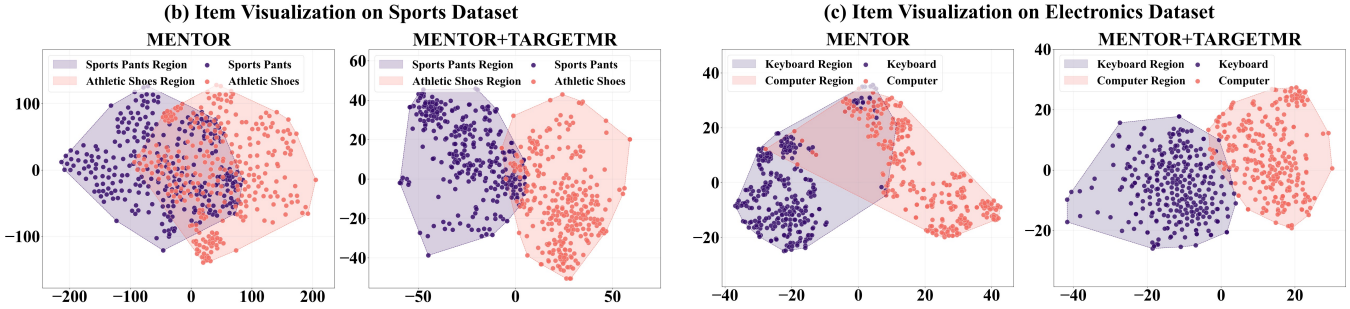

Figure 7: The t-SNE visualization for item representations of two item groups generated by MENTOR and MENTOR+TargetMR on Electronics datasets.

[Instruction]: You are a helpful, respectful and honest assistant.

[Prompt]: As a natural language processing expert, given a item description from the e-commerce domain and a set $O_{v_i}^{txt}$ containing different objects of the item, please select the element from set $O_{v_i}^{txt}$ that best represents the core content of the item based on the overall description.

Requirement: Directly return the element from $O_{v_i}^{txt}$ that you consider most representative of the item's core content.

Item description: Propper men's BDU coat . The coat pairs with the BDU trouser to create a rugged uniform. This four-pocket coat drain holes in the bellowed pockets for quick drying and fused pocket flaps and collar for a clean, pressed look . Other features include a chest pen pocket, two-piece double-reinforced elbows, …

Set $O_{v_i}^{txt}$:{coat, trouser, uniform, holes, pocket, elbows}

[Answer]: coat

Figure 8: The configuration of LLM prompt used in Text Target Selector (Sec.3.2.1).

two item groups are constructed in each dataset by randomly selecting two items with the following target objects: ("Sports Pants", "Athletic Shoes") from the Sports dataset, and ("Keyboard", "Computer") from the Electronics dataset. For the two item groups of each dataset, the multimodal content of items in each group often include auxiliary objects from the other group. The experimental results are shown in Figure 7. We observe that TargetMR effectively reduces the overlap between the representations of the two item groups in both datasets. This is because TargetMR successfully distinguishes between target and auxiliary objects, thereby learning unbiased item representations.

• **The Effect of TargetMR in Distilling Diverse LLMs.** In this part, we extend the experiments of Sec. 4.4.2 on the Sports and Electronics datasets, as demonstrated in Figure 9. The results demonstrate that as the training data ratio increases, the TargetMR's accuracy in identifying text target object gradually improves, eventually converging with LLM performance. Meanwhile, the recommendation performance of TargetMR+MENTOR also progressively improves with the expansion of the training data ratio. In addition, the bar plot in Figure 9 shows that TargetMR achieves a 100% inference in-scope rate due to its discriminative structure, while using significantly fewer parameters than the LLMs. These findings are consistent with those obtained on the Clothing dataset (Sec. 4.4.2), highlighting the lightweight architecture and stability of TargetMR.

Table 6: The cross-dataset transferability experiments for TargetMR, where SD and TD denote the source and target dataset, respectively. The overall highest metric values are shown in bold.

SD → TD	R@10	R@20	N@10	N@20
Sports → Sports	0.0832	0.1221	0.0458	0.0558
Clothing+Sports → Sports	**0.0837**	**0.1229**	**0.0459**	**0.0560**
Electronics+Sports → Sports	0.0819	0.1216	0.0448	0.0551
Clothing + Electronics → Sports	0.0811	0.1206	0.0440	0.0543
Clothing → Clothing	0.0757	0.1078	0.0410	0.0491
Sports+Clothing → Clothing	**0.0760**	**0.1081**	**0.0413**	**0.0494**
Electronics+Clothing → Clothing	0.0733	0.1060	0.0401	0.0480
Sports + Electronics → Clothing	0.0714	0.1045	0.0389	0.0473
Electronics → Electronics	**0.0509**	**0.0751**	**0.0286**	**0.0348**
Sports + Electronics → Electronics	0.0504	0.0742	0.0283	0.0345
Clothing + Electronics → Electronics	0.0502	0.0743	0.0281	0.0343
Sports + Clothing → Electronics	0.0454	0.0665	0.0253	0.0312

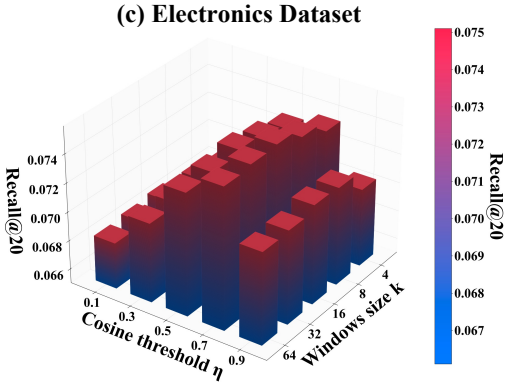

Figure 10: Hyperparameter analysis of cosine threshold η and windows size k on Electronics dataset.

• **The Cross-Dataset Transferability of TargetMR.** To explore the performance of TargetMR in more complex cross-dataset transfer scenarios, we extend the experiments in Sec. 4.5. The results are demonstrated in Table 6. We observe that both Clothing + Sports → Sports and Sports + Clothing → Clothing outperform their non-transfer counterparts (i.e., Sports → Sports and Clothing → Clothing). This outcome arises from the fact that the Clothing

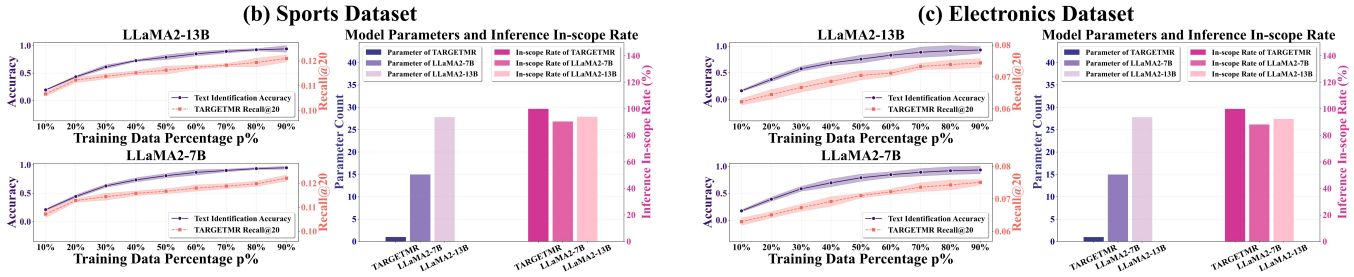

Figure 9: The identification accuracy and Recall@20 of TargetMR based on two LLMs (line chart). The efficiency comparison between TargetMR and two LLMs (bar chart).

and Sports share numerous item categories. Their integration provides TargetMR with more extensive training data, enabling it to infer more effectively on samples for which the LLM fails due to the issue of instruction non-compliance. In contrast, performance degradation occurs when Electronics is used as the target dataset (compared to Electronics → Electronics) in cross-dataset scenarios. This result stems from the fact that item categories of Electronics exhibits a larger semantic gap with the other datasets.

• **Hyperparameter Analysis.** We supplement the hyperparameter experiment on the context windows size k and the cosine similarity threshold η based on Electronics dataset, as shown in Figure 10. The experimental results are consistent with the conclusions presented in Sec. 4.6.1. The optimal performance is achieved with $k = 16$ and $n = 0.7$. This can be attributed to the fact that an excessively large k introduces noise into the text context of modality objects, while an excessively small k leads to insufficient information. In addition, an overly high η fails to capture sufficient latent positive interactions, whereas an overly low η introduces noisy signals.