# The Supplementary Material for RecGOAT

Anonymous Author(s)

## Abstract

Multimodal recommendation systems typically integrates user behavior with multimodal data from items, thereby capturing more accurate user preferences. Concurrently, with the rapid advancement of large models (LMs), recent research has focused on leveraging large language models (LLMs), large vision models (LVMs), and multimodal large language models (MLLMs) to enhance the quality of multimodal representations in recommendation systems. However, existing works overlook the fundamental representational divergence between large models and recommender systems, resulting in incompatible multimodal representations and suboptimal recommendation performance. To bridge this gap, we propose **RecGOAT**, a novel yet simple dual semantic alignment framework for LLM-enhanced multimodal recommendation, which offers theoretically guaranteed alignment capability. RecGOAT first employs graph attention networks to enrich collaborative semantics by modeling item-item, user-item, and user-user relationships, leveraging user/item LM representations and interaction history. Furthermore, we design a dual-granularity progressive multimodality-unique identity (ID) alignment framework, which achieves instance-level and distribution-level semantic alignment via cross-modal contrastive learning (CMCL) and optimal adaptive transport (OAT), respectively. Theoretically, we demonstrate that the unified representations derived from our alignment framework exhibit superior semantic consistency and comprehensiveness. Extensive experiments on three public benchmarks show that our RecGOAT achieves state-of-the-art performance, empirically validating our theoretical insights. Additionally, the deployment on a large-scale online advertising platform confirms the model's effectiveness and scalability in industrial recommendation scenarios. Our code is available at https://anonymous.4open.science/r/RecGOAT-244D.

## CCS Concepts

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

## Keywords

Multimodal Recommendation, Large Language Models, Semantic Alignment, Graph Neural Networks, Optimal Transport

**ACM Reference Format:**
Anonymous Author(s). 2018. The Supplementary Material for RecGOAT. In *Proceedings of 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

In the supplementary materials, we provide additional supporting content, including: (1) a detailed case analysis of LLM-based reasoning for user profile inference in Section 3.1.3, illustrating how large language models interpret and summarize user preferences from behavioral histories; and (2) complete proofs of two key lemmas introduced in Section 3.3.2, offering further theoretical grounding for the proposed framework.

## 1 A Case Study of User Preference Reasoning

As shown in Table 1, we present the preference inference path for user 0. It can be observed that the LLM-generated user preferences accurately capture key attributes of the ground-truth target item, including product category (**Baby products**), brand (**Munchkin**), usage scenario (**home**), and visual preference (**colorful design**). Encoding such information as the user's initial textual representation effectively enhances the model's understanding of user preferences.

## 2 The Detailed Proofs of Lemma

The following sections provide the detailed proofs of Lemma 3.3 (Instance-level Distance Bound) and Lemma 3.4 (Modality-to-Unified Error Bound) from Section 3.3.2 of the paper.

Proof of Lemma 3.3. The cross-modal contrastive loss (InfoNCE) for aligning modality $m$ with the unified representation is defined as

$$\mathcal{L}_{\text{CMCL}} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_i^m, \qquad \mathcal{L}_i^m = -\log\frac{\exp\big(z_i^m\cdot z_i/\tau\big)}{\sum_{j=1}^{B}\exp\big(z_i^m\cdot z_j/\tau\big)},$$

where $z_i^m$ and $z_j$ are $L_2$-normalised ($\|z_i^m\| = \|z_j\| = 1$) and $\tau > 0$ is the temperature.

For a fixed $i$, split the denominator into the positive term and the negative terms:

$$\sum_{j=1}^{B}\exp\big(z_i^m\cdot z_j/\tau\big) = \exp\big(z_i^m\cdot z_i/\tau\big) + \sum_{j\neq i}\exp\big(z_i^m\cdot z_j/\tau\big).$$

Since $z_i^m\cdot z_j \leq 1$ for any $j$ (Cauchy–Schwarz), each negative term satisfies $\exp\big(z_i^m\cdot z_j/\tau\big) \leq \exp(1/\tau)$. Hence

$$\sum_{j\neq i}\exp\big(z_i^m\cdot z_j/\tau\big) \leq (B-1)\exp(1/\tau).$$

Substituting this bound into $\mathcal{L}_i^m$ gives

$$\begin{aligned}\mathcal{L}_i^m &= \log\left(\sum_{j=1}^{B}\exp\big(z_i^m\cdot z_j/\tau\big)\right) - \frac{z_i^m\cdot z_i}{\tau}\\ &\leq \log\Big(\exp\big(z_i^m\cdot z_i/\tau\big) + (B-1)\exp(1/\tau)\Big) - \frac{z_i^m\cdot z_i}{\tau}. \qquad (1)\end{aligned}$$

*SIGIR '26, Melbourne, Australia*

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

**Table 1: Case Study of LLM-based Preference Reasoning on the *Baby* Dataset**

| **Input (Prompt & User Context)** | **Output (LLM Reasoning & Real Target Item)** |
|---|---|
| **User ID:** 0<br><br>**Prompt:** You are a professional data analyst. Your task is to analyze a user's interaction history to infer their preferences.<br>**User Information**<br>• User ID: `{user id}`<br>• Interaction History: `{item ids & text description}`<br>Please conduct a structured reasoning by two steps:<br>1. **Identify Common Attributes Across Items**<br>- Extract shared characteristics from perspectives such as category, function, style, material, usage scenario, and price range. Organize the findings in a list format.<br>2. **Summarize Preferences Across Multiple Dimensions**<br>- Systematically summarize the user's preferences based on the following dimensions, with a brief explanation of supporting evidence:<br>• Style & Aesthetics: e.g., design style, color preference, visual element tendencies<br>• Function & Usability: e.g., functional needs, usage scenarios, interaction experience<br>• Context & Emotion: e.g., emotional value associated with items, social attributes, life scenarios<br>• Other Dimensions: e.g., brand loyalty, price sensitivity, attention to sustainability<br>**Output Format:**<br>`<think> reasoning process here </think>`<br>`<answer> answer here </answer>`<br>Please provide your analysis below: | **LLM Response:**<br>`<think>... ...</think>`<br><br>`<answer>`<br>`1. Identify Common Attributes Across Items`<br>`- **Category**: All items fall under **Baby products**, ...`<br>`- **Brand**: Repeated engagement with trusted brands like **Philips Avent, Munchkin, Britax, and Petunia Pickle Bottom**.`<br>`- **Usage Scenario**: Items cater to **home, travel, and on-the-go use** (e.g., compact breast pump, travel bag).`<br><br>` 2. Summarize Preferences Across Multiple Dimensions`<br>` **Style & Aesthetics**`<br>` - **Design Preference**: Combines **functional simplicity with playful or trendy visuals** (e.g., floral-patterned bottles, colorful silicone spoons).`<br>` - **Color Tendencies**: Neutral and pastel hues (pink, green) alongside vibrant accents (red stroller).`<br>` - **Brand Aesthetics**: Appreciation for **stylish yet practical brands** like Petunia Pickle Bottom (cosmopolitan caharm) and Philips Avent ... ...`<br>` - **Brand Loyalty**: Clear preference for **reputable brands** (Philips Avent appears twice; Britax and Munchkin multiple times).`<br>` ...`<br>` **Key Takeaway**: The user prioritizes **safe, functional, and stylish baby products** from trusted brands, balancing practicality with aesthetic appeal for both home and travel scenarios.`<br>`</answer>`<br><br>`------------------------------------------------------------`<br><br>**Target Item for Next-Buy Prediction (Ground Truth):**<br>`"B000GK7AMU"`: Munchkin Five Sea Squirts.<br>**Brand**: Munchkin.<br>**Category**: Baby.<br>**Description**: Fill them up and squirt them out. They make bath time a real blast for babies. The lovable and colorful sea characters provide creative play and are an amusing way for moms to rinse away soap. Just add them to your water and expect an ocean of fun. |

Because $\|z_i^m\| = \|z_i\| = 1$, we have the identity

$$z_i^m \cdot z_i = 1 - \frac{1}{2}\,\|z_i^m - z_i\|^2. \tag{2}$$

Insert (2) into (1) and denote $t_i = \|z_i^m - z_i\|^2 \geq 0$:

$$\mathcal{L}_i^m \leq \log\Big(\exp\big((1 - t_i/2)/\tau\big) + (B-1)\exp(1/\tau)\Big) - \frac{1 - t_i/2}{\tau}$$

$$\leq \log\big(B\exp(1/\tau)\big) - \frac{1 - t_i/2}{\tau} = \log B + \frac{t_i}{2\tau}.$$

The last inequality uses $\exp\big((1 - t_i/2)/\tau\big) \leq \exp(1/\tau)$.

Rearranging yields

$$\|z_i^m - z_i\|^2 \leq 2\tau\,\mathcal{L}_i^m + 2\tau\log B.$$

Taking expectation over the batch (recall $\mathcal{L}_{\text{CMCL}} = \mathbb{E}_i[\mathcal{L}_i^m]$) gives

$$\mathbb{E}_i\big[\|z_i^m - z_i\|^2\big] \leq 2\tau\,\mathcal{L}_{\text{CMCL}} + 2\tau\log B.$$

Finally, by Cauchy–Schwarz,

$$\big(\mathbb{E}_i[\|z_i^m - z_i\|]\big)^2 \leq \mathbb{E}_i\big[\|z_i^m - z_i\|^2\big] \leq 2\tau\,\mathcal{L}_{\text{CMCL}} + 2\tau\log B.$$

Taking the square root completes the proof:

$$\mathbb{E}_i\big[\|z_i^m - z_i\|\big] \leq \sqrt{2\tau\,\mathcal{L}_{\text{CMCL}} + 2\tau\log B}.$$

□

Proof of Lemma 3.4. From the definitions of $\epsilon_m(f)$ and $\epsilon_F(f)$, and applying the triangle inequality, we have for any fixed user $\boldsymbol{u}$ and a sample $i$:

$$\Big|\,|f(\boldsymbol{u}, \boldsymbol{z}_i^m) - f^*(\boldsymbol{u}, \boldsymbol{v}_i)| - |f(\boldsymbol{u}, \boldsymbol{z}_i) - f^*(\boldsymbol{u}, \boldsymbol{v}_i)|\,\Big| \leq |f(\boldsymbol{u}, \boldsymbol{z}_i^m) - f(\boldsymbol{u}, \boldsymbol{z}_i)|.$$

Taking the expectation with respect to the unified (target) distribution $Q$ yields:

$$\Big|\mathbb{E}_{z_i \sim Q}\big[\,|f(\boldsymbol{u}, \boldsymbol{z}_i^m) - f^*(\boldsymbol{u}, \boldsymbol{v}_i)|\,\big] - \epsilon_F(f)\Big| \leq \mathbb{E}_{z_i \sim Q}\big[\,|f(\boldsymbol{u}, \boldsymbol{z}_i^m) - f(\boldsymbol{u}, \boldsymbol{z}_i)|\,\big]. \tag{1}$$

The difference between the expected error under $Q$ and the true modal error $\epsilon_m(f)$ stems from the distribution shift between $P^m$ and $Q$. Consider the function $g(\boldsymbol{z}) = |f(\boldsymbol{u}, \boldsymbol{z}) - f^*(\boldsymbol{u}, \boldsymbol{v})|$. We show $g$ is $(K + L^*)$-Lipschitz continuous:

$$|g(\boldsymbol{z}_1) - g(\boldsymbol{z}_2)| \leq |f(\boldsymbol{u}, \boldsymbol{z}_1) - f(\boldsymbol{u}, \boldsymbol{z}_2)| \leq K\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|,$$

where the first inequality follows from the reverse triangle inequality, and the second from the $K$-Lipschitz continuity of $f$ (Assumption 3.1). Since $f^*$ is $L^*$-Lipschitz with respect to $\boldsymbol{z}$ (Assumption 3.2), the same bound holds for the composition, confirming $g$ is indeed $(K + L^*)$-Lipschitz.

By the duality property of the 1-Wasserstein distance, for any $(K + L^*)$-Lipschitz function $g$,

$$\left|\mathbb{E}_{P^m}[g(\boldsymbol{z})] - \mathbb{E}_Q[g(\boldsymbol{z})]\right| \leq (K + L^*) \cdot \mathcal{W}_1(P^m, Q).$$

Applying this to $g(\boldsymbol{z}) = |f(\boldsymbol{u}, \boldsymbol{z}) - f^*(\boldsymbol{u}, \boldsymbol{v})|$, we obtain:

$$\left|\epsilon_m(f) - \mathbb{E}_{z_i \sim Q}\left[\,|f(\boldsymbol{u}, \boldsymbol{z}_i^m) - f^*(\boldsymbol{u}, \boldsymbol{v}_i)|\,\right]\right| \leq (K + L^*) \cdot \mathcal{W}_1(P^m, Q). \quad (2)$$

Finally, combining (1) and (2) via the triangle inequality gives:

$$|\epsilon_m(f) - \epsilon_F(f)| \leq (K + L^*) \cdot \mathcal{W}_1(P^m, Q) + \mathbb{E}_{z_i \sim Q}\left[\,|f(\boldsymbol{u}, \boldsymbol{z}_i^m) - f(\boldsymbol{u}, \boldsymbol{z}_i)|\,\right].$$

Using the $K$-Lipschitz continuity of $f$ again, the second term is bounded by $K \cdot \mathbb{E}_i \|\boldsymbol{z}_i^m - \boldsymbol{z}_i\|$. A tighter bound that accounts for the potential shift in the $f^*$ term, consistent with the Lipschitz constant $(K + L^*)$, yields the final result:

$$|\epsilon_m(f) - \epsilon_F(f)| \leq (K + L^*) \cdot \mathcal{W}_1(P^m, Q) + (K + L^*) \cdot \mathbb{E}_i \|\boldsymbol{z}_i^m - \boldsymbol{z}_i\|.$$

This completes the proof. □