MARIO: Modality-Aware Attention and Modality-Preserving Decoders for Multimedia Recommendation

Taeri Kim* Hanyang University

Yeon-Chang Lee* Hanyang University and Georgia Tech

Kijung Shin KAIST

Sang-Wook Kim Hanyang University

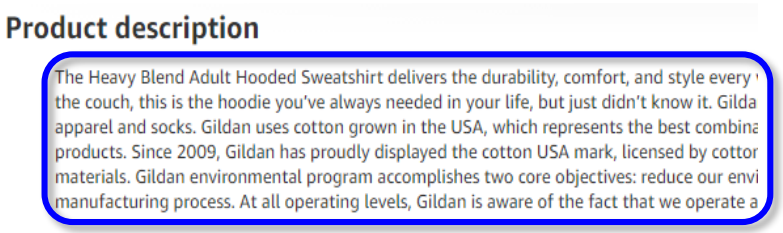
* Equal Contribution

Background: Multimedia Recommendation

Problem definition

- Inputs
  - User-item interaction information
  - Multimedia contents (*e.g.*, image and text information)
- Output: the predicted preference values of the user-item pairs

Image information (Product image)

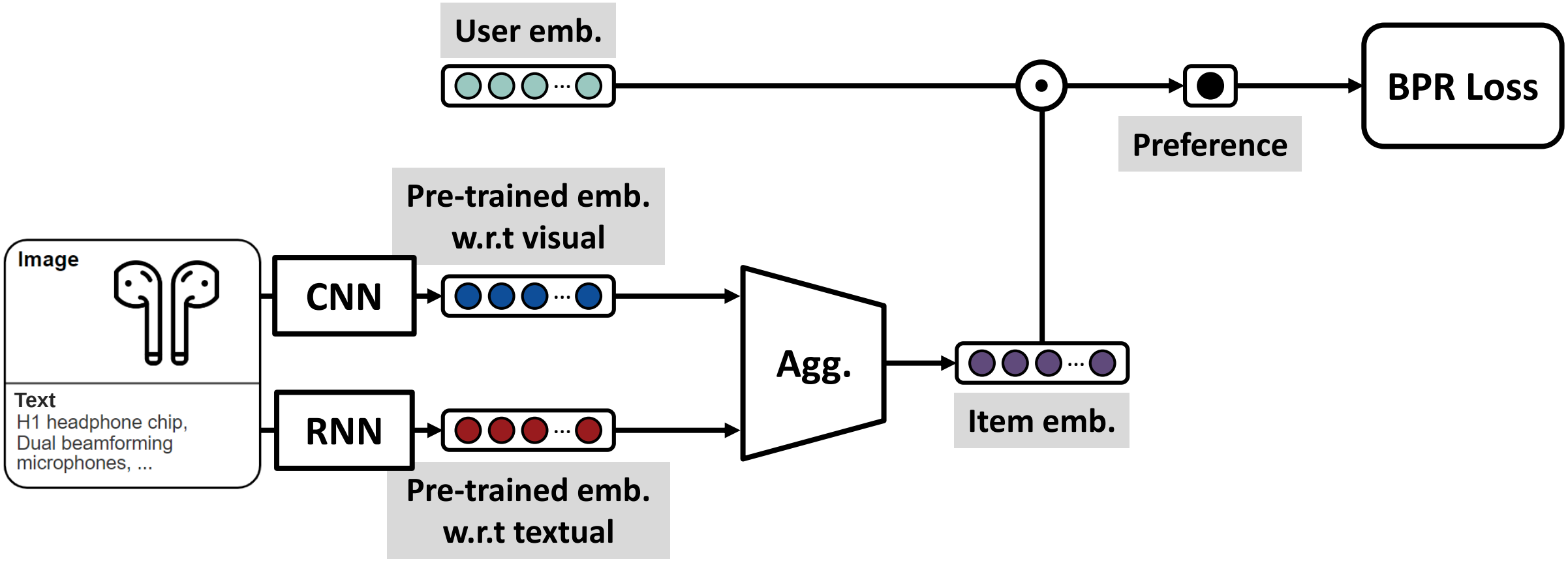


Text information (Product description)

https://www.amazon.com/

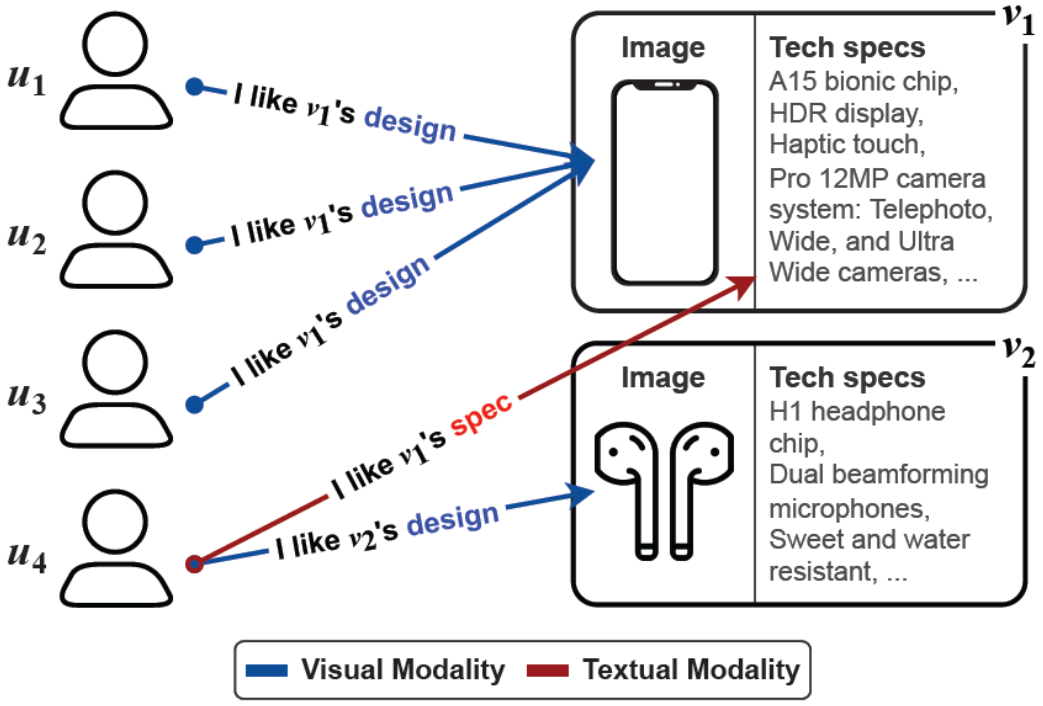
Background: Multimedia Recommendation (cont'd)

Conceptual view of existing work



Motivation 1: Influence of Each Modality at the Interaction Level

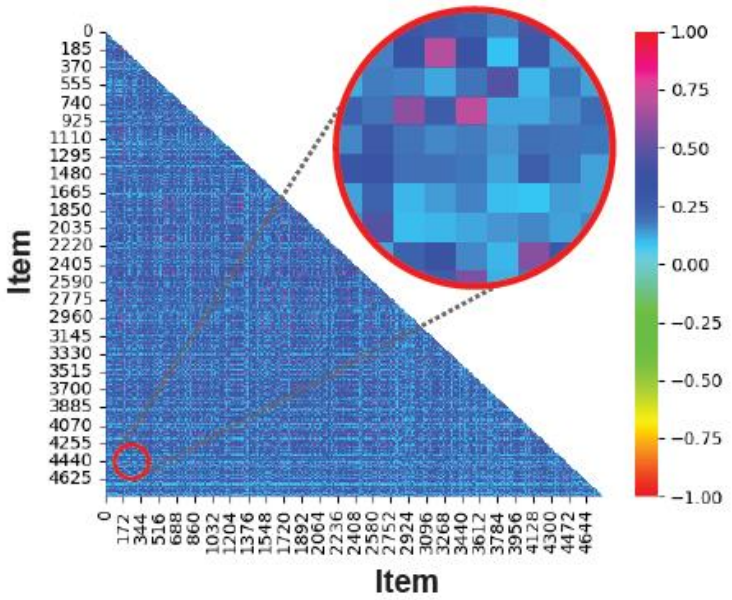
- Recent studies designed an attention mechanism to capture the influence of each modality on the interactions between users and items
- However, none of these methods capture the individual influence of each modality at the interaction level



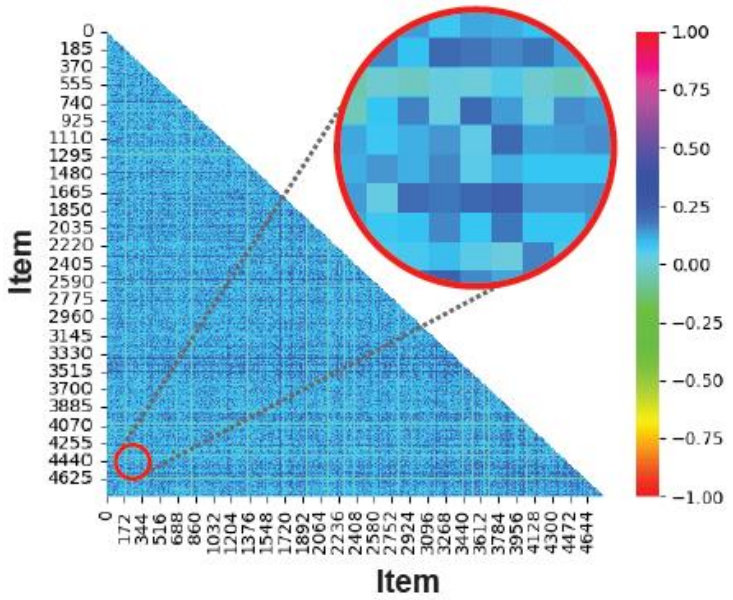
Influence of each modality may differ across interactions!

Motivation 2: Preservation of Items' Modality-Specific Properties

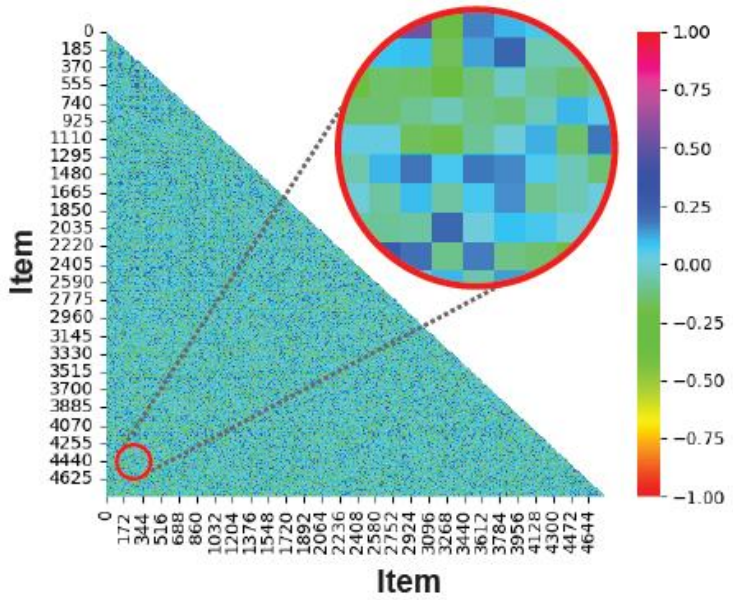
- **Pre-trained embeddings of items per modality reflect their modality-specific properties that cannot be captured by interaction information**



(a) Visual Modality



(b) Textual Modality

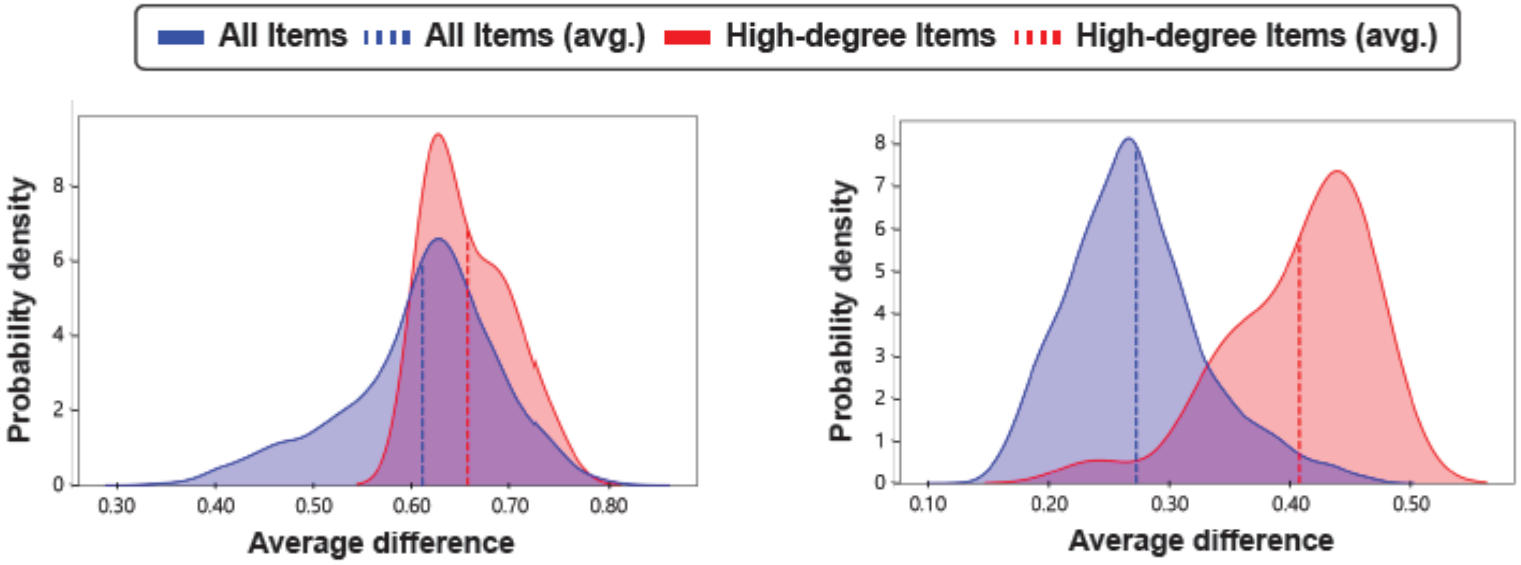


(c) Interaction Modality

Even for the same item pair, the similarities in their visual, textual, and interaction modality differ significantly!

Motivation 2: Preservation of Items' Modality-Specific Properties

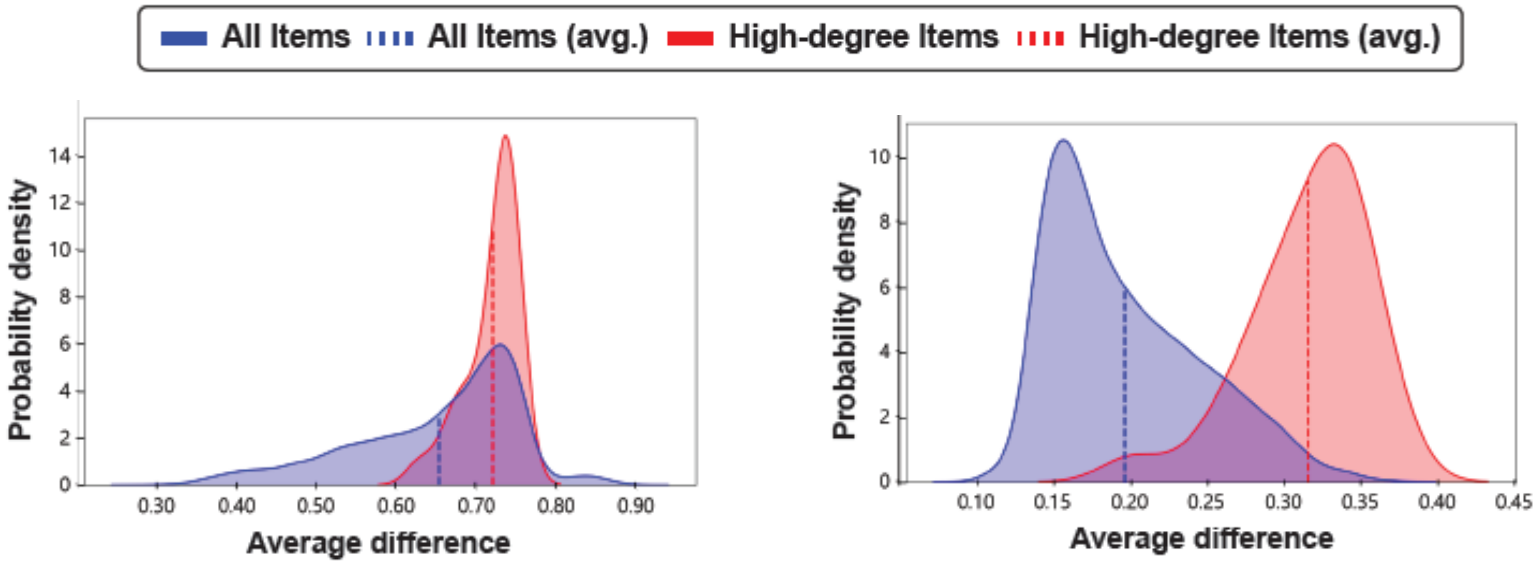
- Existing works learn the final item embeddings to be similar to the embeddings of users who interact with the items
- We observe that such learning procedures of existing works fail to preserve such modality-specific properties in the final item embeddings
  - Visual modality



(a) MMGCN [ACM MM'19] (b) LATTICE [ACM MM'21]

Motivation 2: Preservation of Items' Modality-Specific Properties

- Textual modality



(a) MMGCN [ACM MM'19] (b) LATTICE [ACM MM'21]

Final item embeddings of existing multimedia recommender systems significantly lose the intrinsic modality-specific properties!

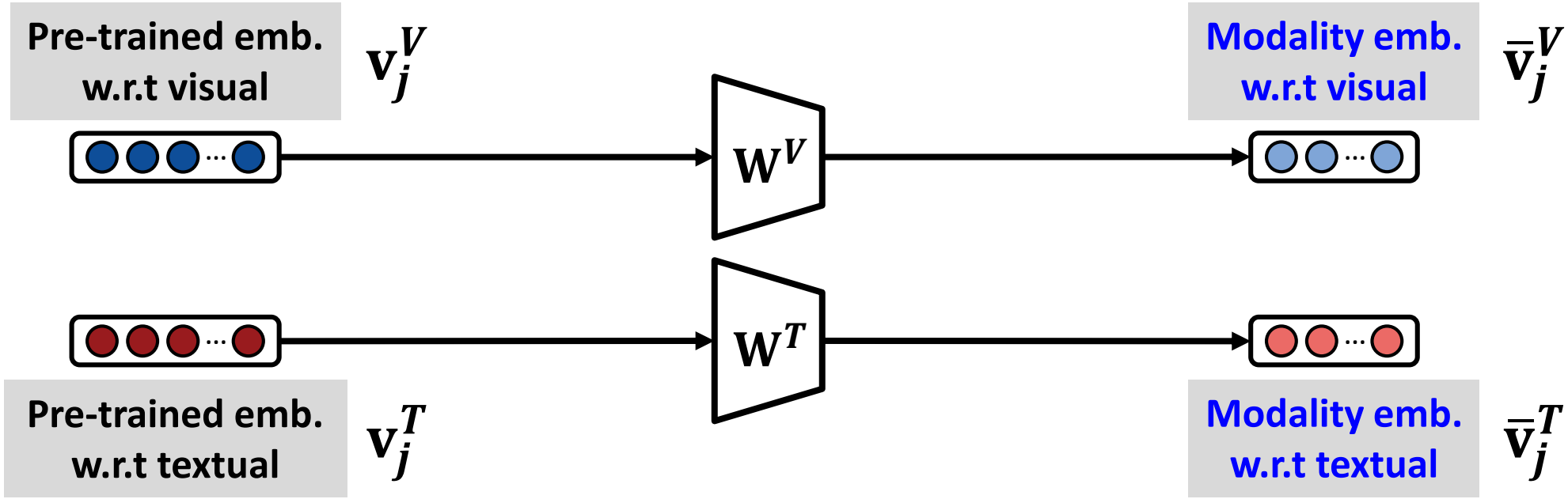
Proposed Method: MARIO

- **Our ideas**
  - (I1) To consider the influence of each modality on each interaction
  - (I2) To obtain item embeddings that preserve the intrinsic modality-specific properties

- **We propose MARIO (Modality-aware Attention and modality-pReservIng decOders) for accurate multimedia recommendation**
  - (C1) Encoders based on interaction and multimodal information
  - (C2) A predictor based on attention networks
  - (C3) Decoders for modality preservation
  - (C4) Training based on BPR loss and modality preservation (MP) loss

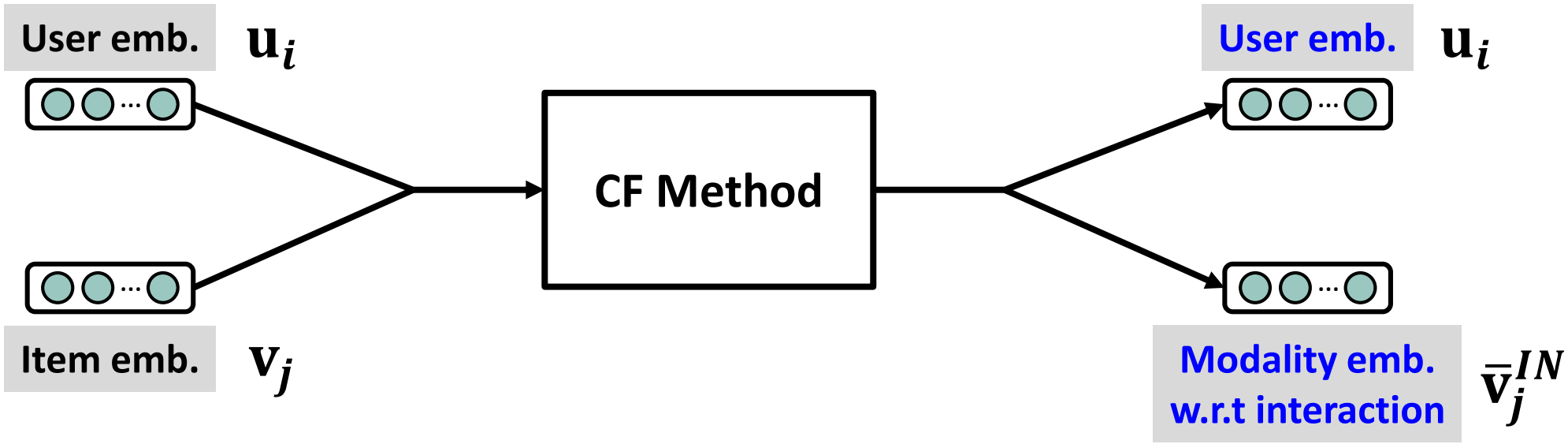
(C1) Encoders

- We obtain the visual and textual-modality embeddings of each item v_j
- To this end, we encode its pre-trained embeddings from visual and textual modalities into embeddings of the same dimensionality



(C1) Encoders (cont'd)

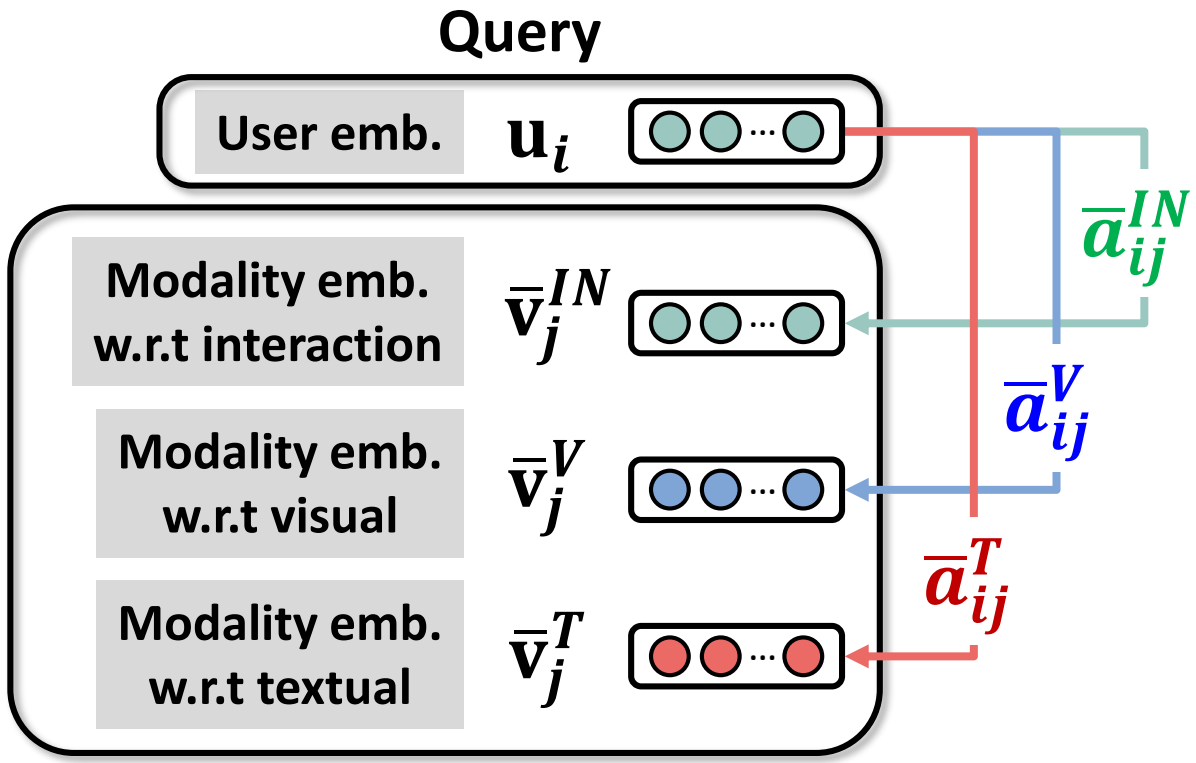
- We obtain the embedding of each user u_i and the interaction-modality embedding of each item v_j by employing *any* CF method



- This design choice makes any CF methods be easily applied to MARIO to utilize items' multimodal features

(C2) A Predictor

- We identify the influence of each modality m on each interaction between u_i and v_j based on a modality-aware attention mechansim



$$\bar{a}_{ij}^{m} = \frac{\exp(a_{ij}^{m})}{\sum_{m \in M} \exp(a_{ij}^{m})},$$

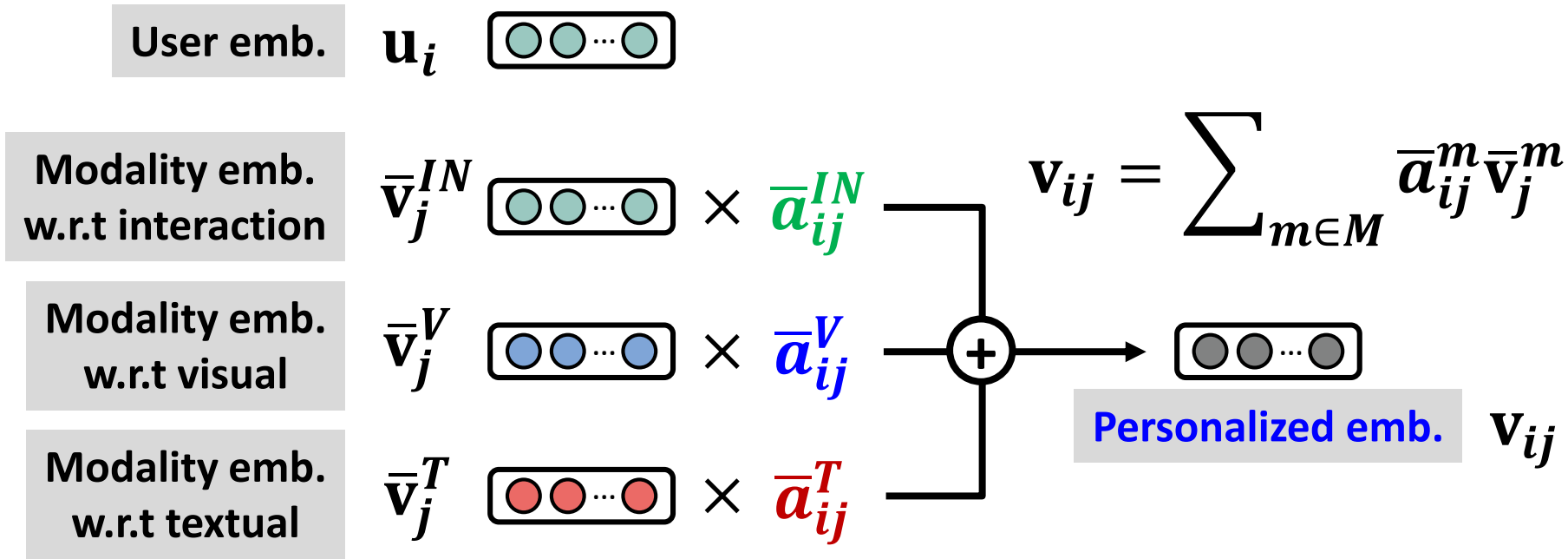
$$\text{where } a_{ij}^{m} = \frac{\mathbf{u}_i \odot \bar{\mathbf{v}}_j^{m}}{\sqrt{d}}$$

$$\text{and } M = \{V, T, IN\}$$

Modality-aware Attention

(C2) A Predictor (cont'd)

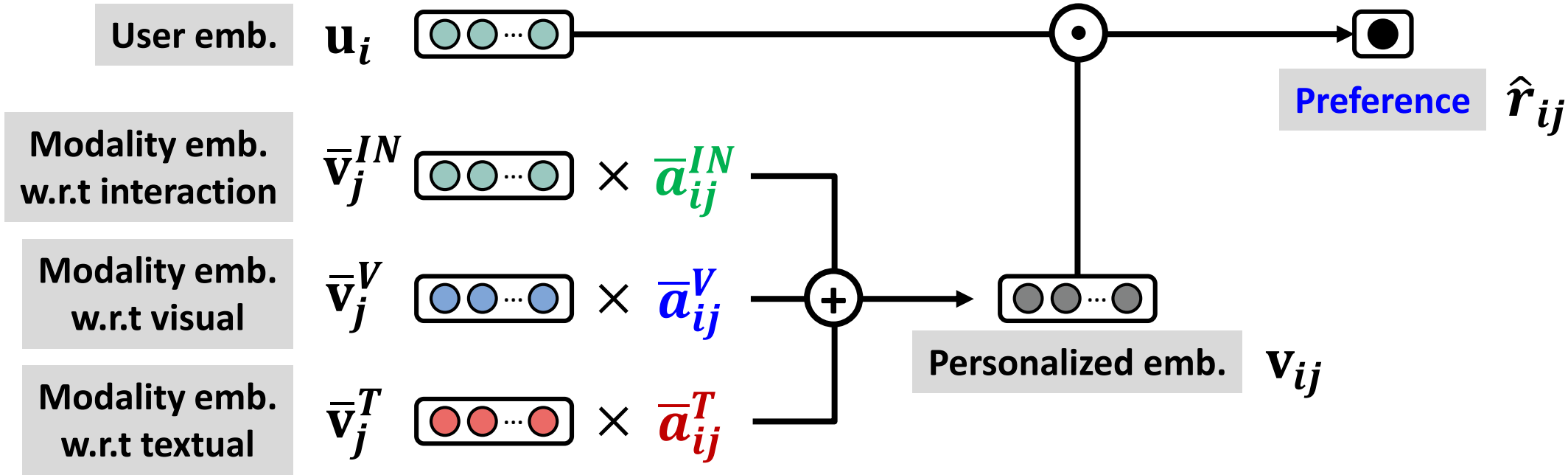
- We obtain each item v_j's personalized embedding w.r.t each user u_i by fusing v_j's modality embeddings based on their influence



(C2) A Predictor (cont'd)

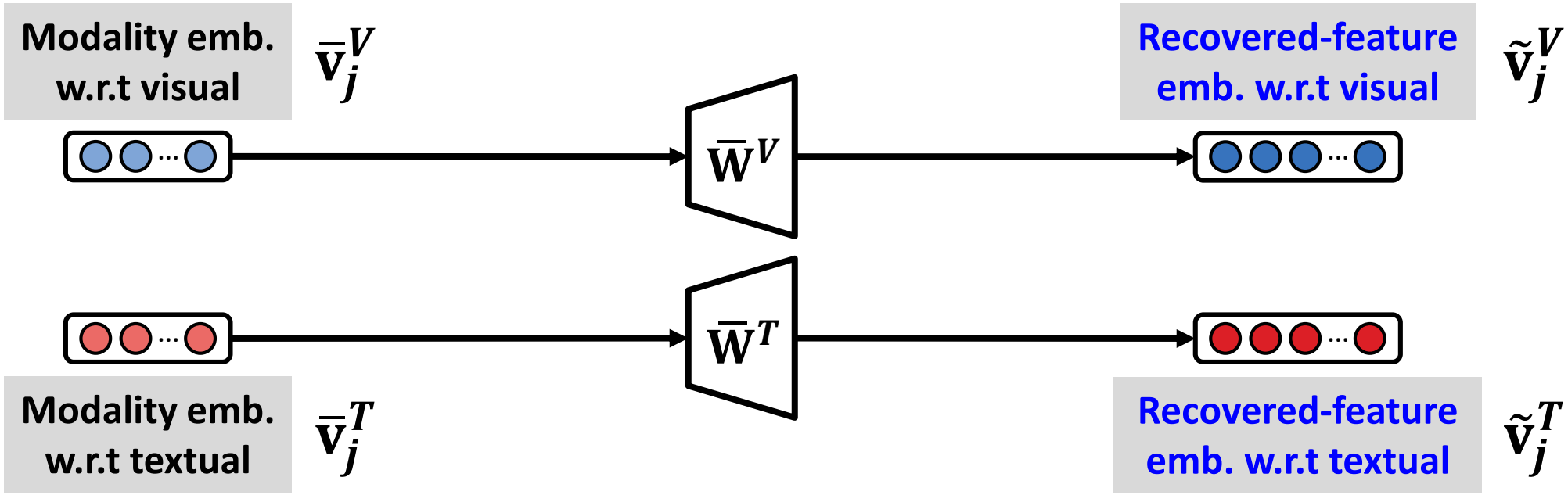
- We predict each user u_i's preference on each item v_j and recommend the most preferred top-*N* items to u_i

$$\hat{r}_{ij} = \mathbf{u}_i \odot \mathbf{v}_{ij}$$



(C3) Decoders for Modality Preservation

- We reconstruct the item v_j's pre-trained embeddings by decoding the modality embeddings obtained in (C1)



- The recovered-feature embeddings will be used to preserve v_j's modality-specific properties in the learning process of MARIO

(C4) Training

- We learn each user u_i's embedding $\mathbf{u}_i$ and each item v_j's modality embeddings $\bar{\mathbf{v}}_j^{IN}$, $\bar{\mathbf{v}}_j^{V}$, $\bar{\mathbf{v}}_j^{T}$

- **Bayesian personalized ranking (BPR)** loss

$$\mathcal{L}_{BPR} = -\sum_{u_i \in U} \sum_{v_j \in N_i} \sum_{v_p \in N_i \setminus I} \ln \sigma(\hat{r}_{ij} - \hat{r}_{ip})$$

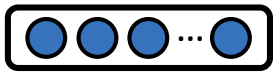
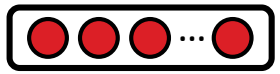
u_i's preference $\hat{r}_{ij}$ on the interacted item v_j is likely to be higher than u_i's preference $\hat{r}_{ip}$ on a (randomly-sampled) non-interacted item v_p

(C4) Training (cont'd)

☐ Modality preservation (MP) loss

$$\mathcal{L}_{MP} = \sum_{m \in M \setminus IN} \sum_{v_j \in I} \sum_{f=1}^{d_m} \left| \tilde{\mathbf{v}}_j^m(f) - \mathbf{v}_j^m(f) \right|$$

Aim to minimize the difference between v_j's recovered-feature embedding $\tilde{\mathbf{v}}_j^V$ / $\tilde{\mathbf{v}}_j^T$ and v_j's pre-trained embedding $\mathbf{v}_j^V$ / $\mathbf{v}_j^T$

Recovered-feature emb. w.r.t visual	$\tilde{\mathbf{v}}_j^V$	Pre-trained emb. w.r.t visual	$\mathbf{v}_j^V$	Recovered-feature emb. w.r.t textual	$\tilde{\mathbf{v}}_j^T$	Pre-trained emb. w.r.t textual	$\mathbf{v}_j^T$
	$\approx$				$\approx$		

(C4) Training (cont'd)

Final loss

$$\mathcal{L} = \mathcal{L}_{BPR} + \mu \mathcal{L}_{MP}$$

Hyperparameter to control the balance between $\mathcal{L}_{BPR}$ and $\mathcal{L}_{MP}$

- By jointly optimizing the two losses, MARIO fully utilizes rich modality-specific semantics in addition to user-item interaction information

Experimental Settings

Datasets

Dataset	# User	# Item	# Interaction	Sparsity	Dim. of V / T
Baby	19,445	7,050	160,792	99.88%	4,096 / 1,024
Clothing	4,955	5,028	32,363	99.87%	4,096 / 1,024
Office	4,874	2,406	52,957	99.55%	4,096 / 1,024
Musical	1,429	900	10,261	99.20%	4,096 / 1,024

Seven competitors

- Three CF methods (*i.e.*, using collaborative information only)
  - BPRMF [UAI'09]
  - NGCF [ACM SIGIR'19]
  - LightGCN [ACM SIGIR'20]
- Four state-of-the-art multimedia recommender systems
  - MAML [ACM MM'19]
  - MMGCN [ACM MM'19]
  - GRCN [ACM MM'20]
  - LATTICE [ACM MM'21]

Evaluation metrics: NDCG, Recall, Precision @ 10/20

Questions to Be Answered

- RO1: Does MARIO provide more-accurate top-*N* recommendation than state-of-the-art multimedia recommender systems?
- RQ2: Does equipping MARIO with different CF methods consistently improve their accuracies?
- RQ3: Are the modality-aware attention networks of MARIO effective for multimedia recommendation?
- RQ4: Does the MP loss of MARIO help modality preservation and accurate multimedia recommendation?

Results for RQ1

Comparison with state-of-the-art multimedia recommender systems

- Top-10 Results

Datasets	Baby			Clothing		
Metrics	NDCG@10	Recall@10	Pre@10	NDCG@10	Recall@10	Pre@10
MAML	–	–	–	0.0226	0.0442	0.0044
MMGCN	0.0187	0.0370	0.0039	0.0133	0.0260	0.0026
GRCN	0.0262	0.0468	0.0050	0.0194	0.0355	0.0036
LATTICE	0.0276	0.0503	0.0053	0.0221	0.0404	0.0040
MARIO	**0.0300***	**0.0539***	**0.0056***	**0.0259***	**0.0484***	**0.0048***
Improvement	**8.58%**	**7.20%**	**6.81%**	**14.61%**	**9.59%**	**9.09%**

Datasets	Office			Musical		
Metrics	NDCG@10	Recall@10	Pre@10	NDCG@10	Recall@10	Pre@10
MAML	0.0505	0.0887	**0.0112**	0.0662	0.1302	0.0134
MMGCN	0.0281	0.0534	0.0067	0.0517	0.0980	0.0101
GRCN	0.0553	0.0845	0.0106	0.0541	0.1068	0.0110
LATTICE	**0.0589**	0.0902	0.0109	0.0769	0.1459	0.0148
MARIO	0.0583	**0.0932***	0.0110	**0.0790***	**0.1513***	**0.0153***
Improvement	-0.89%	**3.38%**	-1.47%	**2.73%**	**3.68%**	**3.30%**

–: out-of-memory

Results for RQ1 (cont'd)

Comparison with state-of-the-art multimedia recommender systems

- Top-20 Results

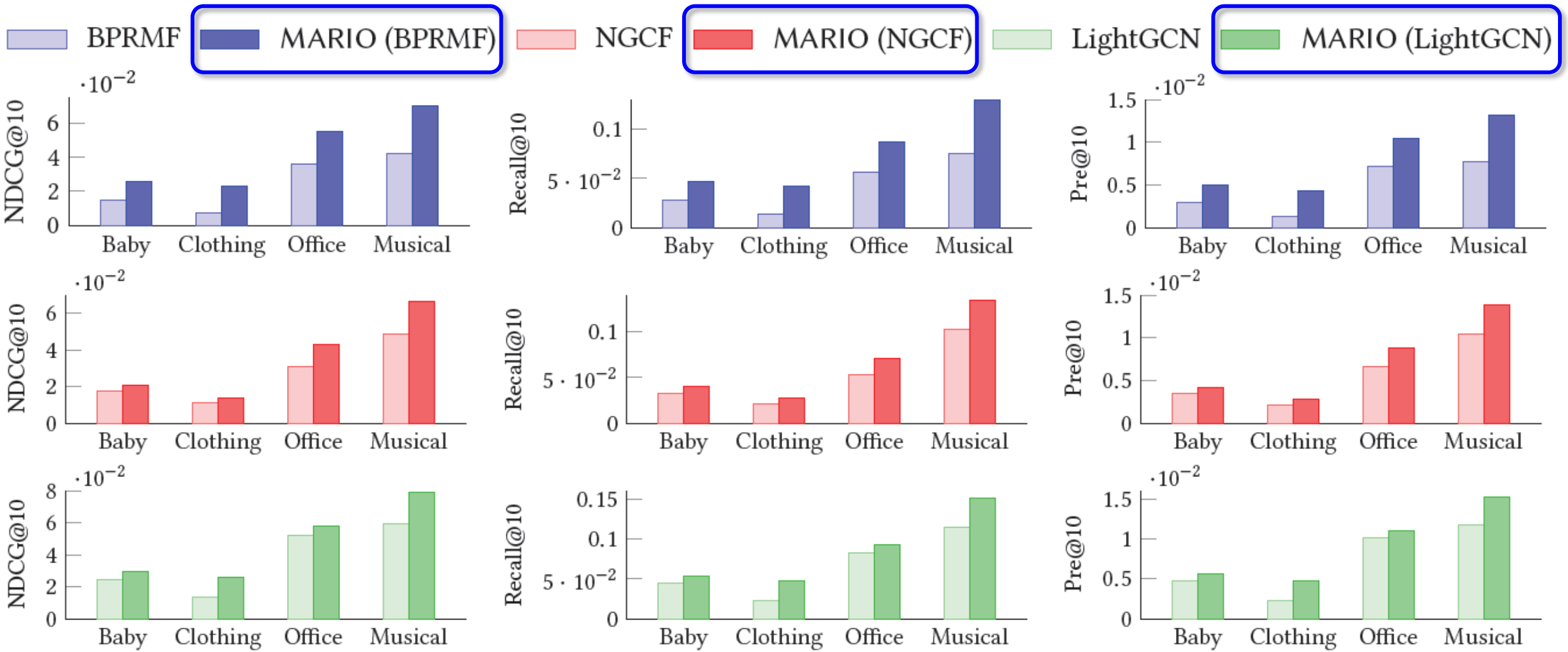
Datasets	Baby			Clothing		
Metrics	NDCG@20	Recall@20	Pre@20	NDCG@20	Recall@20	Pre@20
MAML	–	–	–	0.0289	0.0694	**0.0035**
MMGCN	0.0249	0.0615	0.0033	0.0168	0.0399	0.0020
GRCN	0.0335	0.0743	0.0040	0.0242	0.0544	0.0027
LATTICE	0.0355	0.0804	0.0042	0.0279	0.0634	0.0032
MARIO	**0.0378***	**0.0838***	**0.0044***	**0.0314***	**0.0706***	**0.0035**
Improvement	**6.58%**	**4.13%**	**4.02%**	**8.50%**	**1.79%**	0.00%

Datasets	Office			Musical		
Metrics	NDCG@20	Recall@20	Pre@20	NDCG@20	Recall@20	Pre@20
MAML	0.0628	0.1350	**0.0086**	0.0822	0.1919	0.0099
MMGCN	0.0375	0.0892	0.0056	0.0692	0.1682	0.0086
GRCN	0.0689	0.1326	0.0084	0.0707	0.1726	0.0089
LATTICE	0.0725	0.1360	0.0083	0.0944	0.2126	0.0109
MARIO	**0.0728***	**0.1418***	**0.0086**	**0.0968***	**0.2195***	**0.0113***
Improvement	**0.38%**	**4.26%**	0.00%	**2.53%**	**3.26%**	**3.21%**

–: out-of-memory

Results for RQ2

☐ Effectiveness of MARIOs equipped with three CF methods



Results for RQ3

Effectiveness of our attention network (Clothing dataset)

Model	NDCG@10	Recall@10	Pre@10
MARIO_{mean}	0.0256	0.0478	**0.0048**
MARIO_{max}	0.0232	0.0429	0.0043
MARIO_{fc}	0.0117	0.0214	0.0022
MARIO	**0.0259**	**0.0484**	**0.0048**

- MARIO consistently outperforms MARIO_{mean}, MARIO_{max}, and MARIO_{fc}
- It is most effective to use the proposed attention mechanism to aggregate all modality embeddings of items
  - By considering the individual influence of each modality at the interaction level

Results for RQ4

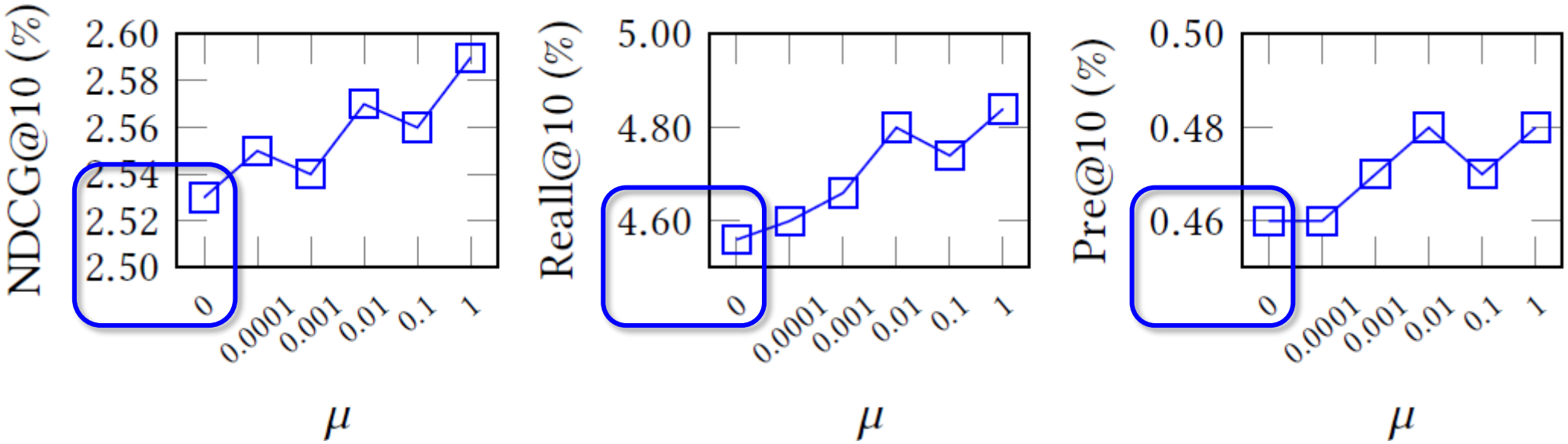
- **Effectiveness of our MP loss in terms of modality preservation (Clothing dataset)**
  - The smaller the differences are, the better the modality-specific properties are preserved

Modality		MMGCN	LATTICE	MARIO
All Items	Visual	0.6111	0.2721	**0.2668**
	Textual	0.6541	0.1961	**0.1842**
Items with High-degree	Visual	0.6576	0.4075	**0.3062**
	Textual	0.7221	0.3155	**0.2739**

  - The MP loss affects modality preservation more for high-degree items than for low-degree items
  - MARIO is most effective in preserving the intrinsic modality-specific properties

Results for RQ4 (cont'd)

- **Effectiveness of our MP loss in terms of recommendation accuracy (Clothing dataset)**



$$\mathcal{L} = \mathcal{L}_{BPR} + \boldsymbol{\mu}\mathcal{L}_{MP}$$

- Accuracy of MARIO is always the worst when $\mu = 0$
- Preserving the modality-specific properties is also effective in improving the accuracy of multimedia recommendation

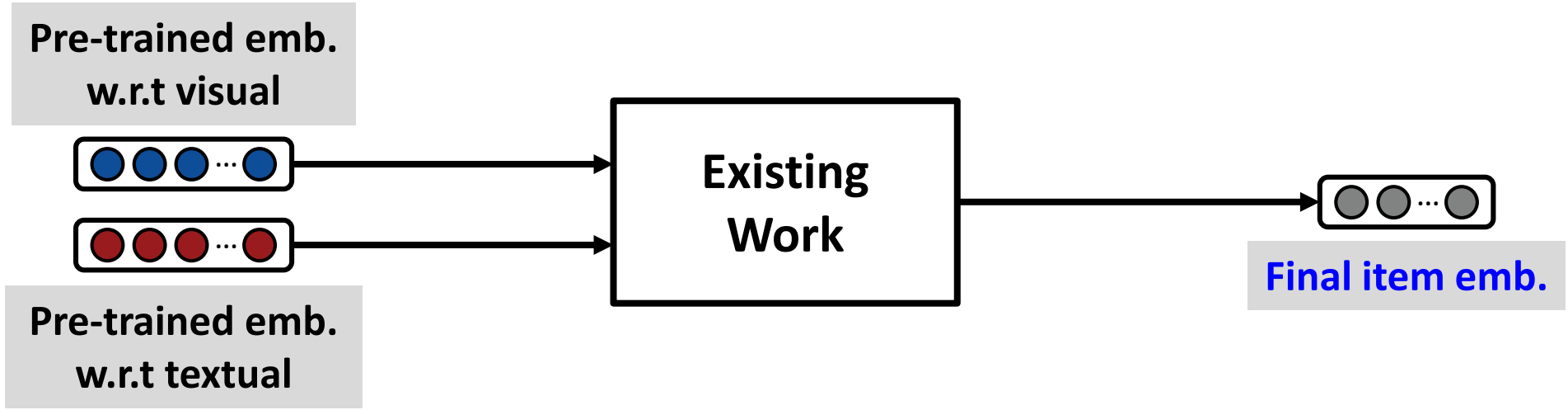
Conclusions

- **(Observation) We pointed out that existing multimedia recommender systems face difficulties in**
  - Capturing the individual influence of each modality at interaction level
  - Exploiting the intrinsic modality-specific properties of items
- **(Framework Design) We proposed MARIO based on modality-aware attention and modality-preserving decoders**
- **(Extensive Evaluation) We demonstrated that MARIO**
  - Outperforms its seven competitors on four real-life datasets
  - Preserves the intrinsic properties of item modalities better

Appendix

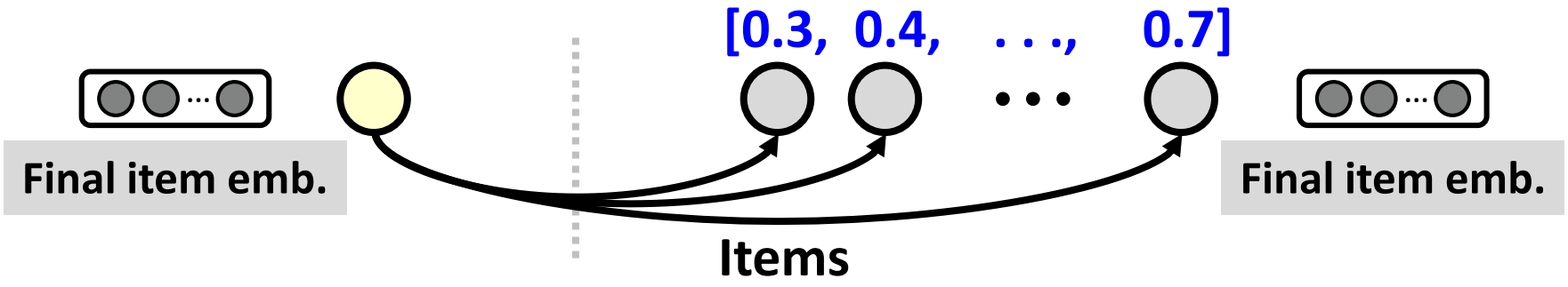
Procedure of Density Function

- (1) Obtaining the final item embeddings from the existing multimedia recommender systems (*e.g.*, MMGCN and LATTICE)

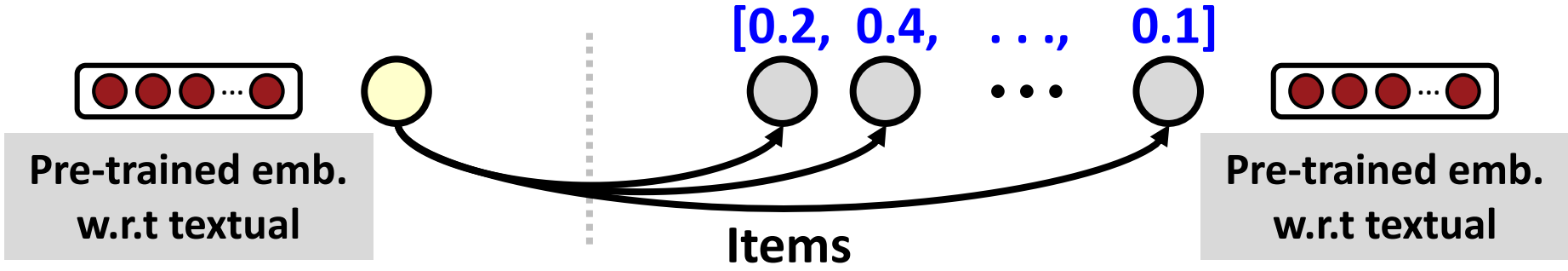


Procedure of Density Function (cont'd)

- (2) For each item, computing the cosine similarity with every other item by using their final embeddings obtained by each method (f-similarities)



- (3) For each item, computing the cosine similarity with every other item by using their pre-trained embeddings (p-similarities) per modality



Procedure of Density Function (cont'd)

- (4) For each item, computing the differences between f-similarity and p-similarity with every other item per modality

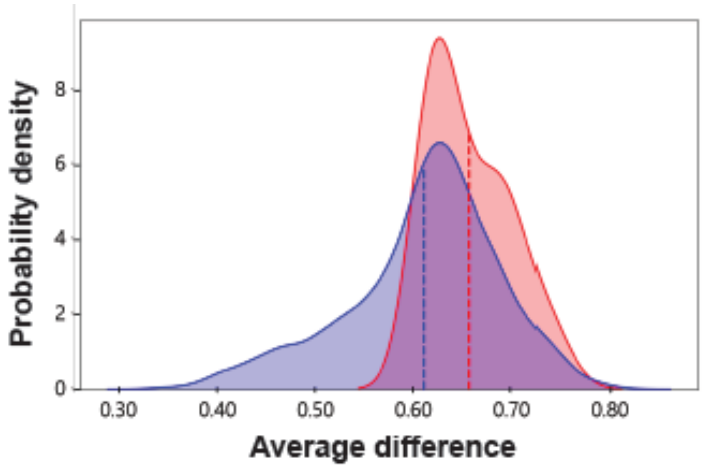
  Visual: $abs($[0.3, 0.4, . . ., 0.7] - [0.1, 0.8, . . ., 0.9]$)$

  Textual: $abs($[0.3, 0.4, . . ., 0.7] - [0.2, 0.4, . . ., 0.1]$)$

- (5) For each item, computing the average difference of f-similarity and those of p-similarity with every other item per modality

  Visual: $avg($[0.2, 0.4, . . ., 0.2]$)$

  Textual: $avg($[0.1, 0.0, . . ., 0.6]$)$



Other Experiment Result

- **The effects of exploiting all the item modalities (Clothing dataset)**
  - MARIO consistently achieves the best accuracy when it exploits all the item modalities

Model	NDCG@10	Recall@10	Pre@10
$\text{MARIO}_{w/o\ \text{V}}$	0.0233	0.0415	0.0042
$\text{MARIO}_{w/o\ \text{T}}$	0.0243	0.0460	0.0046
MARIO	**0.0259**	**0.0484**	**0.0048**