NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

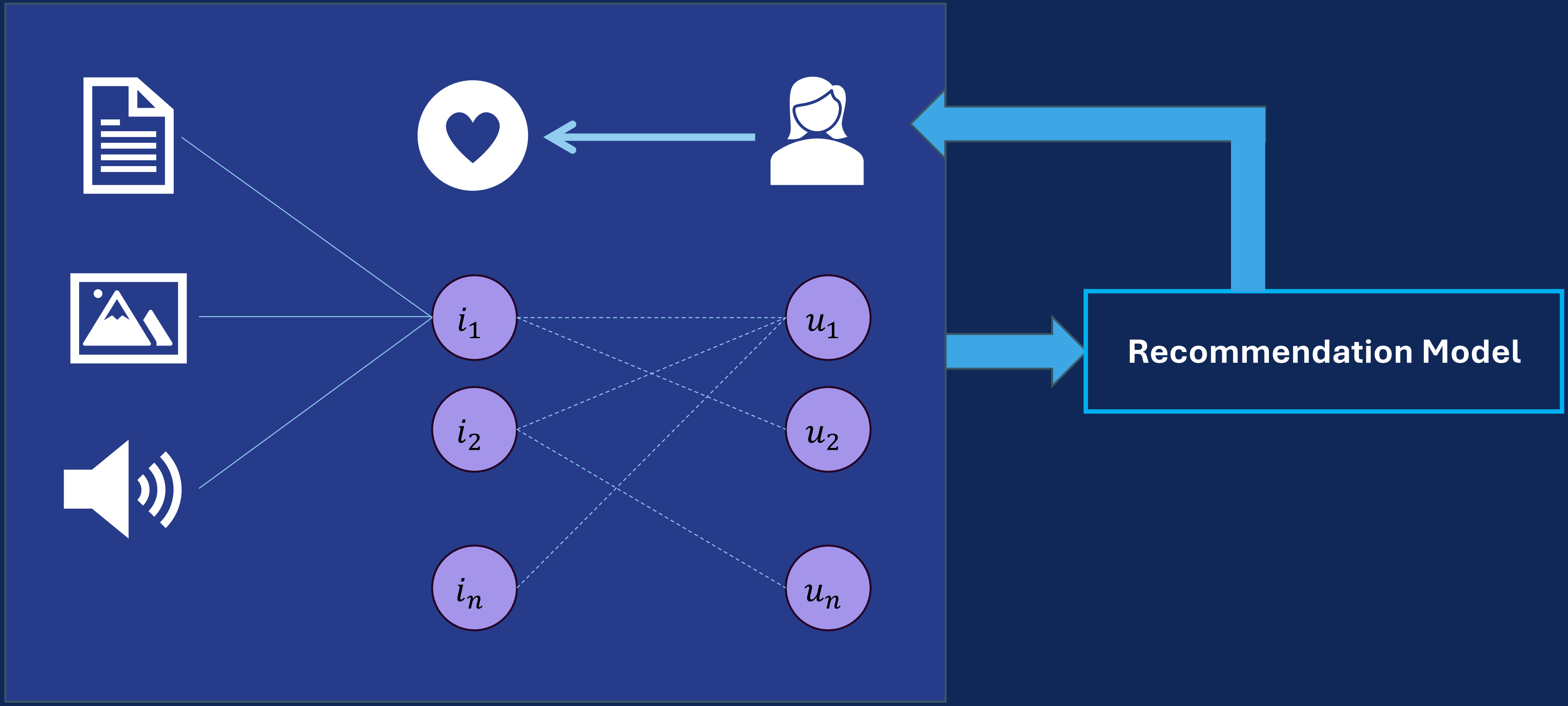
Multimodality In Recommender Systems: Does It Help, and Should We Expect An Answer?

Dr. Aixin Sun
NTU Singapore

@RecSys2025

Data Quality-Aware Multimodal Recommendation

Multimodal Recommender System



Text, Visual, Audio ... Interaction

- Three independent studies
- One central question

We are not short of interesting findings from large-scale evaluations.

Do Recommender Systems Really Leverage Multimodal Content? A Comprehensive Analysis on Multimodal Representations for Recommendation

Claudio Pomo*
claudio.pomo@poliba.it
Politecnico Di Bari
Bari, Italy

Matteo Attimonelli*
matteo.attimonelli@poliba.it
Politecnico Di Bari
Bari, Italy
Sapienza University of Rome
Rome, Italy

Danilo Danese*
danilo.danese@poliba.it
Politecnico Di Bari
Bari, Italy

Fedelucio Narducci
fedelucio.narducci@poliba.it
Politecnico Di Bari
Bari, Italy

Tommaso Di Noia
tommaso.dinoia@poliba.it
Politecnico Di Bari
Bari, Italy

Abstract

Multimodal Recom... tion accuracy by int... and textual metada... their gains stem fro... model complexity. ... item embeddings, ... the representations ... from standard extra...

Are Multimodal Embeddings Truly Beneficial for Recommendation? A Deep Dive into Whole vs. Individual Modalities

Yu Ye
University of Glasgow
Glasgow, United Kingdom
yu.jade.ye@gmail.com

Junchen Fu*
University of Glasgow
Glasgow, United Kingdom
j.fu.3@research.gla.ac.uk

Yu Song
Michigan State University
East Lansing, United States
songyu5@msu.edu

Kaiwen Zheng
University of Glasgow
Glasgow, United Kingdom
k.zheng.1@research.gla.ac.uk

Joemon M. Jose
University of Glasgow
Glasgow, United Kingdom
joemon.jose@glasgow.ac.uk

Abstract

Multimodal recommendation has emerged as a mainstream paradigm, typically leveraging text and visual embeddings extracted from pre-trained models such as Sentence-BERT, Vision Transformers, and ResNet. This approach is founded on the intuitive assumption that incorporating multimodal embeddings can enhance recommendation performance. However, despite its popularity, this assumption lacks comprehensive empirical verification. This presents a critical research gap. To address it, we pose the central research question of this paper: *Are multimodal embeddings truly beneficial for recommendation?*

alone does not. These results offer foundational insights and ... cal guidance for the multimodal recommendation communi... will release our code and datasets to facilitate future resear...

CCS Concepts

• Information systems → Recommender systems.

Keywords

Multimodal Recommendation, Multimodal Embeddings, M... Empirical Study

Does Multimodality Improve Recommender Systems as Expected? A Critical Analysis and Future Directions

HONGYU ZHOU, Nanyang Technological University, Singapore
YINAN ZHANG, Nanyang Technological University, Singapore
AIXIN SUN, Nanyang Technological University, Singapore
ZHIQI SHEN, Nanyang Technological University, Singapore

Multimodal recommendation systems are increasingly popular for their potential to improve performance by integrating diverse data types. However, the actual benefits of this integration remain unclear, raising questions about when and how it truly enhances recommendations. In this paper, we propose a structured evaluation framework to systematically assess multimodal recommendations across four dimensions: Comparative Efficiency, Recommendation Tasks, Recommendation Stages, and Multimodal Data Integration. We benchmark a set of reproducible multimodal models against strong traditional baselines and evaluate their performance on different platforms. Our findings show that multimodal data is particularly beneficial in sparse interaction scenarios and during the recall stage of recommendation pipelines. We also observe that the importance of each modality is task-specific, where text features are more useful in e-commerce and visual features are more effective in short-video recommendations. Additionally, we explore different integration strategies and model sizes, finding that Ensemble-Based Learning outperforms Fusion-Based Learning, and that larger models do not necessarily deliver better results. To deepen our understanding, we include case studies and review findings from other recommendation domains. Our work provides practical insights for building efficient and effective multimodal recommendation systems, emphasizing the need for thoughtful modality selection, integration strategies, and model design.

Before we start

We're Still Doing It (All) Wrong: Recommender Systems, Fifteen Years Later

Alan Said[1], Maria Soledad Pera[2] and Michael D. Ekstrand[3]

[1]University of Gothenburg, Gothenburg, Sweden
[2]TU Delft, Delft, Netherldans
[3]Drexel University, Philadelphia, USA

https://arxiv.org/pdf/2509.09414

On the Generalizability and Predictability of Recommender Systems

NeurIPS'22

Duncan McElfresh*[1], Sujay Khandagale*[1], Jonathan Valverde*[1,3], John P. Dickerson[2,3], Colin White[1]
[1]Abacus.AI, [2]ArthurAI, [3]University of Maryland

What News Recommendation Research Did (But Mostly Didn't) Teach Us About Building A News Recommender

Karl Higley[1], Robin Burke[2], Michael D. Ekstrand[3] and Bart P. Knijnenburg[4]

[1]Department of Computer Science and Engineering, University of Minnesota
[2]Department of Information Science, University of Colorado, Boulder
[3]Department of Information Science, Drexel University
[4]School of Computing, Clemson University

https://arxiv.org/pdf/2509.12361

Table 1: The relative performance of each rec-sys algorithm depends on the dataset and metric. This table shows the mean, min (best) and max (worst) rank achieved by all 20 algorithms over all 85 datasets, over 10 accuracy and hit-rate metrics at all cutoffs tested. This includes metrics NDCG, precision, recall, Prec.-Rec.-Min-density, hit-rate, F1, MAP, MAP-Min-density, ARHR, and MRR.

Rank	Item-KNN	P3alpha	SLIM-BPR	EASE-R	RP3beta	SVD	SLIM-ElasticNet	iALS	NMF	User-KNN	MF-Funk	TopPop	MF-Asy	MF-BPR	Mult-VAE	U-neural	GlobalEffects	CoClustering	Random	SlopeOne
Min.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	9	7
Max.	14	18	14	18	17	16	17	19	14	17	18	19	16	17	20	20	20	19	20	20
Mean	2.3	4.2	4.7	5.3	6	6	7	7	7.1	7.6	9.4	10.4	10.7	11.2	11.7	12.3	13.3	14.9	16.2	16.7

Reproducible Paper Collection

- Papers in the collection → 41 papers
  - Published in 2019 -- 2024 at top-tier venues: SIGIR, WWW, TKDE, CIKM, TOIS, AAAI, TMM, ACM MM.
  - Paper introduces a new technique and tackles issues related to multimodal RecSys.
- Reproducibility →12 papers
  - Code Reproducible: source code is available and functions correctly
  - Dataset Available: publicly accessible, or raw data with the preprocessing code
- Another 7 *code-reproducible* models were also included

Does Multimodality Improve Recommender Systems as Expected? A Critical Analysis and Future Directions

Datasets and Metrics

- Datasets
  - Amazon (Baby, Sports, Clothing, Art, and Beauty) --- E-commerce
  - Taobao dataset --- E-commerce
  - DY dataset --- Short-video
- Dataset split (following original papers' settings)
  - Random split (8:1:1)
  - Leave-one-out + Negative sampling (99 negative samples)

Not perfect setting

- Evaluation metric
  - Recall, HitRate, NDCG

Interaction Only vs. Multimodality

Model	Baby				Taobao				DY			
	Rec@10	NDCG@10	Rec@20	NDCG@20	Rec@10	NDCG@10	Rec@20	NDCG@20	Rec@10	NDCG@10	Rec@20	NDCG@20
ItemKNN	0.0566	0.0327	0.0830	0.0396	0.0554	0.0263	**0.0920**	0.0354	0.2920	0.1960	0.3477	0.2102
UserKNN	**0.0576**	**0.0328**	**0.0841**	**0.0396**	**0.0580**	**0.0277**	0.0908	**0.0360**	**0.2953**	**0.2000**	**0.3488**	**0.2138**
LATTICE	0.0547	0.0292	0.0850	0.0370	-	-	-	-	0.2491	0.1533	0.3247	0.1726
MICRO	0.0569	0.0315	0.0904	0.0401	-	-	-	-	0.2231	0.1332	0.2955	0.1517
BM3	0.0564	0.0301	0.0883	0.0383	0.0461	0.0189	0.0786	0.0270	0.2026	0.1199	0.2831	0.1405
FREEDOM	0.0627	0.0330	0.0992	0.0424	0.0439	0.0187	0.0776	0.0271	0.2162	0.1299	0.2874	0.1481
MGCN	0.0610	0.0328	0.0951	0.0416	-	-	-	-	0.2499	0.1523	0.3221	0.1708
LGMRec	0.0654	0.0353	0.0985	0.0439	0.0490	0.0217	0.0857	0.0309	0.2439	0.1506	0.3144	0.1686
MGCL	0.0678	0.0401	0.1027	0.0499	0.0583	0.0275	0.0974	0.0373	0.2924	0.1961	0.3667	0.2159
MCLN	0.0684	0.0392	0.1028	0.0487	0.0574	0.0254	0.1039	0.0369	0.2306	0.1505	0.3074	0.1709
MGCE	0.0720	0.0421	0.1100	0.0527	0.0612	0.0278	0.1027	0.0381	0.3062	0.2074	0.3777	0.2267
GUME	0.0684	0.0369	0.1040	0.0460	-	-	-	-	0.2711	0.1712	0.3383	0.1884
DiffMM	0.0612	0.0327	0.0933	0.0404	0.0490	0.0220	0.0872	0.0314	0.2244	0.1359	0.2982	0.1548
MENTOR	0.0651	0.0350	0.1027	0.0447	0.0502	0.0226	0.0891	0.0322	0.2416	0.1496	0.3068	0.1663
VBPR	0.0423	0.0223	0.0663	0.0284	0.0494	0.0237	0.0817	0.0318	0.2478	0.1519	0.3211	0.1706
MMGCN	0.0378	0.0200	0.0615	0.0261	0.0396	0.0184	0.0698	0.0259	0.1269	0.0696	0.1882	0.0853
GRCN	0.0539	0.0288	0.0833	0.0363	0.0550	0.0283	0.0890	0.0368	0.2650	0.1671	0.3359	0.1853
DualGNN	0.0448	0.0240	0.0716	0.0309	0.0570	0.0279	0.0973	0.0380	0.2384	0.1474	0.3088	0.1654
SLMRec	0.0529	0.0290	0.0775	0.0353	0.0518	0.0230	0.0858	0.0315	0.2568	0.1580	0.3316	0.1771
LightGT	0.0477	0.0250	0.0753	0.0314	0.0411	0.0186	0.0845	0.0304	0.1119	0.0595	0.1693	0.0745
MMSSL	0.0629	0.0353	0.0948	0.0441	0.0485	0.0210	0.0898	0.0313	0.2525	0.1574	0.3245	0.1760

Random split (8:1:1) | **Worse than interaction only** | **Better than interaction only**

Multimodality vs. Single Modality

Amazon: Text is more important

DY: Visual is more important; but interaction dominates

Model	Ablation Study	Baby w/o T	Baby w/o V	Baby Original	Baby Interaction Only	Taobao w/o V	Taobao Original	DY w/o T	DY w/o V	DY Original	DY Interaction Only
VBPR	✗	**0.0428**	0.0400	0.0423	0.0386	**0.0495**	0.0494	**0.2547**	0.2426	0.2478	0.2510
MMGCN	✓	**0.0384**	0.0365	0.0378	0.0342	**0.0545**	0.0396	0.1203	0.1154	**0.1269**	0.1199
GRCN	✓	0.0488	0.0517	**0.0539**	0.0485	**0.0567**	0.0550	0.2435	0.2367	0.2650	**0.2692**
DualGNN	✓	0.0511	**0.0612**	0.0448	0.0377	0.0329	**0.0570**	0.2430	0.2402	0.2384	**0.2534**
LATTICE	✗	0.0492	0.0546	**0.0547**	0.0469	-	-	**0.2544**	0.2515	0.2491	0.2484
MICRO	✗	0.0487	**0.0580**	0.0569	0.0409	-	-	0.2304	0.2348	0.2231	**0.2393**
SLMRec*	✗	0.0475	0.0495	**0.0529**	0.0476	**0.0548**	0.0518	0.2542	**0.2594**	0.2568	0.2544
BM3	✓	0.0544	**0.0571**	0.0564	0.0561	**0.0476**	0.0461	0.2078	0.2006	0.2026	**0.2082**
FREEDOM	✓	0.0501	0.0622	**0.0627**	0.0443	0.0412	**0.0439**	**0.2228**	0.2119	0.2162	0.2226
MMSSL*	✗	0.0507	0.0613	**0.0629**	0.0462	**0.0525**	0.0485	0.2505	0.2470	**0.2525**	0.2489
LightGT	✗	0.0394	0.0421	**0.0477**	0.0331	0.0281	**0.0411**	**0.2069**	0.0964	0.1119	0.0670
MGCN	✓	0.0528	**0.0640**	0.0610	0.0486	-	-	0.2249	0.2291	0.2499	**0.2585**
MGCL*	✓	0.0613	0.0663	**0.0678**	0.0569	0.0441	**0.0583**	0.2817	0.2886	**0.2924**	0.1940
MCLN	✓	0.0637	**0.0699**	0.0684	0.0461	0.0443	**0.0574**	**0.2686**	0.2576	0.2306	0.1969
MGCE*	✓	0.0634	0.0711	**0.0720**	0.0607	0.0537	**0.0612**	0.3129	**0.3130**	0.3062	0.2785
LGMRec*	✗	0.0499	0.0615	**0.0654**	0.0395	**0.0505**	0.0490	0.2405	**0.2441**	0.2439	0.2368
GUME*	✗	0.0556	0.0597	**0.0684**	0.0523	-	-	0.2643	0.2730	0.2711	**0.2741**
DiffMM*	✗	0.0533	0.0592	**0.0612**	0.0520	0.0475	**0.0490**	0.2273	0.2337	0.2244	**0.2381**
MENTOR*	✗	0.0510	**0.0668**	0.0651	0.0487	0.0479	**0.0502**	0.2543	0.2450	0.2416	**0.2596**

Random split (8:1:1)

Bold: best variant of the method

Gray background: best among all

Further Observations Made (in the paper)

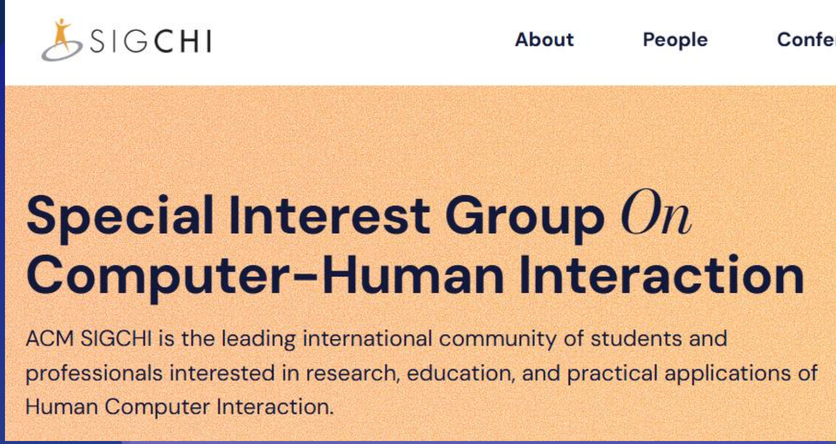
- In e-commerce settings, textual features often play a more important role
- For short video recommendations, visual information tends to be more useful
- Multimodal information tends to be more beneficial in the recall stage than in the re-ranking stage
- Ensemble-Based Learning seems to be more effective than Fusion-Based Learning
  - Fusion-based methods generate a unified embedding by merging modality features and interaction data early in the pipeline.
  - Ensemble-based methods produce separate predictions from each source and combine them at the final stage.

Multimodality In Recommender Systems: Does It Help, and Should We Expect An Answer?

Mixed observations
(from an imperfect setting)

Probably No!
but **WHY?**

RecSys is a conference under



- Significant focus on algorithmic innovation
- Less focus on ***user perspectives and interactions***
- Less focus on the full picture of recommender systems

Prompt to ChatGPT: I will give a talk on recommender system, draw a picture of recommender system for illustration purpose.

User

Interface

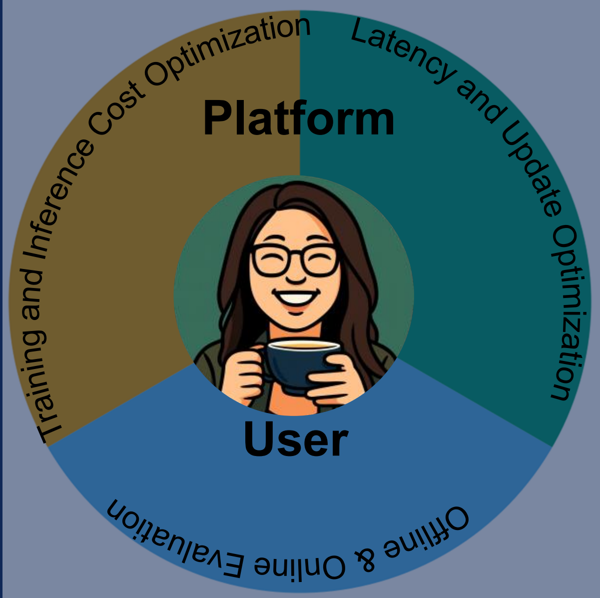
Conservation and Feedback

RecSys

Transaction-Oriented RecSys

- Data sparsity and cold start
- Real-time user preference modeling
- Multi-objective and long-term optimization

Objective: Prompting **transactional actions**, optimizing for conversion, revenue, or purchase likelihood



Content-Oriented RecSys

Video

- Agile capture of short-term interest
- Refined interest modeling
- Real-time interest response

News

- Timeliness and real-time performance
- Deep content understanding
- Multi-objective optimization

Audio

- Cold-start
- Diversity exploration
- Multi-objective and multi-signal selection

Objective: Facilitating **user consumption and engagement**, optimizing for dwell time, clicks, or user satisfaction

Other Recommendation Systems

POI, Socia community, Live streaming
Job opportunities, courses, and others

- Algorithm in a larger picture of RecSys
- Objective differences
- Various challenges
- **User-decision process**
  - Modality in ***right context***
  - E-commerce, short video, news, audio, food delivery, and job

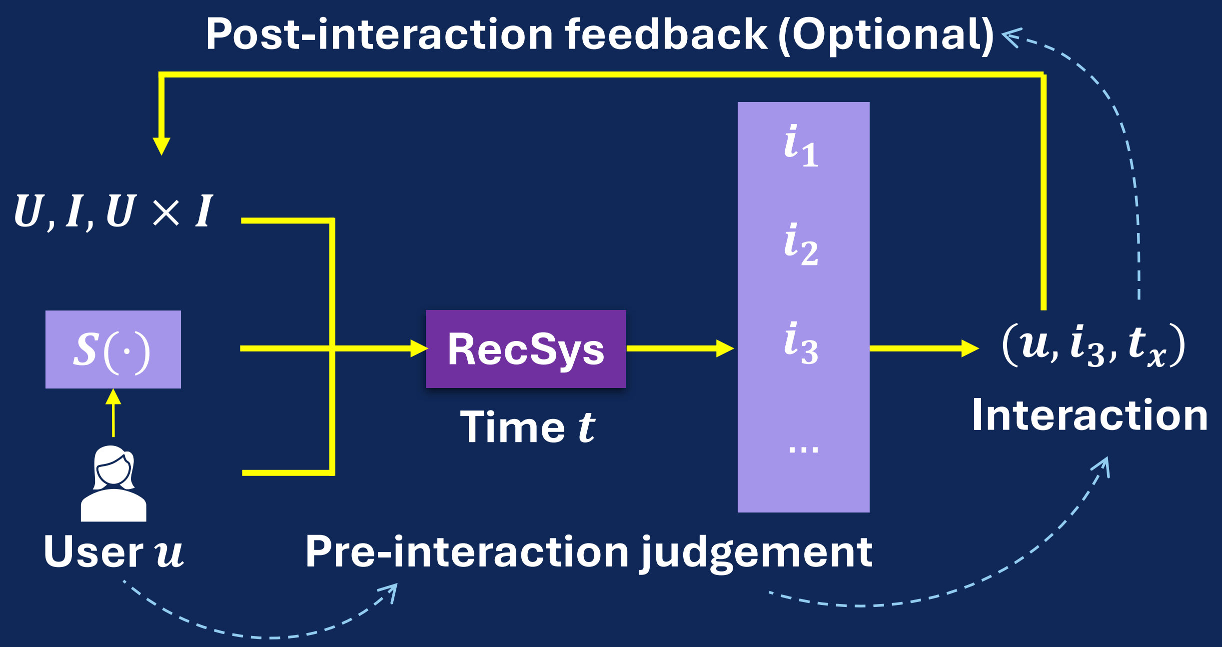
Different modalities: are they important?

- E-Commerce: product with image
- News articles: headline and preview image
  - How many news stories are presented to the users? In a list of grid?
  - Through a mobile screen or PC monitor?

- Short videos: do users even read the text?
  - Do users have a choice of the next recommended video?
  - Do user even see the preview image (except the very first video)?

- Audio
  - User may select the starting song (showing an image), then streaming

User-Item Interaction: The Life Cycle



- **Transaction-Oriented RecSys**
  - Order – delivery – interaction – feedback
  - User purchases an item – $U \times I$?
- **Content-Oriented RecSys**
  - Recommended – watch/listen/read
  - Feedback? → tolerance?

$S(\cdot)$: user specified selection criteria of items

To what extent multimodality impacts the $U \times I$?

Complexity of Interaction

Pre-Interaction Judgment	Recognition of User-Item Interaction

○ Informed vs Uninformed Decision
  ○ User has the knowledge to judge?
○ Item types
  ○ Single type: news, movie, music → similar criteria to judge
  ○ Multiple types: e-commerce → different criteria for different products
	○ Add to cart, payment, delivery, receive the product → CTR, Conversion Rate
	○ Absence of Pre-Interaction Judgment
	  ○ User selects the first item and the following are recommended (as a playlist) → skipping, fast forwarding, or continuing to watch/listen
	○ Unobservable Interaction
	  ○ Job recommendation → verification?

Different modality may play very different roles in different scenarios, and not all modalities may be even visible to users.

Multimodality In Recommender Systems: Does It Help, and Should We Expect An Answer?

We may have an answer for one specific recommendation scenario

Acknowledgement

- Hongyu Zhou
- Yinan Zhang
- Zhiqi Shen
- Kuan Zou

Multimodality evaluation

arXiv:2508.05377 (cs)

[Submitted on 7 Aug 2025]

Does Multimodality Improve Recommender Systems as Expected? A Critical Analysis and Future Directions

Hongyu Zhou, Yinan Zhang, Aixin Sun, Zhiqi Shen

https://arxiv.org/abs/2508.05377

Task classification and challenges

arXiv:2509.06002 (cs)

[Submitted on 7 Sep 2025]

A Survey of Real-World Recommender Systems: Challenges, Constraints, and Industrial Perspectives

Kuan Zou, Aixin Sun

https://arxiv.org/abs/2509.06002

Task formulation and interaction

arXiv:2503.21188 (cs)

[Submitted on 27 Mar 2025 (v1), last revised 15 Apr 2025 (this version, v2)]

Are We Solving a Well-Defined Problem? A Task-Centric Perspective on Recommendation Tasks

Aixin Sun

https://arxiv.org/abs/2503.21188

Thank You!

https://personal.ntu.edu.sg/axsun/