Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Generative Recommendation and Beyond

Tianxin Wei [1] Bowen Jin [1] Ruirui Li [2] Hansi Zeng [3] Zhengyang Wang [2] Jianhui Sun [4] Qingyu Yin [2] Hanqing Lu [2] Suhang Wang [5] Jingrui He [1] Xianfeng Tang [2]

[1]UIUC, [2]Amazon, [3]UMass, [4]UVa, [5]PSU

AIFARMS
Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

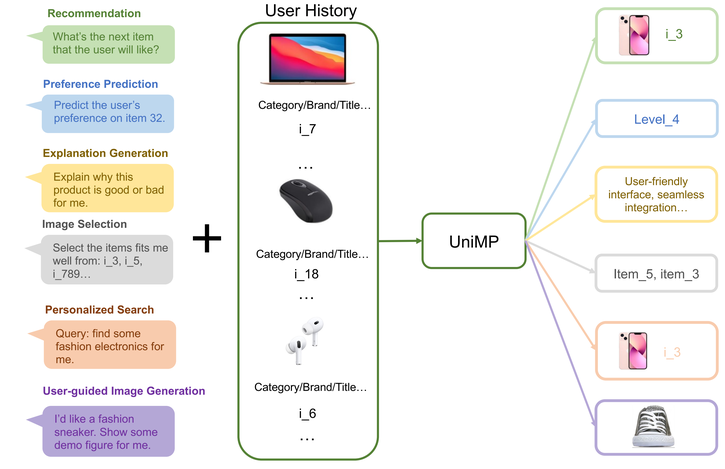
Background

Multi-modal Personalization In today's digitally driven landscape, individuals engage with diverse data types, such as ratings, images, descriptions, and prices, which are crucial for decision-making in visually-driven industries like fashion and retail.

- There is a need for systems that can integrate and harness these diverse data streams for improved personalization.
- Users often require a variety of information systems (search, recommendation, etc.) to meet their different needs.

Question: How to leverage multi-modal data to establish a Unified paradigm for Multi-modal Personalization systems (`UniMP`)?

Unified Multi-Modal Personalization



Decades of research have been dedicated to bolstering the personalization capability of models. However, previous methods suffer from **several key drawbacks**:

- *Struggle with multi-modal data*: Current methods struggle with processing and integrating diverse and complex user raw data including visuals, text, etc.
- *Fail to capture fine-grained preferences*: General corpora-based pre-trained models fail to capture user-specific preferences effectively.
- *Require intensive customization*: Existing approaches require extensive task-specific adjustments, increasing resource demands and reducing versatility.

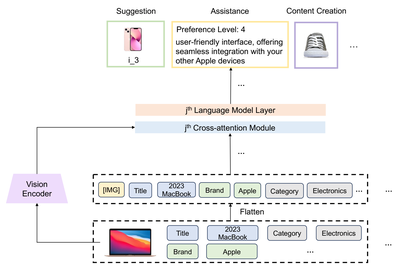
- **Motivation**: Foundational generative models with instructional tuning have demonstrated remarkable generalization abilities on a variety of tasks.
- **Propose**: In light of this, we develop a generic and extensible personalization generative framework to address the aforementioned limitations.

Paper Summary

- **Unified Data Format**: We've developed a format that integrates diverse user history types, supporting both multi-modal inputs and outputs tailored to individual needs.
- **Innovative Architecture**: Our model improves multi-modal information processing for precise preference prediction and includes a multi-task optimization to boost generalization.
- **Benchmarking Personalization Tasks**: We standardize multi-modal personalization tasks to evaluate various needs and conduct extensive tests, showing our model's superiority in performance and transferability.

Proposed Method: `UniMP`

Unified Data Formatting: We introduce a unified data format to seamlessly incorporate various types of user history information. For each item e, an attribute dictionary $\mathcal{D}_e$ is constructed, consisting of heterogeneous key-value attribute pairs $\mathcal{D}_e = \{(\text{"image"}, x), (k_1, v_1), (k_2, v_2), \ldots, (k_m, v_m)\}$. The dictionary is then flattened into a sequence $i_e = \text{Flatten}(\mathcal{D}_e)$ and combined to form a user sentence $u_n = \{[\text{CLS}], i_1, i_2, \ldots, i_n\}$.



Fine-grained Multi-modal User Modeling: UniMP employs a multi-modal model to process the user sequence containing both text and interleaved images. The visual input is processed by a vision encoder f_v, and the textual input is fed into a language model. The visual embeddings $x_v \in \mathbb{R}^{H \times C \times d}$ are used to condition the language model through a cross-attention mechanism at each layer j:

$$\tilde{x}_t^j = \text{Cross-Att}(Q = x_t^j, KV = x_v, M) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} \cdot M\right) V$$

Integration of Personalization Tasks: We formulate each multi-modal personalized task with the task description T as a next-token prediction objective:

$$\min -\log p(y|u) = -\log p(y_{>n}|u_{\leq n} = \{u_n, T\})$$

To unify and optimize multiple A tasks together, the final objective is formulated as joint multi-task learning:

$$\min \sum_{a=1}^{A} \lambda_a \cdot \mathbb{E}_{(u,y)\sim\mathcal{T}_a}\left[-\sum_{\ell=1}^{L} \log p(y_{>n}|u_{\leq n} = \{u_n, T_a\})\right]$$

UniMP employs context reconstruction to enrich the model's learning objectives and token-level re-weighting to adjust the importance of tokens based on their difficulty.

Experiments

We propose multiple personalized multi-modal tasks on multiple domains of Amazon dataset to evaluate the performance of our proposed `UniMP`: (i) Personalized recommendation; (ii) Personalized preference prediction; (iii) Personalized explanation generation; (iv) Personalized multi-modal selection; (v) Personalized multi-modal search; (vi) Personalized user-guided image generation.

Table: Experimental Comparisons on Product Recommendation.

	HR@3	NDCG@3	MRR@3	HR@5	NDCG@5	MRR@5
MF	0.0105	0.0077	0.0065	0.0165	0.0093	0.0078
MACR	0.0110	0.0080	0.0068	0.0170	0.0105	0.0091
LightGCN	0.0142	0.0103	0.0088	0.0206	0.0129	0.0094
UltraGCN	0.0151	0.0111	0.0095	0.0215	0.0134	0.0102
HGN	0.0167	0.0113	0.0113	0.0231	0.0153	0.0117
GRU4Rec	0.0132	0.0101	0.0086	0.0201	0.0128	0.0096
SASRec	0.0189	0.0124	0.0102	0.0276	0.0175	0.0126
S^3-Rec	0.0205	0.0149	0.0133	0.0311	0.0193	0.0156
BERT4Rec	0.0121	0.0092	0.0079	0.0198	0.0127	0.0105
UniSRec	0.0201	0.0146	0.0135	0.0281	0.0191	0.0158
P5	0.0086	0.0056	0.0042	0.0124	0.0074	0.0061
VIP5+	0.0175	0.0125	0.0108	0.0262	0.0163	0.0127
VBPR	0.0114	0.0084	0.0071	0.0181	0.0102	0.0086
CausalRec	0.0143	0.0105	0.0088	0.0229	0.0146	0.0121
MMGCL	0.0151	0.0112	0.0095	0.0241	0.0159	0.0130
MMSSL	0.0181	0.0132	0.0112	0.0281	0.0194	0.0164
UniMP (Ours)	**0.0248**	**0.0194**	**0.0176**	**0.0337**	**0.0231**	**0.0196**

UniMP consistently enhances performance across various tasks, with further details in the paper. We also test on additional datasets, including Netflix and HM.

`UniMP` learns generalizable model

Table: Experiments of Generalization Ability of UniMP.

	HR@5	NDCG@5	MRR@5
P5	0.0119	0.0079	0.0062
VIP5	0.0217	0.0143	0.0124
S^3-Rec	0.0228	0.0140	0.0123
UniSRec	0.0245	0.0152	0.0131
UniMP (Ours)	**0.0364**	**0.0263**	**0.0228**

(a) New User Recommendation

	HR@5	NDCG@5	MRR@5
P5	0.0141	0.0081	0.0062
VIP5	0.0235	0.0167	0.0148
S^3-Rec	0.0226	0.0152	0.0128
UniSRec	0.0278	0.0184	0.0175
UniMP (Ours)	**0.0433**	**0.0289**	**0.0242**

(b) New Domain Recommendation

- We evaluate the generalization capabilities of various methods on new users and domains not included in the training set.
- UniMP shows superior performance in transfer learning due to its effective use of diverse user histories.

Acknowledgement This work is supported by National Science Foundation under Award No. IIS-2117902, and Agriculture and Food Research Initiative (AFRI) grant no.2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.