Multimodal Generation for Recommendation

Rui Zhang

www.ruizhang.info

Multimodal Generation

Prompt: Several giant wooly mammoths approach treading through a snowy meadow, their long wooly fur lightly blows in the wind as they walk, snow covered trees and dramatic snow capped... +

midjourney

https://www.midjourney.com/

Sora

https://openai.com/index/sora/

Can we make them personal?

Table of Contents

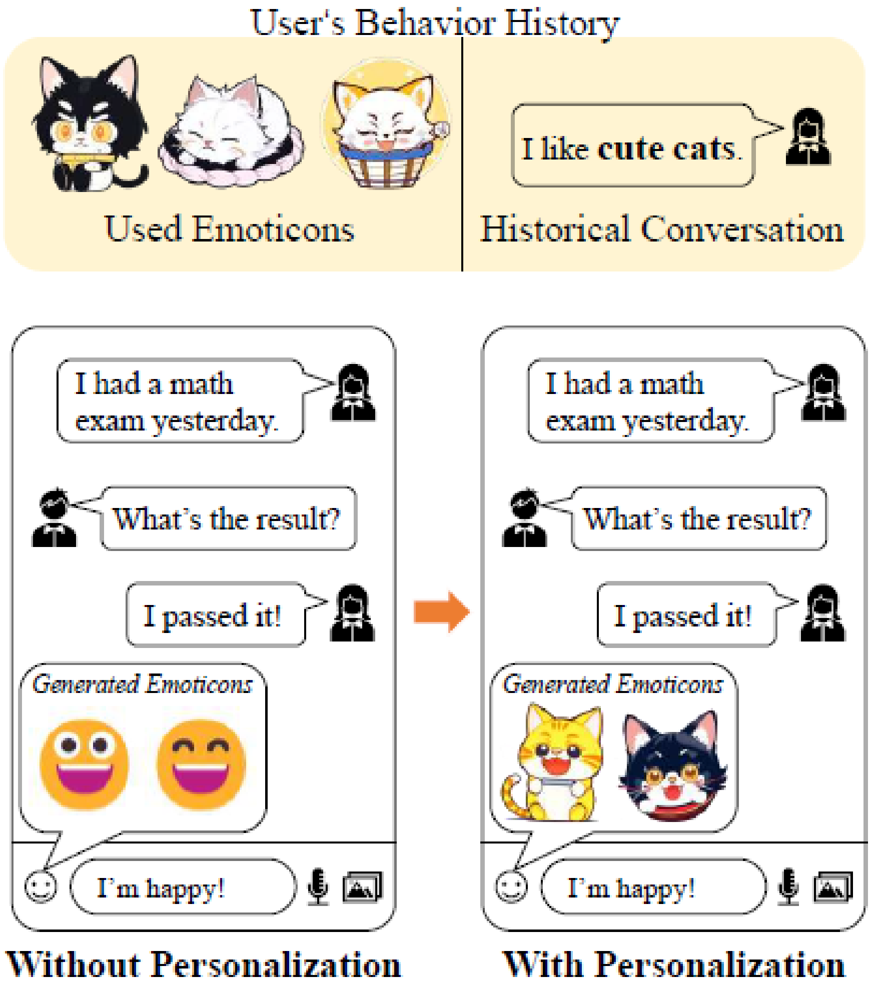
- **PMG (Personalized Multimodal Generation)**
  - PMG for Recommendation: multimodal → image with LLM
  - PMG for Preference Questions: multimodal → multimodal with Vision-Language Model
- **Personalized Generation**
  - Personalized Generation: text → text without LLM
  - Personalized Generation: item → text without LLM
  - Personalized Generation: text → text with LLM
  - Personalized Generation: text → text with LLM & Human
  - (non-Personalized) Multimodal Generation: multimodal → multimodal
- **Other Tasks of Multimodal Generation for Recommendation**
- **What's Next?**

Term: LLM – language models with capabilities similar to chatgpt, such as llama, claude, gemini, etc

Multimodal Pretraining and Generation for Recommendation: A Tutorial, Web Conference 2024
Multimodal Pretraining, Adaptation, and Generation for Recommendation: A Survey, arXiv:2404.00621

PMG for Recommendation: multimodal → image w/ LLM

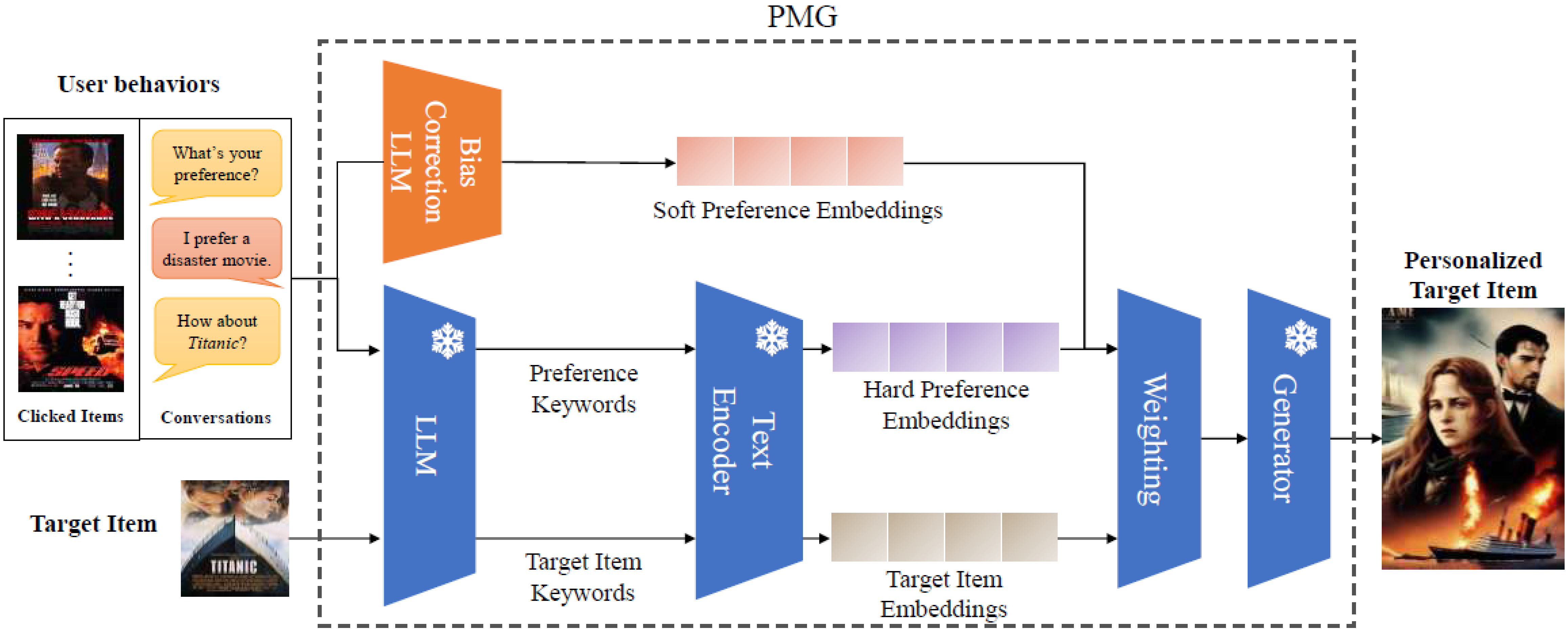
- **PMG: Personalized Multimodal Generation with LLM**
  - Converts user behaviors (conversations, clicks, etc) into natural language
  - Extract user preference descriptions, both hard and soft preference embeddings
  - Preference conditioned multimodal generation
  - Improves 8% in terms of personalization measure



PMG : Personalized Multimodal Generation with Large Language Models, The Web Conference 2024

Friday 17 May 2024: 2:30 - 4pm Poster Session

PMG for Recommendation: multimodal → image w/ LLM



PMG for Recommendation: multimodal → image w/ LLM

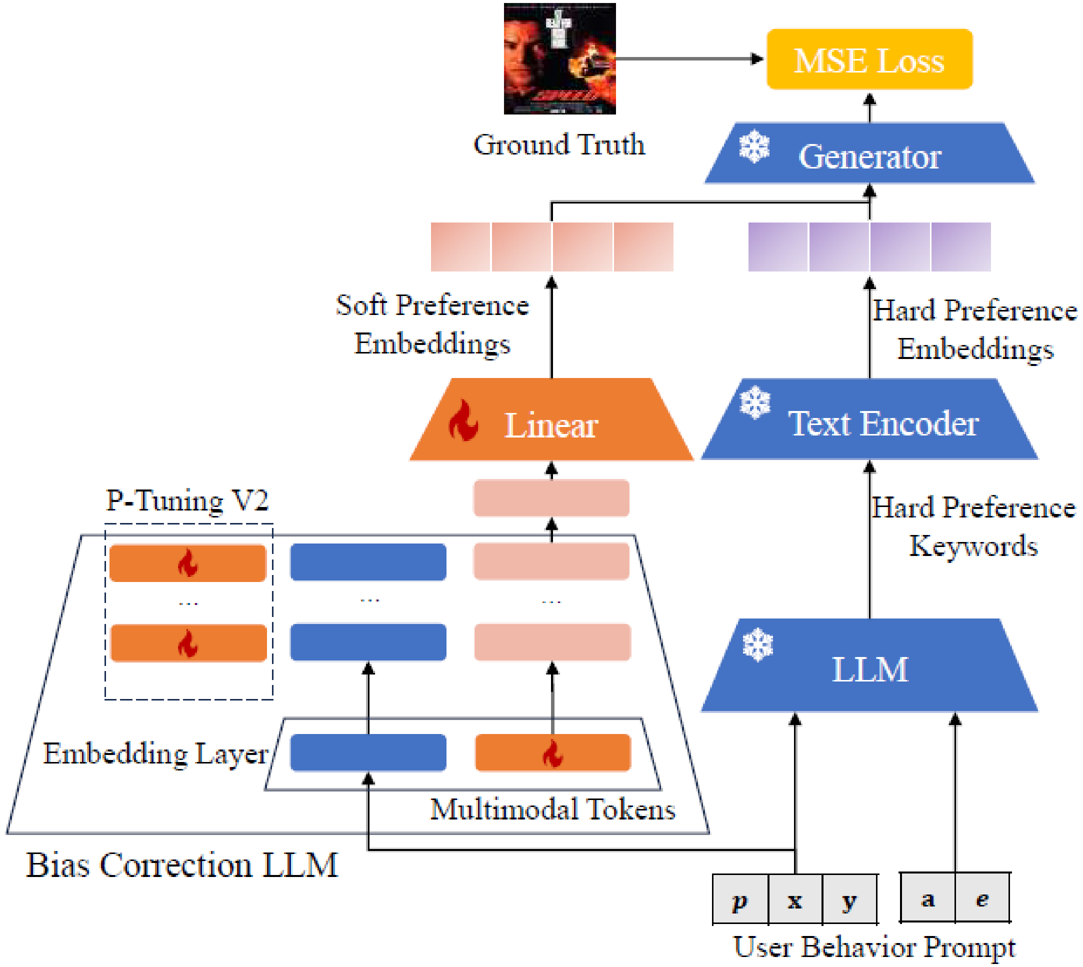


Figure 3: Model designed to train soft preference embeddings.

$$\mathrm{E}^p = concatenate(\mathrm{E}_m, \mathrm{E}_k)$$

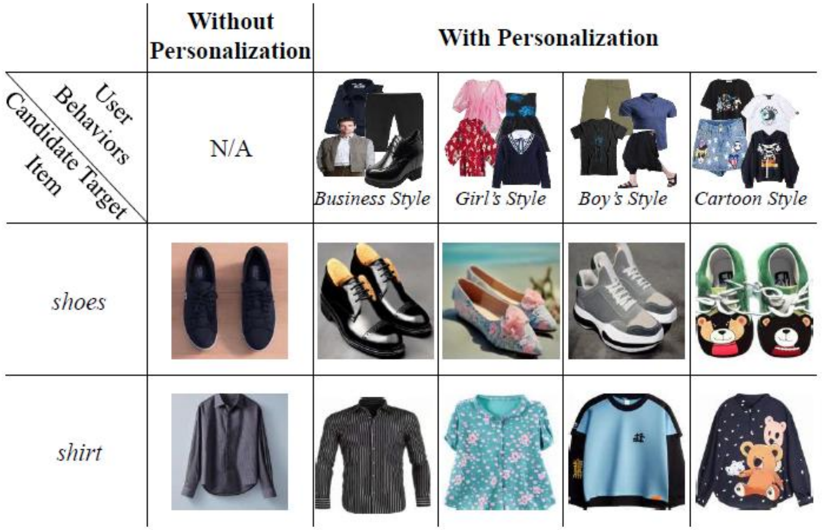
$$M_n = M_s + \epsilon,$$

$$M_d = Unet(\mathrm{E}^p, M_n).$$

The loss is calculated as MSE loss of M_s and M_d:

$$loss = MSE(M_s, M_d).$$

PMG for Recommendation: multimodal → image w/ LLM

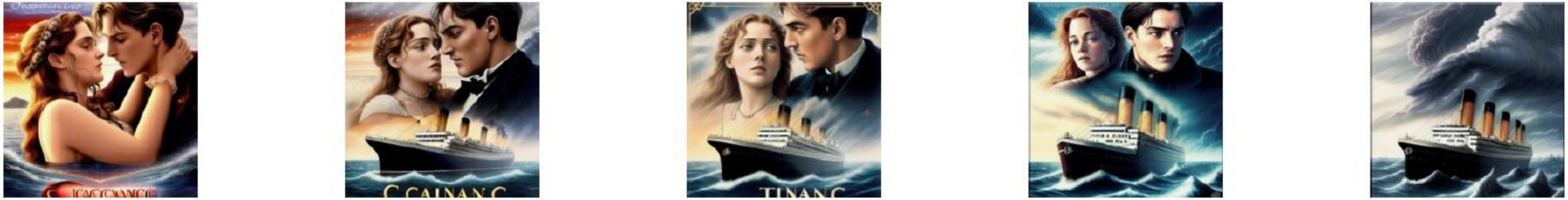


$$d_p = \frac{e_M \cdot e_p}{\|e_M\|_2 \, \|e_p\|_2},$$

$$d_t = \frac{e_M \cdot e_t}{\|e_M\|_2 \, \|e_t\|_2}.$$

Finally, our objective is to optimize the weighted sum of d_p and d_t.

$$z = \alpha \cdot \log d_p + (1 - \alpha) \cdot \log d_t .$$



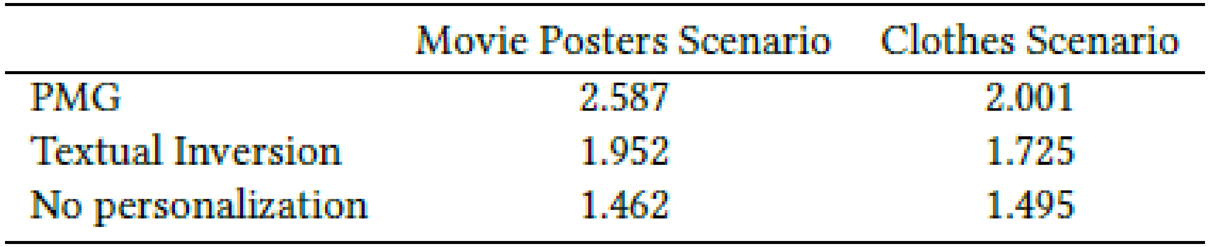
(a) $w_p : w_t = 0 : 4$ (b) $w_p : w_t = 1 : 3$ (c) $w_p : w_t = 2 : 2$ (d) $w_p : w_t = 3 : 1$ (e) $w_p : w_t = 4 : 0$

Figure 7: Generated poster of movie *Titanic* with different weights of conditions. w_p is the weight of preference conditions, which prefer disaster movie. w_t is the weight of target item conditions, which consider it as a romantic movie. When $w_p : w_t = 1 : 3$ it achieves the highest z score and the generated poster is a combination of romance and disaster.

PMG for Recommendation: multimodal → image w/ LLM

- **Data**

1) Generating personalized images of products whose original images are missing according to the historically clicked products of the user. POG dataset, a multimodal dataset of fashion clothes. We selected 2,000 users and 16,100 items for experiments.
2) Generating personalized posters of movies according to historical watched movies of user. MovieLens Latest Datasets, 9,000 movies, 600 users, and 100,000 rating interactions.
3) Generating emoticons in instant messaging according to current conversation and historically used emoticons of the user. We do not train soft preference embeddings and only use keywords to generate images.

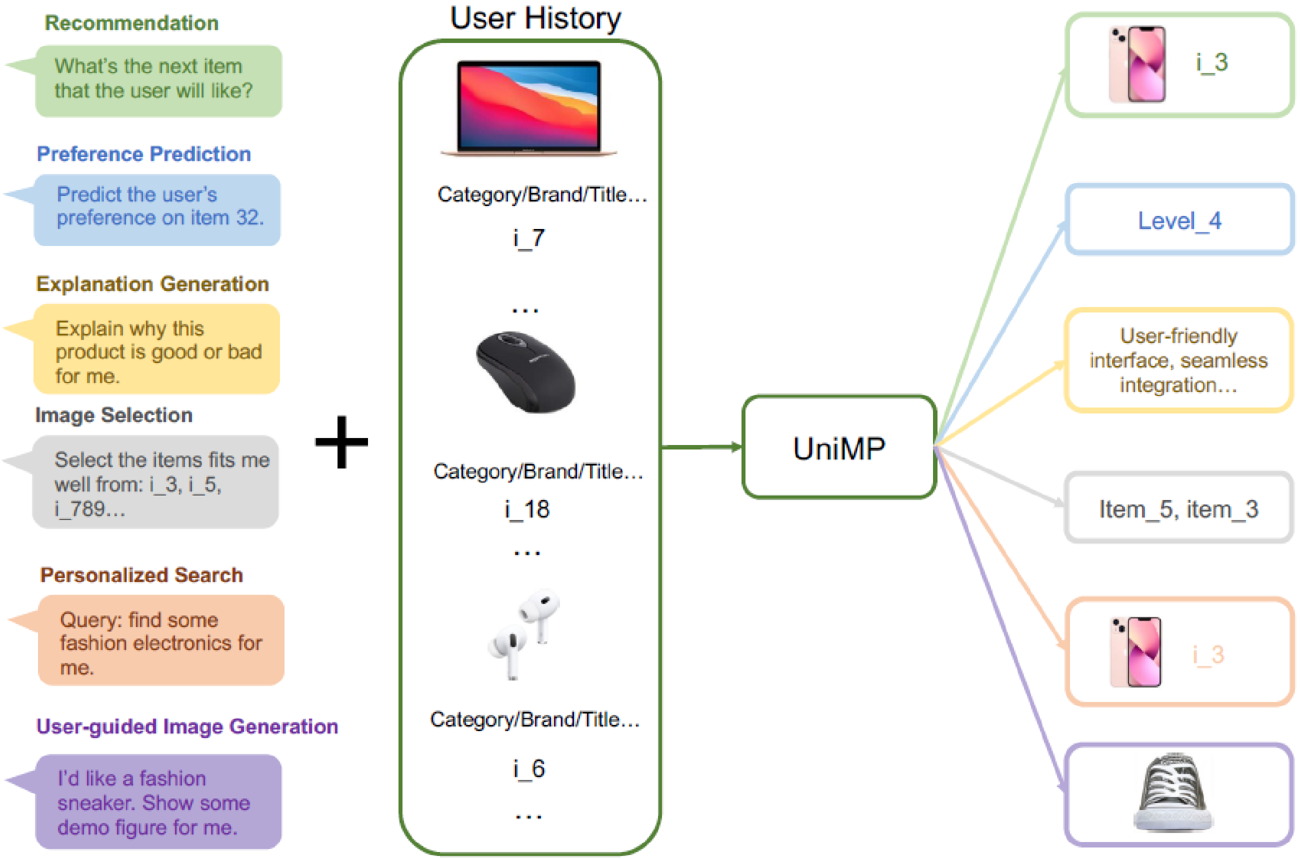


	Movie Posters Scenario	Clothes Scenario
PMG	2.587	2.001
Textual Inversion	1.952	1.725
No personalization	1.462	1.495

Human evaluation score, range (1, 2, 3)

PMG for Preference Questions: multimodal → multimodal w/ V-LM

- **Multi-task Multimodal generation, answering different types of questions**



Recommendation
What's the next item that the user will like?
Preference Prediction
Predict the user's preference on item 32.
Explanation Generation
Explain why this product is good or bad for me.
Image Selection
Select the items fits me well from: i_3, i_5, i_789…
Personalized Search
Query: find some fashion electronics for me.
User-guided Image Generation
I'd like a fashion sneaker. Show some demo figure for me.
+
User History
Category/Brand/Title…
i_7
…
Category/Brand/Title…
i_18
…
Category/Brand/Title…
i_6
…
UniMP
i_3
Level_4
User-friendly interface, seamless integration…
Item_5, item_3
i_3

Figure 2: Through multi-task, multi-modal instruction tuning, the model can adapt to a range of user requirements. By altering the instructions, it can generate diverse responses to suit user needs. For

Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Generative Recommendation and Beyond, ICLR 2024

PMG for Preference Questions: multimodal → multimodal w/ V-LM

- **Item contextual data is serialized and processed through fine-grained cross-modal fusion**

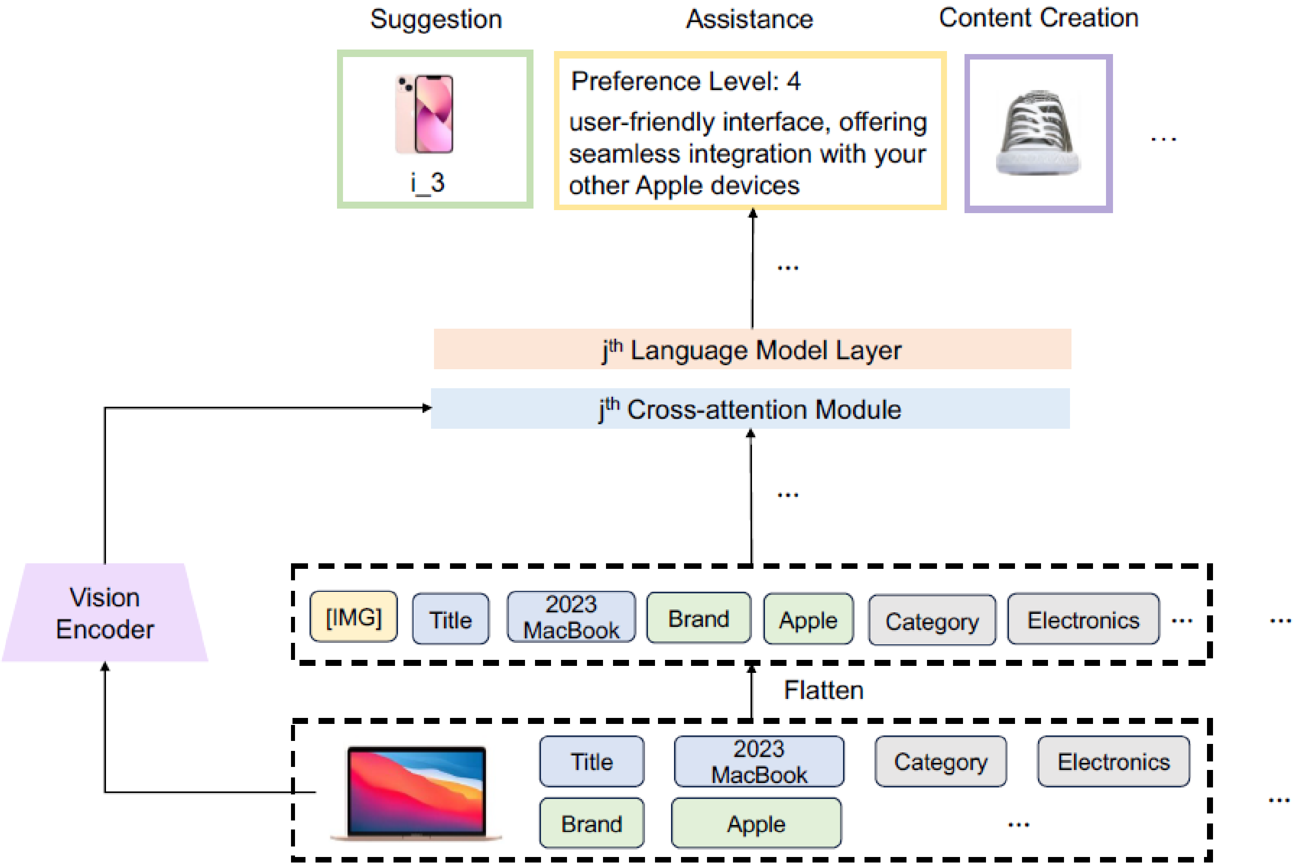
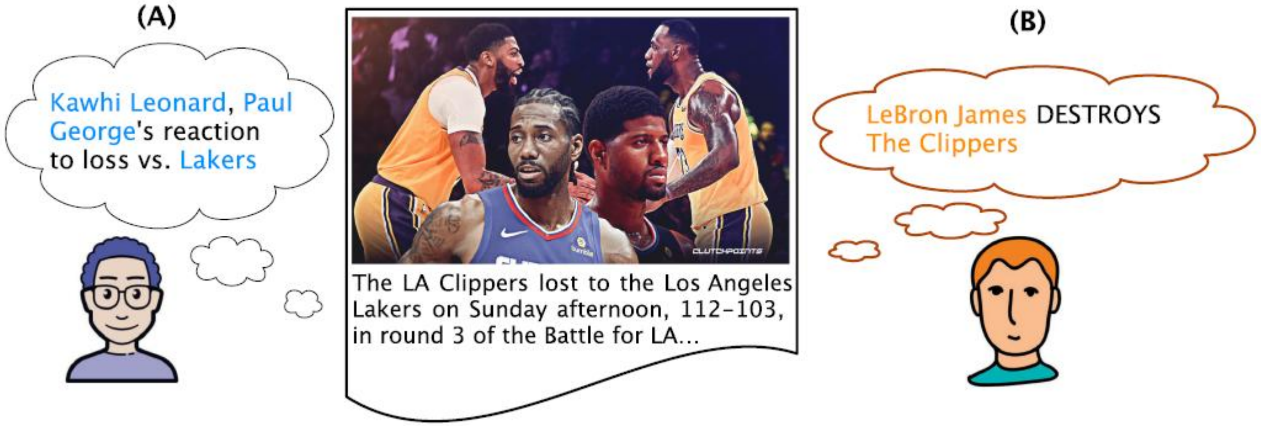


Figure 1: Our proposed UniMP framework operates as follows: Item contextual data is streamlined into a user sequence, which is then processed through fine-grained cross-modal fusion. Depending on the instructions, the output is tailored to produce diverse response types.

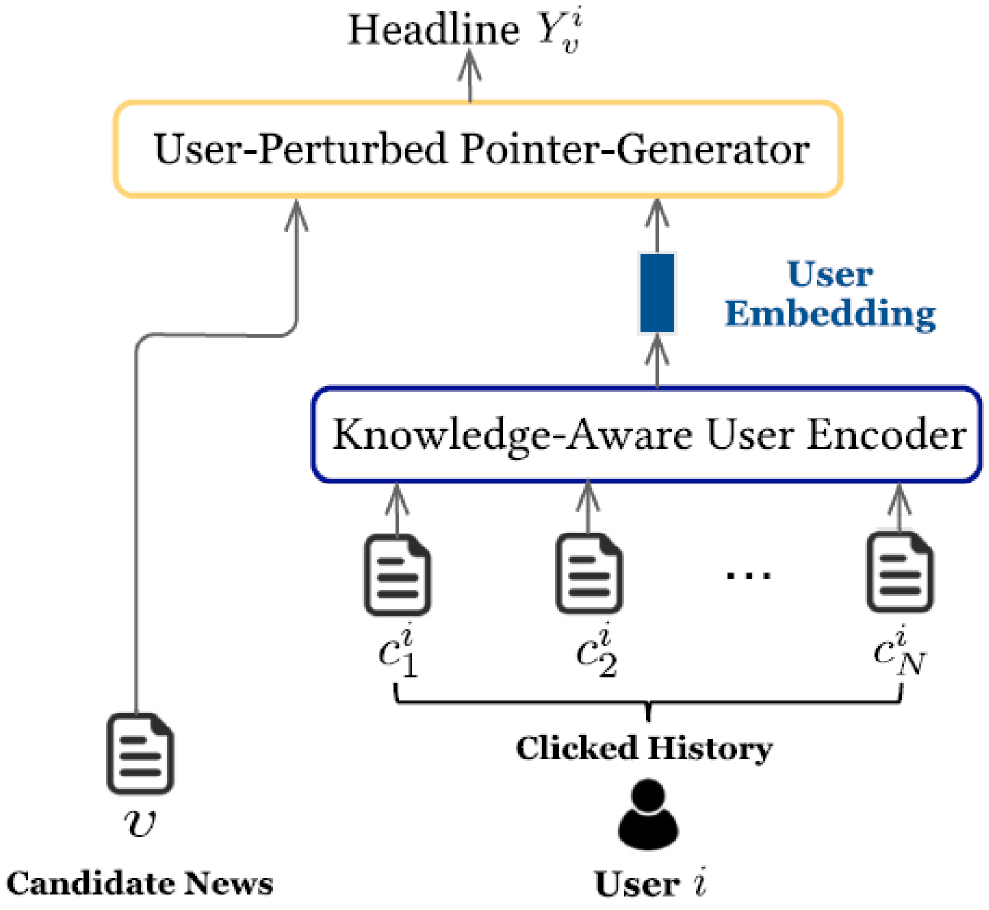
Personalized Generation: text → text w/o LLM

News Headline Generation


(A)
Kawhi Leonard, Paul George's reaction to loss vs. Lakers
The LA Clippers lost to the Los Angeles Lakers on Sunday afternoon, 112-103, in round 3 of the Battle for LA...
(B)
LeBron James DESTROYS The Clippers

Put Your Voice on Stage: Personalized Headline Generation for News Articles, TKDD 2023

- Framework
- Evaluation
  - Automtaic
    - Informativeness: F1 ROUGE
    - Fluency: longest common subsequence (ROUGE-L)
  - Human evaluation

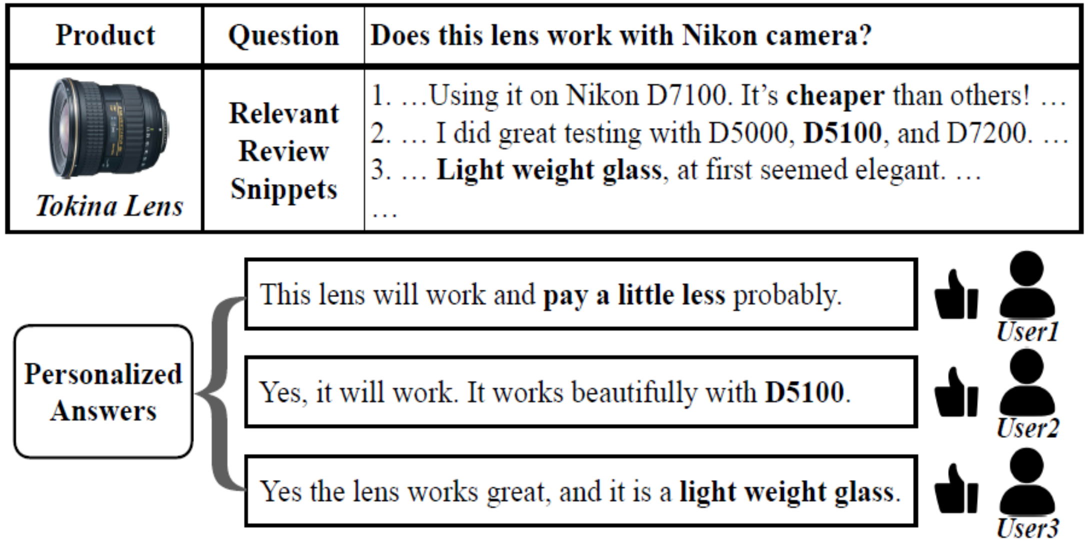


Framework

Put Your Voice on Stage: Personalized Headline Generation for News Articles, TKDD 2023

Personalized Generation: item → text w/o LLM

Personalized Answer Generation in E-commerce



Product	Question	Does this lens work with Nikon camera?
Tokina Lens	Relevant Review Snippets	1. …Using it on Nikon D7100. It's **cheaper** than others! … 2. … I did great testing with D5000, **D5100**, and D7200. … 3. … **Light weight glass**, at first seemed elegant. … …

Personalized Answers

- This lens will work and **pay a little less** probably. — *User1*
- Yes, it will work. It works beautifully with **D5100**. — *User2*
- Yes the lens works great, and it is a **light weight glass**. — *User3*

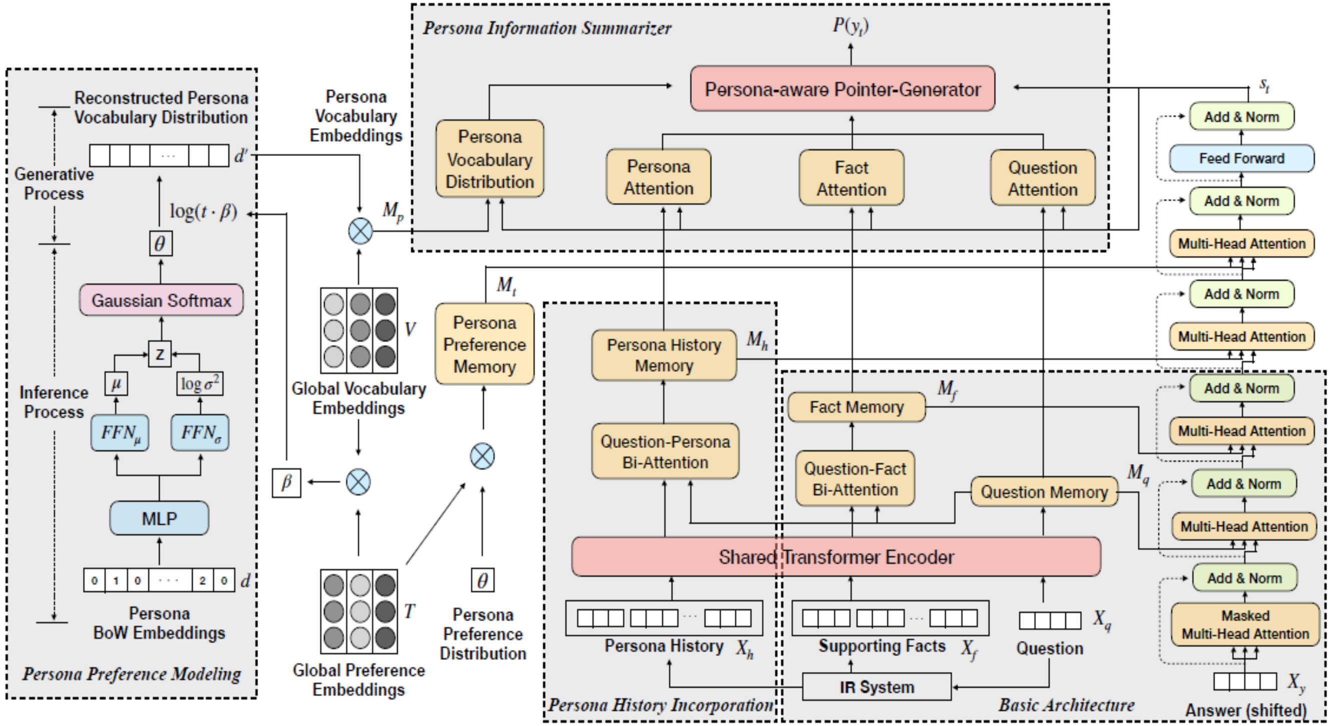


Fig. 3. Overview of the proposed method PAGE, including four components: (1) Basic Encoder-decoder Architecture, (2) Persona History Incorporation, (3) Persona Preference Modeling, and (4) Persona Information Summarizer.

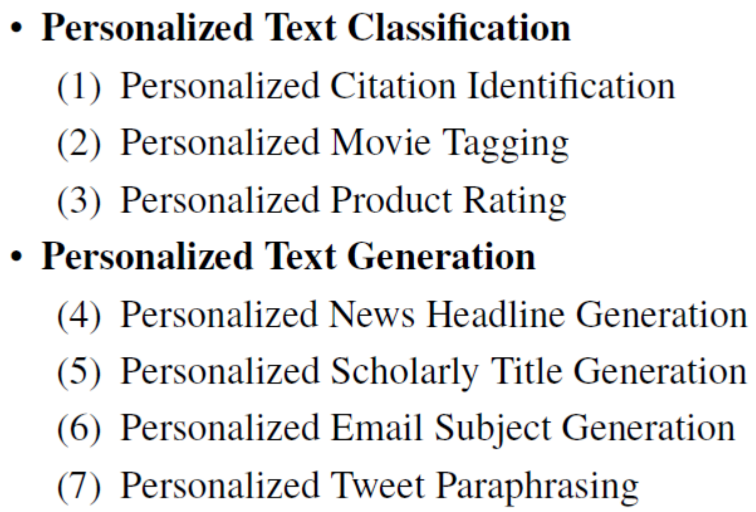
Towards Personalized Answer Generation in E-Commerce via Multi-Perspective Preference Modeling, TOIS 2022

Personalized Generation: text → text w/ LLM

Benchmark, RAG (Retrieval Augmented Generation) paradigm

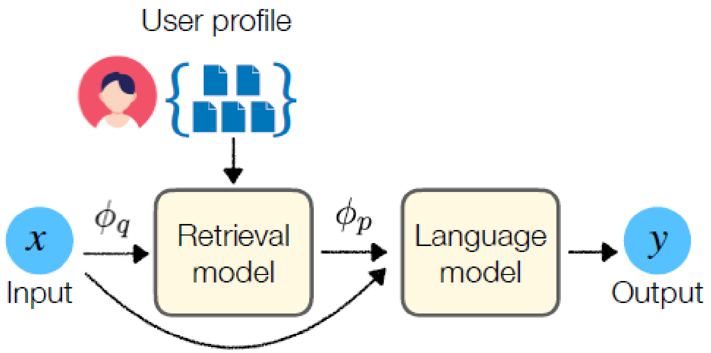
LaMP: When Large Language Models Meet Personalization, arXiv:2304.11406

- 7 Tasks



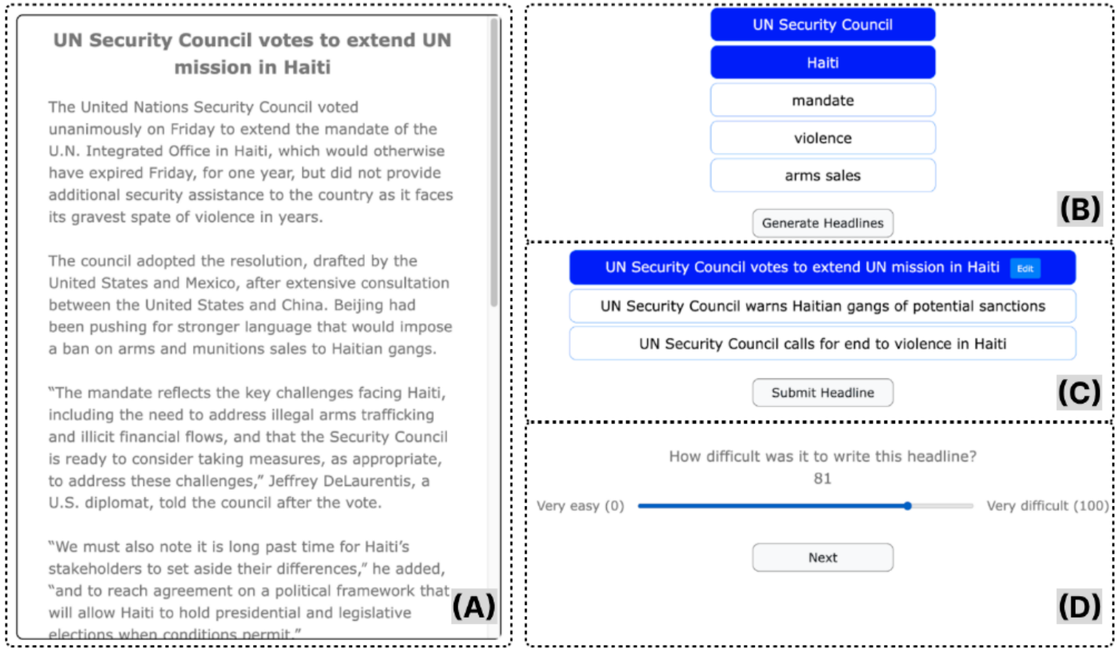
- **Personalized Text Classification**
  (1) Personalized Citation Identification
  (2) Personalized Movie Tagging
  (3) Personalized Product Rating
- **Personalized Text Generation**
  (4) Personalized News Headline Generation
  (5) Personalized Scholarly Title Generation
  (6) Personalized Email Subject Generation
  (7) Personalized Tweet Paraphrasing

- Using RAG paradigm



Personalized Generation: text → text w/ LLM & Human

- **LLM-assisted news headline generation**
  - Human-AI Text Co-Creation



UN Security Council votes to extend UN mission in Haiti

The United Nations Security Council voted unanimously on Friday to extend the mandate of the U.N. Integrated Office in Haiti, which would otherwise have expired Friday, for one year, but did not provide additional security assistance to the country as it faces its gravest spate of violence in years.

The council adopted the resolution, drafted by the United States and Mexico, after extensive consultation between the United States and China. Beijing had been pushing for stronger language that would impose a ban on arms and munitions sales to Haitian gangs.

"The mandate reflects the key challenges facing Haiti, including the need to address illegal arms trafficking and illicit financial flows, and that the Security Council is ready to consider taking measures, as appropriate, to address these challenges," Jeffrey DeLaurentis, a U.S. diplomat, told the council after the vote.

"We must also note it is long past time for Haiti's stakeholders to set aside their differences," he added, "and to reach agreement on a political framework that will allow Haiti to hold presidential and legislative elections when conditions permit."

UN Security Council votes to extend UN mission in Haiti

UN Security Council warns Haitian gangs of potential sanctions

UN Security Council calls for end to violence in Haiti

How difficult was it to write this headline?

81

Figure 2: Interface for human-AI news headline co-creation for *guidance* + *selection* + *post-editing* condition: (A) news reading panel, (B) perspectives (keywords) selection panel (multiple keywords can be selected), (C) headline selection panel with post-editing capability, and (D) difficulty rating slider. Note: (B), (C) and (D) are hidden from the user until the requisite step is finished (e.g., the user does not see the difficulty

Harnessing the Power of LLMs: Evaluating Human-AI Text Co-Creation through the Lens of News Headline Generation, EMNLP 2023

(non-Personalized) Multimodal Generation: multimodal → multimodal

- **Multi-modal News Headline Generation**

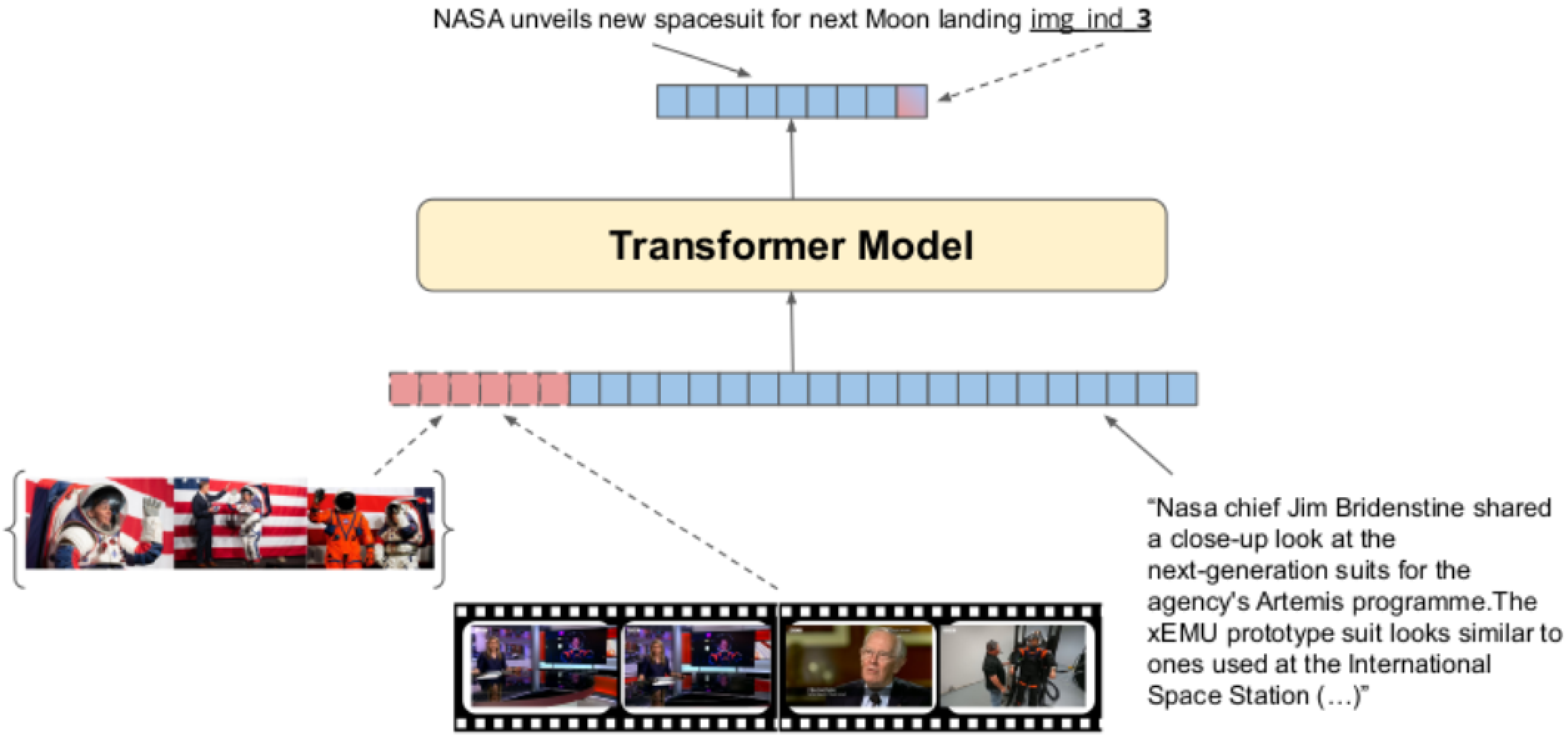

NASA unveils new spacesuit for next Moon landing img_ind_3
Transformer Model
"Nasa chief Jim Bridenstine shared a close-up look at the next-generation suits for the agency's Artemis programme.The xEMU prototype suit looks similar to ones used at the International Space Station (…)"

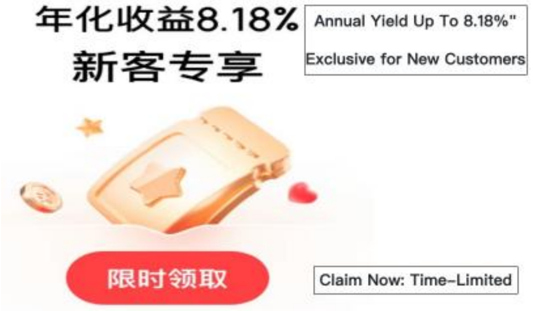
Figure 1: Overview of the proposed unified approach to MSMO. The visual tokens are appended to the text representation. The generated output includes the textual summary and the *index token* that indicates which input image (first, second, third, etc.) is picked as the pictorial summary. During training, a mixture of video-based, image-based, and text-only data is used.

Towards Unified Uni- and Multi-modal News Headline Generation, EACL 2024

Other Tasks of Multimodal Generation for Recommendation

Marketing Copy Generation

- Generate the promotional copy



GCOF: Self-iterative Text Generation for Copywriting Using Large Language Model, arXiv:2402.13667

Explanation Generation

- Generate reasons why an item is recommended

Personalized Reason Generation for Explainable Song Recommendation. TIST 2019

Dialogue Generation

- Generate questions for clarification during conversational search

Zero-shot Clarifying Question Generation for Conversational Search, Web Conference 2023

What’s Next

- **Multimodal → multimodal for Recommendation**
- **Improve the control of correctness (text, image, video, etc)**
- **Include more modalities, such as audio, video**
- **Interactive multimodal generation**

Thanks and Questions?

Hiring junior academics, postdocs, PhD students

Contact email:

rayteam@yeah.net