2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825030

Visual Summary Thought of Large Vision-Language Models for Multimodal Recommendation

Yuqing Liu*†, Yu Wang*†, Yuwei Cao‡, Lichao Sun§, Philip S. Yu†
†University of Illinois at Chicago, Chicago, IL, United States
{yliu363, ywang617, psyu}@uic.edu
‡Meta, Menlo Park, CA, United States
yuweicao@meta.com
§Lehigh University, Bethlehem, PA, United States
lis221@lehigh.edu

***Abstract*—The evolution of large vision-language models (LVLMs) has shed light on the development of many fields, particularly for multimodal recommendation. While LVLMs offer an integrated understanding of textual and visual information of items from user interactions, their deployment in this domain remains limited due to inherent complexities. First, LVLMs are trained from enormous general datasets and lack knowledge of personalized user preferences. Second, LVLMs struggle with multiple image processing, especially with discrete, noisy, and redundant images in recommendation scenarios. To address these issues, we introduce a new reasoning strategy called Visual-Summary Thought (VST) for Multimodal Recommendation. This approach begins by prompting LVLMs to generate textual summaries of item images, which serve as contextual information. These summaries are then combined with item titles to enhance the representation of sequential interactions and improve the ranking of candidates. Our experiments, conducted across four datasets using three different LVLMs: GPT4-V, LLaVA-7b, and LLaVA-13b validate the effectiveness of VST.**

***Index Terms*—Large Vision-Language Models, Multimodal Recommendation, Reasoning Strategy**

I. Introduction

To address the cold-start issues that recommender systems (RSs) lack sufficient records of new items/users, multimodal recommender systems (MMRSs) [1]–[4] are proposed by involving the complementary content of items from multiple perspectives, e.g., textual description and visual illustration, thus enriching the recommender system's knowledge. Traditional MMRSs usually first extract features from various modalities and then use different fusion strategies to combine those features into a unified representation. Although these methods have shown promising results, they encounter challenges in efficiently fusing multimodal knowledge, especially when new modalities are introduced. Ineffective integration and representation learning can further degrade the RS's performance [5]–[8]. Additionally, the product image provided by the seller contains critical marketing highlights that attract buyers, e.g., the game's duration and thematic ambiance, elements that traditional embedding-based MMRSs may struggle to capture effectively.

Meanwhile, the remarkable success of large vision-language models (LVLMs) [9]–[16] offers encouraging solutions to the above issues encountered by traditional MMRSs. LVLMs are proficient in comprehending both textual and visual information about an item since they are trained using vast datasets [17]–[20]. Their ability to distill and adapt item features across modalities into natural language space provides more flexibility to extract information from each modality, thus exhibiting an opportunity for effective knowledge fusion. It is worth mentioning that we choose to explore the inference schema instead of finetuning for two main reasons: (1) Current LVLMs are extremely large, making deployment and finetuning impractical. (2) As LVLMs continuously evolve, extensive finetuning on such large models is computationally expensive and unnecessary. Therefore, we propose a lightweight and plug-and-play reasoning strategy compatible with all LVLM backbones, allowing performance improvements to scale with the ongoing development of LVLMs. Despite the aforementioned strengths, the incorporation of pretrained LVLMs into MMRSs remains an under-explored area. Two possible obstacles may hinder the widespread adoption of LVLMs in MMRSs:

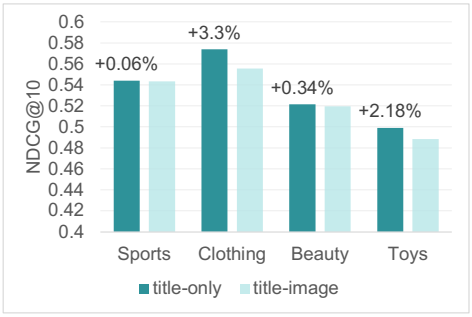

Fig. 1: Performance of GPT4-V on four representative Amazon datasets with title-only and title-image concatenation inputs.

First, *LVLMs are trained from vast general knowledge* and, as such, lack domain-specific knowledge for understanding user preferences revealed through their interactions. This gap results in the under-exploration of LVLMs' capacity in recommendation scenarios. To bridge this gap, it is essential to integrate additional knowledge to inform LVLMs in the context necessary for making appropriate recommendations. This approach, however, introduces the second challenge: *LVLMs'*

*Both authors contributed equally to this research.

inefficiency in processing multiple images. Although models like GPT4-V have been evaluated in video understanding scenarios to examine their capacity in capturing dynamic content across frames [9], [10], [21]–[25], the scenario with MMRSs involves handling multiple, discrete, and noisy images. This complexity can pose a significant challenge even from a human perspective, making it difficult to extract meaningful knowledge from such diverse interactions. Our preliminary experiments on different datasets with various LVLMs indicate that a simple concatenation of multiple images with item titles performs worse than methods relying solely on item titles for recommendations. Figure 1 shows this issue across four representative datasets with powerful GPT4-V. Note that this phenomenon also happens on other datasets using different LVLMs. Furthermore, current reasoning algorithms, e.g., in-context learning (ICL) [26], [27] and chain-of-thought (CoT) [28], [29], are primarily designed for NLP tasks ignoring visual modality. However, the principal challenge in multimodal recommendation is how to effectively leverage image-based knowledge and integrate it into the recommendation process. Thus, effective LVLM-based MMRS requires the design of specific prompting strategies that can utilize their visual comprehension strength without caving to the complexities associated with processing multiple images simultaneously.

Accordingly, we propose a novel **V**isual-**S**ummary **T**hought (VST) reasoning strategy of LVLMs for MMRSs. Our approach includes two primary components: First, we utilize user historical interactions as contextual data for the LVLMs' personalized recommendations. This involves using sequences of both item titles and images as inputs to the LVLMs. Second, to overcome the shortage of handling multiple images, we prompt the LVLMs with one static image to obtain a corresponding textual summary. Then, we construct user history sequences by substituting the images with their textual comprehensions one by one, serving as an intermediate representation for LVLMs during the reasoning phase. This strategy allows for the recommendation based on a more manageable comprehension of user preferences, transitioning from the complex and noisy image sequences to a simpler task of understanding visual-summary enhanced preference dynamics. To validate the efficacy of our proposed reasoning algorithm, we conduct experiments using GPT4-V, LLaVA-7b, and LLaVA-13b as reasoning backbones. We observe consistent improvements over other existing reasoning strategies, such as concatenation, ICL, and CoT. Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first attempt to investigate the reasoning strategies for LVLMs in multimodal recommendation scenarios. This new paradigm embraces the ongoing development and potential of LVLMs, offering a more integrated and effective approach to multimodal recommendation.
- We introduce a novel Visual-Summary Thought (VST) reasoning strategy, specifically designed for the multimodal recommendation context, to harness the proficiency of LVLMs' visual understanding and remedy their deficiency in handling multiple images simultaneously.
- We conduct comprehensive experiments to evaluate VST, utilizing both API-based LVLMs like GPT4-V, and open-source models such as LLaVA-7b and -13b. The consistent improvements observed across these models demonstrate the effectiveness of VST for LVLM-based MMRSs.

II. Methodology

A. Problem definition

In this paper, we follow the problem settings in [30], [31] that use the pretrained LVLMs as reranker to make recommendations to user u via reranking the given n candidate item titles $v = \{v_1, v_2, \ldots, v_n\}$. For each user, we have their historical interactions, which is the sequence of title and image pair of items: $u = \{(t_1, i_1), (t_2, i_2), \ldots, (t_m, i_m)\}$.

B. Preliminary

LVLMs exhibit limitations in handling multiple images. We evaluated the LVLMs' ability to handle multimodal inputs by concatenating the item titles and images of user histories. Surprisingly, leveraging complementary visual information led to poorer results than only using item titles as shown in Figure 1. (An example can be found in section III-D.) This underscores a critical insight: adding more information to the LVLMs' prompt context without a thoughtful design can lead to confusion, especially with discrete and noisy images full of redundancy. To address this challenge, we introduce a novel visual-summary thought of prompting strategy (VST) as shown in Figure 2.

C. Visual-Summary Generation

Existing LVLMs, e.g., GPT4-V and LLaVA, primarily focus on static image understanding scenarios, where LVLMs generate textual descriptions of a given image. However, this paradigm is inefficient for handling multiple images [10]. Existing strategies include concatenating images for LVLM reasoning [10], or adapting LVLMs to video comprehension scenarios via finetuning on video datasets [11]–[13], [22]. Yet, neither approach is suitable for the unique demands of MMRSs, where the image sequence of a user history is discrete and noisy, lacking the continuous nature of video frames and making sequential correlations difficult to discern. To deal with these issues, we propose leveraging LVLMs' strengths in temporal understanding within natural language modality and their capacity for static image interpretation. Instead of processing a sequence of images, we focus on distilling critical marketing highlights from individual image. The prompt can be formalized as: $s_i = summary(i)$ ="*What's in this image?*" For each item, we use one image and get the summarization of each image independently. In this way, we can not only obtain marketing highlights of items via distilling image comprehension from LVLMs but also simplify the temporal user preference understanding from the visual modality to the textual modality, where LVLMs exhibit proficiency.

D. Visual-Summary Thought for MMRSs

After summarizing each item image, we concat the history item titles with their visual summary to construct the prompt for querying user preferences among candidates. The prompt

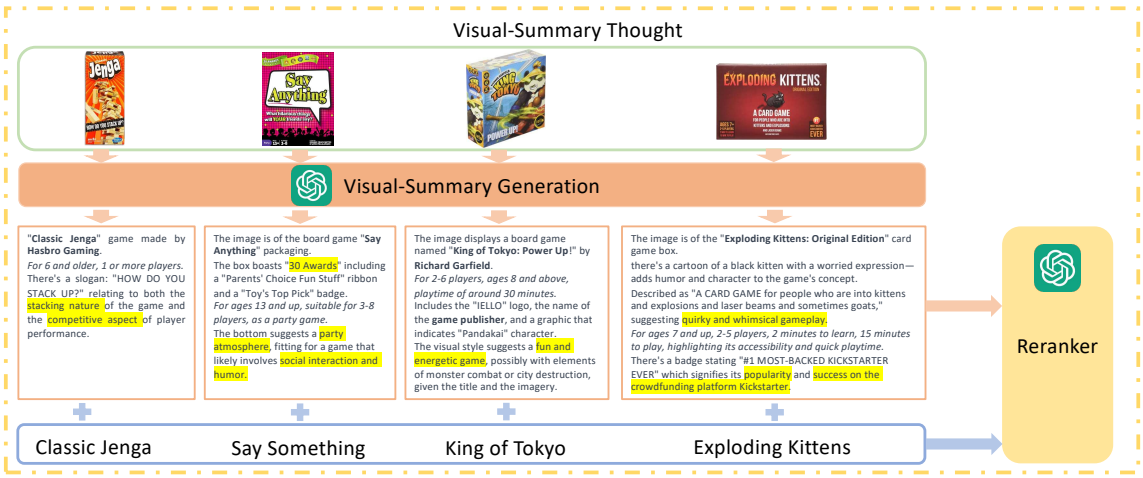

Fig. 2: Framework of Visual-Summary Thought of LVLMs for Multimodal Recommendation. Text in yellow highlights some key features obtained through visual-summary generation.

TABLE I: Statistics of the datasets after sampling.

Datasets	#Users	#Items	#Interactions	Sparsity
Sports	200	1750	2333	99.33%
Clothing	200	1291	1362	99.47%
Beauty	200	2024	2797	99.31%
Toys	200	1684	1967	99.42%

is structured in two parts: the first outlines the user's purchase history in chronological order, demonstrated by each item's title and visual summary. The second segment directs the LVLMs to rerank the candidates represented by their titles. An illustrative prompt might be:

"[Here is a chronological list of my purchase history for some products including the title and the description of each product. $\{(t_1, s_{i_1}), \ldots, (t_m, s_{i_m})\}$*][There are* $|n|$ *candidate products I am considering to buy:* $\{v_1, \ldots, v_n\}$*. Please rank these* $|n|$ *candidate products based on the likelihood I would like to purchase next most according to the given purchase history. You cannot generate products that are not in the given candidate list.]".*

III. Experiments

A. Experimental Settings

a) Dataset: In this paper, we adopt the same dataset as in [32] that uses the Amazon Review datasets for evaluation. Due to the limitation of the inference rate, following the common practice [30], [33], we only sample 200 users for evaluation. We report the statistics of such datasets in Table I.

b) Metrics: Following the leave-one-out evaluation strategy adopted in prior works [34]–[38], we treat the last item in each historical interaction sequence as the ground-truth (target) item. We adopt Recall@K (R@K) and NDCG@K (N@K) to evaluate the ranking performance of the LVLMs over candidate items, which consist of the title of the target item and the 9 random sampled items following [30].

c) Implementation Details: For open-source LVLMs, we use Fastchat to launch models and conduct the model inference on a single GeForce RTX 4090.

d) Baseline Models: As there is no previous work that only utilizes the inference capacity of LVLMs for multimodal recommendation, we adopt the commonly chosen prompting strategies used in NLP tasks: in-context-learning (ICL) and chain-of-thought (CoT) for comparison. **MM**: The plain prompt, using the simple concatenation of the historical item titles and images as the first segment. The second part keeps the same as VST. **MM-ICL**: For ICL, we match each prefix of the user's historical interaction sequence with its corresponding successor as demonstration examples. For example: *"[Here is a chronological list of my purchase history:* $\{(t_1, i_1), \ldots, (t_{m-1}, i_{m-1})\}$*] [Then if I ask you to recommend a new product, you should recommend* t_m*. Now I've just purchased* t_m*, I want to buy a new product...]"*. The remaining part is the same as the second part of VST. **MM-CoT**: For CoT, we adopt zero-shot CoT by adding *"Please think step by step."* to the second part of the prompt, while the first part is the same as MM. For example: *"[Here is a chronological list of my purchase history:* $\{(t_1, i_1), \ldots, (t_m, i_m)\}$*][There are* $|n|$ *candidate products I am considering to buy . . . Please think step by step by considering my preferences based on the given titles and image sequence of the purchased products.. . .]".*

Note that we focus on exploring different reasoning strategies of LVLMs in MMRSs with zero-shot settings. It is out of our scope to compare with traditional full-shot methods that are trained on the target datasets.

B. Overall Performance

To demonstrate the effectiveness of our proposed VST strategy, we employ GPT4-V, LLaVA-7b, and LLaVA-13b as pretrained LVLMs and conduct experiments with four different prompt strategies across four public datasets. The complete experimental results are shown in Table II. From the table, we can observe that our proposed VST reasoning strategy achieves the best or comparable performances across all datasets, demonstrating the effectiveness of our approach. Notably, our approach has a better performance on Sports dataset than others. We observe the characteristic of this dataset

TABLE II: Performance comparison of different prompt strategies. Target items are guaranteed to be included in the candidate sets. We highlight the **best** and the second-best results.

Dataset	Metric	GPT4-V				LLaVA-7b				LLaVA-13b			
		MM	MM-ICL	MM-CoT	VST	MM	MM-ICL	MM-CoT	VST	MM	MM-ICL	MM-CoT	VST
Sports	R@5	0.6900	0.6950	0.5750	**0.7250**	0.1300	0.1900	0.1800	**0.3283**	0.2250	0.3300	0.2300	**0.3750**
	R@10	0.8600	0.8600	0.8150	**0.9000**	0.2950	0.3400	0.3250	**0.5067**	0.3200	0.4850	0.3250	**0.6250**
	R@20	0.8700	0.8650	0.8300	**0.9050**	0.3100	0.3500	0.3550	**0.5117**	0.3400	0.5000	0.3450	**0.6350**
	N@5	0.4880	0.5126	0.4186	**0.5263**	0.0703	0.1138	0.1043	**0.1769**	0.1395	0.2087	0.1393	**0.2244**
	N@10	0.5435	0.5666	0.4961	**0.5834**	0.1243	0.1619	0.1506	**0.2345**	0.1706	0.2598	0.1701	**0.3063**
	N@20	0.5461	0.5678	0.4999	**0.5846**	0.1281	0.1646	0.1580	**0.2357**	0.1755	0.2637	0.1752	**0.3086**
Clothing	R@5	0.6550	**0.7100**	0.6300	0.6950	0.1400	0.1650	0.1700	**0.2800**	0.3650	0.3200	0.2550	**0.3950**
	R@10	0.8950	0.9050	0.8150	**0.9300**	0.2750	0.3100	0.2600	**0.3250**	**0.6700**	0.5450	0.4200	0.6200
	R@20	0.9000	0.9050	0.8200	**0.9350**	0.2900	0.3150	0.2600	**0.3250**	**0.6950**	0.5450	0.4200	0.6250
	N@5	0.4781	**0.5580**	0.4631	0.5322	0.0851	0.1156	0.1086	**0.1875**	0.2248	0.2062	0.1554	**0.2594**
	N@10	0.5555	**0.6205**	0.5238	0.6085	0.1287	0.1633	0.1386	**0.2025**	0.3234	0.2787	0.2058	**0.3329**
	N@20	0.5569	**0.6205**	0.5252	0.6098	0.1326	0.1646	0.1386	**0.2025**	0.3301	0.2787	0.2085	**0.3343**
Beauty	R@5	**0.6300**	**0.6300**	0.5500	0.6200	0.2450	0.1800	0.1450	**0.2750**	0.2650	0.2900	0.2300	**0.3200**
	R@10	0.8450	0.8700	0.6400	**0.9000**	**0.4050**	0.3150	0.1700	0.4000	0.3750	0.4200	0.3200	**0.5500**
	R@20	0.8500	0.8750	0.6500	**0.9000**	**0.4200**	0.3200	0.1750	0.4000	0.3850	0.4200	0.3250	**0.5600**
	N@5	0.4503	0.4395	0.3964	**0.4536**	0.1484	0.1202	0.1006	**0.1769**	0.1641	0.1928	0.1398	**0.2183**
	N@10	0.5197	0.5183	0.4264	**0.5439**	0.1996	0.1641	0.1087	**0.2179**	0.2008	0.2361	0.1692	**0.2942**
	N@20	0.5211	0.5195	0.4290	**0.5439**	0.2035	0.1655	0.1101	**0.2179**	0.2033	0.2361	0.1706	**0.2970**
Toys	R@5	0.5500	**0.6450**	0.4950	0.6300	0.1450	0.1150	0.1300	**0.3000**	0.1875	0.3400	0.2600	**0.3617**
	R@10	0.7650	0.7800	0.6950	**0.8000**	0.2750	0.1450	0.1700	**0.3800**	0.2550	0.4250	0.3800	**0.5150**
	R@20	0.7750	0.7800	0.7050	**0.8000**	0.2850	0.1550	0.1850	**0.3950**	0.2663	0.4350	0.3800	**0.5200**
	N@5	0.4184	**0.4789**	0.3967	0.4399	0.0857	0.0842	0.0835	**0.2035**	0.1389	0.2373	0.1832	**0.2412**
	N@10	0.4883	**0.5227**	0.4349	0.4958	0.1281	0.0941	0.0977	**0.2299**	0.1614	0.2648	0.2228	**0.2919**
	N@20	0.4911	**0.5227**	0.4376	0.4958	0.1305	0.0966	0.1015	**0.2336**	0.1642	0.2672	0.2228	**0.2932**

is that the titles contain much more noise, making the alignment between textual and visual information more challenging for the employed LVLMs. In contrast, through visual-summary generation, VST can better leverage visual modality and capture more relevant information from the image, reducing the impact of the noise from different modalities to some extent. Another observation is that the more powerful the LVLM backbone becomes, i.e., with the evolution from LLaVA-7b to LLaVA-13b to GPT4-V, the better VST performs. This supports the benefit of designing such a lightweight reasoning-only strategy for MMRSs. Moreover, VST outperforms other reasoning strategies regardless of the choice of LVLMs, which supports its effectiveness tailored to MMRS scenarios.

C. Ablation Study

To analyze the effectiveness of the VST reasoning principle, we conduct an ablation study on six variants of the proposed strategy. The results on Toys dataset using LLaVA-13b are shown in Figure 3. The reported results are the average of a minimum of three repeated runs, aimed at minimizing the impact of randomness. **titleSum-VST** refers to the prompt that also lets LVLMs distill information from the title of an item: $s_t = summary(t)$ ="*What information can you get from the title?*", then appended by the summary distilled from the corresponding image. **title-based VST** refers to instructing LVLMs to distill information from an image by taking item title into consideration, where $s_i = summary(i)$ ="*This is an image related to t. Please provide a detailed description of the given image.*"

From the results, we have the following observations: (1) VST can capture more meaningful information from both textual and visual modalities. The results show that VST has the capability to significantly enhance the ranking performance compared to non-VST-based strategies. The improvement stems from VST's proficiency in multimodal understanding and serves better in sequential scenarios, where information from different sources needs to be integrated effectively. (2) Information from the title can boost performance, but it depends on the quality of the title and the alignment between the title and the image. Compared to the results among VST, title-VST, titleSum-VST, and title-based VST, we can observe that adding the title information doesn't yield improvement. This lack of improvement is likely due to the visibility of toy titles in images or the easy identification of entities mentioned in titles from the images themselves. Therefore, combining title information with VST does not provide substantial additional benefits. Whether to include titles during reasoning remains a hyperparameter decision dependent on the quality of titles in each dataset.

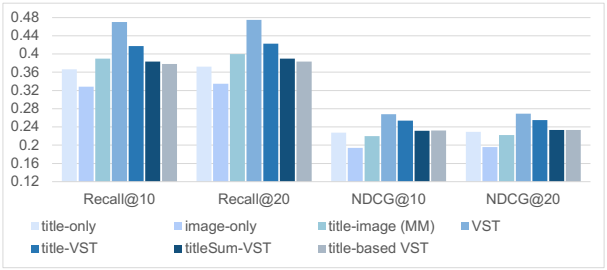

Fig. 3: Ablation study. Performance of LLaVA-13b with different prompts on Toys dataset.

D. Case Study

In this section, we compare the ranking lists generated by LLaVA-13b using VST with title-only and title-image concatenation prompts. The results are shown in Figure 4. Here are our observations from comparing the outputs: Both title-only and VST strategies successfully rank the target item as the first position, while the naive concatenation of title and image places it fourth. This discrepancy suggests that raw images may contain an excess of information, which

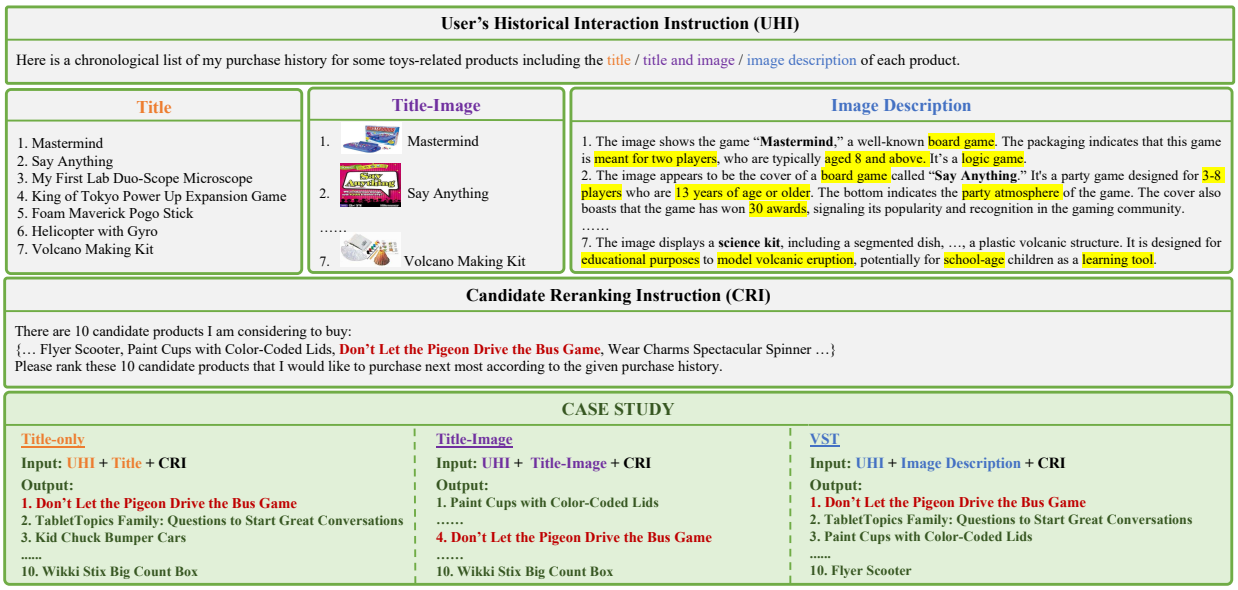

Fig. 4: Case study. Text in red indicates the target item. Text in orange, purple, or blue indicates the pattern to describe the item for the corresponding prompt. Text in yellow highlights some key features obtained through visual-summary generation.

could be perceived as redundant and introduce additional noise into our ranking task. On the other hand, the VST strategy offers a more refined approach. By utilizing VST, we not only incorporate information from the title but also extract richer and more relevant details from the image itself. Such details also align closely with the marketing selling points of the product. Consequently, the VST strategy emerges as a more effective prompt for multimodal recommendation, as it combines textual and visual cues to provide a comprehensive understanding of the item, thereby enhancing the performance of the ranking.

IV. Conclusion

In this work, we investigate the performance of different reasoning strategies for LVLMs in multimodal recommendation scenarios and identify a notable limitation in LVLMs' capability to handle multiple images effectively. To bridge this gap, we propose the Visual-Summary Thought (VST) strategy, which leverages LVLMs' visual understanding to distill information from individual images. Extensive experiments conducted on four real-world datasets using three LVLMs demonstrate the effectiveness of VST. However, our approach has some limitations: it does not integrate the strengths of traditional recommender systems and may have high time complexity, though pre-computing can mitigate this. In the future, we will explore opportunities to combine the strength of both full-shot traditional MMRSs and the inference strategies of LVLMs. Additionally, we will assess the generalization of VST across domains with more complex visual information, such as artworks. Furthermore, future work could refine VST to better capture nuanced sequential correlations among user behaviors.

V. Acknowledgement

This work is supported by the National Science Foundation Grants III-2106758, POSE-2346158, CRII-2246067, ATD-2427915, and Lehigh Grant FRGS00011497.

References

[1] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, and F. Jiang, "Bootstrap latent representations for multi-modal recommendation," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 845–854.

[2] R. He and J. McAuley, "Vbpr: visual bayesian personalized ranking from implicit feedback," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[3] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining latent structures for multimedia recommendation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3872–3880.

[4] K. Liu, F. Xue, D. Guo, L. Wu, S. Li, and R. Hong, "Megcf: Multimodal entity graph collaborative filtering for personalized recommendation," *ACM Trans. Recomm. Syst.*, vol. 41, no. 2, pp. 1–27, 2023.

[5] F. Liu, H. Chen, Z. Cheng, A. Liu, L. Nie, and M. Kankanhalli, "Disentangled multimodal representation learning for recommendation," *IEEE Transactions on Multimedia*, vol. 25, p. 7149–7159, 2023.

[6] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, and L. Wang, "Latent structure mining with contrastive modality fusion for multimedia recommendation," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[7] H. Zhou, X. Zhou, Z. Zeng, L. Zhang, and Z. Shen, "A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions," *arXiv preprint arXiv:2302.04473*, 2023.

[8] H. Liu, Y. Wei, F. Liu, W. Wang, L. Nie, and T.-S. Chua, "Dynamic multimodal fusion via meta-learning towards micro-video recommendation," *ACM Trans. Inf. Syst.*, vol. 42, no. 2, pp. 1–26, 2023.

[9] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, vol. 9, no. 1, p. 1, 2023.

[10] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li, T. MA, Y. Li, L. XU, D. Shang, Z. Zhu, S. Sun, Y. BAI, X. Cai, M. Dou, S. Hu, B. Shi, and Y. Qiao, "On the road with GPT-4v(ision): Explorations of utilizing visual-language model as autonomous driving agent," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[11] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, pp. 543–553.

[12] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-ChatGPT: Towards detailed video understanding via large vision and language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 12 585–12 602.

[13] Z. Wang, L. Wang, Z. Zhao, M. Wu, C. Lyu, H. Li, D. Cai, L. Zhou, S. Shi, and Z. Tu, "Gpt4video: A unified multimodal large language model

for lnstruction-followed understanding and safety-aware generation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, p. 3907–3916.

[14] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[15] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[16] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[17] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, H. Ren *et al.*, "A generalist vision–language foundation model for diverse biomedical tasks," *Nature Medicine*, pp. 1–13, 2024.

[18] Z. Yuan, Z. Li, W. Huang, Y. Ye, and L. Sun, "Tinygpt-v: Efficient multimodal large language model via small backbones," *arXiv preprint arXiv:2312.16862*, 2023.

[19] L. Wei, Z. Jiang, W. Huang, and L. Sun, "Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4," *arXiv preprint arXiv:2308.12067*, 2023.

[20] Z. Yuan, Y. He, K. Wang, Y. Ye, and L. Sun, "Artgpt-4: Towards artistic-understanding large vision-language models with enhanced adapter," *arXiv preprint arXiv:2305.07490*, 2023.

[21] Y. Cao, X. Xu, C. Sun, X. Huang, and W. Shen, "Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead," *arXiv preprint arXiv:2311.02782*, 2023.

[22] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu *et al.*, "Video understanding with large language models: A survey," *arXiv preprint arXiv:2312.17432*, 2023.

[23] R. Wadhawan, H. Bansal, K.-W. Chang, and N. Peng, "Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models," *arXiv preprint arXiv:2401.13311*, 2024.

[24] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, K. Ikamura, G. Gerber, I. Liang, L. P. Le, T. Ding, A. V. Parwani *et al.*, "A foundational multimodal vision language ai assistant for human pathology," *arXiv preprint arXiv:2312.07814*, 2023.

[25] Z. Yan, K. Zhang, R. Zhou, L. He, X. Li, and L. Sun, "Multimodal chatgpt for medical applications: an experimental study of gpt-4v," *arXiv preprint arXiv:2310.19061*, 2023.

[26] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[27] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" in *EMNLP*, 2022, pp. 11 048–11 064.

[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Proceedings of the 36th International Conference on Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[29] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations*, 2023.

[30] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao, "Large language models are zero-shot rankers for recommender systems," in *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024*. Springer, 2024, pp. 364–381.

[31] Y. Wang, Z. Liu, J. Zhang, W. Yao, S. Heinecke, and P. S. Yu, "Drdt: Dynamic reflection with divergent thinking for llm-based sequential recommendation," *arXiv preprint arXiv:2312.11336*, 2023.

[32] S. Geng, J. Tan, S. Liu, Z. Fu, and Y. Zhang, "Vip5: Towards multimodal foundation models for recommendation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 9606–9620.

[33] J. Zhang, Y. Hou, R. Xie, W. Sun, J. McAuley, W. X. Zhao, L. Lin, and J.-R. Wen, "Agentcf: Collaborative learning with autonomous language agents for recommender systems," in *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 3679–3689.

[34] Z. Liu, Z. Fan, Y. Wang, and P. S. Yu, "Augmenting sequential recommendation with pseudo-prior items via reversely pre-training transformer," in *Proceedings of the 44th international ACM SIGIR conference on Research and development in information retrieval*, 2021, pp. 1608–1612.

[35] Y. Wang, H. Zhang, Z. Liu, L. Yang, and P. S. Yu, "Contrastvae: Contrastive variational autoencoder for sequential recommendation," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2056–2066.

[36] Y. Wang, Z. Wang, H. Zhang, Q. Yin, X. Tang, Y. Wang, D. Zhang, L. Cui, M. Cheng, B. Yin *et al.*, "Exploiting intent evolution in e-commercial query recommendation," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 5162–5173.

[37] Y. Wang, Z. Liu, L. Yang, and P. S. Yu, "Conditional denoising diffusion for sequential recommendation," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2024, pp. 156–169.

[38] X. Li, Y. Liu, Z. Liu, and S. Y. Philip, "Time-aware hyperbolic graph attention network for session-based recommendation," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 626–635.