August 3–7, 2025 KDD2025

Adapting Large Vision-Language Models to Visually-Aware Conversational Recommendation

Hyunsik Jeon [1], Satoshi Koide [2], Yu Wang [1], Zhankui He [3], Julian McAuley [1]

[1] UC San Diego [2] Toyota Research [3] Google DeepMind

TOYOTA RESEARCH INSTITUTE

Code & Datasets

Introduction

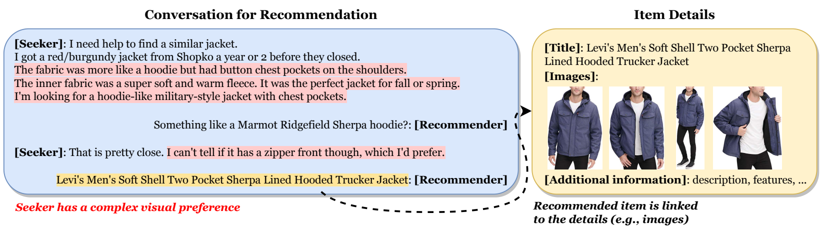
- **Conversational Recommender Systems (CRS)** deliver personalized items through interactive, multi-turn dialogue
- Real user requests frequently include visual requirements (e.g., "*I need a backpack with red straps*"), which text-only CRS cannot resolve reliably.
- We define **Visually-Aware Conversational Recommendation (VACR)**: given the dialogue history and a catalog of candidate items, each with a title and images, select the single item that best satisfies the user's current request.

Key Challenges

- Adapting a large vision-language model to VACR surfaces two practical hurdles:
  - ➢ ***Data scarcity for visual dialogues*** - Natural conversations that explicitly reference item images are still rare
  - ➢ **Context-length pressure** - With a large VLM, each image expands to thousands of tokens; evaluating many candidate items in one shot can overflow the predefined token window (e.g., 4k for LLaVA v1.6)

Reddit-Amazon Dataset

- We open-source **Reddit-Amazon dataset** for VACR benchmark

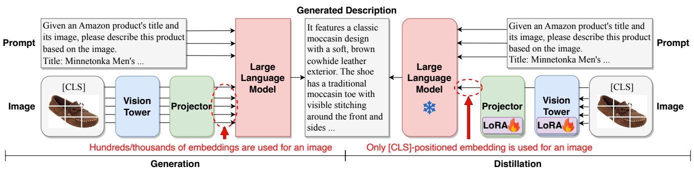


Proposed Framework: LaViC

- We propose **LaViC (Large Vision-Language Conversational Recommendation Framework)**

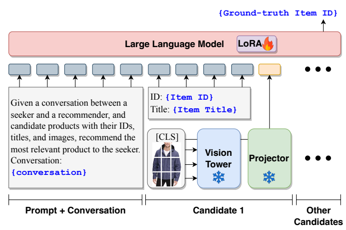
➢ **Visual knowledge self-distillation**

✓ Compress thousands of tokens of each image into 5 [CLS] tokens using self-distillation



➢ **Recommendation fine-tuning**

✓ Feed {ID, Title, $[\text{CLS}]_{1\ldots5}$} along with the dialogue to a Large Vision-Language Model (e.g., LLaVA) and update via LoRA.



Key Results

- **Datasets (Reddit-Amazon)**

Dataset	# Conv.	# Turns	# Items	# Images
Beauty	7,672	22,966	5,433	28,082
Fashion	8,039	21,831	6,716	31,162
Home	3,701	6,675	3,077	18,505

- **Evaluation**
  - ➢ Hit ratio (HR@1) for accuracy
  - ➢ Validation (VR) to detect hallucinations

- **Comparison w/ open-source methods**

7B: Vicuna-v1.5 through LaViC (ours)

Method	*Beauty* HR@1	*Beauty* VR	*Fashion* HR@1	*Fashion* VR	*Home* HR@1	*Home* VR
Retrieval Baselines (item title)						
BM25	0.0169	-	0.0140	-	0.0479	-
SBERT	0.0551	-	0.0681	-	0.2166	-
$\text{RoBERTa}_{\text{large}}$	0.0640	-	0.0631	-	0.1814	-
$\text{SimCSE}_{\text{large}}$	0.0326	-	0.0301	-	0.0957	-
$\text{BLaIR}_{\text{base}}$	0.0371	-	0.0441	-	0.1335	-
Generative Baselines (item title) + SBERT						
Vicuna-v1.5	0.0533	0.9870	0.0481	0.9903	0.1184	1.0000
LLaVA-v1.5	0.0476	0.9896	0.0441	0.9855	0.0932	1.0000
LLaVA-v1.6	0.0770	0.9870	0.0827	0.9867	0.2030	0.9919
Generative Baselines (item title and image) + SBERT						
LLaVA-v1.5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LLaVA-v1.6	0.0584	0.9741	0.0459	0.9843	0.1089	0.9919
Proposed Method (item title and image) + SBERT						
LaViC (ours)	**0.1187**	0.9702	**0.1232**	0.9298	**0.3197**	0.9892
Improvement	**+54.2%**	-	**+49.0%**	-	**+47.6%**	-

- **Comparison w/ proprietary methods**

150-200B: GPT-3.5-turbo through GPT-4o (item title and image); 7B: LaViC (ours)

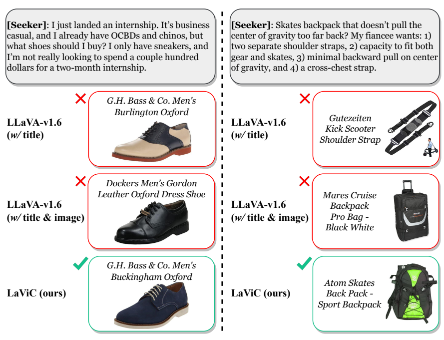
Method	*Beauty* HR@1	*Beauty* VR	*Fashion* HR@1	*Fashion* VR	*Home* HR@1	*Home* VR
Generative Baselines (item title) + SBERT						
GPT-3.5-turbo	0.0968	0.9935	0.0977	0.9903	0.2343	1.0000
GPT-4o-mini	0.1213	1.0000	0.1160	0.9927	0.3258	0.9973
GPT-4o	0.1271	0.9987	0.1278	0.9976	0.3350	1.0000
Generative Baselines (item title and image) + SBERT						
GPT-4o-mini	0.1081	0.9974	0.1098	0.9927	0.2861	0.9946
GPT-4o	0.1160	0.9974	0.1231	0.9939	0.3308	0.9973
Proposed Method (item title and image) + SBERT						
LaViC (ours)	0.1187	0.9702	0.1232	0.9298	0.3197	0.9892

- **Ablation study**

Method	*Beauty* HR@1	*Beauty* VR	*Fashion* HR@1	*Fashion* VR	*Home* HR@1	*Home* VR
Entire tokens (5×577)	0.0256	0.9456	*o.o.m.*		*o.o.m.*	
w/o images	0.0972	0.9767	0.1022	0.9358	0.2944	0.9946
w/o self-distillation	0.0842	0.9793	0.1084	0.9649	0.2861	0.9973
LaViC (ours)	**0.1187**	0.9702	**0.1232**	0.9298	**0.3197**	0.9892

Case Study

- (a) LaViC captures subtle visual attributes (color, design) not evident in the item title
- (b) LaViC captures additional details such as extra straps or shape using compressed image tokens



(a) Same brand but different style (b) Specific requirements

Contribution

- **LaViC** - first unified pipeline that adapts a Large Vision-Language Model to conversational recommendation
- **Token-efficient visual distillation** breaks the context-length barrier
- We release **Reddit-Amazon** visually-aware CRS benchmark to the community
- Achieves **state-of-the-art accuracy** with commodity-scale (7B) parameters