



August 3-7, 2025



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

Hyunsik Jeon¹, Satoshi Koide², Yu Wang¹, Zhankui He³, Julian McAuley¹

¹ UC San Diego, ² Toyota Research, ³ Google DeepMind



Recommender Systems are Everywhere

- They deeply influence our daily choices and experiences

App/Music



 Spotify®

Video

 YouTube

NETFLIX

 prime

Social networks

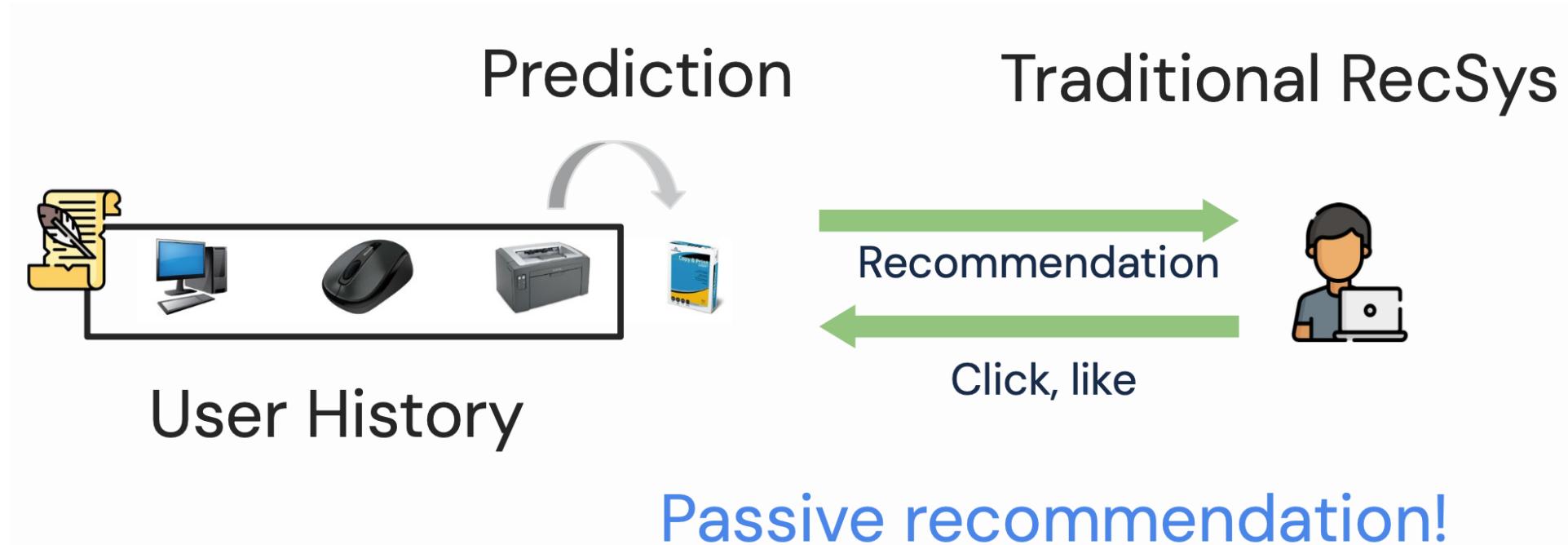
LinkedIn

Instagram

facebook

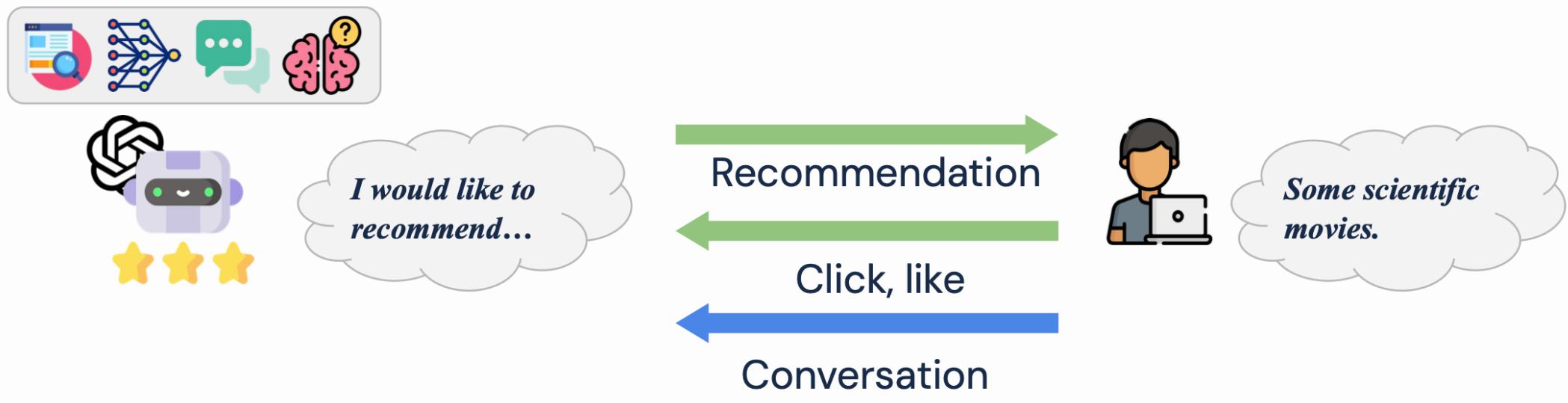
Traditional Recommendation

- Traditional RecSys leverages passive signals



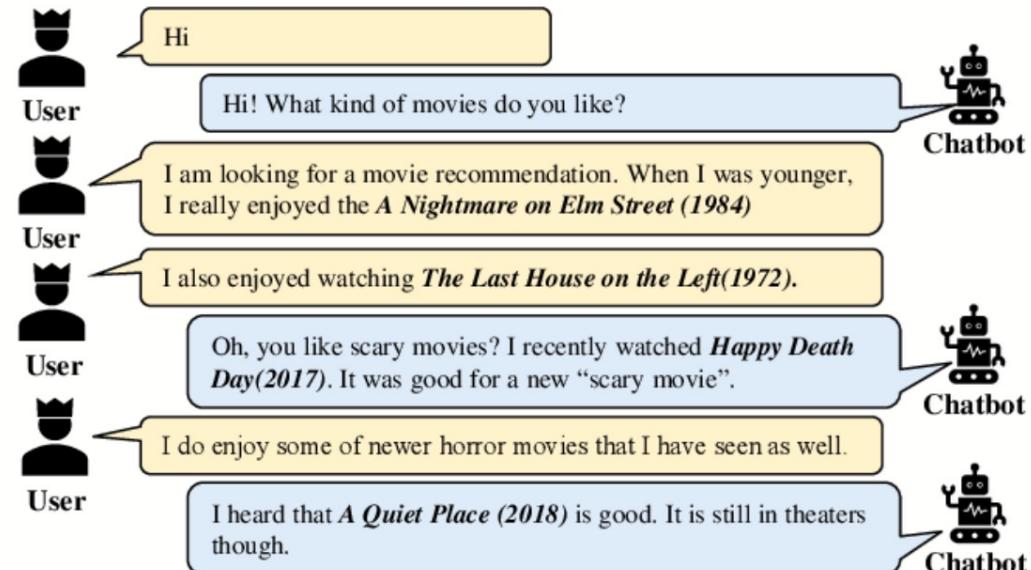
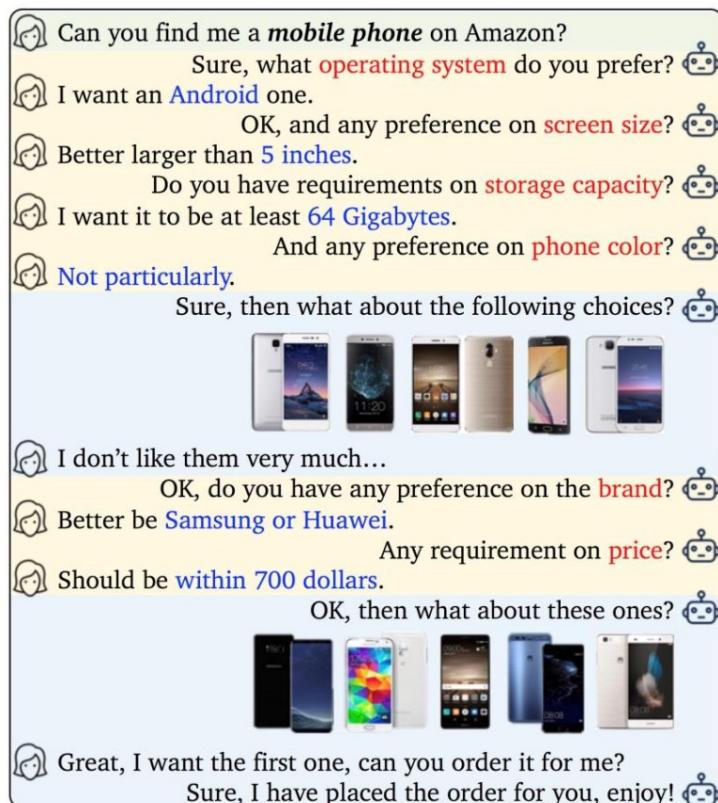
Conversational Recommendation

- **Conversational RecSys** engages users in an interactive loop



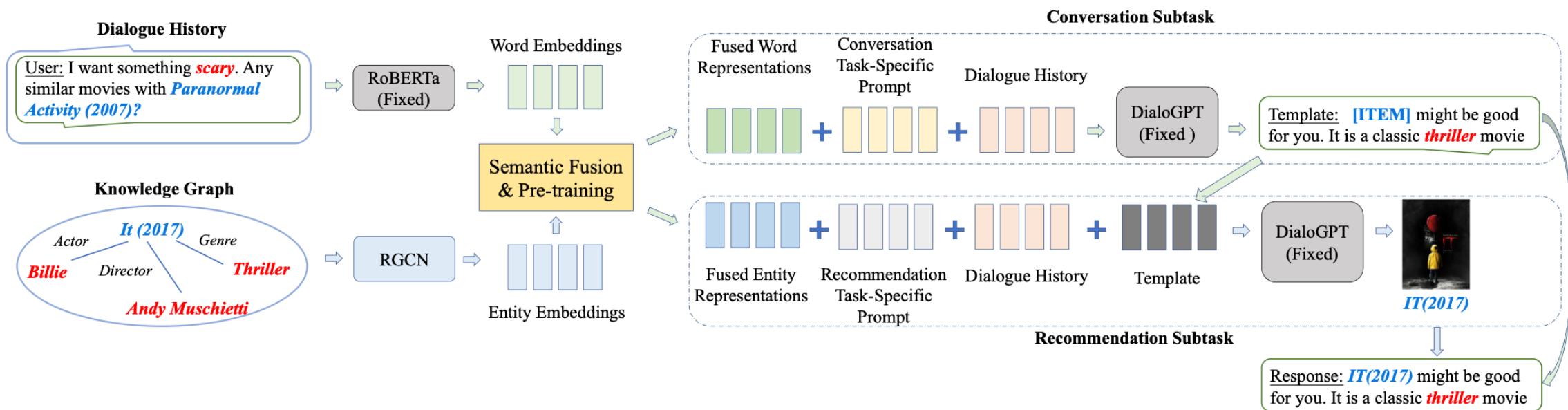
Conversation Brings Revolution

- Interactive recommendation using natural language dialogues



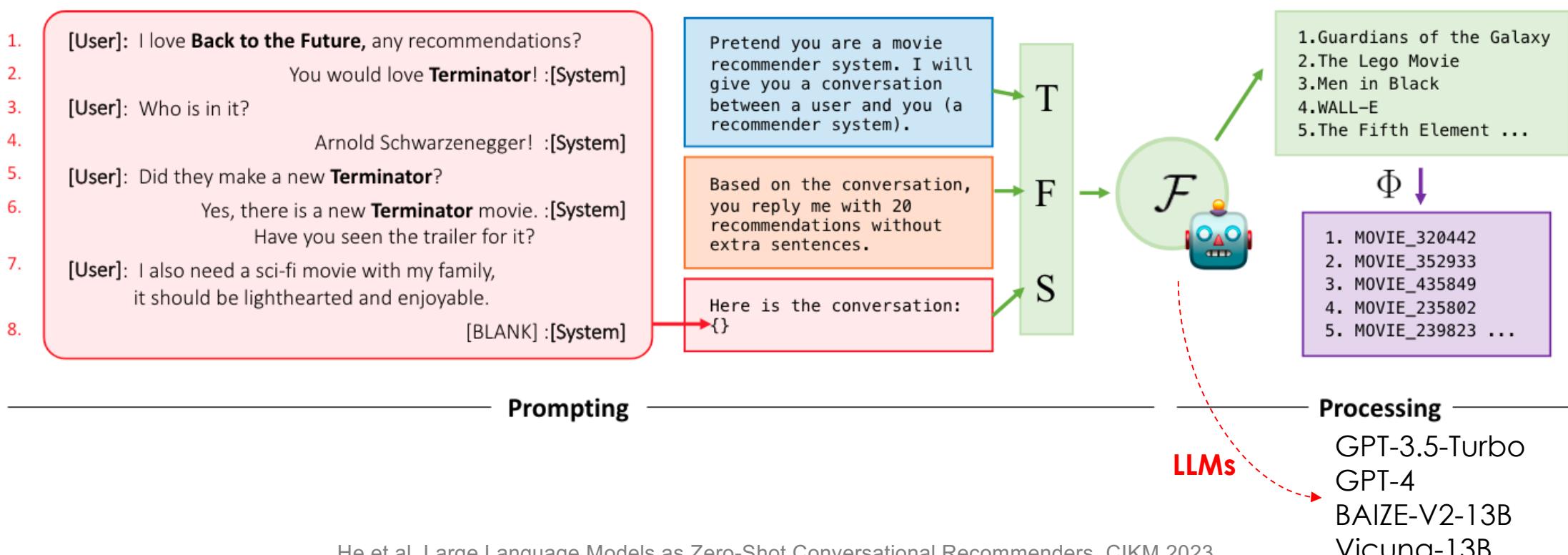
Early Works Before the Era of LLM

- Knowledge graph-based approaches (e.g., UniCRS)
 - Consists of conversation/recommendation modules while utilizing external knowledge



LLM as Conversational Recommender

- **LLM as a zero-shot recommender:** it is a single unified body of conversation, recommendation, and knowledge



Research Motivation

- **Users can request recommendations based on visual features**

[Seeker]: I need help to find a similar jacket.

I got a red/burgundy jacket from Shopko a year or 2 before they closed.

The fabric was more like a hoodie but had button chest pockets on the shoulders.

The inner fabric was a super soft and warm fleece. It was the perfect jacket for fall or spring.

I'm looking for a hoodie-like military-style jacket with chest pockets.

- However, existing CRS cannot provide relevant items unless they have a visual ability

Research Motivation

- **Visual information (image) matters**
 - Images capture crucial details (design, color), especially in visually-driven domains
 - Leveraging images can greatly enhance recommendation performance
 - However, existing CRS rely on textual features (title)

[Title]: Levi's Men's Soft Shell Two Pocket Sherpa Lined Hooded Trucker Jacket

[Images]:



An example of Amazon product

Research Motivation

How can we design a visually-aware CRS solution?

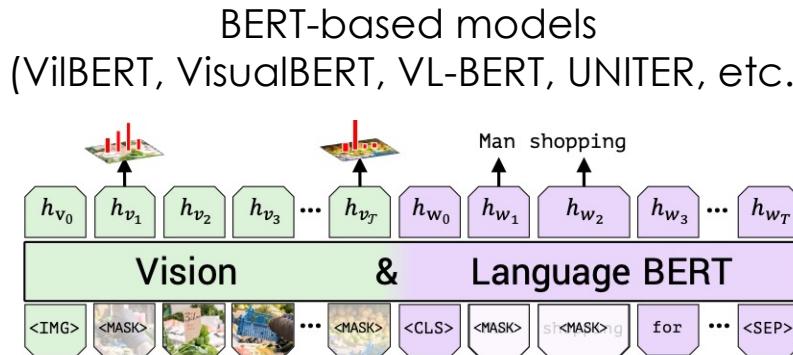
Visually-Aware Conversational Recommendation

- **Given**
 - A multi-turn dialogue with user preferences
 - A set of candidate items, each with textual + visual data (images)
- **Goal**
 - Recommend the relevant item(s) that match user's needs

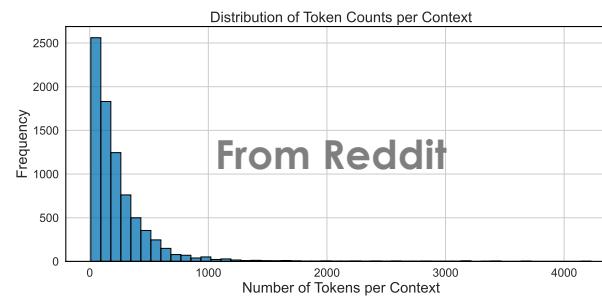
Utilize vision-language models!

Earlier Vision-Language Models

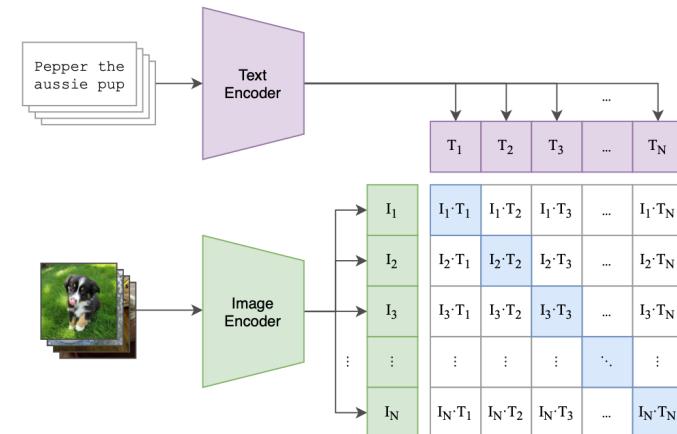
- Two types of early approaches



Mostly trained on image captioning tasks: # tokens < 100
Difficult to apply to long/complex conversations



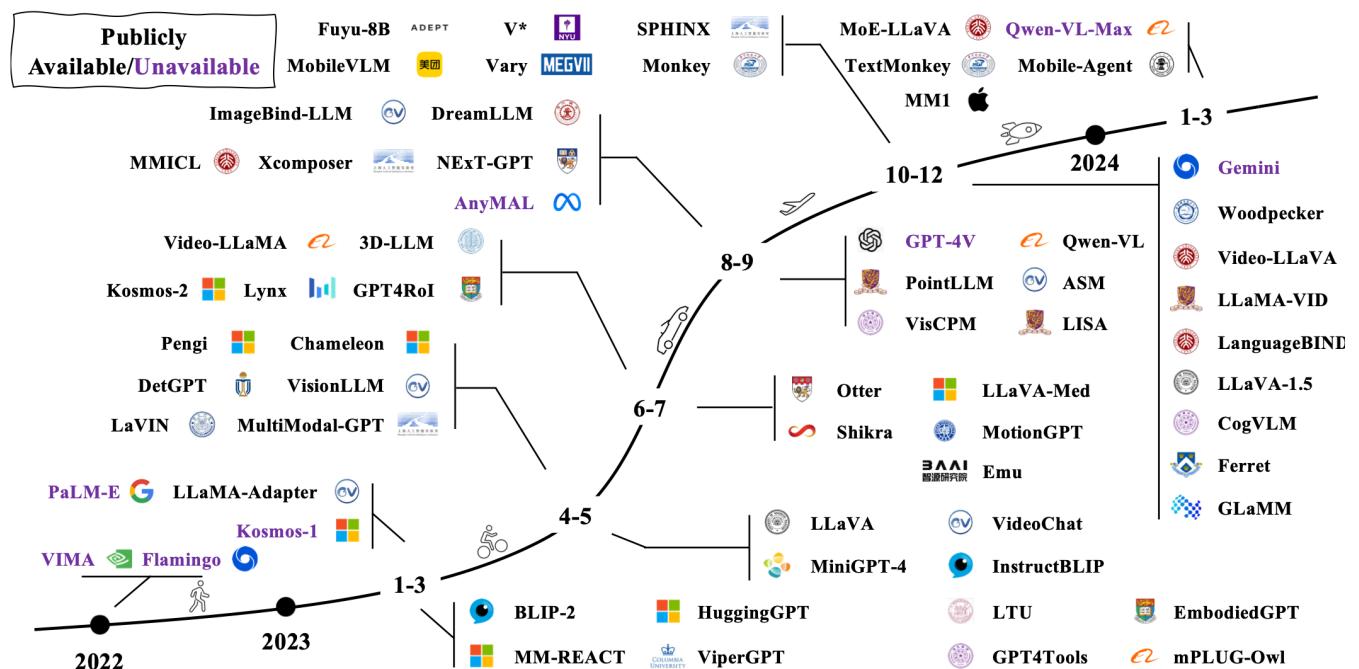
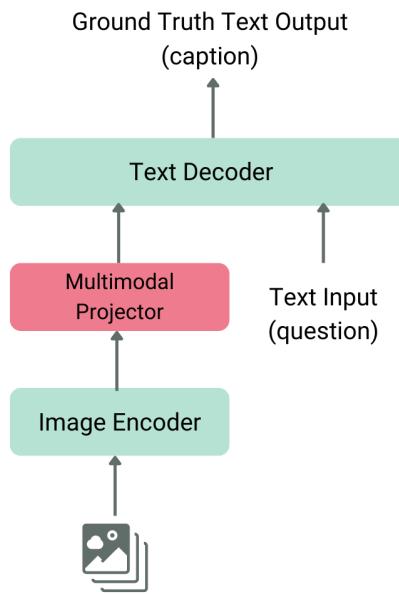
Dual-encoder contrastive models
(CLIP, ALIGN, CoCa, Florence, etc.)



Lu et al. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. NeurIPS 2019.
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

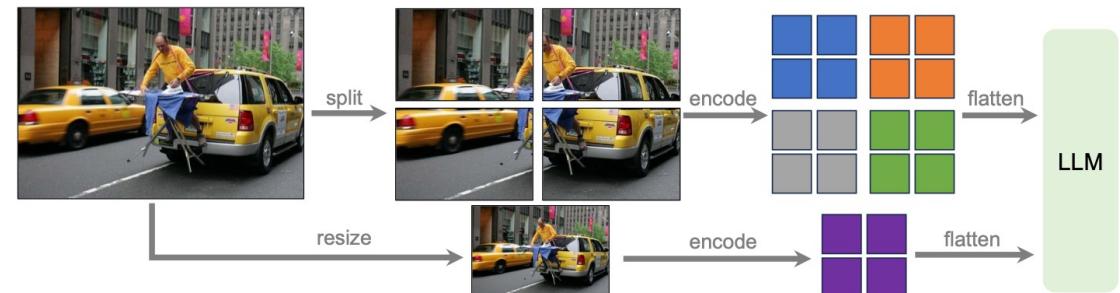
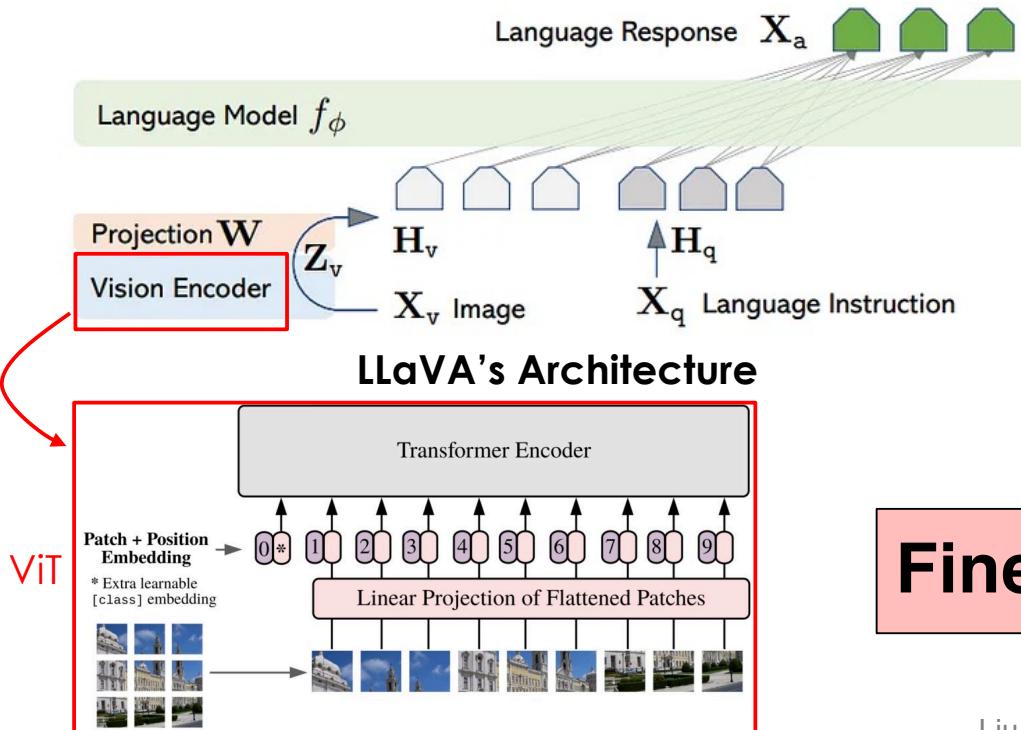
Large Vision-Language Models (VLMs)

- Leveraging LLMs for more complex downstream tasks
 - Vision encoder + LLM backbone



Large Language and Vision Assistant (LLaVA)

- One of the most widely used VLMs in various of downstream tasks
 - Open sourced, strong performance, and handleable models available (7B)



LLaVA-Next further splits each image into multiple sub-images

Fine-tune LLaVA for visually-aware CRS!

Liu et al. Visual Instruction Tuning. NeurIPS 2023.

Liu et al. Improved Baselines with Visual Instruction Tuning. CVPR 2024.

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

Challenges

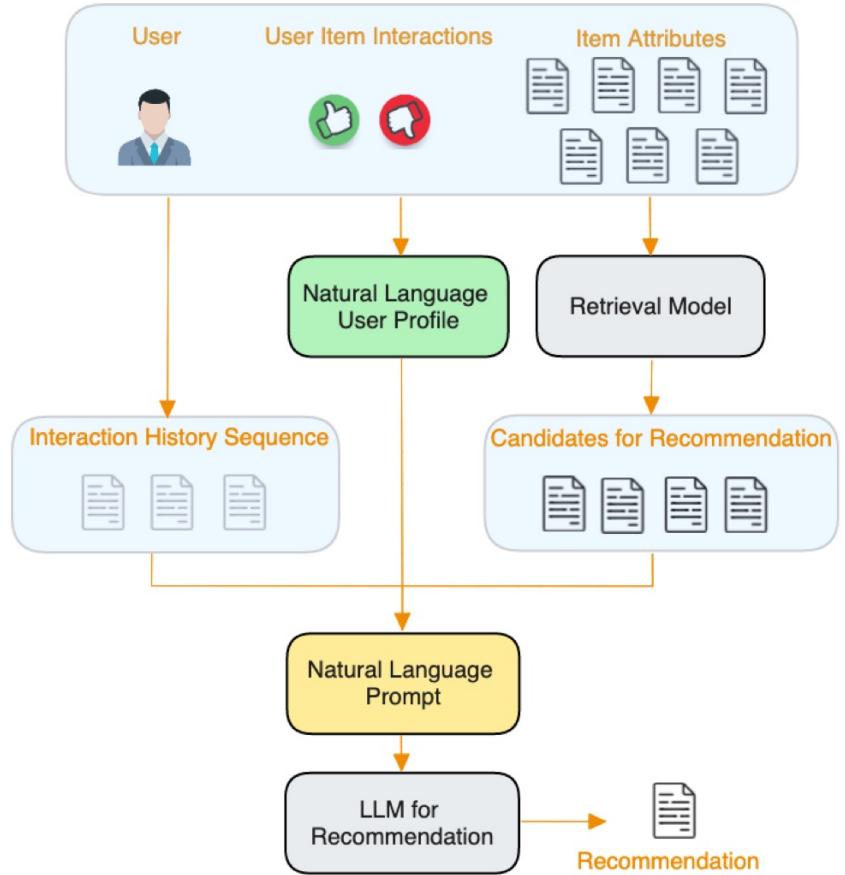
- Absence of adequate **datasets for visually-aware CRS**
 - Ideal: natural conversation ∩ image features

List of CRS datasets

Datasets	#Conv.	#Turns	#Items	Domain	Source
FacebookRec [9]	1M	6M	-	Movies	Synthetic
ReDial [31]	10K	182K	6.2K	Movies	Crowd-sourced
GoRecDial [26]	9K	170K	-	Movies	Crowd-sourced
OpenDialKG [47]	15K	91K	-	Movies, music, etc.	Crowd-sourced
TG-ReDial [70]	10K	129K	-	Movies	Synthetic
DuRecDial 2.0 [44]	16.5K	255K	-	Movies, music, etc.	Crowd-sourced
CCPE-M [49]	502	11K	-	Movies	Crowd-sourced
INSPIRED [15]	1K	35K	1.9K	Movies	Crowd-sourced
Reddit-Movie _{base} [18]	85K	133K	24.3K	Movies	Natural
Reddit-Movie _{large} [18]	634K	1.6M	51.2K	Movies	Natural
U-NEED [43]	7K	53K	-	E-commerce	Natural
E-ConvRec [24]	25K	775K	-	E-commerce	Natural
HOOPS [13]	-	11.6M	-	E-commerce	Synthetic
MGConvRec [63]	7K	73K	-	Restaurant	Crowd-sourced
MMConv [34]	5K	39K	-	Travel	Crowd-sourced
MobileConvRec [45]	12.2K	156K	1.7K	Music, sports, etc.	Synthetic

Challenges

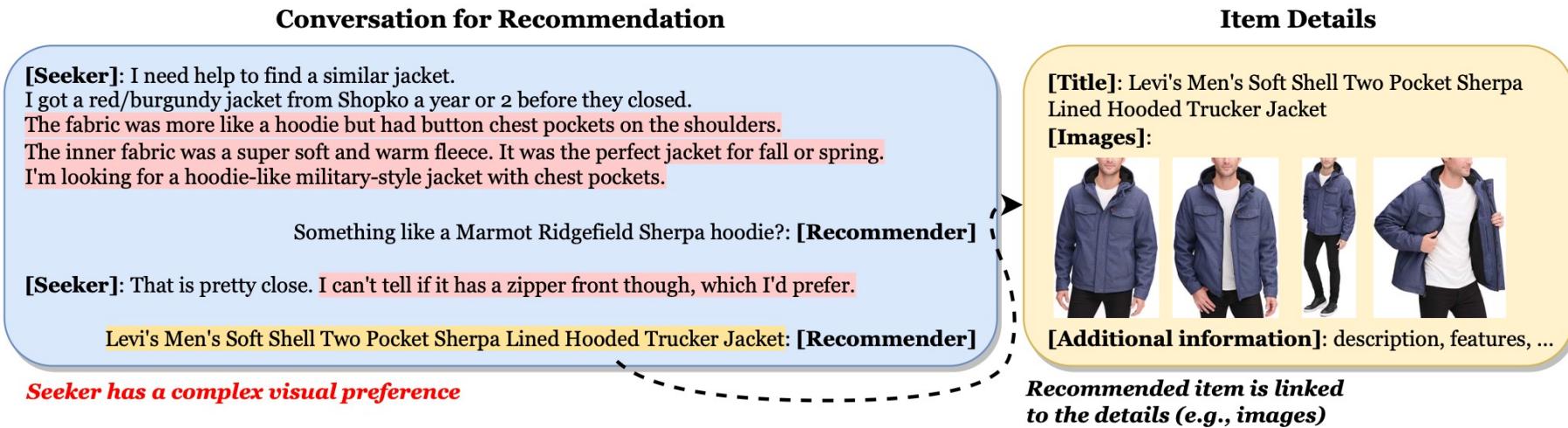
- **Candidate-based recommendation**
 - Effectively narrow down the search space
 - Suitable for domains that LLM is not familiar with
- **Token explosion for multiple images**
 - VLMs convert each image into hundreds/thousands of tokens
 - Multiple candidates cause token explosion
 - LLaVA-v1.6
 - Context length: 4K tokens, image: 2.8K tokens



Candidate-based recommendation

Reddit-Amazon Dataset

- An example



- Statistics

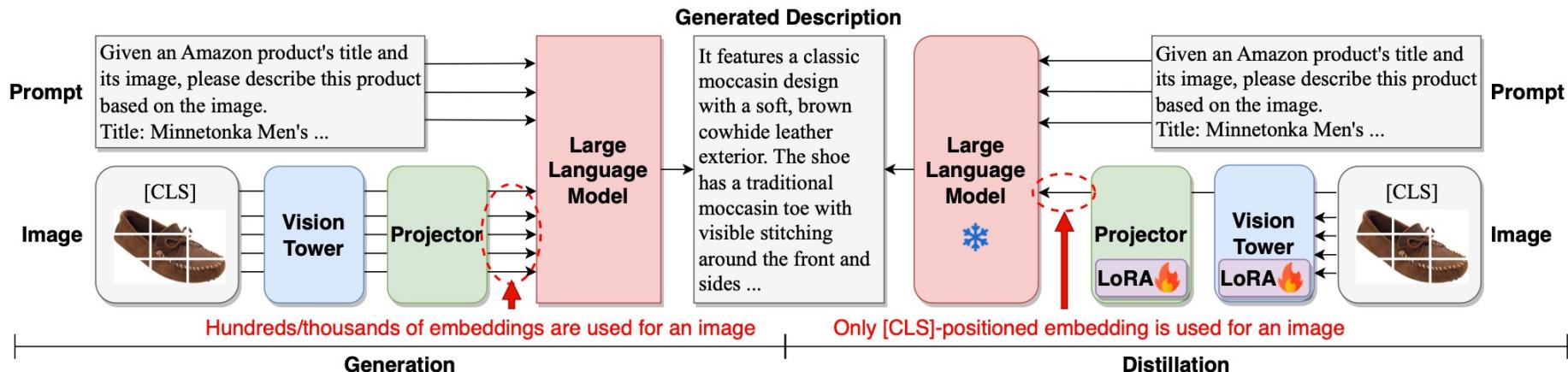
Dataset	# Conv.	# Turns	# Items	# Images
<i>Beauty</i>	7,672	22,966	5,433	28,082
<i>Fashion</i>	8,039	21,831	6,716	31,162
<i>Home</i>	3,701	6,675	3,077	18,505

Proposed Method: LaViC

- Designed **LaViC** (Large Vsion-Language Conversational Recommendation Framework)
- Overview of **LaViC**
 - 1) Visual Knowledge Self-Distillation
 - Compress thousands image tokens into minimal embeddings
 - 2) Recommendation Fine-Tuning
 - Incorporate visual embeddings + dialogue context into the LLMs

Visual Knowledge Self-Distillation

- How to avoid token explosion of using multiple images?
 - LLaVA-V1.6 splits each image into 577×5 tokens
 - Maximum context length: 4K
- Self-distillation of the visual knowledge
 - Compress each image into 5 [CLS] tokens

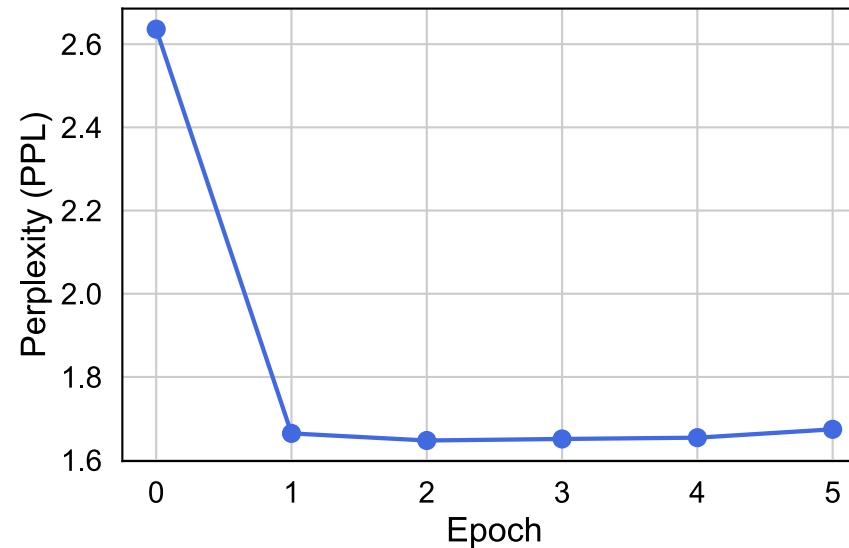


Example of Description Generation

Image	Domain	Title	Description generated by LLaVA-v1.6
	Fashion	LYANER Women's Pajamas Set 4pcs Satin Silk Cami Top Button Down Loungewear Pjs Set	<p>The product is a women's pajama set made of satin silk, consisting of four pieces. The set includes:</p> <ol style="list-style-type: none">1. A button-down long-sleeve top with a collar and black piping details.2. A cami top with thin spaghetti straps.3. A pair of long pants with an elastic waistband.4. A pair of shorts with an elastic waistband. <p>The fabric has a smooth, shiny finish and features a pink base color with a tropical print, including leaves and flowers in shades of blue, white, and black. The overall design is elegant and comfortable, suitable for lounging and sleeping.</p>
	Fashion	Super Mario Bros Classic Retro NES Design Bi-fold Wallet	<p>The product is a bi-fold wallet designed with a retro theme inspired by the classic Super Mario Bros game for the NES (Nintendo Entertainment System). The wallet features a design that mimics the appearance of an NES game cartridge, with the iconic Super Mario Bros label prominently displayed on the front. The label includes pixelated graphics reminiscent of the original game, along with the Nintendo logo and other classic design elements. The wallet appears to be made of a material that resembles the texture and color of an NES cartridge, giving it a nostalgic and vintage look.</p>

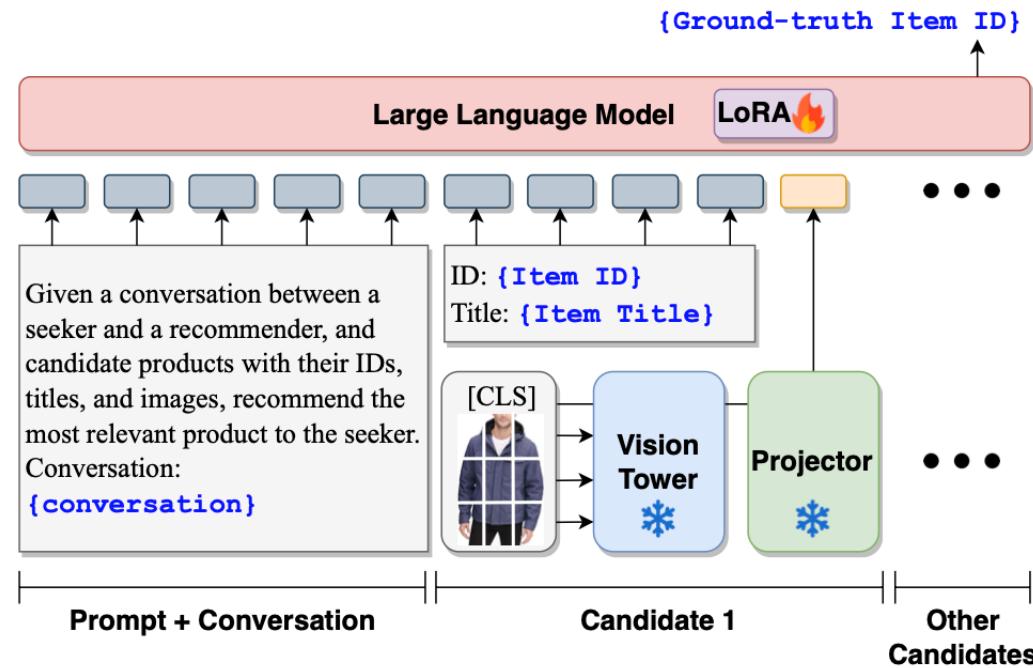
Perplexity on Visual Knowledge Self-Distillation

- The validation PPL reaches a plateau after 1-2 epochs



Recommendation Fine-Tuning

- Candidate-based pipeline
 - Retrieved top-10 items using SBERT
 - Title-dialogue matching
 - For each candidate, we provide {ID, Title, [CLS], ...} as input
 - LLM is trained to answer the ground-truth item ID
- Training only LLM parameters
 - Freeze vision tower and projector
 - LoRA is adapted for LLM



$$\min_{\Omega_{LM}} \sum_{(\mathcal{T}_{conv}, \mathcal{I}_{cand})} -\log P_{\Omega_{LM} + \Omega_{vision}} (ID_{i^*} | \mathcal{T}_{conv}, \{\mathbf{X}_{i_j}\}_{j=1}^{10})$$

Minimize negative log likelihood

Experimental Setup

- **Datasets**
 - Reddit-Amazon (Beauty, Fashion, Home)
 - Each item has title + image(s)
- **Baselines**
 - Retrieval-based: BM25, SBERT, RoBERTa, SimCSE, BLaIR
 - Generative: Vicuna, LLaVA, GPT-3.5, GPT-4o
- **Implementation**
 - Single A100 40GB GPU, LoRA, batch size=4 (distill), 1 (prompt)
- **Evaluation Metrics**
 - HitRatio@1 (HR@1): correct item selection rate
 - ValidRatio (VR): valid item ID output (no hallucination)

Experimental Results

- Comparison w/ open-source methods

Method	Beauty		Fashion		Home	
	HR@1	VR	HR@1	VR	HR@1	VR
<i>Retrieval Baselines (item title)</i>						
BM25	0.0169	-	0.0140	-	0.0479	-
SBERT	0.0551	-	0.0681	-	<u>0.2166</u>	-
RoBERTa _{large}	0.0640	-	0.0631	-	0.1814	-
SimCSE _{large}	0.0326	-	0.0301	-	0.0957	-
BLaIR _{base}	0.0371	-	0.0441	-	0.1335	-
<i>Generative Baselines (item title) + SBERT</i>						
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	<u>0.0770</u>	0.9870	<u>0.0827</u>	0.9867	0.2030	0.9919
<i>Generative Baselines (item title and image) + SBERT</i>						
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
<i>Proposed Method (item title and image) + SBERT</i>						
LaViC (ours)	0.1187	0.9702	0.1232	0.9298	0.3197	0.9892
Improvement	+54.2%	-	+49.0%	-	+47.6%	-

7B parameters

→ Outperforms the others

Experimental Results

- Comparison w/ proprietary methods

Method	Beauty		Fashion		Home	
	HR@1	VR	HR@1	VR	HR@1	VR
<i>Generative Baselines (item title) + SBERT</i>						
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
<i>Generative Baselines (item title and image) + SBERT</i>						
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
<i>Proposed Method (item title and image) + SBERT</i>						
LaViC (ours)	0.1187	0.9702	0.1232	0.9298	0.3197	0.9892

150-200B parameters {

7B parameters ← → Comparable performance

Experimental Results

- **Ablation study**
 - “Entire tokens” use all tokens for each image
 - “w/o images” use only item titles
 - “w/o self-distillation” use [CLS] tokens but without self-distillation

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

Experimental Results

• 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

[Seeker]: I just landed an internship. It's business casual, and I already have OCBDs and chinos, but what shoes should I buy? I only have sneakers, and I'm not really looking to spend a couple hundred dollars for a two-month internship.

LLaVA-v1.6
(w/ title)



LLaVA-v1.6
(w/ title & image)



LaViC (ours)



[Seeker]: Skates backpack that doesn't pull the center of gravity too far back? My fiancee wants: 1) two separate shoulder straps, 2) capacity to fit both gear and skates, 3) minimal backward pull on center of gravity, and 4) a cross-chest strap.

LLaVA-v1.6
(w/ title)



LLaVA-v1.6
(w/ title & image)



LaViC (ours)



(a) Same brand but different style

(b) Specific requirements

Conclusion

- Address the visually-aware conversational recommendation
- Propose a novel framework: LaViC
 - Large vision-language model-based framework
 - Visual knowledge self-distillation
 - Recommendation fine-tuning
- Open-source the benchmark datasets



Thank You!

hyjeon@ucsd.edu