Multimodal Bundle Recommendation and Generation

MA Yunshan

26 Aug, 2025

A Short Bio:

- 2025.1-now: Assistant Professor, School of Computing and Information Systems, Singapore Management University
- 2022.4-2024.12: Postdoctoral Research Fellow in NExT++ Research Center & NCL, NUS
- 2017.8-2022.3: PhD Candidate, School of Computing, NUS, supervised by Prof. Chua Tat-Seng
- Research Interest: Multimodal Event Forecasting, Bundle Recommendation.
- Homepage: https://mysbupt.github.io/

Outline

- What is a bundle and why do we study bundles?
- Bundle recommendation
- Multimodal bundle construction
- Bundle generation
- Open challenges

What is a bundle

Example bundles in various applications

Software

Electronics

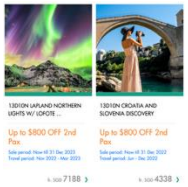
Fashion

Meal

Furniture

Cosmetics

Travel



Video Game

Online Game Shopping

Music Playlist

Some synonyms: next **basket** recommendation, **item set** recommendation

Why do we study bundles

Key characteristics of bundles

- A thematic group, not just a set of scattered items: **bundle ≠ sum of items**
- Construction with certain purposes, e.g., discount, ease of packing/shipping, suitable for situations/occasions, fulfill specific functions

Rationale of Bundling[1]

- Economies of scale: sell more products
- Economies of scope: sell various products
- Lower marginal costs: package, shipment etc.
- Lower production set-up cost
- Lower customer acquisition cost
- Ease the purchase decision making process of customers
- …

Common Bundling Strategies[2]

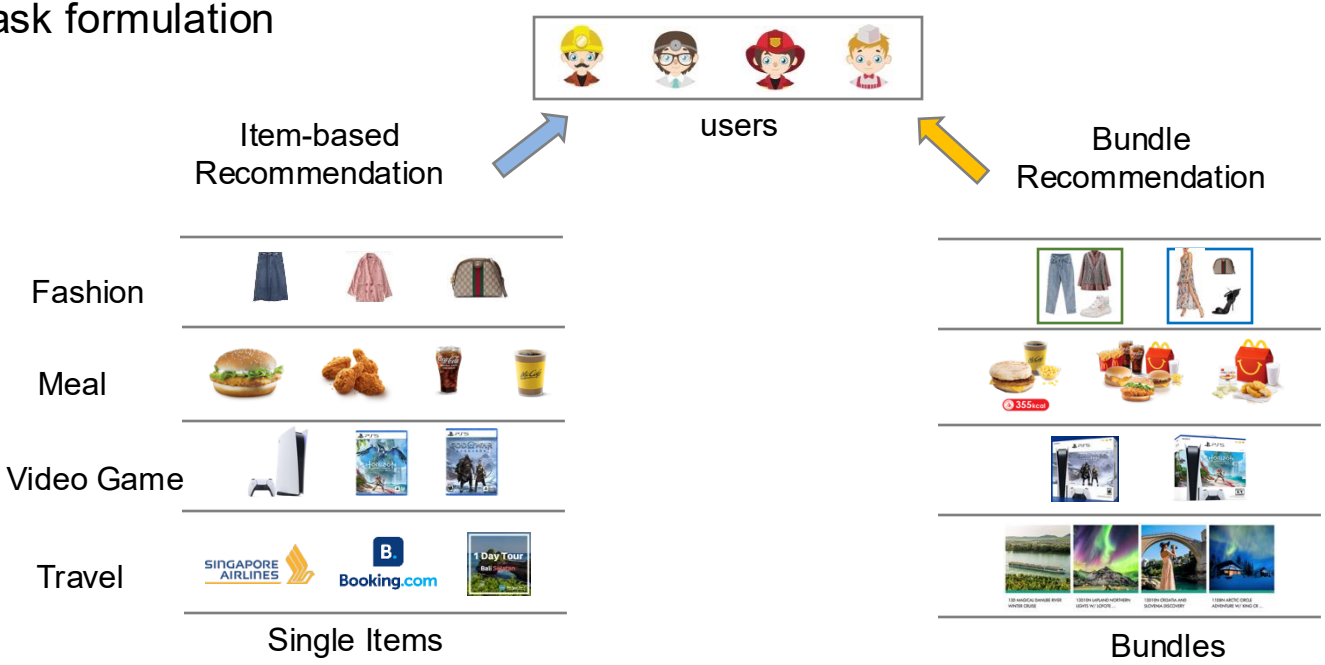
- Pure bundles
- New product bundles
- Mix-and-match bundles
- Cross-sell bundles
- Gifting bundles
- Inventory clearance bundles
- Buy-one-get-one bundles
- …

[1] https://en.wikipedia.org/wiki/Product_bundling

[2] https://www.zoho.com/inventory/guides/what-is-product-bundling.html

Personalized bundle recommendation (1/5)

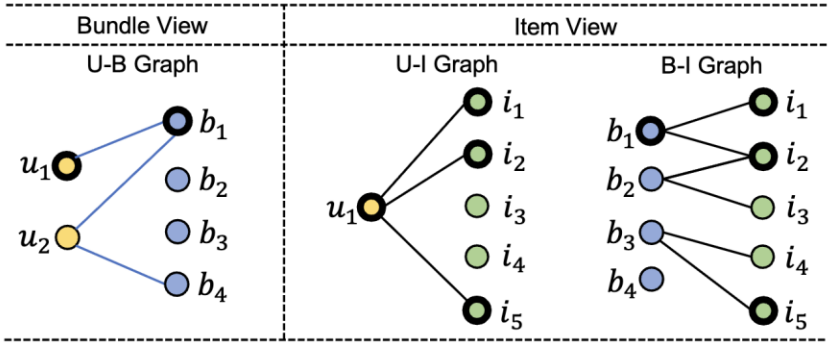
Task formulation

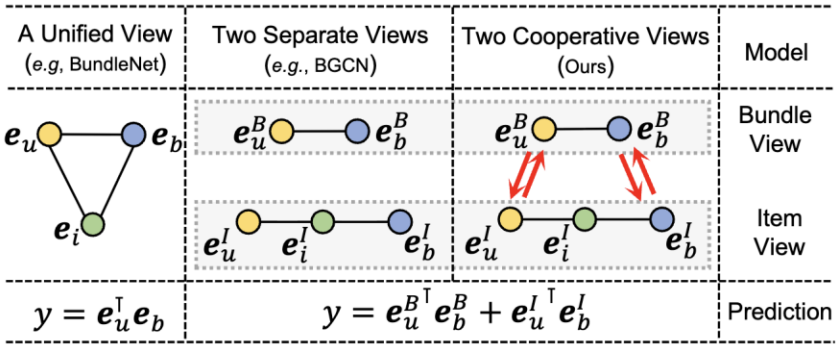


New challenges: need to model both user-item and user-bundle interactions

Personalized bundle recommendation (2/5)

CrossCBR: cross-view contrastive learning





Sources of user preferences:

- Bundle view (U-B graph)
- Item view (U-I graph & B-I graph)

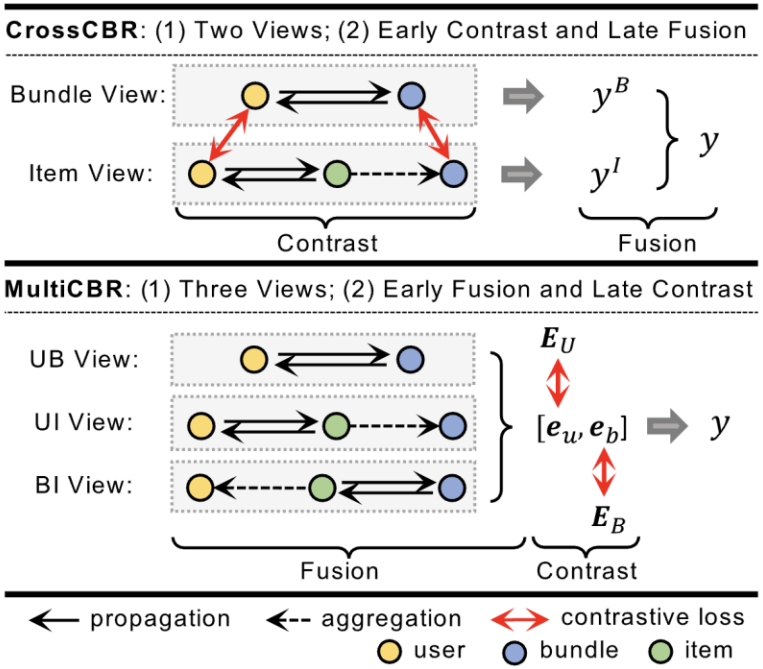
Cooperative association between these two views has been loosely modeled or even overlooked in existing works

Motivation: Modeling the cooperative association between two views is vital to the success of bundle recommendation.

CrossCBR: cross-view contrastive learning for bundle recommendation. Yunshan Ma et al. KDD 2022.

Personalized bundle recommendation (3/5)

MultiCBR: multi-view contrastive learning



Novelties:

1) Introduce the additional third view: B-I view;
2) Adopt the "early fusion and late contrast" paradigm, which can model both cross-view and ego-view user preferences, with few extra computational expenses

$$
\begin{aligned}
y^*_{u,b} &= (\mathbf{e}_u^{UB} + \mathbf{e}_u^{UI} + \mathbf{e}_u^{BI}) \cdot (\mathbf{e}_b^{UB} + \mathbf{e}_b^{UI} + \mathbf{e}_b^{BI}) \\
&= \underbrace{\left(\mathbf{e}_u^{UB} \cdot (\mathbf{e}_b^{UI} + \mathbf{e}_b^{BI}) + \mathbf{e}_u^{UI} \cdot (\mathbf{e}_b^{UB} + \mathbf{e}_b^{BI}) + \mathbf{e}_u^{BI} \cdot (\mathbf{e}_b^{UB} + \mathbf{e}_b^{UI})\right)}_{\text{cross-view preference}} \\
&\quad + \underbrace{(\mathbf{e}_u^{UB} \cdot \mathbf{e}_b^{UB} + \mathbf{e}_u^{UI} \cdot \mathbf{e}_b^{UI} + \mathbf{e}_u^{BI} \cdot \mathbf{e}_b^{BI})}_{\text{ego-view preference}},
\end{aligned}
$$

MultiCBR: multi-view contrastive learning for bundle recommendation. Yunshan Ma et al. TOIS 2024.

Personalized bundle recommendation (4/5)

Some further studies:

- EBRec: we identify that the item representation is insufficiently learned, and enhancing the **item-level representation** will significantly improve the bundle recommendation performance;
- BundleGT: we propose to use **transformer** as backbone to model the bundling strategy of bundles;

Enhancing Item-level Bundle Representation for Bundle Recommendation. Xiaoyu Du et al. TORS 2023.
Strategy-aware Bundle Recommender System. Yinwei Wei et al. SIGIR 2023.

Reflections on bundle recommendation (5/5)

Two unsolved problems:

1. Only interaction data is insufficient
   - Cold-start, long-tail items have few interactions
   - Lack of rich content and semantic information of items

2. Only recommend existing bundle
   - What if there is no pre-constructed bundles?
   - Existing bundles cannot satisfy users' needs
   - New items come but have not been incorporated

Single Modality

Multimodality

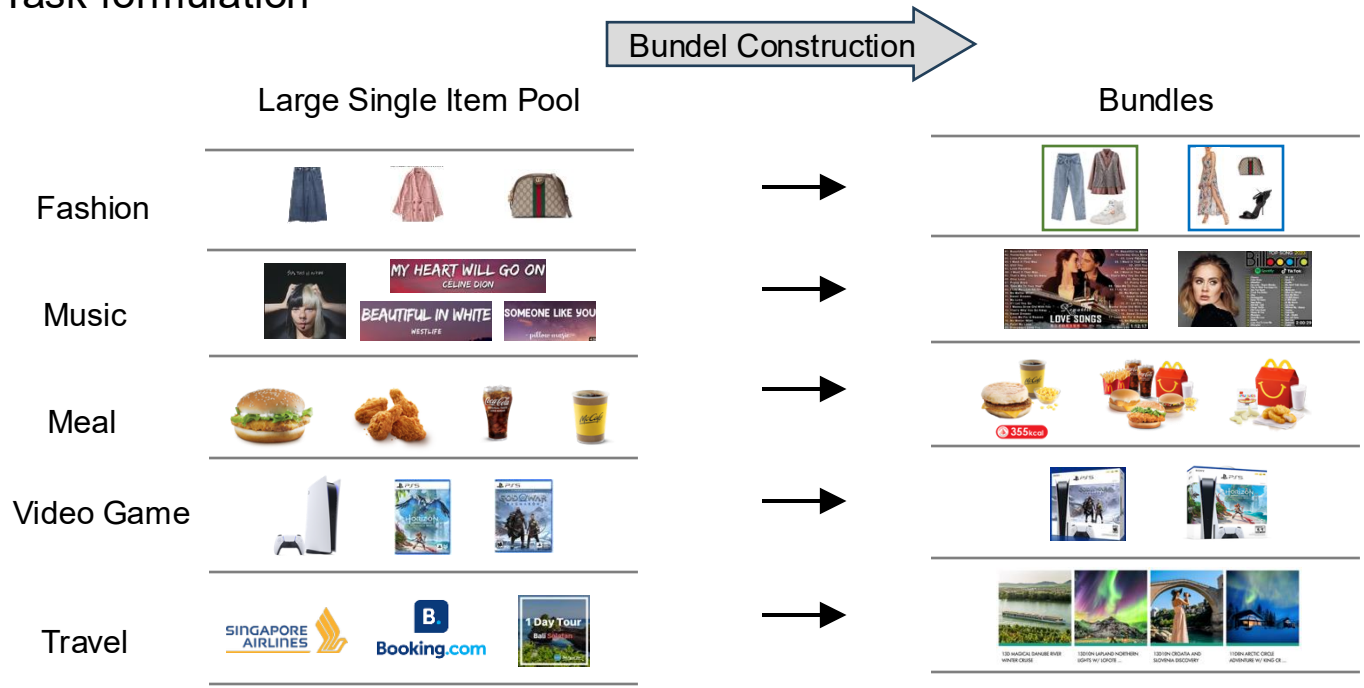
Recommendation

Construction

Multimodal Bundle Construction

Multimodal Bundle Construction (1/7)

Task formulation



Multimodal Bundle Construction (2/7)

Key characteristics

Why are the items included in the same bundle? → the main data sources and key patterns

Multimodal Data Sources

Content Data of Single Item

- Visual: Images, Videos, etc.
- Textual: Title, Lyrics, keywords, etc.
- Audio: Audio track, sounds, etc.

Relational Data Between Items

- Existing Bundles
- User-item Interactions

Items within the same bundle are **Related**

Explicitly Related: Similar on certain Semantic Aspects

- Share keywords, attributes, text spans, etc.
- Share visual objects, concepts, patterns, etc.
- Share instruments, melodies, styles, etc.

Implicitly Related: Co-exist and Co-interact

- Mix-and-Match (co-exist)
- Co-purchase (co-interact)

Multimodal Bundle Construction (3/7)

CIRP: cross-item relational pre-training for multimodal product bundling

Motivation:

Explicitly distill cross-item relations into VLMs.

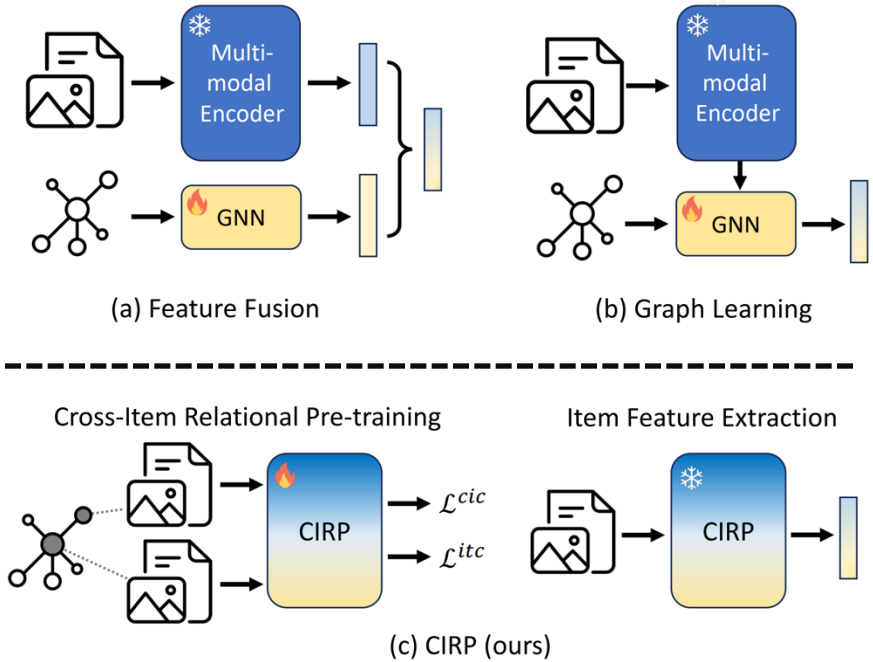
- Cross-Item Relational Pre-training (**CIRP**)

Pre-training Stage:

- Image-text contrastive (ITC) loss retains multimodal understanding.
- Cross-item contrastive (CIC) loss captures cross-item relations.

Inference Stage:

- Generate relation-aware multimodal representations, even for cold-start items.



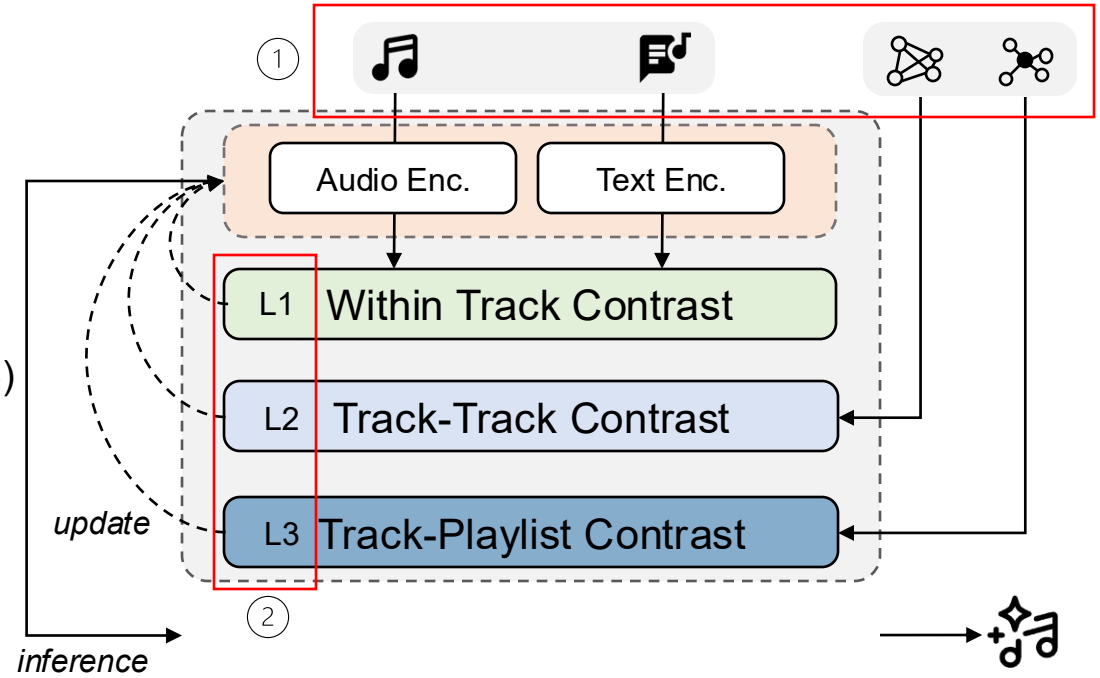
CIRP: Cross-Item Relational Pre-training for Multimodal Product Bundling. Yunshan Ma et al. ACM MM 2024.

Multimodal Bundle Construction (4/7)

LARP: language audio relational pre-training for cold-start playlist continuation

① Three ***modalities*** of data (introduce audio):
- Audio - **new**
- Language
- Relational

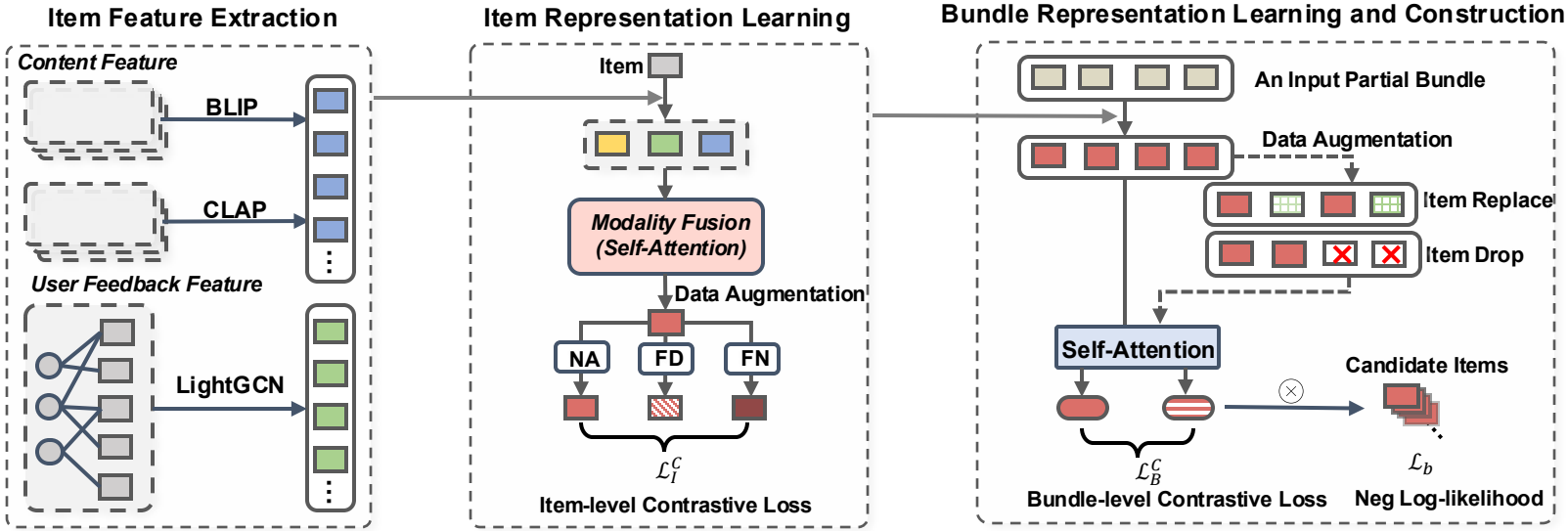
② Extend CIRP from two losses (ITC and CIC) to three ***stages*** of contrastive learning:
- Within Track (WTC) - ITC
- Track to Track (TTC) - CIC
- Track to Playlist (TPC) - **new**



LARP: Language Audio Relational Pre-training for Cold-Start Playlist Continuation. Rebecca Salganik et al. ACM MM 2024.

Multimodal Bundle Construction (5/7)

CLHE: leveraging multimodal features and item-level user feedback for bundle construction



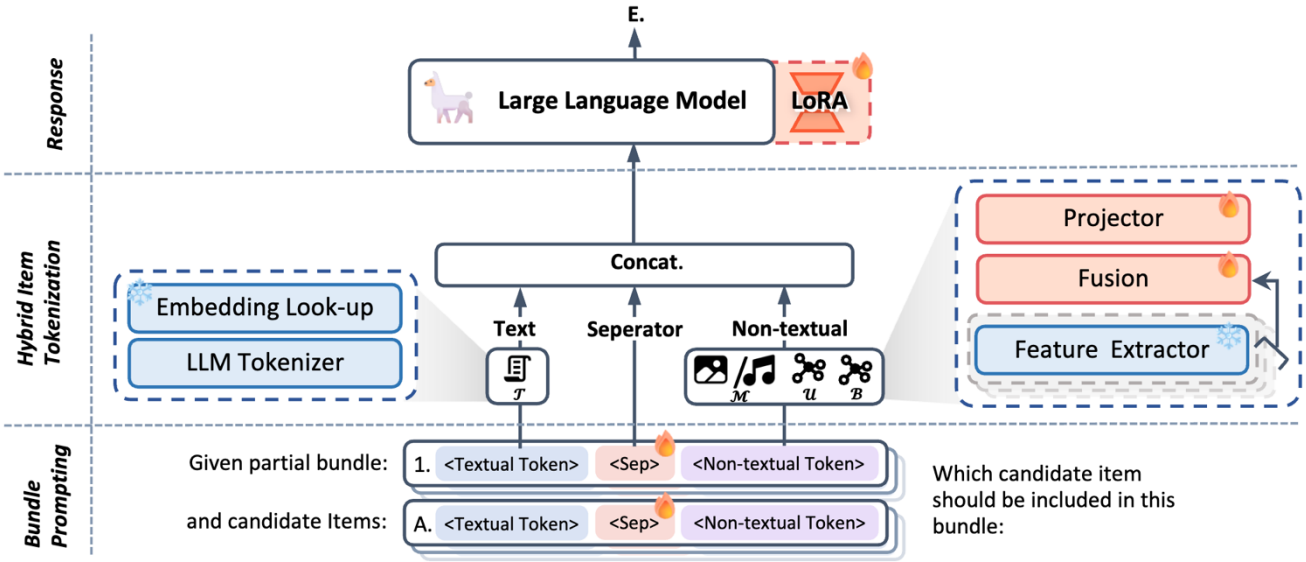
Key Novelty:

1) leverage comprehensive modalities: interactions, text, image, audio;
2) integrate multimodal representation learning into an **end-to-end** bundle construction model (previous work separate the two parts).

Leveraging Multimodal Features and Item-level User Feedback for Bundle Construction. Yunshan Ma et al. WSDM 2024.

Multimodal Bundle Construction (6/7)

Bundle-MLLM: Fine-tuning Multimodal Large Language Models for Product Bundling



Key Novelty:

1) enhanced multimodal understanding capability by using MLLMs, compared with previous CLIP-based backbones;
2) extensive internal knowledge can be utilized for the product bundling task.

Fine-tuning Multimodal Large Language Models for Product Bundling. Xiaohao Liu et al. KDD 2025.

Reflections on bundle construction (7/7)

New problems:

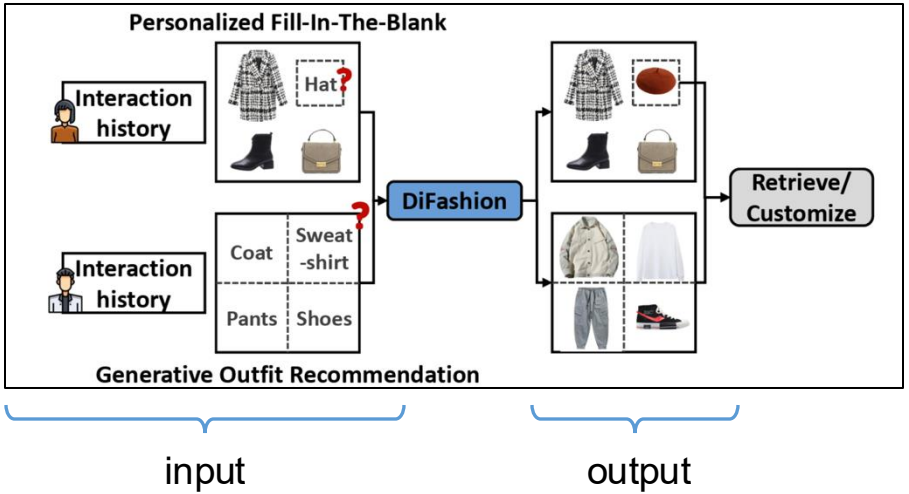
- Only pick items from existing item candidate pool
  - What if there is no available proper item for bundling?
  - What if the items are not in-stock or from other platforms?

🤔 With the substantial progress in GenAI, can we directly generate bundles?

✅ Yes! That is **Bundle Generation**

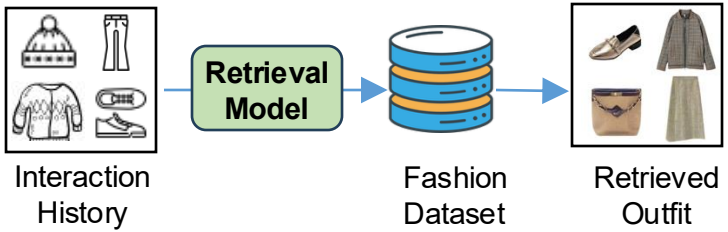
Bundle Generation (1/5)

Task formulation

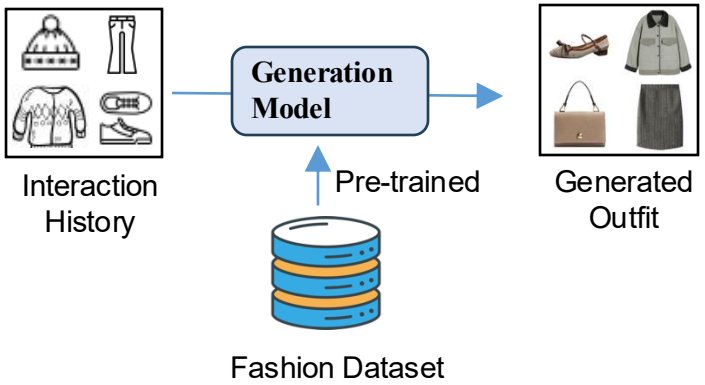


1) Personalized Fill-in-the-Blank (PFITB)
2) Generative Outfit Recommendation (GOR)

Retrieval-based methods

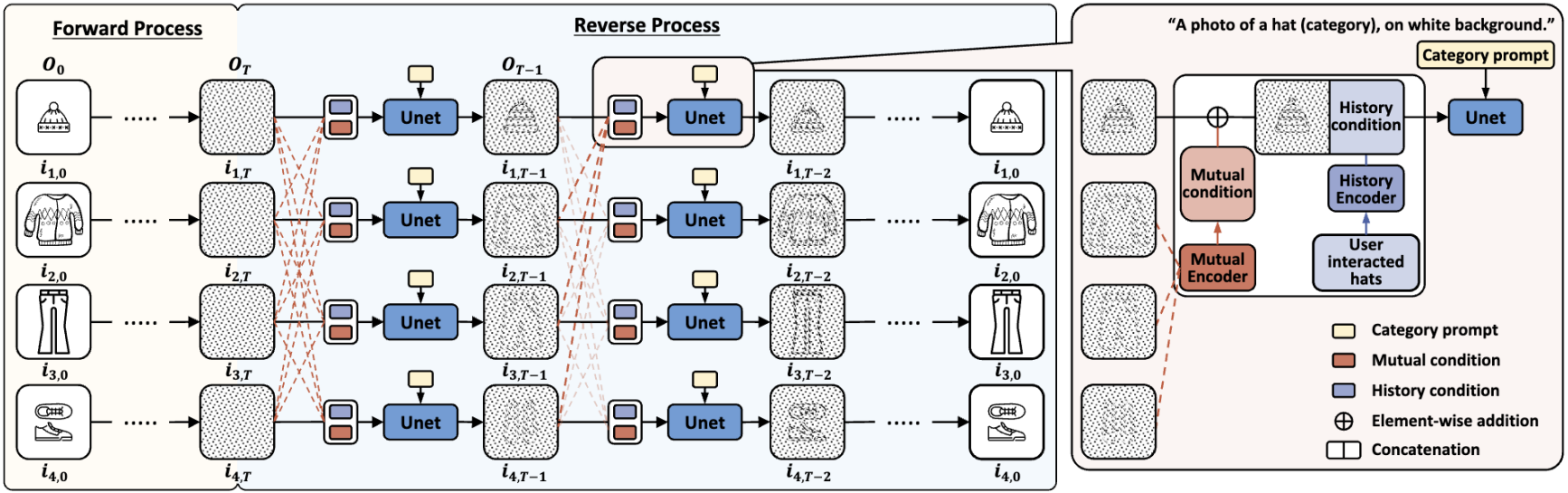


Generation-based methods



Bundle Generation (2/5)

DiFashion: Diffusion Models for Generative Outfit Recommendation

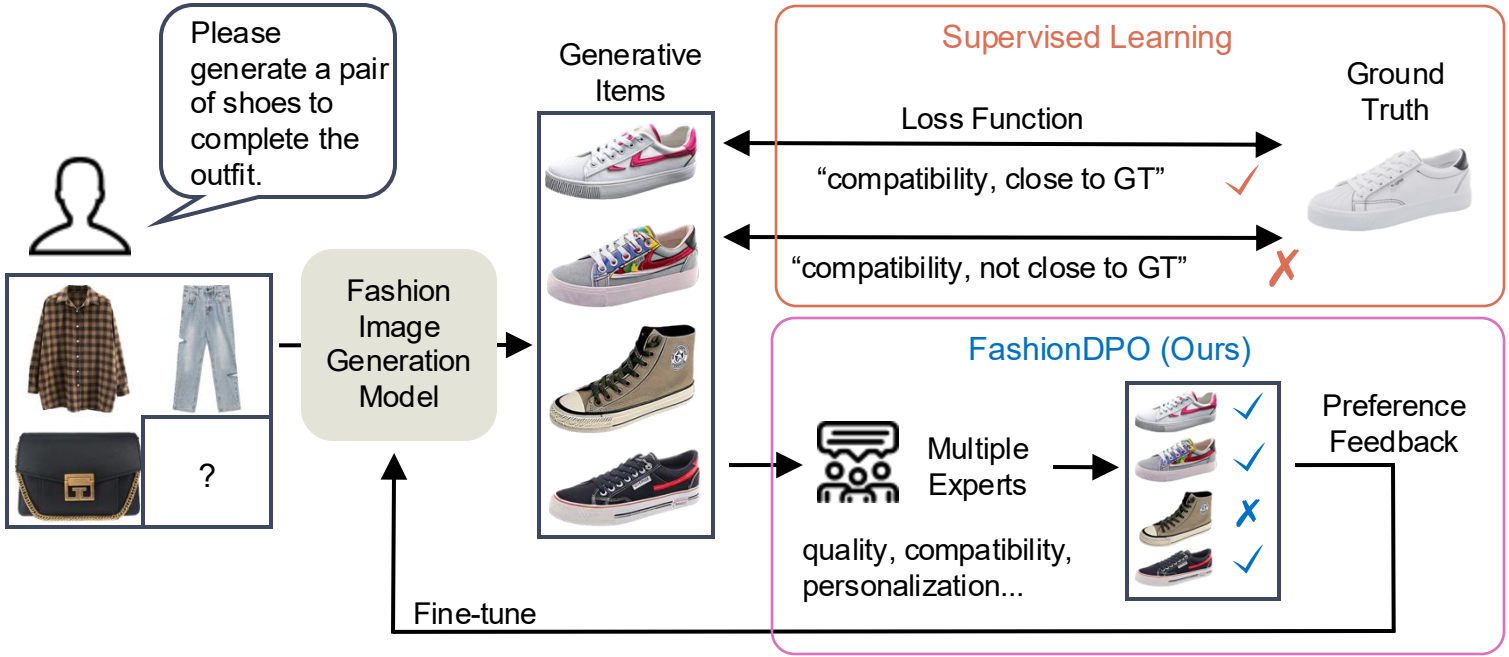


1) Use diffusion model to directly generate fashion item images
2) Condition the generation with both user's historical interaction (personalization) and other items within the bundle (compatibility)

Diffusion Models for Generative Outfit Recommendation. Yiyan Xu et al. SIGIR 2024.

Bundle Generation (3/5)

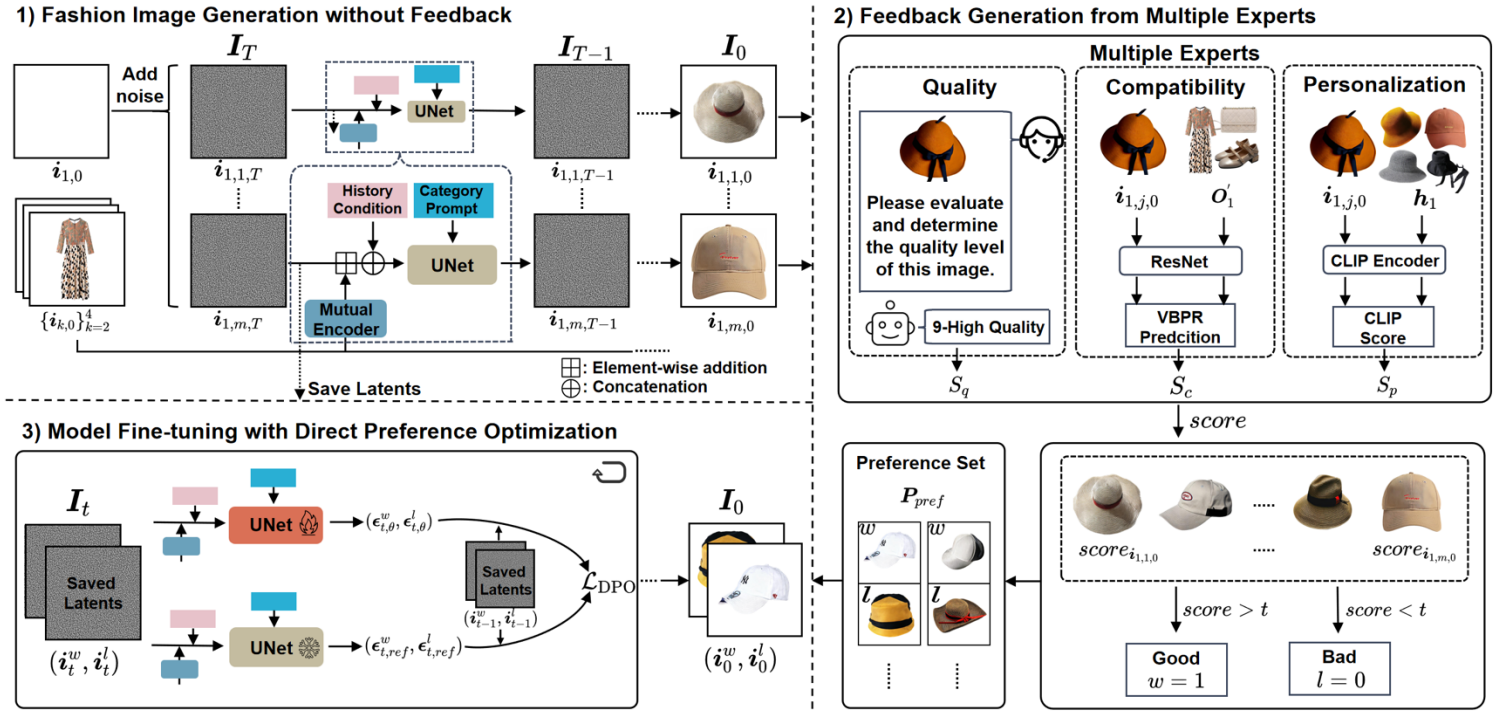
FashionDPO: fine-tune fashion outfit generation model using direct preference optimization



FashionDPO: Fine-tune Fashion Outfit Generation Model using Direct Preference Optimization. Mingzhe Yu et al. SIGIR 2025.

Bundle Generation (3/5)

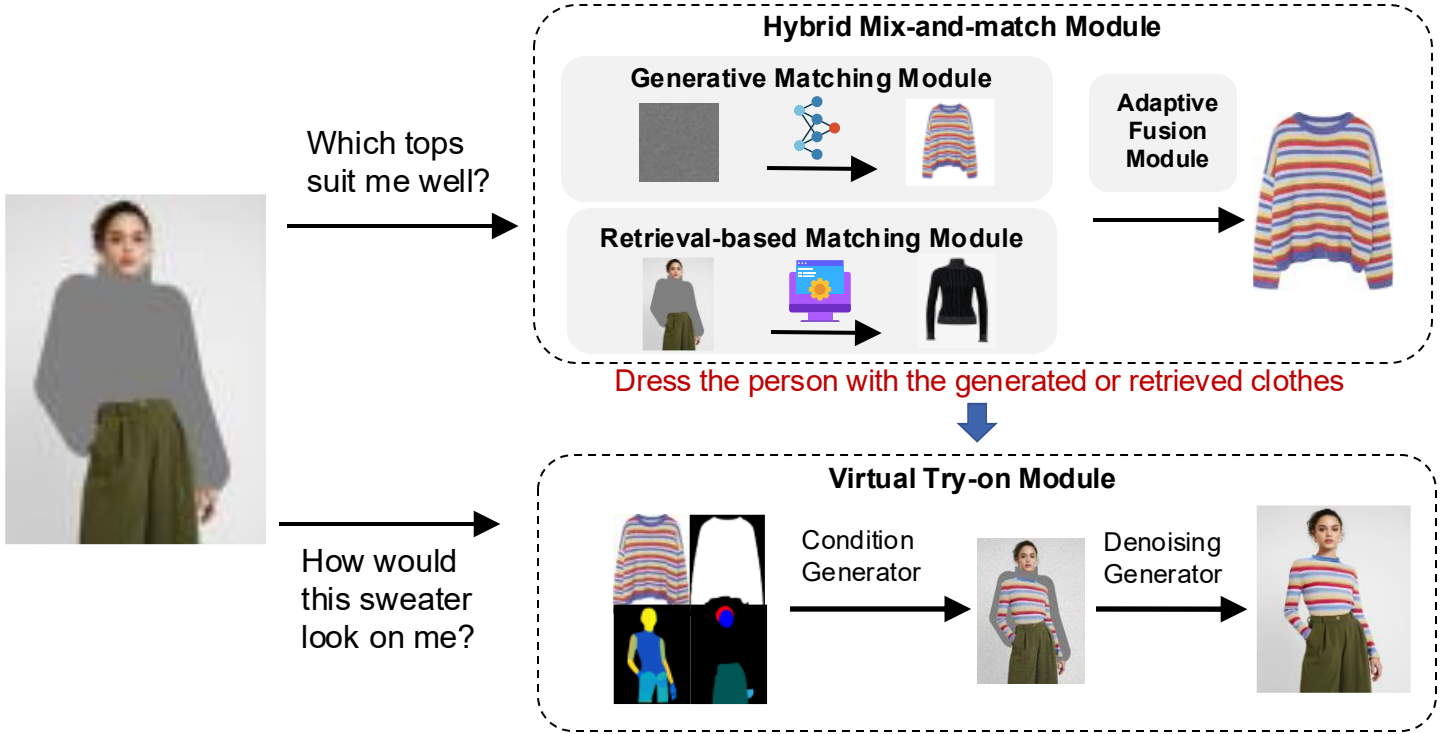
FashionDPO: fine-tune fashion outfit generation model using direct preference optimization



FashionDPO: Fine-tune Fashion Outfit Generation Model using Direct Preference Optimization. Mingzhe Yu et al. SIGIR 2025.

Bundle Generation (4/5)

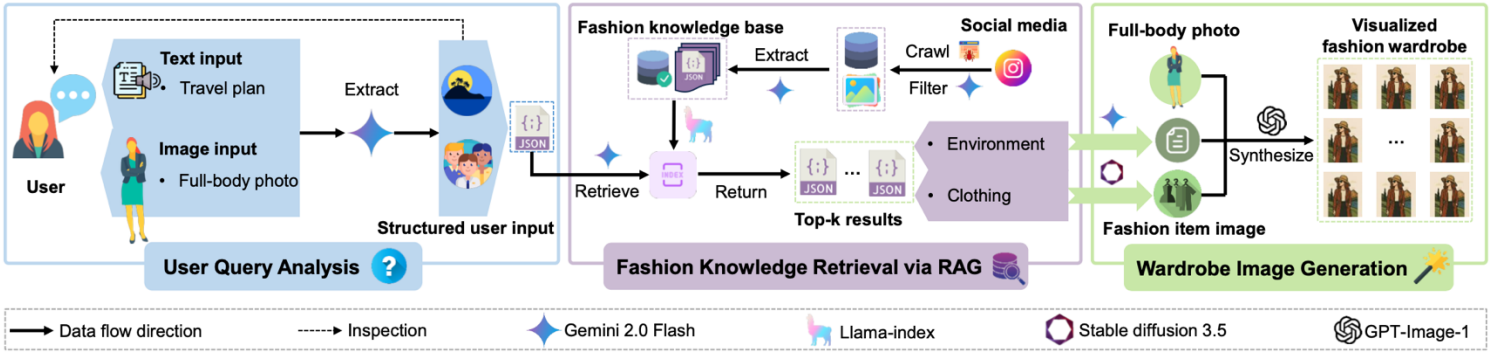
Smart Fitting Room: a one-stop framework for matching-aware virtual try-on



Smart Fitting Room: A One-stop Framework for Matching-aware Virtual Try-on. Mingzhe Yu et al. ICMR 2024.

Bundle Generation (5/5)

GenWardrobe: a fully generative system for travel fashion wardrobe construction



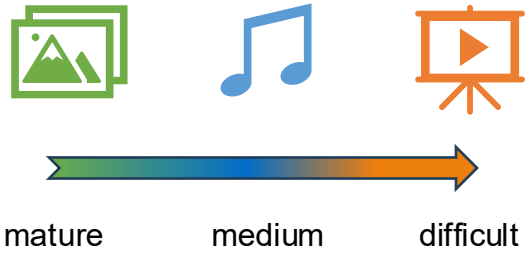
GenWardrobe: A Fully Generative System for Travel Fashion Wardrobe Construction. Peng Jin et al. ACM MM Demo 2025.
https://github.com/ShanFengShanFeng/GenWardrobe/

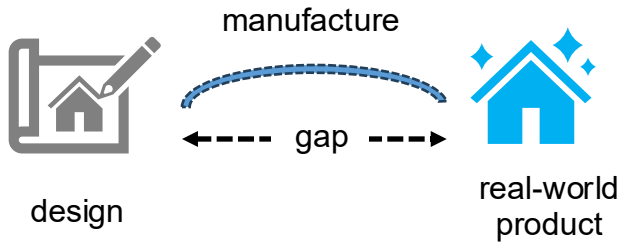
Open challenges (1/5)

Modality generalization:

It heavily depends on the progress of generative models of in certain modalities

- Mature image generation models enable image-centric bundle generation
- Music and video need more time to be readily used
- Gap between virtual design and real-world manufacturing





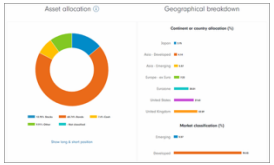
Open challenges (2/5)

Domain generalization:

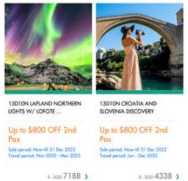
Most of the works are constrained to the fashion domain

- Investigate more domains to enable generalizable bundling algorithm, such as
  - Finance portfolio
  - Travel package
  - Music playlist
  - Video playlist
  - and more

fashion



finance



travel

music

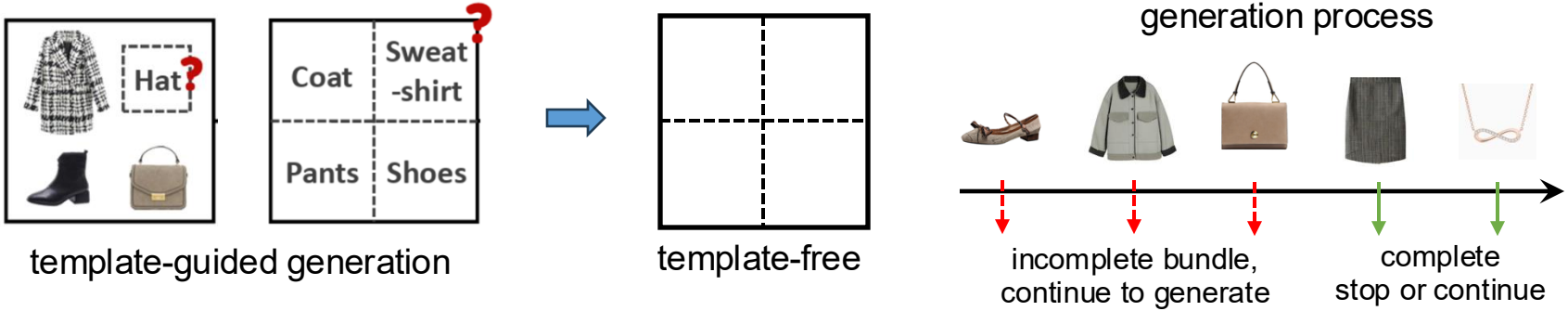
video

Open challenges (3/5)

Template-free entire bundle generation:

Current works need pre-specify the bundle size and category of each item (i.e., template)

- How to directly generate items without the input of any clue of the target item?
- How do bundle generation models determine when to stop?



Open challenges – beyond bundle (4/5)

Large item & bundle space exploration (compositional learning):

The item set is large (N), consequently, the bundle space is huge (C_N^x, x is the average bundle length), most of the bundle space has never been explored

- How to efficiently explore more bundling options?
- How to balance the exploitation and exploration?
- How to evaluate the newly-generated yet unseen bundles (the next challenge)

⇨ The challenges are not only in bundle

# bundle (outfit)	# item	bundle space (3 items per bundle)	% bundles in dataset	# bundles/item
1,013,136	583,464	$3.3*10^{16}$	$3*10^{-7}$	1.74

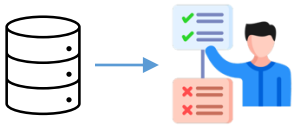
Statistics of the POG fashion outfit dataset

Open challenges – beyond bundle (5/5)

Automatic evaluation of novel bundles (creative generation):

It is easy to generate numerous bundles, how to evaluate them

- We can train an evaluator to do automatic evaluation? – how to guarantee the quality of the evaluator, which is also trained based on existing data
- Multi-agent (expert/aspect) evaluation
- Domain knowledge (aesthetics) injection
- Agentic simulation



data-driven evaluator

multi-expert

domain knowledge

agentic simulation

The evaluation problem is also prevalent in other creative generation problems, while bundle exhibits more unique challenges: large composition space, multimodal, various constraints, etc.

Summary and take-aways

- What is a bundle and why do we study bundles -- background
- Bundle recommendation – graph and contrastive learning
- Multimodal bundle construction – multimodal integration
- Bundle generation – generative models (image generation)
- Open challenges – five open challenges

THANKS and Q&A