Generative Bundle Recommendation

MA Yunshan (马云山)

May 13, 2026

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: Bundle Recommendation, Multimodal Event Forecasting
- Homepage: https://mysbupt.github.io/

What is a bundle

Example bundles in various applications

Software

Electronics

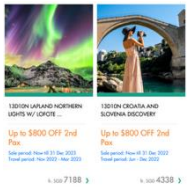
Fashion

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]

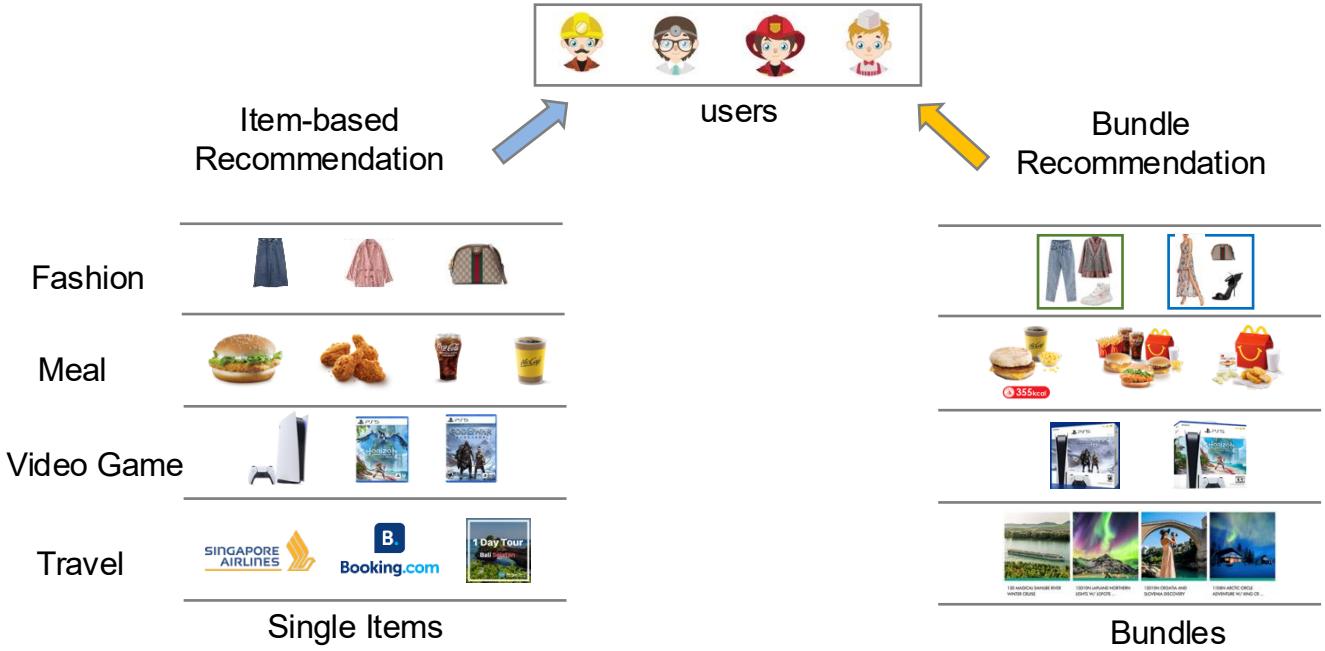
- Pure bundles
- New product bundles
- Mix-and-match bundles
- Cross-sell bundles
- Gifting bundles
- Inventory clearance bundles
- Buy-one-get-one bundles
- …

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

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

Personalized Bundle Recommendation

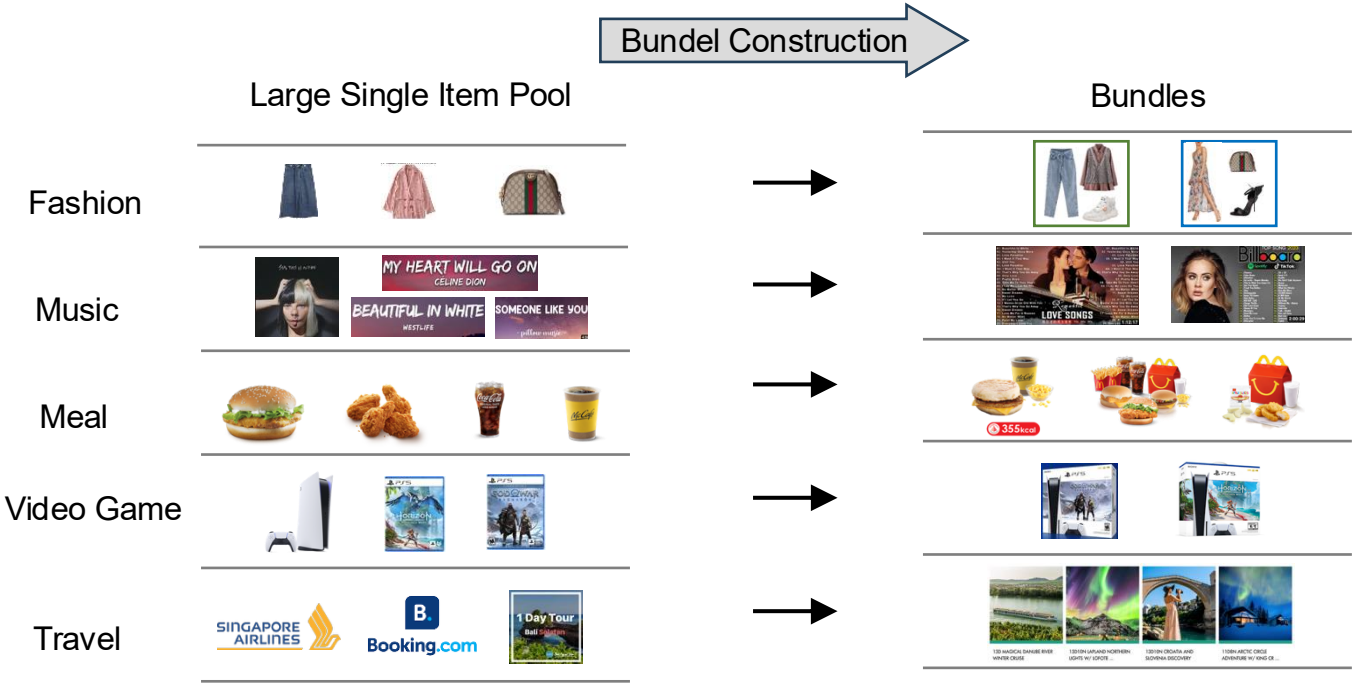
Formulation: from single item recommendation to multiple items (bundle) recommendation



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

Before Recommendation: Bundle Construction!

Formulation: **Select** a small set of items from a large pool of candidate items



The Key Patterns to be Modeled

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

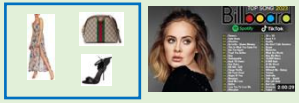
Multimodal Data Sources

Content Data of Single Item

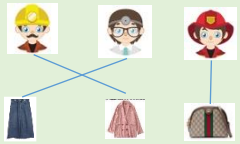
- Visual: Images, Videos, etc.
- Textual: Title, Lyrics, keywords, etc.
- Audio: Audio track, sounds, etc.

Relational Data Between Items

- Existing Bundles



- User-item Interactions



Items within the same bundle are **Related**

Explicitly Related: Similar on certain Semantic Aspects

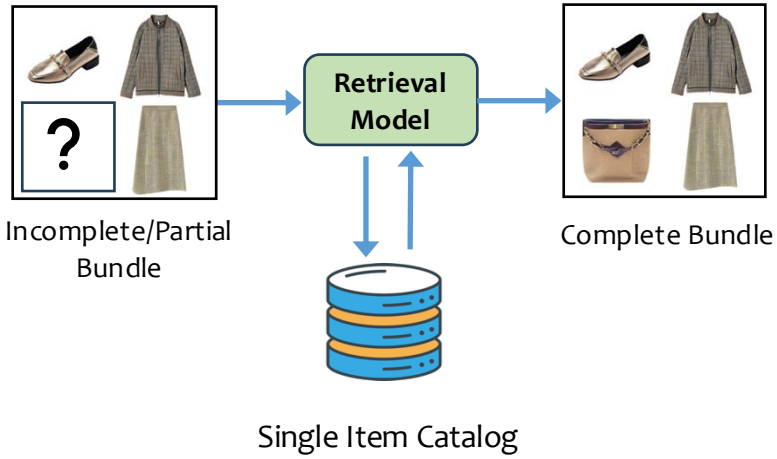
- Share keywords, attributes, text spans, etc.
- Share visual objects, concepts, patterns, etc.
- Share instruments, melodies, styles, etc.

Implicitly Related: Co-exist and Co-interact

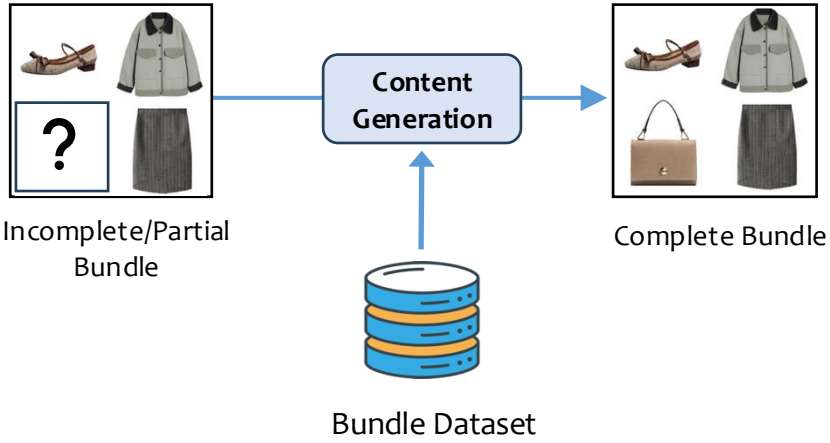
- Mix-and-Match (co-exist)
- Co-purchase (co-interact)

Two Paradigms of Bundle Construction

Paradigm 1: Retrieval



Paradigm 2: Content Generation

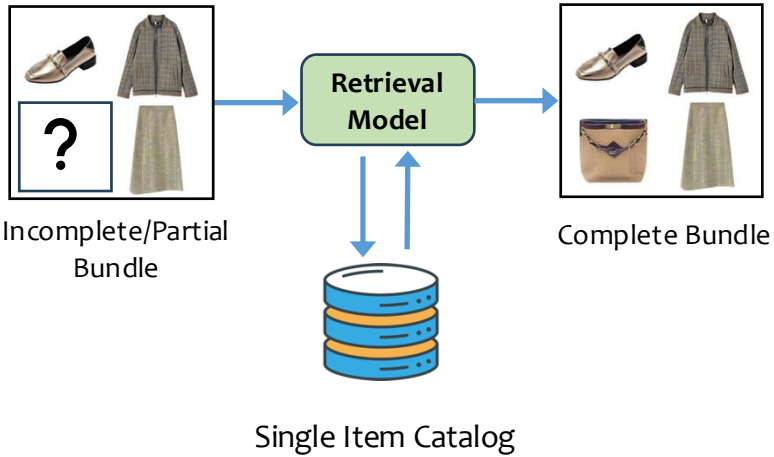


Outline

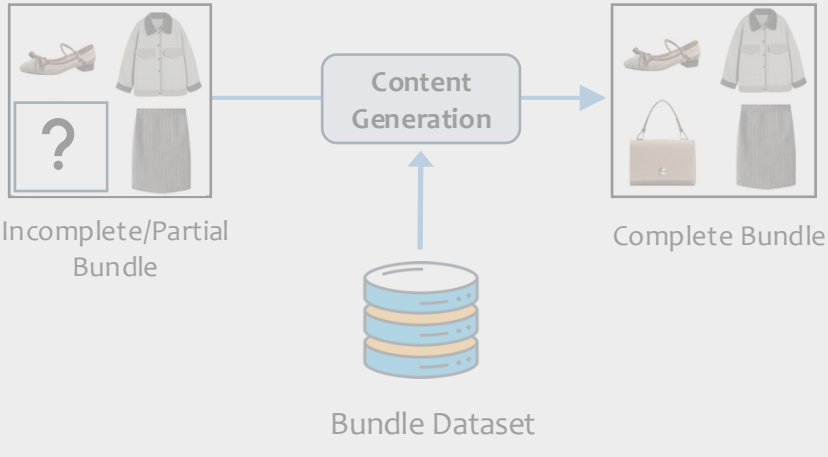
- Background and Preliminary
- Generative Models in
  - Paradigm 1: Discrete Diffusion for Bundle Construction
  - Paradigm 2: Dual-Diffusion for Bundle Generation
- Summary and Future Work

Two Paradigms of Bundle Construction

Paradigm 1: Retrieval



Paradigm 2: Content Generation



Motivation

a) **AR-style** bundle construction inherits a left-to-right factorization,

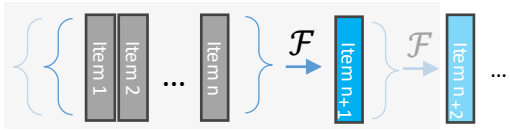
⚠ 1. **Overemphasis on order**: Sparse or even misleading supervision

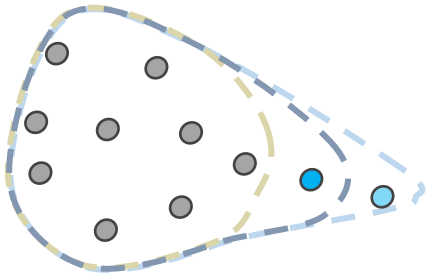
⚠ 2. **Unidirectionally**: Bundles require modeling dense higher-order dependencies

b) **One-Shot-style** adopts an over-simplified dependence factorization,

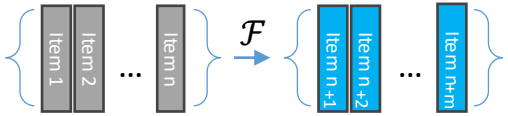
⚠ Omit relations **within the generated items**

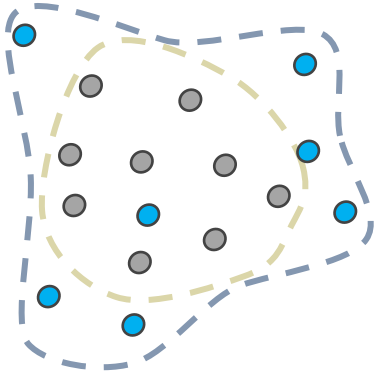
(a) Next-item Prediction





(b) One-Shot Completion



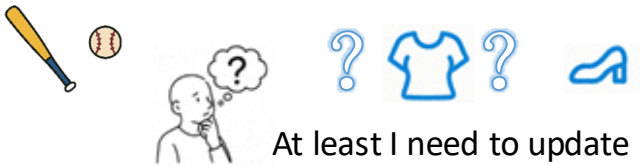


Intuition

①

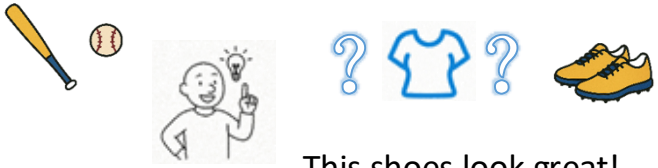
How do I construct my bundle?

②



At least I need to update my shirt and shoes

③

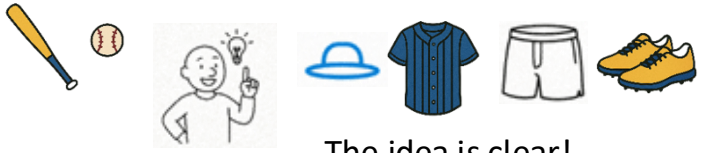


This shoes look great!

④

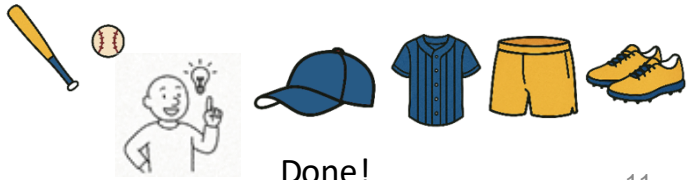
Let me think…
The shirt shall be…

⑤



The idea is clear!

⑥

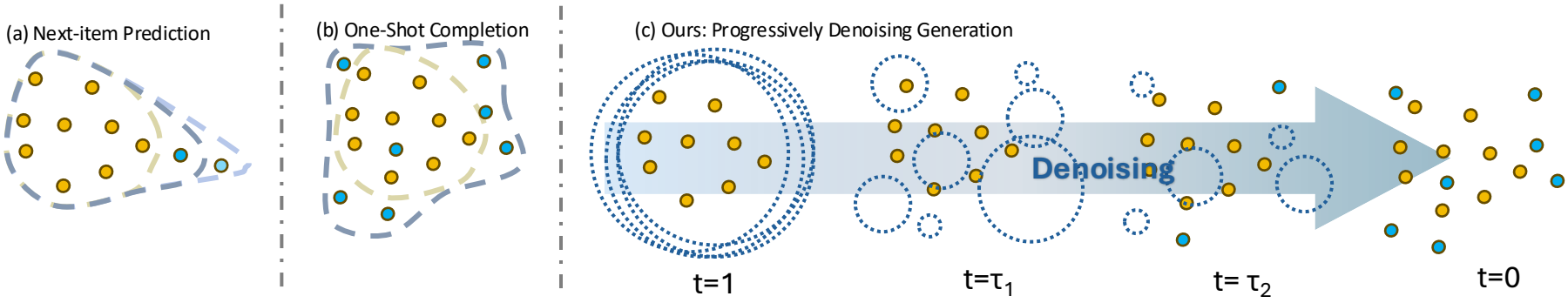


Done!

Key Idea

- We aim for:
  - Need to be order invariance $P(N,k) \longrightarrow C(N,k)$
  - A better modeling of the intra-bundle relations

Diffusion Model could be an option



Key Technical Challenges

Combinatorial Explosion

Although the problem has been simplified from a permutation task to a combination task, it still suffers from an **enormous search space** caused by the **large item catalog size** N.

Much harder than next-item prediction, especially in large catalogs (e.g., playlists)

Dataset	#U	#I	#B	#B-I	#U-I	#Avg.I/B	#Avg.B/I
POG_dense	2,311,431	31,217	29,686	105,775	6,345,137	3.56	3.39
Spotify	118,994	254,155	20,000	1,268,716	36,244,806	63.44	4.99

⚠ Sparse Supervision:
On average, each item appears in only a very limited number of bundles, resulting in **severe supervision sparsity** when training based solely on item IDs.

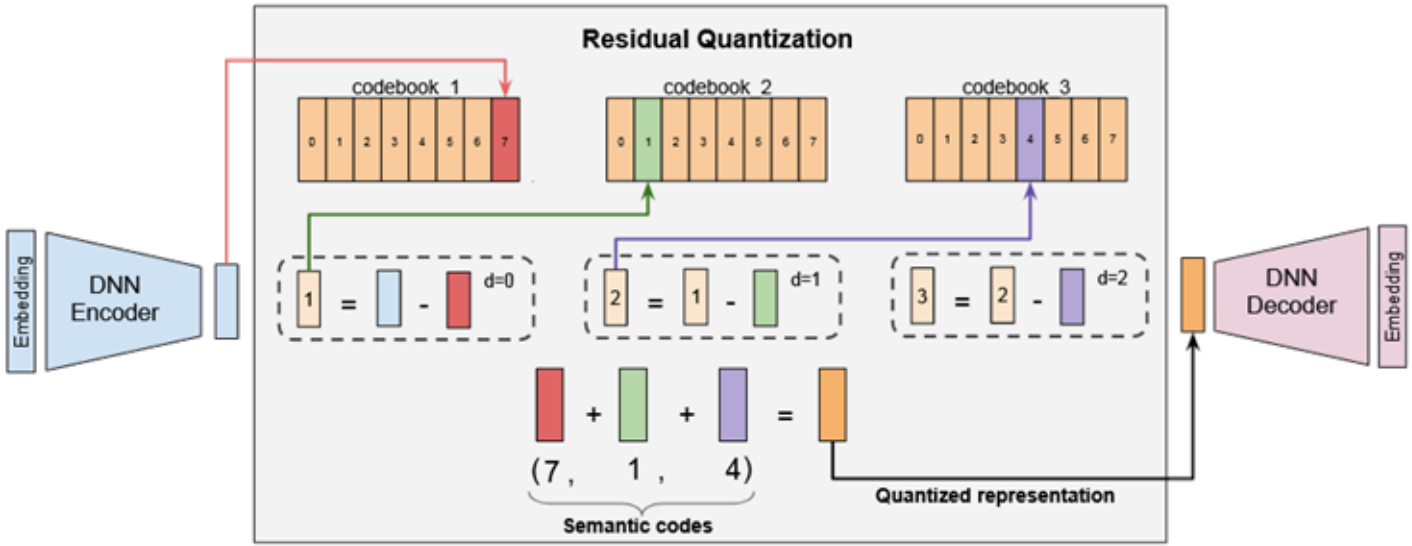
Residual Vector Quantization

Core Idea:

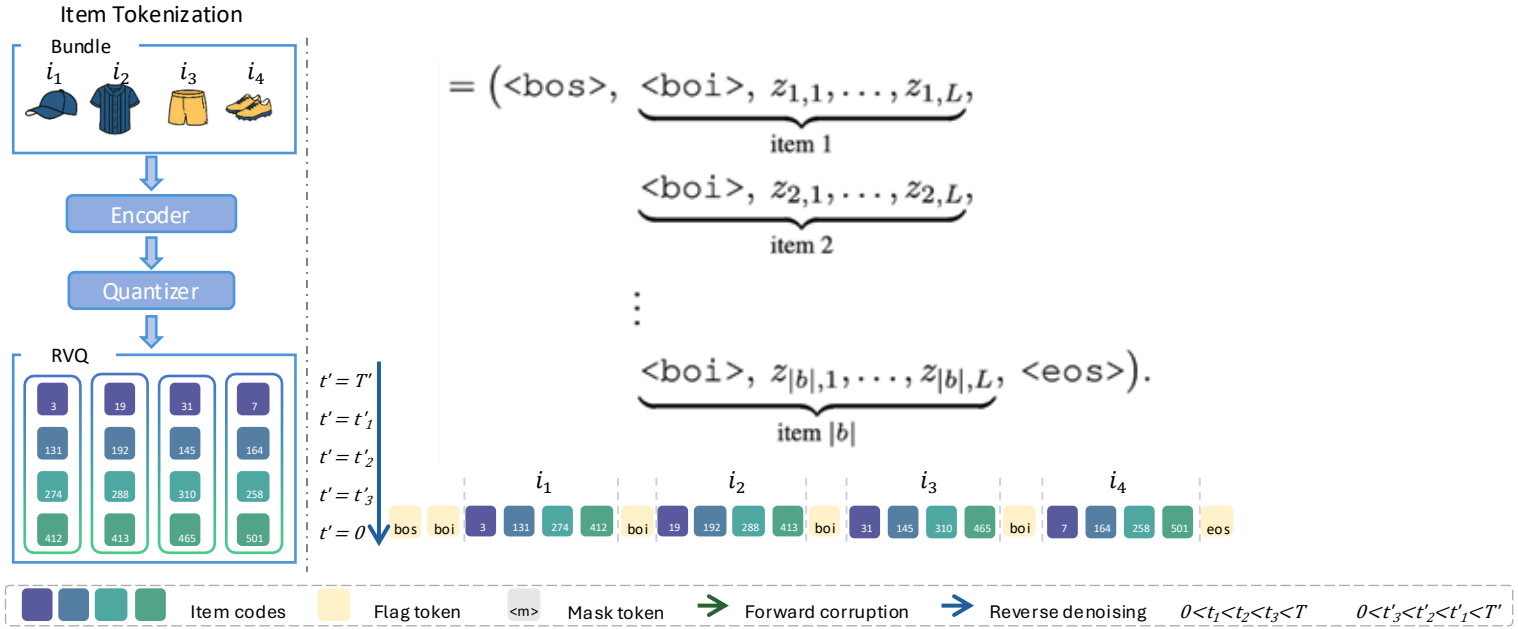
Use **Residual Vector Quantization** to quantize item embeddings into multiple discrete tokes.

Factorizes a large space into smaller codebooks

Reduces per-codebook vocabulary size: $N = 254{,}151$ **v.s.** $V = 128 \times 4$

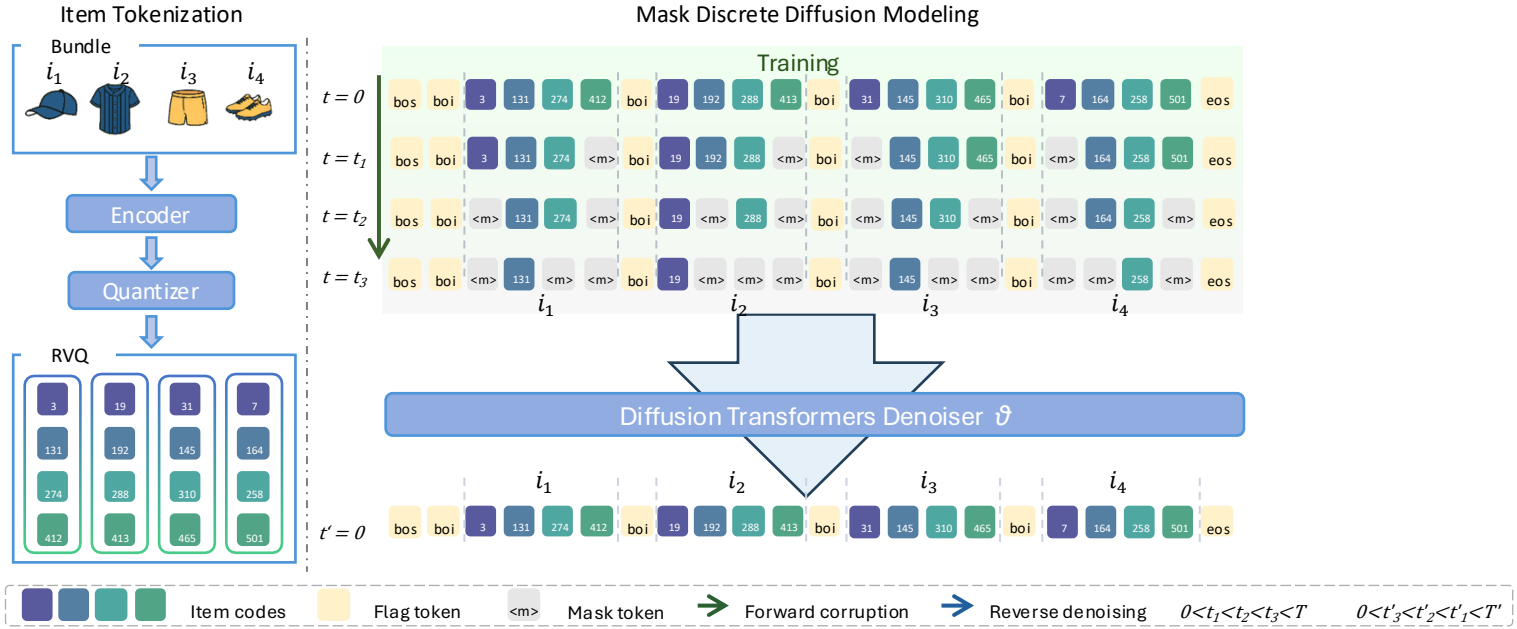


Framework



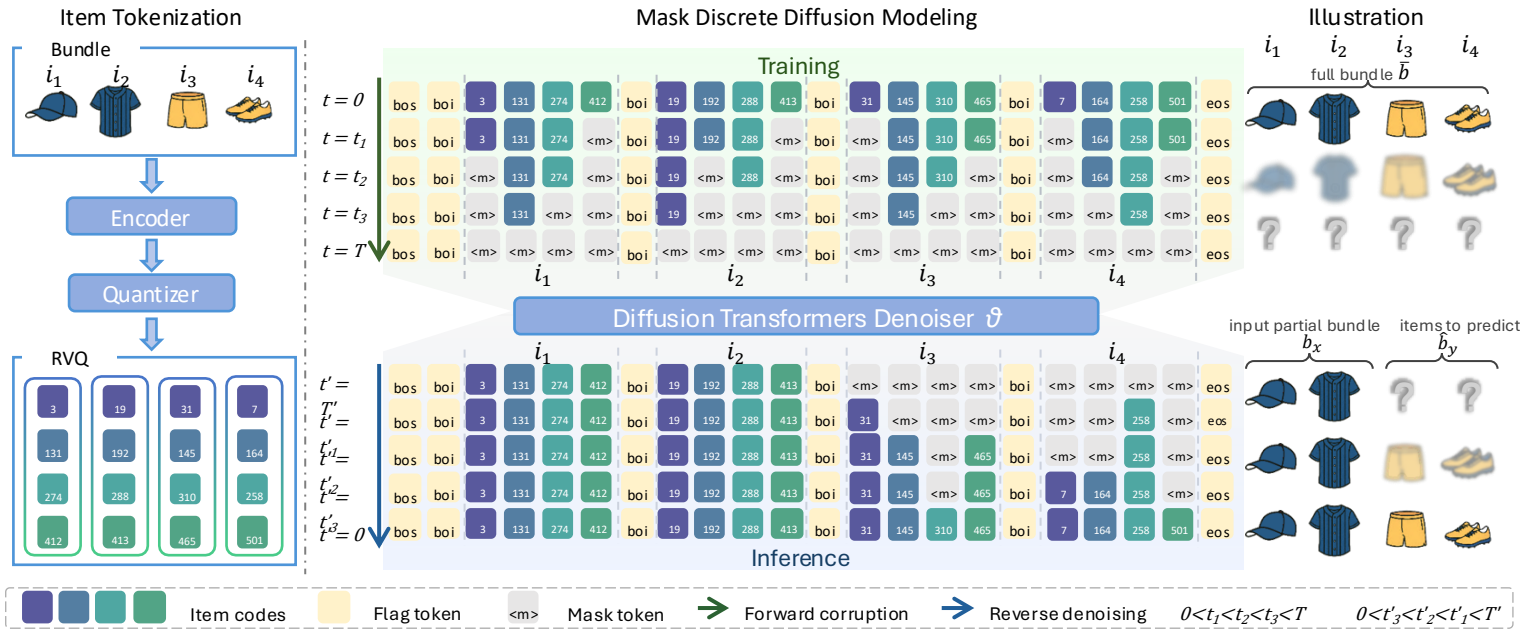
- Backbone: Discrete Diffusion Model (DDM) masked-denoising
- Representation: Residual Vector Quantization (RVQ)

Framework



- How do we achieve approximate order-agnostic modeling?
- The masked denoising process
- + Data augmentation: random item swapping

Framework



Experiment

Table 1: Overall performance comparison between our DDBC and baselines. "%Improv." denotes the relative improvement over the strongest baseline. Best in **bold**, second best underlined.

Model (A/beam)	Spotify$_{k=30}$			Spotify$_{k=60}$			Spotify$_{k=90}$			POG$_{k=4}$		
	F1 ↑	Jacc ↑	OAS ↑	F1 ↑	Jacc ↑	OAS ↑	F1 ↑	Jacc ↑	OAS ↑	F1 ↑	Jacc ↑	OAS ↑
CLHE	0.071	0.039	0.373	0.100	0.054	0.446	0.119	0.065	0.486	0.140	0.096	0.446
Bi-LSTM	0.124	0.071	0.489	0.062	0.034	0.430	0.047	0.025	0.426	0.035	0.024	0.390
SASRec	0.070	0.043	0.318	0.089	0.054	0.310	0.050	0.029	0.285	0.169	0.114	0.468
TIGER	0.093	0.053	0.329	0.129	0.076	0.413	0.123	0.070	0.480	**0.213**	**0.157**	**0.546**
BundleNAT	0.153	0.090	0.454	0.101	0.056	0.438	0.095	0.052	0.446	0.145	0.097	0.462
BundleMLLM	0.046	0.024	0.296	0.045	0.024	0.324	0.052	0.027	0.355	0.070	0.047	0.322
DDBC	**0.282**	**0.178**	**0.618**	**0.296**	**0.185**	**0.668**	**0.287**	**0.177**	**0.684**	0.139	0.098	0.526
%Improv. +	84.3%	97.8%	26.4%	129.5%	143.4%	49.8%	133.3%	152.9%	40.7%	–	–	–

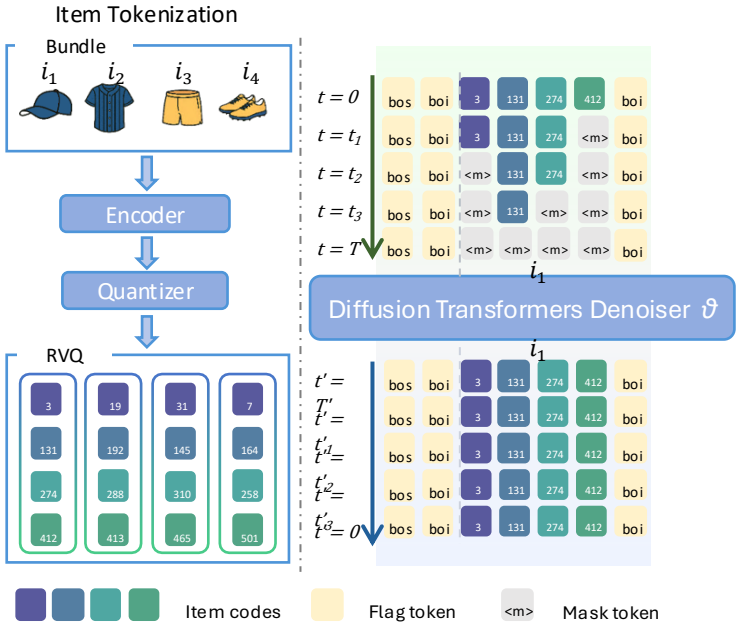
DDBC performance ↑ as bundle length ↑

→ DDBC effectively captures the higher-order intra-bundle item relations, particularly for long-sequence bundles with rich structural dependencies.

Ablation Study

Table 4: Ablation study of key components.

Variant	F1	Jacc	OAS
Our proposed DDBC	0.282	0.176	0.660
w/o RVQ	0.021	0.011	0.557
w/o boi token	0.176	0.104	0.538
w/o data augmentation	0.254	0.158	0.599
w/o token validity filter	0.276	0.173	–



Efficiency

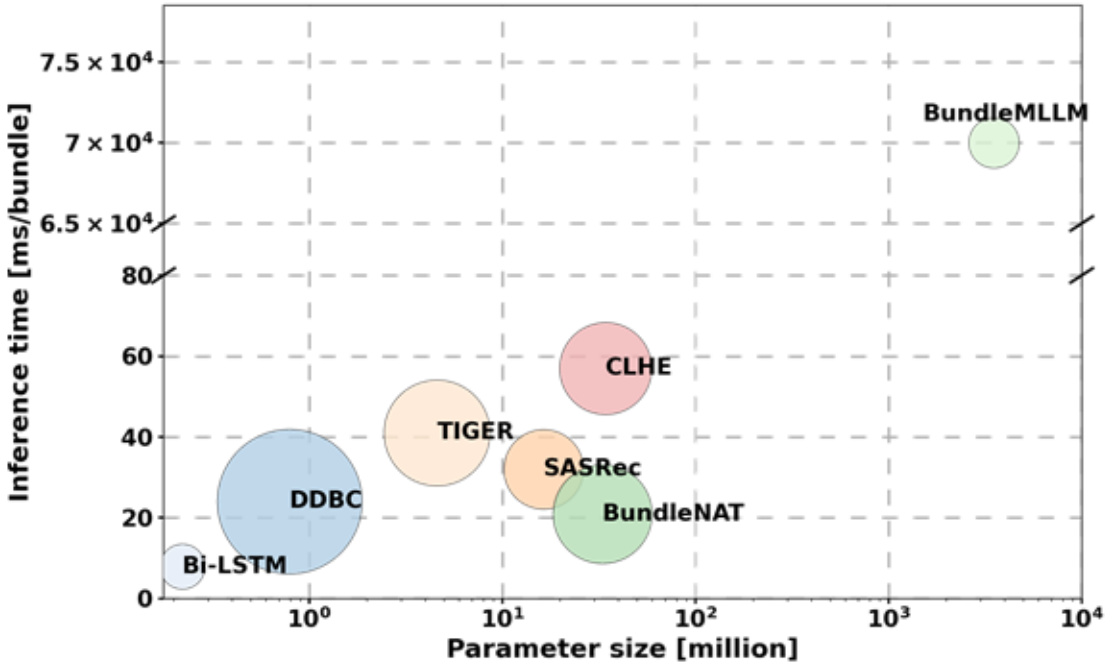
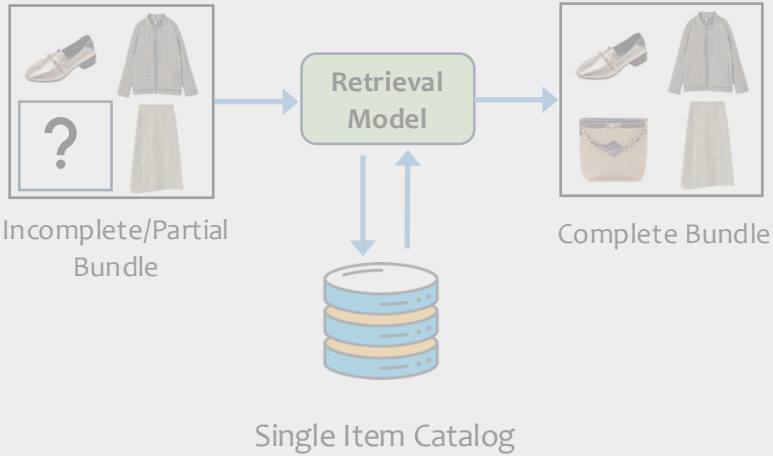


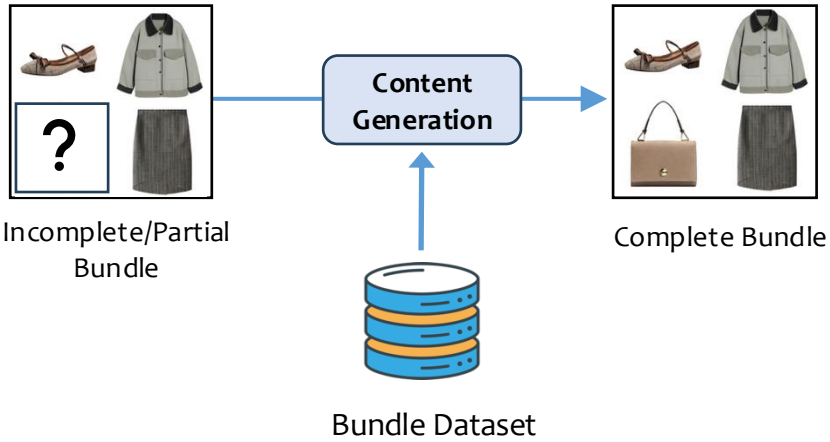
Illustration of the model efficiency comparison.
bubble radius corresponds to overall performance (larger is better).

Two Paradigms of Bundle Construction

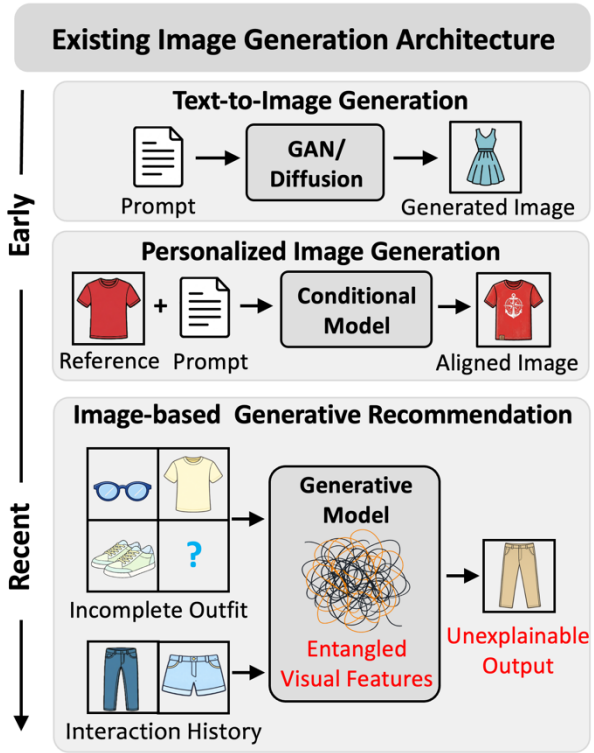
Paradigm 1: Retrieval



Paradigm 2: Content Generation



Motivation



Limitations:

1. User behavior modeling remains insufficient.

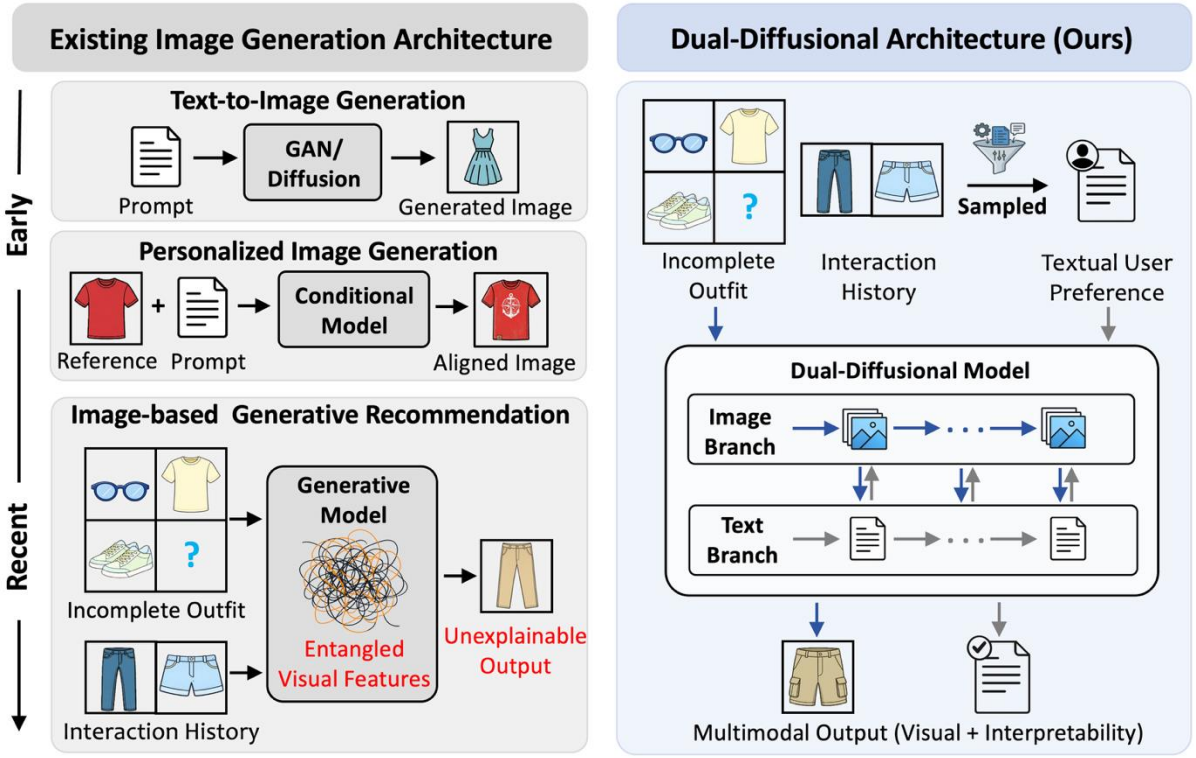
- interacted item images as unimodal inputs
- user–item interactions are inherently sparse
- visual data often contain redundant or preference-irrelevant details

2. Explainability is largely absent.

- generation models treat image generation as the sole output
- solutions are typically add-on patches to image-centric backbones

Common root cause:
the restrictive image-centric architectural design

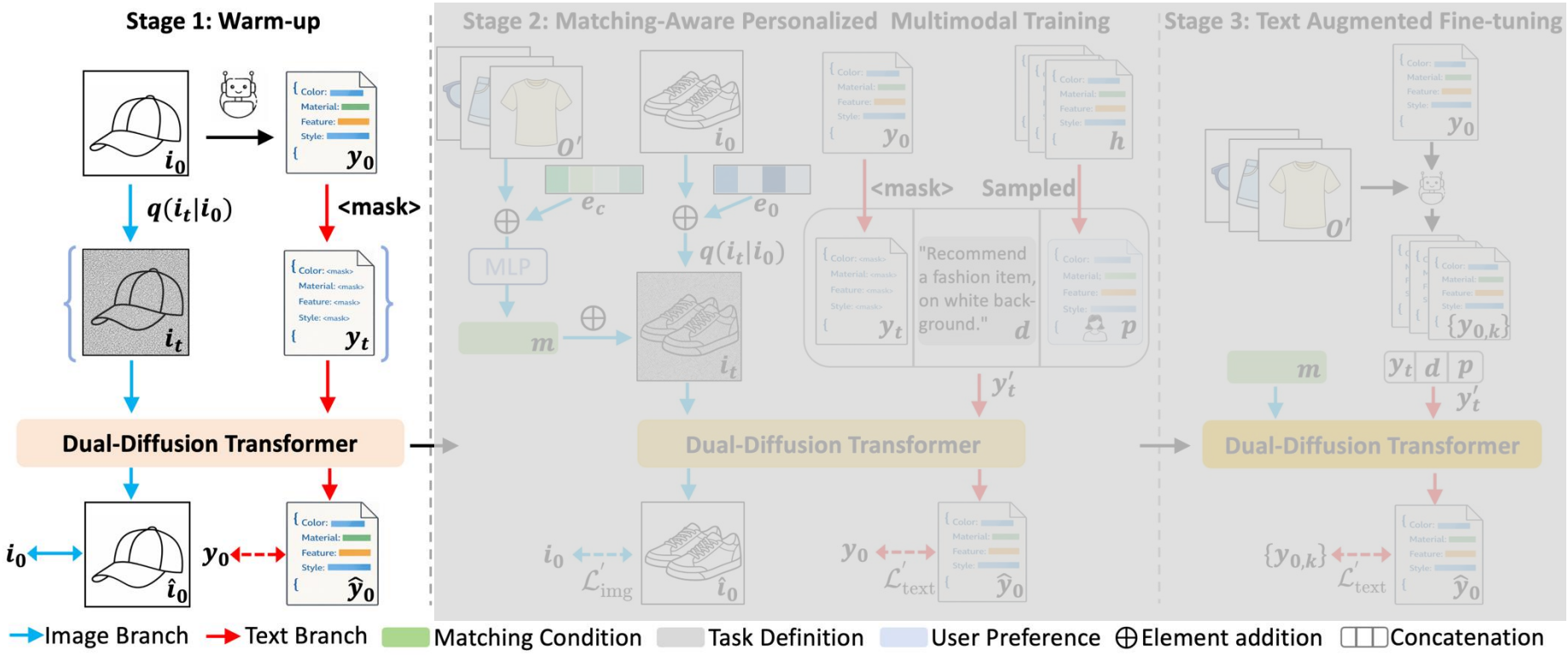
Motivation



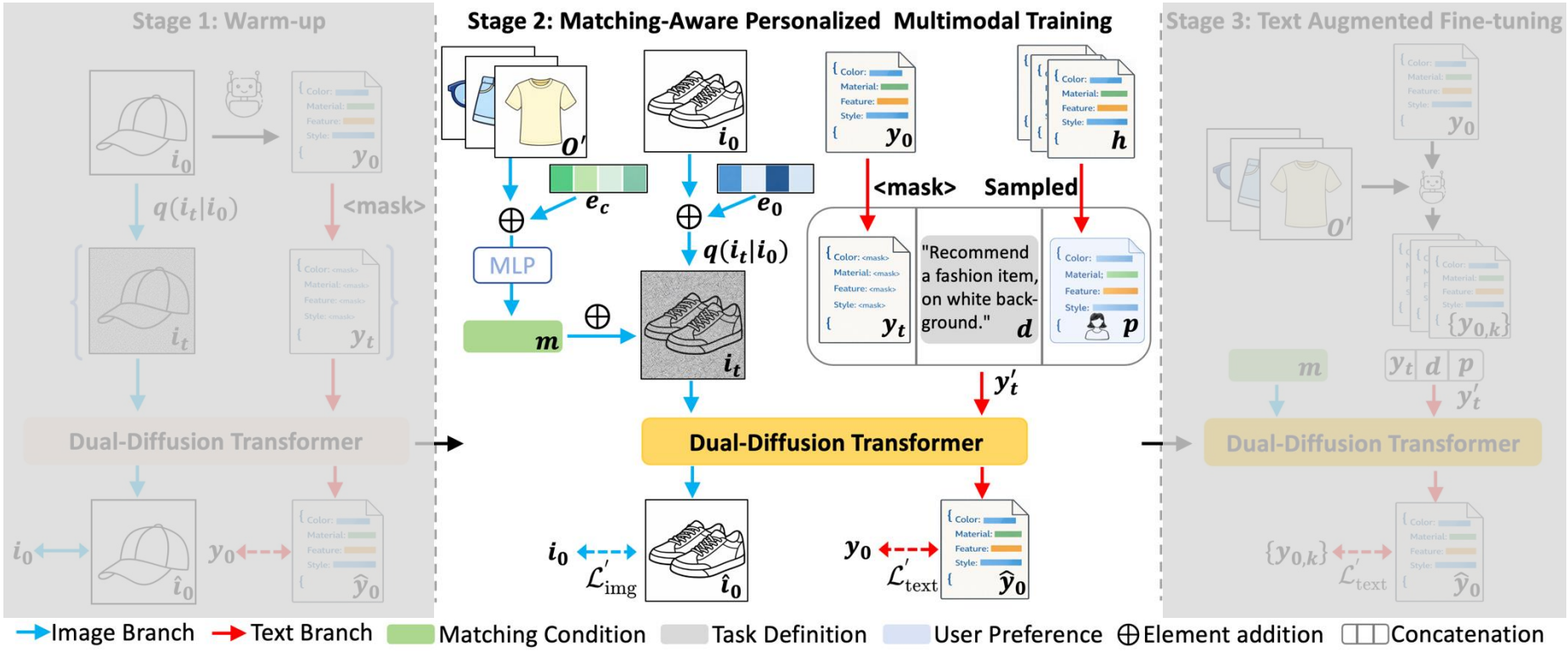
Multimodal architectures

- Sample textual user preference
- Support image–text inputs
- Image Branch
- Text Branch
- Jointly generate visual and textual outputs

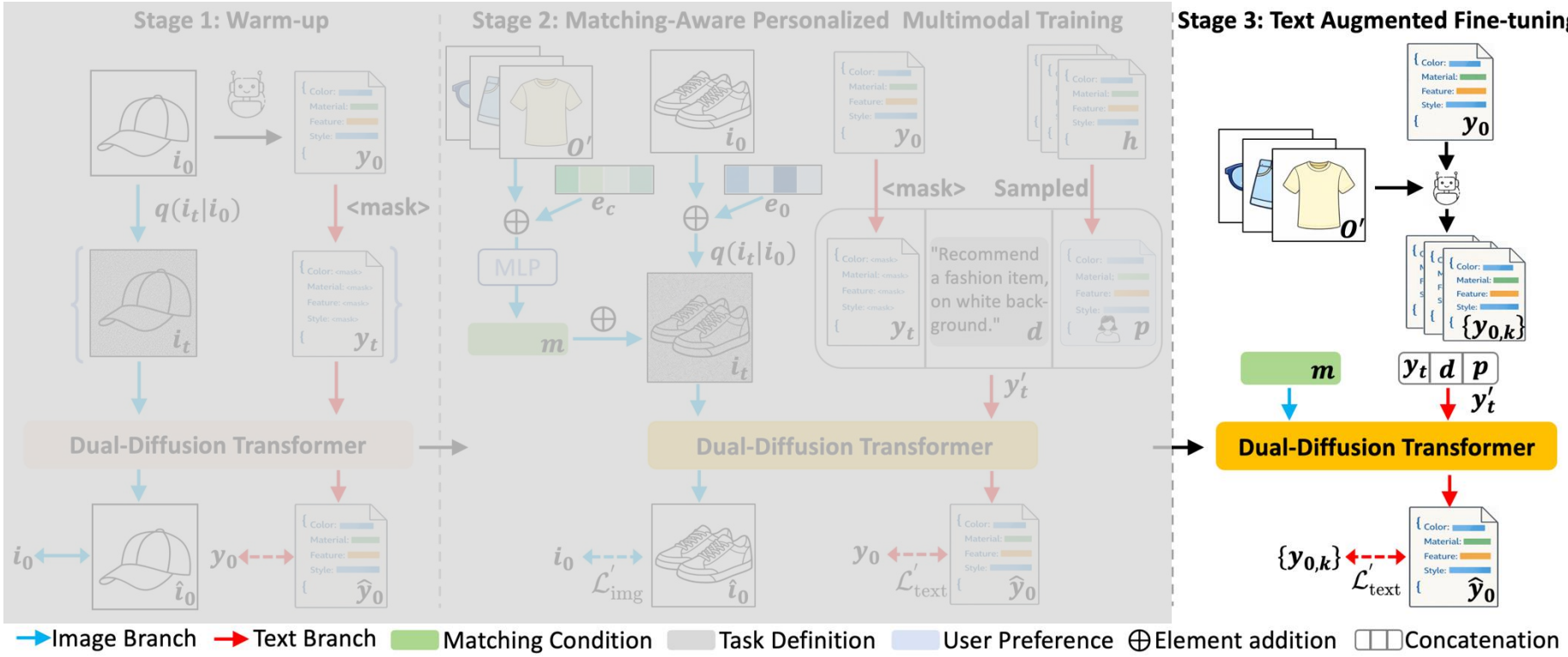
Method



Method



Method



Experiments

Datasets:

iFashion[1]
1.01 million outfits, 583K fashion items

Polyvore-U[2]
21,889 outfits, 261K fashion items

[1] Chen, Wen, et al. "POG: personalized outfit generation for fashion recommendation at Alibaba iFashion." Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019.
[2] Lu, Zhi, et al. "Learning binary code for personalized fashion recommendation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Experiments

Datasets:

iFashion[1]
1.01 million outfits, 583K fashion items

Polyvore-U[2]
21,889 outfits, 261K fashion items

Image Evaluation Metrics:

1) **IS, IS-acc**: The quality of the generated images
2) **Comp.** : The compatibility of the generated images
3) **Per.** : The personalization of the generated images
4) **Div.**: The diversity of the generated images

Text Evaluation Metrics:

1) **Align**: generated texts - generated images
2) **Comp.** : The compatibility between generated texts – outfit captions
3) **Per.** : The personalization between generated texts – user preference
4) **Div.**: The diversity of the generated texts

Experiments

Overall Performance Comparison

Dataset	iFashion									Polyvore-U								
Task	PFITB					GOR				PFITB					GOR			
Evaluation metric	IS	IS-acc	Comp.	Per.	Div.	IS	IS-acc	Per.	Div.	IS	IS-acc	Comp.	Per.	Div.	IS	IS-acc	Per.	Div.
SD-v1.5 [32]*	22.54	0.76	0.08	46.31	0.56	23.20	0.77	46.45	0.57	17.10	0.73	0.70	51.05	0.55	16.95	0.73	50.99	0.52
SD-v2 [32]*	21.66	0.71	0.04	46.60	**0.62**	22.19	0.74	46.60	**0.59**	14.83	0.68	0.60	51.29	**0.60**	14.88	0.67	51.23	**0.65**
SD-v1.5 [32]	26.76	0.83	0.46	53.16	0.50	26.90	0.84	53.24	0.49	17.12	0.72	0.75	58.20	0.50	17.24	0.72	58.16	0.47
SD-v2 [32]	25.85	0.80	0.39	52.99	0.47	25.82	0.82	53.06	0.47	15.59	0.67	0.71	58.79	0.46	16.33	0.70	58.91	0.44
SD-naive [32]	25.45	0.80	0.36	52.95	0.39	25.43	0.81	52.95	0.40	15.45	0.66	0.73	59.24	0.41	15.48	0.67	59.12	0.40
ControlNet [48]	27.76	0.81	0.16	49.90	0.41	28.49	0.82	49.91	0.43	18.93	0.77	0.73	55.44	0.42	19.21	0.77	55.40	0.41
DiFashion [42]	29.99	0.90	0.58	55.86	0.33	30.04	0.90	55.54	0.33	19.67	0.84	0.80	61.44	0.37	18.95	0.83	61.16	0.36
FashionDPO [47]	**33.80**	**0.91**	**0.74**	60.39	0.36	**32.37**	**0.91**	59.98	0.35	24.14	**0.89**	**0.83**	64.67	0.38	**24.93**	0.87	64.79	0.36
DualFashion	31.08	0.88	0.69	**65.17**	0.38	31.53	0.89	**64.05**	0.39	**25.30**	**0.89**	0.81	**67.40**	0.42	24.85	**0.88**	**67.13**	0.40

- Slightly lower scores than FashionDPO in terms of IS, IS-acc, and Comp
- Significant improvement in the Personalization score
- Stable Diffusion models exhibit the highest Diversity scores, but stems from randomness and lack of constraints

Experiments

Ablation Study – text generation

Method	Comp.					Per.					Div.				
	Color	Material	Features	Style	Ave.	Color	Material	Features	Style	Ave.	Color	Material	Features	Style	Ave.
w/o warm-up	0.70	0.71	0.33	0.42	0.54	0.49	0.65	0.50	0.67	0.58	3.31	2.25	4.90	2.82	3.32
w/o learnable embeddings	0.84	0.70	0.37	0.55	0.62	0.56	0.72	0.53	0.74	0.64	3.36	2.28	4.95	2.81	3.35
only image loss	0.63	0.68	0.30	0.34	0.49	0.50	0.57	0.34	0.62	0.51	**4.03**	**4.95**	5.51	**5.38**	**4.97**
only text loss	0.89	0.81	0.40	0.56	0.66	**0.57**	**0.74**	**0.60**	**0.84**	**0.69**	3.21	2.72	**5.52**	2.48	3.48
image and text loss	**0.91**	**0.84**	0.39	0.59	0.68	**0.57**	**0.74**	0.55	0.80	0.67	3.38	2.63	**5.52**	3.03	3.64
w/ text augmented fine-tuning **(Ours)**	**0.91**	0.83	**0.44**	**0.71**	**0.72**	**0.57**	**0.74**	0.57	0.82	0.68	3.47	3.64	**5.52**	4.36	4.25

- Without warm-up results in bad performance in all metrics
- Remove learnable embeddings has more impact on compatibility
- Train with only image loss results in inferior compatibility and higher diversity score
- Incorporate text augmented fine-tuning significantly improve the compatibility and diversity score

Experiments

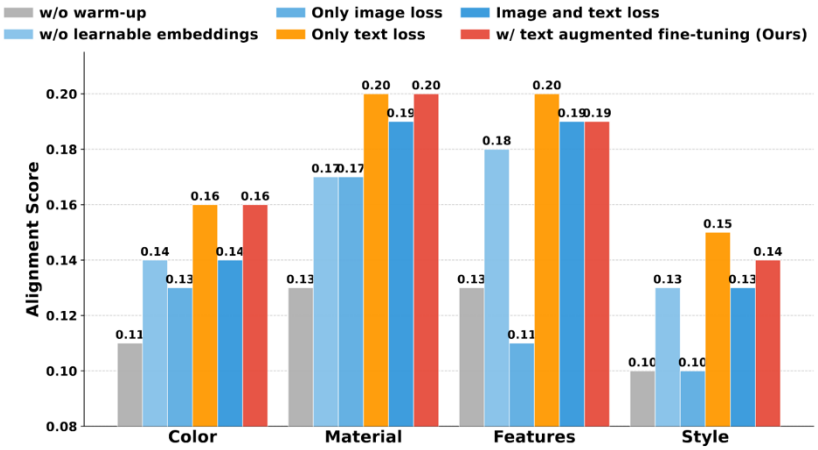
Ablation Study – image generation

Fashion item image generation

Method	IS	IS-acc	Comp.	Per.	Div.
w/o warm-up	28.77	0.80	0.62	62.68	0.31
w/o learnable embeddings	30.23	0.85	0.64	62.93	0.31
only image loss	30.29	**0.93**	**0.71**	63.41	0.36
only text loss	26.57	0.84	0.65	**65.86**	0.33
image and text loss	30.61	0.87	0.69	64.48	0.34
w/ text augmented fine-tuning **(Ours)**	**31.08**	0.88	**0.71**	65.17	**0.38**

- with only image loss, the model achieves significant improvements in IS-acc and Comp. scores
- optimize with only text loss yields higher personalization performance

Image-text alignment



- train with only image loss results in poor alignment
- optimize with only text loss achieves the best performance

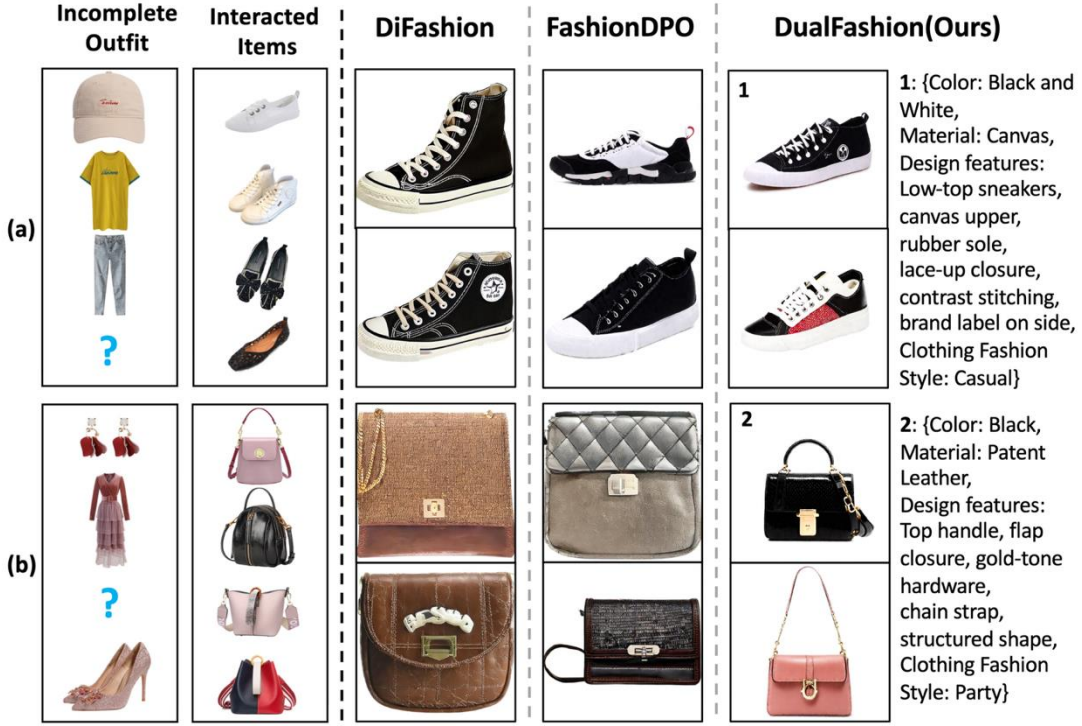
Experiments

Qualitative Results - Generative Outfit Recommendation (GOR)

- **Quality**: all models generate visually plausible items
- **Compatibility**: Baselines generate items that match individual pieces but lack global coherence
- **Personalization**: DualFashion better reflects user preferences observed from interacted items

Experiments

Qualitative Results - Personalized Fill-in-the-Blank (PFITB)

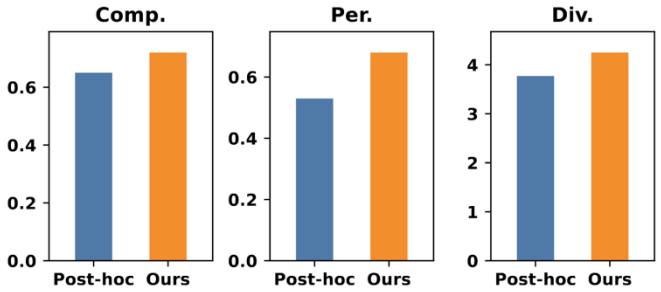


- **Quality**: DiFashion in group (b) produces bags with less structural details, whereas DualFashion maintains clean silhouettes and realistic materials
- **Compatibility**: DualFashion generates items that better complement the incomplete outfit
- **Personalization**: DualFashion generates items that are more strongly aligned with user preferences than the baselines

Experiments

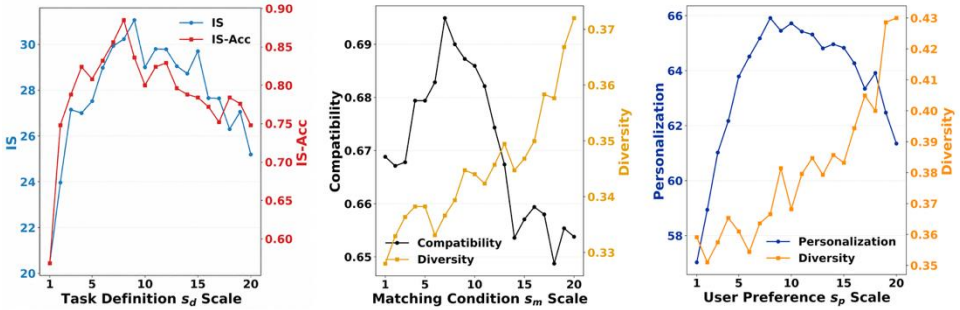
Model Study

a. Caption Interpretability



- Post-hoc: firstly generates the fashion item image, and then applies image-to-text model to generate the post-hoc caption
- the personalization score has significant improvement

b. Hyper-parameter Analysis

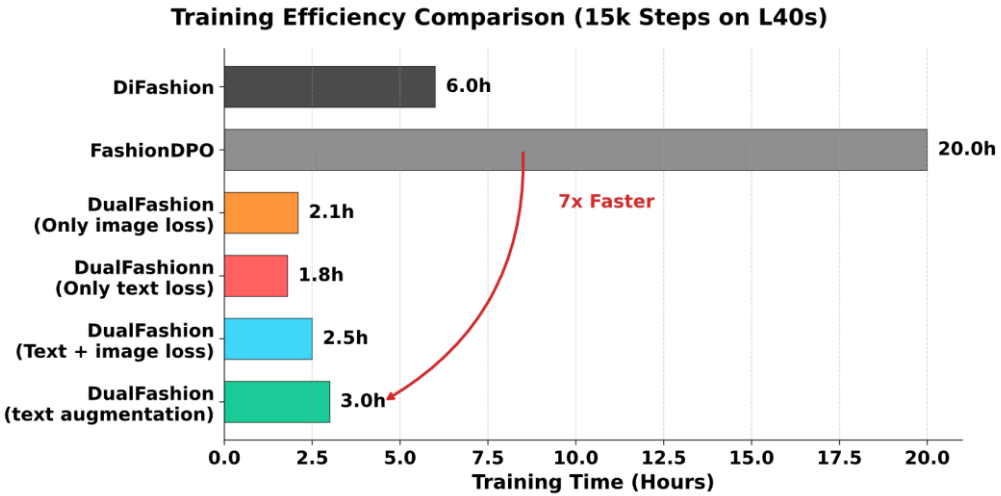


- Increase the task definition scale initially enhances generation quality, but excessive values restrict model flexibility and degrade performance.
- Moderate matching and user preference scales achieve optimal overall results, whereas overly high values sacrifice compatibility and personalization in favor of diversity.

Experiments

Model Study

c. Time Cost Analysis



- DualFashion is approximately **7x faster** than FashionDPO, which significantly reduces the overall training time.
- predict only masked captions and optimizing the text loss incurs the lowest computational cost among all settings

Conclusion and Future Work

- What is a bundle and why do we study bundles
- Discrete diffusion for bundle construction
- Dual-diffusion for bundle generation
- Next steps:
  - Unify the bundle construction with content generation models
  - Extend content generation to more modalities, e.g., music, video, etc.

THANKS and Q&A