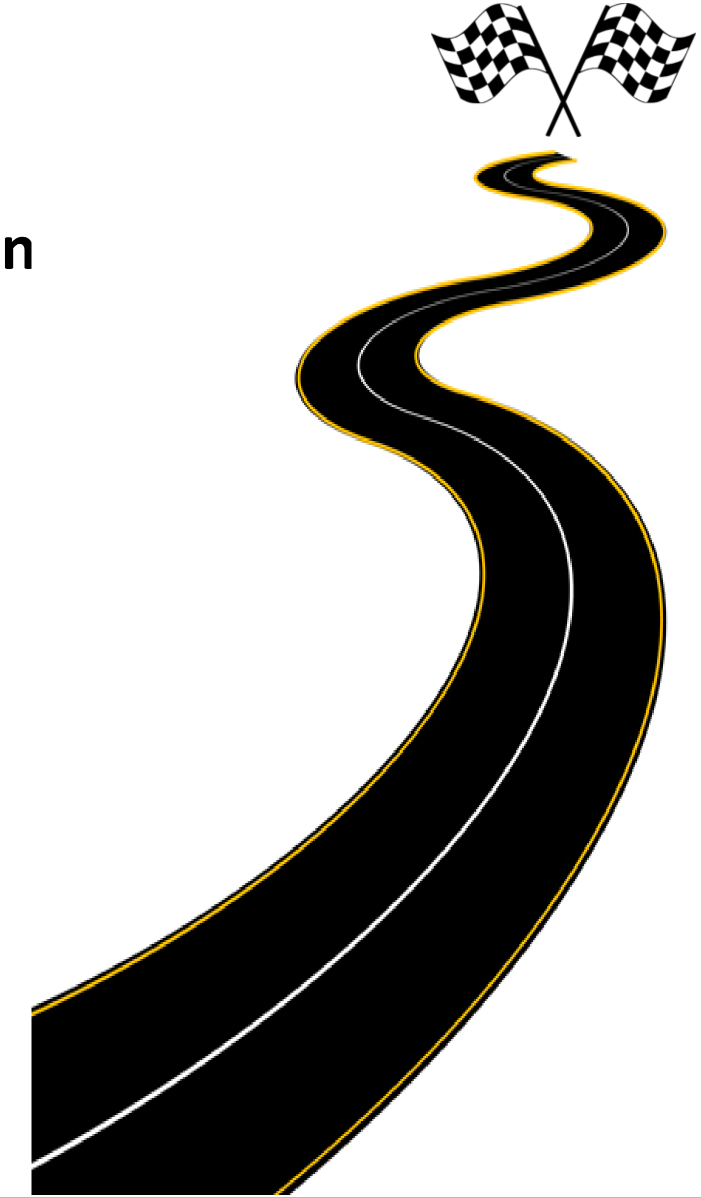
Personalized Parameter-Efficient Fine-Tuning of Foundation Models for Multimodal Recommendation

Sunwoo Kim

Hyunjin Hwang

Kijung Shin

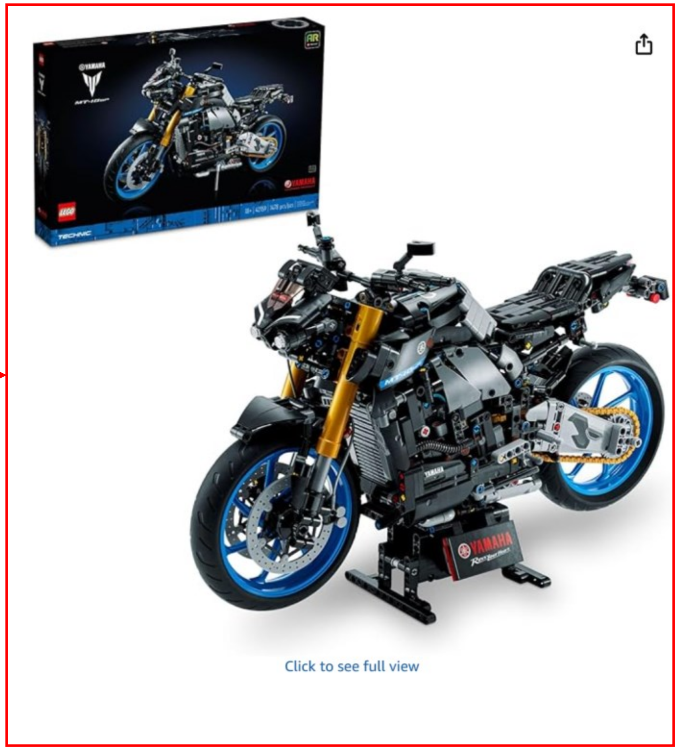
Contents

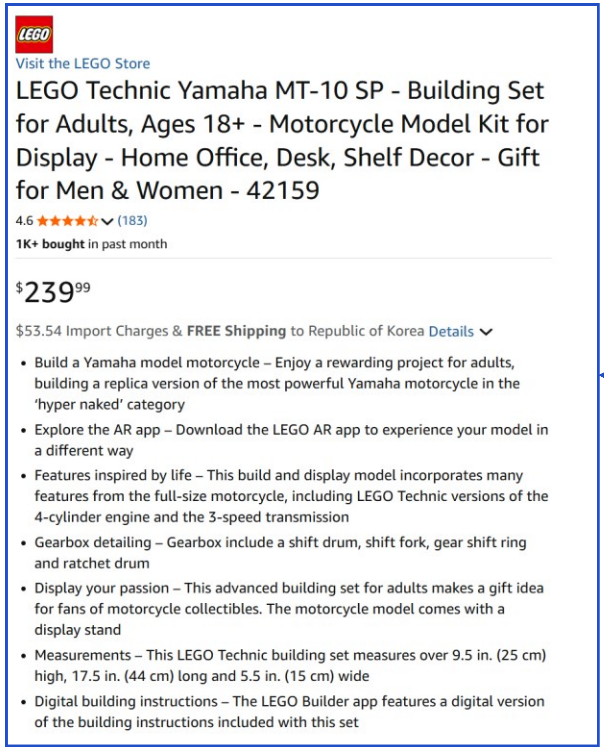


Research background

- **[Multimodal recommendation]** A recommendation task that uses multimodal item information, such as item images and textual descriptions, to recommend relevant items to users.

Item image
(visual information)




LEGO

Visit the LEGO Store

LEGO Technic Yamaha MT-10 SP - Building Set for Adults, Ages 18+ - Motorcycle Model Kit for Display - Home Office, Desk, Shelf Decor - Gift for Men & Women - 42159

4.6 (183)

1K+ bought in past month

$239.99

$53.54 Import Charges & **FREE Shipping** to Republic of Korea Details

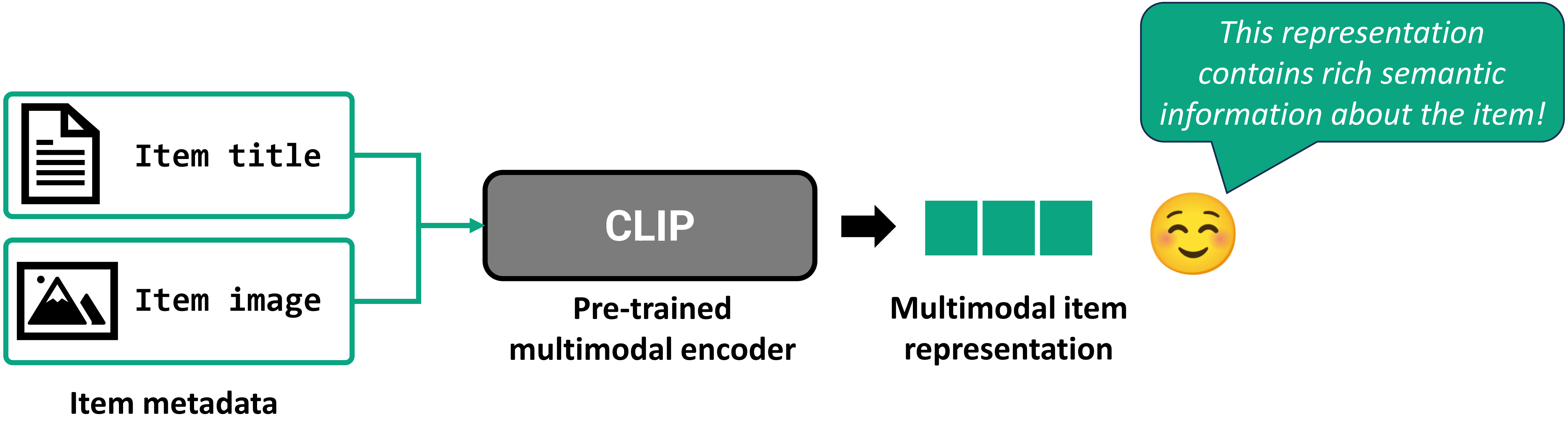
- Build a Yamaha model motorcycle – Enjoy a rewarding project for adults, building a replica version of the most powerful Yamaha motorcycle in the 'hyper naked' category
- Explore the AR app – Download the LEGO AR app to experience your model in a different way
- Features inspired by life – This build and display model incorporates many features from the full-size motorcycle, including LEGO Technic versions of the 4-cylinder engine and the 3-speed transmission
- Gearbox detailing – Gearbox include a shift drum, shift fork, gear shift ring and ratchet drum
- Display your passion – This advanced building set for adults makes a gift idea for fans of motorcycle collectibles. The motorcycle model comes with a display stand
- Measurements – This LEGO Technic building set measures over 9.5 in. (25 cm) high, 17.5 in. (44 cm) long and 5.5 in. (15 cm) wide
- Digital building instructions – The LEGO Builder app features a digital version of the building instructions included with this set

Item description
(textual information)

An example of item metadata from the Amazon website

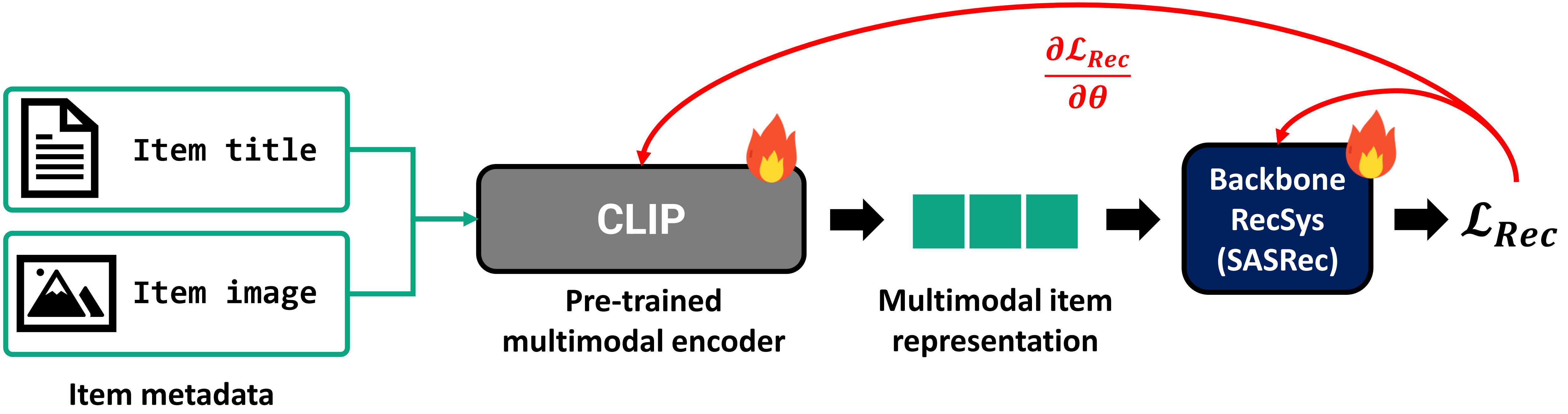
Research background (cont.)

- **[Approach]** One widely used approach is (1) obtaining multimodal representations from multimodal item metadata using a pre-trained multimodal encoder (e.g., CLIP), and (2) using them as item representations.



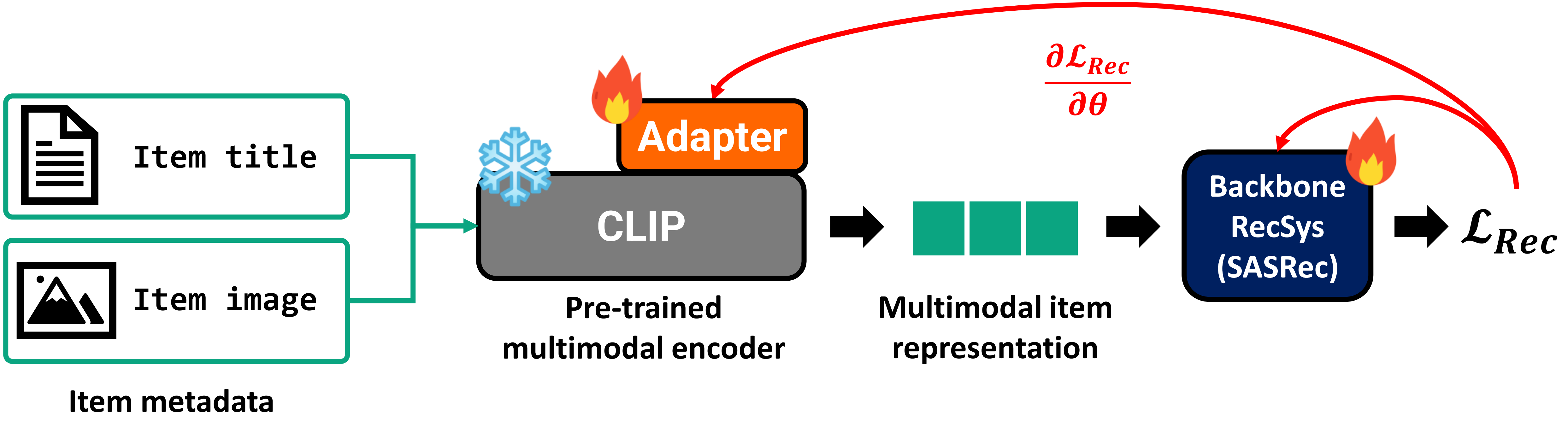
Research background (cont.)

- **[Adaptation to recommendation]** Since common pre-trained multimodal encoders are not specifically trained for recommendation tasks, they are often adapted through fine-tuning.



Research background (cont.)

- **[Scalable adaptation]** Given that pre-trained multimodal encoders often contain a large number of parameters, Parameter-Efficient Fine-Tuning (**PEFT**) is widely used for this adaptation.



Research background (cont.)

- **[Scalable adaptation]** Given that pre-trained multimodal encoders often contain a large number of parameters, **P**arameter-**E**fficient **F**ine-**T**uning (**PEFT**) is widely used for this adaptation.

CLIP

*has **billion**-scale parameters*

Adapter

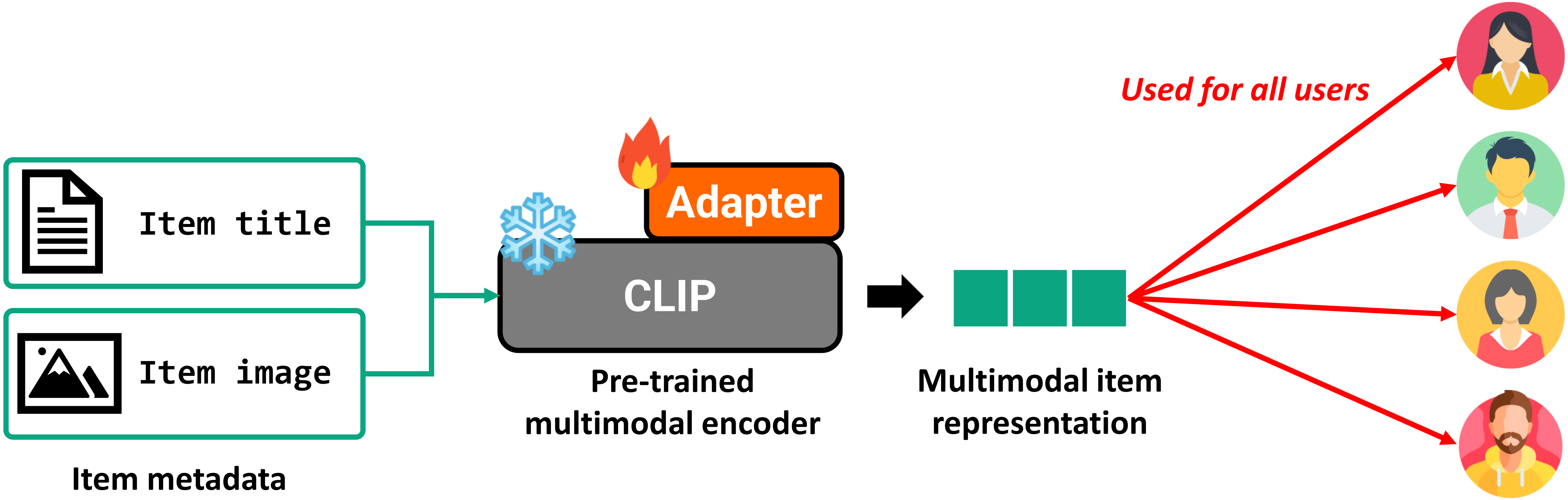
*has **million**-scale parameters*

# of trainable parameters are only ***1%***.

More efficient training

Key research question

- **[Existing work]** They typically employ a single PEFT module (e.g., adapter) across all users.



Key research question (cont.)

- **[Challenge]** Users with different interests focus on different aspects of each item, thus a single shared PEFT module may not effectively capture these diverse item aspects.

Character-figure collectors focus on **colors** and **appearances**.

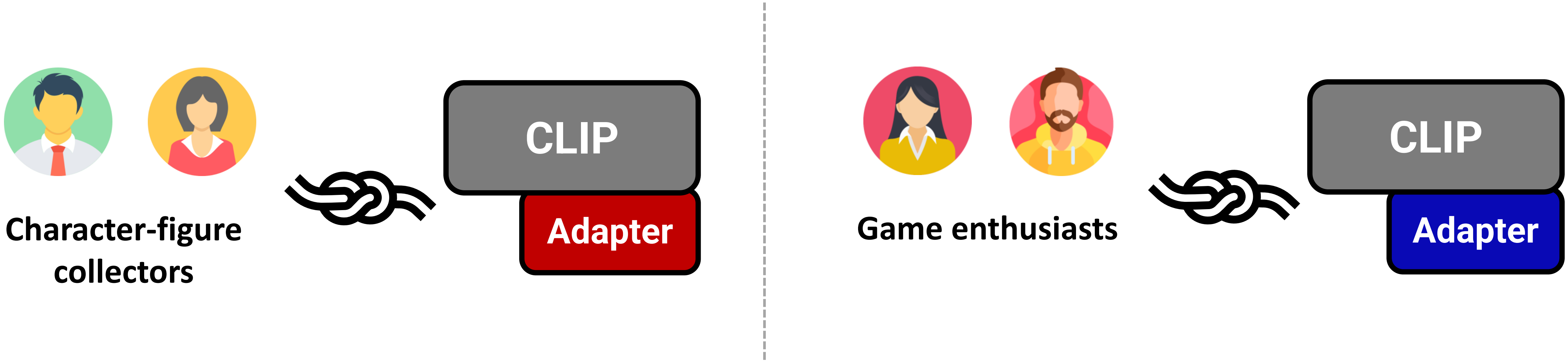
Game enthusiasts focus on **genres** and **platforms**.

It is hard for me to capture all these aspects…

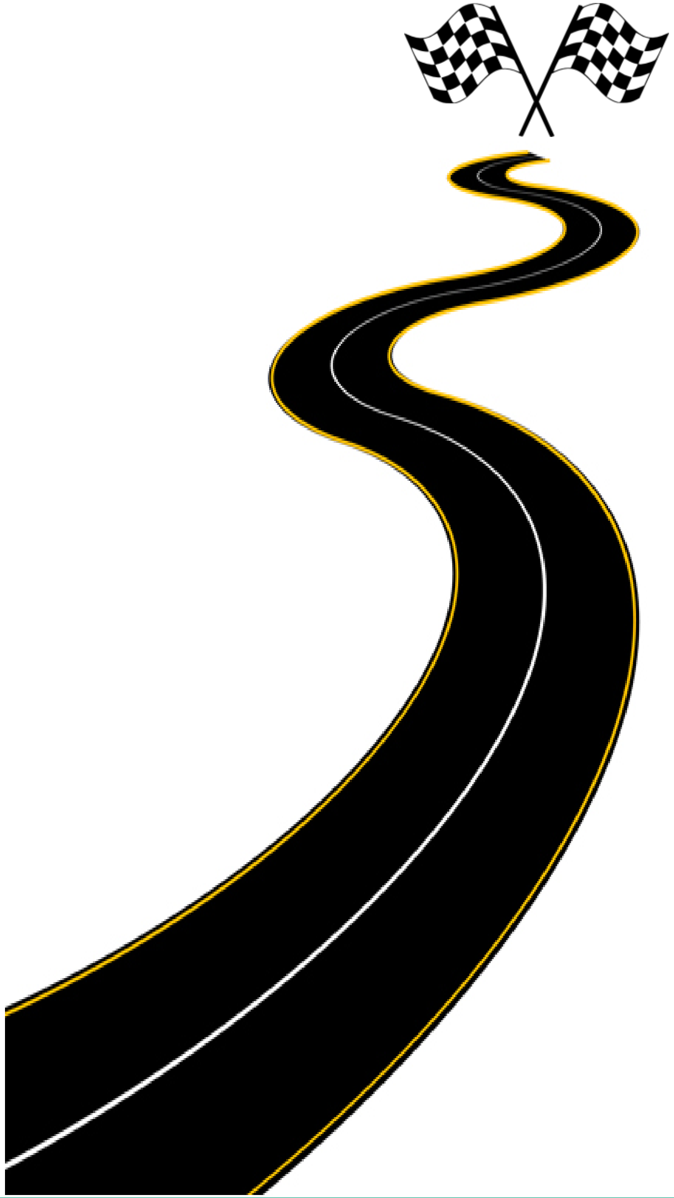
Adapter

Key idea of the proposed method

- **[High-level idea]** To better capture these diverse aspects, we use separate PEFT modules for different user-interest groups, allowing each module to specialize in the preferences of its corresponding group.

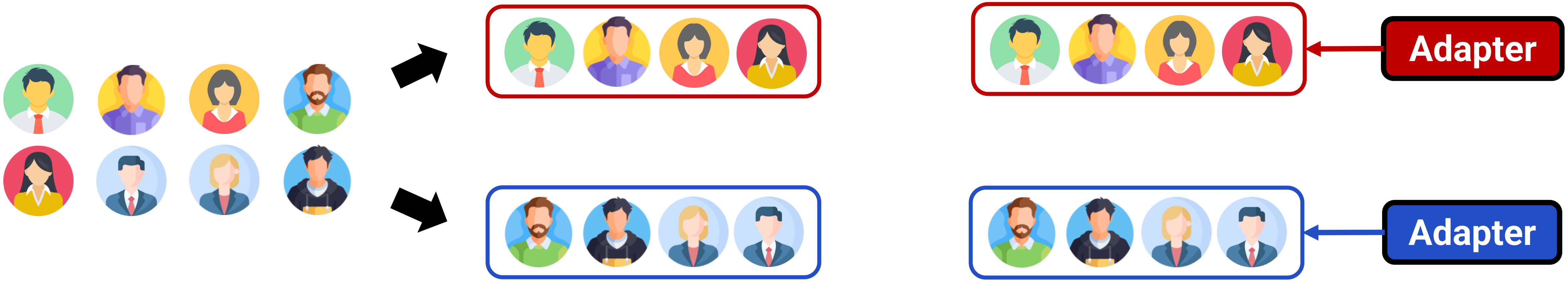


Contents (cont.)



Overview of the proposed method: PerPEFT

- **[Proposed method]** We propose **PerPEFT** (Personalized Parameter-Efficient Fine-Tuning), a personalized PEFT method for multimodal recommendation.
- **[Overview]** PerPEFT consists of two key steps: (1) identifying groups of users with similar interests and (2) training distinct PEFT modules for different user groups.

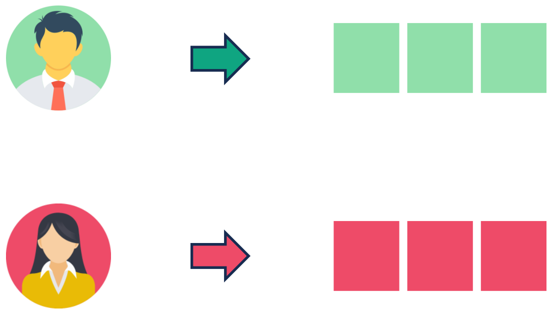


Step 1: User grouping

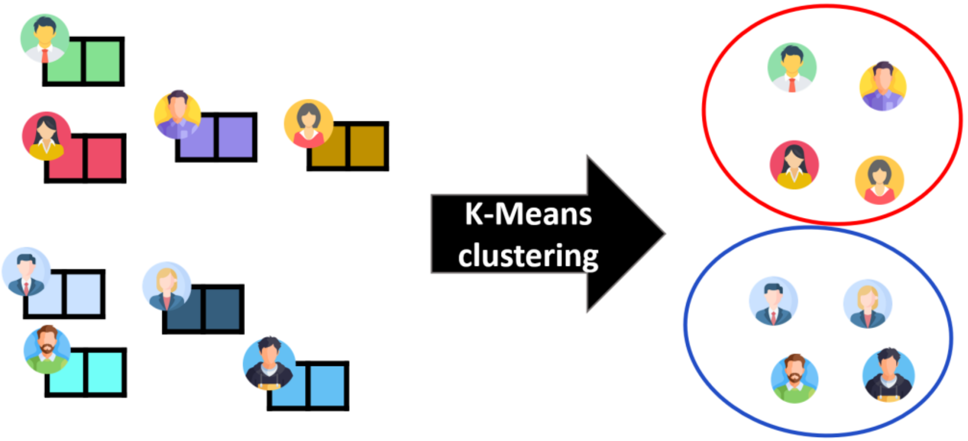
Step 2: Group-wise PEFT

Step 1: User grouping

- **[Key goal]** We aim to group users such that users within the same group share similar interests.
  - **[Step 1.1]** Learn user representations from users' item interaction histories.
  - **[Step 1.2]** Obtain user groups by clustering the obtained user representations.



Obtaining user representations



K-Means clustering

Step 1: User grouping (cont.)

- **[Step 1.1]** Learn user representations from users' item interaction histories.
  - **[Setup]** Consider a user's item interaction history.

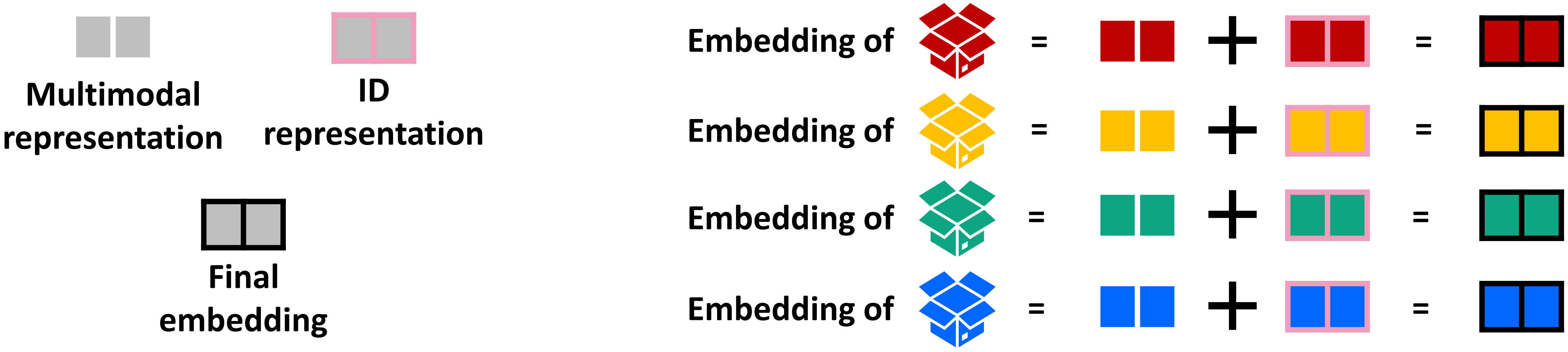
A user's item interaction history, with each item's title and image (item metadata)

Step 1: User grouping (cont.)

- **[Step 1.1]** Learn user representations from users' item interaction histories.
  - **[Item encoding]** We first encode multimodal item metadata using a pre-trained multimodal encoder with a single PEFT module shared across all users.

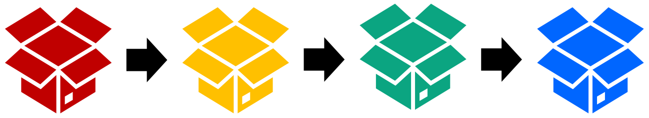
Step 1: User grouping (cont.)

- **[Step 1.1]** Learn user representations from users' item interaction histories.
  - **[Item encoding]** We first encode multimodal item metadata using a pre-trained multimodal encoder with a single PEFT module shared across all users.

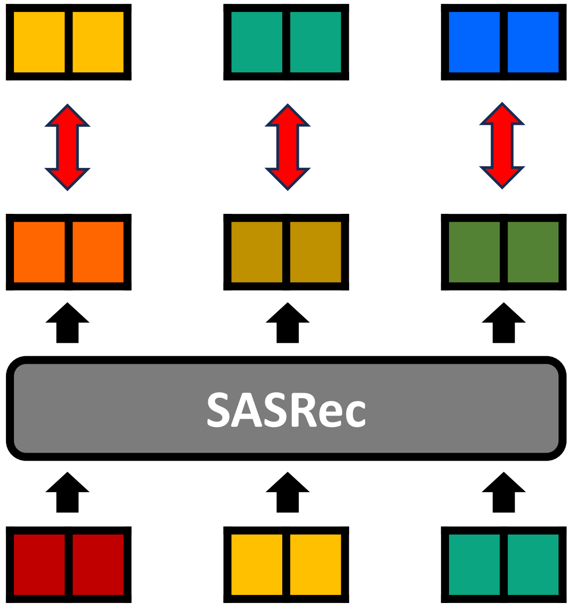


Step 1: User grouping (cont.)

- **[Step 1.1]** Learn user representations from users' item interaction histories.
  - **[Representation model training]** Then, we perform sequential recommendation training by feeding the obtained item embeddings into a sequential recommender system, such as SASRec.



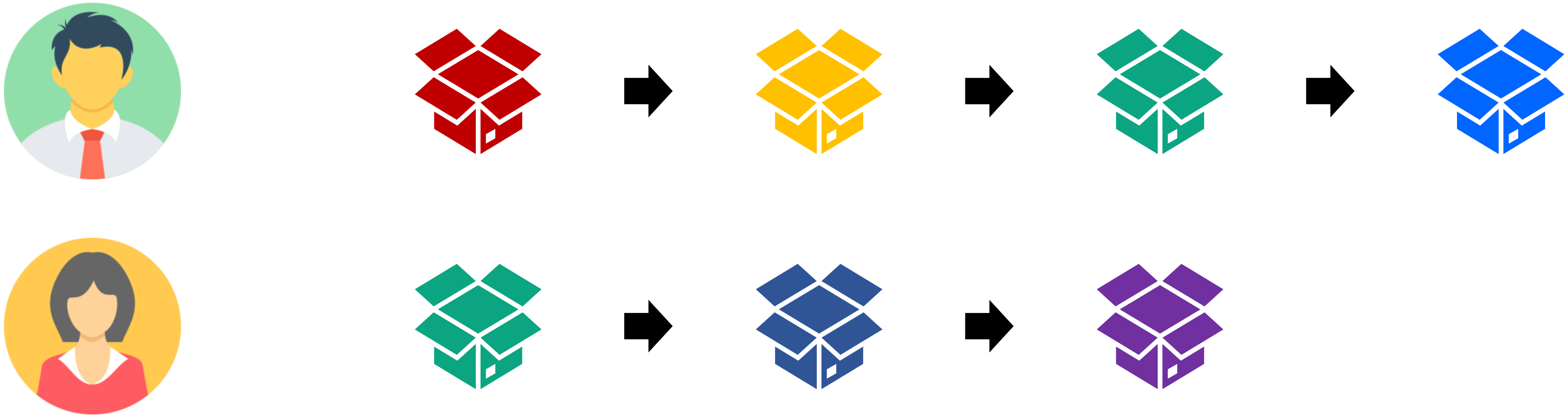
Purchase history



Parameters are updated to make these embeddings similar

Step 1: User grouping (cont.)

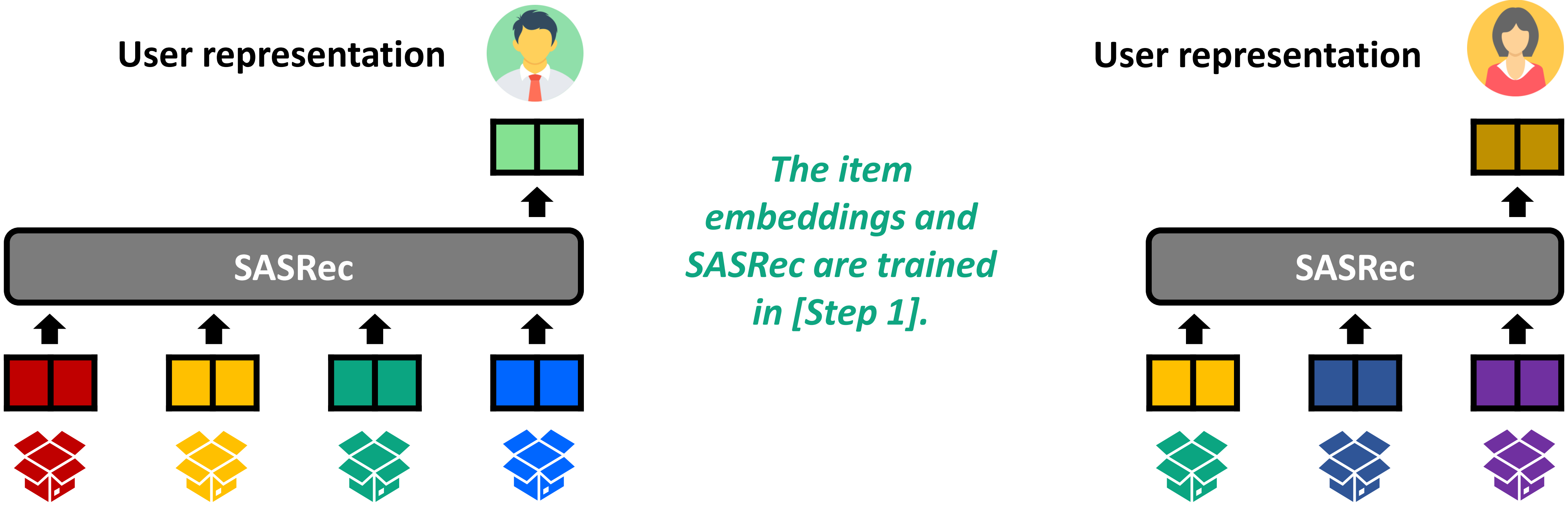
- **[Step 1.1]** Learn user representations from users' item interaction histories.
  - **[Obtaining user representation]** After training, we obtain user representations from the trained recommender system and item embeddings.



Users' item interaction histories

Step 1: User grouping (cont.)

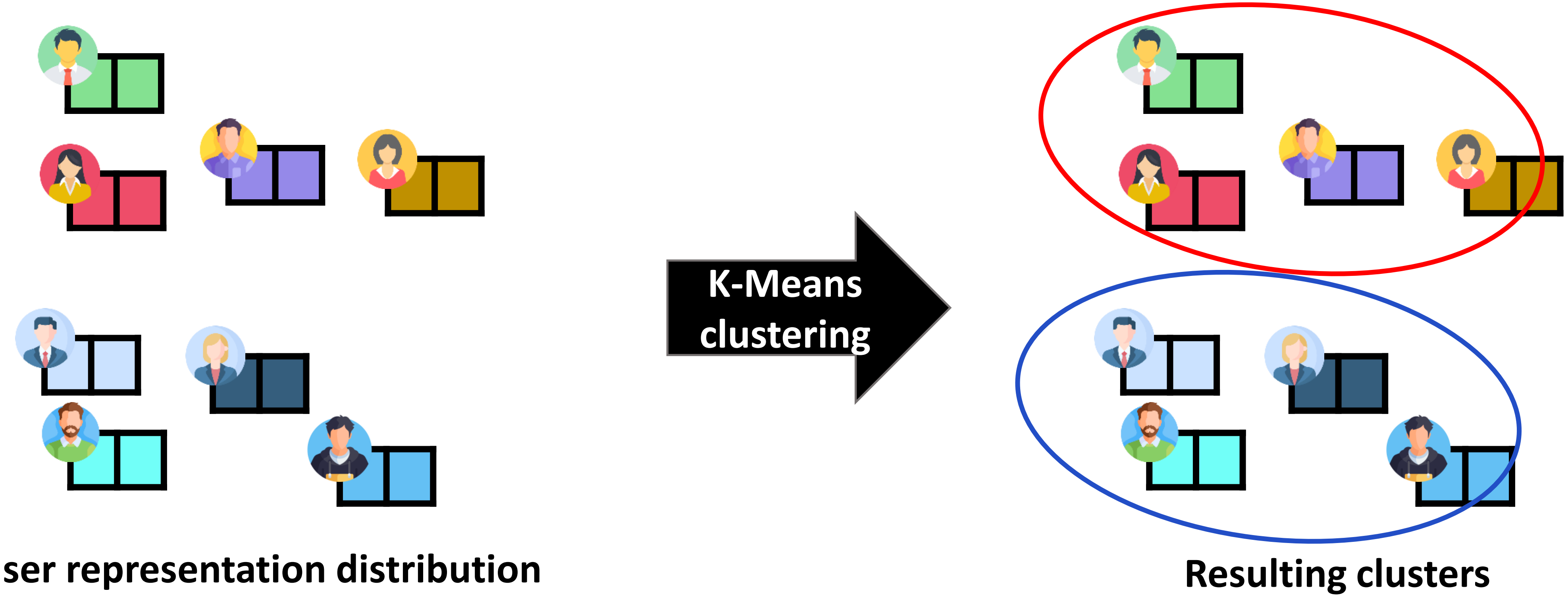
- **[Step 1.1]** Learn user representations from users' item interaction histories.
  - **[Obtaining user representation]** After training, we obtain user representations from the trained recommender system and item embeddings.



The item embeddings and SASRec are trained in [Step 1].

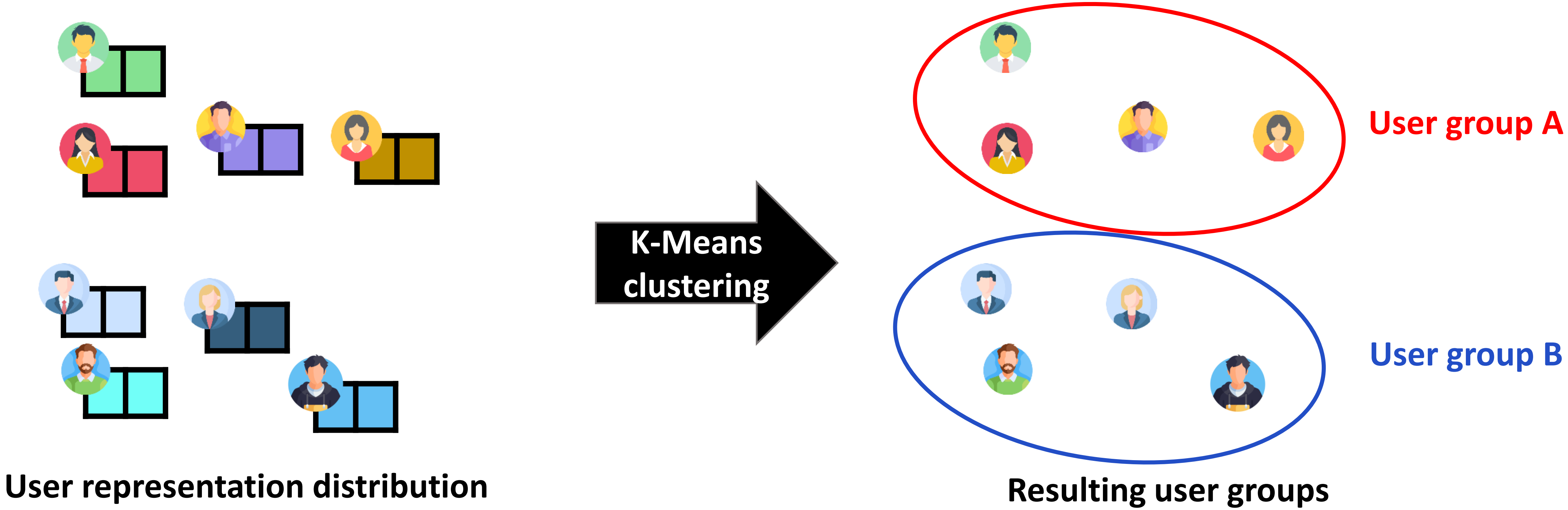
Step 1: User grouping (cont.)

- **[Step 1.2]** Obtain user groups by clustering the obtained user representations.
  - **[Clustering]** We use the K-Means clustering algorithm on the obtained user representations.



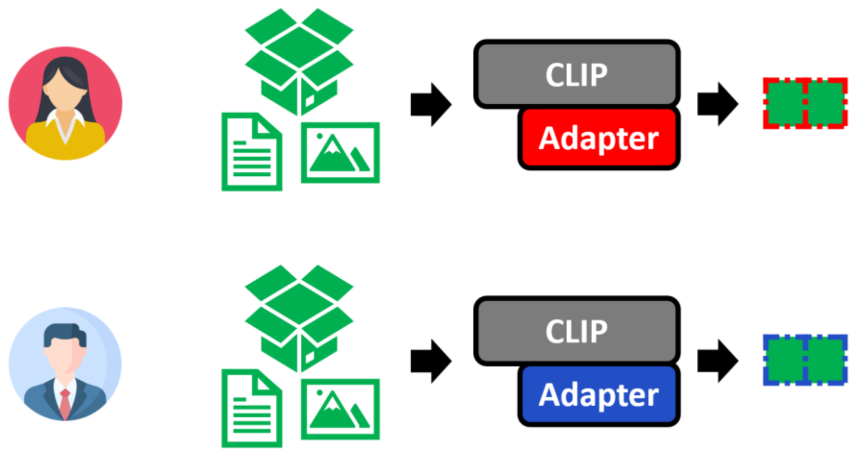
Step 1: User grouping (cont.)

- **[Step 1.2]** Obtain user groups by clustering the obtained user representations.
  - **[Clustering]** We use the K-Means clustering algorithm on the obtained user representations.

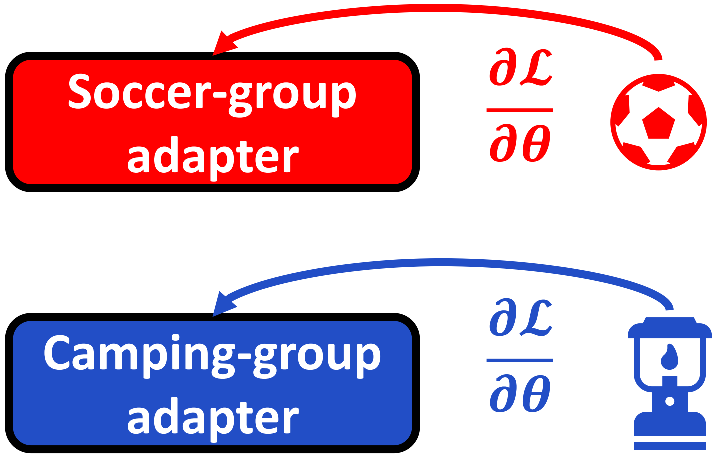


Step 2: Group-wise PEFT

- **[Key goal]** We train a distinct PEFT module for each user group, enabling each module to specialize in its corresponding user group.
  - **[Key idea 1]** Using a group-wise PEFT module.
  - **[Key idea 2]** Using a group-wise negative sampling strategy for training.



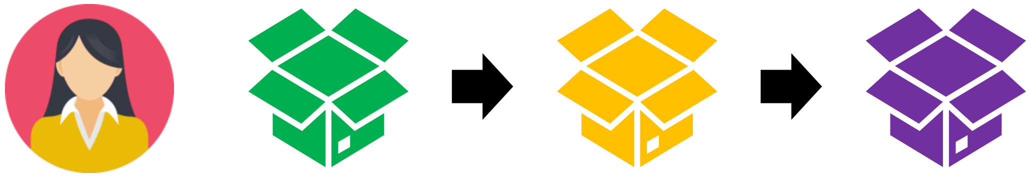
Group-wise PEFT modules



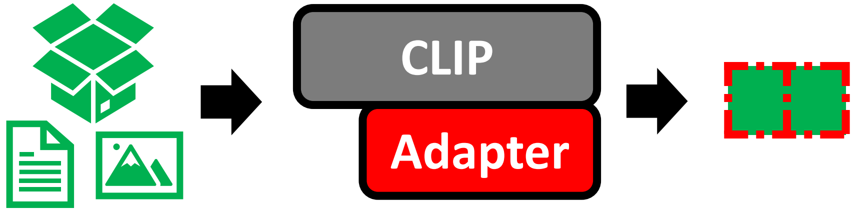
Group-wise negative sampling

Step 2: Group-wise PEFT (cont.)

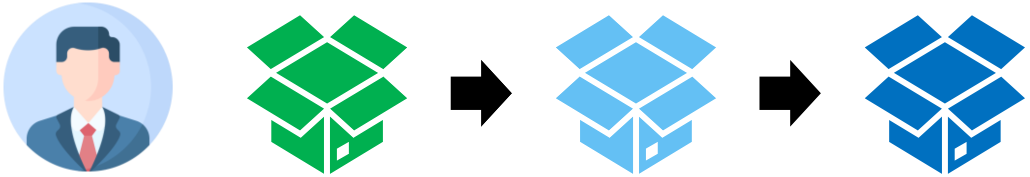
- **[Key idea 1]** Using a group-wise PEFT module.
  - **[Item encoding]** Item embeddings used for a particular user are derived from the PEFT module specialized to the corresponding user group.



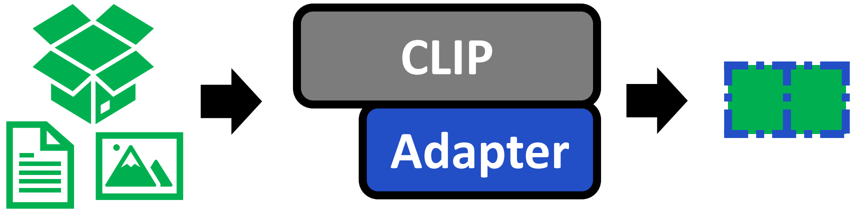
Item interaction history of a Group-A user



Obtaining multimodal item embeddings for Group-B users



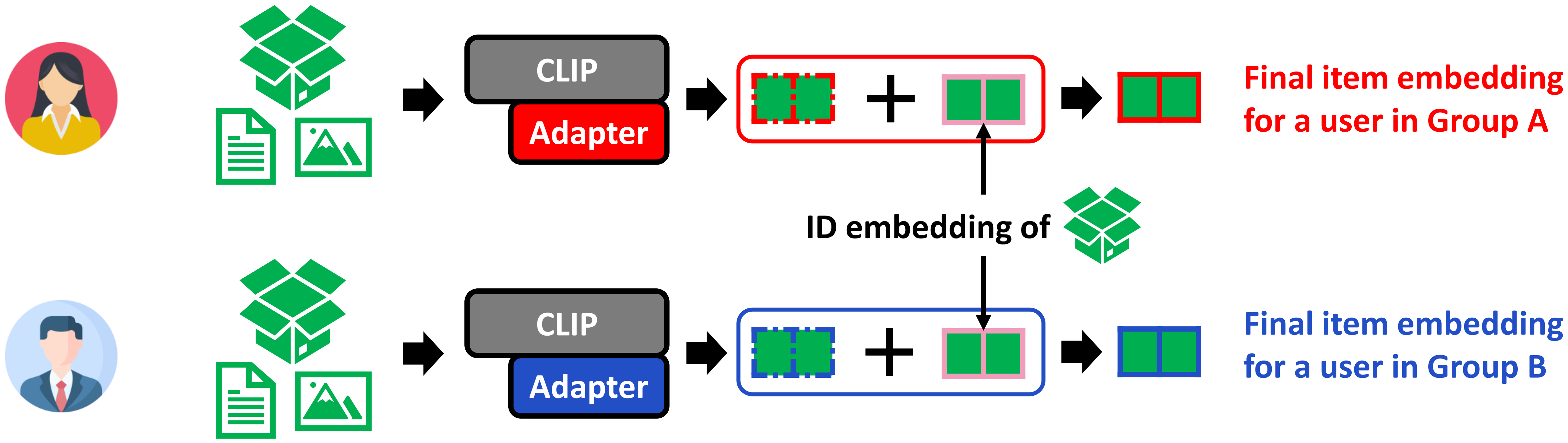
Item interaction history of a Group-B user



Obtaining multimodal item embeddings for Group-B users

Step 2: Group-wise PEFT (cont.)

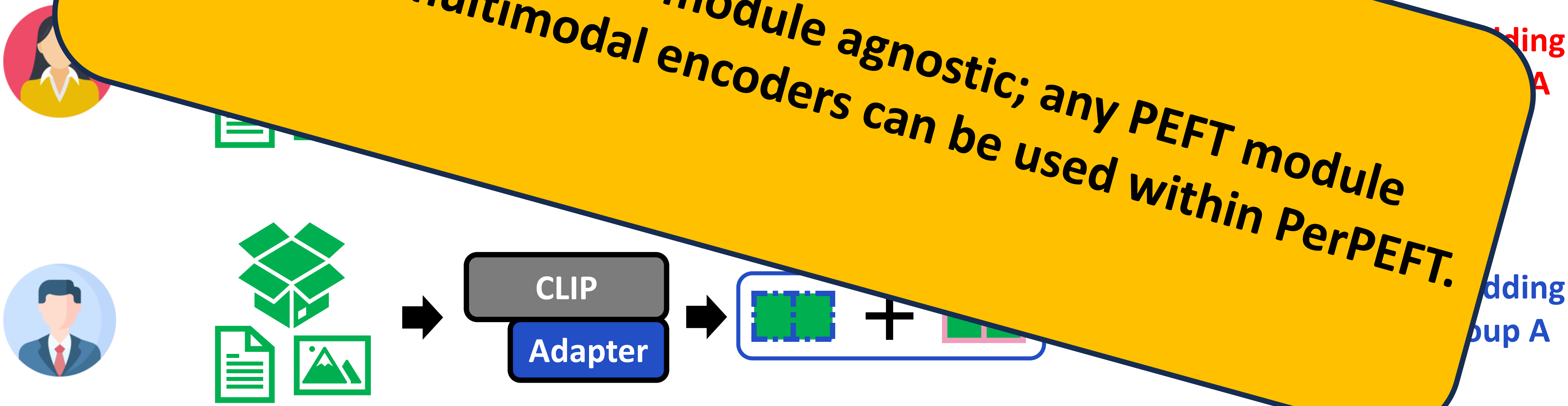
- **[Key idea 1]** Using a group-wise PEFT module.
  - **[Item encoding]** Item embeddings used for a particular user are derived from the PEFT module specialized to the corresponding user group.



Step 2: Group-wise PEFT (cont.)

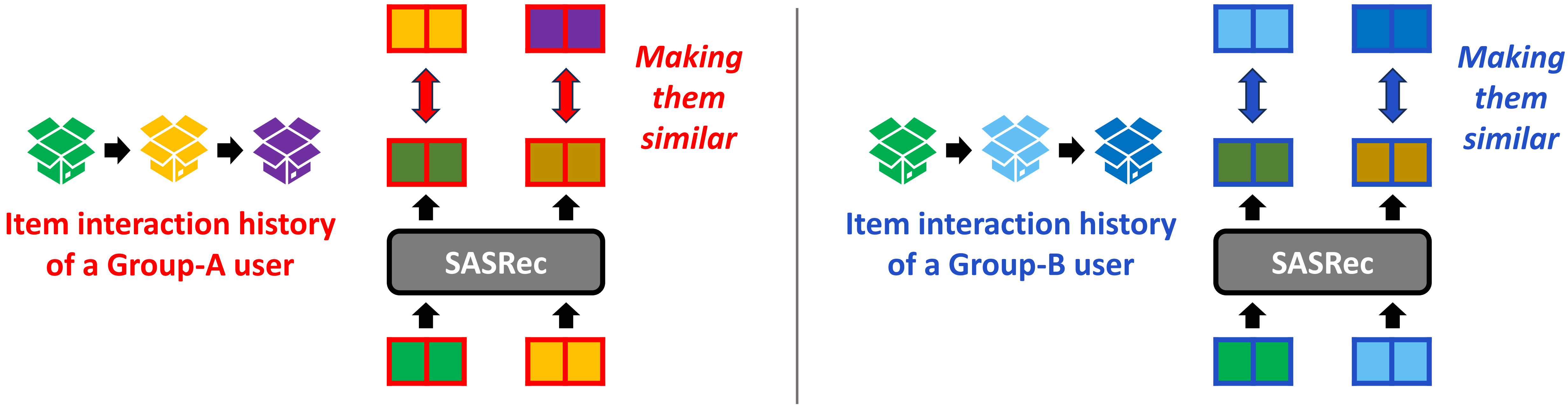
- **[Key i**... module.
  - ...ticular user are derived from the PEFT module

Note. PerPEFT is PEFT-module agnostic; any PEFT module applicable to multimodal encoders can be used within PerPEFT.



Step 2: Group-wise PEFT (cont.)

- **[Key idea 1]** Using a group-wise PEFT module.
  - **[Positive samples]** For each position in a user's interaction sequence, the positive sample is the next item in the sequence, while item representations vary across user groups.

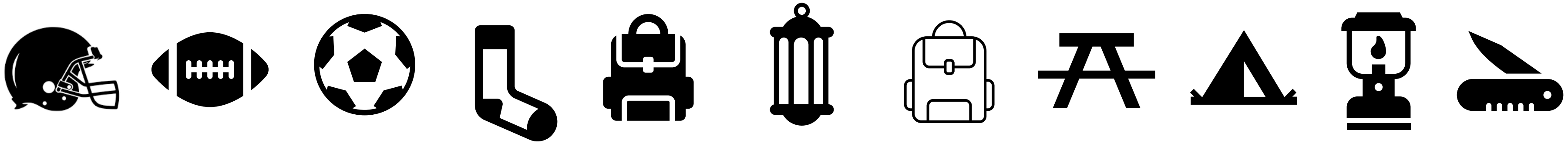


Step 2: Group-wise PEFT (cont.)

- **[Key idea 2]** Using a group-wise negative sampling strategy for training.
  - **[Negative samples]** For users in each group, negative samples are drawn from items observed as positive in the interaction histories of users in that group.

Positive samples for users interested in football and soccer

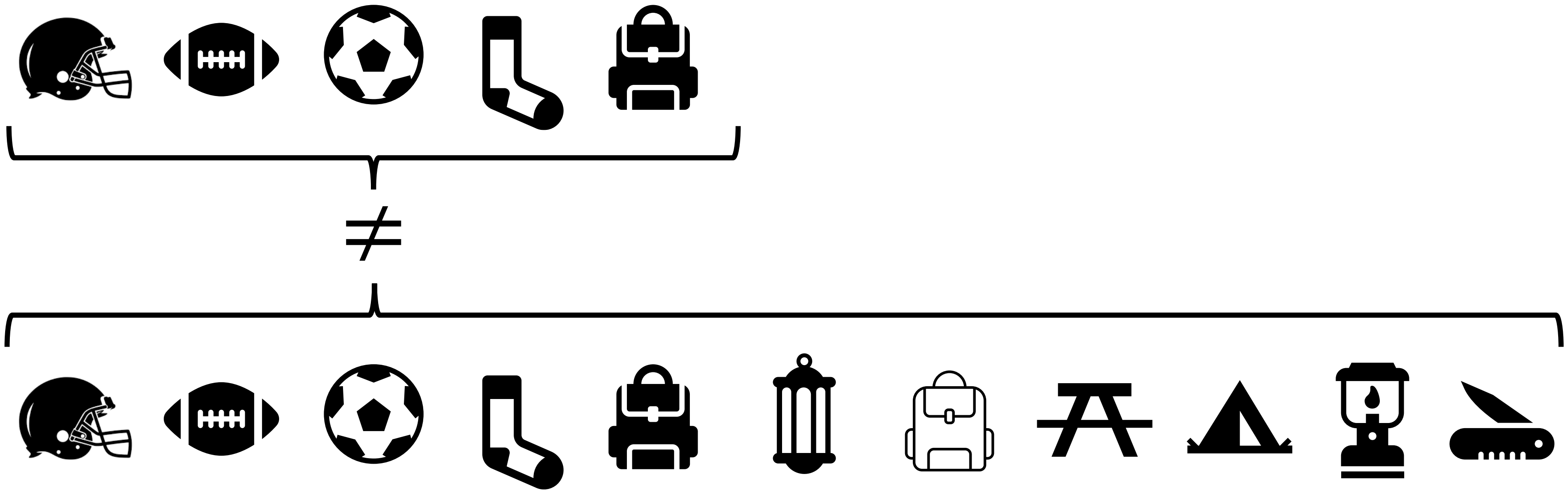
Items in the full item space



Step 2: Group-wise PEFT (cont.)

- **[Key idea 2]** Using a group-wise negative sampling strategy for training.
  - **[Negative samples]** For users in each group, negative samples are drawn from items observed as positive in the interaction histories of users in that group.

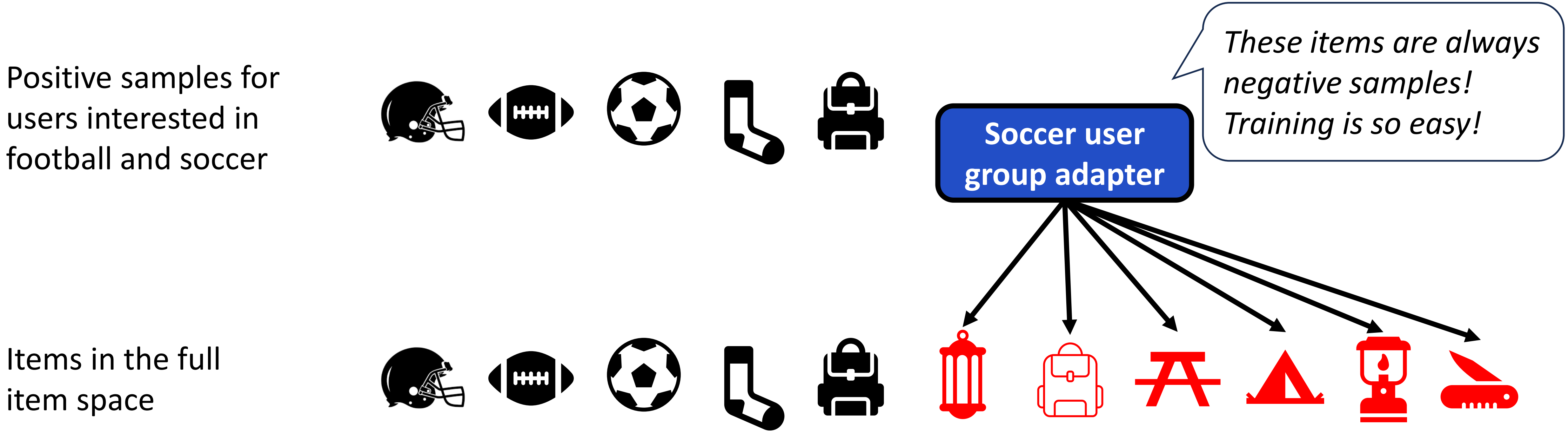
Positive samples for users interested in football and soccer



Items in the full item space

Step 2: Group-wise PEFT (cont.)

- **[Key idea 2]** Using a group-wise negative sampling strategy for training.
  - **[Negative samples]** For users in each group, negative samples are drawn from items observed as positive in the interaction histories of users in that group.



Step 2: Group-wise PEFT (cont.)

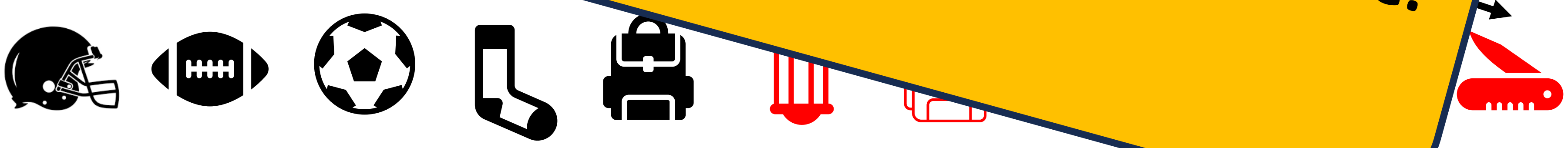
- **[Key i**… tive sampling strategy for training.
  - … tive samples are drawn from items observed as

… se items are always … ples!

Positiv…
users in…
football and so…

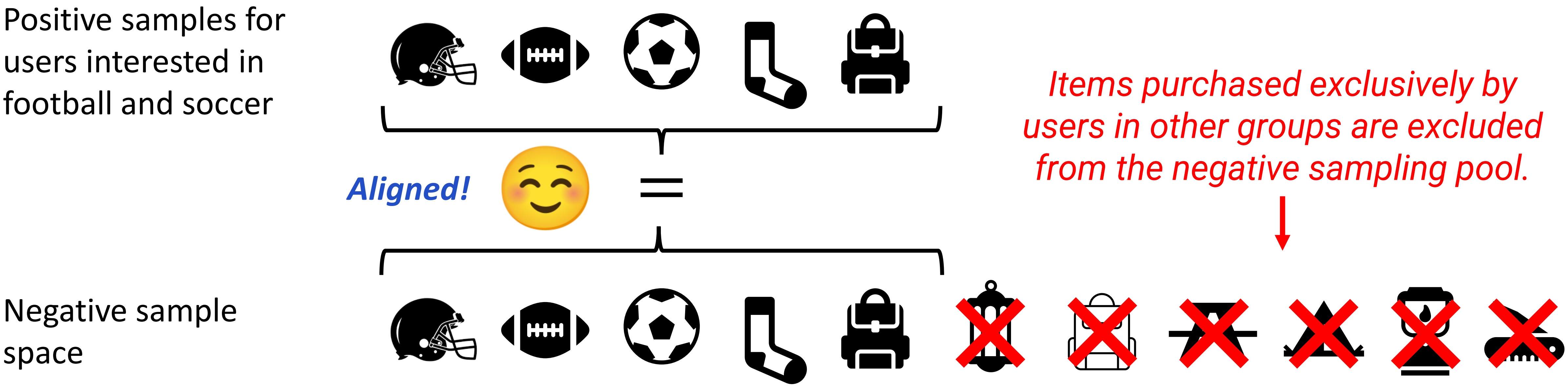
Adapters are not effectively trained to capture fine-grained item aspects, resulting in poor recommendation performance.

Items in the full item space

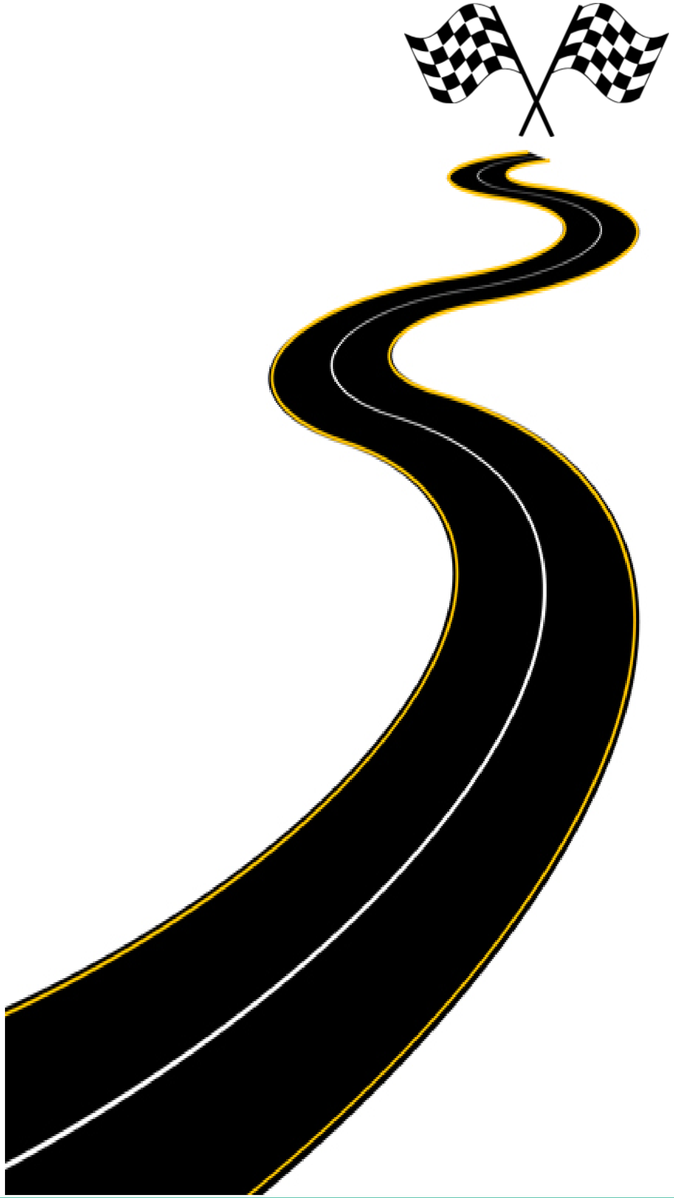


Step 2: Group-wise PEFT (cont.)

- **[Key idea 2]** Using a group-wise negative sampling strategy for training.
  - **[Negative samples]** For users in each group, negative samples are drawn from items observed as positive in the interaction histories of users in that group.



Contents



Key research questions

- **[RQ1. General performance]** How effective is PerPEFT for multimodal recommendation compared with baseline methods?
- **[RQ2. Efficiency analysis]** How efficient is PerPEFT in terms of training time and model size?
- **[RQ3. Design objective]** Does PerPEFT achieve its intended design objective?
- **[RQ4. Ablation study]** Are all key components of PerPEFT necessary for achieving strong performance?

RQ1. General recommendation performance

- **[Result 1.1]** PerPEFT outperforms both (1) a single PEFT module shared across all users (Global PEFT), and (2) existing personalized strategies (User-level and Group-level).

PEFT	Methods	Sports & Outdoors				Toys & Games				Beauty & Personal Care				Arts, Crafts & Sewing			
		H@20	H@30	N@20	N@30	H@20	H@30	N@20	N@30	H@20	H@30	N@20	N@30	H@20	H@30	N@20	N@30
N/A	w/o MM	2.79	3.76	1.08	1.29	2.27	3.04	0.91	1.05	2.60	3.47	1.12	1.35	3.91	5.20	1.48	1.76
	Frozen MM	4.11	5.38	1.71	1.99	3.24	4.34	1.31	1.55	3.60	4.66	1.45	1.67	4.87	6.42	1.89	2.21
LoRA	Global PEFT	4.64	5.97	1.94	2.22	4.19	5.67	1.71	2.02	4.15	5.52	1.72	2.02	6.01	7.81	2.42	2.81
	User-level 1	4.48	5.96	1.86	2.18	4.27	5.67	1.74	2.04	4.19	5.52	1.68	1.95	6.03	7.92	2.44	2.83
	User-level 2	4.51	5.86	1.90	2.19	4.13	5.46	1.67	1.91	4.03	5.43	1.66	1.94	6.05	7.91	2.43	2.82
	Group-level 1	4.44	5.89	1.87	2.18	4.24	5.64	1.76	2.07	4.13	5.45	1.65	1.94	5.98	7.87	2.44	2.82
	Group-level 2	4.56	5.94	1.92	2.21	4.26	5.64	1.75	2.06	4.14	5.47	1.70	1.98	5.96	7.94	2.45	2.88
	PerPEFT (Ours)	4.82 (.04)	6.24 (.03)	2.05 (.02)	2.36 (.02)	4.79 (.03)	6.15 (.04)	1.97 (.03)	2.25 (.02)	4.27 (.03)	5.58 (.03)	1.74 (.01)	2.01 (.01)	6.62 (.11)	8.64 (.06)	2.76 (.04)	3.17 (.02)
$(IA)^3$	Global PEFT	4.55	5.97	1.90	2.21	4.58	5.96	1.89	2.19	4.25	5.61	1.73	2.03	6.24	8.16	2.59	3.00
	User-level 1	4.45	5.92	1.89	2.20	4.41	5.86	1.82	2.13	4.25	5.55	1.76	2.02	6.09	8.06	2.49	2.92
	User-level 2	4.39	5.74	1.81	2.11	3.95	5.38	1.69	1.95	3.64	5.20	1.62	1.89	5.68	7.64	2.32	2.71
	Group-level 1	4.51	5.95	1.91	2.21	4.37	5.75	1.80	2.11	4.16	5.47	1.71	1.99	6.07	7.98	2.49	2.92
	Group-level 2	4.58	5.99	1.93	2.23	4.36	5.87	1.87	2.18	4.14	5.48	1.71	2.00	6.17	8.02	2.52	2.94
	PerPEFT (Ours)	4.82 (.03)	6.30 (.02)	2.01 (.02)	2.32 (.01)	4.81 (.03)	6.20 (.04)	2.01 (.02)	2.30 (.02)	4.33 (.04)	5.70 (.05)	1.78 (.04)	2.06 (.03)	6.56 (.02)	8.50 (.03)	2.67 (.01)	3.11 (.02)
IISAN	Global PEFT	4.77	6.13	2.01	2.30	4.50	5.97	1.89	2.15	4.31	5.71	1.74	2.05	6.02	7.89	2.46	2.85
	User-level 1	4.41	5.84	1.83	2.16	3.75	4.97	1.54	1.84	3.81	5.06	1.52	1.78	5.65	7.63	2.29	2.71
	User-level 2	4.39	5.79	1.82	2.15	4.23	5.54	1.75	2.04	4.03	5.42	1.67	1.95	5.86	7.82	2.41	2.75
	Group-level 1	4.43	5.72	1.84	2.12	4.27	5.49	1.75	2.03	4.17	5.39	1.68	1.95	5.74	7.46	2.32	2.67
	Group-level 2	4.38	5.67	1.84	2.12	4.05	5.38	1.68	1.97	4.14	5.34	1.72	1.97	5.64	7.38	2.27	2.64
	PerPEFT (Ours)	4.81 (.02)	6.22 (.04)	2.00 (.03)	2.31 (.02)	4.70 (.06)	6.10 (.04)	1.95 (.04)	2.24 (.04)	4.30 (.02)	5.69 (.04)	1.77 (.02)	2.07 (.03)	6.37 (.05)	8.17 (.03)	2.59 (.04)	2.98 (.03)

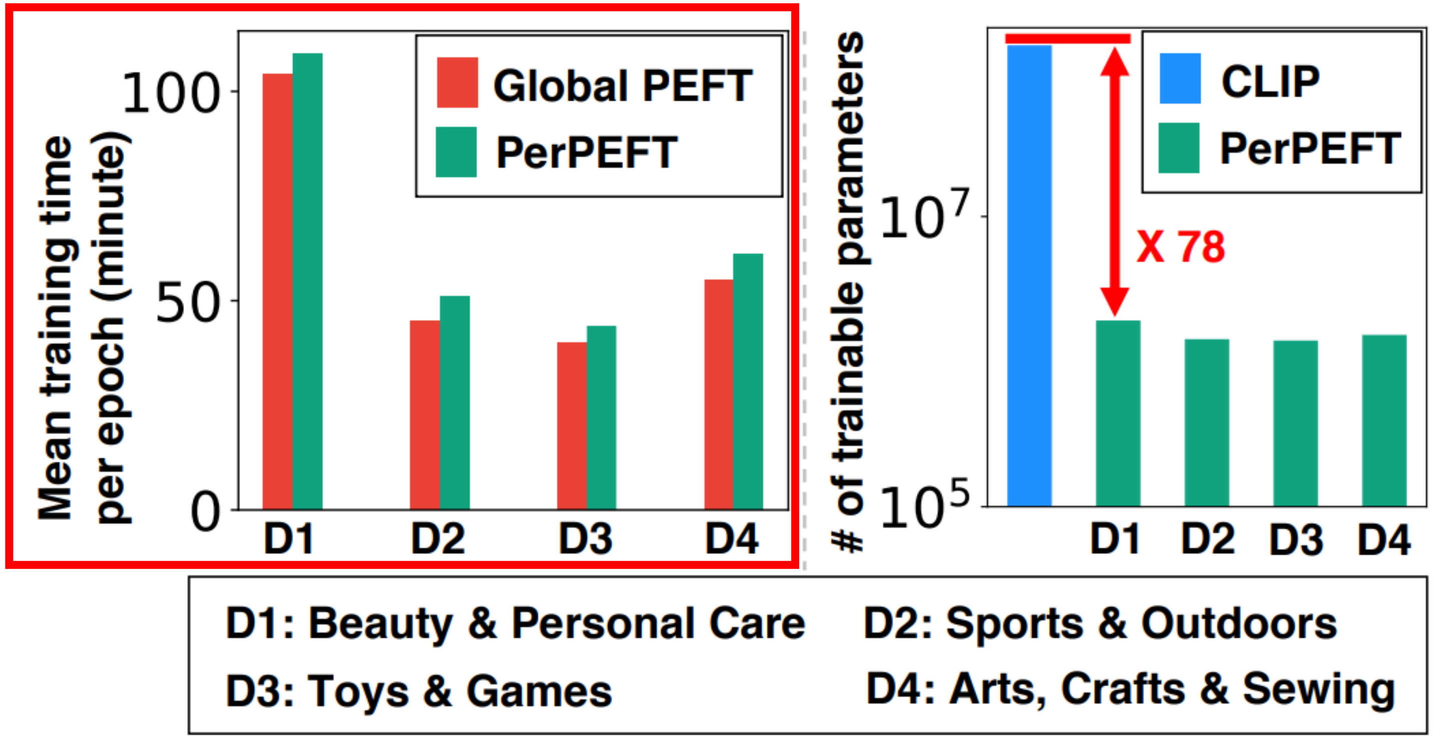
RQ1. General recommendation performance (cont.)

- **[Result 1.2]** The superiority of PerPEFT maintains across diverse PEFT module types, including LoRA, (IA)3, and IISAN.

PEFT	Methods	Sports & Outdoors H@20	Sports & Outdoors H@30	Sports & Outdoors N@20	Sports & Outdoors N@30	Toys & Games H@20	Toys & Games H@30	Toys & Games N@20	Toys & Games N@30	Beauty & Personal Care H@20	Beauty & Personal Care H@30	Beauty & Personal Care N@20	Beauty & Personal Care N@30	Arts, Crafts & Sewing H@20	Arts, Crafts & Sewing H@30	Arts, Crafts & Sewing N@20	Arts, Crafts & Sewing N@30
N/A	w/o MM	2.79	3.76	1.08	1.29	2.27	3.04	0.91	1.05	2.60	3.47	1.12	1.35	3.91	5.20	1.48	1.76
	Frozen MM	4.11	5.38	1.71	1.99	3.24	4.34	1.31	1.55	3.60	4.66	1.45	1.67	4.87	6.42	1.89	2.21
LoRA	Global PEFT	4.64	5.97	1.94	2.22	4.19	5.67	1.71	2.02	4.15	5.52	1.72	2.02	6.01	7.81	2.42	2.81
	User-level 1	4.48	5.96	1.86	2.18	4.27	5.67	1.74	2.04	4.19	5.52	1.68	1.95	6.03	7.92	2.44	2.83
	User-level 2	4.51	5.86	1.90	2.19	4.13	5.46	1.67	1.91	4.03	5.43	1.66	1.94	6.05	7.91	2.43	2.82
	Group-level 1	4.44	5.89	1.87	2.18	4.24	5.64	1.76	2.07	4.13	5.45	1.65	1.94	5.98	7.87	2.44	2.82
	Group-level 2	4.56	5.94	1.92	2.21	4.26	5.64	1.75	2.06	4.14	5.47	1.70	1.98	5.96	7.94	2.45	2.88
	PerPEFT (Ours)	4.82 (.04)	6.24 (.03)	2.05 (.02)	2.36 (.02)	4.79 (.03)	6.15 (.04)	1.97 (.03)	2.25 (.02)	4.27 (.03)	5.58 (.03)	1.74 (.01)	2.01 (.01)	6.62 (.11)	8.64 (.06)	2.76 (.04)	3.17 (.02)
$(IA)^3$	Global PEFT	4.55	5.97	1.90	2.21	4.58	5.96	1.89	2.19	4.25	5.61	1.73	2.03	6.24	8.16	2.59	3.00
	User-level 1	4.45	5.92	1.89	2.20	4.41	5.86	1.82	2.13	4.25	5.55	1.76	2.02	6.09	8.06	2.49	2.92
	User-level 2	4.39	5.74	1.81	2.11	3.95	5.38	1.69	1.95	3.64	5.20	1.62	1.89	5.68	7.64	2.32	2.71
	Group-level 1	4.51	5.95	1.91	2.21	4.37	5.75	1.80	2.11	4.16	5.47	1.71	1.99	6.07	7.98	2.49	2.92
	Group-level 2	4.58	5.99	1.93	2.23	4.36	5.87	1.87	2.18	4.14	5.48	1.71	2.00	6.17	8.02	2.52	2.94
	PerPEFT (Ours)	4.82 (.03)	6.30 (.02)	2.01 (.02)	2.32 (.01)	4.81 (.03)	6.20 (.04)	2.01 (.02)	2.30 (.02)	4.33 (.04)	5.70 (.05)	1.78 (.04)	2.06 (.03)	6.56 (.02)	8.50 (.03)	2.67 (.01)	3.11 (.02)
IISAN	Global PEFT	4.77	6.13	2.01	2.30	4.50	5.97	1.89	2.15	4.31	5.71	1.74	2.05	6.02	7.89	2.46	2.85
	User-level 1	4.41	5.84	1.83	2.16	3.75	4.97	1.54	1.84	3.81	5.06	1.52	1.78	5.65	7.63	2.29	2.71
	User-level 2	4.39	5.79	1.82	2.15	4.23	5.54	1.75	2.04	4.03	5.42	1.67	1.95	5.86	7.82	2.41	2.75
	Group-level 1	4.43	5.72	1.84	2.12	4.27	5.49	1.75	2.03	4.17	5.39	1.68	1.95	5.74	7.46	2.32	2.67
	Group-level 2	4.38	5.67	1.84	2.12	4.05	5.38	1.68	1.97	4.14	5.34	1.72	1.97	5.64	7.38	2.27	2.64
	PerPEFT (Ours)	4.81 (.02)	6.22 (.04)	2.00 (.03)	2.31 (.02)	4.70 (.06)	6.10 (.04)	1.95 (.04)	2.24 (.04)	4.30 (.02)	5.69 (.04)	1.77 (.02)	2.07 (.03)	6.37 (.05)	8.17 (.03)	2.59 (.04)	2.98 (.03)

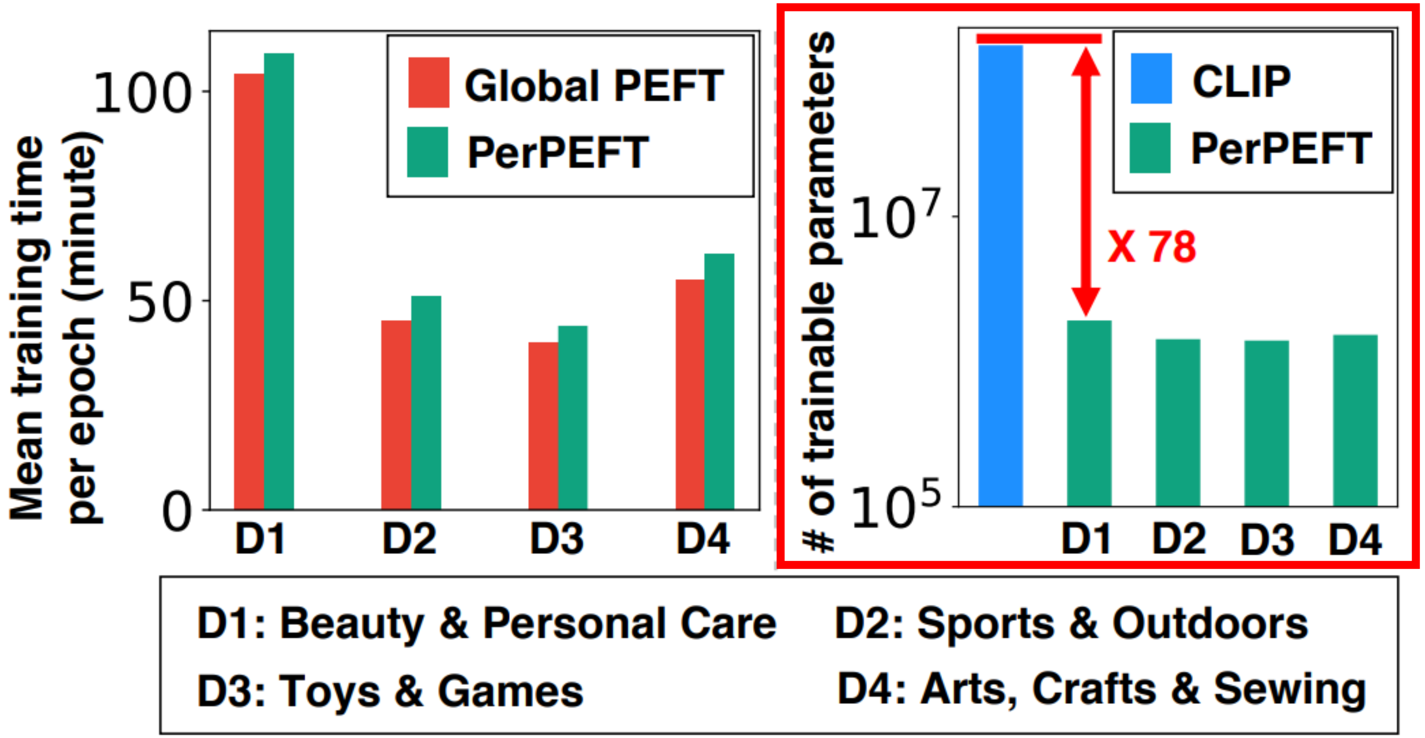
RQ2. Efficiency analysis

- **[Result 2.1]** The training time of PerPEFT is comparable to that of Global PEFT, which uses a single PEFT module shared across ass users.



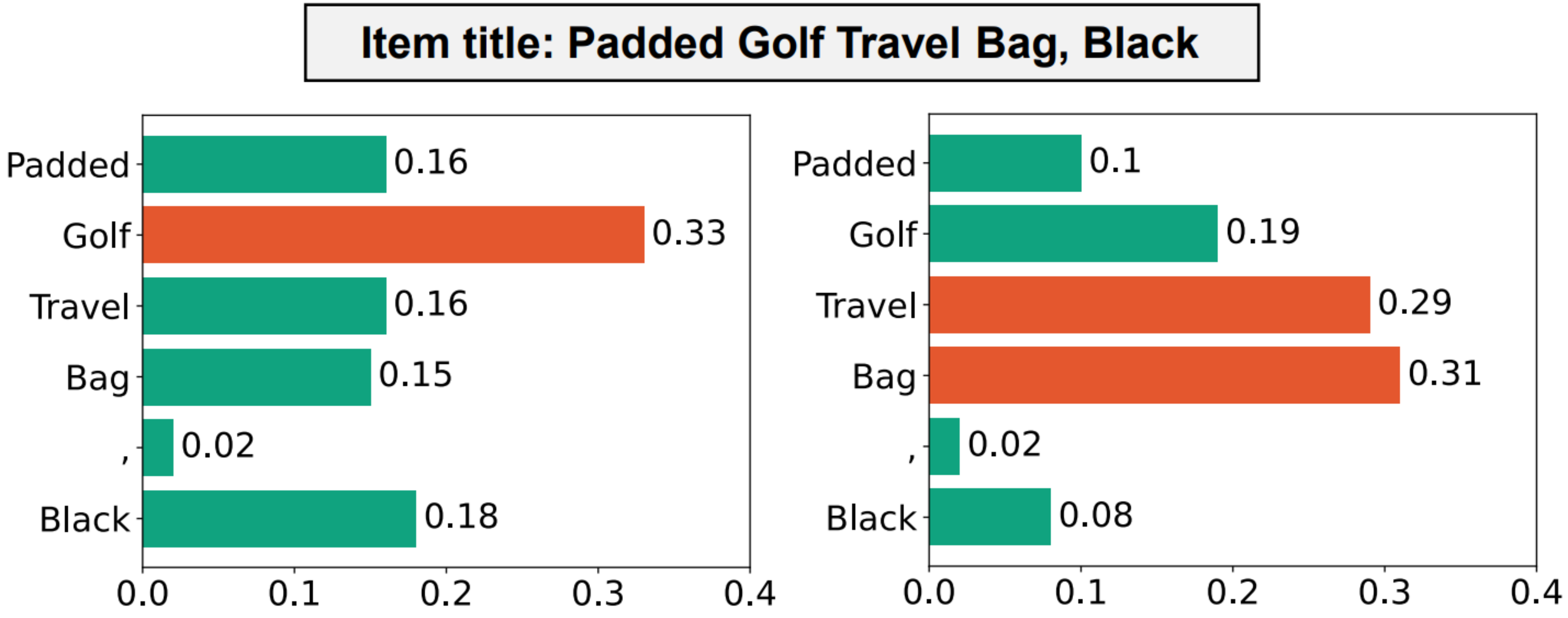
RQ2. Efficiency analysis (cont.)

- **[Result 2.2]** The additional trainable parameters introduced by PerPEFT amount to only 1.3% of the multimodal encoder's (i.e., CLIP's) parameters.



RQ3. Design goal achievement

- **[Result 3]** PerPEFT guides a multimodal encoder to focus on item aspects aligned with each user group's key interests.



Attention weights of CLIP with the golf-group-specialized PEFT

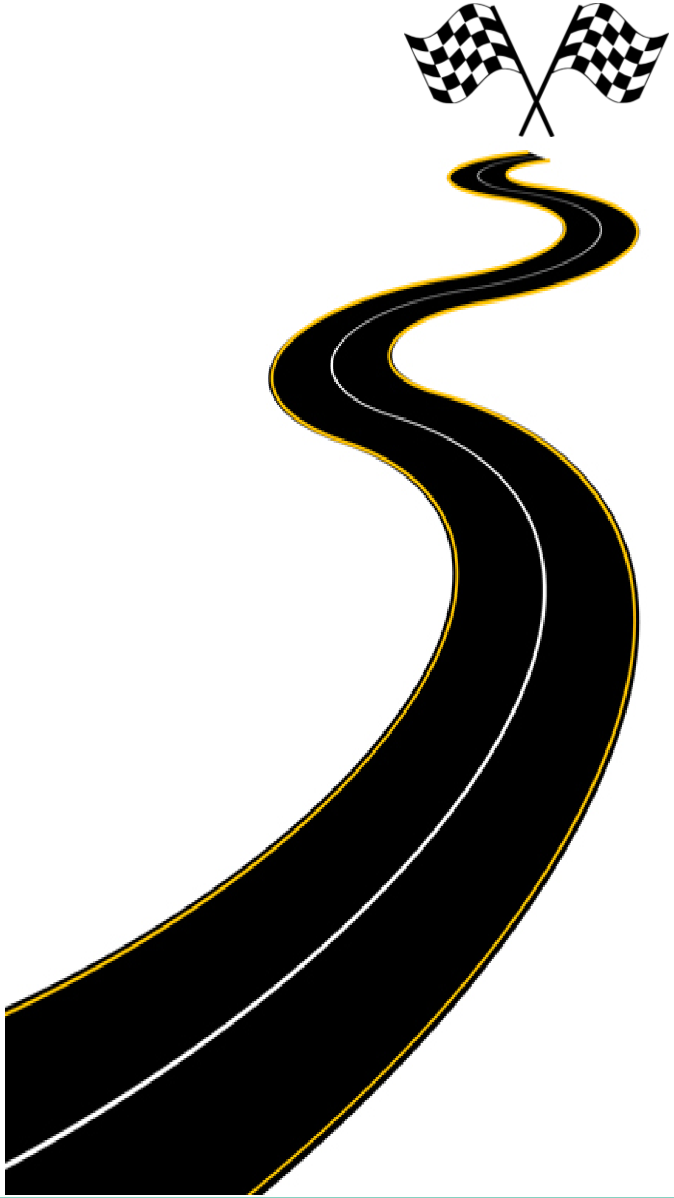
Attention weights of CLIP with the camping-group-specialized PEFT

RQ4. Ablation study

- **[Result 4]** All key components of PerPEFT are necessary for achieving strong performance.
  - **[V1. w/o textual modality]** A variant that does not use textual information.
  - **[V2. w/o visual modality]** A variant that does not use visual information.
  - **[V3. group-specific negatives]** A variant that does not use group-specific negative sampling method.
  - **[V4. Large Global PEFT]** A variant that uses a large-sized single PEFT instead for personalized PEFT.
  - **[V5. Random grouping]** A variant that uses a random grouping instead of interest-based grouping.

Methods	Sports & Outdoors				Toys & Games				Beauty & Personal Care				Arts, Crafts & Sewing			
	H@20	H@30	N@20	N@30	H@20	H@30	N@20	N@30	H@20	H@30	N@20	N@30	H@20	H@30	N@20	N@30
w/o textual modality	3.87	5.15	1.62	1.86	3.85	4.92	1.58	1.81	4.00	5.29	1.61	1.88	4.95	6.55	2.01	2.35
w/o visual modality	4.80	6.16	1.99	2.28	4.48	5.97	1.81	2.12	4.19	5.43	1.68	1.97	6.68	8.57	2.74	3.15
w/o group-specific negatives	4.31	5.73	1.84	2.14	4.30	5.67	1.77	2.06	3.91	5.16	1.61	1.87	6.19	8.32	2.50	2.96
Large Global PEFT	4.51	5.95	1.86	2.17	4.00	5.42	1.67	1.94	4.01	5.32	1.65	1.89	5.94	7.72	2.41	2.78
Random grouping	4.09	5.46	1.71	2.00	3.65	4.84	1.48	1.73	3.82	4.95	1.50	1.74	5.43	7.18	2.17	2.54
PerPEFT (Ours)	4.82 (.04)	6.24 (.03)	2.05 (.02)	2.36 (.02)	4.79 (.03)	6.15 (.04)	1.97 (.03)	2.25 (.02)	4.27 (.03)	5.58 (.03)	1.74 (.01)	2.01 (.01)	6.62 (.11)	8.64 (.06)	2.76 (.04)	3.17 (.02)

Contents



Summary

- **[Contribution 1. New method]** We propose PerPEFT, a personalized PEFT method for multimodal recommendation that applies different PEFT modules to different user groups.
- **[Contribution 2. Training method]** To further improve performance, we introduce a specialized training strategy in which each user group draws negative samples only from items appears in the training data of users in the corresponding group.
- **[Contribution 3. Strong performance]** PerPEFT achieves stronger recommendation performance than baseline methods, and its effectiveness remains consistent across diverse PEFT module types.

Thank you for listening!

Sunwoo Kim

Hyunjin Hwang

Kijung Shin