PAIR: Preference-Aware Image Retrieval

Anonymous ACL submission

Abstract

Despite rapid advances in large language model (LLM) embedding backbones, text-to-image retrieval systems remain fundamentally user-ignorant, returning identical results regardless of the searcher's underlying intent. Bridging this gap requires preference generalization: the ability to surface images semantically aligned with a user's broader tastes rather than strictly matching explicitly stated keywords. To achieve this, we introduce Preference-Aware Image Retrieval (PAIR), a novel task that directly conditions cross-modal retrieval on declarative user profiles. To support this, we release PAIR-Data and PAIR-Bench, a training corpus and evaluation framework featuring implicit target sets designed to explicitly test generalization beyond stated preferences. Methodologically, we propose a parameter-efficient architecture utilizing a small preference encoder to compress user profiles into a soft prefix, steering a frozen MLLM retrieval backbone. Evaluated across two domains and three backbones, PAIR significantly outperforms both text-concatenation and zero-shot LLM-rewriting baselines. Furthermore, our approach substantially enhances retrieval diversity and mitigates popularity bias, ensuring a broader range of items is surfaced across different users. A complementary user study further validates these quantitative gains. Our code and datasets are publicly available **here**.

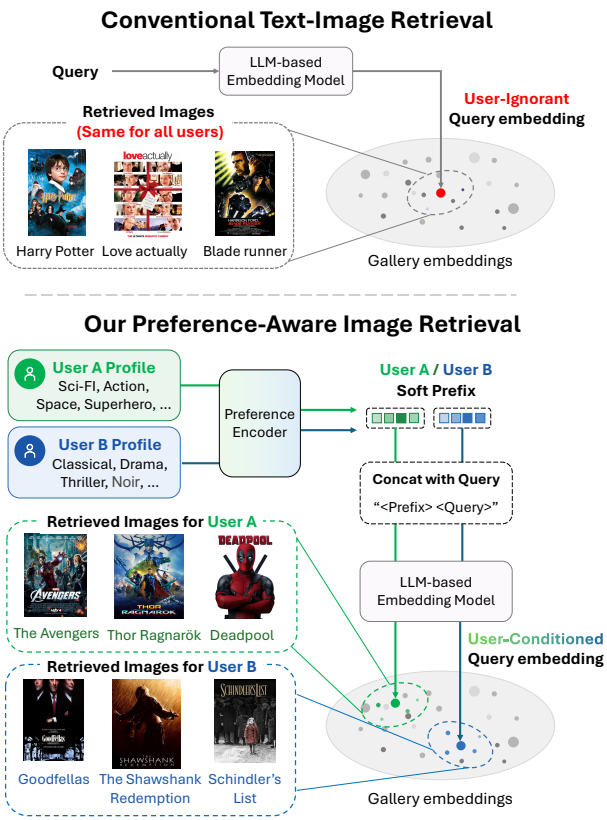


Figure 1: **Preference-Aware Image Retrieval (PAIR)** conditions retrieval on each user's stated preferences. **(Top)** A conventional text-image retrieval returns the same images to every user for a generic query. **(Bottom)** We formulate PAIR and our method compresses the profile through a preference encoder into a soft prefix that conditions the same frozen backbone, steering image retrieval per user.

1 Introduction

Text-to-image retrieval (Radford et al., 2021; Jia et al., 2021) has advanced rapidly with LLM-based embedding models (Jiang et al., 2024b; Zhang et al., 2025; Li et al., 2026) that unify text and image encoding in a single multimodal backbone. Despite this progress, conditioning retrieval on the user behind the query remains underexplored: a generic text query returns the exact same ranked list of images to every user, remaining fundamentally user-ignorant even when the underlying search intent varies contextually.

Bridging this gap requires preference generalization—the ability to surface images semantically aligned with a user's broader tastes rather than strictly matching explicitly stated search keywords. Instead of relying on superficial string alignment or direct keyword overlap, a preference-aware system must perform implicit reasoning to link shared themes, tones, and visual aesthetics across dis-

parate concepts. To operationalize this capability, we argue that retrieval systems must be conditioned on a declarative user profile: a structured, user-editable natural-language description detailing a user's persistent preferences, such as favored genres, previously enjoyed titles, and preferred directors or developers.

In practice, existing cross-modal retrieval models largely overlook such profiling context. Because these backbones are pretrained predominantly on standard caption-image alignment, they lack the intrinsic signaling required for implicit user-taste inference. Furthermore, establishing a rigorous evaluation protocol for this capability is non-trivial, as conventional retrieval benchmarks omit per-user conditioning altogether. To address these limitations, we formalize the task of Preference-Aware Image Retrieval (PAIR). Alongside this task formulation, we introduce PAIR-Data, a curated training corpus, and PAIR-Bench, a robust evaluation benchmark. To explicitly isolate and measure preference generalization, PAIR-Bench constructs per-user implicit target sets composed entirely of compatible items that are strictly filtered to exclude any entity present in the user's explicit profile text.

Methodologically, fine-tuning a massive multimodal backbone to absorb user preference awareness is computationally prohibitive and risks catastrophic forgetting (Kirkpatrick et al., 2017) of pretrained cross-modal alignment signals. We instead propose a parameter-efficient alternative: a lightweight preference encoder that compresses the variable-length textual profile into a compact sequence of continuous soft prefixes. Prepending the user's query at inference time, these learned tokens effectively steer a completely frozen MLLM retrieval backbone. This architectural decoupling ensures that gallery item embeddings can be cached statically, fully preserving the operational and deployment efficiency of the base model.

Evaluated across two domains rich in visual grounding and user preferences (Movies and Games), our experiments demonstrate that PAIR delivers substantial improvements in retrieval accuracy over standard text-concatenation and zero-shot LLM-rewriting baselines. Furthermore, our framework substantially enhances retrieval diversity and mitigates popularity bias, ensuring broader and more personalized catalog coverage across diverse user cohorts. These automated metrics are further validated by a blinded qualitative user study, confirming substantial gains of our approach under human judgment. In summary, our key contributions are:

- **Task formulation.** We formalize Preference-Aware Image Retrieval (PAIR) as a profile-conditioned cross-modal retrieval task requiring preference generalization beyond literal profile text.
- **PAIR-Data and PAIR-Bench.** We release a scalable training corpus and evaluation benchmark with collaborative-filtering-derived target sets that explicitly isolate implicit user tastes.
- **Parameter-Efficient Architecture.** We present a lightweight preference encoder that successfully steers frozen MLLM backbones via soft prefixes, maintaining high deployment efficiency while significantly boosting retrieval diversity and reducing popularity bias.

2 Related Work

Text-to-image retrieval and LLM-based embedding models. Text-to-image retrieval is typically framed as similarity search in a shared embedding space whose text and image towers are trained contrastively against each other (Radford et al., 2021; Jia et al., 2021; Jang et al., 2024). A more recent line replaces the text tower with a multimodal large language model fine-tuned contrastively as a dense embedding model (Jiang et al., 2024b; Zhang et al., 2025; Li et al., 2026), handling longer and more compositional queries. A complementary direction, composed image retrieval (Wu et al., 2021; Liu et al., 2021; Baldrati et al., 2023), conditions retrieval on a reference image plus a text modifier rather than text alone. Across both directions, the conditioning signal describes *what is wanted*, not *who is asking*, and the same query yields the same ranking regardless of user taste.

User modeling. Modeling users from their history or preferences has been studied broadly outside text-to-image retrieval. Recommender systems represent users as latent vectors learned from item-ID interaction histories (Koren et al., 2009; He et al., 2017; Kang and McAuley, 2018; Sun et al., 2019); a parallel line conditions large language models on

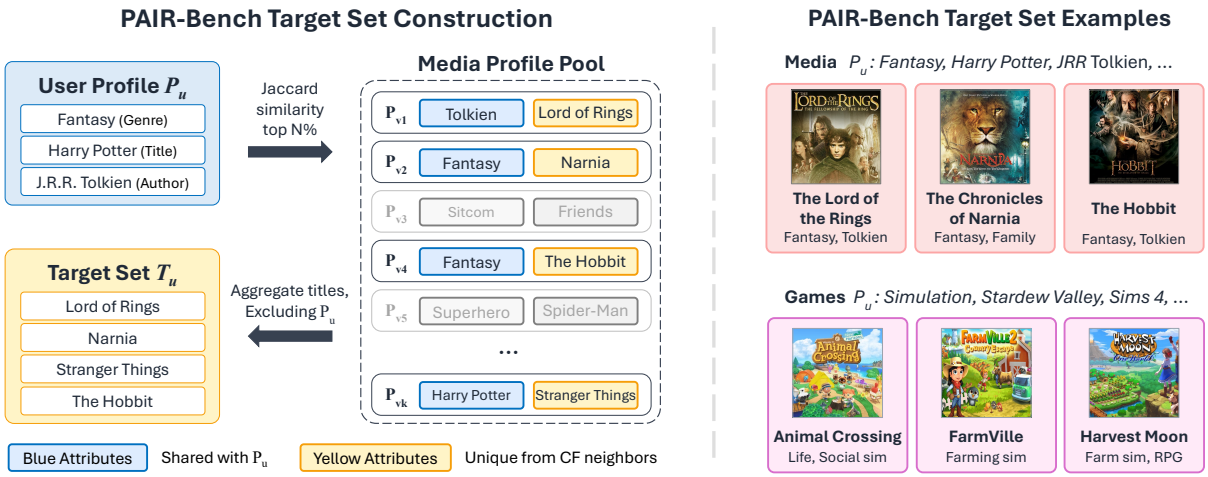


Figure 2: **PAIR-Bench target set construction and examples. (Left)** Shared attributes (blue) identify top CF neighbors via Jaccard similarity. Their unique attributes (yellow) form the implicit target set T_u, strictly excluding items in P_u. **(Right)** Sample targets in T_u for Media and Games. Top-K retrieval against T_u measures preference generalization.

textual user signals to personalize next-item recommendation, dialogue, or summarization (Ning et al., 2024; Doddapaneni et al., 2024; Salemi et al., 2024); and personalized vision-language models bind user-specific concepts to learnable tokens for recognition or VQA (Alaluf et al., 2024; Nguyen et al., 2024). While none of these settings conditions text-to-image retrieval on a persistent user profile, PAIR-Bench utilizes user-based collaborative filtering (Sarwar et al., 2001) to construct each test user's target set from the profiles of similar users, operationalizing the items a user would accept but has not explicitly named.

Parameter-efficient context compression. Our preference encoder builds on parameter-efficient methods that compress variable-length input into soft tokens for a frozen language model. Prefix-tuning (Lester et al., 2021) prepends learned soft tokens to condition a frozen model on a task, with related variants compressing longer contexts into a few reusable tokens (Mu et al., 2023; Ge et al., 2024). PromptEOL (Jiang et al., 2024a) and LLM2Vec (BehnamGhader et al., 2024) turn frozen decoder-only LLMs into text encoders by pooling a prompted input into a single compact vector. PAIR extends this paradigm to a new setting: a per-user textual profile compressed into a soft prefix that conditions a frozen text-to-image retrieval system.

3 PAIR Benchmark

3.1 Task Formulation

In a conventional, user-ignorant text-to-image retrieval setup, a system scores an image I from a gallery $\mathcal{G}$ based solely on a text query q. We formalize Preference-Aware Image Retrieval (PAIR) by introducing a conditioning variable: the declarative user profile $P_u = \{p_1, \ldots, p_n\}$. This profile consists of a structured set of explicit preferences (e.g., preferred genres, favorite titles, and favored directors or developers) typically gathered during initial user onboarding. Given a generic query q, the profile P_u, and an image gallery $\mathcal{G}$, the task requires computing a user-conditioned relevance score $s(I, q, P_u)$ for each image $I \in \mathcal{G}$ to retrieve the top-K visual assets. Crucially, a stated profile P_u captures only a sparse fraction of a user's true taste landscape. The core challenge of PAIR is preference generalization: the system must accurately surface latent target preferences—items the user would highly align with but has not literally named. Because this implicit target set is not directly observable, PAIR-Bench approximates it via collaborative filtering (detailed in §3.3) to establish a rigorous evaluation ground truth.

Domain	# Users	Avg. Prefs/User	# User–Preference Pairs (Train / Val / Test)	Image Gallery
Media	1,000	11.49	2,828 / 411 / 402	5,003
Games	1,000	7.44	2,775 / 411 / 404	4,721

Table 1: **PAIR-Data statistics.** 1,000 curated profiles per domain yield 7,231 total user–preference pairs. The image gallery is built via a modified MMR (see §3.2 and Appendix A).

3.2 PAIR-Data

PAIR-Data is constructed around three requirements: a corpus of declarative user preferences,

domains that are both preference-expressive and visually groundable, and a user pool balanced enough to support collaborative-filtering ground truth (Figure 2). We build on PersonaLens (Zhao et al., 2025), a corpus of 1,500 user personas with declarative preference triples (e.g., favorite titles, preferred genres, favorite actors/directors) across 20 domains.

Domain selection. PersonaLens's domains range from utility-oriented (e.g., Alarm, Hotel) to lifestyle-oriented (e.g., Books, Movies, Games). Preference-aware image retrieval requires domains that are *preference-expressive*—users meaningfully differ from one another—and *visually groundable*—each item maps to a canonical, distinctive image. Utility domains are not preference-expressive, and other lifestyle domains fail one or both criteria. We retain **Media** (movies and TV) and **Games**, which satisfy both, with strong preference signals (genre, cast/director, era, mechanics) and large catalogs of canonical images.

User curation. From PersonaLens's 1,500 personas, we keep only declarative preference fields and use demographic attributes solely for balancing. We drop 266 users whose Media *and* Games preference fields are too sparse for evaluation. The remaining users skew demographically (over 50% UK/US, over 60% male); to prevent collaborative-filtering neighborhoods from reflecting this skew, we downsample at the intersection (male users in the UK or US) to a final pool of 1,000 users per domain (Appendix A).

Image gallery. We build the gallery from *titles*, the only profile level with canonical visuals (genres, directors, etc. stay in P_u as text). For every title named in any user profile, we collect up to five canonical images—movie posters from MovieLens/IMDb (Harper and Konstan, 2015) for Media, and header art and screenshots from the public Steam dataset (Ermilov et al., 2025) for Games. We expand each gallery with additional candidate titles from the domain's full catalog ($\sim$80,000 items) so that retrieval is non-trivial; selection details are in Appendix D. The final galleries contain 5,003 images for Media and 4,721 for Games.

Training pairs. We deduplicate surface-form variants in PersonaLens (e.g., Italian vs. English title spellings) before pairing. Each training instance holds out one *title* preference from the user's profile as the supervisory target, with the remaining profile (other titles, genres, directors, etc.) as model input. This yields 7,231 user–preference pairs across both domains (Table 1).

User-level data split. We partition users disjointly into train/val/test (8:1:1). A random split can leave rare titles, genres, or cast/directors in only one split, confounding preference generalization with profile-vocabulary shift. We instead apply iterative multi-label stratification (Sechidis et al., 2011) over multi-hot profile triples (titles, genres, cast/directors), halving the train–test Jensen–Shannon divergence relative to random splitting (Appendix E).

3.3 PAIR-Bench

Target Set Construction. The held-out items used for training (§3.2) lie inside P_u and therefore only probe recall of stated preferences; preference generalization, by definition, must be measured against items *outside* P_u. We construct such targets at test time via user-based collaborative filtering (CF) (Sarwar et al., 2001; Koren et al., 2009) over the 1,000-user pool per domain. We measure user similarity with the Jaccard index over two users' explicit preference sets,

$$\mathrm{Jaccard}(P_u, P_v) = \frac{|P_u \cap P_v|}{|P_u \cup P_v|}, \tag{1}$$

take u's nearest neighbors $N(u)$ by Eq. 1, and aggregate the titles in their profiles excluding items already in P_u:

$$T_u = \Big(\bigcup_{v \in N(u)} P_v \Big) \setminus P_u.$$

We report three neighborhood cutoffs (top 1%, 2.5%, 5%; Appendix A)

Evaluation Protocol. We report Top-K retrieval metrics against the target set T_u. To evaluate preference generalization, we compute Top-K on the ranked list after removing any image whose underlying item is already in P_u. The resulting score reflects only how well a model surfaces items the user would accept but did not literally name.

Together, PAIR-Data and PAIR-Bench supply an aligned training and evaluation pipeline whose CF-derived target sets explicitly probe preference generalization beyond the stated profile.

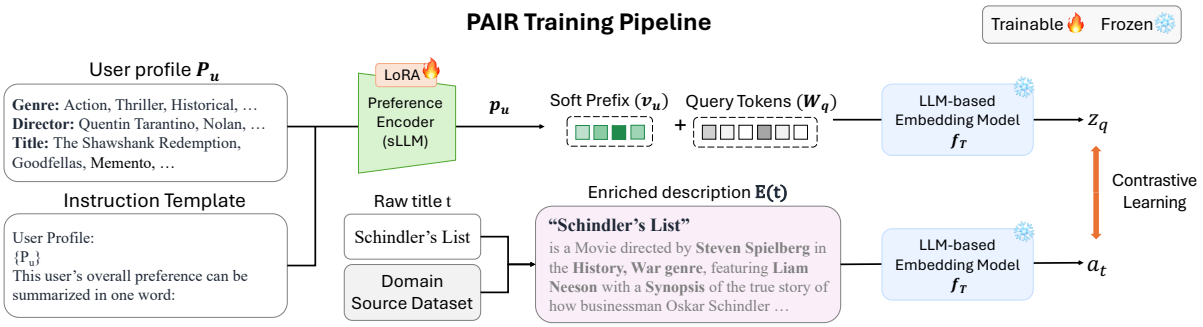


Figure 3: **PAIR training pipeline.** The user profile P_u is *compressed* into a K-token soft prefix $\mathbf{v}_u$ by a LoRA-adapted preference encoder g followed by a projection π. Symmetrically, the target title t is *enriched* into a metadata-rich description $E(t)$ — augmented with genre, cast/director, and synopsis — that grounds the contrastive target. The soft prefix $\mathbf{v}_u$ is prepended to query token embeddings W_q and encoded by the frozen LLM-based embedding backbone f_T to produce the preference-conditioned query embedding $\mathbf{z}_q$; $E(t)$ is encoded by the same f_T to produce the anchor embedding $\mathbf{a}_t$. Training aligns $\mathbf{z}_q$ with $\mathbf{a}_t$ via InfoNCE with in-batch negatives.

4 Method

The PAIR framework augments a frozen text–image embedding backbone with a small preference encoder that maps each user's profile to a soft prefix consumed by the backbone at retrieval time. Two trainable components are added: a preference encoder $g(\cdot)$ that compresses the variable-length profile P_u into a fixed-dimensional vector, and a projection layer π that maps this vector to a soft prefix of K tokens (Figure 3). All backbone parameters remain frozen throughout.

4.1 Architecture

The backbone is a pre-trained multimodal embedding model: a multimodal large language model (MLLM) contrastively fine-tuned to map text and image inputs into a shared embedding space $\mathbb{R}^d$ (Jiang et al., 2024b; Zhang et al., 2025; Li et al., 2026). The text encoder f_T and image encoder f_I share the underlying MLLM weights, differing only in modality-specific input formatting. We use the backbone without modification.

Preference Encoder. We instantiate $g(\cdot)$ as a small instruction-tuned decoder-only language model with PromptEOL pooling (Jiang et al., 2024a), a recent technique that repurposes autoregressive LLMs as text encoders without architectural changes by prompting the model to compress its input into a designated final token. Specifically, we wrap the profile P_u in a fixed instruction template ending with a soft bottleneck phrase (e.g., *"This user's overall preference can be summarized in one word:"*). This forces the LLM to compress the entire profile into its next-token prediction; the hidden state at the final token of the prompt is taken as the profile representation,

$$\mathbf{p}_u = \mathrm{PromptEOL}(g(P_u)) \in \mathbb{R}^{d_{\mathrm{LM}}}. \quad (2)$$

We adapt $g(\cdot)$ with LoRA (Hu et al., 2022) on the attention projections, leaving the base model weights frozen. The exact prompt template used for training and inference is detailed in Appendix B.1.

Projection. The projection π is a small MLP that lifts $\mathbf{p}_u$ into the backbone's embedding dimension, producing a soft prefix of K tokens:

$$\mathbf{v}_u = \pi(\mathbf{p}_u) \in \mathbb{R}^{K \times d}. \quad (3)$$

4.2 Preference-Aware Query Encoding

At retrieval time, the soft prefix $\mathbf{v}_u$ conditions the frozen text encoder on the user's profile. Let $W_q \in \mathbb{R}^{L \times d}$ denote the token embeddings of the query q. We prepend $\mathbf{v}_u$ to W_q and pass the combined sequence through f_T,

$$\mathbf{z}_q(q, P_u) = f_T([\mathbf{v}_u; W_q]), \quad (4)$$

where $[\cdot;\cdot]$ denotes concatenation along the token axis. The self-attention layers of f_T jointly attend over the user's prefix and the query tokens from the first layer. The score thus becomes $s(I, q, P_u) = \mathrm{sim}(\mathbf{z}_q(q, P_u),\ f_I(I))$.

4.3 Training

We train g and π end-to-end on user–preference pairs (u, t) from PAIR-Data. The backbone f_T and f_I remain frozen throughout. A pair (P_u, t) in which t appears verbatim in P_u admits a trivial shortcut: the encoder can copy t from P_u into the prefix rather than infer it from the user's broader

Method	Qwen3-VL (2B)			GME (2B)			E5-V (7B)		
	mAP@5	nDCG@10	R@10	mAP@5	nDCG@10	R@10	mAP@5	nDCG@10	R@10
Media									
Q-only	2.99	3.83	14.50	0.79	1.71	10.06	0.53	1.11	9.47
P-only	5.52	11.30	48.82	4.23	8.13	38.17	6.60	11.33	47.63
P+Q	5.04	10.84	50.30	4.89	8.84	38.76	6.17	10.64	47.04
LLM-summary	5.49	11.43	52.66	5.64	9.82	38.76	6.88	11.92	47.93
LLM-persona	9.19	15.09	54.14	4.35	8.01	36.09	7.53	13.20	46.75
PAIR (ours)	**17.58**	**25.77**	**72.49**	**14.82**	**21.97**	**66.57**	**22.01**	**30.12**	**78.40**
Games									
Q-only	3.80	6.65	38.96	0.71	0.94	3.57	0.32	1.97	16.88
P-only	3.54	7.92	35.06	2.98	5.35	24.68	3.30	7.35	33.12
P+Q	4.34	9.63	42.21	1.35	3.28	21.43	3.62	8.02	35.71
LLM-summary	2.82	7.57	33.12	0.46	1.87	12.99	2.58	6.03	28.57
LLM-persona	4.35	9.36	39.94	1.39	3.22	18.18	0.82	2.28	12.01
PAIR (ours)	**10.50**	**17.47**	**58.77**	**8.59**	**13.69**	**49.35**	**10.15**	**16.62**	**57.14**

Table 2: **Main results on PAIR-Bench** (implicit 1% tier; mAP@5, nDCG@10, R@10 in %). All non-PAIR rows share the same frozen backbone: *textual-concat* variants (Q-only, P-only, P+Q) concatenate profile and query; *LLM-rewriting* variants (LLM-summary, LLM-persona) route the profile through Qwen3.5-9B before encoding (prompts in Appendix B.2). Best per (backbone, domain, metric) in **bold**, second-best underlined.

preferences. Because evaluation targets are by construction absent from P_u, this shortcut would not transfer. We therefore drop the profile sentence that introduces t from P_u before encoding, forcing the model to recover t from the rest of the profile.

Target Embedding. With f_I frozen, all learning signal flows through the text-side anchor $\mathbf{a}_t$, which is cached once before training. Raw titles are short, polysemous, and weakly grounded in visual features, so we augment each title with its genre, cast/director, and synopsis (§3.2) to form an **enriched description** $E(t)$. The anchor embedding is thus computed as:

$$\mathbf{a}_t \;=\; f_T(E(t)). \tag{5}$$

Contrastive Objective. For brevity, write $\mathbf{z}_u := \mathbf{z}_q(q, P_u)$ for the preference-conditioned query embedding for user u. We align $\mathbf{z}_u$ with the corresponding target embedding $\mathbf{a}_t$ via InfoNCE (Oord et al., 2018) with in-batch negatives:

$$\mathcal{L} \;=\; -\frac{1}{|\mathcal{B}|}\sum_{(u,t)\in\mathcal{B}} \log \frac{\exp(\mathbf{z}_u^\top \mathbf{a}_t/\tau)}{\sum_{(t')\in\mathcal{B}} \exp(\mathbf{z}_u^\top \mathbf{a}_{t'}/\tau)}, \tag{6}$$

where τ is a learnable temperature. The architecture is backbone-agnostic; the same g, π design transfers across the backbones we test (§5.1).

5 Experiments

5.1 Experimental Setup

We evaluate PAIR on PAIR-Bench (§3.3) across Media and Games.

Backbones. We instantiate PAIR on three MLLM-based unified embedding backbones: **Qwen3-VL-Embedding-2B** (Li et al., 2026) and **GME-VL-2B-Instruct** (Zhang et al., 2025), and the larger **E5-V** (Jiang et al., 2024b) (LLaVA-NeXT, 7B).

Baselines. We compare PAIR to five baselines that share the same frozen backbone. Three are textual concatenations of the user profile P_u and the generic query q: *Q-only* (query alone), *P-only* (profile alone), and *P+Q* (profile followed by query). Two use an external LLM (Qwen3.5-9B) (Qwen Team, 2026) at inference time as a preference reasoner: *LLM-summary* replaces the profile with an LLM-generated 2–3 sentence summary that is then concatenated with the generic query, and *LLM-persona* replaces both profile and query with a first-person preference articulation generated by the LLM. Exact prompts are in Appendix B.2; note that the LLM-rewriting baselines use a 9B model, larger than PAIR's 1.2B preference encoder.

Metrics and protocol. We evaluate retrieval quality with three standard metrics—mean Average Pre-

Method	K=1	K=5	K=10	K=20
Head-Item Repetition Rate@K (%) ↓				
Frozen (P+Q)	17.3	30.0	34.5	53.6
LLM-summary	25.5	53.6	74.5	90.9
LLM-persona	8.2	10.0	15.5	22.7
PAIR (ours)	**4.5**	**7.3**	**8.2**	**14.5**
Catalog Coverage@K (%) ↑				
Frozen (P+Q)	3.0	7.6	13.6	21.2
LLM-summary	1.4	7.4	12.3	19.0
LLM-persona	3.6	12.0	20.0	29.2
PAIR (ours)	**5.2**	**18.3**	**30.1**	**45.8**

Table 3: **Diversity and coverage** (Qwen3-VL × Media, implicit 1% tier). **Head-Item Repetition Rate@K** — % users with the most-frequent title in their Top-K. **Catalog Coverage@K** — % of 1,413 gallery titles surfaced across users' Top-K.

Anchor description	mAP@5	nDCG@10	R@10
Media			
Title only	13.74	20.83	67.65
+ Genre	17.16 (+3.42)	25.53 (+4.70)	**73.96** (+6.31)
+ Full description	**18.14** (+4.40)	**25.95** (+5.12)	72.49 (+4.84)
Games			
Title only	7.69	12.81	46.21
+ Genre	9.63 (+1.94)	15.87 (+3.06)	**56.82** (+10.61)
+ Full description	**9.75** (+2.06)	**15.93** (+3.12)	55.09 (+8.88)

Table 4: **Anchor enrichment ablation** (implicit 1% tier; metrics in %, averaged across the three backbones). We progressively enrich the textual anchor $E(t)$ from a raw title to a full metadata-augmented description. Subscripts: positive Δ over the **Title only** row.

cision (mAP), normalized Discounted Cumulative Gain (nDCG), and Recall (R)—each computed at cutoffs $K \in \{5, 10\}$ on the implicit 1% tier (§3.3). Items in the user's explicit profile P_u are removed from the ranked list before top-K computation, so each metric reflects only the model's ability to surface unseen but compatible items. Results on the 2.5% and 5% implicit tiers are in Appendix G.

Training. PAIR is trained on PAIR-Data with early stopping on validation within 10 epochs. Implementation details, including LoRA configuration and PAIR's soft-prefix width $K \in \{2, 4\}$, are listed in Appendix F. The generic query q is *"Recommend images related to movies and TV shows"* for Media and *"Recommend images related to video games"* for Games, identical across methods.

5.2 Main Results

PAIR outperforms the strongest frozen-backbone variant on every (backbone, domain) setting by 1.9–2.9× on mAP@5 (Table 2), with consistent gains on nDCG@10 and R@10.

Text concatenation does not exploit profile context. The three textual-concatenation Frozen variants (Q-only, P-only, P+Q) fall within a narrow band on every cell. The frozen backbone, fed the user profile as text, does not extract preference signal at the level the task requires.

External LLM reasoning does not bridge the gap. Routing the profile through a 9B-parameter LLM—as LLM-summary or LLM-persona—does not close this gap. The gain comes from training: PAIR's 1.2B preference encoder is fit jointly with the retrieval objective, which prompting-based interventions cannot match.

Architectural advantages. PAIR keeps the backbone (the LLM-based embedding model) frozen: all training adapts only the external preference module g and the projection π. This has three practical consequences. First, backbone outputs—image gallery embeddings and target embeddings—computed once for the Frozen baseline are reused unchanged at inference under PAIR. Second, a single deployed backbone can serve both standard and preference-aware retrieval without weight-swapping. Third, the same PAIR design transfers across all three backbones we evaluate—including the most recent Qwen3-VL-Embedding model (Li et al., 2026)—and across encoder architectures (§5.4), without architectural changes.

5.3 Diversity and Coverage

Retrieval accuracy alone does not capture personalization quality. A system that always returns the globally most-popular items can still achieve a high mAP@K if those items are also relevant to many users. To probe this dimension, we measure two properties of each method's retrieval results (Table 3). **Head-Item Repetition Rate** identifies each method's *head item*—the title most frequently ranked first across users—and reports the fraction of users whose Top-K retrieval results contain it. Near 100% means everyone gets the same head item; lower means real differentiation. **Catalog Coverage** is the fraction of distinct gallery titles surfaced across users' Top-K retrieval results.

At K=20, Frozen's head item appears in

Preference encoder	mAP@5	nDCG@10	R@10
Frozen (P+Q)	5.04	10.84	50.30
BERT-base-uncased (110M)	16.16	25.31	**73.08**
Llama-3.2-1B-Instruct (1.2B)	**17.58**	**25.77**	72.49

Table 5: **Preference encoder ablation** on Qwen3-VL $\times$ Media (implicit 1% tier; metrics in %). PAIR with different preference encoders g, with the rest of the pipeline held fixed. The Frozen (P+Q) row is the no-encoder baseline.

53.6% of users' Top-20—the head-item dominance flagged in §1. PAIR reduces this to 14.5% (roughly 4$\times$ lower), and Catalog Coverage roughly doubles at the same K. The trend holds at smaller K. This means PAIR retrieves substantively different items for different users; the retrieval set itself shifts with the profile, not just its order.

5.4 Ablations

We ablate two design choices of PAIR: (i) the textual input E(t) representing the contrastive target, and (ii) the architecture of the preference encoder.

Anchor enrichment. The enriched description $E(t)$ serves as the textual input to generate the anchor embedding a_t for contrastive learning. Because f_I is frozen, the target signal relies solely on the text representation through f_T; a raw title alone provides a thin signal. We progressively enrich $E(t)$ to a full metadata-augmented description (Table 4). Averaged across the three backbones, the full anchor improves mAP@5 by +4.40 on Media and +2.06 on Games, showing that target-side gains scale with the information exposed to f_T. Genre delivers most of this lift, with other descriptions adding complementary gains, motivating our fully enriched anchor.

Preference encoder. PAIR's design is agnostic to the choice of preference encoder g. We test this on Qwen3-VL—the most recent multimodal embedding backbone in our evaluation—by replacing the default Llama-3.2-1B-Instruct with a BERT-base-uncased model (110M parameters, mean-pooled), keeping every other component unchanged (Table 5). Both encoders deliver substantial gains over the no-encoder Frozen baseline, and each wins a different subset of metrics: the 110M BERT reaches 16.16 mAP@5, while the 1.2B Llama reaches 17.58, showing PAIR's gain holds across encoders of very different scale and training paradigm.

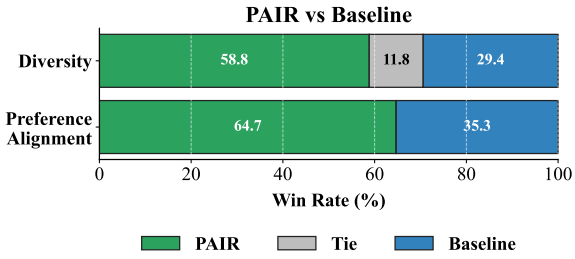


Figure 4: **Blinded pairwise user study** on participants' own Media profiles (n=17). Win/tie/loss rates on the diversity and preference-alignment criteria.

5.5 User Study

We complement the automatic metrics with a blinded pairwise user study comparing PAIR and the Frozen baseline (Figure 4). Seventeen volunteer raters each authored their own Media profile and rated the two systems' Top-5 lists on (i) diversity within the retrieved set and (ii) alignment with the rater's preferences. PAIR's win rate exceeds the baseline's by 2.0$\times$ on diversity and 1.8$\times$ on preference alignment. The consistent direction across both criteria under blinded comparison corroborates the automatic-metric findings. Per-user qualitative comparisons across both domains are provided in Appendix C; full user-study details (recruitment, interface, rating instructions) and examples are in Appendix H.

6 Conclusion

We introduced Preference-Aware Image Retrieval (PAIR), a novel task that conditions text-to-image retrieval on each user's stated profile. To support this, we release PAIR-Data and PAIR-Bench: a training corpus and an evaluation framework whose collaborative-filtering-derived target sets rigorously measure preference generalization beyond explicitly named items. Methodologically, our lightweight preference encoder compresses user profiles into a soft prefix, effectively steering a frozen MLLM embedding backbone. Evaluated across two domains and three backbones, PAIR substantially outperforms text-concatenation and LLM-rewriting baselines. Crucially, our approach expands catalog coverage and successfully mitigates head-item dominance—quantitative gains further validated by a blinded user study. We release our code and datasets to facilitate future research in preference-aware cross-modal retrieval.

7 Limitations

Domain and Modality Scope. PAIR-Data currently focuses on static text-to-image retrieval within two English-language domains (Media and Games) utilizing synthetic user profiles. Validating PAIR on real-world behavioral logs across diverse languages and additional domains (e.g., e-commerce) is necessary to ensure broader generalization. Furthermore, because our formulation is currently restricted to static images, adapting this preference-conditioned setup to text-to-video or text-to-3D retrieval represents an important open direction.

Textual Profile Reliance. Our framework relies entirely on declarative textual profiles to capture user tastes. While effective for genre or entity preferences, users often possess nuanced art styles or visual aesthetics that are inherently difficult to articulate in text. Extending the preference encoder to integrate purely visual reference signals alongside text remains a crucial next step to fully capture a user's taste landscape.

Evaluation Constraints. Our evaluation approximates implicit target sets via collaborative filtering (CF), which risks propagating popularity bias and may underperform for users in sparse regions of the profile graph. Additionally, while our human evaluation (§5.5) validates the automated metrics, it is constrained by a small sample size (n=17). Incorporating large-scale longitudinal logs and broader human annotations would significantly strengthen both the ground-truth signals and the real-world applicability of our findings.

References

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. 2024. Myvlm: Personalizing vlms for user-specific queries. *arXiv preprint arXiv:2403.14599*.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15338–15347.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Sumanth Doddapaneni, Krishna Sayana, Ambarish Jash, Sukhdeep Sodhi, and Dima Kuzmin. 2024. User embedding model for personalized language prompting. *arXiv preprint arXiv:2401.04858*.

Artemiy Ermilov, Arina Nevolina, Artem Pospelov, and Assol Kubaeva. 2025. Steam games dataset 2025.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*.

F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Young Kyun Jang, Junmo Kang, Yong Jae Lee, and Donghyun Kim. 2024. MATE: Meet at the embedding - connecting images with long texts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1659–1672, Miami, Florida, USA. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024a. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024b. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206. IEEE.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3045–3059.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Chen Keqin, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2125–2134.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Jesse Mu, Xiang Li, and Noah Goodman. 2023. Learning to compress prompts with gist tokens. *Advances in Neural Information Processing Systems*, 36:19327–19352.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. 2024. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems*, 37:40913–40951.

Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2024. User-llm: Efficient llm contextualization with user embeddings. *Preprint*, arXiv:2402.13598.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 285–295, New York, NY, USA. Association for Computing Machinery.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III*, ECML PKDD'11, page 145–158, Berlin, Heidelberg. Springer-Verlag.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9274–9285.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025. Personalens: A benchmark for personalization evaluation in conversational ai assistants. *arXiv preprint arXiv:2506.09902*.

A Dataset details

Domains and user pool. PAIR-Bench covers two retrieval domains: **Media** (movies and TV from MovieLens/IMDb) and **Games** (Steam Games Dataset 2025). Both domains share the same 1,000-user pool, with each user contributing a profile in both domains; the user-level 8:1:1 train/val/test split (§E) is applied jointly, so no user appears in more than one split for either domain.

Demographics. User demographic attributes are inherited from the PersonaLens corpus (Zhao et al., 2025); we use them only to describe the resulting pool, not as curation or training signal. Table 6 summarises the distribution. The pool skews English-speaking and young-adult, with the United States, United Kingdom, South Africa, Australia, and New Zealand together accounting for over half of all users.

Gender	
Male	573
Female	404
Non-binary / not stated	23
Age	
18–34	675
35–54	221
55+	103
Not stated	1
Country (top 5)	
United States	168
United Kingdom	153
South Africa	75
Australia	66
New Zealand	63
Other (33 countries)	475

Table 6: User-pool demographics (1,000 users, shared across both domains; attributes inherited from PersonaLens).

Profiles. Each user profile is authored as a list of templated triples `(Domain, Field, Value)`. The fields differ slightly between domains:

- **Media**: *Favorite Media* (titles), *Preferred Genres*, *Favorite Actors and Directors*.
- **Games**: *Favorite Game* (titles), *Preferred Genres*. Games has no cast/director attribute.

A short rendering pass converts the triples into natural-language sentences (e.g., "prefers thriller genre. likes works by Idris Elba. favorite titles include Inception."), used as input to all methods. During training, the sentence mentioning the held-out target title t is removed from P_u to prevent trivial string matching against the target embedding. Listings 1 and 2 show representative profiles for each domain along with the enrichment of one item-level preference into the target embedding's input description.

Statistics. Table 7 summarises the resulting splits and gallery sizes. Preference counts per user are tightly distributed: Media has 9–11 triples per user, Games has 4–5. The implicit target set T_u for each test user (§3.3) is constructed once from the CF (Jaccard) nearest-neighbour graph; see Appendices D and E for the gallery and split construction.

	Media	Games
Users (total)	1,000	1,000
Pairs (train)	2,828	2,775
Pairs (val)	411	411
Pairs (test)	402	404
Gallery: distinct titles	1,413	946
Gallery: image vectors	5,003	4,721

Table 7: Per-domain dataset statistics. The shared 1,000-user pool produces separate (profile, target) training pairs in each domain by holding out one preference triple at a time.

Preference embedding structure. Figure 5 visualises the structure of the preference-embedding space induced by PAIR. Three randomly-sampled user-preference embeddings are plotted alongside the Top-5 gallery images nearest each in cosine similarity; nearby images consistently reflect the dominant fields of each profile (e.g., for a fantasy/action/sci-fi preference, neighbours include *Thor* and *Guardians of the Galaxy*).

Listing 1: Example of a user profile in the Media domain. Each entry encodes either a preferred media title (T), preferred genre (G), or favorite actors and directors (AD). The last entry pairs an original item-level preference with the enriched description produced by our enrichment function.

```
[
  { "user_id": "user720" },
  "Media, T, Frances Ha",
  "Media, AD, Meryl Streep",
  "Media, AD, Greta Gerwig",
  "Media, G, independent",
  "Media, G, comedy",
  "Media, T, Marriage Story",
  "Media, G, drama",
  "Media, AD, Noah Baumbach",
  "Media, T, Little Women",
  "Media, AD, Cate Blanchett",
```

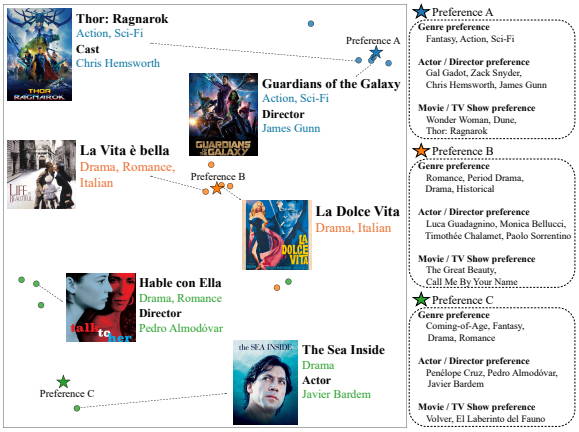


Figure 5: UMAP visualisation (McInnes et al., 2020) of three randomly-chosen user-preference embeddings (stars) and the Top-5 movie/TV show image embeddings nearest each preference (circles). Images near *Preference A* (fantasy/action/sci-fi) include *Thor* and *Guardians of the Galaxy*; images near *Preference B* and *C* similarly reflect each profile's dominant fields.

```
  {
    "Original": "Media, T, Bridesmaids",
    "Enriched": "Bridesmaids is a Movie
        and TV show (Directed by Paul
        Feig) in the Comedy, Romance
        genre, featuring Kristen Wiig,
        Maya Rudolph, Rose Byrne, Chris
        O'Dowd, Wendi McLendon-Covey.
        Synopsis: Annie's life is a mess
        . But when she finds out her
        lifetime best friend is engaged,
         she simply must serve as
        Lillian's maid of honor. Though
        lovelorn and broke, Annie bluffs
         her way through the expensive
        and bizarre rituals. With one
        chance to get it perfect, she'll
         show Lillian and her
        bridesmaids just how far you'll
        go for someone you love."
  }
]
```

Listing 2: Example of a user profile in the Games domain. Each entry encodes either a preferred game genre (G) or a preferred game title (N). The last entry pairs an original item-level preference with the enriched description produced by our enrichment function.

```
[
  { "user_id": "user1300" },
  "Games, G , RPG",
  "Games, N , Stardew Valley",
  "Games, G , simulation",
  "Games, N , Animal Crossing: New
      Horizons",
  "Games, G , adventure",
  {
    "Original": "Games, N , Portal 2",
    "Enriched": "Portal 2 is a video
        game. Tagged as Puzzle,
        Platformer, First-Person, Dark
        Humor, Story Rich, Puzzle
        Platformer, Funny, Female
        Protagonist, 3D Platformer, Sci-
        fi. Summary: Sequel to Portal
        that expands the portal-based
        physics puzzles with a new story
         and full co-op campaign across
        Aperture Science's rebuilt
        facilities."
  }
]
```

B LLM Prompt Examples

B.1 Prompt Template for Preference Encoder

To extract the compressed representation of the user profile P_u, we utilize the PromptEOL pooling technique (Jiang et al., 2024a). This method requires wrapping the input text in a specific instruction template designed to force the autoregressive language model to compress the context into the final token.

The exact prompt template used to instruct the decoder-only language model (g) during both training and inference is provided below. The placeholder $\{P_u\}$ is dynamically replaced with the declarative profile text of each user.

Preference Encoding Template

User profile:
$\{P_u\}$

This user's overall preference can be summarized in one word:

The hidden state corresponding to the colon (':') immediately following the phrase *"in one word"* is extracted as the final profile representation $\mathbf{p}_u$.

B.2 LLM baseline prompts and configuration

The two LLM-rewriting baselines in Table 2 (LLM-SUMMARY and LLM-PERSONA) route the user profile P_u through an external 9B-parameter LLM at inference time before feeding the resulting text to the frozen backbone. Both share the same model and decoding configuration; they differ only in the prompt used.

Model and decoding. We use **Qwen3.5-9B** (Qwen Team, 2026) via vLLM with greedy decoding (`temperature` $= 0$, `max_tokens` $= 128$, `seed` $= 42$). The chat template shipped with the Qwen3.5-9B tokenizer is applied. Because decoding is deterministic, a single generation is produced per (user, prompt); we do not average

over multiple seeds. The same model and decoding configuration are used for both Media and Games.

Domain placeholders. Both prompts contain a {*domain*} placeholder, which is replaced with *"movies and TV shows"* for Media and *"video games"* for Games. The LLM-PERSONA prompt additionally contains an {*item*} placeholder, replaced with *"movie or TV show"* or *"video game"* respectively. The profile placeholder $\{P_u\}$ is filled with the same natural-language rendering of the user's preference triples used by all other baselines (Appendix A).

LLM-SUMMARY. The LLM is prompted to compress P_u into a 2–3 sentence summary describing the user's preference patterns. The generated summary is concatenated with the generic query q (§5.1) and tokenized by the backbone for retrieval.

> **System:** You are a concise preference summarizer for a {*domain*} retrieval system.
>
> **User:** User profile:
> $\{P_u\}$
> In 2–3 sentences, describe this user's overall preference: themes, genres, tone, and stylistic patterns they favor. Output only the summary.

LLM-PERSONA. The LLM is prompted to articulate the user's current retrieval intent in first person. The generated description *replaces* both the profile and the generic query before encoding (i.e., the backbone receives the persona text alone as its text input).

> **System:** You are speaking as the user described below, in first person.
>
> **User:** You are a user with the following preferences in {*domain*}:
> $\{P_u\}$
> In 1–2 sentences, describe the kind of {*item*} you are currently looking for, speaking in first person. Output only the description.

Cost and footprint. Each call generates at most 128 tokens; with greedy decoding the generation is deterministic and can be cached per user. In aggregate, the LLM-rewriting step adds one short forward pass per user per evaluation, which is dwarfed by the gallery-side image-embedding cost of the retrieval pipeline. Note that, unlike PAIR's preference encoder (1.2B parameters), the LLM-rewriting baselines require a separate 9B-parameter inference at retrieval time and gain no parameter sharing with the retrieval backbone.

C Additional qualitative examples

Figures 6 and 7 present additional qualitative examples from the PAIR-Data test split. Each example shows a test user's stated profile P_u together with the Top-K unique retrieval items returned by the frozen E5-V baseline and by PAIR (Media: K=6; Games: K=5). Retrieved items are categorised into three classes: *explicit* (in P_u), *implicit* (in the CF-derived target set T_u at the 1% tier), or *unlabeled* (neither); figure colour coding follows these categories.

In the Media examples (Figure 6), PAIR retrieves items aligned with both explicit and implicit preferences. Notably, among the unlabeled results, PAIR retrieves *The Batman*: although not in P_u or T_u, it is semantically aligned with the implicit preference *The Dark Knight* (same protagonist, similar tone). Similarly, *The Imitation Game* shares its lead actor (Benedict Cumberbatch) with the implicit preference *Sherlock*, suggesting that some unlabeled retrievals reflect plausible preference generalization beyond the CF-derived ground truth.

In the Games examples (Figure 7), PAIR similarly aligns with both preference types. Among the unlabeled results, the model retrieves *Emily's Home Sweet Home*, a time-management game close to the explicit preference *Stardew Valley*, and *FarmVille*, a casual simulation widely cited as a genre antecedent of the same farming mechanics.

D Image gallery construction

For each domain we begin with a large initial candidate pool—approximately 80,000 poster/cover images per domain (Media: IMDb posters across all MovieLens titles; Games: Steam header/screenshot art across the active catalog). We then apply Maximal Marginal Relevance (MMR) selection to distil the pool into a final compact gallery that is both representative (of the catalog as a whole) and visually diverse.

MMR formulation. Given a relevance score $r(i)$ for each candidate i (we use the backbone's image embedding similarity to a domain centroid as a proxy for relevance) and a redundancy score $s(i, j)$

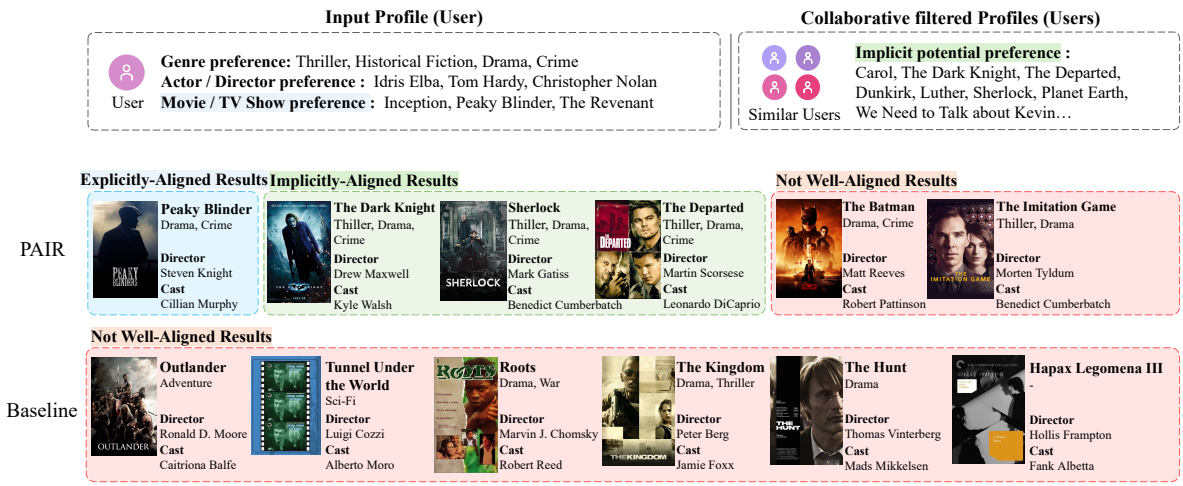


Figure 6: Additional qualitative example, **Media** domain. Given the generic query "Recommend images related to movies and TV shows" and a test user's profile, we compare the Top-6 unique items retrieved by the frozen E5-V baseline and by PAIR from the same fixed gallery. *Explicit* items (in P_u) are highlighted in blue, *implicit* items (in the CF-derived target set T_u at the 1% tier) in green, and *unlabeled* items in red.

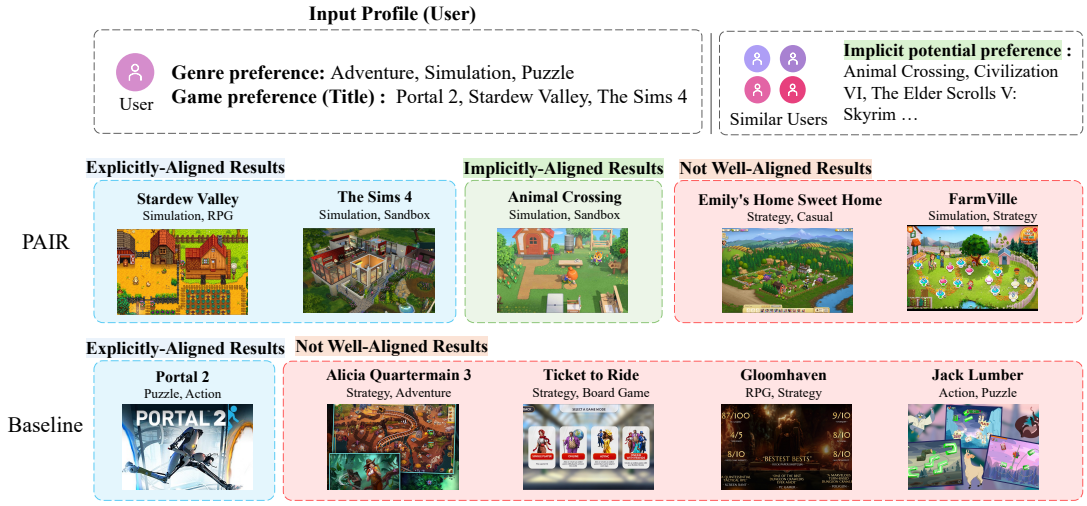


Figure 7: Additional qualitative example, **Games** domain. Given the generic query "Recommend images related to video games" and a test user's profile, we compare the Top-5 unique items retrieved by the frozen E5-V baseline and by PAIR from the same fixed gallery. Colour coding as in Figure 6: *explicit* (P_u) blue, *implicit* (T_u at 1%) green, *unlabeled* red.

between candidates (cosine similarity in image-embedding space), MMR iteratively picks

$$i^* = \arg\max_{i \notin S} \left[\lambda\, r(i) - (1-\lambda) \max_{j \in S} s(i,j) \right]$$

where S is the running selected set and $\lambda = 0.7$ trades off relevance against diversity. Ties are broken deterministically by candidate index.

Final gallery sizes.

- Media: **5,003** image vectors covering **1,413** distinct titles (multiple posters per title where available).
- Games: **4,721** image vectors covering **946** distinct titles.

Reproducibility. The MMR selection is deterministic given (i) the input ordering of the candidate pool, (ii) the relevance/similarity matrices, and (iii) the tie-breaking rule above. No random sampling is involved in the gallery construction itself. The image embeddings used as MMR inputs are produced by the same backbones evaluated in Table 2 (E5-V, GME, Qwen3-VL); the *gallery item set* is the union across backbones (we adopt the same title set per domain across all reported methods to enable cross-method comparison on

identical candidate pools).

E Splits and profile-distribution divergence

Profile-aware split. We partition the 1,000 users in the shared pool (§A) at the user level into 8:1:1 train/val/test, applied jointly across both domains. To minimise distribution shift between splits, we optimise a profile-aware partition: each user is represented as a sparse multi-hot vector over the union of all triples observed in their profile, and users are assigned to splits greedily such that the bag-of-triples distribution is approximately equal across splits. This realises iterative multi-label stratification (Sechidis et al., 2011) with all profile triples (titles, genres, cast/directors) as labels.

Jensen–Shannon divergence. We measure the agreement between the train/val/test profile-feature distributions with the symmetric Jensen–Shannon divergence (JSD; base-2 log, so values lie in $[0, 1]$). Lower is better. Our profile-aware splits achieve:

Domain	JSD (mean across the 3 split pairs)
Media	0.078
Games	0.084

For comparison, random user-level splits typically yield JSD in the 0.15–0.20 range on the same domains. The profile-aware split therefore reduces the train/test feature drift by roughly $2\times$, which is material on a benchmark where many test users have rare-tail genre combinations.

Why profile-aware. The implicit target set T_u (§3.3) is constructed from the CF (Jaccard) nearest-neighbour graph over user profiles. If train and test profile distributions drift, test users' T_u items lie in regions of the CF graph that the model has not seen during training, inflating apparent difficulty and confounding cross-method comparisons. The profile-aware split keeps the benchmark's difficulty curve stable and reproducible.

F Technical Details & Hyperparameters

Table 8 summarises the training hyperparameters used to produce every PAIR cell in the main paper, together with two auxiliary configurations reported in this appendix: a Backbone-LoRA variant on the largest backbone (E5-V), and the BERT preference-encoder ablation of §5.4. All runs share **AdamW** as the optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, gradient clip 1.0, and the learnable logit-scale parameter on a separate parameter group at $0.05\times$ the main LR and no weight decay. Validation criterion: lowest validation cross-entropy loss; the *best_ep* column is the epoch whose checkpoint achieved that minimum and was used for the reported metric.

Token position. When the backbone tokenizer prepends a $\langle$BOS$\rangle$ token to q, we insert v_u immediately after the $\langle$BOS$\rangle$ rather than at position 0, to preserve the BOS-conditioned input distribution the frozen backbone was trained on.

Backbone-LoRA. As a fine-tuning comparison, we additionally trained a Backbone-LoRA (BB-LoRA) variant on the largest backbone (E5-V, 7B): LoRA adapters on the text-attention projections of the frozen backbone, with no preference encoder. We evaluate it on both Media and Games as a sanity check that PAIR's external-adapter approach is competitive with direct backbone fine-tuning. We do not include BB-LoRA in the main paper because modifying the backbone breaks the deployment advantages PAIR is designed for (§5.2): cached gallery embeddings would have to be recomputed, and a single deployed backbone can no longer serve both standard and preference-aware retrieval without weight-swapping. Tier-sensitivity results for the two BB-LoRA cells are reported alongside PAIR and Frozen in Table 9.

Sweep ranges. For PAIR we swept K $\in \{2, 4\}$, lr $\in \{1\text{E-}3, 5\text{E-}4, 2\text{E-}4\}$, projection $\in$ {linear, 2-layer MLP}, and LoRA r $\in \{8, 16\}$ per (backbone, domain) cell. The locked configuration per cell (Table 8) is the one whose checkpoint achieved the best validation loss. We also sweep BB-LoRA lr $\in \{1\text{E-}3, 5\text{E-}4, 2\text{E-}4\}$ for the two E5-V BB-LoRA configurations.

Compute. All training and evaluation runs were performed on a node with $4\times$ NVIDIA A6000 GPUs (48GB each). Each (backbone, domain) training cell completes within standard laboratory compute budgets.

G Implicit-tier sensitivity

Table 2 in the main paper reports the headline 1% CF tier: the implicit target set T_u is the union of explicit preferences of u's top-1% nearest neighbours by profile Jaccard. As a robustness check, we re-evaluate the *same trained checkpoints* against the wider 2.5% and 5% tiers (strictly larger target

Method	Backbone	Domain	Optim.	lr	batch	LoRA r	LoRA α	LoRA targets	K (PAIR)	best_ep
PAIR	E5-V	Media	AdamW	5E-4	16	16	32	q,k,v,o,gate,up,down_proj	2	5 / 10
PAIR	E5-V	Games	AdamW	1E-3	16	8	16	q,v_proj	2	4 / 10
PAIR	GME	Media	AdamW	1E-3	16	8	16	q,v_proj	2	2 / 10
PAIR	GME	Games	AdamW	2E-4	16	16	32	q,k,v,o,gate,up,down_proj	4	2 / 6
PAIR	Qwen3-VL	Media	AdamW	2E-4	16	16	32	q,k,v,o,gate,up,down_proj	4	2 / 6
PAIR	Qwen3-VL	Games	AdamW	2E-4	16	16	32	q,k,v,o,gate,up,down_proj	4	2 / 6
Backbone-LoRA on E5-V — LoRA on the backbone's text-attention projections; no preference encoder.										
BB-LoRA	E5-V	Media	AdamW	1E-4	4	8	16	q,v_proj	—	2 / 5
BB-LoRA	E5-V	Games	AdamW	1E-4	4	8	16	q,v_proj	—	4 / 5
T4 encoder ablation — same PAIR pipeline, different preference encoder.										
PAIR (BERT)†	Qwen3-VL	Media	AdamW	1E-3	16	8	16	q,v_proj (BERT)	1	2 / 10

Table 8: **Training hyperparameters for every cell reported in the main paper.** "best_ep X/Y" = epoch X achieved the lowest validation loss out of a Y-epoch budget. Batch size is per-process. LoRA dropout = 0.05 throughout. † The BERT row uses LoRA adapters on BERT's attention.

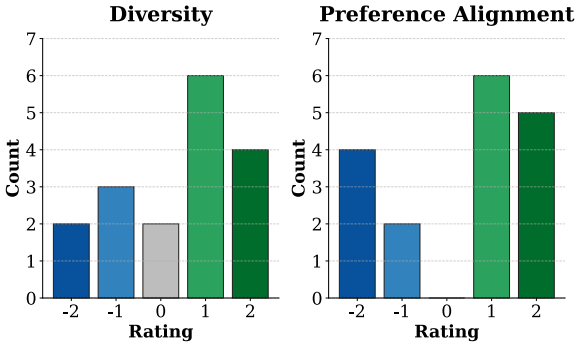


Figure 8: Distribution of signed user-study ratings on the 5-point scale $(-2, -1, 0, 1, 2)$ for the two evaluation criteria. The left panel shows the rating counts for *Diversity*, the right panel for *Preference alignment*. Positive values indicate a preference for PAIR over the frozen E5-V baseline.

sets; some misses at the 1% tier may register as hits at the wider tiers).

H User-study protocol

Recruitment. Seventeen volunteer participants were recruited through internal research-group mailing lists; eligibility criteria were (i) self-reported fluency in English, (ii) at least casual interest in movies, and (iii) no prior exposure to either retrieval system under evaluation. Informed consent was obtained.

Profile collection. Each participant authored a short preference profile using the same templated fields as PAIR-Bench (Appendix A): up to five favourite titles, up to three preferred genres, and (Media only) up to two favourite cast/director names. Participants selected each field value from the catalog autocomplete; no free-text input was permitted to keep profile vocabulary aligned with the test split.

Top-5 list generation. For each participant we generated two retrieval lists using the same **E5-V** backbone, identical generic query, and identical fixed image gallery. The two conditions differ only in how the user's profile is consumed:

- **Condition A — PAIR**: profile encoded by a trained BERT-based preference encoder (the same architecture as the `bert-base-uncased` variant reported in Table 5).
- **Condition B — baseline**: profile concatenated with the generic query as plain text and encoded by the unmodified frozen E5-V text tower (the *Frozen (P+Q)* variant of Table 2).

Each list contains the Top-5 unique titles after removing any title already in the participant's stated profile P_u, matching the implicit-protocol filter used in the main paper (§5.1). Image keys are deduplicated to titles.

Comparison interface. For each (participant, criterion) the two Top-5 image grids were presented side-by-side on a web form (Figure 9). Conditions were randomly assigned to the left/right positions per session and labeled simply "Group A" and "Group B"—participants were blinded to which system produced each list, and to the fact that there were two systems.

Questions. Each participant answered two questions, in randomised order:

Diversity: *"Which image group is more diverse?"*

			2.5% tier			5% tier		
Backbone	Domain	Method	mAP@5	nDCG@10	R@10	mAP@5	nDCG@10	R@10
E5-V	Media	Frozen	7.85	12.59	56.63	12.34	17.75	69.15
E5-V	Media	BB-LoRA	28.37	33.17	82.59	42.18	45.51	89.80
E5-V	Media	PAIR	22.22	27.22	78.86	30.91	36.24	88.31
E5-V	Games	Frozen	4.42	9.31	42.13	10.22	14.36	55.84
E5-V	Games	BB-LoRA	8.78	12.65	43.26	12.25	16.66	58.12
E5-V	Games	PAIR	8.91	13.62	54.78	13.68	17.64	65.48
GME	Media	Frozen	9.50	15.20	61.94	13.85	21.80	78.36
GME	Media	PAIR	18.60	23.06	73.38	24.96	31.56	87.06
GME	Games	Frozen	2.00	4.09	23.60	6.53	8.06	33.25
GME	Games	PAIR	7.14	11.99	42.70	10.04	14.83	53.81
Qwen3-VL	Media	Frozen	9.12	16.15	71.36	14.04	23.27	86.57
Qwen3-VL	Media	PAIR	21.55	27.12	75.62	28.55	34.41	85.82
Qwen3-VL	Games	Frozen	5.53	10.79	49.44	9.73	15.59	64.47
Qwen3-VL	Games	PAIR	11.68	16.90	59.83	15.14	19.99	69.04

Table 9: Tier-sensitivity for the main 12 (backbone, domain, method) cells plus the two auxiliary E5-V BB-LoRA cells. Same checkpoints as Table 2; only the CF tier threshold for target-set construction differs across the two column groups. PAIR's lead over Frozen is preserved at both wider tiers on every cell. On E5-V $\times$ Media, BB-LoRA's lead over PAIR widens at the 2.5%/5% tiers, consistent with the 1% picture: backbone fine-tuning fits Media particularly well on the largest backbone.

Preference alignment: *"Which image group better reflects your preferences?"*

Each question was rated on a five-point signed scale: -2 (strongly prefer left), -1 (slightly prefer left), 0 (tie), $+1$ (slightly prefer right), $+2$ (strongly prefer right). After unblinding, signs were flipped so that positive scores favour PAIR and negative scores favour the baseline.

Aggregation. Per criterion, each participant contributes one signed integer in $\{-2, -1, 0, +1, +2\}$; the distribution of these ratings is shown in Figure 8. We convert to win/tie/loss rates by collapsing $\{+1, +2\}$ to "PAIR win", $\{0\}$ to "tie", and $\{-1, -2\}$ to "baseline win", then divide by the participant count. The resulting rates are:

Criterion	PAIR win	Tie	Baseline win
Diversity	58.8%	11.8%	29.4%
Pref. align	64.7%	0.0%	35.3%

Caveats. With n=17 raters per question, formal statistical inference is limited and we do not report Wilson confidence intervals or null-hypothesis tests. We treat the consistent direction across both criteria under blinded comparison as corroborating evidence for the automatic-metric findings in §5; the automatic-metric tables (Tables 2, 3) carry the main quantitative claim.

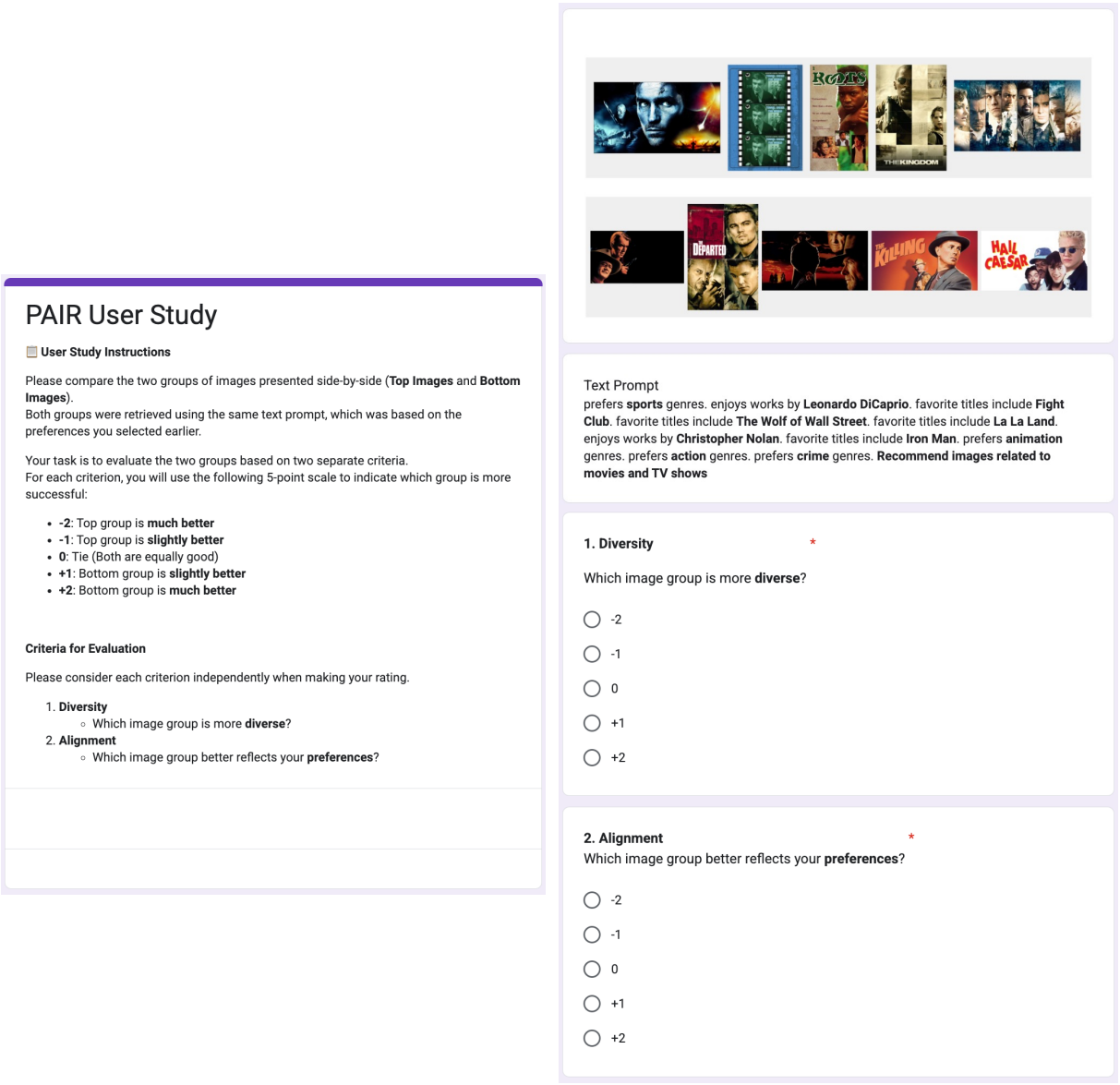
Per-user examples. Figure 10 shows representative per-user comparisons drawn from the study: each panel pairs a participant's own profile with the Top-5 lists returned by the frozen baseline and by PAIR, alongside the participant's signed preference-alignment rating.

Compensation and ethics. Participation was voluntary and unpaid; raters were research-group members blinded to the systems under comparison. The study is a low-risk, internal evaluation of consenting adult volunteers; our institution does not require formal IRB approval for such studies.

I Ethical considerations and disclosures

Potential risks. Preference-aware retrieval may amplify filter-bubble effects if deployed without diversity considerations. PAIR's diversity gains over the frozen baseline (Table 3) suggest that profile conditioning need not collapse retrieval to a popular tail, but production deployment requires ongoing monitoring of catalog reach across user demographics. The benchmark also inherits the demographic skew of PersonaLens (Table 6); models fit on such data may underperform for users from under-represented groups.

Use of AI assistants. We used AI assistants for editorial refinement of the paper text and the preparation of the anonymous GitHub repository.



(a) (b)

Figure 9: (a) The user-study instruction screen, and (b) the side-by-side rating questionnaire used to elicit *Diversity* and *Preference alignment* ratings on a -2 to $+2$ scale.

J Licenses and content

Licenses and intended use. PersonaLens (Zhao et al., 2025) is released for research use; MovieLens (Harper and Konstan, 2015) is provided under a research-only license; the Steam Games Dataset 2025 (Ermilov et al., 2025) is publicly available. Pretrained backbones (E5-V, GME-Qwen2-VL, Qwen3-VL-Embedding) and the preference-encoder backbones (Llama-3.2-1B-Instruct, BERT-base-uncased) are released under their respective open licenses. Our use of all these resources is consistent with their stated research-only or open access terms. We will release PAIR-Data and PAIR-Bench under CC BY-NC-SA 4.0 for research use.

Privacy and content. PersonaLens personas are synthetic and contain no real-user identifying information; demographic attributes (Table 6) are aggregated and not linked to individual personas. Image assets are publicly distributed movie posters and game cover art from the upstream catalogs.

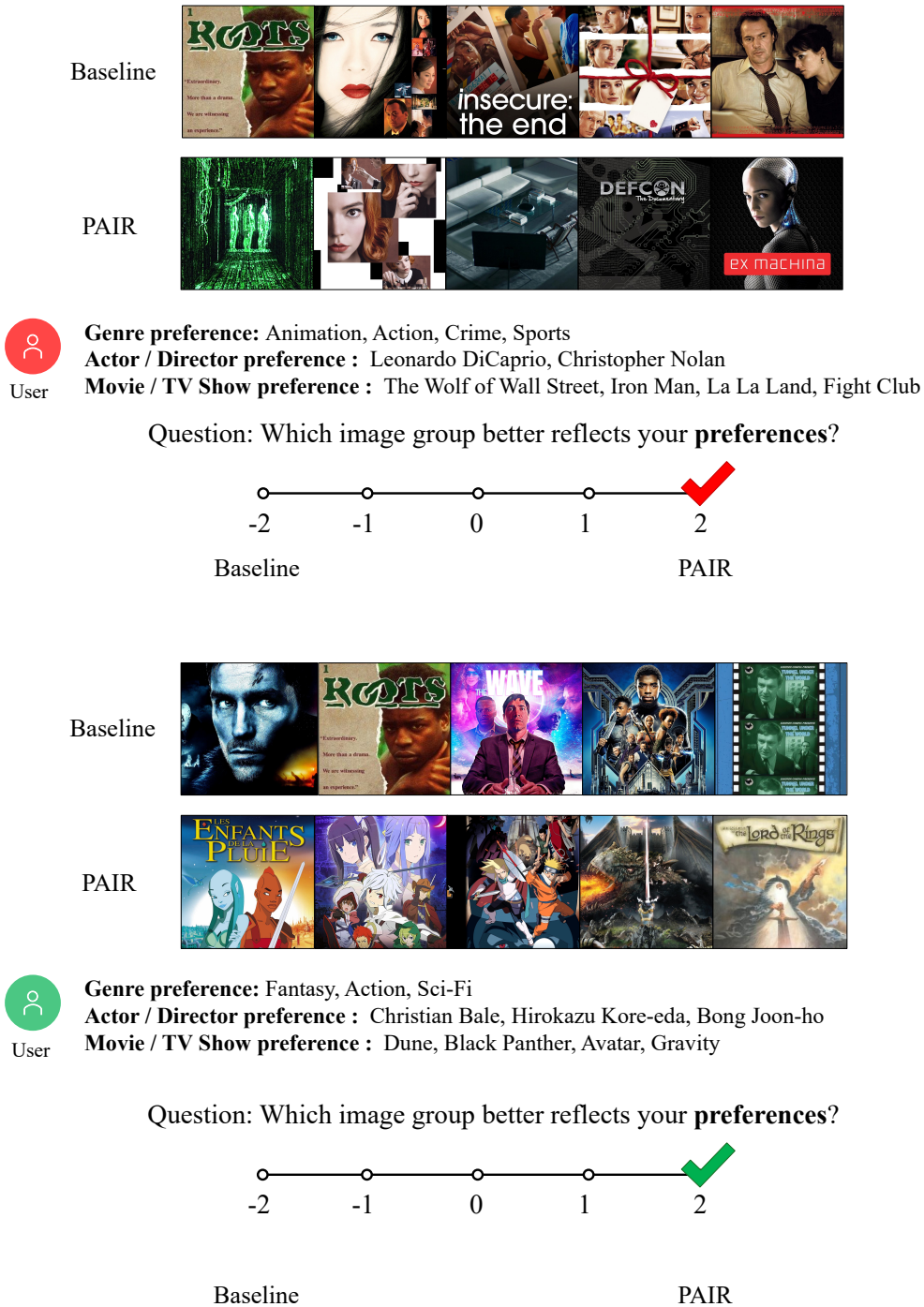


Figure 10: Real user-study examples in the *Media* domain. For each user, we show (top row) the Top-5 images returned by the baseline and (bottom row) those returned by PAIR, given the user's own textual preference profile (genres, actors/directors, and favorite titles). Below each pair of slates the rater's response to "Which image group better reflects your preferences?" is shown on a -2 to $+2$ scale, where positive values favor PAIR.