Powering Video Recommendations with Multimodal Embeddings Guided by LLMs

Andrii Dzhoha[*]
andrew.dzhoha@zalando.de
Zalando SE
Berlin, Germany

Katya Mirylenka[*]
katya.mirylenka@zalando.ch
Zalando Switzerland AG
Zürich, Switzerland

Egor Malykh[*]
egor.malykh@zalando.de
Zalando SE
Berlin, Germany

Marco-Andrea Buchmann
marco.andrea.buchmann@zalando.ch
Zalando Switzerland AG
Zürich, Switzerland

Francesca Catino
francesca.catino@zalando.ch
Zalando Switzerland AG
Zürich, Switzerland

Abstract

In recent years, social media users have spent significant amounts of time on short-form video platforms. As a result, established platforms in other domains, such as e-commerce, have begun introducing short-form video content to engage users and increase their time spent on the platform. The success of these experiences is due not only to the content itself but also to a unique UI innovation: instead of offering users a list of choices to click, platforms actively recommend content for users to watch one at a time. This creates new challenges for recommender systems, especially when launching a new video experience. Beyond the limited interaction data, immersive feed experiences introduce stronger position bias due to the UI and duration bias when optimizing for watch-time, as models tend to favor shorter videos. These issues, together with the feedback loop inherent in recommender systems, make it difficult to build effective solutions. In this paper, we highlight the challenges faced when introducing a new short-form video experience and present our experience showing that, even with sufficient video interaction data, it can be more beneficial to leverage a scalable video retrieval system using a multimodal vision-language model, guided by a Large Language Model for few-shot learning and evaluation, to overcome these challenges. This approach demonstrated greater effectiveness compared to conventional supervised learning methods in online experiments conducted on our e-commerce platform.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

Recommender Systems, Short-Form Video Retrieval, Large Language Models

1 Introduction

Short-form video platforms have rapidly reshaped digital engagement, with users now spending substantial time consuming immersive, vertically-scrolled video feeds. This paradigm shift has prompted established domains, including e-commerce, to experiment with similar experiences in order to capture user attention and drive engagement. However, the success of these platforms is not solely attributed to the content itself, but also to a distinctive UI innovation: rather than presenting users with a list of options, the system actively curates and presents content one item at a time, creating a highly engaging, lean-back experience.

This interaction model introduces new challenges for recommender systems. Unlike traditional settings where users' preferences are inferred from explicit choices among many options, immersive feeds rely on implicit signals such as watch-time and scroll behavior. The sequential, single-item presentation amplifies position bias [12], as users are more likely to engage with content shown earlier in the feed. Moreover, optimizing for watch-time can introduce a strong duration bias, with models tending to favor shorter videos that are more easily completed [20, 21]. These biases are further reinforced by feedback loops [4], making it difficult to ensure fair and relevant recommendations [7], especially in the early stages of a new product where interaction data is limited.

Traditional recommender models trained from scratch, such as collaborative filtering and other supervised approaches [3, 5, 18], are effective in mature platforms with plenty of representative data. However, they often struggle when available data is limited or exhibits strong biases, as is common in new product experiences. Counterfactual learning and bias correction methods have been proposed [9], but they require careful design and large-scale data, which are often unavailable in new experiences. Mitigating such biases is challenging [6], as existing methods often fail to remain robust when data is sparse [17], can show high variance in their results [15], or are affected by interleaving biases [16]. As a result, there is a growing need for approaches that leverage the generalization capabilities of foundation models and can be tailored for specific applications, such as launching a new short-form video experience on an e-commerce platform.

In this work, we highlight the challenges encountered when launching a new short-form video experience in e-commerce. Even with access to video interaction data, conventional methods can be influenced by duration and position biases, which may limit their effectiveness. To address these issues, we present a scalable retrieval system based on a multimodal vision-language model (CLIP [14]) that maps both user history and video content into a shared semantic space. This approach leverages the generalization capabilities of foundation models, enabling robust recommendations even in cold-start scenarios and outperforming conventional supervised

[*] These authors contributed equally to this work.

learning methods in our context. To personalize recommendations, we apply few-shot learning to the CLIP model using interaction data from the main e-commerce catalog, specifically from users' Browse and Search activities, capturing nuanced user preferences. The few-shot learning process is guided by a Large Language Model (LLM) for label refinement. For evaluation, we utilize the expert visual language model Qwen as an LVLM-as-a-judge. Our online experiments demonstrated that this approach increased watch-time completion rates by over 39%, while maintaining a balanced distribution of video duration, popularity, and watch-time.

Our main contributions are:

- We discuss the unique position and duration biases in immersive short-form video feeds, explaining their impact on recommender system performance.
- We share practical lessons from launching a new immersive video product, highlighting real-world challenges and the limitations of standard approaches.
- We present a scalable multimodal retrieval method using vision-language models and LLM-guided evaluation and few-shot learning, which delivers improved personalization and relevance.

2 Background

A recency-based solution is often the preferred initial method for video feeds, as it requires minimal engineering effort and enables rapid prototyping and launch. By surfacing the latest content, it drives engagement and content discovery, particularly in dynamic environments. This approach also minimizes the introduction of biases and the impact of feedback loops, making it a strong, interpretable baseline for evaluating more advanced personalized or multimodal methods.

Once interaction data becomes available, personalization is typically introduced using a conventional two-tower architecture [5, 18], which has become standard for scalable candidate generation. These models learn separate user and item embeddings for efficient retrieval. More advanced user models could represent interaction history as a graph of fashion interests, leveraging Siamese graph neural networks [10]. Given the limitations of early video interaction data, a pragmatic approach is to use non-trainable user embeddings from existing platform models, while training the video tower from scratch on video interactions. Relevance is often defined by videos achieving watch times above a 50% threshold, following industry standards; however, this introduces a bias toward shorter videos [20, 21]. Combined with position bias and feedback loops, the resulting data can be challenging, especially in immersive feed experiences where position bias is amplified [12]. Training typically uses dot-product similarity with sigmoid activation, optimized via binary cross-entropy loss, and evaluated using AUC and NDCG metrics.

Other approaches include reinforcement learning for optimizing retention and watch-time [2], real-time reranking [8], and systems designed for explicit video feedback [11]. While these methods increase modeling complexity or require more data to train user representations, they do not necessarily resolve the challenges posed by bias in immersive video feeds.

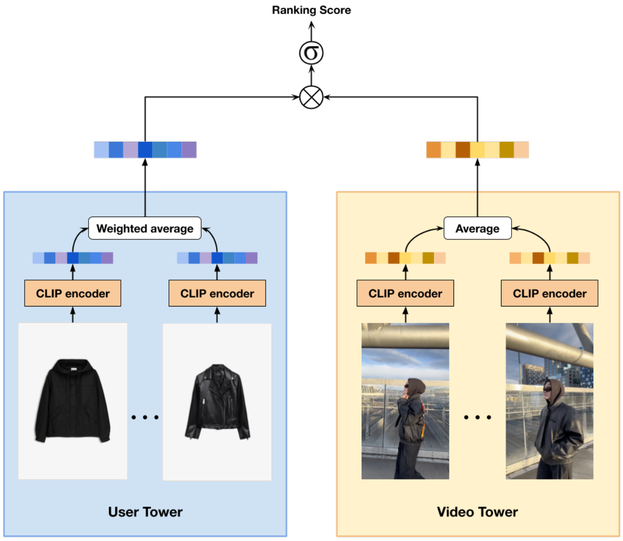

Figure 1: Overview of the multimodal retrieval architecture.

3 Multimodal retrieval approach

Recent advances in multimodal vision-language models, such as CLIP [14], enable robust retrieval by embedding both users and content into a shared semantic space. We leverage this capability with a two-tower architecture to address cold-start and bias challenges in short-form video recommendations for e-commerce.

The video tower computes embeddings by averaging CLIP representations of uniformly sampled video frames, capturing visual and semantic content. The user tower aggregates CLIP embeddings of products from a user's recent interaction history, weighted by recency, to form a personalized profile. This design allows effective matching between users and videos, even without direct user-video interactions. The architecture is depicted in Figure 1.

A key advantage of this approach is its ability to generalize from limited data, providing meaningful recommendations in cold-start scenarios and being less susceptible to duration and position biases present in conventional models.

Further, we utilize a proprietary, adapted version of CLIP within our company, enhanced through few-shot learning on interaction data from users' Browse and Search activities. This enables us to transfer knowledge from the main catalog's interaction data via standard discriminative loss modeling, allowing the model to better approximate a relevance function for video recommendations. Throughout this paper, all references to CLIP refer to our adapted version.

For evaluation, we employ an expert visual language model (Qwen) as an LVLM-as-a-judge, providing external relevance assessments that complement traditional metrics.

3.1 Method

Let $\mathcal{U}$ be the set of users and $\mathcal{V}$ the set of creator videos. For each user $u \in \mathcal{U}$, we have a time-ordered history of product interactions $H_u = [(s_1, t_1), \ldots, (s_n, t_n)]$, where s_i is a product and t_i is the timestamp. Our goal is to learn a scoring function $f : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$

that predicts the relevance of a video v to a user u, producing a ranked list of videos for each user. The scoring function is modeled as the dot product between user and video embeddings in a shared d-dimensional space derived from CLIP:

$$f(u,v) = \mathbf{e}_u^\top \mathbf{e}_v,$$

where $\mathbf{e}_u \in \mathbb{R}^d$ and $\mathbf{e}_v \in \mathbb{R}^d$ are the user and video embeddings, respectively. The video embedding $\mathbf{e}_v$ is computed as the average of CLIP embeddings from m uniformly sampled video frames:

$$\mathbf{e}_v = \frac{1}{m} \sum_{j=1}^{m} \mathbf{E}_{\text{CLIP}} \left(\text{frame}_j\right).$$

Video embeddings are pre-computed and indexed for efficient retrieval. A user's embedding $\mathbf{e}_u$ is dynamically computed online as a weighted average of the CLIP embeddings of products in their recent interaction history. The CLIP product embeddings, $\mathbf{E}_{\text{CLIP}}\left(\text{s}_k\right)$, are comprehensive, incorporating all textual metadata and images associated with the product. To give more importance to recent interactions, we apply an exponential decay weighting based on the time of interaction:

$$\mathbf{e}_u = \frac{\sum_{k=1}^{|H_u|} w_k \mathbf{E}_{\text{CLIP}}(s_k)}{\sum_{k=1}^{|H_u|} w_k},$$

where $w_k = \exp\left(-\lambda(t_{\text{now}} - t_k)\right)$ for decay factor λ. For new users, a global embedding based on popular products is used.

This architecture enables scalable, personalized video retrieval, with embeddings precomputed offline and user profiles computed online for real-time recommendations.

4 Experiments

In this section, we detail our experience personalizing the immersive short-form video feed on a large-scale e-commerce platform. The feed, designed for inspiration and entertainment, allows users to scroll through videos one at a time, similar to popular social media platforms.

4.1 Experimental setup

We began with a recency-based video feed to establish a baseline and collect initial user interaction data. This approach enabled rapid prototyping and provided insights into user engagement with the new experience. During this phase, we intentionally limited traffic to the new feed, allowing us to iteratively collect observations and learnings while minimizing potential risks to the broader user base.

VCG Conventional. For personalization, we first implemented a two-tower architecture. User embeddings were reused from the main e-commerce catalog model, trained on Browse and Search interactions, while the video tower was trained from scratch on video interactions. The final video embedding incorporated metadata, video ID, creator, brand, mean-pooled product and hashtag embeddings, and other relevant features, all projected through multiple non-linear layers. We framed the task as binary classification, predicting whether a video is relevant to a user, with relevance defined as watch time exceeding a 50% threshold. Training used a contrastive loss, with positives and negatives determined by this threshold.

The main e-commerce catalog model, from which we reused user embeddings, is trained on a large sample of catalog sessions. Each session includes articles shown in response to browse or search requests, contextual data (market, device, category), user history (clicks, add-to-cart, wishlist, purchases), and subsequent interactions. The dataset comprises 250 million sessions from 70 million users across 25 markets.

A schematic overview of the scalable two-tower-based Video Candidate Generation (VCG) architecture is provided in Section 2.

VCG Multimodal (CLIP-based). Subsequently, we introduced a multimodal retrieval system based on CLIP embeddings, as described in Section 3. This approach reused the same scalable two-tower architecture, with video embeddings precomputed and stored in an online index for efficient retrieval. User embeddings were dynamically computed from recent product interactions using CLIP representations, enabling real-time personalized video recommendations in a shared semantic space.

4.2 Evaluation protocol

Our evaluation procedure focuses on user engagement and is based on video feed observations. These observations are generated by the existing recency-based solution, meaning we can expect a representative sample of user preferences with fewer sampling and popularity biases that typically arise with Machine Learning models. We use a time-based split where the test (hold-out) data is formed by sequences from the last days. This split corresponds to the actual production setup. The ground truth is derived from the video interaction data, where relevance is modeled as a binary classification problem. A video is considered "watched" (positive example) if its watch time exceeds a 50% threshold. Each data point represents a video feed impression, enriched with a representation of user history.

We evaluate performance primarily using feed-wise ranking metrics to compare our approach with the current recency-based method on feeds containing multiple videos. Additionally, we use video-wise binary metrics for fine-tuning solutions and measuring performance across all feeds, including single-video impressions.

(1) Feed-wise (list-wise) ranking metrics: NDCG, applied to a subset of the test set where each example contains at least one positive and one negative video impressions.
(2) Video-wise (point-wise) binary classification metrics: Accuracy, AUC, precision, and recall.

To account for the position of relevant items, we weight NDCG by inverse propensity scores [13]. Additionally, we monitor the skewness of popularity and watch-time distributions to assess the extent to which the model favors shorter or more popular videos [19].

The ranking metric is specifically used to assess improvements over the current recency-based production solution. The underlying assumption is that an improvement in ranking metrics over the recency-based model should correlate with an enhancement in the retrieval task, while accepting the potential bias towards more active users.

4.3 Visual coherence evaluation

In addition to engagement metrics, our goal is to enhance the visual appeal of the feed in relation to a user's history. To measure this, we define the visual coherence between a user and a video as the dot product of the averaged content-based embeddings of the user's past interactions and the averaged content-based embeddings of the video's associated products. This metric reflects how well a user's history aligns with a video based on content-related features such as brand, color, silhouette, and more. These content-based embeddings are pre-extracted from product image representations, capturing visual attributes such as color, style, and silhouette to enable effective similarity comparisons.

4.4 LVLM-as-a-judge evaluation

To complement standard metrics, we used LVLM-as-a-judge (Large Vision-Language Model) for offline evaluation. Qwen 2.5-VL served as an external judge, rating the relevance of top-k ($k = 5, 10$) recommended videos based on a user's 12 most recent items of interest. Ratings were assigned on a 5-point scale, from 5 (extremely relevant) to 1 (no relevance). This approach provides an external perspective, especially valuable when user behavior data is limited. The prompt used for LVLM-as-a-judge is summarized in Figure 2.

```
You are an AI fashion relevance analyst. Your
primary function is to critically and objectively
evaluate the relevance of video content against a
specific user's fashion history. It is crucial that
you use the defined textual relevance categories
appropriately and avoid defaulting to a generally
positive assessment unless there is substantial,
specific evidence.
<...>
Assign one of the following textual categories for
relevance. Choose the category that most accurately
describes the alignment. Be discerning.
* "excellent_match": <...>
* "good_match": <...>
* "partial_match": <...>
* "poor_match": <...>
* "no_match": <...>
```

Figure 2: LLM prompt for LVLM-as-a-judge evaluation.

4.5 Offline evaluation

Offline evaluation using user behavior data and NDCG with inverse propensity scores showed only modest improvements in watch-time metrics for both VCG approaches compared to the recency baseline, with differences not reaching statistical significance. We hypothesize that position and duration biases were stronger than anticipated in the offline setting. Notably, the VCG Conventional model, trained on video interaction data, exhibited greater skewness in popularity and watch-time distributions than both the recency-based and VCG Multimodal solutions, indicating these biases affected it more. In terms of binary classification metrics, VCG Conventional achieved moderate discriminative power with AUC scores of 0.7.

Table 1: Summary of offline evaluation: VCG Multimodal (CLIP-based) vs recency-based baseline, measured by visual coherence and average LVLM-as-a-judge scores, with standard deviation in parentheses.

		LVLM-as-a-judge	
Method	Visual coherence	top-5	top-10
Recency-based	13.8 (6.48)	2.72 (0.30)	2.43 (0.24)
VCG Multimodal	18.9 (8.23)	3.12 (0.35)	3.09 (0.32)
Improvement	+37%	+14.7%	+27%

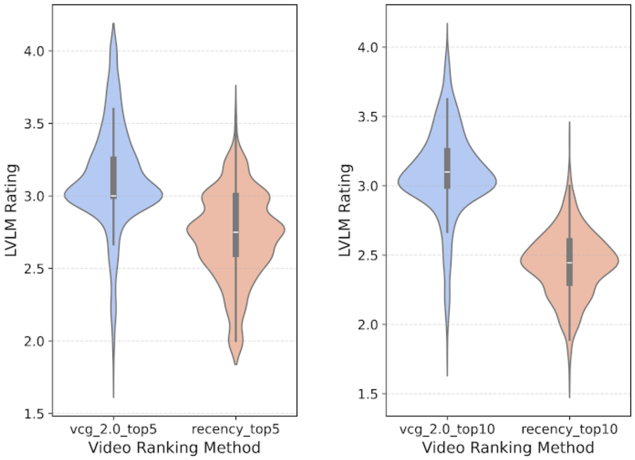

Figure 3: Distribution of LVLM-as-a-judge scores for VCG Multimodal (CLIP-based) and recency-based baseline.

Given the challenges of debiasing [6, 20, 21], we shifted focus in the second iteration with VCG Multimodal to visual coherence and LVLM-as-a-judge metrics, which are especially relevant in new product scenarios with limited user interaction data. Visual coherence gains were substantial, with up to a 37% increase for the top-100 recommended videos. LVLM-as-a-judge consistently assigned higher relevance scores to top-5 and top-10 VCG-ranked videos (see Figure 3). These results are summarized in Table 1, which reports absolute scores and standard deviations.

4.6 Online experiments

We first tested the VCG Conventional approach in an online experiment, which resulted in only modest gains in engagement metrics. However, this model produced a highly skewed popularity distribution (as defined by item co-occurrence) and a strong bias toward short videos, so it was not deployed to production.

In the next iteration, we evaluated the VCG Multimodal (CLIP-based) model. This approach delivered substantial improvements: the number of videos watched for at least 25% of their duration increased by 41% (CI: 21%–61%), and those watched for at least 50% increased by 50% (CI: 22%–78%). The rates of video starts reaching

25% and 50% progress rose by 30% (CI: 17%–44%) and 39% (CI: 17%–61%), respectively. All uplifts were statistically significant, and the model maintained stable core metrics, confirming proper randomization and no negative impact on other KPIs. Importantly, popularity and duration skewness were significantly reduced compared to the VCG Conventional model. Based on these strong results, the VCG Multimodal model was rolled out to production.

5 Conclusion

Immersive short-form video feeds introduce unique challenges for recommender systems, particularly due to strong position and duration biases that can distort relevance and fairness – especially in new product launches with limited interaction data. We found that conventional supervised approaches, while effective in mature settings, are highly susceptible to these biases and may not generalize well. By leveraging a scalable multimodal retrieval system based on vision-language models and guided by LLMs for evaluation, we effectively addressed these challenges. Our approach enabled robust personalization by mapping user history and video content into a shared semantic space, reducing the impact of position and duration biases. Offline and online experiments confirmed substantial improvements in both content relevance and user engagement, with significant gains in watch-time completion rates and a more balanced distribution of video popularity and duration. These results highlight the importance of foundation models and external evaluation frameworks, such as LVLM-as-a-judge, for building fair and effective recommender systems in immersive video environments. Future work will further explore advanced LVLM-based representations to deepen user-video understanding, quantify the reliability of LLM-based judgments [1], and continue mitigating bias in evolving recommendation scenarios.

References

[1] Debarun Bhattacharjya, Balaji Ganesan, Junkyu Lee, Radu Marinescu, Katsiaryna Mirylenka, Michael Glass, and Xiao Shou. 2025. SIMBA UQ: Similarity-Based Aggregation for Uncertainty Quantification in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

[2] Qingpeng Cai, Shuchang Liu, Xueliang Wang, Tianyou Zuo, Wentao Xie, Bin Yang, Dong Zheng, Peng Jiang, and Kun Gai. 2023. Reinforcing User Retention in a Billion Scale Short Video Recommender System. In *Companion Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 421–426. doi:10.1145/3543873.3584640

[3] Marjan Celikik, Jacek Wasilewski, Ana Peleteiro Ramallo, Alexey Kurennoy, Evgeny Labzin, Danilo Ascione, Tural Gurbanov, Géraud Le Falher, Andrii Dzhoha, and Ian Harris. 2024. Building a Scalable, Effective, and Steerable Search and Ranking Platform. arXiv:2409.02856 [cs.IR] https://arxiv.org/abs/2409.02856

[4] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* 41, 3, Article 67 (feb 2023), 39 pages. doi:10.1145/3564284

[5] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 191–198. doi:10.1145/2959100.2959190

[6] Andrii Dzhoha, Alexey Kurennoy, Vladimir Vlasov, and Marjan Celikik. 2024. Reducing Popularity Influence by Addressing Position Bias. arXiv:2412.08780 [cs.IR] https://arxiv.org/abs/2412.08780

[7] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. Understanding Echo Chambers in E-commerce Recommender Systems. 2261–2270. doi:10.1145/3397271.3401431

[8] Xudong Gong, Qinlin Feng, Yuan Zhang, Jiangling Qin, Weijie Ding, Biao Li, Peng Jiang, and Kun Gai. 2022. Real-time Short Video Recommendation on Mobile Devices. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 3103–3112. doi:10.1145/3511808.3557065

[9] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) *(WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 781–789. doi:10.1145/3018661.3018699

[10] Evgeny Krivosheev, Mattia Atzeni, Katsiaryna Mirylenka, Paolo Scotton, Christoph Miksovic, and Anton Zorin. 2021. Business entity matching with siamese graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[11] Zihan Lin, Hui Wang, Jingshu Mao, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and Ji-Rong Wen. 2022. Feature-aware Diversified Re-ranking with Disentangled Representations for Relevant Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 3327–3335. doi:10.1145/3534678.3539130

[12] Qingyun Liu, Zhe Zhao, Liang Liu, Zhen Zhang, Junjie Shan, Yuening Li, Shuchao Bi, Lichan Hong, and Ed H. Chi. 2023. Multitask Ranking System for Immersive Feed and No More Clicks: A Case Study of Short-Form Video Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 4709–4716. doi:10.1145/3583780.3615489

[13] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. Correcting for Selection Bias in Learning-to-rank Systems. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1863–1873. doi:10.1145/3366423.3380255

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

[15] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased Recommender Learning from Missing-Not-At-Random Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 501–509. doi:10.1145/3336191.3371783

[16] Ali Vardasbi, Harrie Oosterhuis, and Maarten de Rijke. 2020. When Inverse Propensity Scoring does not Work: Affine Corrections for Unbiased Learning to Rank. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1475–1484. doi:10.1145/3340531.3412031

[17] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to Rank with Selection Bias in Personal Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 115–124. doi:10.1145/2911451.2911537

[18] Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H. Chi. 2020. Mixed Negative Sampling for Learning Two-tower Neural Networks in Recommendations. In *Companion Proceedings of the Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 441–447. doi:10.1145/3366424.3386195

[19] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the long tail recommendation. *Proc. VLDB Endow.* 5, 9 (may 2012), 896–907. doi:10.14778/2311906.2311916

[20] Ruohan Zhan, Changhua Pei, Qiang Su, Jianfen Wen, Xueliang Wang, Guanyu Mu, Dong Zheng, Peng Jiang, and Kun Gai. 2022. Deconfounding Duration Bias in Watch-time Prediction for Video Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 4472–4481. doi:10.1145/3534678.3539092

[21] Yu Zheng, Chen Gao, Jingtao Ding, Lingling Yi, Depeng Jin, Yong Li, and Meng Wang. 2022. DVR: Micro-Video Recommendation Optimizing Watch-Time-Gain under Duration Bias. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 334–345. doi:10.1145/3503161.3548428