Unbiased Multimodal Reranking for Long-Tail Short-Video Search

Wenyi Xu
Zhejiang University
Hangzhou, China
xuwenyi@zju.edu.cn

Feiran Zhu
Kuaishou Technology
Hangzhou, China
zhufeiran03@kuaishou.com

Songyang Li
Kuaishou Technology
Beijing, China
lisongyang03@kuaishou.com

Renzhe Zhou
Kuaishou Technology
Hangzhou, China
zhourenzhe03@kuaishou.com

Chao Zhang
Kuaishou Technology
Hangzhou, China
zhangchao29@kuaishou.com

Chenglei Dai*
Kuaishou Technology
Hangzhou, China
daichenglei@kuaishou.com

Yuren Mao
Zhejiang University
Hangzhou, China
yuren.mao@zju.edu.cn

Yunjun Gao
Zhejiang University
Hangzhou, China
gaoyj@zju.edu.cn

Yi Zhang
Kuaishou Technology
Beijing, China
zhangyi49@kuaishou.com

Abstract

Kuaishou serving hundreds of millions of searches daily, the quality of short-video search is paramount. However, it suffers from a severe Matthew effect on long-tail queries: sparse user behavior data causes models to amplify low-quality content such as clickbait and shallow content. The recent advancements in Large Language Models (LLMs) offer a new paradigm, as their inherent world knowledge provides a powerful mechanism to assess content quality, agnostic to sparse user interactions. To this end, we propose a LLM-driven multimodal reranking framework, which estimates user experience without real user behavior. The approach involves a two-stage training process: the first stage uses multimodal evidence to construct high-quality annotations for supervised fine-tuning, while the second stage incorporates pairwise preference optimization to help the model learn partial orderings among candidates. At inference time, the resulting experience scores are used to promote high-quality but underexposed videos in reranking, and further guide page-level optimization through reinforcement learning. Experiments show that the proposed method achieves consistent improvements over strong baselines in offline metrics including AUC, NDCG@K, and human preference judgement. An online A/B test covering 15% of traffic further demonstrates gains in both user experience and consumption metrics, confirming the practical value of the approach in long-tail video search scenarios.

CCS Concepts

• **Information system** → **Reranking Optimization**.

*Corresponding author.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Keywords

Long-tail queries, Reranking, Large language models

ACM Reference Format:
Wenyi Xu, Feiran Zhu, Songyang Li, Renzhe Zhou, Chao Zhang, Chenglei Dai, Yuren Mao, Yunjun Gao, and Yi Zhang. 2018. Unbiased Multimodal Reranking for Long-Tail Short-Video Search. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

On leading short-video platforms serving hundreds of millions of searches daily, search quality is paramount. While long-tail queries constitute a significant fraction of this volume—approximately 20%—they are disproportionately responsible for poor user experiences. For these queries, the scarcity of user interaction data creates a vicious cycle known as the Matthew effect [10]. Lacking reliable signals, ranking models often over-rely on superficial cues like sensational thumbnails or catchy titles. Consequently, the results page is frequently dominated by clickbait and shallow content, which steadily erodes user trust and platform reputation [25].

In the long-tail regime, sparse behavioral signals make ranking especially vulnerable to three entangled biases. Position/exposure bias amplifies items merely because they were seen rather than truly preferred, leading models to overfit presentation effects. Popularity bias further distorts supervision by rewarding click-bait and over-exposed creators, especially when feedback is scarce. Finally, cross-modal mismatch, where the cover/title narrative diverges from the actual footage, obscures intrinsic quality. These effects are exacerbated by video's multimodal nature: relevance must be inferred from titles, ASR transcripts, OCR snippets, covers, and key frames, making it hard to disentangle genuine usefulness from noise under weak feedback [14, 29].

Existing remedies are insufficient as they fail to address these coupled challenges: multimodal inconsistency and query-specific relevance. On one hand, methods like query rewriting operate only at the textual level. They are fundamentally blind to the multimodal

nature of video and thus cannot detect critical inconsistencies between a video's title, its spoken content (ASR), and its actual visual footage [17]. On the other hand, while a generic video quality score might assess a video's internal coherence, its assessment is fundamentally query-agnostic. A well-produced video with perfect title-content consistency can still be completely irrelevant or even misleading for a specific long-tail query intent, making this signal unreliable for fine-grained ranking[13, 26].

To address the three sources of long-tail degradation, we propose a unified solution. For popularity bias and cross-modal mismatch, we employ a Large Language Model with world knowledge to perform explicitly query-aware multimodal parsing, integrating titles, ASR, OCR, covers, and key frames. The model assesses intent–content consistency and intrinsic content value, based on which we conduct labeling and training to produce query-specific, comparable scores, thereby suppressing spurious strong signals such as clickbait, shallow content, and over-exposure at the source. For position/exposure bias, we anchor on a user-behavior-agnostic experience score to reconstruct reranking training signals and the page-level reinforcement-learning reward: at the point level, we reconstruct behavior sequences with the experience score to attenuate position effects; at the page level, we construct nDCG-style returns from the experience-induced ideal order, refining the reward function used in reinforcement learning, thus correcting exposure-induced systematic bias. With this design, without relying on user behavior, we ground multimodal consistency and query semantics into a deployable, lightweight reranking signal that mitigates long-tail biases and improves user experience.

The main contributions of this work are listed as follows:

- We analyze the three sources of bias in long-tail queries and propose a user-behavior-agnostic, explicitly query-aware multimodal experience-scoring model.
- We use the model's experience scores to drive label reconstruction and page-level reward optimization in reranking.
- Offline studies and large-scale online A/B tests show stable improvements in user-experience metrics without sacrificing key consumption metrics.

2 Related works

2.1 Architecture for Reranking

Modern search systems typically follow a two-stage cascade: an initial retrieval step rapidly recalls candidate documents, and a subsequent reranking step refines these candidates with more sophisticated models [9, 16]. Within this reranking stage, recent studies have explored large language model approaches from four viewpoints: pointwise, pairwise, listwise, and setwise[1]. RankGPT[22] casts reranking as a sequence-generation problem and already outperforms traditional baselines in zero-shot settings, while setwise prompting scores[18, 32] an entire group of candidates in a single pass, markedly lowering inference cost without sacrificing effectiveness. To remedy the score-incomparability issue of listwise outputs, Self-Calibrated Listwise Reranking[20] introduces a "list view + point view" dual relevance scheme, using the point-view scores to calibrate listwise results and achieve global consistency. The open-source model RankVicuna[19] (7B parameters) further shows that even a relatively small LLM can reach GPT-3.5-level reranking quality in a zero-shot scenario. In addition, novel loss functions have been proposed: Softmax-DPO[6] couples a Plackett–Luce softmax with multiple negatives, diffNDCG[30] directly optimizes ranking metrics such as NDCG, and reinforcement-learning schemes like GRPO align LLM predictions with human preference signals[31]. However, these methods generally assume that labeled data are both plentiful and unbiased; when the supervision itself is noisy or skewed, the advantages of LLM-based rerankers can diminish substantially.

2.2 Personalized Ranking with LLMs

In personalized ranking, researchers focus on injecting user interests and behaviors into LLM-based rankers. For example, PREMIUM [23] encodes preferences via a tag system and lets users iteratively rank model outputs to achieve on-device, personalized fine-tuning of an LLM. For multimodal recommendation, NoteLLM-2[28] enriches item representations with image-and-text inputs so the LLM can capture visual preferences. CHIME[2] encodes a user's entire behavior sequence with an adaptive LLM and, through contrastive learning plus quantization, produces compact long-term interest vectors for holistic preference modeling. HLLM[3] adopts a hierarchical design: one LLM extracts content features from item descriptions, while another predicts future interests from historical interactions, thereby leveraging pre-trained knowledge in sequential recommendation. Work on online feedback loops is emerging as well: Wang[24] decouple novelty and preference with two separate LLMs, then filter novelty-driven recommendations through a preference-aligned model to balance diversity and relevance, significantly boosting user satisfaction and diversity. Collectively, these studies demonstrate how LLMs can learn and adapt to behavioral data to deliver more personalized rankings. However, despite their strong results on platforms with rich profiles and click logs, they depend heavily on high-quality, unbiased user features and thus do not transfer well to long-tail search scenarios.

2.3 Multimodal Alignment and Synthetic Data

In rich-media retrieval, textual, visual, and audio cues often misalign, causing click-bait or content mismatch; meanwhile, long-tail items suffer from sparse labels. Recent work tackles these issues via multimodal alignment and synthetic supervision. RagVL[7] fine-tunes a vision-language model as a reranker and injects visual noise during training to boost robustness for image retrieval. Google Multimodal Reranking[27] for Knowledge-Intensive VQA fuses vision, text, and knowledge vectors through cross-attention, improving both ranking metrics and VQA accuracy. In industry, LLM-Alignment Live-Streaming[12] compresses frames, ASR transcripts, and live comments into a unified embedding to filter modality-inconsistent streams under strict latency budgets. When human labels are scarce, generative models can create auxiliary signals offline. Promptagator[8] uses a handful of seed examples to mass-generate queries, yielding dense retrievers and rerankers that outperform fully supervised baselines. EnrichIndex[5] enriches each document with LLM-generated summaries and Q&A pairs, building a semantically enhanced index that boosts recall and nDCG without extra online inference. Doc2query-style pseudo-query expansion[11, 15]

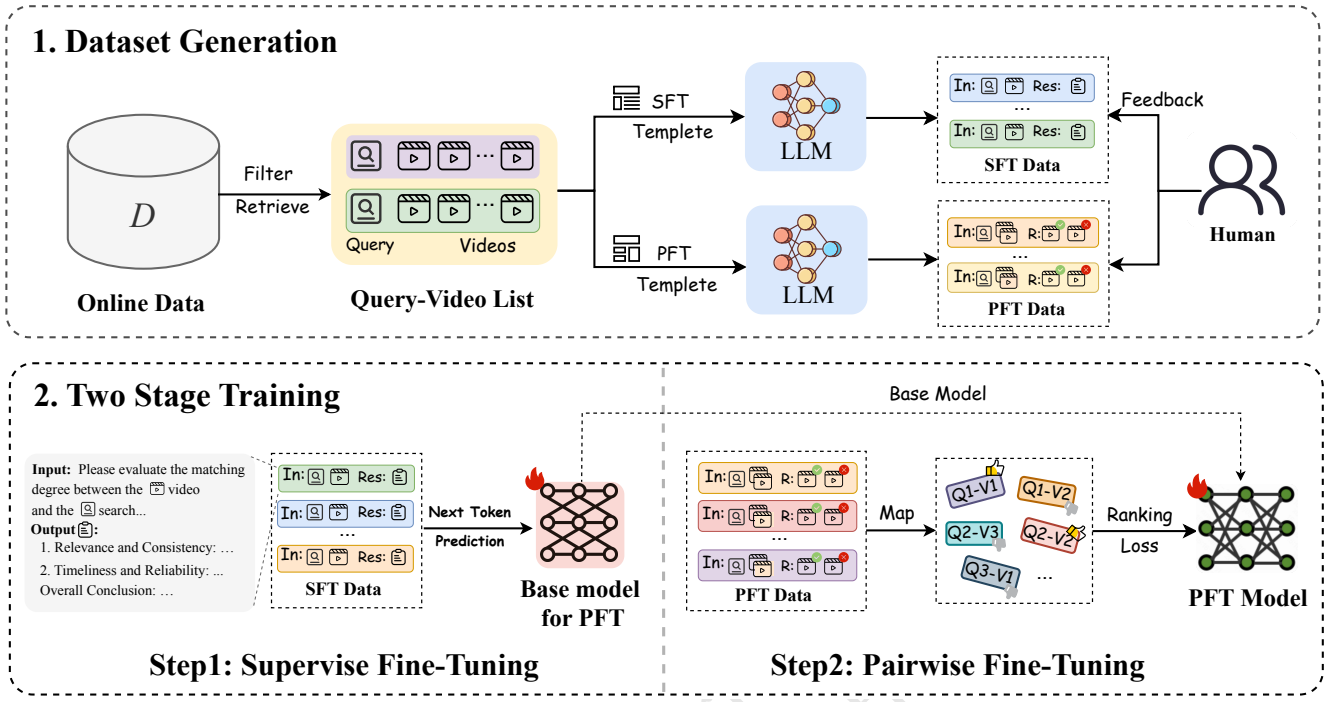


Figure 1: Data construction and two-stage training of the multimodal experience-scoring model, including supervised fine-tuning with multimodal annotations and pairwise fine-tuning with preference data.

is likewise widely adopted. Collectively, multimodal consistency checks and synthetic labels provide the data and modeling capacity needed for sparse, multimodal long-tail retrieval; this requirement is the core motivation for the two-stage training strategy proposed in our work.

3 Methodology

3.1 Framework Overview

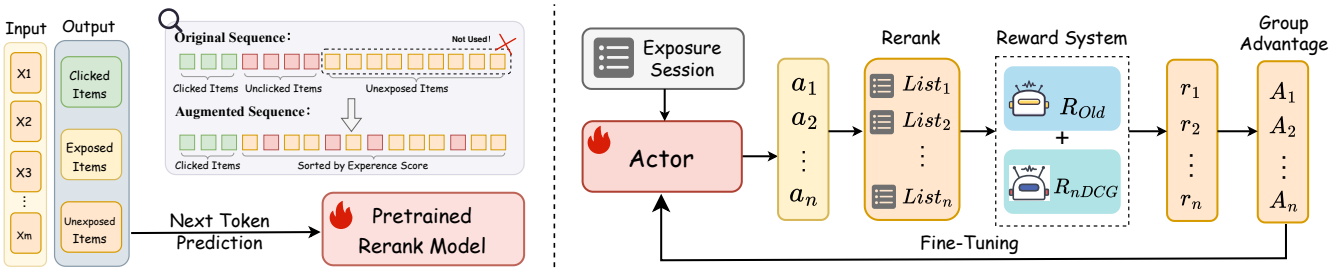
To address the ranking bias and degraded user experience caused by long-tail queries in short-video search, we propose an unbiased multimodal reranking framework driven by large language models. The framework aims to model the alignment between video content quality and user intent based on multimodal evidence, which without relying on user click data, so as to produce comparable and deployable scores that improve ranking quality and user satisfaction.

The overall framework consists of three main stages, where the first two are illustrated in Figure 1 and the third stage in Figure 2. First, in the mutimodal quality alignment stage, we construct a high-quality multimodal annotation dataset. A large language model is prompted to generate and align dimension-wise quality analyses, enabling the model to learn to evaluate content using textual, speech, and visual signals. Second, in the pairwise ranking alignment stage, we introduce intra-query video preference pairs and train the model with a pairwise ranking loss. This allows the model to learn partial orderings among candidates, thereby improving score comparability and mitigating the noise introduced by popularity bias and cross-modal mismatch in behavior-driven models. Finally, in the Integration into the Ranking Pipeline stage, the learned scores are integrated into the production ranking system. Point-wise scores are used to promote high-quality but under-exposed candidates, while sequence-level scores serve as reinforcement learning rewards to optimize page-level ranking strategies and enhance overall user experience.

3.2 Multimodal Quality Alignment

3.2.1 Dataset. Existing open datasets focus mainly on text-only or vision-only retrieval and lack resources that simultaneously cover text, speech, and visual signals for long-tail short-video search. Consequently, models cannot directly learn fine-grained judgments of cross-modal consistency or multi-dimensional content quality. To bridge this gap, we design an LLM-based multi-dimensional annotation pipeline. After obtaining the annotated data, we use it to fine-tune the backbone LLM, effectively enhancing the model's multimodal quality perception and explanatory capabilities.

Specifically, we begin by mining search logs from the past month and filtering out long-tail queries whose seven-day page view (PV) count is below 70. After removing noisy queries, we randomly sample a subset of long-tail queries and retrieve up to ten candidate videos for each query, forming a collection of ⟨query, video⟩ pairs. For each video, we collect multimodal inputs, including textual features (title, description, OCR snippets, and the full ASR transcript) and visual features (the cover image and four key frames). Leveraging the comprehension and generation capabilities of existing multimodal models, we design a prompting template that guides the model to produce dimension-wise reasoning analyses for each



(a) Stage1: Pre-Training Enhancement **(b) Stage II: Page-Level Alignment**

Figure 2: Overview of the two-stage reranking integration. Stage I enhances pre-training with experience-score-ordered supervision; Stage II aligns page-level ranking through GRPO-based reward optimization guided by experience-derived nDCG.

sample across multiple axes, such as relevance and consistency, image safety and age appropriateness, as well as timeliness and credibility. To ensure annotation reliability, we manually spot-check a subset of samples, analyze the causes of errors, and iteratively refine the evaluation dimensions in the prompt. This annotation process yields multidimensional quality assessment data, which serve as supervision signals for subsequent model fine-tuning.

3.2.2 Supervised Fine Tuning. We cast multimodal quality assessment as a standard next-token prediction task for an LLM. Given a multimodal prompt x, which concatenates a fixed system prefix, the user query, and fused textual–audio–visual descriptions, the model is trained to autoregressively generate a complete quality analysis y. Let y consist of T tokens; the conditional likelihood factorises as

$$p_\theta(y \mid x) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, x). \tag{1}$$

The training objective minimises the negative log-likelihood

$$\mathcal{L}_{\text{SFT}} = -\mathop{\mathbb{E}}_{\langle x,y\rangle \in \mathcal{D}_{\text{SFT}}} \Big[\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, x)\Big], \tag{2}$$

where $\mathcal{D}_{\text{SFT}}$ is built by prompting LLM to produce multiple dimension analyses for each ⟨query, video⟩ pair. Because the system-prefix tokens are identical across samples, their prediction losses are masked out during optimisation, and only the analysis portion contributes to $\mathcal{L}_{\text{SFT}}$. After this stage, the model acquires the ability to assess content quality from multiple perspectives conditioned on multimodal evidence.

3.3 Pairwise Ranking Alignment

3.3.1 Dataset. In the pairwise ranking alignment stage, we construct multiple candidate video pairs $\langle d_i, d_j \rangle$ for each long-tail query q based on its exposed videos. A comparison-style prompt template is designed to present the multimodal information of both videos to the large language model and to request a preference decision according to the quality dimensions defined in the first stage. This process produces comparable pairwise preference annotations, enabling the model to learn the relative quality ranking of videos under the same query. To ensure the accuracy and consistency of the annotations, we perform multi-level quality control after labeling. Specifically, we combine rule-based filtering with manual verification to identify and correct partial-order conflicts among videos under the same query, ensuring logical consistency and overall reliability of the final preference labels.

3.3.2 Pairwise Preference Fine Tuning. Building upon the SFT model, we replace the generative head with a *sequence classification head* to output a comparable scalar quality score. Given a query–video pair (q, v), the multimodal backbone first encodes the fused representation $h_{q,v} = \text{Mo}(q, v)$, and the quality score is obtained through a linear projection:

$$s_{q,v} = f_\theta(q, v) = w^\top h_{q,v} + b, \tag{3}$$

where w and b are learnable parameters. For a pair of candidate videos (A, B) under the same query, with annotated preference $A \succ B$, we denote $s^+ = s_{q,A}$, $s^- = s_{q,B}$. The model is trained by minimizing the following objective:

$$\mathcal{L} = -\mathbb{E}\big[\log \sigma(s^+ - s^-)\big] + \lambda\, \mathbb{E}\big[(s^+ + s^-)^2\big], \tag{4}$$

where the first term encourages the model to assign higher scores to preferred videos, and the second term regularizes the overall score distribution to remain centered and stable. Here, $\sigma(\cdot)$ denotes the Sigmoid function, and λ is a balancing coefficient. This unified formulation jointly optimizes relative ranking consistency and score stability, allowing the model to produce comparable point-wise quality scores $f_\theta(q, v)$ for any query–video pair (q, v), which can be directly applied to downstream reranking and evaluation of long-tail queries.

3.4 Integration into the Ranking Pipeline

We integrate the learned experience scores into a two-stage generative reranking pipeline for short-video search to alleviate long-tail bias. In *Stage I – Pre-training Enhancement*, we rewrite the training target sequence used in the autoregressive next-item prediction objective, aligning it with the order induced by the experience scores. This guides the model to prioritize high-quality content during training. In *Stage II – Page-level Alignment*, we construct a page-level consistency signal based on the experience scores and treat it as a label-derived ideal ordering. The sequence policy is then optimized through GRPO, an advantage-weighted reinforcement learning algorithm[21]. The reward function combines the existing behavioral objective with a normalized DCG metric computed against the experience-induced ideal list, thereby preserving

business performance while improving page-level consistency and user experience.

3.4.1 Stage I: Pre-Training Enhancement. For a query q, let $\mathcal{S}$ denote the candidate set for a session and $s_{\text{exp}}(q,i) \in [0,1]$ the point-wise experience score of item $i \in C$. Let C be the set of clicked items, E the set of exposed but not clicked items, and $U = \mathcal{S} \setminus (C \cup E)$ the set of *unexposed* items. We write $\text{sort}(S;\, s_{\text{exp}}\downarrow)$ for sorting a set S in descending order by s_{exp}. We construct the training target sequence by concatenating two segments:

$$\mathbf{y}_{\text{aug}} = \big[\text{sort}(C;\, s_{\text{exp}}\downarrow) \triangleright \text{sort}(E \cup U;\, s_{\text{exp}}\downarrow)\big]. \tag{5}$$

That is, all clicked items are placed in front and internally ranked by s_{exp}, while the exposed-but-not-clicked and unexposed items are merged and jointly sorted by the same score. This preserves the business constraint "clicked-first" while allowing high-quality, previously unexposed items to appear early in the supervision sequence. The loss remains the standard autoregressive next-token (next-item) negative log-likelihood:

$$\mathcal{L}_{\text{stage1}} = \sum_{t=1}^{|\mathbf{y}_{\text{aug}}|} -\log \pi_\theta(y_t \mid q,\ y_{<t},\ C). \tag{6}$$

No auxiliary terms or structural changes are introduced in Stage I; only the target sequence is replaced by (5).

3.4.2 Stage II: Page-Level Alignment. From the learned experience scores, we derive an ideal page-level ranking order that serves as the reference for supervision. Let C denote the candidate set for a given query, and $s_{\text{exp}}(i)$ the experience score assigned to item $i \in C$. We sort the candidates in descending order of their experience scores to obtain the ideal ranking list:

$$\mathbf{y}^s = \text{sort}(C;\, s_{\text{exp}}\downarrow). \tag{7}$$

The position of item i in this ideal list is denoted as $\text{rank}_{\mathbf{y}^s}(i)$. Based on this order, we define a graded relevance value for each item as:

$$\text{rel}_{\mathbf{y}^s}(i) = K - \text{rank}_{\mathbf{y}^s}(i) + 1, \tag{8}$$

where K is the maximum list length. This assignment ensures that items ranked higher in $\mathbf{y}^s$ receive larger scores. Given a generated list $\mathbf{y} = (y_1, \ldots, y_K)$ produced by the model, we compute its discounted cumulative gain (DCG) and normalized DCG (nDCG) with respect to the ideal ranking $\mathbf{y}^s$ as:

$$\begin{aligned} \text{DCG}_{@K}(\mathbf{y} \mid \mathbf{y}^s) &= \sum_{t=1}^{K} \frac{\text{rel}_{\mathbf{y}^s}(y_t)}{\log_2(1+t)}, \\ \text{nDCG}_{@K}(\mathbf{y}, \mathbf{y}^s) &= \frac{\text{DCG}_{@K}(\mathbf{y} \mid \mathbf{y}^s)}{\text{DCG}_{@K}(\mathbf{y}^s \mid \mathbf{y}^s)}. \end{aligned} \tag{9}$$

Here, the numerator measures the quality of the generated ranking, while the denominator normalizes the score using the ideal order, thus constraining $\text{nDCG}_{@K}$ within the range $[0,1]$.

To align the model's generation policy with this page-level consistency, we design a GRPO-style sequence-level reward that interpolates the existing behavioral objective with the experience-based page utility:

$$r(q,\mathbf{y}) = \alpha\, R_{\text{old}}(q,\mathbf{y}) + \beta\, \text{nDCG}_{@K}(\mathbf{y}, \mathbf{y}^s), \tag{10}$$

where q denotes the search query, $R_{\text{old}}(q,\mathbf{y})$ represents the pre-existing reward computed according to business rules or models, and $\alpha, \beta > 0$ are mixing coefficients that control the trade-off between behavioral and page-level objectives. The policy π_θ is then optimized using GRPO updates over trajectory-level returns defined by Eq. 10. Intuitively, the first term preserves historical behavioral performance, while the second term drives the policy toward consistency with the experience-derived ideal ordering (Eq. 7), thereby mitigating click-bias amplification and improving long-tail experience.

4 Experiments

This section reports offline experiments only. Under a setup that mirrors real short-video search and emphasizes long-tail queries, we fix recall and ranking and evaluate only the reranking stage. We then present the experimental setup and results: first verifying the effectiveness of the two-stage experience-scoring model, then assessing its deployment effect in reranking, followed by minimal ablation analyses as needed.

4.1 Experimental Setup

4.1.1 Dataset. Before model training, we constructed a large-scale annotated dataset with strict temporal splitting and rule-based deduplication to ensure the reliability and reproducibility of our experiments. During the Supervised Fine-Tuning (SFT) stage, we collected and generated approximately 187,000 <Query, Video> samples from search logs within an earlier time window, covering 31,392 long-tail queries with an average of 5.96 videos per query. Long-tail queries are defined as those with fewer than 70 page views within any consecutive seven-day window. The training, validation, and test sets are strictly disjoint at both the query and video levels, with all queries normalized before splitting. In the subsequent Pairwise Preference Fine-Tuning (PFT) stage, we further expanded the dataset to 348,000 cleaned preference pairs, spanning 42,308 long-tail queries with an average of 8.22 pairs per query.

Table 1: Statistics of the Training Datasets Used for SFT and PFT Stages.

Dataset	Sample Size	Queries	Avg. per Query
SFT Data	187,150	31,392	5.96
PFT Data	347,935	42,308	8.22

4.1.2 Metrics. We evaluate the proposed experience score model and its downstream reranking framework entirely through offline metrics that capture ranking accuracy, listwise quality, and perceptual consistency. Specifically, we first assess the model's intrinsic ability to distinguish preferred items using Pairwise Accuracy (Acc) or AUC on a balanced preference test set, which measures whether the model consistently assigns higher scores to preferred videos; under this balanced setup, PairAcc is equivalent to AUC and is reported as accuracy (%). We further employ NDCG@K as the primary metric, to evaluate listwise ranking quality by verifying whether high-quality items are ranked toward the top. In addition,

Table 2: Comparison of baseline methods on the human-labeled query set: NDCG@{1,5,10}.

Method	NDCG@1	NDCG@5	NDCG@10
GPT-4o (T-P)	0.793	0.824	0.905
GPT-4o (VL-P)	0.811	0.837	0.921
GPT-4o (T-L)	0.687	0.752	0.885
RankGPT	0.759	0.801	0.904
BGE-m3	0.612	0.721	0.864
ExpModel(Ours)	**0.849**	**0.854**	**0.930**

a human preference evaluation (GSB) provides a perceptual measure of quality, where annotators compare the model-generated ranking against a baseline as Good, Same, or Bad, and we summarize the advantage rate to quantify net human preference. Beyond evaluating the model itself, we also examine its contribution when incorporated into the two-stage reranking pipeline, applying the same offline metrics on the reranked results. This setup allows us to determine whether the experience model improves overall ordering consistency and page-level quality. Together, these evaluations provide a comprehensive view of the model's capacity to produce human-aligned experience scores and its effectiveness as a scoring component in the reranking stage.

4.1.3 Baselines. We evaluate our method against several categories of reference baselines under a controlled zero-shot setting. The baselines fall into three groups: (i) API-based large language model references, (ii) listwise rerankers based on LLM prompting, and (iii) dual-encoder retrieval models. All methods operate on the same per-query candidate pool with matched input budgets, including truncated ASR/OCR text, a fixed number of visual frames or cover images, and identical evaluation lists. This ensures that performance differences reflect genuine ranking quality rather than variations in recall. All external LLMs are kept strictly in zero-shot mode. This zero-shot configuration mirrors realistic deployment settings of API-based models, providing a capacity-oriented reference rather than a task-optimized competitor.

API Models. For the API-based LLM references, we employ several GPT-4o[1] configurations to measure capacity-oriented zero-shot performance. Specifically, **GPT-4o (T-P)** denotes the text-based pointwise setting, where GPT-4o receives each query along with the candidate's textual metadata and outputs a scalar score for that pair; the ranked list is obtained by sorting candidates in descending order of these scores. **GPT-4o (T-L)** refers to the text-based listwise variant, which processes the entire candidate set within a single prompt and directly outputs a global ranking order. **GPT-4o (VL-P)** indicates the multimodal pointwise variant that extends the same procedure by additionally providing visual cues such as cover images and key frames, serving as an approximate upper bound for zero-shot multimodal ranking.

Listwise Rerankers. The listwise reranking baseline, RankGPT [22] performs zero-shot listwise permutation generation: the LLM is prompted to output an ordering of a group of candidates; due to context limits, RankGPT applies a sliding-window strategy that reranks the tail window first and then moves back-to-first, merging window-level permutations into a final list.

[1]https://platform.openai.com/docs/models/gpt-4o

Table 3: Performance of integrate the experience scores into two-stage reranking pipeline on the human-labeled query set: NDCG@{1,5,10}.

Model	NDCG@1	NDCG@5	NDCG@10
Exposure Seq	0.707	0.754	0.884
CTR score	0.610	0.687	0.854
Relevance score	0.611	0.722	0.866
Quality score	0.564	0.673	0.843
Base rerank	0.763	0.794	0.891
Base + S1	0.782	0.816	0.903
Base + S1&S2	0.785	0.822	0.910
ExpModel	**0.849**	**0.854**	**0.930**

Dual-Encoder Retrieval Models. The dual-encoder BAAI/bge-m3(dense mode)[4] represents a traditional retrieval paradigm. It encodes both the query and each candidate's textual fields, computes cosine similarity between normalized embeddings, and sorts candidates by similarity in descending order to obtain the final ranked list.

4.2 Evaluation Result

4.2.1 Effectiveness of the experience-scoring model. We first assess the ranking capability of the two-stage experience-scoring model (ExpModel) on the human-labeled query set. As shown in Table 2, ExpModel achieves the best NDCG@1,5,10 scores, outperforming baselines across all cutoffs. The gain @1 is particularly notable, indicating stronger discrimination among top-ranked items and more effective promotion of high-quality, query-consistent results.

Comparisons show that zero-shot GPT-4o variants have strong generalization but lack task-aligned, explicit quality modeling for short-video search. RankGPT is single-modal and cannot exploit visual evidence, which limits its performance. BGE-m3 emphasizes textual relevance and is less expressive for cross-modal consistency and content quality. Our two-stage approach aligns multimodal evidence, learns pairwise preferences for query-aligned quality, yielding more robust gains across cutoffs, particularly at the top ranks.

4.2.2 Effectiveness of reranking deployment. To assess the effect of injecting the two-stage experience score into reranking on user experience and consumption behavior, we conduct two complementary experiments using Tables 3 and 4. First, in Table 3, we keep recall and ranking fixed and vary only the reranking signals and strategies: *Base*, *Base+S1* which adds the pre-training enhancement, and *Base+S1&S2* which adds the page-level alignment; we report listwise experience metrics such as NDCG@{1,5,10} to test whether the pre-training enhancement and the page-level alignment yield perceptible top-of-page gains. Second, in Table 4, under the same reranking settings, we evaluate business-critical consumption proxies, including long-play ratio and click metrics, to determine whether experience optimization produces only controllable substitution effects on consumption behavior.

Table 4: Evaluation of models on long-play and click-through metrics.

Model	Long-play		Click	
	AUC	NDCG@10	AUC	NDCG@10
Exposure Seq	0.595	0.879	0.612	0.883
CTR Score	0.762	0.918	0.863	0.920
Relevance score	0.587	0.797	0.609	0.805
Quality score	0.572	0.829	0.663	0.830
Base Rerank	**0.696**	**0.877**	**0.680**	**0.898**
Base + S1	0.694	0.872	0.679	0.883
Base + S1&S2	0.691	0.866	0.675	0.879
ExpModel	0.654	0.863	0.673	0.868

In Table 3, when the reranking side is progressively enhanced from *Base* to *Base+S1* and *Base+S1&S2*, the experience metrics exhibit a monotonic improvement. This indicates that the two-stage training yields a comparable and de-biased pointwise experience score that consistently pushes high-quality, query-consistent content toward more prominent positions, while the page-level alignment further optimizes the relevance and diversity at the top ranks, with particularly pronounced gains in long-tail scenarios. In contrast, the other signals in the table each have limitations: *CTR score* reflects popularity but remains influenced by historical behavior and position/exposure propensity; *Quality score* emphasizes content quality but lacks an explicit constraint for query alignment, so it is difficult to support ranking on its own; *Relevance score* provides strong semantic matching but does not explicitly address exposure/position bias or multimodal inconsistency, and it also lacks page-level governance; *Exposure Seq* denotes the current online ordering visible to users, determined jointly by base rerank and downstream pipeline strategies, and it tends to preserve the existing exposure structure rather than actively upgrading items to the very top. Using the experience model directly for ranking (*ExpModel*) achieves the best overall results, which reflects the upper bound of the two-stage model capacity and validates the effectiveness of injecting its scores into the reranker.

In the offline evaluation of Table 4, with recall and ranking held fixed, the long-play and click metrics for *Base*, *Base+S1*, and *Base+S1&S2* remain in a similar range. After injecting the experience scores into pre-training lables and adding page-level alignment, we observe only a slight and controllable decline relative to *Base*, with no significant degradation in consumption. This pattern aligns with the gains on the experience side: the page prioritizes higher-quality, query-consistent content, which moderately reduces reliance on historical clicks and exposure tendencies and thus introduces small, acceptable fluctuations in consumption.

4.2.3 Human Preference Evaluation. We assess user satisfaction via GSB (Good/Same/Bad) pairwise judgments against the base rerank. As shown in Table 5, using the two-stage experience model directly for ranking (ExpModel) yields a strong net advantage (Adv +11.7%), representing an upper bound of model capacity. When its scores are injected into the production reranker with page-level alignment (Base + S1&S2), a clear positive advantage remains (Adv +5.5%) with a stable proportion of "Same," indicating that the model's scoring ability translates into tangible page-level gains, while the magnitude is moderated by deployment constraints. This aligns with our offline metrics: experience gains are stable and the perturbation to behavior is controlled.

Table 5: Human GSB evaluation vs. base rerank. Annotators compare each method's top-K list with the base and label Good/Same/Bad. Adv = $(G-B)/(G+S+B)\times 100\%$.

Method	Good	Same	Bad	Adv
ExpModel	48	114	26	**+11.70%**
Base + S1&S2	39	133	28	+5.50%

4.3 Ablation study

4.3.1 Impact of Two-stage training. We first assess the necessity of two–stage training by comparing *pairwise fine–tuning only* with the complete *SFT → PFT* pipeline under both text–only and multimodal settings. With identical model capacity and data, inserting the supervised fine–tuning stage raises ACC from 79.9% to 80.7% on the text model and from 82.2% to 84.6% on the multimodal model, demonstrating that evidence-grounded supervision provides a stronger foundation upon which pairwise preference learning can build.

4.3.2 Impact of Modalities and Training Strategy on Reranking. As shown in Figure 3 (a) and (b), at either stage the text-only input underperforms the multimodal setting (text + vision), with a larger gap on experience-oriented metrics such as *human*. On behavior-driven proxies (*click*, *long*, *dur*), the difference is smaller. This indicates that visual cues complement textual signals, enhance query–content consistency, and help surface high-quality items at the top ranks. The pattern is consistent in both pre-training enhancement and page-level alignment.

As shown in Figure 3 (c) and (d), training with PFT only is overall weaker than the two-step scheme SFT then PFT; the advantage is more stable on *human* and long-play related metrics. Establishing a comparable, de-biased base score with SFT and then refining the relative order with PFT better captures the notion of "query-aligned quality." This conclusion holds across both stages.

5 Online Applications

5.1 Deployment Details

To validate the feasibility and effectiveness of our proposed experience score–driven two-stage reranking framework in real-world search scenarios, we deployed the method in Kuaishou's production search system and conducted a large-scale A/B test. The experiment targeted the long-tail query segment, defined as queries with fewer than 70 page views over a 7-day window. We sampled 5% of total search traffic for the test bucket, within which long-tail queries accounted for approximately 15.4%. The experiment lasted two weeks and handled over 50 million queries per day, ensuring statistical significance and broad representativeness.

The reranking model was implemented using the TensorFlow framework, with a compact architecture of only 4 million parameters, enabling efficient online inference and smooth deployment.

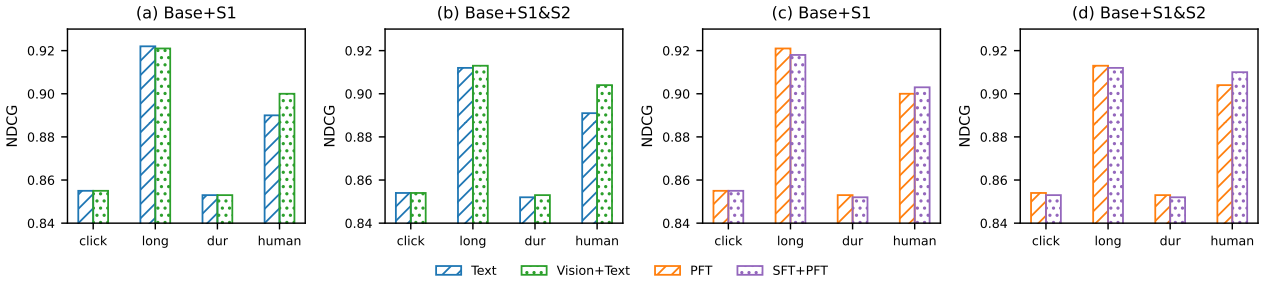


Figure 3: Ablations within the two-stage optimization in reranking. Bars report NDCG under four label types: click: click-based labels, long = long-play labels, dur = watch-duration labels, human = human-preference labels.

Table 6: Accuracy of different training schemes and modalities.

Scheme	Modality	Model	Accuracy (%)
PFT	text	Qwen3-4B	79.9
SFT + PFT	text	Qwen3-4B	80.7
PFT	text+image	Qwen2.5-VL-3B	82.2
SFT + PFT	text+image	Qwen2.5-VL-3B	84.6

Online inference was performed on 2 nodes, each equipped with 2 NVIDIA A10 GPUs. For training, the model was optimized using our constructed multimodal long-tail annotation dataset, on a distributed setup with 6 nodes, each using 2 Tesla T4 GPUs (16 GB). The first stage employed supervised fine-tuning, with a per-node batch size of 1024, using the AdamW optimizer and a learning rate of 5×10^{-6}. The second stage incorporated a page-level reinforcement learning strategy based on GRPO, using 64-sample rollouts and the same learning rate. The training pipeline followed a batched update schedule, with model weights refreshed every 12 hours to adapt to evolving data distributions.

5.2 Performance of A/B Test

As shown in Table 7, the online A/B test results confirm the effectiveness of our experience score–driven two-stage reranking framework in improving both user experience and ranking quality. In the long-tail query segment, all three core metrics show significant gains: the IQRR rate dropped by 1.28%, indicating that users were more satisfied with the first-page results and less likely to reformulate their queries; the CTR increased by 1.24%, suggesting better surfacing of relevant candidates; and the LVR rose by 1.67%, reflecting higher content quality and user retention. These improvements demonstrate that our model can reliably identify and promote high-quality videos even in low-feedback scenarios. Importantly, these benefits also generalized to overall traffic. Across all queries, the IQRR rate decreased by 0.11%, while CTR and long-view ratio increased by 0.13% and 0.19%, respectively. Although the absolute gains are smaller, they remain statistically meaningful given that long-tail queries account for only 15.4% of total volume. This suggests that improving long-tail ranking not only enhances sparse-query performance but also contributes to global ranking

Table 7: Online A/B test results on Kuaishou Search. "All queries" refers to the entire test bucket; "Long-tail queries" are defined as those with 7-day PV < 70. IQRR denotes the *Intent Query Reformulation Rate*, CTR is the *Click-Through Rate*, and LVR is the *Long-View Ratio*.

Metric Name	IQRR	CTR	LVR
all queries	-0.11%	+0.13%	+0.19%
long-tail queries	-1.28%	+1.24%	+1.67%

stability without disrupting consumption behavior. Notably, the increase in online consumption metrics is not at odds with the slight declines observed in offline evaluation: offline labels are derived from historical logs with a fixed page composition and thus measure a static substitution effect. After deployment, the experience score places higher-quality, query-consistent results more prominently, which dynamically reshapes users' browsing paths and viewing intent, thereby translating into a natural overall rise in consumption.

6 Conclusion

This paper presents an unbiased multimodal reranking framework for long-tail short-video search. By combining multimodal evidence–aligned supervised fine-tuning with pairwise preference optimization, the model learns a comparable and de-biased experience score that captures both query alignment and content quality. Injecting this score into the production reranker, together with pointwise enhancement and page-level alignment, yields stable improvements in top-of-page experience metrics without disrupting consumption behavior. Offline results show consistent gains and confirm the model's ability to identify high-quality, query-consistent content. Large-scale online A/B tests further validate the deployability and scalability of the framework, achieving statistically significant improvements in both user experience and consumption behavior. For future work, we plan to further refine the two-stage training mechanism by exploring direct listwise modeling to better capture holistic ranking dependencies. We also aim to incorporate richer audio semantics and session-level feedback, extending the framework's generalization and stability to more diverse, multimodal, and long-tail retrieval scenarios.

References

[1] Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. 2025. How Good are LLM-based Rerankers? An Empirical Analysis of State-of-the-Art Reranking Models. arXiv:2508.16757 [cs.CL]

[2] Yong Bai, Rui Xiang, Kaiyuan Li, Yongxiang Tang, Yanhua Cheng, Xialong Liu, Peng Jiang, and Kun Gai. 2025. Chime: A compressive framework for holistic interest modeling. *arXiv preprint arXiv:2504.06780* (2025).

[3] Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. 2024. Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling. *arXiv preprint arXiv:2409.12740* (2024).

[4] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216* (2024).

[5] Peter Baile Chen, Tomer Wolfson, Michael Cafarella, and Dan Roth. 2025. EnrichIndex: Using LLMs to Enrich Retrieval Indices Offline. *arXiv preprint arXiv:2504.03598* (2025).

[6] Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024. On softmax direct preference optimization for recommendation. *Advances in Neural Information Processing Systems* 37 (2024), 27463–27489.

[7] Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439* (2024).

[8] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755* (2022).

[9] Mathew Jacob, Erik Lindgren, Matei Zaharia, Omar Khattab, and Andrew Drozdov. 2025. Drowning in Documents: Consequences of Scaling Reranker Inference. *arXiv preprint arXiv:2411.11767* (2025).

[10] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Model. User Adapt. Interact.* 34, 5 (2024), 1777–1834. doi:10.1007/S11257-024-09406-0

[11] Haoran Li, Zhiming Su, Junyan Yao, Enwei Zhang, Yang Ji, Yan Chen, Kan Zhou, Chao Feng, and Jiao Ran. 2025. Semi-Supervised Synthetic Data Generation with Fine-Grained Relevance Control for Short Video Search Relevance Modeling. *arXiv preprint arXiv:2509.16717* (2025).

[12] Yueyang Liu, Jiangxia Cao, Shen Wang, Shuang Wen, Xiang Chen, Xiangyu Wu, Shuang Yang, Zhaojie Liu, Kun Gai, and Guorui Zhou. 2025. LLM-Alignment Live-Streaming Recommendation. *arXiv preprint arXiv:2504.05217* (2025).

[13] Sebastian Lubos, Alexander Felfernig, and Markus Tautschnig. 2024. An overview of video recommender systems: state-of-the-art and research issues. *Frontiers Big Data* 6 (2024). doi:10.3389/FDATA.2023.1281614

[14] Wajiha Naveed, Zartash Afzal Uzmi, and Zafar Ayyub Qazi. 2025. Thumbnail-Truth: A Multi-Modal LLM Approach for Detecting Misleading YouTube Thumbnails Across Diverse Cultural Settings. *arXiv preprint arXiv:2509.04714* (2025).

[15] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* 6, 2 (2019).

[16] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. *arXiv preprint arXiv:1910.14424* (2019).

[17] Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. 2024. Large Language Model based Long-tail Query Rewriting in Taobao Search. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw (Eds.). ACM, 20–28. doi:10.1145/3589335.3648298

[18] Jakub Podolak, Leon Perić, Mina Janičijević, and Roxana Petcu. 2025. Beyond reproducibility: Advancing zero-shot llm reranking efficiency with setwise insertion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3205–3213.

[19] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088* (2023).

[20] Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. 2025. Self-calibrated listwise reranking with large language models. In *Proceedings of the ACM on Web Conference 2025*. 3692–3701.

[21] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[22] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542* (2023).

[23] Yihang Sun, Tao Feng, Ge Liu, and Jiaxuan You. 2025. Premium: Llm personalization with individual-level preference feedback. (2025).

[24] Jianling Wang, Yifan Liu, Yinghao Sun, Xuejian Ma, Yueqi Wang, He Ma, Zhengyang Su, Minmin Chen, Mingyan Gao, Onkar Dalal, et al. 2025. User Feedback Alignment for LLM-powered Exploration in Large-scale Recommendation Systems. *arXiv preprint arXiv:2504.05522* (2025).

[25] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1288–1297. doi:10.1145/3404835.3462962

[26] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. 2021. Rich Features for Perceptual Quality Assessment of UGC Videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 13435–13444. doi:10.1109/CVPR46437.2021.01323

[27] Haoyang Wen, Honglei Zhuang, Hamed Zamani, Alexander Hauptmann, and Michael Bendersky. 2024. Multimodal reranking for knowledge-intensive visual question answering. *arXiv preprint arXiv:2407.12277* (2024).

[28] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. 2025. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 2815–2826.

[29] Gengyuan Zhang, Mang Ling Ada Fok, Jialu Ma, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. 2025. Localizing Events in Videos with Multimodal Queries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 3339–3351. doi:10.1109/CVPR52734.2025.00317

[30] Jiacong Zhou, Xianyun Wang, and Jun Yu. 2024. Optimizing Preference Alignment with Differentiable NDCG Ranking. *arXiv preprint arXiv:2410.18127* (2024).

[31] Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034* (2025).

[32] Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 38–47.

A Prompt Template

A.1 Prompt Template for SFT Data Generation.

Task: Judge the match between the search query and the video, and the overall viewing experience, using the provided evidence (query, title, cover, key frames, ASR, OCR).

Instructions:
Analyze each dimension below. If evidence is missing or inconclusive, write *"Insufficient evidence"* and make a conservative judgment. Keep reasoning concise and tied to the cited evidence (e.g., refer to title/ASR/OCR/frame).

Evaluation dimensions & rules:
- Relevance & Consistency: Check whether title, cover, key frames, and ASR consistently address the query intent...
- Image Safety & Age-Appropriateness: Inspect cover and key frames for lowbrow/violent/gory/sexual/illegal elements or suggestive thumbnails that mislead...
- Timeliness & Trustworthiness: If the query is time-sensitive, judge whether content appears outdated or conflicts with common facts...
- xxx: ...

Output format:
- Relevance & Consistency: xxx
- Image Safety & Age-Appropriateness: xxx
- Timeliness & Trustworthiness: xxx
- xxx...
- Overall verdict: Summarize the match between the query and the video and the expected user experience.

Input:
– Search query: {search_term}
– Video info: {video_info}

A.2 Prompt Template for PFT Data Generation.

Task: Given a search query and two candidate videos (A and B), compare their overall match to the query and expected user experience, using all available multimodal evidence (title, cover, key frames, ASR, OCR). Decide which video provides a better experience for the query.

Instructions:
Analyze both videos along the following dimensions, referencing specific evidence when possible (e.g., title, ASR/OCR content, visual cues). If evidence is missing or unclear, write *"Insufficient evidence"* and make a conservative judgment. Keep reasoning concise and comparative. At the end, provide a clear verdict choosing the preferred video.

Evaluation dimensions & rules:
- Relevance & Consistency: Check whether title, cover, key frames, and ASR consistently address the query intent...
- Image Safety & Age-Appropriateness: Inspect cover and key frames for lowbrow/violent/gory/sexual/illegal elements or suggestive thumbnails that mislead...
- Timeliness & Trustworthiness: If the query is time-sensitive, judge whether content appears outdated or conflicts with common facts...
- xxx: ...

Output format:
- Overall Experience: [A better / B better / Tie] + brief reasoning
- Final Verdict: Preferred video = A / B.

Input:
– Search query: {search_term}
– Video A info: {video_A_info}
– Video B info: {video_B_info}

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009