

Query-aware Multi-modal based Ranking Relevance in Video Search

Chengcan Ye^{1*} Ting Peng^{1*} Tim Chang^{2*} Zhiyi Zhou¹ Feng Wang¹

Tencent Video, PCG

¹{chengcanye, penneypeng, liuxizhou, feynmanwang}@tencent.com

²timchang2022@163.com

Abstract

Relevance ranking system plays a crucial role in video search on streaming platforms. Most relevance ranking methods focus on text modality, incapable of fully exploiting cross-modal cues present in video. Recent multi-modal models have demonstrated promise in various vision-language tasks but provide limited help for downstream query-video relevance tasks due to the discrepancy between relevance ranking-agnostic pre-training objectives and the real video search scenarios that demand comprehensive relevance modeling. To address these challenges, we propose a **Q**Uery-Aware pre-training model with multi-modal**L**ITy(**Q**UALITY) that incorporates hard-mined query information as alignment targets and utilizes video tag information for guidance. **Q**UALITY is integrated into our relevance ranking model, which leverages multi-modal knowledge and improves ranking optimization method based on ordinal regression. Extensive experiments show our proposed model significantly enhances video search performance.

1 Introduction

Video search has become a prevalent method for users to identify relevant content in response to text queries on video streaming platforms. Relevance ranking is crucial in video search (Pang et al., 2017), as it determines the relevance degree of a video concerning a given query. Pointwise loss (e.g., binary cross-entropy loss), ranking loss (e.g., hinge loss) and **Combined-Pair** loss (a linear combination of pointwise and pairwise loss) (Zou et al., 2021) are commonly used to optimize relevance ranking task. However, these methods fail to balance calibration ability (globally stable prediction with good interpretability) and ranking ability (prediction can lead to a correct ranking) (Sheng et al., 2022). At the same time, the transformer architecture’s recent success (Vaswani et al., 2017) in

computer vision and natural language processing has led to pre-trained language models achieving promising results in retrieval and ranking tasks (Zou et al., 2021; Nogueira et al., 2019; Liu et al., 2021). However, most existing approaches primarily focus on text modality and alternative methods which integrate large-scale Vision-and-Language Pre-training (VLP) models, such as CLIP (Radford et al., 2021) and ALBEF (Li et al., 2021), into video search engines face two key challenges: (1) Images typically align with verbose and detailed video texts, providing limited assistance for modeling matching relationship between visual signals and concise queries in downstream relevance tasks. (2) Most VLP models are trained on single-frame images and texts, neglecting video information such as keyframes and tag data, rendering them unsuitable and inadequate for video search engines.

To address these challenges, we propose a query-aware, multi-modal relevance ranking model for real video search systems within a two-step framework, as depicted in Fig.1.

Query-aware Pre-training Model with Multi-modality. We present a real-world query-aware pre-training model that simultaneously aligns image features with video text features and query features. Additionally, we propose a hard query mining strategy to effectively exploit query knowledge. Inspired by CLIP4CLIP (Luo et al., 2022) and TABLE (Chen et al., 2023), we introduce a local tag-guided attention network to extract features from sequential frames, rather than a single image. To preserve pre-trained knowledge to the greatest extent and accelerate the training process, we employ an adapter-tuning strategy

Ranking Relevance. Following the approach in (Bo et al., 2021), we model relevance ranking under the pre-training and fine-tuning paradigm, utilizing various handcrafted features (e.g., BM25 (Robertson and Walker, 1994), click similarity (Yin et al., 2016), term weight) and pre-trained representations

*Equal contribution

of query and video within a wide and deep network architecture. We enhance ranking performance by incorporating multi-modal knowledge and proposing an ordinal regression based approach for joint optimization of ranking and calibration in relevance prediction.

In summary, this paper makes the following contributions:

- We introduce a novel query-aware pre-training model tailored for real-world applications, aligning image with both title and query. This approach effectively utilizes video modality information and exhibits improved adaptability to downstream tasks.
- We propose an innovative relevance ranking optimization method based on ordinal regression, balancing calibration and ranking abilities effectively.
- We present a novel approach for applying pre-trained VLP models to online relevance ranking tasks in real industrial video search scenarios. Comprehensive offline and online evaluations demonstrate that the proposed techniques significantly enhance relevance ranking performance.

2 Methodology

In this section, we describe the details of our multi-modal-based ranking-relevance approach. The overall architecture of our methodology is illustrated in Fig.1, comprising a query-aware pre-training multi-modal model and a ranking-relevance model that utilizes both visual and textual information.

2.1 Query-aware Pre-training Model with Multi-modality

As illustrated in Fig.1(a), our **QUery-Aware Pre-training Model with Multi-modality**(referred to as **QUALITY**), is composed of a query tower, a video visual tower, and a video text tower, which is an extension of the dual-tower structure of the image-level ALBEF model.

Model Input. Given an input video v and an input query q , we employ a 12-layer visual transformer ViT-B/16 (Dosovitskiy et al., 2020) to encode N frames uniformly sampled from the input video, and a shared 12-layer textual encoder, BERT-base (Devlin et al., 2018), to encode the

title and tags of the input video and the input query. The above frame-level visual encoder and the textual encoder are initially pre-trained using the CLIP approach on industrial video-search log data. To accelerate the training process of the **QUALITY** model and prevent catastrophic forgetting (Sharkey and Sharkey, 1995) of the uni-modal pre-trained encoders, we follow the AdaptFormer (Chen et al., 2022a) method that a trainable and lightweight down-up bottleneck module is added to feed-forward parts of transformer blocks within our pre-trained encoders and meanwhile, all the other parameters within the pre-trained encoders are frozen, significantly reducing trainable parameters and enhancing the training efficiency.

Tag Guidance. Video tags are widely present on video-sharing platforms, which are usually keywords and phrases that facilitate video content understanding. To gain a better understanding of the video content rather than merely relying on low-level visual features, a tag-guided cross-attention network is designed to align semantic information with visual signal. Specifically, given the visual representation generated by visual encoder $\{f_{cls}, f_1, f_2, \dots, f_N\}$ and tag representation generated by textual encoder with M tokens $\{g_{cls}, g_1, \dots, g_M\}$, a 3-layer transformer with 8 cross-attention heads (as displayed in purple color in Fig.1) is used to align visual information with semantic tags, then we retain the tag-guided visual part as $\{v_{cls}, v_1, \dots, v_N\}$ for the subsequent query-awareness computation.

Query-awareness. Previous VLP models such as ALBEF and CLIP4CLIP, have focused on modeling the relationship between visual signals and their corresponding text descriptions. However, in real-world search scenarios, how video content is described and how users express their input queries can differ significantly. Moreover, text descriptions of the video content often fail to summarize the video content adequately. Thus the obtained representations by these methods may offer limited assistance for search tasks. To better adapt to our downstream video-search tasks, we explicitly model the matching relationship between the query tower and vision tower (i.e., video frames) through vision-query contrastive learning (VQC) task, a shared cross-modal cross-attention encoder (as displayed in cadet blue color in Fig.1) and vision-text matching (VQM) task, while also maintain the matching modeling between vision and title towers

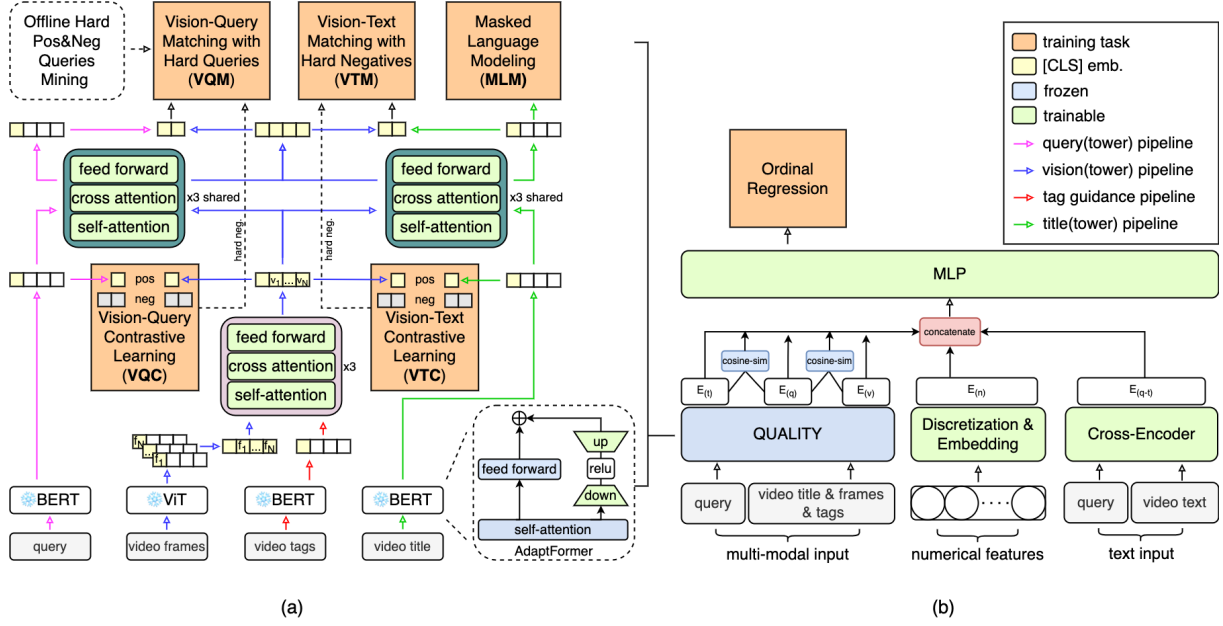


Figure 1: Model architecture. (a) QUALITY model. (b) Multi-modal-based ranking relevance model.

through vision-text contrastive learning (VTC) task, the same shared encoder and vision-text matching (VTM) task. The shared cross-modal encoder is composed of a 3-layer transformer with 8 cross-attention heads. Our query-awareness strategy can alleviate the issue of mismatching purely based on text information in the downstream ranking-relevance task.

2.2 Pretraining Objectives

QUALITY is pre-trained using the following five objectives: Vision-Query Contrastive Learning (VQC) and Vision-Text Contrastive Learning (VTC) applied to uni-modal encoders, as well as Vision-Query Matching (VQM), Vision-Text Matching (VTM), and Masked Language Modeling (MLM) applied to multi-modal encoders. The performance of VQM and VTM is enhanced through online contrastive hard negative mining. Additionally, VQM is further improved by employing offline hard query mining.

Vision-Query Contrastive Learning aims to align the visual signal v_{cls} and query q_{cls} prior to fusion. We define a function $s(v_{cls}, q_{cls}) = h_v(v_{cls})^\top h_q(q_{cls})$ to calculate the similarity between the visual signal and the query. Here, $h_v(\cdot)$ and $h_q(\cdot)$ are linear layers that project the [CLS] embeddings into a shared semantic space and normalize them. We express the vision-query contrastive loss with a trainable temperature parameter τ and batch size B as follows:

$$L_{v2q} = -\frac{1}{B} \sum_i \log \frac{\exp(s(v_{i_{cls}}, q_{i_{cls}}) / \tau)}{\sum_{j=1}^B \exp(s(v_{i_{cls}}, q_{j_{cls}}) / \tau)},$$

$$L_{q2v} = -\frac{1}{B} \sum_i \log \frac{\exp(s(v_{i_{cls}}, q_{i_{cls}}) / \tau)}{\sum_{j=1}^B \exp(s(v_{j_{cls}}, q_{i_{cls}}) / \tau)},$$

$$L_{vqc} = \frac{1}{2} (L_{v2q} + L_{q2v}) \quad (1)$$

Vision-Text Contrastive Learning seeks to align the visual signal v_{cls} and video title t_{cls} prior to fusion. Analogous to the vision-query task, the vision-text contrastive loss with a trainable temperature parameter μ can be defined as follows:

$$L_{v2t} = -\frac{1}{B} \sum_i \log \frac{\exp(s(v_{i_{cls}}, t_{i_{cls}}) / \mu)}{\sum_{j=1}^B \exp(s(v_{i_{cls}}, t_{j_{cls}}) / \mu)},$$

$$L_{t2v} = -\frac{1}{B} \sum_i \log \frac{\exp(s(v_{i_{cls}}, t_{i_{cls}}) / \mu)}{\sum_{j=1}^B \exp(s(v_{j_{cls}}, t_{i_{cls}}) / \mu)},$$

$$L_{vtc} = \frac{1}{2} (L_{v2t} + L_{t2v}) \quad (2)$$

Vision-Query Matching aims to predict whether a pair of vision and query is matched or not. We concatenate the [CLS] embeddings of the visual-text multi-modal encoder, into which the vision and query signals are fed. A fully-connected layer is then employed to generate the two-class probability of matching, denoted as p^{vqm} . The vision-query matching loss can be defined as:

$$L_{vqm} = -\frac{1}{J} \sum_i y_i^{vqm} \log_2(p^{vqm}(v_i, q_i)) \quad (3)$$



Title: 速度与激情系列最佳飙车场景 (Best racing scenes in Fast & Furious saga)
 Tags: 速度与激情, 飙车 (Fast & Furious, Racing)
 Hard positive query: 赛车电影 (Racing movie)
 Hard negative query: 极品飞车 (Need for Speed)

Figure 2: Random sampled video example with key frames, title and tags.

Here, y^{vqm} represents the ground-truth label, and J is the total number of vision-query pairs for this task. In addition to the online hard negative mining strategy employed by ALBEF, an embedding-based offline strategy is also designed to mine both hard positive and negative queries. Specifically, we first derive query and video embeddings from a query-video click graph, utilizing lightweight graph embedding algorithms such as item2vec (Barkan and Koenigstein, 2016) and DeepWalk (Perozzi et al., 2014). Then for a given video, a query is chosen as a hard positive if the cosine-similarity, based on the graph embedding, between the query and this video exceeds a predetermined threshold. Conversely, if the cosine-similarity between them is below the threshold, the query is considered a hard negative. The threshold is an empirical hyperparameter.

Vision-Text Matching aims to predict whether a pair of vision and title originates from the same video. Analogous to the vision-query matching, we define the vision-text matching loss as:

$$L_{vtm} = -\frac{1}{O} \sum_i^O y_i^{vtm} \log_2 (p^{vtm}(v_i, t_i)) \quad (4)$$

Here, p^{vtm} represents the prediction of matching, y^{vtm} is the ground-truth label, and O is the total number of vision-text pairs for this task.

Masked Language Modeling aims to predict masked video title tokens using both visual and textual signals. Video title tokens are randomly masked with a 15% probability and replaced with the special token [MASK]. Let \hat{T} denote the masked token, and $p^{mlm}(I, \hat{T})$ denote the probability of a masked token. We define the masked language modeling loss as:

$$L_{mlm} = -\frac{1}{R} \sum_i^R y_i^{mlm} \log_2 (p^{mlm}(I_i, \hat{T}_i)) \quad (5)$$

Here, R represents the total number of masked tokens, and y^{mlm} is the ground-truth label indicating

whether a token is masked.

The total loss function for our model is:

$$L_{pre} = L_{vqc} + L_{vtc} + L_{vqm} + L_{vtm} + L_{mlm} \quad (6)$$

2.3 Multi-Modal Based Ranking Relevance

2.3.1 Ranking Relevance Model

As illustrated in Fig.1(b), the proposed multi-modal based ranking relevance model comprises four major components: our pre-trained QUALITY which produces query embedding $E_{(q)}$, video textual embedding $E_{(t)}$ and video visual embedding $E_{(v)}$ based on the multi-modal input; a discretization and embedding learning module (Guo et al., 2021) that extracts representation $E_{(n)}$ from hand-crafted numerical features (e.g., BM25, click similarity, term weight); a pre-trained transformer-based cross-encoder that takes query and video text as input, where the video text includes title, actors, uploader name and tags; a multilayer perceptron (MLP) module which produces a relevance score between the query and video. Provided with $E_{(q)}$, $E_{(t)}$ and $E_{(v)}$ generated by QUALITY, we further compute the cosine-similarity of the query embedding with the text or visual embedding to obtain query-text similarity and query-visual similarity, respectively. The cross-encoder is pre-trained following multi-stage training paradigm (Zou et al., 2021) and the representation of the [CLS] token, as well as mean and max pooling of the final layer of the cross-encoder, are concatenated to obtain a presentation of semantic relevance $E_{(q,t)}$. Finally, the concatenation of the outputs of $E_{(q,t)}$, $E_{(q)}$, $E_{(t)}$, $E_{(v)}$, $E_{(n)}$ and the derived query-text similarity and query-visual similarity is fed into MLP module to conduct relevance score between a query and a video.

2.3.2 Ranking Loss Function

The relevance ranker can be considered as a scoring function $f_{\theta}(q, v)$ for a query q and a candidate

video v , and θ denotes the trainable model parameters. In order to ensure both calibration and ranking abilities of the predicted scores, we model the ranking problem as a K -grade ordinal regression problem that accommodates both labeled order $y \in 1, 2, \dots, K$ and a set of thresholds $\rho_1, \dots, \rho_{K-1}$ with the property that $\rho_1 < \rho_2 < \dots < \rho_{K-1}$. Specifically, the final output of the model $f_\theta(q, v)$ is considered as an observed ordinal variable, with its cumulative probability given by the sigmoid function, denoted as σ (Bürkner and Vuorre, 2019). The set of thresholds, which can be optimized during the model training process, divides $f_\theta(q, v)$ into K disjoint segments. In our setting, the probability Pr of relevance k can be formulated as follows:

$$g_k = \sigma(\rho_k - f_\theta(q, v))$$

$$Pr(f_\theta(q, v) = k) = \begin{cases} g_k, & \text{if } k = 1 \\ g_k - g_{k-1}, & \text{if } 1 < k < K \\ 1 - g_k, & \text{if } k = K \end{cases} \quad (7)$$

The corresponding ordinal regression loss function is defined in Equation 8. Besides, A binary cross-entropy loss with binary label $y_b \in \{0, 1\}$, denoted as L_{binary} in Equation 9, is also employed with the purpose of enhancing the differentiation between relevant and irrelevant candidates more accurately. A rating $k \leq K/2$ is considered irrelevant, while a rating $k > K/2$ is deemed relevant. The final ranking loss can be written as Equation 10:

$$L_{ordinal} = -\log(Pr(f_\theta(q, v) = y)) \quad (8)$$

$$L_{binary} = -y_b \log\left(\sum_{k=1}^{K/2} Pr(f_\theta(q, v) = k)\right) + (1 - y_b) \log\left(1 - \sum_{k=1}^{K/2} Pr(f_\theta(q, v) = k)\right) \quad (9)$$

$$L_{final} = \alpha L_{ordinal} + (1 - \alpha) L_{binary} \quad (10)$$

where α is a hyper-parameter that balances the importance of two different loss functions. In order to anchor the predicted probability to a meaningful range, the ranking score is computed as:

$$score = \sum_{k=1}^K \left(\frac{k-1}{K-1}\right) Pr(f_\theta(q, v) = k) \quad (11)$$

3 Experiments

3.1 Datasets

As for training our QUALITY, we construct a dataset consisting of high-quality and diverse

videos sourced from Tencent Video, a prominent Chinese video streaming platform. This dataset contains 10 million videos, including keyframes, video titles, and over 15,000 labeled tags. An example of a video accompanied by hard-mined queries is shown in Fig 2, we explicitly model the matching relationship between vision and concise queries. As for ranking relevance, we manually annotate query-video pairs sampled from video search logs to construct train and test datasets, resulting in a training dataset of 270,000 query-video pairs and a test dataset of over 90,000 items. Annotators judge each query-video pair and assign a label with a relevance grade from **1** to **4**, corresponding to the relevance levels of **Bad**, **Less**, **Good**, **Excellent**, respectively. Apparently, **Excellent** / **Bad** means most relevant / irrelevant video for the given query.

3.2 Evaluation Metrics

We use AUC (Area Under the Curve) and PNR (Positive Negative Ratio) as offline evaluation metrics. For the AUC metric, labels **1** and **2** are considered negative, while labels **3** and **4** are considered positive. The PNR metric considers the partial order between labels and measures the consistency of prediction results and ground truth. As for online evaluation, we employ Average Watch Time (AWT) to quantify user preference on video search results. The **Good vs. Same vs. Bad** (GSB) metric compares two systems in a side-by-side manner, and we utilize ΔGSB (Zou et al., 2021) to assess the satisfaction gain achieved by a new system.

3.3 Offline Performance

We first evaluate the effectiveness of the QUALITY model. Since the original ALBEF is specifically designed for images rather than videos, we have extended it to a video version for a fair comparison, which we refer to as Video-ALBEF. The baseline Video-ALBEF model utilizes a transformer-like pooling strategy inspired by CLIP4CLIP to aggregate keyframe embeddings and then generates a video-level visual embedding. The remaining components remain identical to those in the original ALBEF setup. As shown in Table.1, our method achieves an AUC of 0.683 and a PNR of 2.180, beating the baseline Video-ALBEF model with an absolute 10.5% AUC improvement and a relative 42.2% PNR improvement. Meanwhile, as shown in Table.2, compared to the baseline text-based relevance model, we find that the performance of this baseline can be enhanced by introducing mul-

timodal embeddings via either the method QUALITY or Video-ALBEF, demonstrating the effectiveness of the multi-modal information that can alleviate the issue of mismatching purely based on text information. Furthermore, our method outperforms Video-ALBEF by 0.3% in AUC and 3.2% in PNR respectively, suggesting that explicitly modeling the matching relationship between the query tower and vision tower can help the downstream relevance model.

Table 1: Offline comparison results of multi-modal pre-training models and ablation study of QUALITY. QUALITY outperforms Video-ALBEF and each technical components brings it’s separate gain independently.

Models	AUC	PNR
Video-ALBEF baseline	0.578	1.533
QUALITY	0.683	2.180
(w/o) query tower	0.623	1.752
(w/o) title tower	0.678	2.127
(w/o) hard pos/neg query mining	0.670	2.081
(w/o) tag guidance	0.665	2.046
(w/o) AdaptFormer	0.653	1.881

Table 3 provides the performance of our ordinal regression-based ranking loss. We observe that the proposed ranking loss outperforms the pointwise loss and the **Combined-Pair** ranking loss, by relative improvements of 28.9% and 4.1% on PNR, respectively. We also notice that the pointwise-based model achieves the highest AUC of 0.925, but the lowest PNR of 6.911. This outcome indicates that pointwise loss only focuses on the calibration ability and neglects the ranking ability.

Table 2: Offline comparison results of ranking relevance models and ablation study on technical components of QUALITY.

Models	AUC	PNR
Text-based baseline	0.917	8.425
Text-based + Video-ALBEF	0.918	8.637
Text-based + QUALITY	0.921	8.914
(w/o) query tower	0.920	8.840
(w/o) title tower	0.920	8.853
(w/o) hard pos/neg query mining	0.920	8.819
(w/o) tag guidance	0.920	8.866
(w/o) AdaptFormer	0.919	8.785

3.4 Ablation Study

Effects of Query-awareness. As depicted in Table.1, our QUALITY model achieves an absolute

Table 3: Offline comparison of ranking relevance model performances for different ranking loss functions.

Rank loss	AUC	PNR
Pointwise	0.925	6.911
Combined-Pair	0.918	8.565
Ours	0.921	8.914

6.0% AUC improvement and a relative 24.4% PNR improvement compared to the model without the query tower. Consequently, as shown in Table.2, our model gains improvements of 0.1% on AUC and 0.8% on PNR. We attribute this significant performance boost to two primary factors. First, aligning query and visual signals makes the pre-training task more adaptable to downstream relevance tasks. Second, query information is more concise compared to video titles, increasing the efficacy of contrastive learning due to harder alignment, as evidenced by model without title tower outperforms model without query tower by 5.5% on AUC and 21.4% on PNR in Table.1. We introduce an embedding-based strategy for hard pos/neg query mining in the VQM task, suggesting that it is more effective than the online hard negative mining approach employed by ALBEF. As shown in Table 1, in comparison to the model without hard pos/neg query mining, our strategy yields improvements of 1.3% on AUC and 4.8% on PNR. Consequently, as illustrated in Table 2, our model achieves a PNR improvement of 1.1%.

Effects of Tag Guidance. In our work, we employ tag information modality as explicit guidance. As illustrated in Table.1, our QUALITY model achieves an absolute 1.8% AUC improvement and a relative 6.5% PNR improvement compared to the model without tag guidance. Consequently, as presented in Table.2, our model attains improvements of 0.1% on AUC and 0.5% on PNR. Tag information encapsulates the core entity knowledge of a video, enabling the visual signal to develop a semantic-level understanding of the video, rather than being confined to the low-level visual signal. In summary, incorporating tag information proves beneficial for query-vision relevance tasks.

Effects of AdaptFormer. We employ 12-layer pre-trained uni-modal encoders for efficient training in real-world industry applications and the preservation of pre-trained knowledge. To this end, we utilize an adapter-tuning strategy. As shown in Table.1, our QUALITY model outperforms the fully fine-tuned model by achieving an absolute 3.0% improvement on AUC and a relative 15.9% improve-


Query	Document	Human labeling	Score (w/o QUALITY)	Score (w/ QUALITY)
灰姑娘2 : 美梦成真 (Cinderella 2 Dreams Come True)	 《美梦成真》灰姑娘以为遇上了富二代男友，谁知道小伙是个穷光蛋 ("Dreams Come True" Cinderella thought she had met a rich second-generation boyfriend, who knew that the guy was a pauper)	Bad	0.52	0.25
	 灰姑娘当上王妃，改变皇宫所有规矩 《仙履奇缘2美梦成真》 (Cinderella became the princess and changed all the rules of the palace. "Cinderella 2 Dreams Come True")	Good	0.65	0.67
王鸥主演的惊蛰 (Awakening of Insects starring Wang Ou)	 余小晚意外发现张离的秘密，两人重归于好，确定是真爱了 (Yu Xiaowan accidentally discovered Zhang Li's secret, the two got back together, and it was definitely true love)	Good	0.36	0.71
	 王鸥早起素颜曝光，镜头拉进那刻颜值太抗打，哪里像是37岁 (Wang Ou woke up early and exposed without makeup. The moment the camera was pulled in, his appearance was too resistant. He didn't look like 37 years old)	Less	0.283	0.296

Figure 3: Cases of video search. "score (w/o QUALITY)" / "score (w/ QUALITY)" represents the prediction score of relevance ranking model with / without QUALITY.

ment on PNR. We suggest that freezing the primary parameters of uni-modal encoders mitigates the issue of catastrophic interference. Moreover, the adapter training method is 3.4 times faster than the fully fine-tuned approach. Consequently, as shown in Table.2, our model attains improvements of 0.2% on AUC and 1.5% on PNR. Overall, AdaptFormer proves advantageous for both training effectiveness and efficiency.

3.5 Case Study

Apart from the above quantitative analysis, we conduct qualitative analysis based on some cases in real world video search scenario, as shown in Fig 3. For instance, given the query "Cinderella2 Dreams Come True", we observe a video whose title includes the keywords of query, but the content of the video is a Thai romantic comedy, not Disney's Cinderella. This video was initially predicted as rate "Good" with a score 0.52. After incorporating QUALITY, the prediction score decreases to 0.25. Through the analysis of these cases, we empirically conclude that incorporating multi-modal features extracted from QUALITY can significantly enhance the discriminative power of the relevance ranking model.

3.6 Deployment & Online A/B Testing

To evaluate the effectiveness of our proposed method in our real-world video search engine, we deploy the proposed model to our online system and compare it with online baseline models which are mainly text-based baselines like BERT and BM25. Following a week-long observation, A/B

test results demonstrate that query-aware multi-modal-based ranking relevance model outperforms the online baseline models, achieving a 2.1% improvement on AWT. Furthermore, we conduct manual GSB evaluation on the final search results, and our proposed model contributes to a 5.7% improvement in ΔGSB .

4 Conclusion & Limitations

In this study, we introduce QUALITY, a query-aware pre-training model that leverages multi-modal information, including queries, video frames, tags, and titles. QUALITY is integrated into our ordinal regression-based ranking relevance model. Through extensive experiments conducted on real-world data, we demonstrate the effectiveness of our proposed method.

Our method relies on a graph mining strategy that utilizes search log data to identify previously unobserved query-video pairs, thus alleviating the Matthew Effect problem in search engines. Nonetheless, the accuracy of our approach may be influenced by noise during graph construction. Consequently, we recommend investigating alternative hard mining strategies or visual debiasing strategy (Chen et al., 2022b) to enhance performance.

References

Oren Barkan and Noam Koenigstein. 2016. Item2vec: neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

- Lin Bo, Liang Pang, Gang Wang, Jun Xu, XiuQiang He, and Ji-Rong Wen. 2021. Modeling relevance ranking under the pre-training and fine-tuning paradigm. *arXiv preprint arXiv:2108.05652*.
- Paul-Christian Bürkner and Matti Vuorre. 2019. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.
- Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022a. Adapterformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678.
- Si Chen, Chen Lin, Wanxian Guan, Jiayi Wei, Xingyuan Bu, He Guo, Hui Li, Xubin Li, Jian Xu, and Bo Zheng. 2022b. Visual encoding and debiasing for ctr prediction. *arXiv preprint arXiv:2205.04168*.
- Yizhen Chen, Jie Wang, Lijian Lin, Zhongang Qi, Jin Ma, and Ying Shan. 2023. Tagging before alignment: Integrating multi-modal tags for video-text retrieval. *arXiv preprint arXiv:2301.12644*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Huifeng Guo, Bo Chen, Ruiming Tang, Weinan Zhang, Zhenguo Li, and Xiuqiang He. 2021. An embedding learning framework for numerical features in ctr prediction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2910–2918.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.
- Yiqun Liu, Kaushik Rangadurai, Yunzhong He, Siddarth Malreddy, Xunlong Gui, Xiaoyi Liu, and Fedor Borisjuk. 2021. Que2search: fast and accurate query and document understanding for search at facebook. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3376–3384.
- Huashao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.
- Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.
- Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. 2017. DeepRank: A new deep architecture for relevance ranking in information retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 257–266.
- Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.
- 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*, pages 8748–8763. PMLR.
- Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.
- Noel E Sharkey and Amanda JC Sharkey. 1995. An analysis of catastrophic interference. *Connection Science*, 7(3-4):301–330.
- Xiang-Rong Sheng, Jingyue Gao, Yueyao Cheng, Siran Yang, Shuguang Han, Hongbo Deng, Yuning Jiang, Jian Xu, and Bo Zheng. 2022. Joint optimization of ranking and calibration with contextualized hybrid model. *arXiv preprint arXiv:2208.06164*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly, Mi-wei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, et al. 2016. Ranking relevance in yahoo search. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 323–332.
- Lixin Zou, Shengqiang Zhang, Hengyi Cai, Dehong Ma, Suqi Cheng, Shuaiqiang Wang, Daiting Shi, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained language model based ranking in baidu search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4014–4022.

A Implementation Details

Our QUALITY model comprises a BERT-base with 124M parameters, a ViT-B/16 with 86M parameters, a vision-tag multi-modal encoder with 2M parameters, and a vision-text multi-modal encoder with 2M parameters. BERT-base and ViT-B/16 are pre-trained as a CLIP model on 20M video cover-title pairs from our search log. We uniformly sample 5 keyframes for each video and resize them to a resolution of 224×224 . For online usage convenience, the embedding size of image, query, tag, and title modalities is reduced from 768 to 64 using projection layers. We train the models for 1 million steps on 4 NVIDIA A100 GPUs, with an initial learning rate of $1e^{-4}$ for the first 10,000 steps, which is then gradually decayed to $5e^{-5}$.

We use a hierarchical learning rate for the relevance ranking model, setting $1e^{-5}$ for pre-trained cross-encoder layers and $5e^{-4}$ for other layers. Notably, the pre-trained cross-encoder is based on a single-layer transformer network distilled from BERT-base, featuring an embedding size of 64 and a hidden layer size of 64. Regarding the thresholds of ordinal regression, we initialize them with $-5, 0, 5$. Besides, we set hyper-parameter α in final ranking loss as 0.5.