Multi-task Ranking with User Behaviors for Text-video Search

Peidong Liu[1,2], Dongliang Liao[2,†], Jinpeng Wang[1,2]
Yangxin Wu[2], Gongfu Li[2], Shu-Tao Xia[1,3,†], Jin Xu[4]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Wechat Group, Tencent Inc.
[3]Research Center of Artificial Intelligence, Peng Cheng Laboratory
[4]School of Future Technology, South China University of Technology
{lpd19,wjp20}@mails.tsinghua.edu.cn
{brightliao,yangxinwu,gongfuli}@tencent.com
xiast@sz.tsinghua.edu.cn,jinxu@scut.edu.cn

ABSTRACT

Text-video search has become an important demand in many industrial video sharing platforms, *e.g.*, YouTube, TikTok, and WeChat Channels, thereby attracting increasing research attention. Traditional relevance-based ranking methods for text-video search concentrate on exploiting the semantic relevance between video and query. However, relevance is no longer the principal issue in the ranking stage, because the candidate items retrieved from the matching stage naturally guarantee adequate relevance. Instead, we argue that boosting user satisfaction should be an ultimate goal for ranking and it is promising to excavate cheap and rich user behaviors for model training. To achieve this goal, we propose an effective Multi-Task Ranking pipeline with User Behaviors (*MTRUB*) for text-video search. Specifically, to exploit the multi-modal data effectively, we put forward a *Heterogeneous Multi-modal Fusion Module (HMFM)* to fuse the query and video features of different modalities in adaptive ways. Besides that, we design an *Independent Multi-modal Input Scheme (IMIS)* to alleviate competing task correlation problems in multi-task learning. Experiments on the offline dataset gathered from WeChat Search demonstrate that *MTRUB* outperforms the baseline by 12.0% in mean gAUC and 13.3% in mean nDCG@10. We also conduct live experiments on a large-scale mobile search engine, *i.e.*, WeChat Search, and *MTRUB* obtains substantial improvement compared with the traditional relevance-based ranking model.

CCS CONCEPTS

• **Information systems** → **Content ranking**.

KEYWORDS

Text-video Search, Ranking Model, Multi-task Learning, User Behaviors, Multi-modal Fusion

† Corresponding author.
This work is supported in part by the National Natural Science Foundation of China under Grant 62171248, and the PCNL KEY project (PCL2021A07).

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04.
https://doi.org/10.1145/3487553.3524207

ACM Reference Format:
Peidong Liu[1,2], Dongliang Liao[2,†], Jinpeng Wang[1,2], Yangxin Wu[2], Gongfu Li[2], Shu-Tao Xia[1,3,†], Jin Xu[4] . 2022. Multi-task Ranking with User Behaviors for Text-video Search. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487553.3524207

1 INTRODUCTION

With the development of artificial intelligence, researchers concentrate on different topics, such as video understanding [15], neural architecture search [9], intelligence system [23], etc. Text-video search, one of the important tasks in video understanding, aims to retrieve the most user-satisfactory videos given a text query. This task has attracted much attention from researchers around the world for its practical use in the industrial video-sharing platforms, *e.g.*, YouTube, TikTok, and WeChat Channels.

In recent years, the video understanding community has witnessed substantial progress in text-video search. The common practices for text-video search focus on exploiting the semantic relevance between video and query with concept-based methods, cross-modal fusion methods, and latent space-based methods. Concept-based methods [1, 7, 14, 18, 21] depend on representing the video with a set of concepts, which are used to match with query text. The challenges of these methods lie in how to select relevant and detectable concepts for both video and query. As for cross-modal fusion methods [24–26], they design a cross-modal fusion subnetwork that takes text and video as input and directly produces similarity between them. Although these methods are effective, their search efficiency is low in reality as video and text should be coupled together before feeding to the network. Instead, latent space-based methods [4, 12, 17, 22] propose to encode video and query and then map them into a common latent space, which is commonly used because of their great ranking performance and high efficiency.

In practice, text-video search can be typically decomposed into two stages, namely the matching stage and ranking stage. In the matching stage, we train the model with semantic relevance between the video and the query text in the usual way to guarantee that the fetched videos are semantically relevant. However, in the ranking stage, we argue that semantic relevance is no longer the principal issue because the candidate items retrieved from the matching stage naturally guarantee adequate relevance. In some cases, only considering semantic relevance is not sufficient to provide user-satisfactory videos given a query text. For example, when a user searches with a query text *LeBron James*, a traditional

relevance-based search engine would rank the videos about *LeBron James playing a basketball game* without distinction, which may bring bad experience to users because most of them actually want to watch the top-10 exciting *LeBron James*'s dunks of the night. To improve user experience in the search engine, we focus on the ranking stage and exclusively utilize user behaviors, *e.g.*, user feedback and interactions, instead of the relevance scores, to catch user interest on videos given a query. Compared with the video recommendation methods [8, 10, 11], we concentrate on understanding the query texts instead of using user portrait features (*e.g.*, whether a user like cartoons) or context features (*e.g.*, time and location) because query texts reflect user intention directly and contain sufficient semantic information to match with video features.

In this work, we propose an effective Multi-Task Ranking pipeline with User Behaviors (*MTRUB*) for text-video Search. To utilize the multi-modal data effectively, we propose *Heterogeneous Multi-modal Fusion Module (HMFM)* to fuse the query and video features in different manners. In addition, we design *Independent Multi-modal Input Scheme (IMIS)* to mitigate the complex and even competing task correlation problems in multi-task learning.

Experiments on the offline dataset gathered from WeChat Search show that *MTRUB* outperforms the baseline by 12.0% in mean gAUC and 13.3% in mean nDCG@10. We further conduct live experiments with *MTRUB* on a large-scale mobile search engine, *i.e.*, WeChat Search, and obtain great improvement over the relevance-based ranking model. The offline and live experiments consistently demonstrate the effectiveness of our proposed method.

In summary, our main contributions are as follows:

- We propose the *MTRUB* pipeline, the first study to exclusively explore user behaviors to learn a ranking model for text-video, which directly studies user satisfaction towards videos.
- To utilize the multi-modal data effectively, we put forward *Heterogeneous Multi-modal Fusion Module (HMFM)* to match query with different modalities in adaptive ways.
- We design the *Independent Multi-modal Input Scheme (IMIS)* to alleviate the complex and even competing task correlation problems in multi-task learning.
- Both offline and live experiments on WeChat Search show that our *MTRUB* outperforms the competitive baseline substantially.

2 THE PROPOSED APPROACH

2.1 MTRUB Pipeline

Basically in the ranking model for text-video search, we can leverage both text information and visual information from videos. The text data involves the video title, the texts obtained from Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR), while the visual data contains the video cover image and key frames. To model the relationship between the query text and multi-modal video information, we propose the MTRUB pipeline that exploits the readily available user behaviors to learn a ranking model for text-video search. Figure 1 is the overview of the MTRUB.

For the query text and the text information of the video, *i.e.*, OCR, ASR, and the video title, we apply a pre-trained BERT [3] encoder from the CLIP [19] model to obtain both sentence-level embeddings and word-level embeddings. For the visual information of the video, *i.e.*, video cover and key frames, we apply a pre-trained ViT [5] from the CLIP [19] model to acquire image embeddings. Note that the parameters of the text encoder and image encoder are fixed to only provide compact semantic vectors for query and video information. We match embeddings from the query and video with the proposed Heterogeneous Multi-modal Fusion Module, which excavates diverse information from multi-modal data.

As the tasks are complex and even competing with each other, we propose a novel input scheme for multi-modal data, *i.e.*, Independent Multi-modal Input Scheme, in expert-based models, where different tasks can exploit distinct modal data explicitly. Besides that, this scheme improves model interpretability by analyzing the weights in the gate mechanisms to obtain the importance of distinct modal data towards different tasks.

2.2 Heterogeneous Multi-modal Fusion Module

Previous multi-modal fusion methods [8, 11, 12, 16] fail to consider the heterogeneous structures of text and image, which limit the matching performance. To further consider distinct characteristics of text and visual data, we propose a Heterogeneous Multi-modal Fusion Module (HMFM) to fuse the query and video features in different manners. On the one hand, we match the query with the image data, *i.e.*, video cover and key frames, in a coarse-grained way by concatenating the image embeddings and query sentence-wise vectors. On the other hand, HMFM fuses the query with video text data, *i.e.*, OCR, ASR, and title, in a fine-grained way by modeling their semantic relationship with Conv-KRNM [2] for word-wise text embeddings. Conv-KRNM is a popular convolutional kernel-based neural ranking model that models n-gram soft matches between query and video text in a unified embedding space. Given a query word-wise embedding and video text word-wise embedding, Conv-KRNM generates n-gram vectors with a CNN layer, cross-matches query n-gram vectors and video text n-gram vectors of different lengths in the n-gram embedding space to get translation matrix and apply kernel pooling layer to get the soft-TF features, which describes the distribution of match scores between query and video text. These obtained soft-TF features introduce more semantic information by fusing with the original embeddings. With HMFM, the query can match with different modal data using adaptive manners, *i.e.*, coarse-grained matching for video visual features and fine-grained matching for video text features, to excavate diverse semantic information from multi-modal data.

2.3 Independent Multi-modal Input Scheme

Multi-task ranking for text-video search encounters complex and even competing task correlation problems as in multi-task video recommendation systems [13, 20]. To address this issue, we propose a novel Independent Multi-modal Input Scheme (IMIS), where different tasks can explicitly utilize distinct modal data. IMIS decouples the multi-modal input by feeding only one modal data, instead of all modal data, to each expert. In addition, the gate mechanisms in the ranking model can reflect the importance of distinct modal input towards different tasks with learned weights, where large values indicate a higher degree of importance. Therefore, IMIS improves multi-task model interpretability to a large extent.

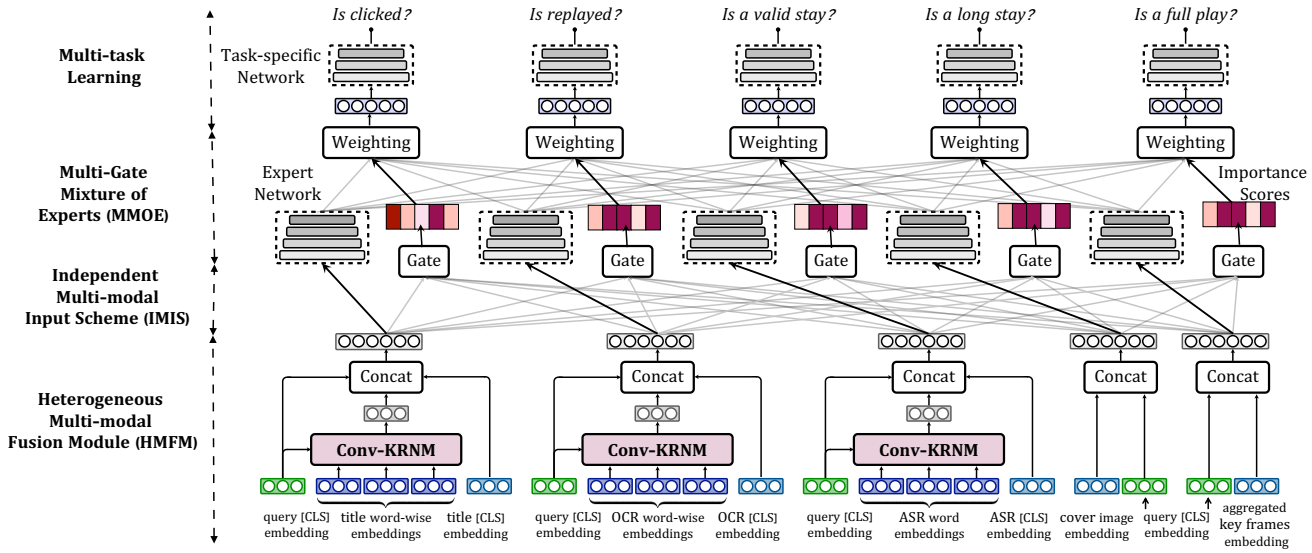

Figure 1: An overview of the proposed *MTRUB* pipeline. For the multi-modal input fusion, we propose a *Heterogeneous Multi-modal Fusion Module (HMFM)*, which considers different characteristics of text and visual data. HMFM matches query with each of multiple modal data in adaptive manners, *i.e.*, coarse-grained matching for video visual features and fine-grained matching for video text features. In addition, to alleviate the competing task correlation problems, we design an *Independent Multi-modal Input Scheme (IMIS)*, which decouples the multi-modal input by feeding only one modal data, instead of all modal data, to each expert. The gating mechanism can learn the importance of distinct modal input towards different tasks. Note that the five small squares gained from the gates indicate importance scores, where dark red means high importance and light red means low importance. We discover that in *isClick* task, cover image of the video and title play more important roles, while OCR, ASR and video key frames contribute more to tasks that are related to user staying time, *e.g.*, *isReplay*, *isLongStay*, etc.

3 EXPERIMENTS

3.1 Datasets and Metrics

3.1.1 Datasets. For offline experiments, we collect 800k samples from the Wechat Search engine, each of which contains a text query, a set of candidate videos, and the user behavior log about this query. Each query is associated with around 10 video candidates, including both positive candidates (exposed and clicked) and negative ones (exposed but not clicked). For each video, we extract the video cover image and 30 key frames as the visual information. Besides, we obtain the texts from Optical Character Recognition (OCR), Automated Speech Recognition (ASR), and video title to serve as the text information for each video. We extract the user behavior labels from the logs, including: whether a user clicks a video (*isClick*), whether a user replays a video (*isReplay*), whether a user plays a video for more than 10 seconds (*isValidStay*), whether a user plays a video for more than 30 seconds (*isLongStay*), and whether a user plays a video for more than 90 percent of a video duration (*isFullPlay*). We split the whole dataset into a training set of 560k samples and a test set of the rest 240k samples.

3.1.2 Metrics. We adopt group-wise Area Under the ROC Curve (gAUC) and normalized version of Discounted Cumulative Gain (nDCG@k) as the performance metrics, where gAUC measures the likelihood that a relevant item is ranked higher than an irrelevant item and nDCG@k is truncated at a particular rank level k to emphasize the importance of the top-k searched videos.

3.2 Implementation Details

To alleviate the learning difficulty and save computational cost, we pre-extract the text and visual features before forwarding them to the models. Specifically, we adopt a BERT [3] encoder to extract the text features and a ViT [5] model to obtain image embeddings. The parameters of BERT and ViT are from the pre-trained CLIP [19]. We adopt the Adam [6] optimizer with a cyclic cosine annealing learning rate schedule and a weight decay of 10^{-2}. We train the models for 100 epochs with a batch size of 1024. We apply a binary cross-entropy loss for each task. For the MMOE [13], we use 16 experts and 128 hidden units. To facilitate fair comparison, we keep the parameters for all compared models approximate (in 60M±2M). To find an appropriate learning rate, we try several settings (*i.e.*, 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}) for each experiment, and report the models with the best performance metric. We run 5 times with different random seeds for each experiment and report the average values.

3.3 Ablation Studies on Offline Dataset

3.3.1 Effectiveness of Mixture-Of-Experts. To verify the effectiveness of Mixture-Of-Experts (MOE) in the ranking model, we construct two simple MEE [16] baselines with (*i.e.*, MEE+MMOE) and without MMOE (*i.e.*, MEE+Shared-Bottom) for comparison. As shown in Table 1, MOE enhances the model capability to capture the task differences, thereby boosting the performances on different tasks over the shared-bottom counterpart.

	isClick		*isReplay*		*isValidStay*		*isLongStay*		*isFullPlay*		Mean	
	gAUC	nDCG	gAUC	nDCG	gAUC	nDCG	gAUC	nDCG	gAUC	nDCG	gAUC	nDCG
MEE+Shared-Bottom	65.6	57.1	67.9	58.4	66.6	57.4	67.0	57.2	67.8	58.0	66.9	57.6
MEE+MMOE	68.6	60.4	71.7	63.2	69.4	61.1	70.2	62.0	70.6	62.4	70.1	61.8
MEE+MMOE with ***IMIS***	68.1	59.8	72.7	64.2	69.0	60.8	71.0	62.8	72.1	63.6	70.6	62.2
HMFM+MMOE with ***IMIS***	**78.6**	**70.8**	**79.3**	**71.1**	**78.6**	**70.6**	**78.7**	**70.7**	**79.2**	**71.2**	**78.9**	**70.9**

Table 1: Ablation studies on offline Wechat Search dataset. "nDCG" indicates nDCG@10 for short. All models are trained with multi-task learning and we report the results of gAUC and nDCG@10 for all tasks, *i.e.*, *isClick*, *isReplay*, *isValidStay*, *isLongStay*, and *isFullPlay*. The best value of each column is in bold font. Our proposed MTRUB shows substantial improvements over the baselines in terms of gAUC and nDCG@10, demonstrating that *IMIS* is an effective input scheme for ranking model and *HMFM* is a better multi-modal fusion scheme than the MEE [16].

3.3.2 Effectiveness of Independent Multi-modal Input Scheme. We evaluate the Independent Multi-modal Input Scheme (IMIS) that makes each expert focus on one of the input modalities. Table 1 shows that IMIS boosts the performance in terms of the mean gAUC and nDCG@10 on most tasks. In addition, by analyzing the gate weights in the MOE, where a larger value indicates higher importance, we discover that cover image and video title play more important roles in the *isClick* task, while OCR, ASR, and video key frames contribute more to tasks that are related to the user staying time, *e.g.*, *isValidStay*, *isLongStay*, etc.

3.3.3 Effectiveness of the Heterogeneous Multi-modal Fusion Module. Here we investigate the effectiveness of the Heterogeneous Multi-modal Fusion Module (HMFM), which leverages the text and visual information in videos in a heterogeneous manner. As shown in table 1, HMFM consistently improves the gAUC and nDCG@10 by large margins over other baselines with MEE. HMFM provides more adaptive interactions between the query and each of various video modalities, thus contributing to a better information fusion.

3.3.4 Effectiveness of multi-task learning for MTRUB. We further compare the performance of MTRUB under the single-task and multi-task learning manners. In single-task learning, we apply the proposed *MTRUB* with only one single task tower and report the performance metric of the same task. As illustrated in figure 2, compared to single-task learning, multi-task learning enables different tasks to share knowledge together and hence boost the search performance for all tasks, especially those tasks that perform poorly in single-task learning, *e.g.*, isFullplay, isReplay, etc.

3.4 Online Experiment Result

In this section, we conduct a large-scale online evaluation on the Wechat Search engine. We randomly sample three user groups from Wechat Search, each of which contained about 1.4 million users. We implement 7 days control variable experiments on different groups to evaluate different ranking models, with the same video index, retrieval methods, and front-end interface. We evaluated three different ranking strategies: ranking based on relevance model trained with the human-annotated dataset, ranking based on single-task learning user behavior-based (*isClick*) model, ranking based on the proposed multi-task learning user behavior-based model. Especially,

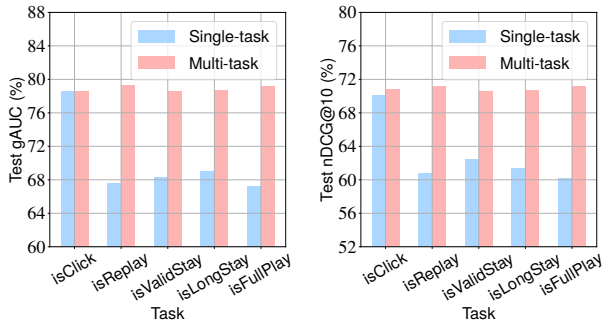

Figure 2: Comparison between MTRUB and single-task version MTRUB. Note that in single-task version MTRUB, we only apply one single task tower for each experiment and report the performance metric of the same task. The results on gAUC and nDCG@10 are listed on the left and right, respectively. Experiments on different tasks consistently indicate the superiority of MTRUB with multi-task learning in user behavior modeling-based text-video search.

we generate ranking scores by exponential fusion of all user behaviors, *i.e.*, $Score = P_{click} * P^{\alpha}_{fullplay} * P^{\beta}_{replay} * P^{\gamma}_{validstay} * P^{\delta}_{longstay}$ and employ Pareto Optimality for searching parameters $\alpha, \beta, \gamma, \delta$. Comparing with relevance ranking, ranking by click-through rate (*isClick*) prediction improve 3.3% clicked query views and 3.7% *isClick* rate relatively. Furthermore, comparing with single-ctr-task, ranking by multi-task learning user behavior-based model further improve online *isClick* rate by 1.3%, *isReplay* rate by 3.6%, *isLongStay* rate by 1.7%, *isValidStay* rate by 3.4% and *isFullPlay* rate by 1.9%.

4 CONCLUSIONS

In this work, we propose an effective Multi-Task Ranking pipeline with User Behaviors (*MTRUB*) for text-video search. To exploit the multi-modal input data, we put forward *Heterogeneous Multi-modal Fusion Module (HMFM)* to fuse the query and video features of different modalities in adaptive manners. In addition, we design *Independent Multi-modal Input Scheme (IMIS)* to alleviate competing task correlation problems in a multi-learning scheme. Both offline and live experiments on WeChat Search show that *MTRUB* outperforms the baseline by a large margin.

REFERENCES

[1] Konstantinos Avgerinakis, Anastasia Moumtzidou, Damianos Galanopoulos, Georgios Orfanidis, Stelios Andreadis, Foteini Markatopoulou, Elissavet Batziou, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, et al. 2018. ITI-CERTH participation in TRECVID 2018.. In *TRECVID*.

[2] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *WSDM*. 126–134.

[3] 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* (2018).

[4] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2018. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia* 20, 12 (2018), 3377–3388.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

[6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[7] Duy-Dinh Le, Sang Phan, Vinh-Tiep Nguyen, Benjamin Renoust, Tuan A Nguyen, Van-Nam Hoang, Thanh Duc Ngo, Minh-Triet Tran, Yuki Watanabe, Martin Klinkigt, et al. 2016. NII-HITACHI-UIT at TRECVID 2016.. In *TRECVID*, Vol. 25.

[8] Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. 2021. SEMI: A Sequential Multi-Modal Information Transfer Network for E-Commerce Micro-Video Recommendations. In *KDD*. 3161–3171.

[9] Peidong Liu, Gengwei Zhang, Bochao Wang, Hang Xu, Xiaodan Liang, Yong Jiang, and Zhenguo Li. 2020. Loss Function Discovery for Object Detection via Convergence-Simulation Driven Search. In *ICLR*.

[10] Qi Liu, Ruobing Xie, Lei Chen, Shukai Liu, Ke Tu, Peng Cui, Bo Zhang, and Leyu Lin. 2020. Graph neural network for tag ranking in tag-enhanced video recommendation. In *CIKM*. 2613–2620.

[11] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-video co-attention network for personalized micro-video recommendation. In *WWW*. 3020–3026.

[12] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*.

[13] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *KDD*. 1930–1939.

[14] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. 2017. Query and keyframe representations for ad-hoc video search. In *ICMR*. 407–411.

[15] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516* (2018).

[16] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516* (2018).

[17] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*. 2630–2640.

[18] Phuong Anh Nguyen, Qing Li, Zhi-Qi Cheng, Yi-Jie Lu, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. 2017. VIREO@ TRECVID 2017: Video-to-Text, Ad-hoc Video Search, and Video hyperlinking.. In *TRECVID*.

[19] 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. *arXiv preprint arXiv:2103.00020* (2021).

[20] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *RecSys*. 269–278.

[21] Kazuya Ueki, Koji Hirakawa, Kotaro Kikuchi, Tetsuji Ogawa, and Tetsunori Kobayashi. 2017. Waseda_Meisei at TRECVID 2017: Ad-hoc Video Search.. In *TRECVID*.

[22] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. 2019. A graph-based framework to bridge movies and synopses. In *ICCV*. 4592–4601.

[23] Xiaojun Yang, Lunjia Liao, Qin Yang, Bo Sun, and Jianxiang Xi. 2021. Limited-energy output formation for multiagent systems with intermittent interactions. *Journal of the Franklin Institute* 358, 13 (2021), 6462–6489.

[24] Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *ECCV*. 471–487.

[25] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*. 3165–3173.

[26] Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *CVPR*. 8746–8755.