Query-Oriented Micro-Video Summarization

Mengzhao Jia, Yinwei Wei, *Member, IEEE*, Xuemeng Song, *Senior Member, IEEE*, Teng Sun, Min Zhang, *Member, IEEE*, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*—Query-oriented micro-video summarization task aims to generate a concise sentence with two properties: (a) summarizing the main semantic of the micro-video and (b) being expressed in the form of search queries to facilitate retrieval. Despite its enormous application value in the retrieval area, this direction has barely been explored. Previous studies of summarization mostly focus on the content summarization for traditional long videos. Directly applying these studies is prone to gain unsatisfactory results because of the unique features of micro-videos and queries: diverse entities and complex scenes within a short time, semantic gaps between modalities, and various queries in distinct expressions. To specifically adapt to these characteristics, we propose a query-oriented micro-video summarization model, dubbed QMS. It employs an encoder-decoder-based transformer architecture as the skeleton. The multi-modal (visual and textual) signals are passed through two modal-specific encoders to obtain their representations, followed by an entity-aware representation learning module to identify and highlight critical entity information. As to the optimization, regarding the large semantic gaps between modalities, we assign different confidence scores according to their semantic relevance in the optimization process. Additionally, we develop a novel strategy to sample the effective target query among the diverse query set with various expressions. Extensive experiments demonstrate the superiority of the QMS scheme, on both the summarization and retrieval tasks, over several state-of-the-art methods.**

***Index Terms*—Video summarization, query suggestion, micro-video retrieval.**

I. Introduction

As ONE of the most widespread media types, micro-videos are able to bring viewers incredible audio-visual enjoyment in a short time [1], [2], [3]. In recent years, the volume of user-generated micro-videos uploaded on various platforms, including Tiktok[1] and Kwai,[2] has witnessed an explosive surge. Taking Kwai as an example, roughly tens of millions of micro-videos are recorded and published every day.[3]

Facing the huge number of micro-videos, users on Kwai platform conduct more than 300 million querying requests per day to obtain the desired micro-videos.[4] One effective approach for micro-video retrieval is to evaluate similarities between the user-generated video summarizations and queries [4]. However, the summarizations associated with micro-videos are prone to be overlooked by users when uploading due to the inconvenience on mobile devices, which leads to suboptimal performance in micro-video retrieval.

Many endeavors have been devoted to the video summarization task in recent years. These earlier studies primarily focus on multi-modal content understanding [5] and natural language generation. For instance, Liu et al. [6] explored a multistage fusion network for capturing the fine-grained interactions between visual and textual signals to better understand the multi-modal inputs. Shang et al. in [7] proposed a time-aware transformer to utilize the information in the video timestamps, therefore facilitating the integration of multi-modal signals. In addition to enhancing multi-modal information comprehension, some studies also aim to improve the generation process of video summarization. For example, Amirian et al. [8] devised a two-stage generation system, which first generated captions for videos and then fed the captions into a uni-modal document summarizer to obtain the abstract of the video. Recently, Yu et al. [9] have leveraged the advantages of pre-trained generative language models for the video summarization task. Specifically, by incorporating visual information into pre-trained models, they enhanced the multi-modal summarization with their acquired knowledge of language generation.

Despite their remarkable performance, we argue that these approaches are inapplicable to our micro-video summarization task due to the following two reasons: 1) existing methods are designed for the traditional long videos that differ from the micro-video in duration and quality; and 2) the target summaries in prior studies are created and refined by human experts, which is not feasible for the scale of micro-video summarization. Therefore, we collect the query-microvideo pairs from the Kwai platform and resort to the queries as supervision, to implement a query-orient summarization model for micro-videos. To distinguish the traditional video summarization, we formulate the query-orient summarization by jointly considering the micro-video content and users' intents hidden in their queries. However, it is

Manuscript received 11 April 2023; revised 30 December 2023; accepted 13 January 2024. Date of publication 18 January 2024; date of current version 7 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62236003 and 62376137, in part by Shandong Provincial Natural Science Foundation under Grant ZR2022YQ59, and in part by Shenzhen College Stability Support Plan under Grant GXWD20220817144428005. Recommended for acceptance by M. Wang. *(Corresponding authors: Yinwei Wei; Liqiang Nie.)*

Mengzhao Jia, Xuemeng Song, and Teng Sun are with the Department of Computer Science and Technology, Shandong University, Qingdao 250100, China (e-mail: jiamengzhao98@gmail.com; sxmustc@gmail.com; stbestforever@gmail.com).

Yinwei Wei is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: weiyinwei@hotmail.com).

Min Zhang and Liqiang Nie are with the School of Computing, Harbin Institute of Technology, Shenzhen 150001, China (e-mail: zhangminmt@hotmail.com; nieliqiang@gmail.com).

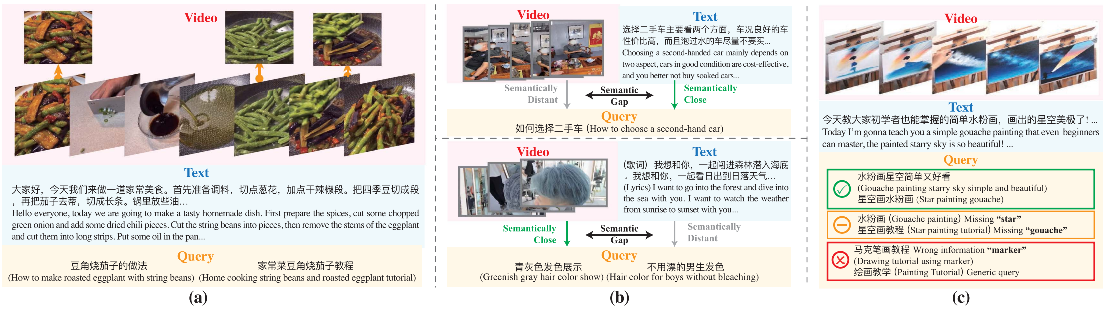
Digital Object Identifier 10.1109/TPAMI.2024.3355402

[1] [Online]. Available: https://www.tiktok.com.

[2] [Online]. Available: https://www.kwai.com.

[3] [Online]. Available: https://tinyurl.com/mr962p5x.

[4] [Online]. Available: https://tinyurl.com/3z4wvtva.



大家好，今天我们来做一道家常美食。首先准备调料，切点葱花，加点干辣椒段。把四季豆切成段，再把茄子去蒂，切成长条。锅里放些油…
Hello everyone, today we are going to make a tasty homemade dish. First prepare the spices, cut some chopped green onion and add some dried chili pieces. Cut the string beans into pieces, then remove the stems of the eggplant and cut them into long strips. Put some oil in the pan...

豆角烧茄子的做法 (How to make roasted eggplant with string beans)
家常菜豆角烧茄子教程 (Home cooking string beans and roasted eggplant tutorial)

选择二手车主要看两个方面，车况良好的车性价比高，而且泡过水的车尽量不要买...
Choosing a second-handed car mainly depends on two aspect, cars in good condition are cost-effective, and you better not buy soaked cars...

如何选择二手车 (How to choose a second-hand car)

(歌词) 我想和你，一起闯进森林潜入海底。我想和你，一起看日出到日落天气…
(Lyrics) I want to go into the forest and dive into the sea with you. I want to watch the weather from sunrise to sunset with you...

青灰色发色展示 (Greenish gray hair color show)
不用漂的男生发色 (Hair color for boys without bleaching)

今天教大家初学者也能掌握的简单水粉画，画出的星空美极了! ...
Today I’m gonna teach you a simple gouache painting that even beginners can master, the painted starry sky is so beautiful! ...

水粉画星空简单又好看
(Gouache painting starry sky simple and beautiful)
星空画水粉画 (Star painting gouache)

水粉画 (Gouache painting) Missing **“star”**
星空画教程 (Star painting tutorial) Missing **“gouache”**

马克笔画教程 Wrong information **“marker”**
(Drawing tutorial using marker)
绘画教学 (Painting Tutorial) Generic query

Fig. 1. Illustration of three features in micro-videos and queries. (a) Micro-videos involve a wide range of content in a short time. In contrast, the users’ search intention is narrowed to specific key entities. (b) Large semantic gaps between different modalities. (c) A micro-video may link to different queries, which involve diverse expression manners and search intents.

non-trivial to develop a new query-orient summarization model for micro-videos due to the following challenges.

- C1: The duration is a significant difference between micro-videos and long videos. Owing to the advanced editing software, users can seamlessly transition between various scenes. Despite the short length, micro-videos cover a wide range of scenes and diverse entities to convey the complete theme. Thus, only limited information can be provided to understand each scene. As illustrated in Fig. 1(a), in some cases, certain entities may only appear for a few frames, making it even more difficult to recognize and understand the scene. Therefore, recognizing entities and identifying the critical ones under limited information is the first challenge.
- C2: The semantic gap between the different modalities should be considered in the micro-video understanding. As shown in Fig. 1(b), the keyframes visually depict a scene with three men drinking tea, while the audio reflects a conversation about vehicle purchases. To empirically explore this problem, we conduct a comparison experiment detailed in Section IV-A2 to compare the cross-modal semantic gap of micro-video with that of long video [10]. The experiment results clearly illustrate that the semantic gap in micro-videos is more significant than in long videos. This observation is consistent with the previous studies [1], [11], [12]. Hence, the semantic gap challenges us to capture the essential topic from the multi-modal information of micro-video.
- C3: Regarding the collected queries, it can be observed that a single micro-video can be linked with a variety of diverse and possibly noisy queries. This may be attributed to the distinct intents, modes of expression, and search behaviors. As exemplified in Fig. 1(c), queries in the green rectangle precisely grasp two key subjects (*star painting* and *gouache*). However, the queries in the yellow rectangle lack one essential subject, leading to a less comprehensive understanding of the topic. Moreover, the queries in the red rectangle provide either erroneous information (*marker*) or are too vague to reveal the specific topic. Therefore, how to optimize the model under the situation of noisy targets is the third challenge we are facing.

To address these challenges, we develop a novel **Q**uery-oriented **M**icro-video **S**ummarization model, dubbed QMS, as shown in Fig. 2. Based on the encoder-decoder structure, the proposed model can learn multi-modal representations from the micro-videos and generate query-oriented summaries for them. In particular, we propose a novel entity recognition module that efficiently identifies and highlights the key entities from the limited information in the micro-video. Additionally, we introduce an optimization method that effectively mitigates the challenges encountered during the learning process. The proposed method assigns confidence scores to different modalities based on their semantic content to process valid information from the multi-modal features. Furthermore, it employs a query sampling strategy that takes into account the amount of information provided by the queries to address the issue of noisy queries.

To acquire the video-summary pairs for supervised training, we resort to the real-world search logs to construct a dataset named QMV-Kwai, which contains 60,096 micro-videos and 331,414 queries. In the dataset, several diverse queries of different qualities correspond to a single micro-video. We evaluate the effectiveness of our method on the QMV-Kwai dataset. Extensive experiments demonstrate that QMS outperforms state-of-the-art baselines on the query-oriented micro-video summarization task. We released the codes and the trained model checkpoints to facilitate other researchers in this community.[5]

We summarize our main contributions into threefold:

- We thoroughly analyze the differences between the traditional video summarization task and the query-oriented micro-video summarization task. Meanwhile, we tailor the encoder-decoder-based model toward the unique characteristics of micro-videos and users’ search queries.
- We propose a query-oriented micro-video summarization model. By highlighting critical entities, selecting informative target queries, and dynamically assigning confidence scores to modalities, QMS can generate concise and retrieval-friendly summaries under sophisticated micro-video searching scenarios.

[5] [Online]. Available: https://projs2release.wixsite.com/qms-page.

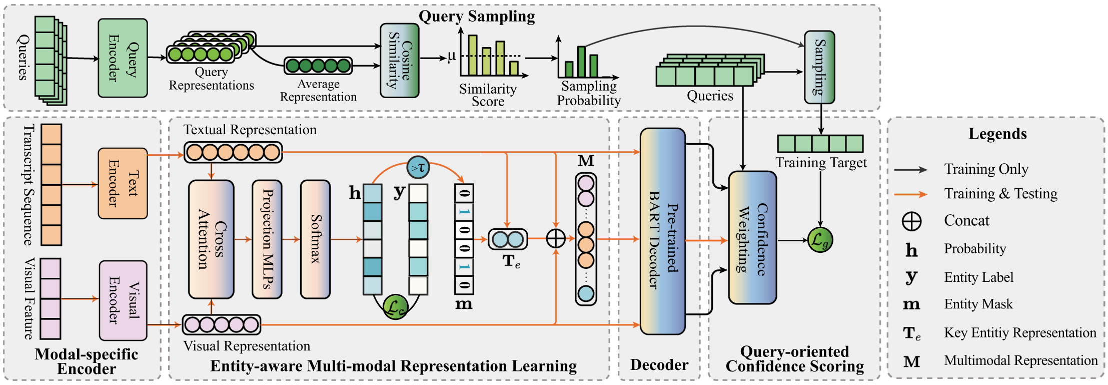


Fig. 2. Schematic illustration of our proposed framework.

- We collect video-summary pairs from search logs to construct a large-scale dataset to evaluate the performance of the proposed model. Experiments validate that our proposed model outperforms existing state-of-the-art baselines on both generative and retrieval metrics.

II. Related Work

A. Abstractive Summarization

Abstractive summarization is an important task in the field of natural language generation, which can summarize raw data (i.e., news articles [13], speech [14], videos [15]) into a condensed text with high information density. Early work focuses on abstractive summarization based on textual contents and has made remarkable advances [16]. In recent years, the multi-source content on the Internet has been growing exponentially, necessitating summarization methods and techniques that harness the power of diverse modalities. Multi-modal video abstractive summarization has become an emerging research field. Palaskar et al. [17] pioneered the focus on the summarization of multimodal instructional videos. They collected the How2 dataset and proposed a hierarchical attention-based structure to fuse and integrate multi-modal signals. Thereafter, Liu et al. [6] built a multistage fusion network with a fusion forget gate module to gain fine-grained multimodality interactions, which remedied the problem of redundancy in long multi-modal sequences. Apart from the methods trained from scratch, Yu et al. [9] investigated the potential of leveraging uni-modal pre-trained generative models to boost the multi-modal summarization task. However, they mostly focused on the traditional long videos with clear topics. Directly adapting these designs to the micro-video summarization may harm the performance. As such, we conducted an analysis of the unique characteristics of micro-videos and developed a novel model for micro-video summarization.

B. Query-Oriented Generation

The retrieval algorithm is the heart of a search engine. Although prior researchers have significantly improved the performance of the Information Retrieval (IR) system [18], [19], with the increasing amount of source content to be retrieved, neural-based approaches suffer from high computational cost [20], which is unacceptable for real-time search engines. Alternatively, researchers turned to generating query-like summaries for source documents to improve retrieval effectiveness while preventing prohibitive runtime overhead. Many studies are devised for query-oriented generation task [21], [22] Specifically, Nogueira et al. [23] adopted a sequence-to-sequence model to generate likely-queries according to the textual documents. Thereafter, Kim et al. [4] first extended the query-oriented generation task on multi-modal documents. They proposed a multi-task model by considering queries as the third modality, which can generate likely queries for offline indexing. More recently, Lien et al. [24] investigated query-oriented generation for the E-commerce area in a zero-shot manner.

However, they seldom design the query-oriented generation method for micro-video consisting of complicated scenes and entities. Hence, we adopted an entity-aware multi-model learning module to capture the fine-grained entity information, in order to capture critical topics for the summarization.

C. Learning Under Noisy Targets

Learning under noisy targets has gained significant attention as it is broadly applicable in real-world scenarios, particularly for data obtained from crowd-sourcing or web-crawling. Various approaches have been proposed to tackle this problem [25], [26], including robust architecture design [27], noisy robust regularization [28], loss function adjustment [29], and sample selection strategy [30]. Among the four techniques, we focused on the sample selection strategy which is the most relevant to our model. It concentrates on selecting true-labeled samples from noisy ones. In general, sample selection methods can be categorized into multi-network learning and multi-round learning. The former employs multiple DNNs in a cooperative manner, such as MentorNet proposed by Jiang et al. [31], which leveraged a pretrained mentor network to supervise the training of a student network collaboratively. The latter category involves multiple training rounds to refine the selected samples. For instance, Wu et al. [32] proposed a unified graph-based framework, NGC, which constructs a nearest neighbor graph iteratively using

TABLE I
KEY NOTATIONS AND DESCRIPTIONS

Notation	Description
$\mathcal{I}$	Micro-video training set.
x_i	i-th micro-video in $\mathcal{I}$.
$\mathcal{G}_i$	Set of queries associated with x_i.
C	Number of micro-videos in $\mathcal{I}$.
J_i and g_j^i	Number of queries in $\mathcal{G}_i$ and j-th query.
d_v and d_t	Dimension of visual features and text embeddings, respectively.
$\lvert v_i\rvert$ and $\lvert t_i\rvert$	Key frame number and transcript length.
$\mathbf{F}_v^i \in \mathbb{R}^{d_v \times \lvert v_i\rvert}$	Visual features.
$\mathbf{F}_t^i \in \mathbb{R}^{d_t \times \lvert t_i\rvert}$	Text embeddings.
$\mathbf{V}$, $\mathbf{T}$, and $\mathbf{M}$	Visual, textual, and multi-modal representation matrix, respectively.
U_v and U_t	Confidence score of visual and textual modalities, respectively.

latent representations of training samples, and then performs soft pseudo-label propagation to aggregate information from neighborhoods to correct the noisy labels.

However, in the context of query-oriented micro-video summarization, effective sampling strategies are necessary to preserve the diversity of users' search intents and expression manners, while filtering out noisy targets. Therefore, specialized sampling strategies are required to accommodate the unique demands of micro-video search scenarios.

III. Methodology

A. Problem Formulation

Given a training set $\mathcal{I} = \{(x_1, \mathcal{G}_1), (x_2, \mathcal{G}_2), \ldots, (x_C, \mathcal{G}_C)\}$ of C micro-videos and their corresponding queries, we denote the ith micro-video as x_i. For $(x_i, \mathcal{G}_i) \in \mathcal{I}$, $\mathcal{G}_i = \{g_1^i, g_2^i, \ldots, g_j^i, \ldots, g_{J_i}^i\}$ represents J_i queries that associate to x_i. Therein, g_j^i denotes the jth query offered by users, reflecting their interests in the content of ith micro-video. We use $g_{j,l}^i$ to represent the lth word in query g_j^i. Beyond the associating queries of micro-video, we capture one frame per second from each micro-video and obtain $|v_i|$ keyframes. Then we extract the corresponding feature vectors, denoted as $\mathbf{F}_v^i \in \mathbb{R}^{d_v \times |v_i|}$, where d_v represents the dimension of features. Considering the effectiveness of audio, we employ automatic speech recognition (ASR) to extract the transcript from the soundtrack of micro-video. Following the typical text processing techniques, we tokenize and embed the transcript into a sequence of text embeddings, i.e., $\mathbf{F}_t^i = [\mathbf{t}_1^i, \mathbf{t}_2^i, \ldots, \mathbf{t}_k^i, \ldots, \mathbf{t}_{|t_i|}^i]$, where $\mathbf{t}_k^i \in \mathbb{R}^{d_t}$ denotes the kth word embedding in the transcript. d_t and $|t_i|$ are the embedding dimension and length of the transcript, respectively. Formally, given a micro-video x_i, we aim to design a model leveraging its visual and acoustic information to generate a summary g that can precisely summarize its crucial subject or main content. For simplicity, we temporarily omit the index (i.e., i) of each training sample. Table I demonstrated the notations and descriptions in detail.

B. Query-Oriented Micro-Video Summarization

1) Modal-Specific Encoder: To acquire the informative signal from the off-the-shelf features, we devise two different encoders on visual and textual modalities, respectively. In particular, for textual modality, we opt to use the pre-trained BART [33] encoder since it achieves prominent performance in many text generation tasks [34], [35]. Specifically, the encoder contains L Transformer [36] layers, each of which consists of a multi-head self-attention sub-layer and a fully connected feed-forward sub-layer. As to the visual encoder, we employ a randomly initialized Transformer encoder with the same architecture as the BART encoder. Formally, we have

$$\begin{cases} \mathbf{V} = \mathrm{Enc}_v(\mathbf{F}_v | \Theta_v), \\ \mathbf{T} = \mathrm{Enc}_t(\mathbf{F}_t | \Theta_t), \end{cases} \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{d_v \times |v|}$ and $\mathbf{T} \in \mathbb{R}^{d_t \times |t|}$ are visual and textual representation matrices, respectively. Enc_v and Enc_t refer to the visual and textual encoders with trainable parameters Θ_v and Θ_t, respectively.

2) Entity-Aware Multi-Modal Representation Learning: To generate the query-orient video caption, we not only capture the informative features from content information, but identify the entities concerned by users in their queries. After explicitly capturing the entities, we yield the entity-aware multi-modal representation of micro-video.

Entity Identification: It is not trivial to recognize various entities and distinguish the critical ones due to the limited information conveyed by the micro-video. Intuitively, entities that are frequently and simultaneously included in the textual and visual signals have a higher probability of being related to the crucial topic. Based on that assumption, we jointly analyze the multi-modal information to discover key entities and highlight them as a hint for the summarization process. To fulfill this, we resort to the cross-attention mechanism, which is able to highlight the entity words in the transcript under the supervision of visual features. Thereafter, we fuse the entity representation along with the visual and textual representation for the generation phase. In particular, the query vector and the key-value vector pair for the attention are derived from textual representation $\mathbf{T}$ and visual representation $\mathbf{V}$, respectively. Linear projections are used to project features into latent space first. We use three matrices $\mathbf{W}^q$, $\mathbf{W}^k$ and $\mathbf{W}^v$ to project the representations. The projection can be written as

$$\begin{cases} \mathbf{Q}^a = (\mathbf{T})^T \mathbf{W}^q, \\ \mathbf{K}^a = (\mathbf{V})^T \mathbf{W}^k, \\ \mathbf{V}^a = (\mathbf{V})^T \mathbf{W}^v, \end{cases} \tag{2}$$

where $\mathbf{W}^q \in \mathbb{R}^{d_t \times d_a}$ is used to project the textural representation. We also apply two matrices $\mathbf{W}^k \in \mathbb{R}^{d_v \times d_a}$ and $\mathbf{W}^v \in \mathbb{R}^{d_v \times d_a}$ to project the visual representation into the key and value spaces, respectively. The dimension of the hidden space of cross-attention is denoted as d_a. The projected matrices for the cross-attention function are denoted as $\mathbf{Q}^a$, $\mathbf{K}^a$, and $\mathbf{V}^a$. Thereafter, the cross-attention operation is given as follows,

$$\mathbf{E} = \mathrm{Softmax}\left(\frac{\mathbf{Q}^a (\mathbf{K}^a)^T}{\sqrt{d_a}}\right) \mathbf{V}^a, \tag{3}$$

where $\mathbf{E} \in \mathbb{R}^{|t| \times d_a}$ denotes the weighted representation through cross-attention.

With the attentive representations, we learn a function to filter out the entities irrelevant to users' concerns. Formally, we implement the function with a multi-layer perceptron (MLP) equipped with the Sigmoid activation function as,

$$\mathbf{h} = \mathrm{Sigmoid}\left(\mathrm{MLP}\left(\mathbf{E}\right)\right), \tag{4}$$

whereinto, $\mathbf{h} \in \mathbb{R}^{|t|*1}$ scores words in the transcript, measuring how likely the corresponding entity will be included in users' queries. Thereafter, it is able to obtain a mask for selecting the highlighted entities based on the above scores, formally as,

$$m_k = \mathbb{1}(h_k) = \begin{cases} 1, & \text{if } h_k \geq \tau, \\ 0, & \text{else } h_k < \tau, \end{cases} \tag{5}$$

where h_k denotes the kth score in $\mathbf{h}$. $m_k = 1$ means the kth word in the transcript should be highlighted under the threshold τ, and vice versa. We let $\mathbf{m} = [m_1, \ldots, m_{|t|}]$ denote a vector that contains $|t|$ masks.

To fulfill the emphasis procedure, we apply the derived mask vector $\mathbf{m}$ onto $\mathbf{T}$ as follows,

$$\mathbf{T}_e = \mathbf{T} \odot \mathbf{m}, \tag{6}$$

where $\odot$ is the element-wise multiplication operation and $\mathbf{T}_e$ stands for the entity representation. Notably, we remove all zero columns in $\mathbf{T}_e$ for computation efficiency. The left columns can be viewed as the selected entity representation.

Information Fusion: Although we capture the entities concerned by users, it is hard to generate human-friendly summaries merely upon them due to insufficient information. Therefore, we integrate the fine-grained features to supply the cues for summarization. In particular, we concatenate the visual features (i.e., $\mathbf{V}$), textual features (i.e., $\mathbf{T}$), and fine-grained information of entities (i.e., $\mathbf{T}_e$) as,

$$\mathbf{M} = [\mathbf{V}; \mathbf{T}; \mathbf{T_e}], \tag{7}$$

where $[;;]$ denotes the concatenation operation and $\mathbf{M}$ is the enriched representation including multi-modal and multi-grained information.

3) Decoder: After encoding the content information in a compressed vector, we establish a decoder based on the pre-trained BART model, in order to generate the summary for the micro-video. In particular, the decoder is composed of L transformer layers, each of which contains the following three sub-layers: 1) a multi-head self-attention layer, 2) a fully connected feed-forward layer, and 3) a cross-attention layer that performs multi-head attention with the encoded representations. We apply the standard decoder architecture proposed in [36] to generate the summary sequence $\hat{g}$ in an auto-regressive way conditioned on the encoded multi-modal information. Specifically, for each time step l, the decoder attends over the previously generated sequence $\hat{g}_{<l}$ and the encoder outputs, then puts out the probability distribution of the next word $\hat{g}_l$. Formally, we feed the multi-modal features of micro-videos into the decoder and generate their corresponding summaries as,

$$p\left(\hat{g}_l \mid \hat{g}_{<l}, \mathbf{M}\right) = \mathrm{Dec}(\hat{g}_{<l}, \mathbf{M}|\Theta_d), \tag{8}$$

wherein, Dec represents the decoder with the to-be-learned parameters Θ_d, and $\hat{g}$ is the generated summary.

C. Optimization

To learn the trainable parameters, we adopt a multi-task optimization strategy on our proposed QMS model. In particular, under the supervision of entity identification and summary generation tasks, our proposed model is capable of understanding the content of micro-video and generating the corresponding summary in natural language.

1) Entity Identification: To conduct the entity identification task detailed in Section III-B2, we first manually construct a label vector for each micro-video, formally as,

$$\mathbf{y} = [y_1, y_2, \ldots, y_{|t|}], \tag{9}$$

where $\mathbf{y} \in \{0, 1\}^{|t|}$ denotes the distribution of entities users concern. More specifically, each entry in the vector reflects whether the corresponding entity in the transcript may be queried by users. To be implemented, we collect the entities issued in the queries. Thereafter, we set y_k as 1 if the kth word in the transcript is included in the collected query entities. Otherwise, the y_k is labeled as 0.

With the label mentioned above, we employ the binary cross entropy loss as the objective function, formally as,

$$\mathcal{L}_c = \min_{\Theta_c} -[\mathbf{y}\log(\mathbf{h}) + (\mathbf{1} - \mathbf{y})\log(\mathbf{1} - \mathbf{h})], \tag{10}$$

where $\mathbf{1} \in \mathbb{R}^{|t|}$ is an all-one vector. $\mathbf{h}$ is the entity classification scores derived from (4). Θ_c represents all to-be-learned parameters in the classification task, which includes visual encoder parameters Θ_v, textual encoder parameters Θ_t, and parameters in the cross-attention process and the $\mathrm{MLP}(\cdot)$ operation in (4).

2) Summary Generation: Query Sampling: In the setting of query-oriented micro-video summarization, query-video pairs are collected from search logs, which record users' interaction with micro-videos (e.g., clicking, watching, and rating). By aggregating multiple search logs, we can gather a set of diverse queries with different intents and expression manners that are associated with one micro-video. Under this real-world scenario, it is inevitable that a portion of queries are insufficient in the query-oriented summary generation. Such a training sample may cause bias in the model optimization [37]. To address this problem, we devise a sampling strategy to locate the effective samples for optimization. Instead of randomly choosing the target query for training, we propose to select more semantically relevant queries as the training target, which ensures the summarization optimization is conducted in the higher-quality data.

In particular, we use a pre-trained BERT model [38] to gain the representation of each query. For each micro-video, we can represent its corresponding queries and gather them as $\mathcal{G}^{rep} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_j, \ldots, \mathbf{g}_J\}$, where $\mathbf{g}_j$ is the representation vector of the jth query. With the obtained representations of queries, we conduct average operations on them as,

$$\bar{\mathbf{g}} = \frac{1}{J}\sum_{j=1}^{J}\mathbf{g}_j, \tag{11}$$

where $\bar{\mathbf{g}}$ is the averaged representation among all queries of the micro-video and we use $\bar{\mathbf{g}}$ to represent the whole query set.

Despite variations in expression, the effective queries for the same micro-video are based on the content of the micro-video and thus their expressions are prone to be centralized. Therefore, we can calculate the distance between each query with the average representation of the query set to discover the noisy queries. Formally, we score the semantic similarity between each query representation vector $\mathbf{g}_j$ and the overall query representation $\bar{\mathbf{g}}$ with cosine similarity as follows,

$$s_j = \frac{\mathbf{g}_j \cdot \bar{\mathbf{g}}}{\|\mathbf{g}_j\| \|\bar{\mathbf{g}}\|}, \tag{12}$$

where $\| \cdot \|$ stands for the norm of a vector. s_j is the cosine similarity score between the jth query and the average query representation. Analogously, we score the semantic similarities of all queries to the average representation, represented as $\mathbf{s} = [s_1, s_2, \ldots, s_J]$.

Thereafter, we deploy the filtering process as follows,

$$s_j^m = \begin{cases} s_j, & \text{if } s_j \geq \mu, \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where μ denotes the threshold hyperparameter. The similarity score after filtering out jth query is denoted as s_j^m.

After that, the remaining queries are still diverse in focus and expression manners. Among them, some generic queries (i.e., *How to cook dish* and *Funny videos*) frequently appear in the dataset, leading the model to overfit these universal queries if selecting the training target under the uniform probability distribution. However, these queries hardly reveal the micro-video's main idea, which is insufficient for the micro-video summarization task. Consequently, we devise a strategy to alleviate the effect of generic queries by suppressing their sampling probabilities. To be more specific, generic queries tend to have large similarity scores with the overall query representation since they contain comprehensive information. In contrast, the particular focus of specific queries varies, resulting in a low similarity score with the overall query representation. We then sample queries with higher similarity in a lower probability to encourage the model to learn more specific and detailed information. Formally, we derive the following probability distribution for sampling the query utilized in the learning process as,

$$p_j = \frac{(s_j^m)^{-1}}{\sum_{j=1}^{J} (s_j^m)^{-1}}, \tag{14}$$

where p_j is the probability of the jth query being selected for training. Large s_j^m results in low sampling probability p_j. Let $\mathcal{P} = [p_j]_{j=1}^{J}$ denote the distribution of the sampling probability of query set $\mathcal{G}$. In this way, we establish a negative correlation between the sampling probability and the similarity score of the query. For each training step, we sample a query g as the supervision from the query set $\mathcal{G}$ under the probability distribution $\mathcal{P}$.

Query-Oriented Confidence Scoring: As one of their characteristics, micro-videos tend to have large semantic gaps across modalities. In some cases, different modalities contribute to the summarization goal in various degrees. Under this condition, treating both modalities equally during training is suboptimal. In contrast, we propose to explicitly measure the contribution of each modality by calculating confidence scores. The quality of generated summaries can be estimated by counting the overlap between the summarization and the ground-truth. Consequently, to assess the importance of each modality in query-oriented summarization, we calculate the overlap between unimodal generated and the target summaries. Empirically, we utilize $\mathbf{V}$ and $\mathbf{T}$ to acquire the uni-modal generation results $\hat{g}_v$ and $\hat{g}_t$, respectively. Notably, the uni-modal generation process does not involve multi-modality interaction i.e., entity identification. Thereafter, we calculate the word-level overlap ratio between the generated summary and target queries using *Sørensen–Dice* coefficient [39] as follows,

$$\begin{cases} U_v = \max \frac{2N_s^v}{|g_j| + |\hat{g}_v|}, \ 1 \leq j \leq J, \\ U_t = \max \frac{2N_s^t}{|g_j| + |\hat{g}_t|}, \ 1 \leq j \leq J, \end{cases} \tag{15}$$

where g_j is the jth query in the query set. J stands for the total number of the corresponding queries. $|\hat{g}_v|$ and $|\hat{g}_t|$ refer to the number of words in $\hat{g}_v$ and $\hat{g}_t$, respectively. The overlap word number between $\hat{g}_v$ and g_j is denoted as N_s^v. Similarly, N_s^t is the overlap word number between $\hat{g}_t$ and g_j. We regard U_v and U_t as the confidence score of visual and textual modalities, respectively, in the summarization procedure.

After that, we assign different weight coefficients during training. To achieve that, besides using multi-modal representation to calculate the generation objective function, we supervise the training of uni-modal generation. The objective functions are written as follows,

$$\begin{cases} \mathcal{L}_g = \min\limits_{\Theta_v, \Theta_t, \Theta_c, \Theta_d} -\frac{1}{|g_p|} \sum_{l=1}^{|g_p|} \log p\left(g_{p,l} \mid g_{p,<l}, \mathbf{M}\right), \\ \mathcal{L}_g^v = \min\limits_{\Theta_v, \Theta_d} -\frac{1}{|g_p|} \sum_{l=1}^{|g_p|} \log p\left(g_{p,l} \mid g_{p,<l}, \mathbf{V}\right), \\ \mathcal{L}_g^t = \min\limits_{\Theta_t, \Theta_d} -\frac{1}{|g_p|} \sum_{l=1}^{|g_p|} \log p\left(g_{p,l} \mid g_{p,<l}, \mathbf{T}\right). \end{cases} \tag{16}$$

Wherein, the target summary is referred to as g_p, which is sampled from the query set $\mathcal{G}$ under the probability derived in (14), and $g_{p,l}$ stands for its lth words. $\mathcal{L}_g$ is the multi-modal generative loss. $\mathcal{L}_g^v$ and $\mathcal{L}_g^t$ denote the generative losses for visual and textual inputs, respectively. Θ_v, Θ_t, Θ_c, and Θ_d represent the trainable parameters in the visual encoder, textual encoder, multi-modal representation learning module, and decoder, respectively.

Thereafter, with the confidence scores derived above, we apply U_v and U_t as the weight coefficients for the visual and textual modalities in training. Ultimately, the objective function for generation is written as follows,

$$\bar{\mathcal{L}}_g = \frac{1}{3}[(U_v \cdot \mathcal{L}_g^v) + (U_t \cdot \mathcal{L}_g^t) + \mathcal{L}_g], \tag{17}$$

where $\bar{\mathcal{L}}_g$ stands for the generative loss derived from the weighting procedure.

3) Overall Objective: To conclude the training process, we combine the entity identification (10) and the summary generation (17) objective functions as the final objective function of query-oriented micro-video summarization, formally as,

$$\mathcal{L} = \bar{\mathcal{L}}_g + \lambda \cdot \mathcal{L}_c + \beta \cdot ||\Theta||_F, \tag{18}$$

TABLE II
STATISTICS OF OUR DATASET

	Train	Val	Test
#Micro-video	50,096	5,000	5,000
#Query	239,015	37,964	37,921
#V-Q Pair	427,963	42,854	42,683

#Micro-video, #Query, and #V-Q pair denote the number of micro-videos, queries, and video-query pairs, respectively.

where λ is a hyperparameter to control the combination ratio of two objective functions. We denote the regularization term as $\beta||\Theta||_F$, where $\Theta = \{\Theta_v, \Theta_t, \Theta_c, \Theta_d\}$ and β controls the strength of the regularization.

D. Generation Process

In the generation phase, the visual and textual encoders first process the multi-modal signals separately to derive two representations $\mathbf{V}$ and $\mathbf{T}$. After that, the visual and textual representations go through the entity-aware multi-modal representation learning module to locate the key entity information. We then derive the fused representation $\mathbf{M}$ through multi-modal integration, which is passed into the decoder to generate the query in an auto-regressive way. Specifically, the decoder predicts a probability distribution of the same size as the vocabulary table. After that, the word with the largest probability is sampled in each time step. The generation process terminates once an end-of-sentence (EOS) token is reached.

IV. Experiments

A. Dataset

1) Collection: To the best of our knowledge, there is no publicly available dataset tailored for the query-oriented micro-video summarization. Therefore, we constructed a large-scale query-oriented micro-video summarization dataset, dubbed QMV-Kwai, to train and evaluate our proposed QMS model. We collected the real-world queries associated with recalled micro-videos from the Kwai platform, instead of the manual annotation.

Specifically, we analyzed the user behaviors, like query-and-click operation, in the historical search logs and filtered more than 1,000,000 query-microvideo pairs as the candidates. This not only reflect the users' interest but also guarantee the relevance between the content of micro-video and their queries. To ensure the relevance between queries and micro-videos in our dataset, we retained only those pairs where the query was searched and the micro-video was clicked by more than 50 users. Then, we excluded pairs with an average user viewing time less than one-third of the total duration to mitigate the adverse effects of clickbait. Additionally, to capture a comprehensive perspective of micro-videos, we aggregated multiple search logs for each video, requiring a minimum of five matching queries. Micro-videos with fewer queries were excluded to enhance result reliability and validity. Overall, we totally achieved 687,125 microvideo-query pairs comprising 60,096 micro-videos and 331,414 queries, as listed in Table II.

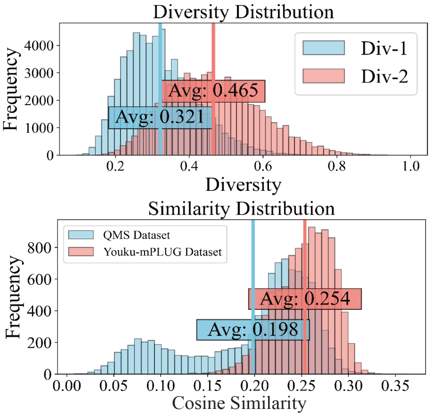


Fig. 3. Upper: Statistics of diversity metrics distribution for different micro-videos' query sets. Lower: Statistics of semantic similarities distribution for two video datasets.

2) Empirical Analysis: With the obtained dataset, we conduct extensive experiments on it to study the micro-videos' dense entities, cross-modal semantic gap, and diverse queries.

To analyze the entities in micro-videos, we extract their transcripts and adopt the entity recognition model to capture the entities mentioned in each micro-video. Based on these operations, we achieve the corpus of 60,096 textual transcripts comprising a total of 1,154,270 entities, with an average of 19.2 entities per transcript. By contrast, the queries in the dataset encompass a total of 927,719 entities across 427,963 entries, and on average, each query contains 2.2 entities. From the different numbers of entities in micro-videos and queries, we can find that only a few entities will be concerned by the users' queries. It justifies that it is necessary to recognize entities and identify the critical ones for micro-video understanding.

For the cross-modal semantic gap in micro-video, we resort to measuring the similarity between the visual and acoustic features. As a comparison, we perform the same operation on a large Chinese language video dataset, named Youku-mPLUG, to compare the semantic gaps between long videos and micro-videos. For fairness, we randomly sample 10,000 pairs from two datasets. Extracting frames and transcripts from each pair, we use the pre-trained visual and textual encoders from the CLIP model to learn feature vectors. Subsequently, we calculate cosine similarities to assess the cross-modal semantic gap. By separately performing the average operations on the similarity values of videos and micro-videos, we achieved an average semantic similarity of 0.254 in Youku-mPLUG and 0.198 in the QMS dataset. We suggest that this difference can verify our argument in Challenge 2. Moreover, we illustrate the similarity distribution of long videos and micro-videos in Fig. 3. Observing the results, we find that the similarity distribution of micro-videos is more uniform than that of long videos. Particularly, a substantial portion of micro-videos exhibits similarity centered around 0.08. We believe this distribution further validates the significant semantic gap in micro-video.

In addition, we also analyzed the diversity of the query set corresponding to each micro-video. Following [40], we measured the ratio of distinct n-grams per query set as the diversity score (Div-$\{1, 2\}$). The distribution of micro-videos' query set diversity score is illustrated in Fig. 3. As shown in Fig. 3, the Div-1 metric is concentrated in the range of 0.2 to 0.4, with a mean value of 0.32. The dist-2 metric has a wider distribution, with a mean value of 0.47. These statistics exhibit the difference in queries provided by different users and show the rationality of our concerns about noisy queries.

3) Pre-Processing: To represent the visual and acoustic information of the micro-videos, we conducted pre-trained deep learning models to extract their features. The micro-videos are processed using FFmpeg[6] to split into frames, capturing one frame per second [41], with a maximum of 32 keyframes per video. The CLIP [42] model is then used to extract visual features from these keyframes. Apart from the visual modality, the acoustic signal of the micro-video contains rich linguistic information, such as voiceovers, providing abundant cues for identifying and comprehending the micro-video. Instead of directly analyzing soundtracks, audio is converted to text transcripts using ASR tools from Kwai company, resulting in 59,074 transcripts averaging 183.1 words. These transcripts are standardized to a length of 184 words through truncation or padding. They are then encoded using a pre-trained BART [33] encoder, which is fine-tuned during training to enhance performance and adaptability. Moreover, to measure the semantic information of different queries (detailed in Section III-C2), we applied the pre-trained BERT [38] model to embed the micro-video query into a dense vector. Specifically, for each query, a sequence of representations for all words is generated and then averaged over the sequence length to produce a single vector representation.

B. Experimental Settings

1) Evaluation Settings: We evaluated the performance of our model from two perspectives: generation and retrieval.

Generation Evaluation: We adopted several mainstream natural language generation evaluation metrics, i.e., BLEU-{1,2,3,4} (B-{1,2,3,4}) [43], ROUGE-L (R-L) [44] and CIDEr (C) [45]. We used the official evaluation script[7] to obtain the results of the three metrics. The reported performance values are averaged over all samples in the testing set.

Retrieval Evaluation: Beyond the generation evaluation, we also tested whether the queries generated by our model can improve the performance in micro-video retrieval. For this purpose, we conducted the micro-video text-to-text retrieval experiment and compared the performance between QMS and several state-of-the-art baselines. To be specific, we used the user's queries to search for micro-videos by matching them with the generated summaries. It is worth noting that such a textual-matching-based retrieval algorithm is widely used in real scenarios as a coarse-grained searching operation [4] due to its efficiency. As to evaluate the retrieval performance, we used two standard retrieval metrics Recall@k (R@k) and Precision@k (P@k). Besides, we also adopted Accuracy@k (Acc@k) to measure the top-k micro-video retrieval accuracy, which denotes the proportion of queries whose top k retrieved results contain at least one of the target micro-video. We reported the evaluation results of $k \in \{1, 5, 10, 30\}$. (The Acc@1 result is omitted since it equals P@1).

2) Baselines: To validate our proposed model, we compared its performance with eight representative multi-modal natural language generation methods. These methods include four classical non-pretrained methods (Lead, Seq2Seq, Transformer, and MDVC) and four state-of-the-art pretrained methods (BART (Text), BART (MM), VG-BART (DP), and VG-BART (MH)). Our motivation for comparing both the non-pretrained models and the pretrained ones includes the following objectives: 1) demonstrating that our proposed model outperforms the non-pretrained models and pretrained models; 2) exploring the impact of pretrained language model on the multi-modal natural language generation task; and 3) validating the effectiveness of the designs in the QMS model by comparing it with the baselines built upon the same BART structure. **Lead [46]** directly captures the first P characters from the transcript. The average query length of the training set of our dataset equals 7.5, thus we took $P = 8$. **Seq2Seq [47]** utilizes a standard sequence-to-sequence architecture, which employs LSTMs [48] with a global attention mechanism for both the encoder and decoder. **Transformer [36]** is a classical structure in the natural language generation field. **MDVC [49]** is a video captioning approach that employs multiple Transformers to produce internal representations of multi-modal features by concatenation. **BART (Text) [33]** is a well-known generative pre-trained language model. The model follows the general transformer-based encoder-decoder architecture and only takes textual signals as inputs. **BART (MM) [33]** shares the same architecture with the former BART (Text) model, but it integrates the multi-modal signals by concatenation to perform multi-modal generation. **VG-BART (DP) [9]** is the state-of-the-art method for video summarization. It adds extra layers and employs fusion techniques based on dot product attention to adjust a uni-modal textual pre-trained BART model to leverage multi-modal inputs. **VG-BART (MH) [9]** is a variant of the former one, where the difference lies in the multi-head cross-attention mechanism used for multi-modal fusion.

3) Implementation Detail: Regarding the training details, the Lead model is a training-free method. Seq2Seq, Transformer, and MDVC are trained from scratch using the training set of our QMS dataset. As for the BART-based baselines, they are initialized with parameters pre-trained on the same large language corpus, and subsequently fine-tuned using our QMS dataset. Moreover, the hyper-parameters (e.g., learning rate and batch size) for all baselines are tuned based on their performance on the validation set. For the BART model, a variant[8] pre-trained on Chinese corpus is used. Similarly, a Chinese version of BERT[9] is employed. Both encoder and decoder have $L = 6$ layers, with representation dimensions (d_v, d_t, and d_a) equal to 768. Queries are padded to a fixed length of 20 characters to form mini-batches, covering the longest query length in

[6] [Online]. Available: https://ffmpeg.org.

[7] [Online]. Available: https://github.com/tylin/coco-caption.

[8] [Online]. Available: https://huggingface.co/fnlp/bart-base-chinese.

[9] [Online]. Available: https://huggingface.co/bert-base-chinese.

Algorithm 1: Training Procedure.

```
Input: training set $\mathcal{I}$ for optimizing the model,
  hyperparameters $\{\lambda, \beta\}$.
Output: Parameters $\Theta = \{\Theta_v, \Theta_t, \Theta_c, \Theta_d\}$.
Initialize all model parameters in $\Theta$;
for each sample $(x, \mathcal{G})$ do
  Generate entity classification label $\mathbf{y}$ w.r.t. (9).
  Calculate the target sampling probability distribution
    $\mathcal{P}$ w.r.t. (14).
end for
repeat
  Randomly sample a batch from $\mathcal{I}$.
  for each sample $(x, \mathcal{G})$ do
    Calculate $\mathcal{L}_c$ with $\mathbf{y}$ w.r.t. (10).
      Obtain $\hat{g}_v = \mathrm{Dec}(\mathbf{V}|\Theta_v, \Theta_d)$,
      $\hat{g}_t = \mathrm{Dec}(\mathbf{T}|\Theta_t, \Theta_d)$.
      $g \sim \mathcal{P}(\mathcal{G})$
      Obtain $U_v$ and $U_t$ w.r.t. (15).
      Apply $U_v$ and $U_t$ as the weight coefficients in $\bar{\mathcal{L}}_g$
      w.r.t. (17)
      Obtain overall objective
      $\mathcal{L} = \bar{\mathcal{L}}_g + \lambda \cdot \mathcal{L}_c + \beta \cdot ||\Theta||_F$
  end for
  Update $\Theta$ by optimizing $\mathcal{L}$.
until Converges.
```

the training set. Hyperparameters (λ, β, τ, μ) are determined using a grid search on the validation set, testing values in the set $0.25, 0.5, 0.75, 1$. Ultimately, λ and β are set to 0.5, while two threshold parameters, τ and μ, are set to 0.5 and 0.75, respectively. To optimize the models, we adopted Adam [50] with the initial learning rate $5e^{-5}$. A learning rate scheduler named cosine annealing [51] is used to reduce the learning rate progressively. The mini-batch sizes for training and generation are both 64. It takes around 15 epochs for the training stage to get the peak performance on the validation set. During the generation stage, we used the beam search with a beam size of 4 for all models to generate the queries. The overall algorithm of the optimization is briefly summarized in the Algorithm 1.

The experiment codes were implemented with the Pytorch [52] deep learning platform on version 1.8.1. We used an NVIDIA Tesla V100 32 GB graphics card for all of the experiments.

C. Performance Comparison

We first conducted a comparison between our proposed model and the baselines on the constructed datasets, respectively. Specifically, we listed their results *w.r.t* BLEU-{1,2,3,4}, Rouge-L, and CIDEr in Table III, where $Improvement.\%$ represents the relative improvement of the best-performing method (bolded) over the strongest baselines (underlined). According to the results, we gain the following observations.

TABLE III
EXPERIMENTAL RESULTS ON THE QMS AND BASELINE METHODS

Model	B-1	B-2	B-3	B-4	R-L	C
Lead	18.73	13.98	10.82	8.43	15.67	46.51
Seq2Seq	56.04	47.72	39.57	33.31	42.29	131.29
Transformer	58.73	51.15	44.12	38.76	46.30	164.95
MDVC	59.30	52.38	45.81	40.50	47.13	174.75
BART (Text)	52.68	47.45	42.51	38.53	42.36	176.06
BART (MM)	64.44	58.07	51.90	46.87	51.43	208.86
VG-BART (DP)	65.78	59.16	52.82	47.73	51.91	211.51
VG-BART (MH)	66.19	59.40	52.91	47.68	52.39	212.29
QMS	**67.38**	**61.29***	**55.35***	**50.50***	**54.25***	**226.24***
Improvement. %	1.80	3.18	4.61	5.80	3.55	6.57

The symbol * means statistically significant improvement over the strongest baseline with $p < 0.05$.

- As expected, the QMS model significantly outperforms the baselines on the constructed dataset. In particular, our proposed model outperforms the strongest baselines *w.r.t.* BLEU-4 by 5.9%, ROUGE-L by 3.6%, and CIDEr by 6.5%, respectively. It demonstrates the effectiveness of our proposed model. More importantly, among the improvements *w.r.t.* BLEU-{1,2,3,4} over the strongest baselines, the QMS model gains the largest increase *w.r.t.* the 4-gram score (BLEU-4), which indicates that QMS is superior on fluency and accuracy, especially in terms of longer phrases.
- Comparing with the non-pretrained models, pretrained models, i.e., BART (MM), VG-BART (DP), VG-BART (MH), and ours achieve better results across all six metrics. It verifies that the pretrained language model can provide the prior knowledge to help the multi-modal micro-video understanding and natural language generation. Nevertheless, we also find that the pretrained BART (Text) performs worse than the non-pretrained models in most cases. It indicates that in the multi-modal summarization, relying solely on a pre-trained language model is insufficient. Understanding visual information and effectively fusing multi-modal information are equally vital. This further demonstrates the rationale and effectiveness of our proposed model.
- Although the VG-BART (DP) and VG-BART (MH) employ distinct multi-modal fusion techniques, the performances of the two baselines are comparable. One possible reason is that various fusion methods directly applied to raw multi-modal features cannot eliminate the semantic gaps between multi-modalities, which is one of the micro-videos' inherent characteristics. In contrast, by measuring the semantic contribution and calculating confidence scores of different modalities, the QMS model gains considerable improvement under this situation.
- The QMS model outperforms the BART (MM) model despite they share the same Transformer-based encoder-decoder architecture. It implies that the concatenation of multi-modal representations cannot precisely integrate the complex entity information, thus failing to satisfy the query-oriented summarization task. In contrast, the QMS model is equipped with the entity-aware multi-modal representation learning module, which helps the model to capture the critical entity information based on the fine-grained intention embodied in the queries.

TABLE IV
ABLATION STUDY RESULTS OF THE PROPOSED QMS

Model	B-1	B-2	B-3	B-4	R-L	C
w/o-Visual	51.37	46.55	42.06	38.40	41.85	173.90
w/o-Textual	50.80	43.56	36.89	31.80	41.13	128.62
w/o-Scores	64.95	58.55	52.36	47.31	51.89	209.85
w/o-Entity	65.44	59.09	52.93	47.96	52.65	215.54
w/o-Qsample	66.12	59.86	53.84	48.92	52.99	218.64
QMS	**67.38**	**61.29**	**55.35**	**50.50**	**54.25**	**226.24**

The best results are highlighted in bold.

D. Ablation Study

1) Impact of Components: To verify that all the designed modules and strategies in our model are indispensable for the query-oriented micro-video summarization, we compared the original QMS model with the following derivatives:

- *QMS-w/o-T:* and *QMS-w/o-V*. In these derivatives, we only kept the visual frames or the textual transcripts of the micro-video to verify the indispensability of textual and visual modality, respectively. To accomplish this, we fed the single modal features into the model and replaced the cross-attention with self-attention in the model.
- *QMS-w/o-Scores:* In this derivative, we disabled the query-oriented confidence scoring module during optimization by directly using the multi-modal generative loss $\mathcal{L}_g$.
- *QMS-w/o-Entity:* In this derivative, we discarded the entity-aware multi-modal representation learning module by utilizing the original textual representation $\mathbf{T}$ without fine-grained entity information. We omitted the entity identification objective function $\mathcal{L}_c$ in (18) as well.
- *QMS-w/o-Qsample:* In this derivative, we discarded the query sampling strategy and randomly selected a query as the training target g. In other words, the probability is equal for all corresponding queries to be chosen as the target.

We summarised the experimental results of the ablation study in Table IV and obtained the following observations: (1) QMS outperforms all three derivatives that omit one part in the model, i.e., QMS-w/o-Scores, QMS-w/o-Entity, and QMS-w/o-Qsample, across different evaluation metrics, which indicates that each derived module and strategy in the proposed QMS model is indispensable in terms of the query-oriented micro-videos summarization task. (2) QMS exceeds both QMS-w/o-V and QMS-w/o-T, demonstrating the necessity of multi-modal information, and removing either modality significantly drops the performance of generation. And (3) the model without the query-oriented confidence scoring module gains the most unsatisfactory result compared to other multi-modal derivatives. One possible explanation is that the semantic gap in the different modalities largely affects the learning of the model, which cannot be remedied by the other parts of the model.

2) Impact of Backbone: It is consensus that the model's performance is greatly influenced by its backbone, especially in the era of the recent flourishing pre-training. Therefore, we explored the effect of using different backbones: *Transformer* without pre-training. **mT5 [53]** is a large-scale pre-trained multilingual text-to-text transformer model. Specifically, we chose a version that is trained on an abstractive summarization task [54] to close the gap with the micro-video summarization task. **BART [33]** is pretrained on Chinese corpus, including Chinese Wikipedia and WuDaoCorpus [55].

Table VI shows the comparison of using different backbones. Without any doubt, the Transformer backbone is surpassed by the other two pre-trained backbones, reflecting the necessity of pre-training. As to the comparison between mT5 and BART, the latter exhibits slightly better performance. One conjecture is that the BART model is trained only on Chinese corpus, resulting in a stronger ability when dealing with Chinese domain data. In contrast, the pre-training corpus for mT5 model covers 101 languages, which might constrain its competence in Chinese data. Thus, we chose BART model as the default backbone of QMS.

3) Impact of Visual Feature Extractor: We studied the impact of using different visual feature extractors and chose a prominent one to obtain a better visual feature. We tested the following three extractors: **Resnet50** [56] is a residual learning method and it wildly serves as the backbone in many video-related tasks [57], [58]. **I3D [59]** is an inflated 3D ConvNet model. We adopted the parameters that are pre-trained on the Charades [60] video dataset. **CLIP [42]** is a dual model containing an image encoder and a text encoder. It is pre-trained on 400 million image-text pairs obtained from the Internet.

Table VII shows the effectiveness of various feature extractors in processing visual content. The CLIP feature extractor stands out due to its large pre-training dataset and relevance to web images and micro-videos. Besides, I3D features underperform Resnet features. One possible explanation is that static information in video frames (e.g., entities, items) often conveys more meaning in micro-videos than action information. Hence, frame-based feature extractors like Resnet and CLIP achieve better results.

E. Retrieval Evaluation

In this section, we validated how the generated summaries improve the retrieval task. We compared the retrieval evaluation results based on the generated summaries by seven baselines and the QMS model. Specifically, we adopted two well-known textual-based retrieval methods: BM25 [61] and Dense Passage Retriever (DPR) [62]. *BM25* is a lexical-based matching method. It uses an inverted index to score the similarities between each query and the generated summaries of micro-videos. And *DPR* is an up-to-date dense retriever. It first leverages the standard pre-trained BERT model and a dual-encoder architecture to embed queries and summaries into dense representations. Thereafter, the similarity scores are calculated with the maximum inner product. Notably, for DPR, we adopted two versions: a pre-trained version[10] (PT), and a fine-tuned (FT) one using our micro-video dataset.

For both retrieval methods, we took the micro-videos with the Kth ($K \in \{1, 5, 10, 30\}$) highest similarity scores as the retrieval results. Observing the retrieval results listed in Table V, we obtain the following insights:

[10] [Online]. Available: http://tinyurl.com/bdj95dua.

TABLE V
PERFORMANCE COMPARISON FROM THE RETRIEVAL PERSPECTIVE

Retriever	Model	R@1	R@5	R@10	R@30	P@1	P@5	P@10	P@30	Acc@5	Acc@10	Acc@30
BM25	Seq2Seq	8.86	16.88	20.52	25.38	17.88	7.19	4.53	1.99	38.40	47.42	60.58
	Transformer	14.30	25.20	29.53	34.60	27.82	10.66	6.49	2.68	48.82	57.34	67.90
	mdvc	14.23	25.87	30.59	36.55	27.22	11.23	6.96	2.89	50.26	58.34	68.24
	BART (Text)	19.71	19.12	31.79	35.54	33.94	11.09	6.33	2.45	50.80	55.98	63.24
	BART (MM)	19.93	29.98	33.58	37.97	34.58	11.54	6.77	2.73	55.70	63.00	71.58
	VG-BART (DP)	19.97	31.90	36.20	40.89	35.20	12.79	7.66	3.06	57.46	64.74	73.48
	VG-BART (MH)	19.62	31.04	35.43	40.47	35.50	12.59	7.48	3.01	57.32	63.98	73.34
	QMS	**21.28**	**32.50**	**36.99**	**42.04**	**37.32**	**12.85**	**7.73**	**3.12**	**59.92**	**66.86**	**75.20**
	Improvement. %	6.56	1.88	2.18	2.81	5.13	0.47	0.91	1.96	4.28	3.27	2.34
DPR (PT)	Seq2Seq	2.83	8.90	14.04	24.87	6.46	4.76	4.04	2.62	18.30	25.98	40.20
	Transformer	5.76	16.00	21.72	33.07	13.18	8.29	6.07	3.38	30.32	38.22	50.84
	MDVC	4.31	12.84	18.95	30.86	10.18	6.68	5.08	3.03	25.30	34.18	48.62
	BART (Text)	8.55	17.64	22.44	29.31	15.54	7.48	5.07	2.36	29.10	35.78	45.22
	BART (MM)	8.84	19.83	26.00	37.27	17.08	9.40	6.75	3.58	33.10	41.24	53.66
	VG-BART (DP)	8.80	19.88	26.33	37.81	17.04	9.43	6.71	3.60	35.54	43.54	56.78
	VG-BART (MH)	8.48	20.01	25.87	36.87	16.70	9.54	6.70	3.55	34.30	42.30	54.82
	QMS	**9.56**	**21.94**	**28.76**	**39.81**	**18.70**	**10.43**	**7.34**	**3.78**	**36.90**	**45.38**	**57.44**
	Improvement. %	8.14	9.65	9.23	5.29	9.48	9.33	8.74	5.00	3.83	4.23	1.16
DPR (FT)	Seq2Seq	7.68	20.33	27.86	39.60	15.80	10.11	7.46	3.87	35.90	44.34	56.86
	Transformer	13.22	28.60	35.45	46.97	25.84	13.70	9.20	4.45	46.76	54.48	65.70
	MDVC	13.08	28.65	35.80	47.33	25.24	13.56	9.08	4.44	47.00	54.52	64.58
	BART (Text)	17.42	30.35	34.92	41.55	30.30	12.80	7.83	3.33	46.88	51.96	59.16
	BART (MM)	17.99	33.26	39.99	50.60	31.72	14.67	9.65	4.55	52.28	59.18	68.42
	VG-BART (DP)	18.25	35.13	42.08	53.11	33.06	15.82	10.28	4.81	53.52	60.34	69.52
	VG-BART (MH)	17.63	34.30	41.37	52.16	31.98	15.43	10.01	4.74	53.50	60.46	69.80
	QMS	**19.69**	**37.36**	**44.15**	**54.32**	**35.40**	**16.70**	**10.72**	**4.86**	**56.40**	**62.96**	**71.74**
	Improvement. %	9.45	6.35	4.92	2.28	7.08	5.56	4.28	1.04	5.38	4.13	2.78

We use bold font to highlight the best results in each group of different retrieval settings.

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT BACKBONES

Model	B-1	B-2	B-3	B-4	R-L	C
Transformer	62.20	55.21	48.66	43.47	49.99	192.24
mT5	65.98	59.73	53.73	48.89	53.23	220.31
QMS	**67.40**	**61.30**	**55.30**	**50.50**	**54.30**	**226.20**

The best results are highlighted in bold.

TABLE VII
PERFORMANCE COMPARISON OF DIFFERENT VISUAL FEATURE EXTRACTORS

Model	B-1	B-2	B-3	B-4	R-L	C
Resnet	63.71	57.51	51.61	46.86	51.38	211.45
I3D	53.76	47.28	41.40	36.89	43.85	163.64
CLIP	**67.40**	**61.30**	**55.30**	**50.50**	**54.30**	**226.20**

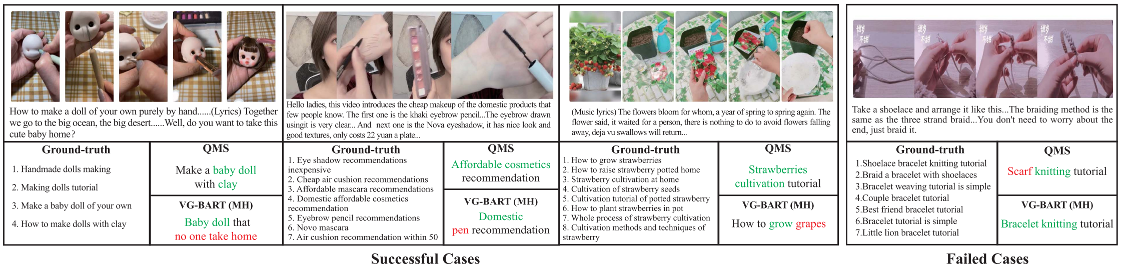
The best results are highlighted in bold.

(1) Under all three retrieval settings, QMS outperforms the baselines *w.r.t.* P@{1, 5, 10, 30}, R@{1, 5, 10, 30}, and Acc@{5, 10, 30} by a margin, revealing the superiority of the QMS model in facilitating the micro-video retrieval task. (2) Different retrieval methods affect the retrieval performance significantly. For R@1, P@1, and Acc@$\{5, 10, 30\}$ metrics, all models gain their best performance under the BM25 retriever. While for other metrics, DPR (FT) retriever attains higher performance than using the BM25 or DPR (PT) retrieval methods. Besides, fine-tuning the neural-based DPR retriever can improve performance by a large margin. These observations reveal the importance of adopting suitable retrieval settings regarding specific requirements. (3) Although the VG-BART(MH) model outperforms or equals the VG-BART(DP) model regarding all generative-based metrics, this advantage seems hard to be transferred to the retrieval task. Instead, under the BM25 retrieval setting, the VG-BART(DP) model outperforms the VG-BART(MH) model, except for only one retrieval metric (i.e., P@{1}). And similar observation can be seen under the DPR(FT) retrieval setting, where the VG-BART(DP) exhibits better performance in most retrieval metrics. One possible explanation is that the summary generated by the VG-BART(MH) model contains more generic words (i.e., *teach, what, how*) and therefore performs better in the generation metrics. In contrast, the VG-BART(DP) model generates more discriminative summaries, which are more beneficial for the retrieval task. To take a step further, the QMS model is equipped with the entity-aware multi-modal representation learning module, whose role is to identify the specific and concrete key entities in the transcript and reflect it in the generated summary. Thus, the summaries generated by QMS can better empower the retrieval task.

F. Case Study

To qualitatively demonstrate our proposed model, we illustrated several micro-videos associated with the generated summaries. In particular, we randomly sampled six micro-videos from the testing dataset and conducted QMS and VG-BART (MH) models on them. In addition, we exhibited the micro-videos and the extracted transcripts in Fig. 4.

- From the first sample, we observe that QMS can precisely understand the content of the micro-video and highlight the key entity information (*baby doll, clay*) by leveraging the visual and textual cues. Therefore it generates a reasonable summary. While the counterpart is confused by the diverse entities and expressions in the transcript (*take the baby home*) and mistakenly generates an irrelevant phrase (*no one take home*), thus failing to capture the critical message.
- The second sample illustrates a micro-video with a diverse query set, which includes various search intents including



How to make a doll of your own purely by hand......(Lyrics) Together we go to the big ocean, the big desert......Well, do you want to take this cute baby home?

Ground-truth
1. Handmade dolls making
2. Making dolls tutorial
3. Make a baby doll of your own
4. How to make dolls with clay

QMS
Make a baby doll with clay

VG-BART (MH)
Baby doll that no one take home

Hello ladies, this video introduces the cheap makeup of the domestic products that few people know. The first one is the khaki eyebrow pencil...The eyebrow drawn usingit is very clear... And next one is the Nova eyeshadow, it has nice look and good textures, only costs 22 yuan a plate...

Ground-truth
1. Eye shadow recommendations inexpensive
2. Cheap air cushion recommendations
3. Affordable mascara recommendations
4. Domestic affordable cosmetics recommendation
5. Eyebrow pencil recommendations
6. Novo mascara
7. Air cushion recommendation within 50

QMS
Affordable cosmetics recommendation

VG-BART (MH)
Domestic pen recommendation

(Music lyrics) The flowers bloom for whom, a year of spring to spring again. The flower said, it waited for a person, there is nothing to do to avoid flowers falling away, deja vu swallows will return...

Ground-truth
1. How to grow strawberries
2. How to raise strawberry potted home
3. Strawberry cultivation at home
4. Cultivation of strawberry seeds
5. Cultivation tutorial of potted strawberry
6. How to plant strawberries in pot
7. Whole process of strawberry cultivation
8. Cultivation methods and techniques of strawberry

QMS
Strawberries cultivation tutorial

VG-BART (MH)
How to grow grapes

Take a shoelace and arrange it like this...The braiding method is the same as the three strand braid...You don't need to worry about the end, just braid it.

Ground-truth
1.Shoelace bracelet knitting tutorial
2.Braid a bracelet with shoelaces
3.Bracelet weaving tutorial is simple
4.Couple bracelet tutorial
5.Best friend bracelet tutorial
6.Bracelet tutorial is simple
7.Little lion bracelet tutorial

QMS
Scarf knitting tutorial

VG-BART (MH)
Bracelet knitting tutorial

Fig. 4. Illustration of several examples to show the qualitative results of QMS compared with the strongest baseline. Some vital words that imply the main semantic or search intentions are highlighted in green. We also mark the generation errors in red font.

eye shadows, air cushion, cosmetics, mascara and *eyebrow pencil*. The QMS model generates an adequate summary that aggregates the intention information of the micro-video i.e., *affordable cosmetics*. While the summary from the baseline model, i.e., *domestic pen recommendation*, is deficient in summarizing the primary idea of the micro-video.

- When the semantics in the visual and textual modalities differ, as shown in the third sample, the visual modality demonstrates strawberry cultivation skills. While the textual modality contains the lyrics of a song about flowers. Owning to the confidence scores assignment in optimization, the QMS model can identify that visual modality contributes more to the summary. However, the counterpart, which is confused by the semantic gap of two modalities, mistakenly generates irrelevant information (*grapes*).
- We also show some failed cases in the last column. Under the visual demonstration and the textual description of knitting, our model directly associates it with *scarf* knitting and makes the wrong generation.

With these samples, we demonstrate that the proposed QMS model can capture the essential information from complex micro-video scenarios and generate queries that concisely summarize users' search intentions.

V. Conclusion and Future Work

In this paper, we propose a query-oriented micro-video summarization model to generate summaries for micro-videos, named QMS, which can precisely summarize the critical content of the micro-video and further boost the retrieval performance. To verify the effectiveness of the proposed scheme, we first collect a dataset from logs of the real-life micro-video search engine, and then conduct extensive experiments, demonstrating that QMS outperforms all baselines on all generation metrics. More importantly, QMS also shows superiority in retrieval-based metrics, suggesting the potential to facilitate searching efficiency. In the future, we plan to expand the model to relevant tasks like micro-video query refining and suggestion.

References

[1] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Trans. Image Process.*, vol. 29, pp. 1–14, 2019.

[2] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1437–1445.

[3] H. Jiang, W. Wang, Y. Wei, Z. Gao, Y. Wang, and L. Nie, "What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3487–3495.

[4] K. Kim, K. Lee, S. Hwang, Y. Song, and S. Lee, "Query generation for multimodal documents," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 659–668.

[5] Y. Wang, "Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 17, no. 1s, pp. 10:1–10:25, 2021.

[6] N. Liu, X. Sun, H. Yu, W. Zhang, and G. Xu, "Multistage fusion with forget gate for multimodal summarization in open-domain videos," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 1834–1845.

[7] X. Shang, Z. Yuan, A. Wang, and C. Wang, "Multimodal video summarization via time-aware transformers," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1756–1765.

[8] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "Automatic generation of descriptive titles for video clips using deep learning," in *Proc. Adv. Artif. Intell. Appl. Cogn. Comput.*, 2021, pp. 17–28.

[9] T. Yu, W. Dai, Z. Liu, and P. Fung, "Vision guided generative pre-trained language models for multimodal abstractive summarization," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 3995–4007.

[10] H. Xu et al., "Youku-mPLUG: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks," 2023, *arXiv:2306.04362*.

[11] L. Nie et al., "Enhancing micro-video understanding by harnessing external sounds," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1192–1200.

[12] J. Zhang, L. Nie, X. Wang, X. He, X. Huang, and T. Chua, "Shorter-is-better: Venue category estimation from micro-video," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1415–1424.

[13] C. Zhu, Z. Yang, R. Gmyr, M. Zeng, and X. Huang, "Leveraging lead bias for zero-shot abstractive news summarization," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1462–1471.

[14] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 368–378, Sep. 2003.

[15] N. Liu, X. Sun, H. Yu, W. Zhang, and G. Xu, "D-MmT: A concise decoder-only multi-modal transformer for abstractive summarization in videos," *Neurocomputing*, vol. 456, no. 16, pp. 179–189, 2021.

[16] J. Kupiec, J. O. Pedersen, and F. Chen, "A trainable document summarizer," in *Proc. 18th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1995, pp. 68–73.

[17] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for how2 videos," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6587–6596.

[18] A. Miech, J. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Thinking fast and slow: Efficient text-to-visual retrieval with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9826–9836.

[19] Y. Wang, X. Lin, L. Wu, and W. Zhang, "Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1393–1404, Mar. 2017.

[20] B. Qian, Y. Wang, R. Hong, and M. Wang, "Adaptive data-free quantization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7960–7968.

[21] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, and N. Bhamidipati, "Context- and content-aware embeddings for query rewriting in sponsored search," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2015, pp. 383–392.

[22] W. U. Ahmad, K. Chang, and H. Wang, "Context attentive document ranking and query suggestion," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 385–394.

[23] R. Nogueira, W. Yang, J. Lin, and K. Cho, "Document expansion by query prediction," 2019, *arXiv:1904.08375*.

[24] Y. Lien, R. Zhang, F. M. Harper, V. Murdock, and C. Lee, "Leveraging customer reviews for e-commerce query generation," in *Proc. Eur. Conf. Inf. Retrieval*, 2022, pp. 190–198.

[25] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.

[26] M. Xie and S. Huang, "Partial multi-label learning with noisy label identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3676–3687, Jul. 2022.

[27] K. Lee, S. Yun, K. Lee, H. Lee, B. Li, and J. Shin, "Robust inference via generative classifiers for handling noisy labels," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 3763–3772.

[28] A. K. Menon, A. S. Rawat, S. J. Reddi, and S. Kumar, "Can gradient clipping mitigate label noise?," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–26.

[29] Y. Yao et al., "Dual T: Reducing estimation error for transition matrix in label-noise learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 7260–7271.

[30] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13723–13732.

[31] L. Jiang, Z. Zhou, T. Leung, L. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2309–2318.

[32] Z. Wu, T. Wei, J. Jiang, C. Mao, M. Tang, and Y. Li, "NGC: A unified framework for learning with open-world noisy data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 62–71.

[33] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[34] Y. Liu et al., "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, no. 47, pp. 726–742, 2020.

[35] Y. Liu, Y. Wan, L. He, H. Peng, and P. S. Yu, "KG-BART: Knowledge graph-augmented BART for generative commonsense reasoning," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 6418–6425.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[37] S. Song, J. Liu, L. Lin, and Z. Guo, "Learning to recognize human actions from noisy skeleton data via noise adaptation," *IEEE Trans. Multimedia*, vol. 24, pp. 1152–1163, 2022.

[38] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[39] T. A. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter*, vol. 5, pp. 1–34, 1948.

[40] J. Aneja, H. Agrawal, D. Batra, and A. G. Schwing, "Sequential latent spaces for modeling the intention during diverse image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4260–4269.

[41] Y. Wei, X. Wang, L. Nie, X. He, and T.-S. Chua, "Graph-refined convolutional network for multimedia recommendation with implicit feedback," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3541–3549.

[42] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[43] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[44] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

[45] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.

[46] L. Xing, W. Xiao, and G. Carenini, "Demoting the lead bias in news summarization via alternating adversarial learning," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 948–954.

[47] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[49] V. Iashin and E. Rahtu, "Multi-modal dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 4117–4126.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[51] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–16.

[52] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[53] L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2021, pp. 483–498.

[54] T. Hasan et al., "XL-sum: Large-scale multilingual abstractive summarization for 44 languages," in *Proc. Annu. Meeting Assoc. Comput. Linguistics, Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4693–4703.

[55] Y. Shao et al., "CPT: A pre-trained unbalanced transformer for both chinese language understanding and generation," 2021, *arXiv:2109.05729*.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[57] J. Wang, J. Jiao, L. Bao, S. He, W. Liu, and Y. Liu, "Self-supervised video representation learning by uncovering spatio-temporal statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3791–3806, Jul. 2022.

[58] Z. Wu, H. Li, C. Xiong, Y. Jiang, and L. S. Davis, "A dynamic frame selection framework for fast video recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1699–1711, Apr. 2022.

[59] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[60] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 510–526.

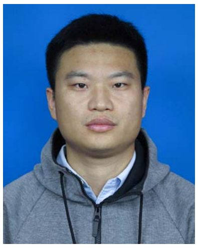
[61] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. Text REtrieval Conf.*, 1994, pp. 109–126.

[62] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 6769–6781.

Mengzhao Jia received the BE degree from the School of Information Science and Engineering, Shandong University, Qingdao, in 2020. She is currently working toward the graduate degree with the School of Computer Science and Technology, Shandong University. Her research interests include multimodal learning and natural language generation.

Yinwei Wei (Member, IEEE) received the MS degree from Tianjin University, and the PhD degree from Shandong University, respectively. Currently, he is a research fellow with Monash University. His research interests include multimedia computing and recommendation. Several works have been published in top forums, such as ACM MM, *IEEE Transactions on Multimedia*, and *IEEE Transactions on Image Processing*. He has served as the PC member for several conferences, such as MM, AAAI, and IJCAI, and the reviewer for *IEEE Transactions on Multimedia*, *IEEE Transactions on Knowledge and Data Engineering*, and *IEEE Transactions on Image Processing*.

Xuemeng Song (Senior Member, IEEE) received the BE degree from the University of Science and Technology of China, in 2012, and the PhD degree from the School of Computing, National University of Singapore, in 2016. She is currently an associate professor with Shandong University, China. She has published several papers in the top venues, such as ACM SIGIR, MM, and *ACM Transactions on Information Systems*. Her research interests include information retrieval and social network analysis. She has served as a reviewer for many top conferences and journals. She is also an AE of *IEEE Transactions on Circuits and Systems for Video Technology* and *IET Image Processing*.



Teng Sun received the master's degree from the School of Computer Science and Technology, Shandong University, Shandong, in 2020. He is currently working toward the PhD degree with the School of Computer Science and Technology, Shandong University. His research interests include multimedia computing and cross-modal information retrieval. Several works have been published in top forums, such as ACM MM and *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Min Zhang (Member, IEEE) received the bachelor's and PhD degrees in computer science from the Harbin Institute of Technology, Harbin, in 1991 and 1997, respectively. He is currently a professor with the Harbin Institute of Technology (Shenzhen), China. His current research interests include natural language processing, multi-model, artificial intelligence and cryptology.

Liqiang Nie (Senior Member, IEEE) received the BEng and PhD degrees from Xi'an Jiaotong University and National University of Singapore (NUS), respectively. He is IAPR fellow, is currently the dean with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen campus). His research interests lie primarily in multimedia content analysis and information retrieval. He has co-/authored more than 100 CCF-A papers and 5 books, with 16 k plus Google Scholar citations. He is an AE of *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Information Science*. Meanwhile, he is the regular area chair or SPC of ACM MM, NeurIPS, IJCAI and AAAI. He is a member of ICME steering committee. He has received many awards more than the past three years, like ACM MM and SIGIR best paper honorable mention, in 2019, the AI 2000 most influential scholars 2020, SIGMM rising star, in 2020, MIT TR35 China 2020, DAMO Academy Young Fellow, in 2020, SIGIR best student paper, in 2021, first price of the provincial science and technology progress award in 2021 (rank 1), and provincial youth science and technology award, in 2022. Some of his research outputs have been integrated into the products of Alibaba, Kwai, and other listed companies.