Personalized Multi-modal Video Retrieval on Mobile Devices

Haotian Zhang
haotian.zhang@samsung.com
Samsung AI Center - Toronto
Canada

Allan D. Jepson
allan.jepson@samsung.com
Samsung AI Center - Toronto
Canada

Iqbal Mohomed
i.mohomed@samsung.com
Samsung AI Center - Toronto
Canada

Konstantinos G. Derpanis
k.derpanis@samsung.com
Samsung AI Center - Toronto
Canada

Ran Zhang
ran.zhang@samsung.com
Samsung AI Center - Toronto
Canada

Afsaneh Fazly
a.fazly@samsung.com
Samsung AI Center - Toronto
Canada

ABSTRACT

Current video retrieval systems on mobile devices cannot process complex natural language queries, especially if they contain personalized concepts, such as proper names. To address these shortcomings, we propose an efficient and privacy-preserving video retrieval system that works well with personalized queries containing proper names, without re-training using personalized labelled data from users. Our system first computes an initial ranking of a video collection by using a generic attention-based video-text matching model (i.e., a model designed for non-personalized queries), and then uses a face detector to conduct personalized adjustments to these initial rankings. These adjustments are done by reasoning over the face information from the detector and the attention information provided by the generic model. We show that our system significantly outperforms existing keyword-based retrieval systems, and achieves comparable performance to the generic matching model fine-tuned on plenty of labelled data. Our results suggest that the proposed system can effectively capture both semantic context and personalized information in queries.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Retrieval models and ranking**;

KEYWORDS

Video Retrieval; Personalization; Multi-modal Learning

ACM Reference Format:
Haotian Zhang, Allan D. Jepson, Iqbal Mohomed, Konstantinos G. Derpanis, Ran Zhang, and Afsaneh Fazly. 2021. Personalized Multi-modal Video Retrieval on Mobile Devices. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3474085.3481545

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3481545

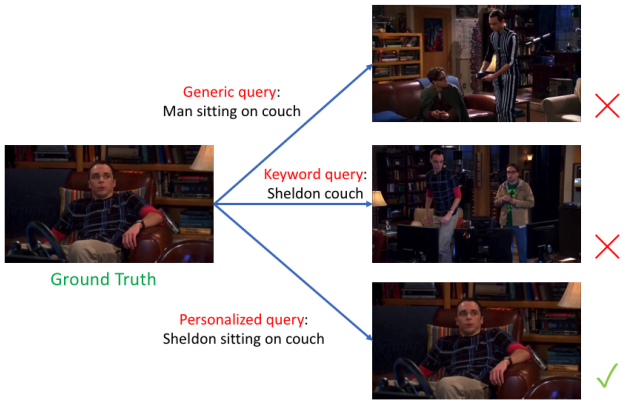


Figure 1: Importance of addressing personalized queries. Whereas the keyword query lacks enough semantic information, the generic query is missing personalized details.

1 INTRODUCTION

Current on-device video retrieval systems on constrained computing devices (e.g., smart phones and tablets) are mainly keyword-based and often fail to capture the rich semantic information contained in natural language queries. Language-based video retrieval is an active research area in the multimedia retrieval community [3], where videos and text queries are projected into a joint embedding space [5, 10, 11, 13], enabling search and retrieval across the different modalities. Despite these advances in cross-modal retrieval, most such techniques focus on handling generic queries (e.g., "Man sitting on the couch"), and as such cannot understand personalized queries such as those containing proper names (e.g., "Sheldon sitting on the couch"). We posit that not being able to handle personalized queries on-device (i.e., without resorting to cloud access) is a key barrier to the adoption of semantic approaches on today's smart phones. We refer to the problem of handling personalized information (e.g., proper names) in natural language queries as the *personalized video retrieval problem*; see Figure 1 for an illustration of this problem. We emphasize that this is distinct from, e.g., personalized recommendation systems, where users are presented with customized results according to their stored preferences. We use the term *personalized* here since our retrieval system again needs to draw on information that is unique to specific users, e.g., the names and faces of their family members.

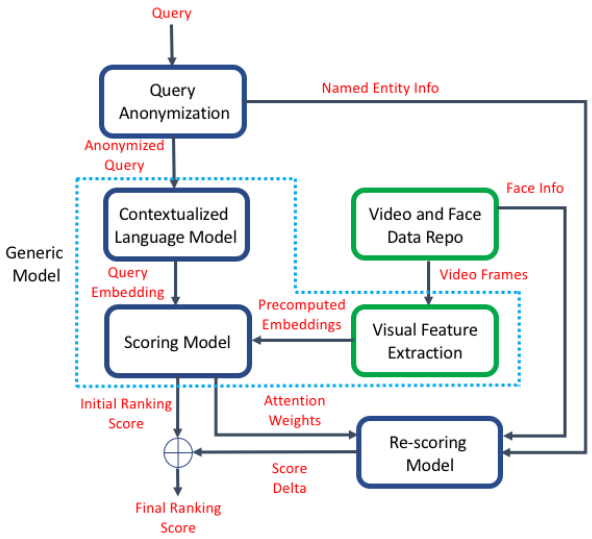


Figure 2: Architecture of mySCAN. Green and blue blocks correspond to offline and online components, respectively.

The key challenge of adapting current retrieval approaches to personalized video retrieval is the need to re-train the retrieval system with personalized training data. Specifically, current approaches require, for each user, a retrieval system to be trained (or fine-tuned) on a combination of generic training data and annotated personal data from that user. This annotated data is needed to visually ground the personal concepts indicated by the user's use of proper nouns like "Sheldon" in the example above. The process of collecting sufficient personalized training data for each user is awkward and practically infeasible (specifically, one would require assembling a dataset of video-text captions that include proper nouns/personalized entities); and moreover, sending such personal data to the cloud for training endangers user privacy.

We address the above-mentioned challenges by proposing an efficient and privacy-preserving system for on-device personalized video retrieval, without requiring any extra training data from users. Specifically, we focus on the case where queries may contain person names, and assume that our system is equipped with an on-device face detector. While our proposed techniques can be applied to queries containing other types of proper names (as long as the corresponding detector is available), we focus on this case because queries with person names are prevalent, and the supporting face detector is a common built-in component of software on most mobile devices today. Figure 2 shows an overview of how our system works (at retrieval/inference time). Our system, **mySCAN**, consists of three main components: query anonymization, generic scoring and personalized re-scoring model. To guarantee privacy and run-time efficiency, only the generic model is trained in the cloud (and later pushed to the device); the other two components are running on device.

At a high level, the system works as follows: First, given a personalized query (e.g., "Sheldon sitting on the couch"), we obtain an anonymized version of it (e.g., "Man sitting on the couch") by replacing person names with a generic noun "man" or "woman". Note that here we assume the name *Sheldon* is registered in our system and the corresponding gender information is available (either provided by the user or detected by the face detector); if such gender information is not available, we can replace the name with more generic noun like "person". Next, the anonymized query is fed into a generic video-text matching model (one that can be trained on generic data and in the cloud) to obtain an initial ranking score over user videos. Finally, the re-scoring model combines together these initial ranking scores, person name information, and the face information (i.e., face labels and spatial coordinates provided by a face detector) to generate the final score.

Our contributions in this paper are threefold.

- First, we introduce the new problem of personalized video retrieval, and we believe addressing this problem will greatly improve the user experience of mobile video search.
- Secondly, we propose an efficient technique that achieves personalization by reasoning over information from a generic model and a face detector, without requiring any re-training (or fine-tuning) on personal user data.
- Third, we show through experiments that our proposed system has superior performance compared to several alternative and baseline techniques.

2 RELATED WORK

Multi-modal Image Retrieval. Multi-modal retrieval aims to index and retrieve multimedia contents using natural language queries, which relies on learning a joint representation for data from different domains (e.g., vision, text and audio) [3]. For image retrieval, there has been a trend to explore more fine-grained semantic alignment between vision and language [14, 18, 25]; among these approaches, the SCAN model [18] achieves superior performance by utilizing full latent alignment between image regions and sentence words. Another trend is the rising of the Transformer model [29] as a building block, inspired by its success in various NLP applications. To leverage the power of Transformers, the Transformer-based models normally contain cross-modal layers [15, 22, 27, 30], which requires early fusion of the vision and language branches [31] (i.e., the computation of the visual or text embeddings depends on information from the other branch). However, compared to late fusion approaches (e.g., SCAN), early fusion prevents us from (independently) precomputing the visual features and results in heavier computation burdens at inference time, especially for mobile search.

Multi-modal Video Retrieval. Videos are natural sources of multi-modal information and have become prevalent on mobile devices. Besides representing videos as single embedding vectors, researchers have explored to unfold the rich semantic information of videos along different dimensions, including the temporal dimension [11, 13] and modality (e.g., appearance, faces and speech) dimension [10, 21]. A limitation of these approaches, especially in the context of mobile computing, is that they rely on heavy models and computing expensive visual features over moving windows. Thus, instead of leveraging on temporal information in videos, we decide to process each frame separately and focus more on fine-grained spatial domain information over image regions within each frame.

Personalization. In applications such as search and recommendation, personalization normally refers to the ability of a system to return customized results for different users, even with the same query [4, 32]. In contrast, the term personalization in this paper refers to the modelling of proper names, conditioned on each user's personal information. A relevant work using the same definition of personalization, but for different applications, is [23], where the authors developed models to learn user-conditioned representations of named entities for better user understanding.

3 SYSTEM DESIGN

3.1 Generic Video-Text Matching Model

For the generic model, we propose videoSCAN, a cross-modal retrieval model based on the Stacked Cross Attention Network (SCAN) originally designed for image retrieval [18]. While any multi-modal video retrieval approach can serve as our generic model to provide the initial ranking score, in order to introduce personalized information, we need to utilize the fine-grained spatial information per-frame (specifically, the cross attention between query tokens and image regions, as we will discuss later in Section 3.2), and thus we choose to adopt an image retrieval model as the base component. Moreover, due to limited computation and storage resources on mobile devices, the expensive features across the temporal dimension (e.g., motion features) are not used in videoSCAN, and each frame of the video is processed independently and only the per-frame ranking scores are fused to obtain a final video-query similarity.

Among different image-text models, we adopt SCAN because: 1) SCAN utilizes cross-attention to model alignments between image regions and query words, and these fine-grained attention weights learned over image regions can be easily matched against the output of the face detector, i.e., labelled face bounding boxes, allowing for a seamless integration with face detection; 2) compared to other cross-attention models (e.g., various Transformer-based models [15, 22, 27, 30]), SCAN is more efficient due to *late fusion* of visual and query embeddings (i.e., in the process of computing visual and query embeddings, information is not shared between the visual and language modalities), and thus we can pre-compute the expensive visual features; 3) compared to other late fusion approaches, SCAN achieves competitive image retrieval performance on various benchmark datasets.

Given a video I and an anonymized query Q, videoSCAN relies on SCAN to first obtain a frame-query score for each frame $I_t \in I$. These scores are then aggregated to get the initial video-query score. To ground the query with a single frame I_t, the basic idea of SCAN is: 1) decompose the query and frame into word tokens and regions, respectively; 2) for each token, compute a "customized" weighted average of the frame region features (i.e., apply the attention mechanism); and 3) compute the similarity between token and weighted region features and aggregate the scores.

The generic videoSCAN model itself consists of three components (see Figure 2): a contextualized language model, a visual feature extraction component and a scoring model. For the contextualized language model, we represent each word/token of the query with a one-hot vector and use a bi-directional GRU [2] to get the sequence of contextualized word embeddings. Note that we can also consider other contextualized language models (e.g., BERT [8]); we choose bi-GRU here due to efficiency considerations. Let the number of tokens of the query be n and denote the set of token features to be $E = \{e_1, \cdots, e_n\}$.

For visual feature extraction, following [18], we adopt the bottom-up attention model [1] to extract a set of k salient regions and their corresponding features from each frame. Due to the late fusion property of SCAN, these visual features can be pre-computed offline and stored for online inference. In the following, we focus on a single frame I_t and discuss how the scoring model works.

Denote the set of image region features (corresponding to I_t) to be $V = \{v_1, \cdots, v_k\}$. Let the cosine similarity between region i and token j be

$$s_{ij} = \frac{v_i^T e_j}{\|v_i\|\|e_j\|}. \tag{1}$$

We further normalize and apply rectifier $[x]_+ = max(0, x)$ to the similarities as

$$\bar{s}'_{ij} = \frac{[s_{ij}]_+}{\sqrt{\sum_{j=1}^{n}[s_{ij}]_+^2}}. \tag{2}$$

The weighted combination of frame region features is defined as

$$a_j^v = \sum_{i=1}^{k} \alpha'_{ij} v_i \tag{3}$$

where

$$\alpha'_{ij} = \frac{\exp(\lambda_1 \bar{s}'_{ij})}{\sum_{i=1}^{k} \exp(\lambda_1 \bar{s}'_{ij})}. \tag{4}$$

The similarity between a_j^v and e_j is

$$R(e_j, a_j^v) = \frac{e_j^T a_j^v}{\|e_j\|\|a_j^v\|}. \tag{5}$$

Then the frame-query similarity score is defined as

$$S(I_t, Q) = \frac{\sum_{j=1}^{n} R(e_j, a_j^v)}{n}. \tag{6}$$

Finally, the frame-query scores are aggregated to obtain the video-query score $S(I, Q)$, i.e.,

$$S(I, Q) = AGG\left(\{S(I_t, Q) | I_t \in I\}\right) \tag{7}$$

where $AGG(\cdot)$ is the aggregation function. Note that $AGG(\cdot)$ can simply take the form of max or average pooling, or we can learn a weighted average of the frame-query scores.

Triplet loss is a commonly used ranking loss for deep metric learning and remains competitive compared to more complicated losses proposed in recent years [24]. Thus, to learn the parameters of videoSCAN, we adopt triplet loss, and following [9, 18], we utilize the hardest negatives in a mini-batch. Specifically, for a positive pair of video and query (I, Q), we define the hardest negatives (within a mini-batch) to be

$$Q_{\text{neg}} = \arg\max_{Q' \neq Q} S(Q', I), \tag{8}$$

$$I_{\text{neg}} = \arg\max_{I' \neq I} S(Q, I'), \tag{9}$$

and we use the following loss:

$$\mathcal{L}(I, Q) = [\Delta - S(I, Q) + S(I, Q_{\text{neg}})]_+ + [\Delta - S(I, Q) + S(I_{\text{neg}}, Q)]_+ \tag{10}$$

where Δ is the margin. Note that the videoSCAN model of our system is pre-trained using the above loss in the cloud and does not need any extra training data from users.

3.2 Re-scoring Model

The re-scoring model is designed to modify the initial frame-query similarity scores by reasoning over the information provided by videoSCAN and the face detector. Instead of relying on videoSCAN, a straightforward way to use the face information is to simply increase the frame-query similarity score whenever the face of a mentioned person is detected in the frame. However, this solution ignores the context information provided by the query; for example, considering the image corresponding to the generic query in Figure 1, although the person "Sheldon" is detected, he is not conducting the action described in the query. Moreover, for common individual users, the same group of people (e.g., family members) appear frequently in all videos, which introduces significant noise to this straightforward solution.

Denote the set of names of all known identities to be $\mathcal{P}$. We first focus on a single person name token $P \in \mathcal{P}$. Assume P is the z-th token in the sequence, i.e., e_z is the (contextualized) embedding of the word "man" or "woman" (depending on the gender of the person with name P). The intuition behind our re-scoring model is that if the region that e_z most attends to (captured by α'_{iz}) shares the same label (provided by the face detector) with P (the original person name), then we increase the image-query similarity score. In other words, we rely on the attention weight α'_{iz} to determine whether region i is corresponding to the embedding e_z of the person token P. Moreover, based on the observation that the region e_z most attends to may not necessarily contain the face (e.g., could only focus on the body part), a more robust way to utilize the attention information is to look at the top m regions with highest attention values (i.e., highest α'_{iz} values). Note that m here is a hyperparameter. Specifically, let $\sigma_m(P)$ be the set of identifiers of the regions that e_z most attends to, i.e.,

$$\sigma_m(P) = \underset{s \subset \{1,\cdots,n\}, |s|=m}{\arg\max} \sum_{i \in s} \alpha'_{iz}. \tag{11}$$

Let the image region corresponding to identifier σ be R_σ, and let the face region corresponding to the name P be $F(P)$. The intersection-over-face (IoF) between $F(P)$ and R_σ is defined as

$$IoF(F(P), R_\sigma) = \frac{|F(P) \cap R_\sigma|}{|F(P)|}, \tag{12}$$

where $|\cdot|$ is the size of the region bounding box. Then for each $P \in \mathcal{P}$, we define the identity confidence score $h(P)$ as

$$h(P) = \max_{\sigma \in \sigma_m(P)} IoF(F(P), R_\sigma), \tag{13}$$

and we modify the image-query similarity score as

$$\hat{S}(I_t, Q) = S(I_t, Q) + \sum_{P \in \mathcal{P}} h(P)\delta, \tag{14}$$

where $\delta > 0$ is some constant hyperparameter (e.g., $\delta = 0.1$). Note that although the face information on its own is not sufficient for personalization, we can still apply a *face filter* to the frame-query matching score, i.e., if any face of the named entity in the query is not detected in current frame, then we will not consider its score.

Finally, using the aggregation function as in Equation 7, the modified video-query similarity is

$$\hat{S}(I, Q) = AGG\left(\{\hat{S}(I_t, Q) | I_t \in I\}\right). \tag{15}$$

4 EXPERIMENTS

4.1 Dataset and Implementation Details

Dataset. We evaluate our proposed algorithm on the TV show Retrieval (TVR) dataset [19], which is a large scale video retrieval dataset with queries containing proper names. Specifically, we use the *Big Bang Theory (BBT)* portion of the TVR validation dataset [19], which we refer to as *BBT*. We split BBT into 1600 training examples (*BBT-train*), and 100 testing examples (*BBT-test*). We choose the seven main characters (i.e., Sheldon, Leonard, Penny, Howard, Raj, Amy and Bernadette) in BBT as known identities and all other people as unknown.

Face Detection. We pre-process video frames in the BBT dataset to extract bounding boxes, labels, and corresponding features for their face regions. We adopt the multitask cascaded convolutional networks (MTCNN) [33] to conduct face detection and alignment, and FaceNet [26] to extract embedding features. We then collect a dataset for the known (and unknown) identities, and train a SVM model for face detection based on FaceNet embeddings.

Implementation Details. We choose the aggregation function in Equation 7 to be max-pooling, and set $m = 3$ (i.e., considering the top 3 most attended regions). For visual feature extraction, following [18], we use the ResNet-101 based Faster R-CNN model pretrained by Anderson et al. [1] on Visual Genome [17]. The number of image regions extracted from each frame is $k = 36$. The dimension of the word token embeddings and image region embeddings is set to be 128. The videoSCAN model is pretrained on MS-COCO [20] and ActivityNet Captions [16].

Baselines and Alternatives. Our baseline is a *keyword-based system*, which is our faithful reproduction of existing keyword-based video search system commonly used in mobile devices [6, 12]. For a single frame, we have a synonym-expanded set of tags generated by our trained face detector, an image classifier (EfficientNet [28] pretrained on ImageNet [7]), and a scene classifier (official model released by the Places dataset [34]). Then for each video, we do symbolic matching between query tokens and frame-wise tags, and aggregate the matching score temporally through max-pooling.

To benchmark the proposed personalized re-scoring model, we also consider two alternative methods which only consist of the generic videoSCAN model: *anonymized-videoSCAN*, which expects anonymized generic queries as input (i.e., like in mySCAN, the person names in queries are replaced with "man" or "woman") and is thus not personalized, and *personalized-videoSCAN* that explicitly adds person names to the model vocabulary (i.e., the person names are not anonymized). To be clear, *personalized-videoSCAN* is merely an illustrative benchmark as a gold standard. Unlike our approach, which is feasible in a real-world deployment context on-device, this gold standard requires practically infeasible labeled data (video-caption correspondences where the text caption includes personalized entities) and is merely shown for comparision purposes. We fine-tune both these models on two datasets: the BBT-train dataset, and a randomly chosen subset of BBT-train (consisting

Method	Semantic	Personalized	Fine-tuned on	R@1	R@5	R@10	Average
keyword-based	✗	✓	✗	15	35	53	34.3
anonymized-videoSCAN	✓	✗	✗	15	34	55	34.7
anonymized-videoSCAN	✓	✗	BBT-small	19	45	60	41.3
anonymized-videoSCAN	✓	✗	BBT-train	25	50	72	49
personalized-videoSCAN	✓	✓	BBT-small	16	46	59	40.3
personalized-videoSCAN	✓	✓	BBT-train	<u>30</u>	**66**	**79**	**58.3**
mySCAN (ours)	✓	✓	✗	**33**	<u>62</u>	<u>76</u>	<u>57</u>

Table 1: Results on BBT-test. Columns 2–4 specify if a method is based on semantic embeddings or keywords, uses personalized information, and is fine-tuned on additional data. Best performance is in boldface; second best performance is <u>underlined</u>.

of around 350 training examples) which we refer to as *BBT-small*. Note that for anonymized-videoSCAN, the performance gain of fine-tuning mainly comes from closing the domain gap between training data and BBT-test data, and for personalized-videoSCAN, the gain is mainly from achieving some level of personalization by using labelled fine-tuning data.

4.2 Quantitative Results

Table 1 shows the experimental results on BBT-test. We use standard evaluation metrics used in information retrieval, namely Recall@K, which captures the fraction of times a correct video was found among the top K choices; we also present the average of all Recall@K values. As illustrated in Table 1, the different methods differ across three dimensions: 1) whether the method is based on semantic embeddings or keyword matching; 2) whether the method is using personalized information (i.e., person names); and 3) whether the method is fine-tuned on additional labelled data.

According to the third dimension, we first consider the group of approaches that do not require fine-tuning on additional in-domain data, namely, the keyword-based system, the anonymized-videoSCAN model without fine-tuning, and mySCAN. An important characteristic of these approaches is that they do not need to upload local on-device data to the cloud, since labelled training data from users is not required. Therefore, these approaches are not only cost-effective in terms of bandwidth and computation, but also preserve the privacy of user data. Table 1 shows that anonymized-videoSCAN without fine-tuning (which captures the context information but does not utilize any personalized information) performs similarly to the keyword-based system (which represents the performance of existing mobile video search systems), and that mySCAN significantly outperforms both. In particular, for the metric of average recall, mySCAN achieves 66% improvement comparing to the keyword-based system, which shows the effectiveness of mySCAN in terms of utilizing both context and personalized information.

Next we compare mySCAN to the group of approaches that have been fine-tuned, i.e., the fine-tuned anonymized-videoSCAN and personalized-videoSCAN methods. Table 1 shows that when fine-tuned on a small dataset like BBT-small, anonymized-videoSCAN and personalized-videoSCAN achieve similar performance; in this case, there is not enough data for the personalized model to learn useful embeddings for the newly added person names. In contrast, when trained on a larger dataset like BBT-train, the performance of both models improves and personalized-videoSCAN substantially outperforms its anonymized counterpart. Moreover, we can see that personalized-videoSCAN fine-tuned on BBT-train achieves the best performance (in terms of average recall), and that mySCAN achieves comparable performance. Informally speaking, the sizes of BBT-train and BBT-small can be interpreted as "upper-bound" and (loose) "lower-bound" on the amount of fine-tuning data needed for personalized-videoSCAN to achieve the performance of mySCAN. However, it is not even practical to collect on-device training data with the lower-bound size (i.e., several hundred training examples as in BBT-small). To summarize, compared to the fine-tuned approaches, mySCAN achieves substantially better or comparable performance *without requiring labelled video-text correspondence data* from users, which shows the effectiveness of the proposed re-scoring model in terms of combining the detected face information and semantic information in attention maps.

4.3 Visualizations

In this subsection, we present visualizations to qualitatively demonstrate the retrieval performance of mySCAN. Figure 3a shows a positive example (i.e., the ground-truth video is among the top five selected candidates) for a variant of our algorithm where the face filter is not applied. We can see that the ground-truth video is correctly ranked as Top-0.[1] Moreover, all top results capture the context information that two person are conducting the action *eating* in the scene of *cafeteria* or *restaurant*. This indicates the capability of the model to capture semantic meaning of the query.

Figure 3b shows the retrieval results with face filter applied for the same query as in Figure 3a. Due to the face filter, we can see that only frames containing both Leonard and Howard are among the top results. Although the ground-truth video is ranked as Top-1 (instead of Top-0 as in Figure 3a), we can see that the query is also a reasonable description for the Top-0 frame.

Finally, we present some qualitative results to illustrate the main failure modes of mySCAN:

- Face detection error, i.e., the face detector is not able to accurately recognize the face of the person in the query. For example, see Top-2 frame in Figure 4 (where Sheldon is

[1] Note that the rank of the videos starts from 0.

Figure 4: Failure mode: face detection error. The top row contains the query, the rank of ground-truth (GT) video in the retrieval list, and the most representative frame of the query manually selected from GT video. The bottom two rows contain highest scored frames from the top five ranked videos in the retrieval list.

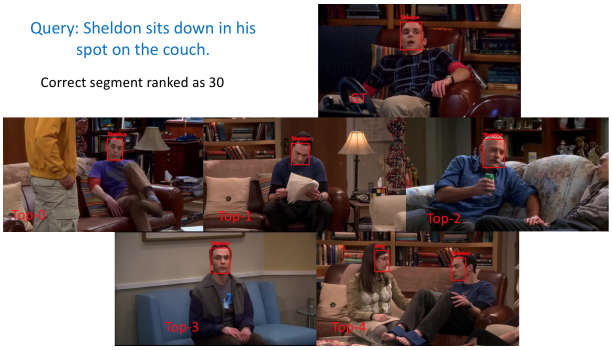


Figure 5: Failure mode: uninformative query and face detector error. The top row contains the query, the rank of ground-truth (GT) video in the retrieval list, and the most representative frame of the query manually selected from GT video. The bottom two rows contain highest scored frames from the top five ranked videos in the retrieval list.

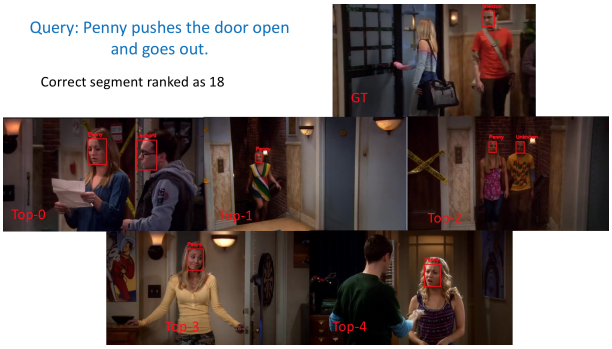
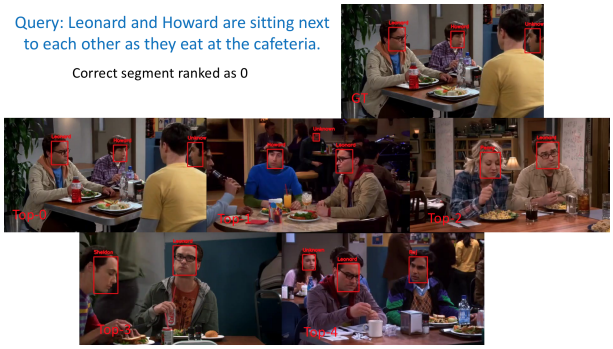
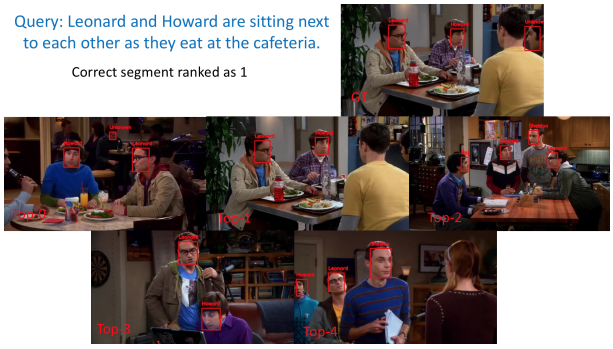


Figure 6: Failure mode: face of identity not visible. The top row contains the query, the rank of ground-truth (GT) video in the retrieval list, and the most representative frame of the query manually selected from GT video. The bottom two rows contain highest scored frames from the top five ranked videos in the retrieval list.



(a) Without face filter.



(b) With face filter.

Figure 3: Illustration of the retrieval results of mySCAN. In each subfigure, the top row contains the query, the rank of ground-truth (GT) video in the retrieval list, and the most representative frame of the query manually selected from GT video. The bottom two rows contain highest scored frames from the top five ranked videos in the retrieval list.

misclassified as Leonard) and Top-2 frame in Figure 5 (where an unknown person is misclassified as Sheldon).

- The query is not informative. For example, in Figure 5, the query is "Sheldon sits down in his spot on the counch" and the described scenario appears very often in the dataset; the top results in Figure 5 (except the misclassified one) are all valid choices.
- Face of known identity in the query is not visible. For example, in the GT video in Figure 6, when Penny conducts the action *push the door open*, her face is not visible.

5 CONCLUSIONS

We have introduced the personalized video retrieval problem, and designed an efficient and privacy-preserving system to address this problem by integrating face detection and cross-modal attention. We show through experiments that the proposed system outperforms existing keyword-based mobile search system significantly, and matches the performance of a gold standard generic model

(i.e., videoSCAN) fine-tuned on plenty of labelled data. We have therefore demonstrated that the developed techniques can improve mobile video search in a practically feasible way.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, 6077–6086.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*. San Diego, CA.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[4] Keping Bi, Qingyao Ai, and W Bruce Croft. 2020. A Transformer-based embedding model for personalized product search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1521–1524.

[5] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 10638–10647.

[6] Jeffrey Dalton, James Allan, and Pranav Mirajkar. 2013. Zero-shot video retrieval using content and concepts. In *Proceedings of the Conference on Information and Knowledge Management*. ACM, 1857–1860.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[9] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*.

[10] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision*. Springer, 214–229.

[11] J. Gao, C. Sun, Z. Yang, and R. Nevatia. 2017. TALL: Temporal activity localization via language query. In *Proceedings of the The International Conference on Computer Vision*. IEEE, 5267–5275.

[12] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. 2014. Composite concept discovery for zero-shot video event detection. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 17–24.

[13] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the The International Conference on Computer Vision*. 1247–1257.

[14] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal LSTM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2310–2318.

[15] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv:2004.00849* (2020).

[16] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the The International Conference on Computer Vision*. IEEE, 706–715.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*. 201–216.

[19] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. TVR: A Large-scale dataset for video-subtitle moment retrieval. In *Proceedings of the European Conference on Computer Vision*. Springer, 447–463.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[21] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. In *Proceedings of the British Machine Vision Conference*.

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Conference on Neural Information Processing Systems*.

[23] Levi Melnick, Hussein Elmessilhy, Vassilis Polychronopoulos, Gilsinia Lopez, Yuancheng Tu, Omar Zia Khan, Ye-Yi Wang, and Chris Quirk. 2020. Privacy-aware personalized entity representations for improved user understanding. In *First Workshop on Privacy in Natural Language Processing at the Conference on Empirical Methods in Natural Language Processing*.

[24] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision*. Springer, 681–699.

[25] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 299–307.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 815–823.

[27] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[28] Mingxing Tan and Quoc V Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*. 6105–6114.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems*. 5998–6008.

[30] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 10941–10950.

[31] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33.

[32] Han Zhang, Songlin Wang, Kang Zhang, Zhiling Tang, Yunjiang Jiang, Yun Xiao, Weipeng Yan, and Wen-Yun Yang. 2020. Towards personalized and semantic retrieval: An end-to-end solution for E-commerce search via embedding learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2407–2416.

[33] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[34] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2017), 1452–1464.