Personalised CLIP or: how to find your vacation videos

Bruno Korbar
https://kor.bar
Andrew Zisserman
https://www.robots.ox.ac.uk/~az

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

Abstract

In this paper, our goal is a person-centric model capable of retrieving the image or video corresponding to a personalized compound query from a large set of images or videos. Specifically, given a query consisting of an image of a person's *face* and a text *scene description* or *action description*, we retrieve images or video-clips corresponding to this compound query. We make three contributions: (1) we propose CLIP-PAD, a model that is able to retrieve images/video given a personalized compound-query. We achieve this by building on a pre-trained CLIP vision-text model that has compound, but general, query capabilities, and provide a mechanism to personalize it to the target person specified by their face; (2) we share a new *Celebrities in Action* (CiA) dataset of movies with automatically generated annotations for identities, locations, and actions that can be used for evaluation of the compound-retrieval task; (3) we evaluate our model's performance on two datasets: Celebrities in Places for compound queries of a celebrity and a scene description; and our new CiA for compound queries of a celebrity and an action description. We demonstrate the flexibility of the model with free-form queries and compare to previous methods.

1 Introduction

Suppose that you want to find the video of your vacation where *you* ran in a red and white striped shirt in *front of the Parthenon* in Athens amongst all videos on your phone. Or imagine that you cannot remember the name of the movie where *Ben Stiller* runs *on a boat*. Or maybe you don't even know Ben Stiller's name, but have seen him in another movie and can find his picture. On its own, finding out if you (or Ben Stiller) are in the video or if the video contains a boat or Parthenon (single-query) is a well-researched problem. Searching with multiple (and potentially multi-modal) *compounded* queries in large databases, however, is still a challenging and under-researched proposition. In this paper, we aim to tackle this problem and deliver a model that is capable of precise retrieval for compound queries looking for specific people in specific places or performing a specific action.

Vision-language models such as CLIP [35] and ALIGN [14] have transformed the performance of many visual-language tasks. These models use a dual encoder and are trained on large-scale datasets [29, 38] using a contrastive loss. In particular, they can be used to retrieve an image or video given a free-form text description. However, this freedom of using a sentence to describe the sought visual content is also a limitation. How can a query

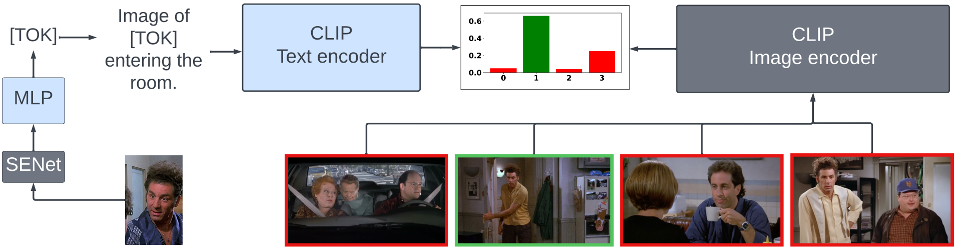

Figure 1: An outline of the CLIP-PAD architecture, when it is queried by a target face image (in this case of Michael Richards) and a sentence. Gray modules (dark background) are pre-trained and frozen, whilst blue ones (light background) are learned. The model correctly retrieves a video clip that matches the text description (specified by the sentence) personalized to Michael Richards (specified by his face).

also include specific *visual* content such as a particular person (specified by their face), or a particular object instance (specified by an image of it)? If the model doesn't have the notion of identity, how can it find a conceptual difference between different instances ("Brad Pitt" vs "George Clooney")? In this work, we propose a simple addition to the foundation models that would allow them to do just that.

The key idea is to provide a mechanism to adapt a face image (that specifies the identity) to 'act as' a text token that describes the identity within the query, as illustrated in Fig. 1. We show that making the model identity-aware works remarkably well compared to a zero-shot model. It significantly improves the retrieval performance on the Celebrities-in-Places (CiP) [45] compound-query retrieval dataset. Furthermore, we show that we can use our personalised model even in video scenarios. To this end, we annotate a human action movie dataset with person-specific labels, which we call *Celebrities in Action* (CiA), and evaluate performance compared to existing retrieval methods.

This person adaptive model, which we refer to as CLIP-PAD(Person ADaptive), has applications in real-world video-retrieval applications, such as searching a video archive for historical celebrities performing actions or in particular places. For example, it would enable a broadcaster such as the BBC or a stock company such as Shutterstock to carry out a personalized compound query, using free-form text, to search the archive on their visual content, without requiring any text annotation of the archive. The scheme is clearly applicable also for seaching personal images and videos.

In section 3 we outline the CLIP-PAD model, in particular the simple adaptor mechanism for the CLIP model that allow us to retrieve images (and video) given a personalized compound-query. In section 4 we describe the CiA dataset, a new benchmark for video compound retrieval, and how it is generated. Finally, in section 5 we demonstrate the high performance of the model against baselines on both CiP and CiA under various different retieval scenarios. In the supplementary material, we additionally show a real-world retrieval example from various episodes of the Seinfield TV Show.

2 Related work

Foundation video-text models. Ever since AlexNet [17], pre-trained models have been used to boost performance or bootstrap models on downstream tasks [36]. Using video-to-text correspondence to develop strong pre-trained models is not a novel phenomenon [10, 26, 27], however, recently the scale of training data became so large that a new generation of pre-trained (sometimes also called foundation) models was developed with the capacity of implicitly solving even the tasks they weren't trained for explicitly (CLIP [35], ALIGN [14], BLIP [20], FILIP [43], to name just a few). Following these advances, we use CLIP as a backbone embedding for our model and use a trainable-prompting paradigm to achieve our tasks.

Text-to-video retrieval. Video-text retrieval has long been a fruitful research direction, with a plethora of datasets available [12, 16, 18, 37, 46]. Due to the processing costs, most models were traditionally developed on top of feature “experts" [3, 6, 9, 21, 25, 26]. First, the features are extracted from models that were pre-trained for a specific task (and often combined), and then the retrieval model was trained. With the rise of large-scale models that use video-to-text correspondence [7, 8, 10, 19, 26, 27], the focus switched to direct similarity metrics between retrieval text queries and videos. Foundation models dominate the video-text retrieval leader boards, even in the zero-shot setting [1, 2, 22, 32, 35, 41]. This task is closely related to the tasks of text-to-video-localization and (corpus) moment retrieval [44], for which various architectures have been proposed [40, 42].

Compound (person-specific) query retrieval. Text-to-video retrieval is a notably different task to compound (person-specific) query retrieval as it requires much less specificity. Traditional text-to-video retrieval datasets are often depersonalized (e.g. in LSMDC, all names are substituted by “someone" [37], so a query "Kramer enters the apartment", and "George enters the apartment" would be indistinguishable). In the case of compound retrieval, the focus is on specificity. Despite it being a common everyday task, very few datasets have been released and we believe it is an under-explored task. We address this by presenting a new dataset based on the existing Hollywood2 benchmark [24] and High-Five human interaction dataset [33], annotated with identities to enable compound-queries and responses.

CLIP and its shortcomings. CLIP is a dual-encoder foundation model introduced by Radford et al. [35]. The premise is rather simple, given an image processed by a visual encoder and corresponding text processed by a text encoder train the model contrastively (using symmetric contrastive loss), and train it on an unprecedented scale (over 400M labeled images). Since its publication and release, CLIP models have been used in a myriad different ways for a plethora of tasks, often unrelated to their original training task [23, 28, 35, 39, 41]. To harness emerging properties of these models, CLIP-based models are used in a zero-shot manner [34], simply by modifying the prompt to match the desired output [28, 35]. If we wanted to know if the image contains a bird or a dog, we'd simply feed an image and two text prompts (“Image of a dog”, “Image of a bird”), and then find which text prompt has higher similarity to the image. Other common ways of “adapting” clip are via learning models on top of the visual embedding [23, 35, 39], or via learnable prompting [15]. Despite the size and the variety of the training dataset CLIP is trained on, there are some tasks that it is inherently less suited to – e.g. tasks containing actions. This shortcoming is not surprising as the model is trained on images alone, and as recent works have shown, adapting it to the video domain does take additional ingenuity [1, 23, 34]. The most prevalent approaches are training specific prompts to feed into the model [15], augmenting the model architecture to accept different embedding types [1], or training aggregation models on top of it [23]. Our approach roughly falls amongst “learnable prompting" approaches.

3 Personalising CLIP: CLIP-PAD

To personalise CLIP, we adopt a prompt-learning method in order to adapt the model for our task. We wish to use a free-form text query but adapted to the specific target person. To achieve this we fine-tune the CLIP text encoder to “recognise” the target person given a text query and a prompt starting from a loose crop of the person’s face.

Architecture overview. The architecture is illustrated in figure 1. The target face image is processed by a pre-trained face encoder, and passed through a Multi-Layer Perception (MLP) to adapt the face encoding to the space of the word encodings. The entire compound query is then encoded using the CLIP text encoder. So, for example, to find an image of Tom Cruise running, the face image would be of Tom Cruise, and the text query would be “An image of *TOK* running”, where *TOK* is the output of the MLP adaptor. Note, the only trainable components of the architecture are the MLP adaptor, and in case of text-queries, the CLIP Text encoder. All the other modules: ConvNet face encoder, and CLIP image encoder are pre-trained and frozen. The details of each module are given in the implementation details.

Dataset retrieval. In order to perform a text-to-image retrieval, embeddings are generated for every image in the dataset using the CLIP image encoder. The match for a given (compound) query is then obtained by finding the image representation with the highest cosine similarity to the CLIP embedding of the query text. Query text can be formed of text only (“An image of Tom Cruise running”), text with a visual query (“An image of *TOK* running”), or a combination (“An image of Tom Cruise *TOK* running”).

Training. We start from the pre-trained weights for CLIP [35], and keep as many parameters fixed as possible, training only the MLP adapter and the CLIP text encoder. The purpose of training the model is to make it identity-aware. To achieve this, we train the MLP adaptor and the CLIP text encoder to be able to discriminate amongst different identities. To do so, we fine-tune our model on a *person-recognition dataset*. During training, each batch contains 60% of queries containing the image token, 20% being names of celebrities as strings, and the rest a combination of both. We found empirically that this ratio yields the optimal performance in the general case where we might or might not have a visual query. The model is trained using the symmetric contrastive loss as proposed in [35] until it can perform the person-classification task sufficiently well, with the criterion being different for each evaluation dataset as outlined in the implementation details.

3.1 Implementation details

Model details. Our model starts from an unmodified pre-trained CLIP dual-encoder architecture: the visual encoder is a ViT-B/32 transformer, and the text encoder is a 3M-parameter 12-layer 512-wide model with 8 attention heads as outlined in [35]. The face encoder is a SE-ResNet-50-128D model pre-trained on the person dataset (from [5]). It is a ResNet50 [11] with Squeeze and Excitement (SE) layers [13] that outputs 128-dimesional vector for each person. The output of the face encoder is processed via a two-layer MLP, taking the dimension from 128 to 256 and 256 to 512 respectively with ReLu non-linearity. This 512 vector is then used as an input embedding to a CLIP text encoder of the same dimension. This embedding is then processed as if it was a standard input to a CLIP text encoder – i.e. position embeddings are added following the protocol in Radford et al. [35] before it is ingested by the encoder. Note that on the video data, we employ mean pooling similarity calculator to aggregate visual embeddings from different frames of the video on top of the CLIP vision encoder as proposed in Luo et al. [23].

Training details. The training procedure and dataset depend on the target dataset. In general, we would want as many known celebrities to be a part of our model and thus ideally we would want to train it on a large VGGFace2 [6] dataset. This dataset however contains some of the test 'unseen' faces from the Celebrities in Places (CiP) dataset, and having an embedding trained on these images would yield an unfair comparison to the baseline [45]. Therefore, for evaluation on CiP, the model is trained on the loose crops of the VGGFace dataset [31] – it being the same pre-training dataset used by Zhong et al. [45]. We choose to use loose crops (as opposed to traditionally used tight crops) to minimise the domain gap to the target images that will often show entire body as we are not concerned by a potential impact on the overall person-classification task.

The model is trained for 10 epochs, evaluating it on a held-out validation set from *VGGFace2* at the end of every epoch. We early stop the model if it achieves 85% or 95% accuracy on validation when trained on VGGFace and VGGFace2 datasets respectively.

For CiA, we train the model on loose crops of VGGFace2 [6] following the same procedure as above. We additionally fine-tune it for 5 epochs on the training movies of CiA, to account for the fact that CLIP might not have been trained with action classes in mind and that actors might not be present in the VGGFace2 dataset. We keep the ratio of the person queries constant during the fine-tuning process.

The hyper-parameters and training ratios of queries were found via a linear search based on model performance on the CiP held-out validation set. The optimal parameters were selected based on maximum average performance when using text queries, text+image queries, or image only queries. We then use the same hyper-parameters for all other datasets.

4 Celebrities in Action Dataset

In this section we describe how we annotate the Hollywood 2.0 dataset for person-centric retrieval. The Hollywood 2.0 dataset consists of 12 classes of human actions and 10 classes of scenes distributed over 3669 video clips from 69 movies (33 training, 36 testing) [24]. Examples of actions include 'eating' and 'running', whilst the scenes include the 'office', 'shop', and 'car'. Test sets for both actions and scenes have been manually cleaned, whilst the training set has been annotated automatically and has inherent noise (upon manual inspection of randomly selected 50 clips, 2 were unclear). To better evaluate classes with human interactions, we also include 172 clips from the High-Five human interaction dataset [33] which has collaborative actions such as 'hugging', 'kissing' and 'giving a high five'. The limitation for person-centric queries is that although the clips are labelled with the actions performed, the person performing the action is not labelled.

split	clips	% in VGGFace2	person (per clip)	actions (per actor)	scenes (per actor)
train	1308	23.6	2.5	4.1	3.1
val	328	23.6	3.1	4.6	2.9
test	1052	38.2	2.0	3.9	3.1

Table 1: CiA stats. We report the total number of clips in every split, percentage of annotated celebrities in the split clips that are present in the fine-tuning VGGFace2 dataset [5]. Furthermore, we show how many people are on average present in each video clip, and we try to find if a given actor on average performs different actions or appears in multiple scenes in the data.

In order to form the *Celebrities in Action* (CiA) dataset [1], we automatically annotate the clips with the person performing the action. To achieve this we use the automatic video annotator by Brown *et al*. [4]. In brief, this method uses the IMDB cast list from the movie to obtain face images for each actor in order to classify their occurrences. On the video side, faces are detected and tracked in each clip, and then an identity is associated with the face track if it is classified as one of the known actors from that film. In the case of multiple actors getting annotated (in 47% of video clips), we select the most confident one as our final annotation. This can potentially cause incorrect or ambiguous examples as seen in the last two columns of figure 2. We do not address this issue as ambiguous or incorrect labels are rare. We manually verify the label correctness on a randomly selected 100 clips from the test set and find the actor annotations to be correct for 97 of them (i.e. annotated actors are visible in the video clip). We separate the training set into the training, and held-out validation set (for model development) – not according to movies but rather, according to the clips in the training data. We only use the training data to fine-tune CLIP for better action performance.

To form queries from the data, we form template sentences in three ways: 1) "`{celebrity}` is doing `{action}`", 2) "`{celebrity}` in `{place}`", and 3) "`{celebrity}` in `{place}` doing `{action}`". For the "`{celebrity}`" token, we consider both text, image, and a combination.

Statistics of the dataset can be found in Table 1, and some example from the annotated dataset are given in Figure 2.

5 Experiments

In this section we evaluate the CLIP-PAD model for two person-centric retrieval tasks using compound queries. The first is 'a person in a place', and for this we evaluate on the existing *Celebrities in Places* benchmark dataset where images are annotated with both the celebrity and the place. The second task is 'a person doing something', and for this we evaluate on the 'Celebrities in Action' dataset described in section 4.

Since CLIP has been trained on millions of images, it is likely that it will have seen examples of some of the celebrites labelled with their identity. However, the long-tailed nature of the "celebrity" classes makes it unlikely that the model would have seen *all* or even most of them. Similarly, is likely that CLIP has seen all the classes of places in the Celebrities in Places dataset. For this reason we include zero-shot baselines where we simply evaluate the retrieval performance given the celebrity's name. For example, for the 'Celebrities in Places' dataset we evaluate both for the celebrity alone with queries such as "An image of

[1] https://www.robots.ox.ac.uk/ vgg/research/celebrities-in-action/

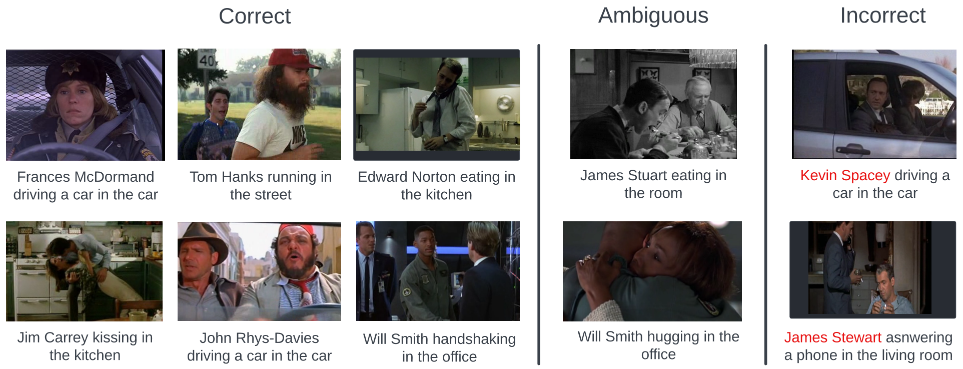
Correct
Ambiguous
Incorrect
Frances McDormand driving a car in the car
Tom Hanks running in the street
Edward Norton eating in the kitchen
James Stuart eating in the room
Kevin Spacey driving a car in the car
Jim Carrey kissing in the kitchen
John Rhys-Davies driving a car in the car
Will Smith handshaking in the office
Will Smith hugging in the office
James Stewart asnwering a phone in the living room

Figure 2: Examples of correctly labeled, ambiguous, and incorrectly labeled (all regarding the person annotation) from CiA.

Tom Hanks", as well as for celebrities in a place with queries such as "An image of Tom Hanks in the supermarket".

5.1 Celebrities in Places

Celebrities in places (CiP) is an established benchmark for a person-centric compound retrieval. Its test set contains 15.1k images with 2.3k celebrities in 16 places. This dataset contains a wide variety of celebrities from the VGGFace dataset ('seen' – 0.6k), and some that were not included in VGGFace ('unseen' – 1.7k). Places include visually different scenarios such as the 'beach', the 'stage', and the 'golf course' to name a few. Note that the dataset is not class-balanced, so the most common place ('stage') has over 2k examples, while the least common place ('desert') has only 102 examples. Similarly, the most common celebrity is present in over 60 images, whilst the least common ones are present in only 1. The model is evaluated by forming a query such as '*Celebrity* in *place*' (where the provided queries cover only a subset of the 2.3k celebrities: 792 unseen and 223 seen), and 'place' is 1 of 16 possible places, and retrieving the correct example amongst 15.1k annotated images, and an additional annotated distractor-images provided sampled from other datasets (36k images in total for the annotated test set and distractors).

In table 2, we first compare the performance of our model to other baselines on *non-compound* queries. For example, we would query our model with `'Image of John Wayne'` for a face retrieval, or `'Image of the golf course'` for the place retrieval. We compare our model to the compound query retrieval baseline by Zhong et al. [45] which employs two separate ConvNets to extract face and place embedding respectively and trains a joint embedding used for retrieval with a multiclass hinge loss. We also compare our model to a zero-shot CLIP [35] model evaluated in the same way as our model with the *txt* query. We can observe that the original CLIP model can very accurately retrieve places without fine-tuning, however, it lacks the capability of retrieving people with high precision. The difference in performance between the seen and unseen classes is negligible. This is likely explained by the fact that CLIP's visual descriptor is highly discriminative 'as-is' and is likely to have seen some of the celebrities belonging to the 'unseen' set. Lastly, we can

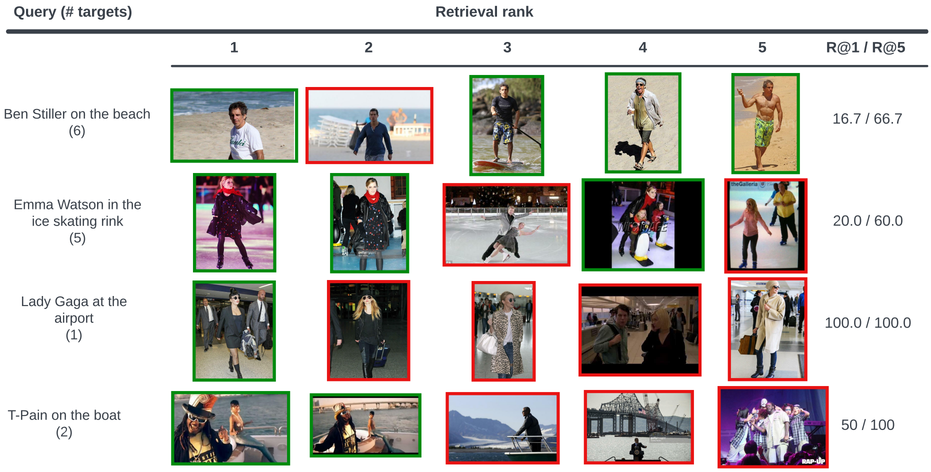

Figure 3: Qualitative retrieval examples for various queries on CiP [45] dataset sorted by retrieval rank. A green boarder indicate the correctly retrieved class, and a red one indicates an incorrect one.

see that personalising our CLIP text encoder doesn't impact the performance of the retrieval for 'places' (−1.6mAP) while it dramatically improves the performance on the face-retrieval (+22.0mAP).

In table 3 we compare the *compound-query retrieval* performance. As we did not observe large difference between the 'seen' and 'unseen' faces, we average the results weighted by the number of queries. The personalised model performs significantly better than both baseline results [45] and zero-shot CLIP [35]. Qualitative examples can be found in figure 2.

Method	mAP (face:unseen)	mAP (face:seen)	mAP (places)
[45]	74.1	73.7	51.1
CLIP [35]	60.1	59.1	83.1
text	82.1	82.0	81.5
img	84.0	83.7	–
text+img	86.7	86.6	–

Table 2: Baselines results for retrieving non-compound query-classes of the CiP dataset as compared to our model (bottom section) with different person prompts.

Method	mAP	R@5
[45]	63.4	82.2
CLIP [35]	60.6	79.5
text	79.5	94.5
img	77.7	93.0
text+img	79.5	94.5

Table 3: Compound-retrieval results on the CiP dataset. Note that we average results on seen/unseen faces.

5.2 Celebrities in Action

We first evaluate three zero-shot baselines in Table 4; we measure if the model can solve actor classification, action classification and scene classification on the training set in a zero-shot setting. If the task is to classify an actor, we follow the original CLIP zero-shot protocol [35] and for a given video-clip we compute similarities to text queries using the actor names as classes (e.g. "Image of Tom Cruise") and pick the most similar one as a predicted class. Classification for actions and scenes is done in an analogous way. Note again, that a zero-shot CLIP model can classify both action and scene with a very high degree of accuracy, however where it fails is the classification of the actor. Our model alleviates this issue to a large extent. Furthermore, we see a notable improvement in classification accuracy (6.2%) when querying our model with images. This is likely due to the fact that only a small proportion of faces can be found in VGGFace2 dataset (see Table 1), thus the additional information does help significantly.

Method	Actor classification	Action classification	Scene classification
Random	0.7	8.3	10
CLIP [35]	9.1	73.1	91.6
CLIP-PAD-*text*	26.5	75.7	91.6
CLIP-PAD-*img*	32.7	–	–
CLIP-PAD-*text+image*	33.8	75.7	91.6

Table 4: Baseline zero-shot classification results on CiA can be found in the top section. Ours are in the bottom section, with modalities used as a query appended. Results are given in % accuracy.

In order to compare our results with a more-traditional retrieval models, we compare them to two modern baselines: a mixture of experts model [21], and CLIP in a zero-shot setting using the straight-CLIP protocol [34, 35]. We report recall at ranks 1 and 5 (R@1, R@5). The results can be seen in the table 5. Two things stand out immediately: First is the discrepancy between the 'action' and 'place' retrieval of zero-shot CLIP: even on 'simple' action classes, it does demonstrably worse when compared to the compound retrieval on 'places', which it recognises well. Second is the performance increase coming from our personalised model when compared to zero-shot CLIP. Note that the baseline models cannot query the model by using multi-modal queries. To overcome this limitation, we present a two-stage baseline in the supplementary material.

Method	query		action		place		action + place	
	text	rgb	R@1	R@5	R@1	R@5	R@1	R@5
CE [21]	✓	*	35.3	65.1	36.7	65.3	32.1	62.5
CLIP [35]	✓		42.4	73.5	58.2	78.1	39.9	72.0
CLIP-PAD	✓		62.1	81.3	67.8	87.0	64.2	81.1
CLIP-PAD		✓	64.5	83.7	68.2	87.3	64.9	81.8
CLIP-PAD	✓	✓	65.0	85.4	71.5	88.6	66.3	82.7

Table 5: Retrieval results on CiA with celebrities doing 'action', in 'place', or combined respectively. Our model's results are presented in the lower part of the table. 'query' column refers to the modality of the person-query used. CLIP has not been fine-tuned or modified in any way, and we train the CE model on the training set with experts and parameters given in the supplementary. '*' denotes that while CE does not use face embedding as a query, to make the comparison as fair as possible we include the face query embedding as an additional 'expert' input to the model.

5.3 Real-world retrieval example

In order to present a real-world example of personalised retrieval, we apply our model on the entire Season two of the Seinfeld TV show. We want to see how many occurrences of "Michael Richards entering the room" we can correctly retrieve from a total of 483 video clips. By our count, Richards is portrayed entering the room 26 times in the season. 21 of these were correctly retrieved in the top-25. The top-25 retrieved clips are presented in Fig. 4.

For more information about the clip extraction process, as well as expanded results with different models and in a different clip extraction regime, we refer the reader to supplementary material.

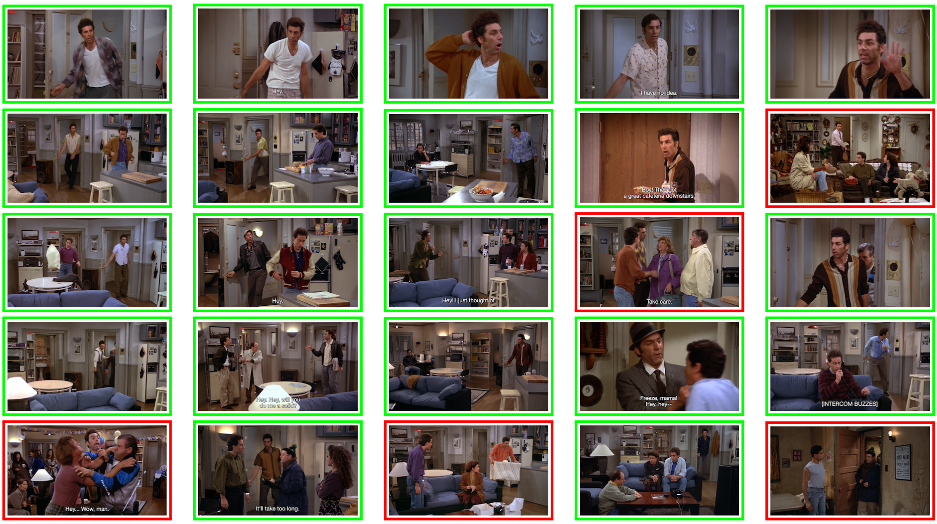

Figure 4: Center frames of top-25 retrieved clips from Seinfield Season 2, sorted from left to right and from top to bottom (top left is rank 1, bottom right is rank 25). Correctly retrieved examples have a green border, whilst incorrectly retrieved examples have a red border. Figure best seen in colour.

6 Conclusion

We have shown how CLIP can be modifed for person-centric retrieval using a face image as a form of prompt engineering for a free-form text query. The retrieval performance has been demonstrated on public datasets for celebrities. However, the same method could be applied to search personal image and video collections in order to find images of your father, say, in a particular place or doing a particular action.

More generally, this idea of prompt engineering using an image can be extended beyond faces to a particular instances of objects, for example a particular building or a particular car. The prompt does not necessarily have to be an image either – we could equally prompt the model using other modalities such as audio in order to add instance recognition in text-to-audio retrieval [30].

Acknowledgements

We would like to thank the reviewers and the area chairs of BMVC2022 for their insightful questions and suggestions. This research is supported by of the EPSRC programme Grant VisualAI EP/T028572/1 and Royal Society Research Professorship RP/R1/191132.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021.

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval. *CoRR*, abs/2205.08508, 2022.

[3] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, volume 12354 of *Lecture Notes in Computer Science*. Springer, 2020.

[4] Andrew Brown, Ernesto Coto, and Andrew Zisserman. Automated video labelling: Identifying faces by corroborative evidence. In *Multimedia Information Processing and Retrieval (MIPR)*, 2021.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 2018.

[6] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021.

[7] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[9] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, 2020.

[10] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.

[12] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with temporal language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2018.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

[15] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *CoRR*, abs/2112.04478, 2021.

[16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 2012.

[18] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *CoRR*, abs/2107.09609, 2021.

[19] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[21] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019.

[22] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[23] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[24] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[25] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a Text-Video Embedding from Imcomplete and Heterogeneous Data. In *arXiv*, 2018.

[26] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019.

[27] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020.

[28] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021.

[29] Edwin G. Ng, Bo Pang, Piyush Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. 2021.

[30] Andreea-Maria Oncescu, A. Sophia Koepke, Joao F. Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. In *INTERSPEECH*, Annual Conference Series, 2021.

[31] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[32] Mandela Patrick, Po-Yao Huang, Yuki Markus Asano, Florian Metze, Alexander G. Hauptmann, João F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[33] Alonso Patron-Perez, M. Marszałek, Andrew Zisserman, and Ian D. Reid. High five: Recognising human interactions in TV shows. In *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings*, 2010.

[34] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using CLIP. In *Pattern Recognition - 13th Mexican Conference, MCPR 2021, Mexico City, Mexico, June 23-26, 2021, Proceedings*, 2021.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

[36] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015.

[37] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 2017.

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[39] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? 2022.

[40] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[41] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 2021.

[42] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[43] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. 2022.

[44] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[45] Yujie Zhong, Relja Arandjelovic, and Andrew Zisserman. Faces in places: compound query retrieval. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[46] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.