Cross Modal Retrieval with Querybank Normalisation Supplementary Material

Simion-Vlad Bogolin[1,2,*] Ioana Croitoru[1,2,*]
Hailin Jin[3] Yang Liu[1,4,†] Samuel Albanie[1,5,†]
[1]Visual Geometry Group, University of Oxford [2]Inst. of Mathematics of the Romanian Academy [3]Adobe Research
[4]Wangxuan Inst. of Computer Technology, Peking University [5]Department of Engineering, University Cambridge

In this appendix, we provide additional information and ablation studies relating to QB-NORM. We begin by providing more details about the datasets used for each task (Sec. 1). We then provide ablation studies that investigate: (1) The influence of the k hyperparameter on the proposed DIS normalisation scheme (Sec. 2); (2) Whether effective querybanks can also be constructed from the training set using IS normalisation, rather than DIS normalisation (Sec. 3); (3) How embedding dimensionality influences the effectiveness of QB-NORM (Sec. 4). Next, we present comparisons on additional datasets for the text-video retrieval task (Sec. 5). In Sec. 6 we discuss the complexity of each normalisation technique. In Sec. 7 we present a comparison with CENT [36] normalisation. Then, we provide details on the *skewness* metric reported in the submission (Sec. 8), offer a more complete set of metrics across ablations (Sec. 9) and give more details about the text and video experts used in this work (Sec. 10). Finally, we report metrics indicating how QB-NORM performs on video-text retrieval (Sec. 11) and provide some additional qualitative results (Sec. 12).

1. Dataset details

In this section, we describe the splits and datasets employed for all tasks considered in this work.

1.1. Text-video retrieval

For the task of text-video retrieval we test our approach on seven current benchmarks.

MSR-VTT [43] contains around 10k videos, each having 20 captions. For the task of text-video retrieval, we follow prior works [9, 23] and we report results on the official split (`full`) which contains 2,990 videos for testing and 497 for validation. Since a number of recent works [9, 15, 23, 29] also report results on the 1k-A split, we compare against these method on this split as well. The 1k-A split contains 1,000 videos for testing and around 9,000 for training. We use the same videos and captions as defined in [23] which are used by other works [15, 29, 45] for evaluation. We report the results using models trained for 100 epochs.

MSVD [5] has 1,970 videos and around 80k captions. We report results on the standard split using in prior works [9,23,39,44] which consists of 1,200 videos for training, 100 for validation and 670 for testing.

DiDeMo [1] has 10,464 videos. They are collected from a large-scale creative commons collection [37] and are varied in content (concerts, sports, pets etc.). For each video, there are 3-5 pairs of descriptions. For the task of text-video retrieval, we use the paragraph video retrieval protocol as defined in prior works [9, 23, 46]. This means that we the split consisting of 8,392 for training, 1,065 validation and 1,004 test videos.

LSMDC [32] contains 118,081 short video clips extracted from 202 movies. Each clip has a textual description which consist in a caption which is extracted either from the movie script or transcribed from descriptive video services (DVS) for the visually impaired. We use the official splits as defined in the Large Scale Movie Description Challenge (LSMDC). The testing split contains 1,000 videos.

VaTeX [41] contains 3,4911 videos and has multilingual captions in Chinese and English. Each video has 10 captions for each language. As for the other datasets, we follow the same protocol as defined in prior works [6, 9, 29] and use 1,500 videos for testing, while there are 1,500 videos for validation. Please note that in this work, we use only the English annotations.

QuerYD [27] has 1,815 videos for training, 388 for validation and 390 for testing. The videos are extracted from YouTube and are varied in content. The dataset has 31,441 textual descriptions. 13,019 of these are precisely localized in the video with start time and end time annotations while the other 18,422 are coarsely localized. In this work, we do not use the localization annotations and report results on the official splits following prior work on text-video retrieval [9].

ActivityNet [3] contains 20k videos and has around

[*]Equal contribution. [†]Corresponding authors.

Querybank Source Data	Topk	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
No querybank	-	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.5_{\pm 0.1}$	$10.0_{\pm 0.0}$
In Domain					
MSR-VTT	1	$17.0_{\pm 0.1}$	$41.3_{\pm 0.1}$	$54.1_{\pm 0.1}$	$8.6_{\pm 0.5}$
MSR-VTT	2	$17.1_{\pm 0.1}$	$41.7_{\pm 0.1}$	$54.5_{\pm 0.1}$	$8.0_{\pm 0.0}$
MSR-VTT	3	$17.1_{\pm 0.1}$	$41.8_{\pm 0.1}$	$54.6_{\pm 0.1}$	$8.0_{\pm 0.0}$
MSR-VTT	5	$17.1_{\pm 0.1}$	$41.9_{\pm 0.1}$	$54.7_{\pm 0.1}$	$8.0_{\pm 0.0}$
MSR-VTT	10	$17.1_{\pm 0.1}$	$41.9_{\pm 0.1}$	$54.7_{\pm 0.1}$	$8.0_{\pm 0.0}$
Far Domain					
LSMDC	1	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.2_{\pm 0.1}$	$10.0_{\pm 0.0}$
LSMDC	2	$14.8_{\pm 0.0}$	$38.0_{\pm 0.0}$	$51.0_{\pm 0.0}$	$10.0_{\pm 0.0}$
LSMDC	3	$14.7_{\pm 0.0}$	$37.9_{\pm 0.0}$	$50.9_{\pm 0.0}$	$10.0_{\pm 0.0}$
LSMDC	5	$14.6_{\pm 0.0}$	$37.8_{\pm 0.0}$	$50.8_{\pm 0.0}$	$10.0_{\pm 0.0}$
LSMDC	10	$14.5_{\pm 0.0}$	$37.5_{\pm 0.0}$	$50.4_{\pm 0.0}$	$10.0_{\pm 0.0}$

Table 1. **The influence of the k hyperparameter on DIS normalisation.** Performance is reported on MSR-VTT `full` split [43], while querybanks of 5,000 samples are sampled from the training sets of different datasets. We observe that for *Far Domain* querybanks, $k = 1$ performs the best, while retaining good performance for *In Domain* querybanks.

100K descriptive sentences. The videos are extracted from YouTube. We use a paragraph video retrieval as defined in prior works [9, 23, 46]. We report results on the `val1` split. The training split consists of 10,009 videos, while there are 4,917 videos for testing.

1.2. Text-image retrieval

For text image retrieval, we report results on the **MSCoCo** [7] dataset. It consists of 123k images with 5 captions for each sentence. We report results for the 5k test split.

1.3. Text-audio retrieval

For text audio retrieval, we report results on the **AudioCaps** [22] dataset which comprises sounds with event descriptions. We use the same setup as prior work [28] where 49,291 samples are used for training, 428 for validation and 816 for testing.

1.4. Image-to-image retrieval

CUB-200-2011 [40] contains 11,788 images with 200 classes. The training split consist of the first 100 classes (5,863 images) while the testing split contains the remaining classes (5,924 images). We use the same setup as used in prior work [33].

Stanford Online Products [35] contains 120,053 images with products from 22,634 classes. We use the provided train and test splits containing 59,551 and 60,502 images respectively, as used in prior works [33, 35].

2. The influence of the Top-k hyperparameter on DIS normalisation

In Tab. 1 we show the influence of k in the Top-k selection employed when constructing the gallery activation set (introduced in Sec. 3.4 of the main paper). We observe that choosing $k = 1$ offers a good trade-off between good performance when constructing *In Domain* querybanks and robustness when constructing *Far Domain* querybanks. We therefore use $k = 1$ for all reported experiments.

Querybank Source	Size	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
No querybank	-	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.5_{\pm 0.1}$	$10.0_{\pm 0.0}$
Training set	60k	$17.3_{\pm 0.1}$	$42.1_{\pm 0.2}$	$54.9_{\pm 0.1}$	$8.0_{\pm 0.0}$
Val set	10k	$16.7_{\pm 0.1}$	$41.2_{\pm 0.1}$	$54.0_{\pm 0.1}$	$8.7_{\pm 0.5}$
Test set	60k	$17.5_{\pm 0.0}$	$42.4_{\pm 0.1}$	$55.1_{\pm 0.0}$	$8.0_{\pm 0.0}$

Table 2. **Effective querybanks can be constructed from the training set.** Performance is reported on MSR-VTT `full` split [43] using IS normalisation. We observe that a querybank of 60K samples from the training set performs comparably to a test set querybank.

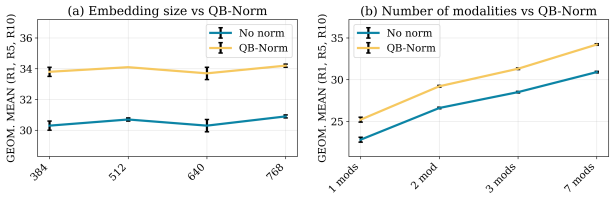

Figure 1. *(Left)*: **The influence of embedding dimension on QB-NORM effectiveness**. We observe that QB-NORM brings a large increase in performance in all cases *(Right)*: **The influence of number of used video embeddings on QB-NORM effectiveness.** We observe that our method is more effective with an increased number of modalities.

3. Can effective querybanks can be constructed from the training set with IS normalisation?

In the main submission, we showed that effective querybanks can be constructed from the training set when employing DIS normalisation. Here, we show that this property also applies to IS normalisation, supporting our hypothesis that Querybank Normalisation has the general property of not requiring concurrent access to multiple test queries for appropriate normalisation strategies. In Tab. 2 we report the results of selecting queries from training, validation or testing split to form the querybank when employing IS normalisation. Similarly to DIS, we observe that training set querybanks perform comparably to test set querybanks for IS normalisation.

4. The influence of embedding dimensionality on the effectiveness of QB-NORM

Radovanovic et al. [31] posit that hubness is a phenomenon that is: (i) inherent to high dimensional spaces; (ii) heavily influenced by the *intrinsic dimensionality* of the data. To investigate these perspectives, we study the improvement yielded by QB-NORM over embeddings of different dimensionality, reporting results in Fig. 1 (left).

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
Dual [11]	7.7	22.0	31.8	32.0
HGR [6]	9.2	26.2	36.5	24.0
MoEE [26]	$11.1_{\pm 0.1}$	$30.7_{\pm 0.1}$	$42.9_{\pm 0.1}$	$15.0_{\pm 0.0}$
CE [23]	$11.0_{\pm 0.0}$	$30.8_{\pm 0.1}$	$43.3_{\pm 0.3}$	$15.0_{\pm 0.0}$
CE+ [9]	$14.4_{\pm 0.1}$	$37.4_{\pm 0.1}$	$50.2_{\pm 0.1}$	$10.0_{\pm 0.0}$
CE+ (+QB-NORM)	$16.4_{\pm 0.0}$	$40.3_{\pm 0.1}$	$53.0_{\pm 0.1}$	$9.0_{\pm 0.0}$
TT-CE+ [9]	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.5_{\pm 0.1}$	$10.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$17.3_{\pm 0.0}$	$42.1_{\pm 0.1}$	$54.9_{\pm 0.1}$	$8.0_{\pm 0.0}$
CLIP4Clip [24]‡	27.9	52.7	63.6	5.0
CLIP4Clip (+QB-Norm)	**29.6**	**54.5**	**65.3**	**4.0**

Table 3. **MSR-VTT full split: comparison to state of the art.** ‡ denotes results obtained training using the official code.

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
MoEE [26]	$16.1_{\pm 1.0}$	$41.2_{\pm 1.6}$	$55.2_{\pm 1.6}$	$8.3_{\pm 0.5}$
CE [23]	$17.1_{\pm 0.9}$	$41.9_{\pm 0.2}$	$56.0_{\pm 0.5}$	$8.0_{\pm 0.0}$
TT-CE	$21.0_{\pm 0.6}$	$47.5_{\pm 0.9}$	$61.9_{\pm 0.5}$	$6.0_{\pm 0.0}$
Frozen [2]	31.0	59.8	72.4	3.0
CLIP4Clip [24]	**43.4**	70.2	80.6	**2.0**
CE+ [9]	$18.2_{\pm 0.2}$	$43.9_{\pm 0.9}$	$57.1_{\pm 0.8}$	$7.9_{\pm 0.1}$
CE+ (+QB-NORM)	$20.7_{\pm 0.6}$	$46.6_{\pm 0.2}$	$59.8_{\pm 0.2}$	$6.3_{\pm 0.5}$
TT-CE+ [9]	$21.6_{\pm 0.7}$	$48.6_{\pm 0.4}$	$62.9_{\pm 0.6}$	$6.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$24.2_{\pm 0.7}$	$50.8_{\pm 0.7}$	$64.4_{\pm 0.1}$	$5.3_{\pm 0.5}$
CLIP4Clip [24]‡	43.0	70.5	80.0	**2.0**
CLIP4Clip (+QB-NORM)	43.3	**71.4**	**80.8**	**2.0**

Table 4. **DiDeMo: Comparison to state of the art methods**. ‡ denotes results obtained training using the official code.

We observe that QB-NORM brings around the same gain when changing the embedding size. We can interpret this finding within the framework of [31] as making the statement that changing the shared embedding dimensionality *does not* influence intrinsic dimensionality. To provide further analysis, we make a crude approximation to increasing/decreasing intrinsic dimensionality by increasing/decreasing the number of modalities employed in the video embedding. Intuitively, since audio provides a different "view" of a sample to visual data, we expect a joint embedding with access to more modalities to exhibit higher intrinsic dimensionality than one with only visual cues. We plot the effect of these changes in Fig. 1 (right). We observe a slight increase in performance gain when applying QB-NORM with an increased number of modalities, which accords with the Radovanovic [31] hypothesis.

5. Additional text-video retrieval results

In Tab. 3, 6 we report additional comparisons with state of the art on the MSR-VTT `full` split as well as ActivityNet [3]. In both cases, we observe that QB-NORM yields improvements. We also explore the use of QB-NORM with CLIP4Clip [24]—for this, we train models using the code made available by the authors. For CLIP4Clip experiments, we use a β value of 0.45 with the exception of LSMDC where β is 1.26^{-1}.

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
MoEE [26]	$12.1_{\pm 0.7}$	$29.4_{\pm 0.8}$	$37.7_{\pm 0.2}$	$23.2_{\pm 0.8}$
CE [23]	$12.4_{\pm 0.7}$	$28.5_{\pm 0.8}$	$37.9_{\pm 0.6}$	$21.7_{\pm 0.6}$
MMT [15]	$13.2_{\pm 0.4}$	$29.2_{\pm 0.8}$	$38.8_{\pm 0.9}$	$21.0_{\pm 1.4}$
Frozen [2]	15.0	30.8	39.8	20.0
CLIP4Clip [24]	21.6	**41.8**	**49.8**	**11.0**
CE+ [9]	$14.9_{\pm 0.6}$	$33.7_{\pm 0.2}$	$44.1_{\pm 0.6}$	$15.3_{\pm 0.5}$
CE+ (QB-NORM)	$16.4_{\pm 0.8}$	$34.8_{\pm 0.4}$	$44.9_{\pm 0.9}$	$14.5_{\pm 0.4}$
TT-CE+ [9]	$17.2_{\pm 0.4}$	$36.5_{\pm 0.6}$	$46.3_{\pm 0.3}$	$13.7_{\pm 0.5}$
TT-CE+ (QB-NORM)	$17.8_{\pm 0.4}$	$37.7_{\pm 0.5}$	$47.6_{\pm 0.6}$	$12.7_{\pm 0.5}$
CLIP4Clip [24]‡	21.3	40.0	49.5	**11.0**
CLIP4Clip (+QB-NORM)	**22.4**	40.1	49.5	**11.0**

Table 5. **LSMDC: Comparison to state of the art methods**. ‡ denotes results obtained training using the official code.

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@50 \uparrow$	$MdR \downarrow$
MoEE [26]	$19.7_{\pm 0.3}$	$50.0_{\pm 0.5}$	$92.0_{\pm 0.2}$	$5.3_{\pm 0.5}$
CE [23]	$19.9_{\pm 0.3}$	$50.1_{\pm 0.7}$	$92.2_{\pm 0.6}$	$5.3_{\pm 0.5}$
HSE [46]	20.5	49.3	–	–
MMT [15]	$22.7_{\pm 0.2}$	$54.2_{\pm 1.0}$	$93.2_{\pm 0.4}$	$5.0_{\pm 0.0}$
SSB [29]	26.8	58.1	93.5	3.0
CLIP4Clip [24]	40.5	**72.4**	**98.1**	**2.0**
TT-CE+ [9]	$23.5_{\pm 0.2}$	$57.2_{\pm 0.5}$	$96.1_{\pm 0.1}$	$4.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$27.0_{\pm 0.2}$	$60.6_{\pm 0.4}$	$96.8_{\pm 0.0}$	$4.0_{\pm 0.0}$
CLIP4Clip [24]‡	36.3	65.9	96.8	3.0
CLIP4Clip (+QB-Norm)	**41.4**	71.4	97.6	**2.0**

Table 6. **ActivityNet: Comparison to state of the art methods**. ‡ denotes results obtained training using the official code.

6. The computational complexity of normalisation strategies

As discussed in the main paper in Sec.3.4, we use various normalization techniques in conjunction with QB-NORM. In this section, we describe the computational cost of each technique in the context of its influence on inference time. For clarity of exposition, we consider exact similarity searches, but note that in practice approximate nearest neighbour implementations are employed for large-scale deployments [21]. All strategies incur an initial cost that corresponds to pre-computing the similarity between a test query and all the videos from the gallery, $\mathcal{O}(N)$, where N represents the number of videos in the gallery. We further assume that we have pre-computed and stored all similarities between each query in the querybank and videos from the gallery. This assumption incurs both computational and storage costs of $\mathcal{O}(NM)$, where M represents the number of queries in the querybank.

Globally-Corrected (GC) retrieval [10] involves determining the rank of the test query with respect to the querybank for each gallery item. Since we assume that we have pre-computed similarities between the querybank and the gallery, we also pre-compute an initial ranking over querybank elements for each gallery item. For each test query, we establish its rank amongst the querybank for every target item by performing a binary search over the sorted list of pre-computed similarities. This incurs an inference time cost of $\mathcal{O}(N \log M)$.

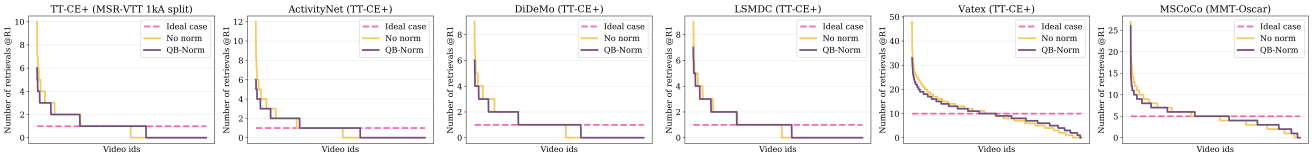

Figure 2. **Distribution of number of times each video is retrieved before and after applying QB-NORM.** We observe that QB-NORM reduces the maximum number of retrievals for any individual video. Furthermore, we note that with QB-NORM, previously unretrieved videos become possible to retrieve.

Cross-Domain Similarity Local Scaling (CSLS) [8] consists of finding the most similar queries from the querybank for each gallery video and finding the K gallery videos (here K is a hyperparameter of CSLS) that are most similar to the test query. For the former, we can pre-compute, for each video in the gallery, the K most similar queries from the querybank and store the average similarity into a vector of size N. For the latter, we must compute (during inference) the average similarity of the K most similar items among the gallery to our test query. Using *quickselect*, this can be done in $\mathcal{O}(N)$ time on average (note that we do not require the top K element similarities to be sorted, since they will be averaged).

Inverted Softmax (IS) [34] involves normalizing the final similarity by the sum of the similarities given the querybank. However, the softmax denominator can be pre-computed by summing the querybank similarities for each gallery item and storing the results into a vector of size N. During inference the similarities are divided by this pre-computed sum, which adds only constant-time overhead. Pre-computing the sum in this manner also reduces the storage cost associated with the querybank from $\mathcal{O}(NM)$ to $\mathcal{O}(N)$ (since we can discard the memory allocated to store the similarities between each query in the querybank and each video in the gallery).

Dynamic Inverted Softmax (DIS). Since DIS involves applying IS dynamically, the computation of the normalization for each test query is done in constant time as described above for IS. The additional gallery activation set employed by DIS can be pre-computed and stored for an additional $\mathcal{O}(N)$ storage cost. There is an additional cost during inference: the top-1 search to determine the video originally retrieved by the test query (which determines whether normalisation is performed). This can be done in linear time ($\mathcal{O}(N)$).

7. Comparison to CENT

In Tab. 7 we show how CENT [36] normalisation performs in comparison to an unnormalised baseline and Querybank Normalisation with DIS. Since we found CENT to consistently harm performance for cross-modal retrieval, we did not include it in all experiments in the main paper.

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
Baseline	15.0	38.4	51.5	10.0
CENT [36]	14.4	37.2	50.2	10.0
DIS	17.3	42.1	54.9	8.0

Table 7. **MSR-VTT full split** Comparison with CENT for a seed of TT-CE+ [9] model.

8. Hubness and Skewness

We use the skewness metric as defined in [31] to measure hubness:

$$S_{N_k} = \frac{E(N_k - \mu_{N_k})^3}{\sigma_{N_k}^3} \tag{1}$$

where μ_{N_k} and σ_{N_k} are the mean and standard deviation of N_k. N_k represents the k-occurrence distribution and is defined as follows $N_k(\mathbf{x}) = \sum_{i=1}^{n} p_{i,k}(\mathbf{x})$ where

$$p_{i,k}(\mathbf{x}) = \begin{cases} 1, \text{if } \mathbf{x} \text{ is among the } k \text{ nearest neighbours of } q_i \\ 0, \text{otherwise.} \end{cases} \tag{2}$$

Here $\mathbf{x}$ represents a video embedding and $q_i \in Q$ a set of queries. To compute these statistics, In practice, we use the we use $k = 10$, following [13] for the k-occurences distribution, employing the implementation of [14].

As shown in Tab. 3 in the main paper, skewness and hence hubness is reduced after applying QB-NORM. The same can be seen in Fig. 2 which depicts the distribution of number of times each video is retrieved before and after using QB-NORM. We observe that the maximum number of times a video is retrieved is reduced, indicating a hubness reduction.

9. Additional ablations on other metrics

In the main paper, to maintain conciseness we report ablation plots for the influence of querybank size and inverse temperature using the geometric mean of R1, R5 and R10. For completeness, in this section we show results on each metric individually. As seen in Fig. 3, the individual metrics reflect the trend shown for the geometric means, aligning with the results shown in the main paper.

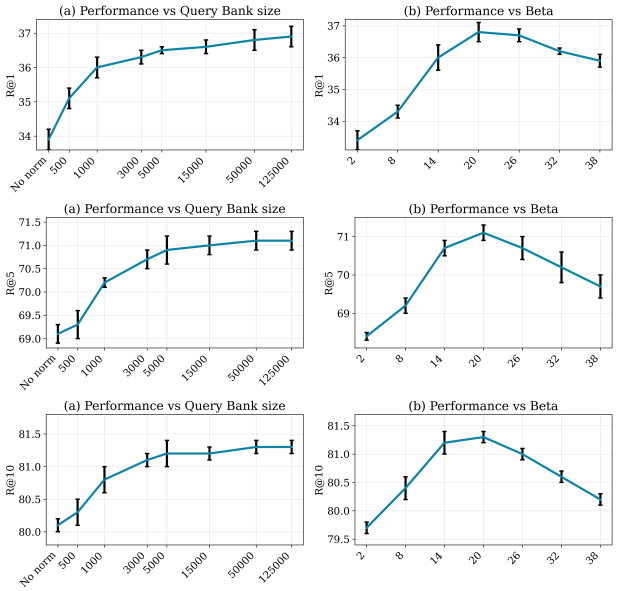

Figure 3. Retrieval results reported for a TT-CE+ [9] model on the MSR-VTT [43] `validation` split in terms of R@1, R@5 and R@10. *Left: The influence of querybank size on retrieval performance.* We observe that performance grows steadily with increasing querybank size, but saturates. *Right: The influence of inverse temperature, β.* Performance varies smoothly with inverse temperature, peaking at a value of 20.

10. Video and text embeddings (experts) description used for video retrieval

For this work, we used the pretrained weights provided by TT-CE+ [9] and CE+ [23] (`https://github.com/albanie/collaborative-experts`). These models use a set of pretrained experts. Below, we summarise how these experts were extracted.

- Two action experts are used: *Action(KN)* and *Action(IG)*. *Action(KN)* is a 1024-dimensional embedding produced by an I3D architecture trained on Kinetics [4]. The embeddings are extracted from frame clips at 25fps and center cropped to 224 pixels. For *Action(IG)* the model is a 34-layer R(2+1)D [38], trained on IG-65m [17]

- Two forms of object experts: *Obj(IN)* and *Obj(IG)*. For extracting *Obj(IN)* a SENet-154 [19] model trained on ImageNet was used. For extracting *Obj(IG)* a ResNext-101 [42] model trained on Instagram data with weakly labelled hashtags [25] was used. Both of the embeddings are extracted at 25fps.

- For producing an audio expert a VGGish model trained from audio classification on the YouTube-8m dataset [18] was used.

Model	Task	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
CE+ [9]	v2t	$22.7_{\pm 0.5}$	$52.6_{\pm 0.6}$	$66.3_{\pm 0.2}$	$5.0_{\pm 0.0}$
CE+ (+QB-NORM)	v2t	$28.6_{\pm 0.4}$	$58.9_{\pm 0.5}$	$71.4_{\pm 0.5}$	$4.0_{\pm 0.0}$
TT-CE+ [9]	v2t	$24.6_{\pm 0.3}$	$54.1_{\pm 0.3}$	$67.5_{\pm 0.5}$	$4.7_{\pm 0.5}$
TT-CE+ (+QB-NORM)	v2t	$30.1_{\pm 0.4}$	$61.4_{\pm 0.4}$	$73.2_{\pm 0.4}$	$3.0_{\pm 0.0}$

Table 8. **MSR-VTT full split: Comparison to state of the art - v2t task.**

- For the scene expert a DenseNet-161 [20] pretrained on Places365 [47] was used. The scene embedding has a 2208 dimension.

- For the speech expert, the Google Cloud API (to transcribe the speech content) is used.

- For the text we use GPT2-xl [30] finetuned as provided by the authors. The size of the final pre-trained embedding is 1600.

For CLIP2Video [12] we used the model as it is provided online `https://github.com/CryhanFang/CLIP2Video`. The model receives as input the raw frames and raw queries. For CLIP4Clip [24], we use the online code `https://github.com/ArrowLuo/CLIP4Clip` and re-train the model for each dataset where we present results since weights are not available online. For the other tasks we followed the instructions given on the official repositories. For MMT-Oscar [16] we used the pretrained weights and the features provided at `https://github.com/UKPLab/MMT-Retrieval`. For RDML [33] we used the models provided at `https://github.com/Confusezius/Deep-Metric-Learning-Baselines`. For audio retrieval [28] we used the pretrained weights and models provided at `https://github.com/oncescuandreea/audio-retrieval`.

11. v2t performance metrics

In Tab. 8 we report metrics indicating the performance of QB-NORM on the reverse task of video-text retrieval (in which videos are used as queries to retrieve descriptions). We apply QB-NORM with DIS normalisation using all videos from the training split to construct the querybank. We observe that QB-NORM yields a striking boost in performance.

12. Qualitative results

In Fig. 4 we provide some qualitative examples, illustrating cases for which the QB-NORM model correctly retrieves videos that are not retrieved without QB-NORM. Examining failure cases, we found qualitative examples for which the retrieval ranking produced with QB-NORM was more "reasonable" (as shown in the bottom set of Fig. 4). However, in line with prior work [31] suggesting that hubness is a property of the distribution (rather than driven by

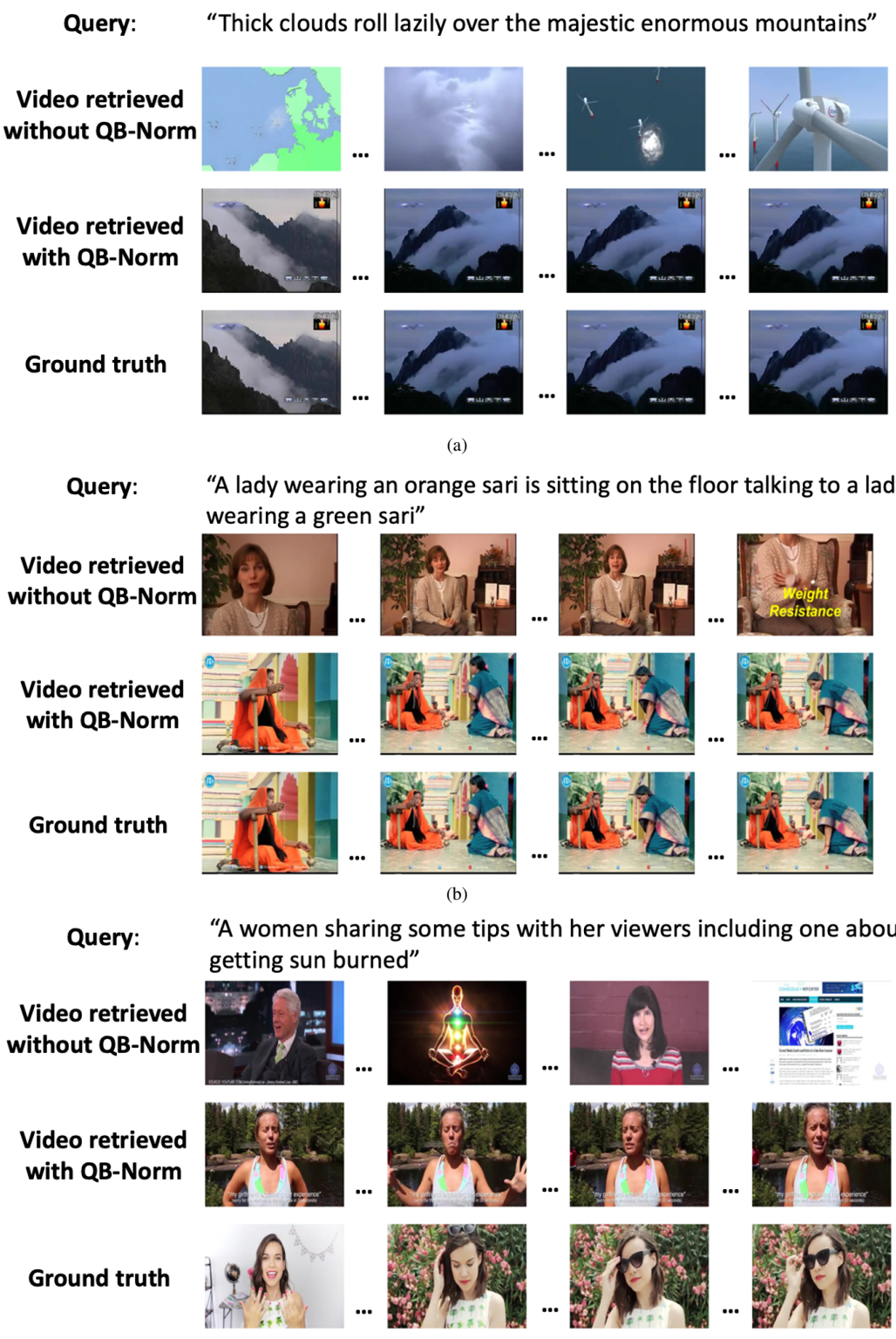

Figure 4. **Qualitative results for the text video retrieval task.** We show queries and frames from the retrieved videos. For the first two example queries, we observe that the use of QB-NORM leads to the retrieval of the correct target video. The third query represents a failure case in which the target video is not retrieved. However, we nevertheless observe qualitatively that for this example, the video retrieved with QB-NORM is more related to the query than the video retrieved without QB-NORM.

individual samples), we did not observe consistent, obvious qualitative trends among the samples that were corrected, or remaining failure cases. As an example, we observed gains for queries with both shorter and longer, highly descriptive captions.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. 2021. 3

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1, 3

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[5] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011. 1

[6] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 1, 3

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015. 2

[8] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. 4

[9] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 1, 2, 3, 4, 5

[10] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. 3

[11] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, and Xun Wang. Dual dense encoding for zero-example video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[12] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 5

[13] Roman Feldbauer, Maximilian Leodolter, Claudia Plant, and Arthur Flexer. Fast approximate hubness reduction for large high-dimensional data. *2018 IEEE International Conference on Big Knowledge (ICBK)*, pages 358–367, 2018. 4

[14] Roman Feldbauer, Thomas Rattei, and Arthur Flexer. scikit-hubness: Hubness reduction and approximate neighbor search. *Journal of Open Source Software*, 5(45):1957, 2020. 4

[15] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. *European Conference on Computer Vision*, 2020. 1, 3

[16] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *arXiv preprint arXiv:2103.11920*, 2021. 5

[17] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019. 5

[18] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. 5

[19] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 5

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019. 3

[22] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL*, 2019. 2

[23] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 1, 2, 3, 5

[24] 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. *arXiv preprint arXiv:2104.08860*, 2021. 3, 5

[25] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 5

[26] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 3

[27] Andreea-Maria Oncescu, Joao F. Henriques, Yang Liu, Andrew Zisserman Zisserman, and Samuel Albanie. Queryd: a video dataset with high-quality textual and audio narrations. *arXiv preprint arXiv:2011.11071*, 2020. 1

[28] Andreea-Maria Oncescu, A Koepke, João F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. *Interspeech*, 2021. 2, 5

[29] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 1, 3

[30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. 5

[31] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531, 2010. 2, 3, 4, 5

[32] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 1

[33] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pages 8242–8252. PMLR, 2020. 2, 5

[34] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017. 4

[35] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016. 2

[36] Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Marco Saerens, and Kenji Fukumizu. Centering similarity measures to reduce hubs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 613–623, 2013. 1, 4

[37] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1

[38] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 5

[39] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 1

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[41] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4581–4591, 2019. 1

[42] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[43] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 2, 5

[44] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 1

[45] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 1

[46] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018. 1, 2, 3

[47] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5