UATVR: Uncertainty-Adaptive Text-Video Retrieval
Supplementary Material

Bo Fang[1*] Wenhao Wu[2,3*] Chang Liu[4*] Yu Zhou[1†] Yuxin Song[3]
Weiping Wang[1] Xiangbo Shu[5] Xiangyang Ji[4] Jingdong Wang[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences [2]The University of Sydney
[3]Baidu Inc. [4]Tsinghua University [5]Nanjing University of Science and Technology

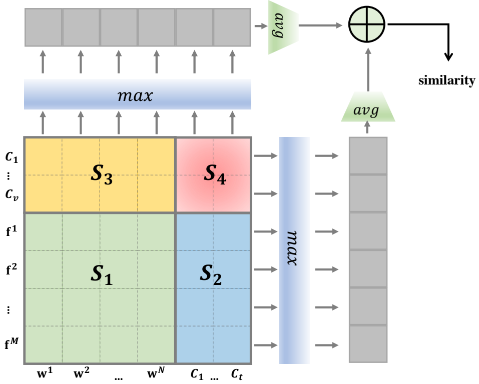


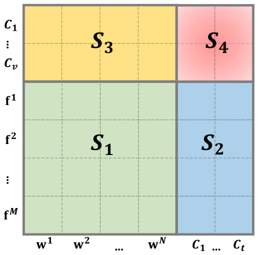
Figure A.1: Illustration of token-wise matching pipeline with additional learnable tokens appended. $\mathbf{f}^i, \mathbf{w}^i$, C_i denote frame, word, and additional tokens.

We include additional materials in this document. We describe further details on the DSA module (Sec. A) and the DUA module (Sec. B) to complement the main paper. We provide ablation studies and comparisons on coefficients (Sec. C) and post-processing operations (Sec. D). Finally, more visualizations are shown for qualitative analysis, *c.f.* Sec. E.

A. Additional Learnable Tokens

A.1. The DSA pipeline

We give particular token-wise matching procedures with our DSA tokens involved, *c.f.* Fig. A.1. S_1 denotes the token-wise matching baseline w/o appending additional tokens. And S_4 area indicates only extra learnable tokens. For each frame token $\mathbf{f}_i$ or extra video token C_i, the text token with the maximum inner production with $\mathbf{f}_i$ or C_i is selected and vice versa. Then we average the score for each modality to calculate the final similarity.



Method	R@1	
	t2v	*v2t*
$\{S_1\}$	49.6	46.8
$\{S_4\}$	45.7	42.3
$\{S_1, S_2\}$	49.5	47.9
$\{S_1, S_3\}$	50.6	47.3
$\{S_1, S_2, S_3, S_4\}$	**50.8**	**48.1**

Figure A.2: **Left**: Illustration of DSA token-wise matching partition. $\mathbf{f}^i, \mathbf{w}^i$, C_i denote frame, word, and additional tokens. **Right**: Retrieval performance on additional learnable tokens (with DUA and KL branches added).

A.2. Token-wise Matching Partition

Fig. A.2 shows token-wise matching partition results with a schematic plot. The retrieval performance is reported with DUA and KL modules added. Suppose that only extra (video/text) tokens are employed for matching, *i.e.* $\{S_4\}$, UATVR still obtains a decent 45.7% *t2v* R@1 and 42.3% *v2t* R@1. It at least demonstrates that these tokens contain meaningful knowledge for cross-modal retrieval. The baseline $\{S_1\}$ outperforms $\{S_4\}$ significantly, showing the superiority of vanilla token-wise matching. Eventually, when combining all video, text, and additionally appended tokens, *i.e.* $\{S_1, S_2, S_3, S_4\}$, UATVR gets the best results. This verifies the effectiveness of additional tokens in that they can aggregate high-level semantic information adaptively for flexible retrieval when necessary.

A.3. Additional Text Token Number C_t

In the main paper, Sec.4.2, we set the final additional appended frame token number $C_v = 3$ and word token number $C_t = 2$ in subsequent experiments. In Tab.3, $C_v = 3$ gives the best *t2v* (text-to-video) retrieval performance on the MSRVTT 1k-A test set, but $C_t = 2$ word tokens do not affect the *t2v* results much (with even little worse performance). The reason for setting extra 2 word tokens is shown

Tokens	#	**Video → Text**				
		R@1	R@5	R@10	MdR↓	MnR↓
baseline	0	45.7	74.8	82.8	2.0	9.7
C_t $(C_v = 3)$	1	47.4	75.2	84.4	2.0	8.2
	2	**47.8**	75.6	**84.7**	2.0	**8.1**
	3	47.3	**76.3**	84.6	2.0	**8.1**
	4	47.3	75.1	84.2	2.0	8.3

Table A.1: Ablation study for the number of extra text tokens. ***v2t*** retrieval results on MSR-VTT are reported.

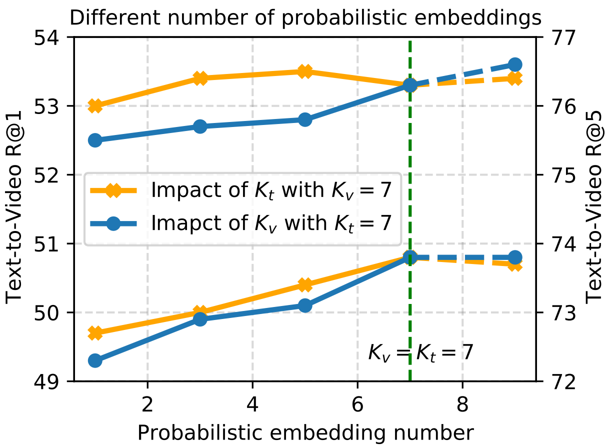


Figure B.3: Impact of sampling different numbers of (video/text) probabilistic embeddings. The video embedding number K_v is 7 when exploring the text embedding number K_t and vice versa. We report *t2v* R@1 (bottom lines) and R@5 (upper lines) results.

in Tab. A.1. We consider the *v2t* (video-to-text) retrieval direction as well and find that $C_t = 2$ gives the best *v2t* retrieval results upon the same benchmark, significantly outperforming the baseline not using extra word tokens (47.8% *vs.* 45.7%). Considering both *t2v* and *v2t* retrieval performance, $C_v = 3$ and $C_t = 2$ are finally adopted. Since the *t2v* application is more commonly used in a real scenario, we say extra C_v frame tokens are more critical than the extra C_t text tokens.

Combining both Tab.3 and Tab. A.1, we conclude that an appropriate number of extra frame tokens benefits the *t2v* retrieval most, while right amount of extra word tokens boosts the *v2t* retrieval more. This phenomenon is in line with our expectations. In the text-to-video direction, video information is dispersed into sequential frames or aggregated by extra frame tokens. The matching process thus can be determined by the most powerful *frames* in optimal granularities given a text query. It is the same for video-to-text retrieval, in which text information is better fused for a video query. This phenomenon might also explain how and why additional appended tokens could work.

B. Adaptive Distribution Matching

B.1. More details

Head modules. To convert deterministic video and text embeddings into probabilistic Gaussian distributions, we require specific head modules to formulate the video's or the text's mean and variance. Following previous text-image PVSE [10] and PCME [3], we employ basically the same probabilistic head modules in UATVR. However, vision features are extracted by Transformers in UATVR while CNNs in PVSE and PCME. Thus we pool all frame token embeddings and retain the local attention branch before distribution modeling.

Soft contrastive loss. In the main paper Tab.1, we also implement the DUA module with soft contrastive loss via Monte-Carlo estimation. Specifically, HIB [7] formulates a soft version of the contrastive loss widely used for training deep metric embeddings. For a pair of samples, the probability that the pair is matching can be defined as:

$$p(m|z_1, z_2) := \sigma(-a||z_1 - z_2||_2 + b), \tag{1}$$

where a, b are scalar parameters and $\sigma(\cdot)$ is the sigmoid function. Given the match probability $p(m|z_1, z_2)$, the soft contrastive loss is formulated as:

$$\mathcal{L}_{\text{softcon}} = \begin{cases} -\log p(m|z_1, z_2) & \text{if } z_1, z_2 \text{ is a match,} \\ -\log(1 - p(m|z_1, z_2)) & \text{otherwise,} \end{cases} \tag{2}$$

[7] has factorized Eq.1 into the match probability based on stochastic embeddings. This is done via Monte-Carlo estimation:

$$p(m|x_1, x_2) \approx \frac{1}{K^2} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} p(m|z_1^{(k_1)}, z_2^{(k_2)}), \tag{3}$$

where $z^{(k)}$ are probabilistic embeddings, which are sampled from the embedding distribution $p(m|x)$ with gradient propagating.

B.2. Probabilistic embeddings

The number of probabilistic embeddings K is 7 in the main paper Tab.4. Here, we study the impact with a different number of video and text probabilistic embeddings. We set $K_v = 7$ when changing text embedding numbers and $K_t = 7$ when adjusting video embedding numbers. As shown in Fig. B.3, we see generally performance increasing as K_v and K_t increase, during which video embedding number K_v has a more significant influence. When the

Coefficient	Value	R@1
α	0.1	50.3
	0.01	**50.8**
	0.001	49.9
β	0.001	50.1
	0.0001	**50.8**
	0.00001	50.2

Table C.2: The impact of weight coefficients α and β.

number is larger than 7, the performance starts to saturate, although further improvements can be obtained. Considering the computational overhead, we set final $K_v = K_t = 7$ (the green dash line in the figure).

C. Weight Coefficients

The total objective of UATVR is defined as $\mathcal{L}_{\text{UATVR}} = \mathcal{L}_{\text{DSA}} + \alpha \cdot \mathcal{L}_{\text{DUA}} + \beta \cdot \mathcal{L}_{\text{KL}}$. The three terms are strictly controlled by two weight coefficients, *i.e.*, α and β. In Tab. C.2, we report R@1 retrieval on MSRVTT 1k-A test set with different α and β settings. Since dynamic semantic adaptation DSA and distribution-based uncertainty adaptation DUA are jointly optimized in complementary deterministic and probabilistic views, the dramatic magnitude discrepancy of three loss terms requires low weights α and β for balance. When α=1e-1, 1e-2, 1e-3 and β=1e-4, R@1 on MSRVTT are 50.3, **50.8**, 49.9; when α=1e-2 and β=1e-3, 1e-4, 1e-5, R@1 on MSRVTT are 49.0, **50.8**, 50.2. We thus set final $\alpha = 0.01$ and $\beta = 0.0001$ in all experiments.

D. Post-Processing Operations

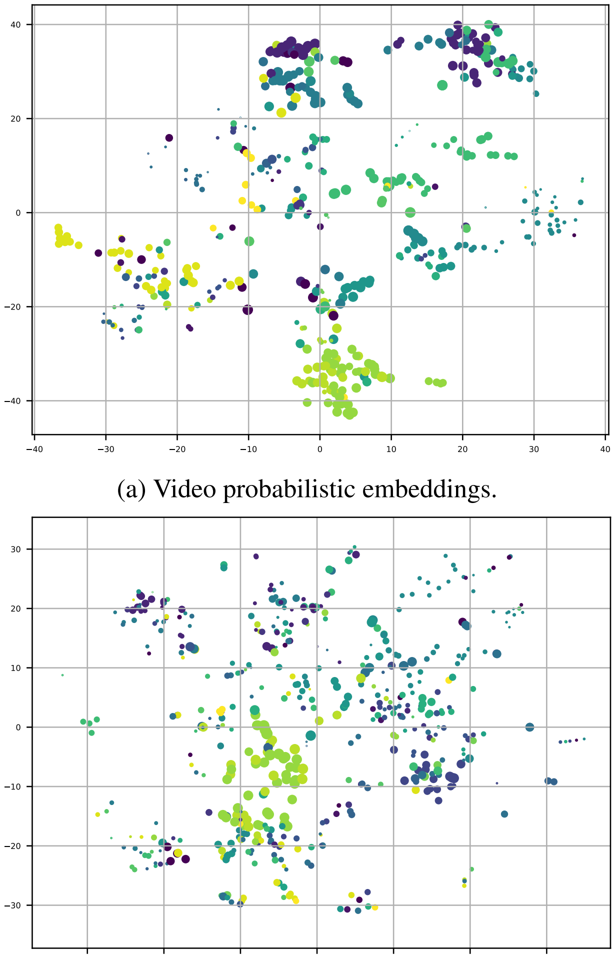
Typically, there are two common post-processing operations for text-video retrieval, *i.e.*, QB-Norm[1] and Inverted Softmax[9]. To tackle the hubness problem[8], which refers to the phenomenon that a small number of gallery points form the nearest k neighbors of many queries, QB-Norm employs a querybank to normalize similarities for reducing the similarity of a hub to the query. CAMoE[2] improves the inverted softmax by proposing Dual Softmax Loss (DSL), in which a prior matrix is introduced to revise the similarity score. In Tab. D.3, we compare UATVR with other methods using post-pocessing operations. With the same ViT-B/16 encoder and the same post operation DSL, UATVR again surpasses CLIP2TV[4] and TS2-Net[5] on most indicators. It further demonstrates that UATVR captures representative features.

Methods	R@1	R@5	R@10	MdR	MnR
QB-Norm[1]	47.2	73.0	83.0	2.0	-
CAMoE[2]	47.3	74.2	84.5	2.0	**11.9**
UATVR (ViT32)	**49.8**	**76.1**	**85.5**	**2.0**	12.9
CLIP2TV[4]	52.9	78.5	86.5	1.0	12.8
TS2-Net[5]	**54.0**	79.3	87.4	1.0	11.7
UATVR (ViT16)	53.5	**79.5**	**88.1**	**1.0**	**10.2**

Table D.3: Retrieval performance comparisons with post-processing operations.

E. More Visualization

In Fig. E.4, we visualize the video and text probabilistic embeddings extracted by our UATVR model. MSR-VTT `test 1k-A` set, about 1k text-video pairs with 20 categories, are sampled for qualitative analysis. Moreover, we present some retrieval results compared with CLIP4Clip [6] and the token-wise baseline in Fig. E.5 and some failure examples in Fig. E.6.



(a) Video probabilistic embeddings.

(b) Text probabilistic embeddings.

Figure E.4: Visualization of the probabilistic embeddings.

[1] Querybank Normalisation

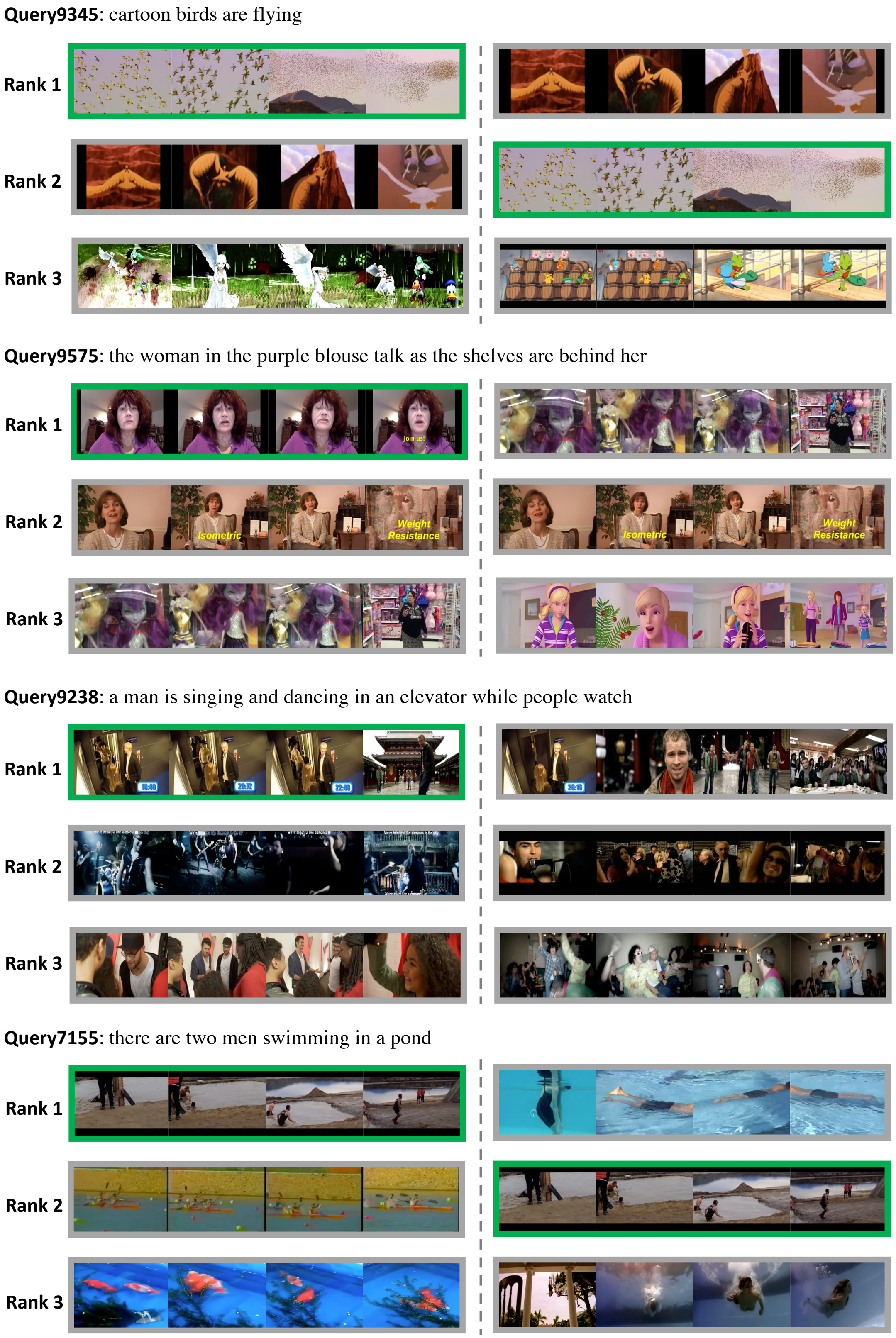

Query9345: cartoon birds are flying
Rank 1
Rank 2
Rank 3
Query9575: the woman in the purple blouse talk as the shelves are behind her
Rank 1
Join us!
Rank 2
Isometric
Weight Resistance
Isometric
Weight Resistance
Rank 3
Query9238: a man is singing and dancing in an elevator while people watch
Rank 1
Rank 2
Rank 3
Query7155: there are two men swimming in a pond
Rank 1
Rank 2
Rank 3

Figure E.5: Examples of text-to-video results on the MSRVTT 1k-A test set. The left are the videos ranked by our UATVR. And the right are results from the baseline model.

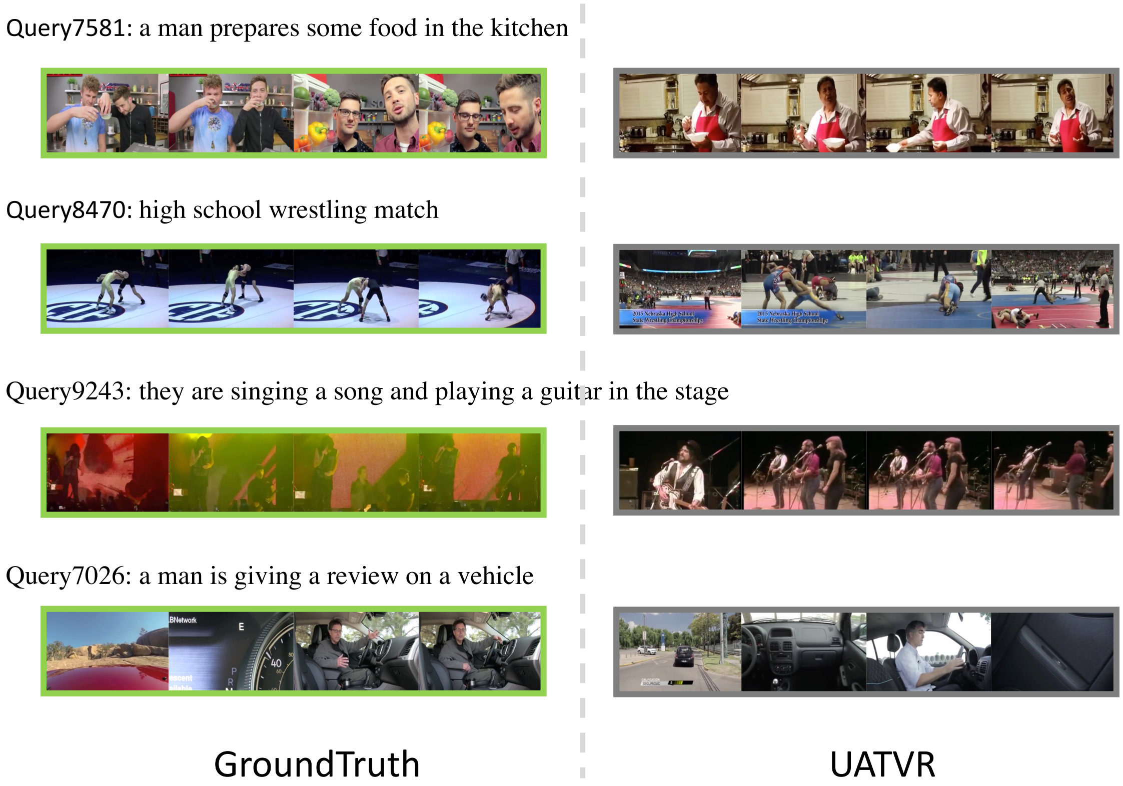

Query7581: a man prepares some food in the kitchen
Query8470: high school wrestling match
Query9243: they are singing a song and playing a guitar in the stage
Query7026: a man is giving a review on a vehicle
GroundTruth
UATVR

Figure E.6: Some retrieval failures. The left are the videos annotated by the ground-truth, and the right are the results retrieved by our UATVR. UATVR selects similar or even more appropriate videos for a given text query.

References

[1] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *CVPR*, pages 5194–5205, 2022.

[2] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

[3] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, pages 8415–8424, 2021.

[4] Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 2021.

[5] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, pages 319–335. Springer, 2022.

[6] 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.

[7] Seong Joon Oh, Kevin P Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C Gallagher. Modeling uncertainty with hedged instance embeddings. In *ICLR*, 2018.

[8] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010.

[9] 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.

[10] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, pages 1979–1988, 2019.