Supplementary Material of "DUQ: Dual Uncertainty Quantification for Text-Video Retrieval"

Author Name
Affiliation
email@example.com

The appendix provides additional details, including model performance (Sec. A), ablation studies (Sec. B), and visualizations (Sec. C).

A Performance

A.1 Performance on LSMDC

In Tab. 1, we report the text-to-video retrieval performance of the proposed DUQ framework using CLIP-ViT-B/32 and CLIP-ViT-B/16 as backbones. Additionally, we provide the results after post-processing with DSL [Cheng *et al.*, 2021], which are marked with ‡. Notably, our model, combined with the post-processing technique, achieves significantly improved retrieval performance.

A.2 Performance of Post-Processing Operations

In Tab. 2, we provide text-to-video post-processing DSL [Cheng *et al.*, 2021] retrieval results to further explore the performance of our method. A consistent performance boost can be achieved across different datasets. For example, DSL enables a 10.9% improvement in R@1 for MSRVTT with ViT-B/32 and an 11.4% improvement with ViT-B/16.

B Ablation Studies

B.1 Ablation of Video Pooling

Since the dimensionality of video features depends on the number of frames, pooling aggregation must be applied to the video to compute similarity with text features. To investigate the impact of different pooling operations, we compare four architectures, *i.e.*, "Mean", "MLP", "SA", and "CA". Fig. 1 illustrates the four pooling heads for video and Tab. 3 presents an ablation comparison of different video pooling methods. Notably, the importance of incorporating text as a feature is highlighted, as the cross-attention pooling method shows a significant improvement over the mean pooling method (+11.8%, R@1).

B.2 Ablation of Evidence-Generation Functions

How to generate evidence e from the similarity matrix $s \in \mathbb{R}^{B\times B}$ between the text features and the video features is crucial for uncertainty modeling. Following [Sensoy *et al.*, 2018], three optional functions, including the ReLU function, the softplus function, and the exponential function, are used to generate evidence as follows:

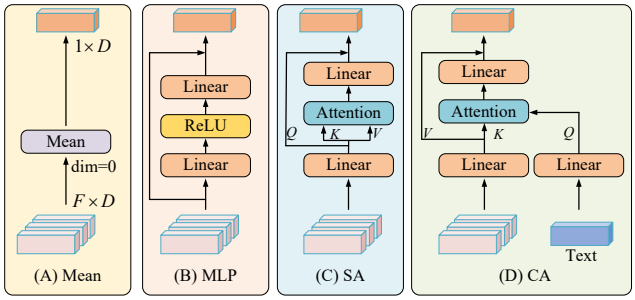

Figure 1: **The structure of the video pooling module.** We list four several popular structures, *i.e.*, "Mean", "MLP", "SA" and "CA". F and D represent the number of video frames, the feature dimension, respectively.

ReLU Function:

$$e(s) = \max(0, s). \tag{A}$$

Softplus Function:

$$e(s) = \begin{cases} \frac{1}{\gamma}\log(1+\exp(\gamma s)), & \text{if } \gamma s \le \theta \\ s, & \text{otherwise} \end{cases}, \tag{B}$$

where θ denotes a threshold for reverting to a linear function, with $\gamma = 1$ and $\theta = 20$ set in the experiments.

Exponential Function:

$$e(s) = \exp(s/\tau), \tag{C}$$

where $\tau = 5$ is a scale parameter of the exponential function. Tab. 4 compares retrieval performance without uncertainty confidence learning and with three confidence generation functions. We observe that confidence estimation improves retrieval performance, with Exponential being the optimal choice. We report ReLU retrieval results throughout the paper.

B.3 Ablation of Frame Numbers

Tab. 5 shows comparison results with different sampled frames. Notably, our model's performance improves as the number of frames increases.

*Corresponding author.
†Our code will be available soon.

Methods	LSMDC (CLIP-ViT-B/32)					LSMDC (CLIP-ViT-B/16)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CE [Liu *et al.*, 2019]	11.2	26.9	34.8	25.3	-	-	-	-	-	-
MMT [Gabeur *et al.*, 2020]	12.9	29.9	40.1	19.3	75.0	-	-	-	-	-
CLIP4Clip [Luo *et al.*, 2022]	22.6	41.0	49.1	11.0	61.0	-	-	-	-	-
TS2-Net [Liu *et al.*, 2022]	23.4	42.3	50.9	9.0	56.9	-	-	-	-	-
X-Pool [Gorti *et al.*, 2022]	25.2	43.7	53.5	8.0	53.2	26.1	46.8	56.7	7.0	47.3
DiffusionRet [Jin *et al.*, 2023c]	24.4	43.1	54.3	8.0	40.7	-	-	-	-	-
CLIP-VIP [Xue *et al.*, 2022]	25.6	45.3	54.4	8.0	-	29.4	50.6	59.0	5.0	-
CLIP-VIP ‡ [Xue *et al.*, 2022]	26.0	46.4	54.9	8.0	-	30.7	51.4	60.6	5.0	-
MSIA [Chen *et al.*, 2024]	19.7	38.1	47.5	12.0	-	-	-	-	-	
T-Mass [Wang *et al.*, 2024]	28.9	48.2	57.6	6.0	43.3	30.3	52.2	61.3	5.0	40.1
T-Mass ‡ [Wang *et al.*, 2024]	30.5	51.4	60.6	**5.0**	**40.6**	31.5	53.9	63.0	**4.0**	36.6
DUQ (Ours)	28.5	48.2	58.0	6.0	41.2	30.5	52.6	62.8	5.0	35.9
DUQ (Ours)‡	**30.6**	**51.5**	**60.0**	**5.0**	**40.6**	**33.6**	**54.8**	**64.1**	**4.0**	**33.9**

Table 1: **Text-to-video retrieval performance on the LSMDC.**

Methods	MSRVTT (Text-to-Video)					DiDeMo (Text-to-Video)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP-ViT-B/32										
UATVR [Fang *et al.*, 2023]	47.5	73.9	83.5	2.0	12.3	43.1	71.8	82.3	2.0	15.1
UATVR‡ [Fang *et al.*, 2023]	49.8	76.1	85.5	2.0	12.9	49.8	76.1	85.5	2.0	12.9
CLIP-VIP [Xue *et al.*, 2022]	50.1	74.8	84.6	**1.0**	-	48.6	77.1	84.4	2.0	-
CLIP-VIP‡ [Xue *et al.*, 2022]	55.9	77.0	86.8	**1.0**	-	53.8	79.6	86.5	**1.0**	-
T-Mass [Wang *et al.*, 2024]	50.2	75.3	85.1	**1.0**	11.9	50.9	77.2	85.3	**1.0**	12.1
T-Mass‡ [Wang *et al.*, 2024]	52.7	80.3	87.3	**1.0**	10.0	55.0	80.9	87.5	**1.0**	9.7
DUQ (Ours)	51.2	77.3	86.1	**1.0**	10.8	51.8	77.9	86.5	**1.0**	10.6
DUQ (Ours)‡	**56.8**	**81.6**	**89.1**	**1.0**	**8.2**	**58.6**	**80.9**	**87.7**	**1.0**	**8.3**
CLIP-ViT-B/16										
UATVR [Fang *et al.*, 2023]	50.8	76.3	85.5	**1.0**	12.4	45.8	73.7	83.3	2.0	13.5
UATVR‡ [Fang *et al.*, 2023]	53.5	79.5	88.1	**1.0**	10.2	53.5	79.5	88.1	**1.0**	10.2
CLIP-VIP [Xue *et al.*, 2022]	54.2	77.2	84.8	**1.0**	-	50.5	78.4	87.1	**1.0**	-
CLIP-VIP‡ [Xue *et al.*, 2022]	57.7	80.5	88.2	**1.0**	-	55.3	82.0	89.3	**1.0**	-
T-Mass [Wang *et al.*, 2024]	52.7	77.1	85.6	**1.0**	10.5	53.3	80.1	87.7	**1.0**	9.8
T-Mass‡ [Wang *et al.*, 2024]	55.9	80.4	89.6	**1.0**	9.7	58.0	82.8	88.9	**1.0**	7.5
DUQ (Ours)	55.9	81.0	88.6	**1.0**	8.4	55.8	80.0	87.4	**1.0**	8.9
DUQ (Ours)‡	**62.3**	**84.4**	**90.8**	**1.0**	**7.5**	**60.7**	**83.4**	**89.9**	**1.0**	**7.1**

Table 2: **Text-to-video post-processing retrieval performance.** ‡ denotes using inverted dual softmax for post-processing.

Video Pooling	MSRVTT (Text-to-Video)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Mean	45.8	70.3	81.2	3.0	15.7
MLP	47.2	71.4	82.3	3.0	14.9
SA	49.3	74.7	84.6	2.0	13.7
CA	**51.2**	**77.3**	**86.1**	**1.0**	**10.8**

Table 3: **Ablation study for video pooling.** We select the text-supported cross-attention "CA" as the frame aggregation head.

B.4 Ablation of Probabilistic Distances

To demonstrate the advantage of the intra-pair minimum and inter-pair maximum distance method in handling multiple probabilistic embeddings for a single feature, Tab. 6 presents a comparative experiment with traditional distance-solving methods. We observe that our model demonstrates strong retrieval performance when handling multiple probabilistic embeddings ($K \geqslant 4$). For these traditional distance-solving methods, a brief introduction is provided for comparison:

Monte-Carlo [Oh *et al.*, 2018] employs a soft cross-modal contrastive loss via Monte-Carlo estimation for distribution alignment, primarily optimizing the loss function $\mathcal{L}_{\text{softcon}}$:

$$\mathcal{L}_{softcon} = \begin{cases} -\log p(m|z_t, z_v) & \text{if } m = 1 \\ -\log(1 - p(m|z_t, z_v)) & \text{if } m = 0 \end{cases}, \quad \text{(D)}$$

where $p(m|z_t, z_v) = \sigma(-a\|z_t - z_v\|_2 + b)$, with a and b being scalar parameters, and $\sigma(\cdot)$ denoting the sigmoid function.

MIL [Song and Soleymani, 2019] refers to Multiple Instance

Methods	Eq.	MSRVTT (Text-to-Video)					MSRVTT (Video-to-Text)				
		R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Baseline	-	46.9	72.8	82.2	2.0	14.3	44.4	73.3	84.0	2.0	9.0
ReLU	Eq. (A)	51.2	**77.3**	86.1	**1.0**	10.8	50.4	79.2	87.5	**1.0**	6.4
Softplus	Eq. (B)	51.7	77.0	85.6	**1.0**	10.6	**51.1**	**79.6**	88.1	**1.0**	6.4
Exponential	Eq. (C)	**51.9**	77.1	**86.6**	**1.0**	**10.4**	50.1	79.2	**88.2**	**1.0**	**6.2**

Table 4: **Ablation study of evidence generation functions.**

$F =$	MSRVTT (Text-to-Video)					DiDeMo (Text-to-Video)				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
$F = 4$	47.9	73.5	83.2	2.0	13.2	47.2	71.0	81.2	3.0	15.4
$F = 8$	50.3	75.6	85.4	**1.0**	12.1	51.4	76.1	84.6	2.0	14.1
$F = 12$	51.2	77.3	**86.1**	**1.0**	**10.8**	51.8	77.9	86.5	**1.0**	10.6
$F = 16$	50.5	77.6	85.9	**1.0**	**10.8**	51.5	**78.7**	**87.4**	**1.0**	**10.4**
$F = 20$	**51.6**	**77.7**	85.9	**1.0**	10.9	**52.4**	77.8	85.9	**1.0**	10.8

Table 5: **Ablation study of different sampling frame numbers F.**

Methods	Eq.	$K = 4$			$K = 7$			$K = 10$		
		R@1↑	R@5↑	MnR↓	R@1↑	R@5↑	MnR↓	R@1↑	R@5↑	MnR↓
Baseline	-	44.8	73.5	16.2	46.9	74.5	14.3	45.9	75.2	15.2
Cosine Distance	-	45.1	71.2	15.6	47.1	73.6	13.4	46.1	75.1	15.2
Monte-Carlo [Oh *et al.*, 2018]	Eq. (D)	46.7	72.4	15.3	47.3	70.5	13.1	46.2	74.9	14.9
MIL [Song and Soleymani, 2019]	Eq. (E)	47.2	74.3	14.8	47.9	73.5	13.1	45.4	73.8	15.4
Unif Loss [Chun *et al.*, 2021]	Eq. (F)	48.1	75.4	14.1	48.7	74.2	12.6	48.3	74.9	13.2
KL-divergence	-	48.3	75.3	13.9	48.1	75.4	12.1	47.8	75.1	13.1
2-Wasserstein	-	47.1	74.1	14.9	48.2	75.7	12.3	47.6	74.1	13.1
UDA [Fang *et al.*, 2023]	Eq. (G)	47.5	74.4	12.9	47.5	73.9	12.3	48.0	74.0	12.9
UDA‡ [Fang *et al.*, 2023]	Eq. (G)	48.9	75.4	**12.3**	50.0	76.1	11.5	50.6	76.0	12.9
DUQ (Ours)	-	**49.5**	**76.5**	12.5	**51.2**	**77.3**	**10.8**	**50.7**	**76.7**	**10.8**

Table 6: **Ablation study of different probabilistic distance.**. ‡ indicates the incorporation of the EDL uncertainty model.

Learning loss, which extends cosine distance by introducing a learning constraint specifically designed for the cross-modal retrieval scenario:

$$\mathcal{L}_{mil} = \max\left(0, \rho - \min_{m,n} d(z_t^{i,m}, z_v^{j,n}) + \min_{m,n} z_t^{i,m}, z_v^{i,n}\right), \tag{E}$$

where ρ is a margin parameter.

Uniformity Loss ensures that feature vectors are uniformly distributed on the unit hypersphere, improving the representation quality for L2-normalized features:

$$\mathcal{L}_{\text{unif}} = \sum e^{-2\|z_t - z_v\|_2^2}. \tag{F}$$

UDA [Fang *et al.*, 2023] represents distribution-based uncertainty adaptation. This method modifies the Multi-Instance InfoNCE loss [Oh *et al.*, 2018], which treats all probabilistic embeddings from a matched text-video pair as positive samples to simulate one-to-many cross-modal matching. Specifically, given the probabilistic embedding $z^{i,m}$, we construct the positive set $\mathcal{Z}^i = \{z^{i,m}\}_{m=1}^K$ and the negative set $\tilde{\mathcal{Z}}^i = \{z^{j,n}\}_{j\neq i}$. The distribution-based uncertainty adaptation loss from z_t to z_v is then defined as:

$$\mathcal{L}_{z_t \to z_v} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{z_t^i \in \mathcal{Z}^i} \exp(s(z_t^i, z_v^i))}{\sum_{z_t^j \in \mathcal{Z}^i \cup \tilde{\mathcal{Z}}^i} \exp(s(z_t^i, z_v^j))}, \tag{G}$$

where $s(\cdot)$ denotes the cosine similarity calculation.

C Visualization

In Fig. 2, we visualize similarity s, similarity uncertainty u, and probabilistic distance d during the text-video retrieval process. In Fig. 3, we visualize several seemingly failed retrieval queries. The root cause of this issue lies in the inherent defects of the dataset itself. In contrast, we intuitively argue that our retrieval results align better with the corresponding query conditions.

Figure 2: **Qualitative retrieval results**. Examples of text-to-video retrieval results on the MSRVTT.

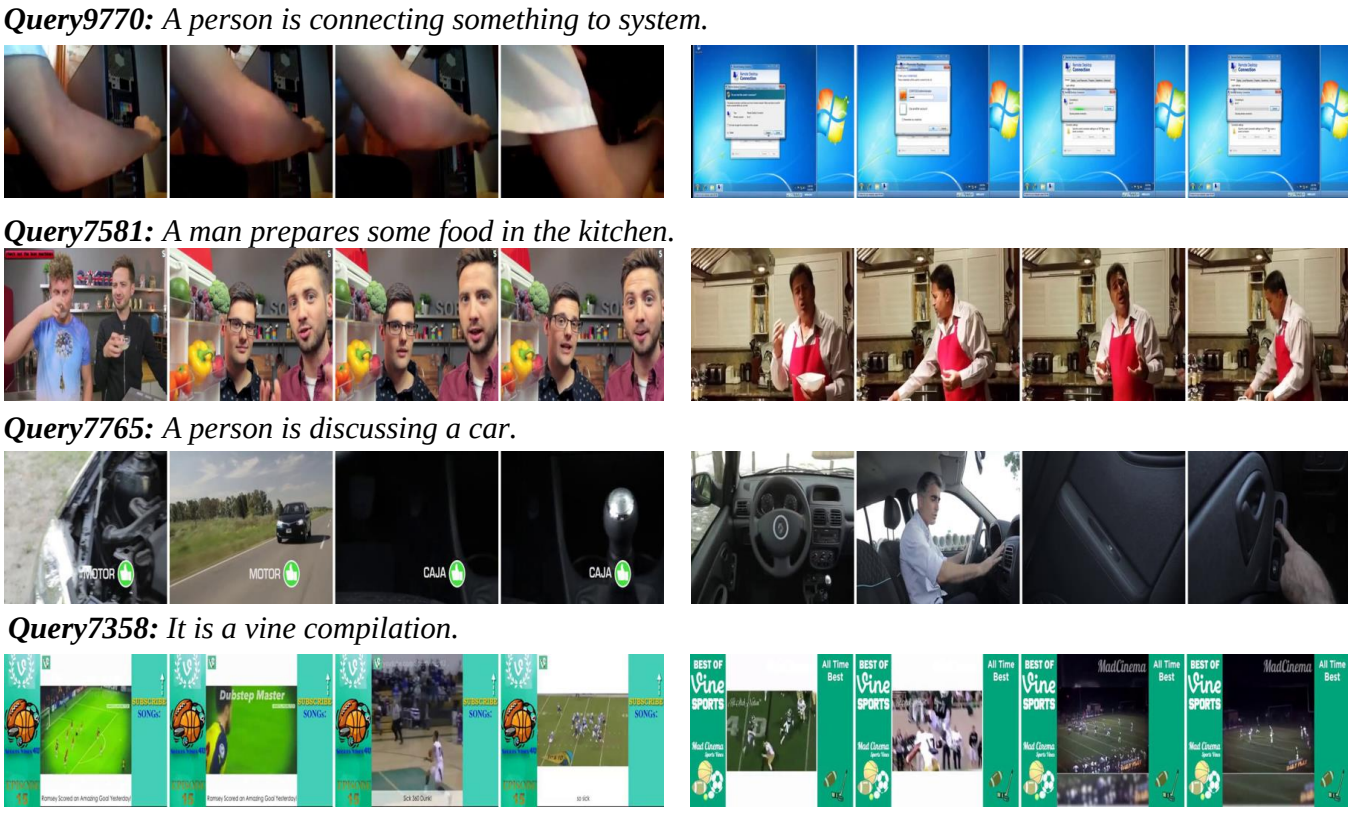

Figure 3: **Failed retrieval case**. The left side shows the ground-truth retrieval results labeled in the dataset, while the right side displays the results obtained by our model. Compared to the ground-truth, our retrieval results are more accurate.

References

[Anne Hendricks *et al.*, 2017] 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.

[Chen *et al.*, 2020] 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.

[Chen *et al.*, 2024] Lei Chen, Zhen Deng, Libo Liu, and Shibai Yin. Multilevel semantic interaction alignment for video–text cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Cheng *et al.*, 2021] 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.

[Chun *et al.*, 2021] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021.

[Fang *et al.*, 2023] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13723–13733, 2023.

[Gabeur *et al.*, 2020] 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 16*, pages 214–229. Springer, 2020.

[Gorti *et al.*, 2022] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022.

[Jin *et al.*, 2022] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in neural information processing systems*, 35:30291–30306, 2022.

[Jin *et al.*, 2023a] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482, 2023.

[Jin *et al.*, 2023b] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 938–946. International Joint Conferences on Artificial Intelligence Organization, 2023.

[Jin *et al.*, 2023c] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2470–2481, 2023.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[Li *et al.*, 2023] Hao Li, Peng Jin, Zesen Cheng, Songyang Zhang, Kai Chen, Zhennan Wang, Chang Liu, and Jie Chen. Tg-vqa: ternary game of video question answering. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1044–1052, 2023.

[Li *et al.*, 2024] Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. *Advances in Neural Information Processing Systems*, 36, 2024.

[Lin *et al.*, 2022] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *European Conference on Computer Vision*, pages 413–430. Springer, 2022.

[Liu *et al.*, 2019] 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.

[Liu *et al.*, 2022] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, pages 319–335. Springer, 2022.

[Luo *et al.*, 2022] 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.

[Oh *et al.*, 2018] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. *arXiv preprint arXiv:1810.00319*, 2018.

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

[Piergiovanni *et al.*, 2022] AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In *European Conference on Computer Vision*, pages 76–94. Springer, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rohrbach *et al.*, 2015] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[Sigurdsson *et al.*, 2016] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[Song and Soleymani, 2019] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[Torabi *et al.*, 2016] Atousa Torabi, Niket Tandon, and Leon Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv:1609.08124*, 2016.

[Wang *et al.*, 2019] 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/CVF international conference on computer vision*, pages 4581–4591, 2019.

[Wang *et al.*, 2023] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2816–2827, 2023.

[Wang *et al.*, 2024] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16551–16560, 2024.

[Wu *et al.*, 2023] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713, 2023.

[Xu *et al.*, 2016] 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.

[Xue *et al.*, 2022] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[Yang *et al.*, 2021] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021.

[Yu *et al.*, 2018] 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.

[Zeng *et al.*, 2022] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022.