Supplementary materials: Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval

Jianfeng Dong[1,2] , Minsong Zhang[1*], Zheng Zhang[1*], Xianke Chen[1]
Daizong Liu [3], Xiaoye Qu [4], Xun Wang [1,2], Baolong Liu [1,2 †]
[1]Zhejiang Gongshang University, [2]Zhejiang Key Lab of E-Commerce
[3]Peking University, [4]Huazhong University of Science and Technology

Due to the limited space, we here report more experimental results and technical details which are not included in the paper:

- The generalizability of the proposed dynamical knowledge distillation strategy on the pre-trimmed short-duration video dataset (Section 1).
- Further investigations on the performance of the feature-based knowledge and similarity-based knowledge distillations, respectively (Section 2).
- Studies on dual learning in terms of the hyper-parameter β and complementarity analysis (Section 3).
- Comparison in terms of time complexity and memory consumption (Section 4).
- More implementation details of our method (Section 5).

1. Resutls on pre-trimmed short-duration video dataset

Recall that the previous experiments are all conducted on untrimmed long videos. To verify the generalizability of the proposed dynamical knowledge distillation strategy, we conduct experiments on a pre-trimmed short-duration dataset, *i.e*. MSR-VTT [11], in the context of traditional T2VR. We adopt Dual Encoding [3] as our baseline, considering its source code released and has been widely used as the baseline in T2VR. As in Table 1, by using our proposed dynamical knowledge distillation strategy, Dual Encoding [3] achieves a further performance gain. Besides, our dynamical knowledge distillation still outperforms the fixed counterpart that uses a fixed weight during the distillation. The results not only verify the generalizability of the proposed dynamical knowledge distillation strategy, and again confirm its advantage over the fixed distillation.

*Both authors contributed equally to this work.
†Corresponding author: Baolong Liu (liubaolongx@gmail.com)

Table 1. The effectiveness of our dynamical knowledge distillation strategy for traditional T2VR. We adopt Dual Encoding as the baseline method.

Distillation	R@1	R@5	R@10	R@100	SumR
✗	10.5	28.4	38.9	89.5	167.2
✓(Fixed)	11.1	30.0	41.0	89.2	171.3
✓(Dynamic)	**11.7**	**30.9**	**42.0**	**91.8**	**176.4**

Table 2. Performance comparison of similarity-based distillation and feature-based distillation with various losses, *i.e*. MSE and KL divergence loss on ActivityNet. The similarity-based distillation with the KL loss performs the best.

		ActivityNet-Captions				
		R@1	R@5	R@10	R@100	SumR
Similarity-based	**MSE**	6.9	22.5	35.3	77.1	142.1
	KL	**8.0**	**25.0**	**37.5**	**77.1**	**147.6**
Feature-based	**MSE**	6.6	22.2	34.5	**75.7**	138.9
	KL	**6.8**	**22.3**	**34.6**	75.6	**139.2**

Table 3. Performance comparison of similarity-based distillation and feature-based distillation with various losses on TVR.

		TVR				
		R@1	R@5	R@10	R@100	SumR
Similarity-based	**MSE**	13.7	33.4	44.6	83.6	175.4
	KL	**14.4**	**34.9**	**45.8**	**84.9**	**179.9**
Feature-based	**MSE**	13.6	32.7	43.9	83.2	173.4
	KL	**14.0**	**33.3**	**44.2**	**83.6**	**175.0**

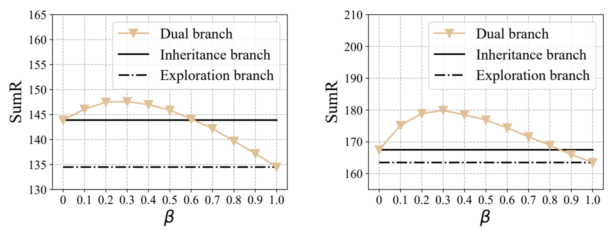
2. Similarity-based Distillation vs. Feature-based Distillation

Although knowledge distillation is conducted in many tasks [6,10] mainly by transferring features from the teacher model to the student model. Our experiment results show that, in PRVR, guiding the student model by constraining the consistent semantic similarity distributions between the teacher-student models is more effective.

Table 2 and Table 3 summarize the comparison results of similarity-based and feature-based distillations on ActivityNet [7] and TVR [8]. It is noticed that the similarity-

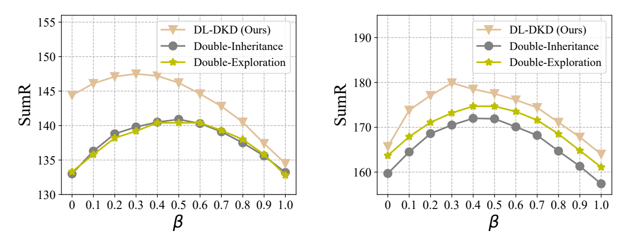
Table 4. Complexity comparison in terms of computation overhead at the inference stage and memory consumption at the training stage. All models utilize the same vision backbone. Note that the computation cost excludes the vision backbone and RoBERT. For a comprehensive comparison, we also report the performance on ActivityNet and TVR.

Text backbone	**Model**	**FLOPs** (G)	**Memory (MiB)**	**ActivityNet**	**TVR**
Graph-based	HGR [1]	2.96	8555	107.0	50.1
	DL-DKD (Ours)	1.01	4229	125.4	103.6
Bi-GRU	DE [3]	5.24	5837	121.7	123.4
	DE++ [4]	5.30	3515	121.7	128.3
	DL-DKD (Ours)	0.98	4057	137.3	145.1
RoBERTa	RIVRL [5]	8.64	4809	117.8	135.6
	XML [8]	0.80	2451	128.4	155.1
	ReLoCLNet [12]	0.96	2673	126.6	157.1
	MS-SL [2]	1.22	5349	140.1	172.4
	DL-DKD (Ours)	1.04	4455	147.6	179.9



(a) ActivityNet (b) TVR

Figure 1. The influence of hyper-parameter β that balances the similarities obtained by the inheritance and exploration branches.



(a) ActivityNet (b) TVR

Figure 2. Performance comparsion to two-branch baselines of varying hyper-parameter β.

based distillation outperforms the feature-based distillation with clear margins. Additionally, for distillation losses, we compare Kullback–Leibler (KL) divergence loss we used to the Mean Square Error (MSE) loss. As shown in the two tables, the KL divergence loss consistently performs better than that of MSE loss. On the whole, we employ similarity-based distillation and the KL loss in this work.

3. Studies on dual learning

Influence of the hyper-parameter for two-branch inference Fig. 1 shows the influence of hyper-parameter β in Eq.9 of our paper. Note that $\beta = 0$ indicates that only the similarity obtained inheritance branch is employed, while $\beta = 1$ indicates using the exploration branch counterpart. The inheritance branch consistently outperforms the exploration one on both datasets, showing the benefit of learning knowledge from the teacher model. The best performance is achieved at β of 0.3, where the inheritance branch is dominated for the final similarity. It is worth noting that our proposed model is not very sensitive to beta, as we find that the proposed method holds state-of-the-art (SOTA) results on both ActivityNet and TVR datasets when beta is in the range from 0 to 0.7 (as shown in Figure 1).

Comparison to two-branch baselines. In this experiment, we tune β (Eqn. 9) for the double-branch baselines, *i.e*., Dual-exploration and Dual-inheritance. The results are summarized in Fig. 2. With the sameβ, our model consistently performs better, which further demonstrates the benefit of using two hybrid branches.

4. Model Complexity

Table 4 summarizes the model complexity comparison in terms of time complexity and memory consumption, and their corresponding performance on ActivityNet and TVR. For a specific method, its time complexity is measured as FLOPs it takes to encode a given video-text pair. For a more fair comparison, we compare previous works using the same text backbone. Our model has comparable FLOPs and memory usage, but gives better performance.

Additionally, it is worth noting that although our model has two branches, they utilize the same vision and text backbones, and only the encoder headers are doubled. Moreover, as the majority of the computation cost is in the vision (222G FLOPs) and text backbones (79G FLOPs), the computation cost only increases 0.3% when extending from one branch to two branches.

5. Implementation Details

The dimension sizes of the video features extracted by the pre-trained CNN model of the ActivityNet-Captions and TVR datasets are 1024 and 3072, respectively. The dimensions of all the above features are linearly reduced to 384 further for the convenience of the Transformer's (384 hidden sizes, 4 attention heads) feature encoding. For a textual query, we employ the pre-trained Roberta [9] to extract a feature with 1024 dimension firstly, reduce the dimension of the feature to 384 further, and then feed the feature to a Transformer (384 hidden sizes, 4 attention heads) for feature encoding. For different decay strategies, we empirically set the k to 0.95 and 800 for the exponential decay and Sigmoid decay, respectively. For the linear decay strategy function, we set the parameters k and b to -0.01 and 1, respectively. We utilize an Adam optimizer with a mini-batch size of 128 for model training.

References

[1] 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. 2

[2] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257, 2022. 2

[3] 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*, pages 9346–9355, 2019. 1, 2

[4] 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*, 2022. 2

[5] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2

[6] Shitian He, Huanxin Zou, Yingqian Wang, Runlin Li, Fei Cheng, Xu Cao, and Meilin Li. Enhancing mid–low-resolution ship detection with high-resolution feature distillation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 1

[7] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1

[8] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020. 1, 2

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[10] Siyu Ren and Kenny Q Zhu. Leaner and faster: Two-stage model compression for lightweight text-image retrieval. *arXiv preprint arXiv:2204.13913*, 2022. 1

[11] 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

[12] 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*, pages 685–695, 2021. 2