#6148: "DUQ: Dual Uncertainty Quantification for Text-Video Retrieval"

R#1 Q1. non-CLIP architectures. We choose CLIP backbones following previous work [1]. Besides, our work focuses on investigating the uncertainty in feature interactions, rather than developing an advanced backbone.

R#2 Q1. Lack of empirical evidence. 1) The empirical evidence can be represented by performance R@1 for retrieval tasks. 2) And it is affected by the uncertainty mass u in Eq. 3 and Eq. 4 during sampling. Experiments on MSRVTT show that the uncertainty mass u can enhance the empirical evidence, e.g., R@1 improves from 46.9% to 51.2% in Tab. 3. 3) To enhance the understanding of the gain in empirical evidence before and after applying DUQ, we visualize this change in Fig. 3. The left side shows the retrieval failures by X-Pool based on similarity s, while the right side shows the successful results after re-ranking by DUQ based on similarity s', validating the effectiveness of DUQ.

R#2 Q2. Failed retrieval case. The proposed method fails when video (NOT-GT) with more matching query appear. In ReFig. 1, the query is "A man prepares some food in the kitchen," but the GT video contains two men in the kitchen. In contrast, our method which retrieves a video with one man in the kitchen is regarded as a failed retrieval. More failed cases refer to Fig. 3 in supplementary material.

ReFig. 1: Failed retrieval case. (zoom in to view)

R#2 Q3. In which case is deterministic modeling more effective than probabilistic? We believe deterministic modeling outperforms probabilistic modeling on datasets with fewer false positives. If we remove a large number of false positives similar to GT, we can establish highly matched text-video relationships. Meanwhile, in this case, deterministic modeling does not require complex network designs to construct diverse probabilistic embeddings and exhibits excellent time efficiency, as shown in ReTab. 1.

R#2 Q4. Computational cost of DUQ. We conduct additional experiments on the MSRVTT to evaluate the computational cost, considering time complexity, parameters, training time, and sample time in ReTab. 1. Although our method is slightly inferior to existing methods in terms of computational cost, it achieves the best performance reflected by R@1.

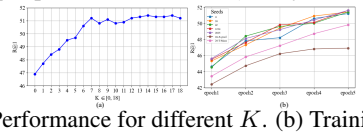
Method	Complexity	Params	TrainTime	SampleTime	R@1↑
Clip4clip	$\mathcal{O}(B^2)$	162.3M	15.79h	168.4s	44.5
X-Pool	$\mathcal{O}(B^2)$	**152.6M**	**14.32h**	**157.4s**	46.9
DUQ	$\mathcal{O}(KB^2)$	165.7M	18.31h	200.3s	**51.2**

ReTab. 1: Computational cost of different methods on the MSRVTT.

R#5 Q1. Lack of hyper-parameter sensitivity analysis. 1) We analyze the sensitivity of the hyper-parameters α, β, and γ by setting different values for them in additional experiments. In ReTab. 2, the variation of these hyper-parameters does not significantly affect our results, indicating that their sensitivity is relatively low. 2) We further explore the rationale behind the choice of K in the ablation study by varying its value from 0 to 18. In ReFig. 2 (a), the model's retrieval performance begins to converge when K reaches 7. We choose K=7 due to low computational cost.

α	R@1↑	β	R@1↑	γ	R@1↑
0.00	46.9	0.0000	50.3	0.00	50.7
0.01	50.0	0.0001	**51.2**	0.01	50.9
0.05	50.5	0.0005	51.0	0.05	51.0
0.10	**50.9**	0.0010	50.5	0.10	**51.2**
0.20	50.7	0.0050	50.4	0.50	51.1

ReTab. 2: Hyper-parameter α, β, and γ analysis on the MSRVTT.



ReFig. 2: (a) Performance for different K. (b) Training stability for different random seeds on the MSRVTT. (zoom in to view)

R#5 Q2. Lack of variance analysis. 1) Intuitively, training instability arises from the addition of $\epsilon \sim \mathcal{N}(0, I)$ to the variance in the probabilistic embedding in Eq. 10. 2) To prove that the results obtained by DUQ are not affected by the variance of sampled features, we can set different random seeds and compare them with the baseline performance. 3) ReFig. 2 (b) shows a comparison of the results at each epoch with the baseline X-Pool under multiple random seeds. As the epochs of DUQ increase, R@1 also grows, demonstrating the stable performance improvement of DUQ.

R#5 Q3. Unclear optimality of loss design. 1) We use $\mathcal{L}_S$ as our baseline, consistent with previous studies. To demonstrate the importance of $\mathcal{L}_S$, we conduct ablation studies based different combination of $\mathcal{L}_S$ and $\mathcal{L}_S^U$. As shown in ReTab. 3, model using both $\mathcal{L}_S$ and $\mathcal{L}_S^U$ outperforms those utilizing only $\mathcal{L}_S$ or $\mathcal{L}_S^U$ individually. Besides, we conduct ablation studies to highlight the effectiveness of ISUM and IDUM, as shown in ReTab. 3. The results confirm the necessity of ISUM. More ablation studies are done for demonstrating the necessity and complementarity of each loss component in Tab. 3 of the original paper.

R#5 Q4. Lack of integration. Good suggestion! In the proposed method, we only integrate the components during the sampling phase via the similarity s' calculated in Eq. 17. For improving the integration strategy, we use s' to integrating the components in the training phase at each batch. Since the training and sampling objectives are the same, the experimental results show that the model achieves an R@1 of 51.5% on MSRVTT. We will explore more effective integration strategies in future work.

R#5 Q5. Design of two loss objectives. 1) The ISUM loss objective is our original design, and its effectiveness is demonstrated through ablation studies as shown in Tab. 3 of the original paper. 2) Good suggestion! We admit that maximizing the maximum distance is a rational choice. We further conduct an ablation study on MSRVTT by designing a variant model with maximizing the maximum distance setting. The variant model achieves an R@1 score of 50.9%. Compared to our approach, which utilizes IDUM and achieves an R@1 score of 51.2%, the performance gap is minimal. Therefore, we believe both approaches are viable.

$\mathcal{L}_S$	$\mathcal{L}_S^U$	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
✓		46.9	**74.5**	82.2	**2.0**	14.3
	✓	47.1	73.6	83.3	**2.0**	12.4
✓	✓	**47.5**	73.9	**83.5**	**2.0**	**12.3**
ISUM	IDUM	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
✓		47.5	73.9	83.5	2.0	12.3
	✓	48.3	74.6	84.8	2.0	12.4
✓	✓	**51.2**	**77.3**	**86.1**	**1.0**	**10.8**

ReTab. 3: Comparison of different losses and modules individually.

[1] *Xiaobo Shen et al. Contrastive Transformer Cross-Modal Hashing for Video-Text Retrieval. IJCAI, 2024*