# Augmenting Moment Retrieval: Zero-Dependency Two-Stage Learning

## Supplementary Material

## 1. Additional Implementation Details

Our model architecture employed a transformer-based design with a hidden dimension of 256 and a feed-forward network (FFN) that expanded this dimension to 4× (1024 units) using ReLU activation. It incorporated 8 parallel attention heads and 10 queries for each group, with a dropout rate of 0.1 applied within transformer layers for regularization. During training, the batch size is set to 32. Notably, no post-processing techniques such as non-maximum suppression were applied.

## 2. Additional Ablation Study

| $\lambda_{\text{disc}}$ | R1 | | mAP | | |
|---|---|---|---|---|---|
| | @0.5 | @0.7 | @0.5 | @0.75 | Avg. |
| 0.3 | 69.81 | 55.81 | 69.93 | 52.93 | 51.13 |
| 0.5 | **70.13** | **56.65** | **70.28** | **53.20** | **51.66** |
| 0.7 | 69.29 | 55.26 | 69.41 | 52.17 | 50.84 |
| 0.9 | 69.87 | 56.19 | 69.84 | 52.72 | 50.95 |

Table 1. Ablation study according to the parameter $\lambda_{\text{disc}}$ on the QVHighlights validation set.

| $\lambda_{\text{dill}}$ | R1 | | mAP | | |
|---|---|---|---|---|---|
| | @0.5 | @0.7 | @0.5 | @0.75 | Avg. |
| 0.3 | 69.42 | 54.52 | 69.69 | 52.50 | 50.58 |
| 0.5 | 70.13 | 56.65 | **70.28** | **53.20** | **51.66** |
| 0.7 | **70.45** | **56.71** | 70.12 | 52.91 | 51.25 |
| 0.9 | 69.35 | 55.23 | 68.42 | 50.86 | 49.79 |

Table 2. Ablation study according to the parameter $\lambda_{\text{dill}}$ on the QVHighlights validation set.

**Impact of $\lambda_{\text{disc}}$** The results in Table 1 demonstrate that the choice of $\lambda_{\text{disc}}$ significantly influences the model's performance. As $\lambda_{\text{disc}}$ increases from 0.3 to 0.5, there is a noticeable improvement across all evaluation metrics, with the highest values observed at $\lambda_{\text{disc}} = 0.5$. Beyond this point, performance slightly degrades, suggesting that an excessive weight on this parameter may hinder the model's ability to generalize effectively. This indicates that moderate regularization through $\lambda_{\text{disc}}$ is beneficial for achieving optimal results.

**Impact of $\lambda_{\text{dill}}$** As shown in Table 2, the parameter $\lambda_{\text{dill}}$ also plays a crucial role in model performance. The best results are obtained when $\lambda_{\text{dill}} = 0.5$, with a peak in both R1 and mAP scores. Increasing $\lambda_{\text{dill}}$ to 0.7 slightly enhances R1 but leads to a minor drop in mAP, suggesting a trade-off between precision and recall. Further increasing $\lambda_{\text{dill}}$ to 0.9 results in a notable decline across all metrics, indicating that excessive reliance on this parameter negatively impacts the model's effectiveness.

These results collectively highlight the robustness of our model, achieving state-of-the-art performance across different parameter choices, demonstrating its adaptability to various hyperparameter settings.

## 3. Additional Qualitative Results

As illustrated in Fig. 1, we conducted additional qualitative comparisons on the QVHighlights [1] validation set against UVCOM [2] and TaskWeave [3]. The results demonstrate that our method significantly outperforms previous models, particularly exhibiting superior capabilities in accurate semantic comprehension and precise boundary localization. These findings substantiate the effectiveness of our proposed approach.

## 4. Limitations

Despite the significant progress of our framework in video moment retrieval tasks, there are still some limitations and potential directions for improvement. Firstly, even though the Splice-and-Boost augmentation strategy improves boundary and semantic discrimination, we can explore more realistic data augmentation methods. This will help models adapt more smoothly to the distribution of real data during the two-stage training. Secondly, in the two-stage training, while we've introduced the Discriminative Contrastive Loss, more in-depth research is needed on how to better utilize the knowledge from the base model to improve semantic discrimination. Future research should focus on developing more effective knowledge transfer methods and refining the training framework to address these limitations and enhance the overall performance of moment retrieval models.

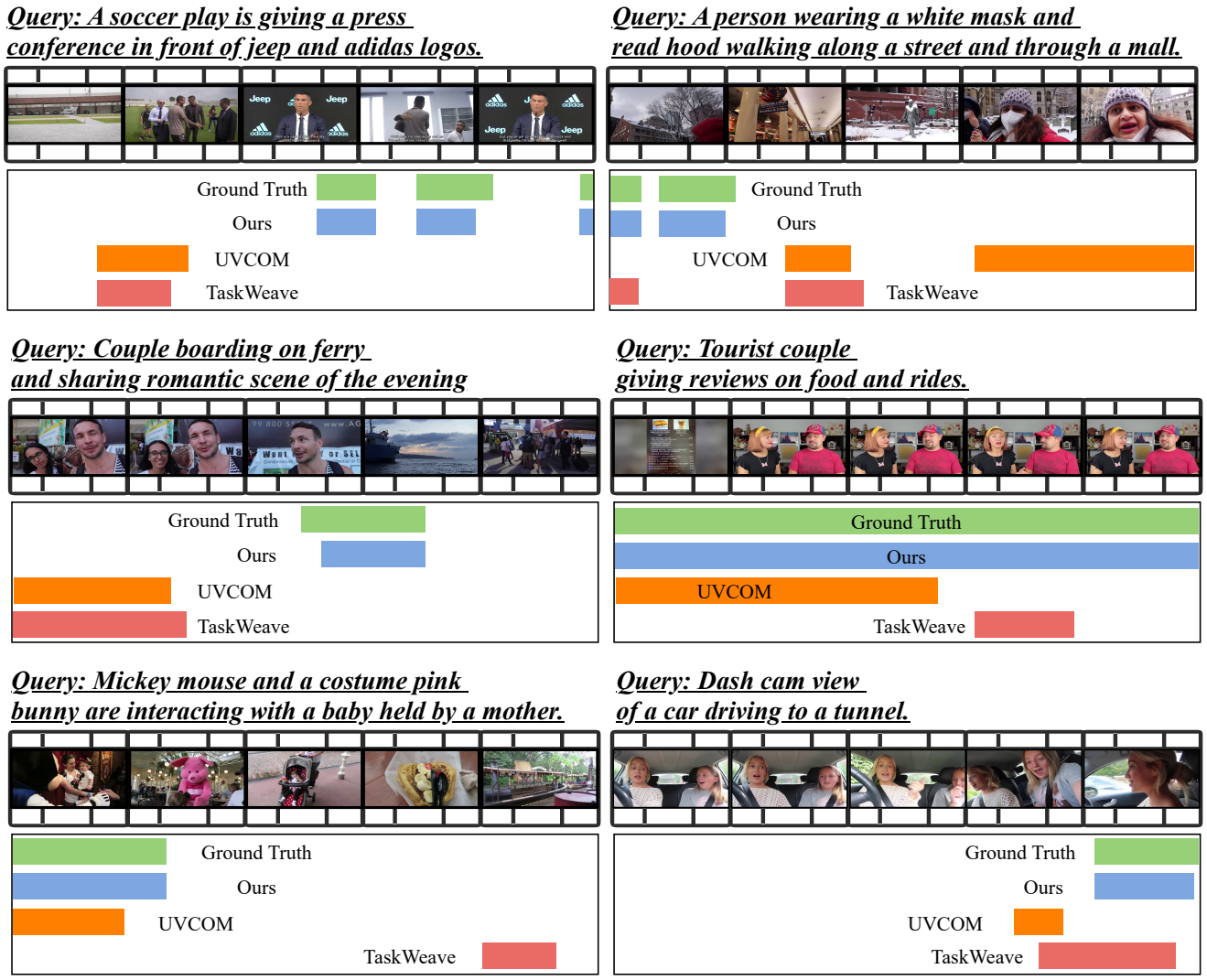


Figure 1. Additional qualitative results on QVHighlights validation set, comparing our method with UVCOM and TaskWeave.

## References


[1] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Neurips*, pages 11846–11858, 2021. 1

[2] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18709–18719, 2024. 1

[3] Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18308–18318, 2024. 1