–Appendix–
Iterative Proposal Refinement for Weakly-Supervised Video Grounding

Meng Cao[1], Fangyun Wei[2], Can Xu[3], Xiubo Geng[3], Long Chen[4],
Can Zhang[1], Yuexian Zou[1], Tao Shen[3], Daxin Jiang[3*]

[1]School of Electronic and Computer Engineering, Peking University [2]Microsoft Research Asia
[3]STCA, Microsoft [4]The Hong Kong University of Science and Technology

This appendix is organized as follows. First, we present more details about the baseline mentioned in the main paper. Then we report additional experimental results to further validate our network design. At last, some qualitative results are shown to provide more insights into our IRON.

1. Illustrations of baseline

The schematic illustration of baseline is illustrated in Figure 1. It contains four consecutive procedures, *i.e.*, feature extraction, proposal generation, confidence score generation, and grounding module. Most of the settings have been illustrated in the main paper and we briefly state them here again for completeness.

Feature Extraction. The encoded video feature is represented as $\boldsymbol{v} \in \mathbb{R}^{T \times C}$, where T is the number of sampled frames and C is the feature dimension. The query embedding is represented as $\boldsymbol{q} \in \mathbb{R}^{S \times C}$, where S denotes the total word length.

Proposal Generation. We follow [11, 12] to conduct the proposal generation by predicting upon the video-language fusion results. Firstly, the proposal generation module integrates the text feature $\boldsymbol{q}$ and the video feature $\boldsymbol{v}$ with a vanilla Transformer [8]. Then, a set of proposals $\boldsymbol{u} \in \mathbb{R}^{N \times 2}$ is predicted, where N denotes the proposal number. The corresponding proposal features $\boldsymbol{p} \in \mathbb{R}^{N \times C}$ are generated by RoI Align.

Confidence Score Generation. We simply use MLPs activated by `sigmoid` function to generate proposal-wise confidence scores $\boldsymbol{e} \in \mathbb{R}^{N \times 1}$.

Grounding Module. The baseline model is compatible with both MIL-based and reconstruction-based grounding modules. The MIL-based method learns a joint space by attracting the aligned video-query pairs while repelling the unmatched pairs. The reconstruction-based method evaluates each proposal by appraising how well it reconstructs the entire query. Refer to Sec. 3.3 and Figure 4 of the main paper for detailed descriptions.

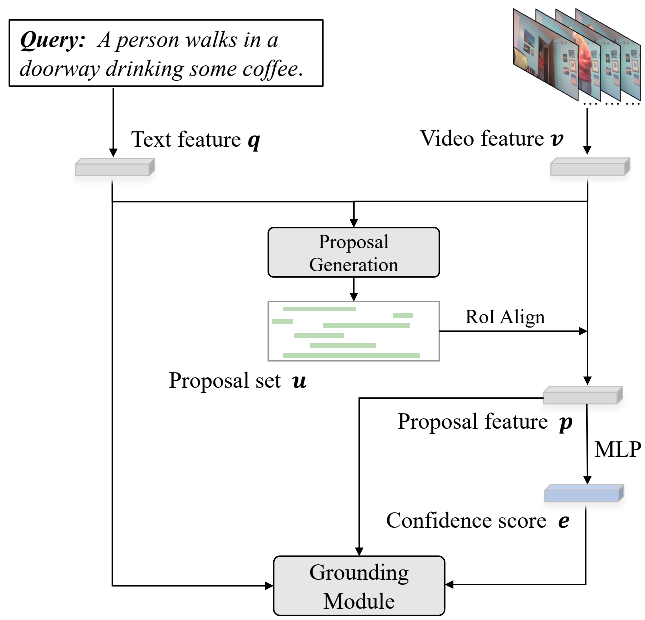

Figure 1. **An overview of baseline.** The proposal generation module firstly integrates the text feature $\boldsymbol{q}$ and the video feature $\boldsymbol{v}$ with a vanilla Transformer. Then, a set of proposals $\boldsymbol{u}$ is predicted and the corresponding proposal features $\boldsymbol{p}$ are generated. Based on this, proposal-wise confidence scores $\boldsymbol{e} \in \mathbb{R}^{N \times 1}$ are simply predicted via MLPs. Finally, the grounding module takes confidence scores $\boldsymbol{e}$, proposal feature $\boldsymbol{p}$, and text feature $\boldsymbol{q}$ as input. It can be implemented with either MIL or query reconstruction (*cf*. Figure 4 of the main paper).

2. More Experiments

Ablation on Semantic & Conceptual Score Generation. In Eq.(1) of the main paper, the semantic & conceptual scores are generated via two MLPs and then multiplied by

*Corresponding Author

Table 1. **Ablations on semantic & conceptual generation on Charades-STA dataset.** *w/* multiplication denotes the semantic & conceptual scores are multiplied by the confidence score.

Exp	*w/* multiplication	R1@0.3	@0.5	@0.7
#1	✓	**70.71**	**51.84**	**25.01**
#2	✗	68.32	50.50	24.11

the confidence score. Here we ablate to cancel the multiplication of the confidence score and modify Eq.(1) as follows.

$$\begin{aligned} \boldsymbol{e}^k &= \mathrm{Sigmoid}\left(\boldsymbol{p}\cdot\mathbf{W}_{\mathrm{e}}^k\right), \\ \boldsymbol{s}^k &= \mathrm{Sigmoid}\left(\boldsymbol{p}\cdot\mathbf{W}_{\mathrm{s}}^k\right), \\ \boldsymbol{c}^k &= \mathrm{Sigmoid}\left(\boldsymbol{p}\cdot\mathbf{W}_{\mathrm{c}}^k\right), \end{aligned} \tag{1}$$

where $\mathbf{W}_{\mathrm{s}}^k, \mathbf{W}_{\mathrm{e}}^k \in \mathbb{R}^{C\times 1}$ and $\mathbf{W}_{\mathrm{c}}^k \in \mathbb{R}^{C\times M}$ are learnable parameters in the k^{th} iteration as defined in the main paper.

We list the comparison results with (*w/*) and without (*w/o*) multiplication on Charades-STA dataset in Table 1. As expected, the variant with multiplication leads to better performance. This may be because the confidence score is the direct basis for selecting the proposal during the inference process. Therefore, directly multiplying confidence scores with the semantic & conceptual score is conducive to generating proposals with both high confidence scores and high semantic & conceptual scores.

Ablation on Language Encoder. Besides using DistilBERT [6], we also conduct experiments using Glove [5] as the language encoder. The comparison results on Charades-STA and ActivityNet Captions datasets are summarized in Table 2 and Table 3, respectively. We can draw the following conclusions: **1)** Compared to GloVe, DistilBERT is a better language feature encoder in most cases. For example, when using MIL for grounding on Charades-STA dataset, IRON with DistilBERT surpasses the GloVe counterpart by 0.32% absolute improvement on R1@0.3 (69.43% *v.s.* 69.11%). **2)** With the same Glove language feature encoder, our IRON still outperforms the previous state-of-the-art methods (*e.g.*, CPL [12] and CNM [11]). Since both CPL and CNM are reconstruction-based methods, we compare them with our reconstruction-based version. For example on R1@0.5 of Charades-STA dataset, our IRON outperforms CPL by 2.09% (51.33% *v.s.* 49.24%), demonstrating the superiority of our method.

Ablations on concept number M. As shown in Figure 2a, the performance of our IRON is not much sensitive to the concept number M, and the best performance is achieved at a medium value ($M = 30$).

Ablations on iteration number. Here we discuss the influence of the iteration number K. The results in Figure 2b show that the performance saturates at $K = 4$.

Ablations on proposal number. We conduct the ablation studies on the proposal number N in Figure 2c. As shown,

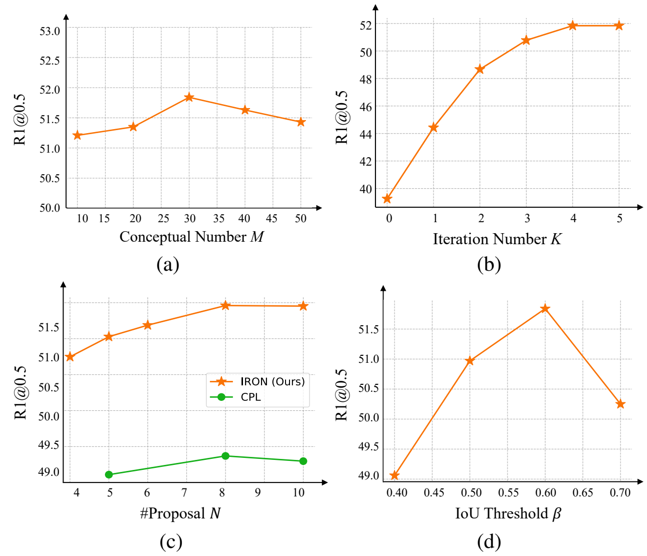

Figure 2. **Ablations** on (a) conceptual number M; (b) iteration number K; (c) proposal number N; and (d) IoU threshold β.

the performance of our IRON reaches the bottleneck when $N > 8$. For comparison, we also list the performance of CPL [12] with the number of proposals. The results show that our IRON performs better at different values of N.

Ablations on IoU similarity threshold. We ablate on the IoU similarity threshold β. In Figure 2d, we can see that setting β to 0.6 obtains the best performance. Too small β value will result in overabundant proposals being marked as positive, *i.e.*, generating false positive samples. Similarly, too large β value leads to false negative results.

3. Visualizations

Concept Set Visualizations. The used concept set of Charades-STA and ActivityNet Captions datasets are shown in Table 4 and Table 5, respectively.

Long-tailed Distribution Visualizations. We found that a potential advantage brought by semantic distillation is that it can alleviate the phenomenon of long-tailed distribution. To demonstrate this, we select the thirty most frequent verbs, and separately evaluate the performance of the query sentences containing them. In Figure 3, we list the per-action R1@0.5 values on Charades-STA [7] test set. The actions (*i.e.*, verbs) are sorted according to their frequency. As shown, IRON without semantic distillation shows a typical long-tailed distribution, where the low frequency actions have much low performance. In contrast, our IRON leads to a relatively more even distribution.

Besides, we also visualize the ground truth distributions and prediction results for model variants with and without semantic distillation loss, respectively. Specifically, we visualize four high frequency actions ("`open`", "`put`", "`take`", and "`eat`") in Figure 4 and four low frequency

Table 2. **Comparison results (%) with DistilBERT [6] and Glove [5] language encoder on Charades-STA dataset.** IRON* uses MIL for grounding and IRON follows the reconstruction strategy.

Method	Text Encoder	R1@0.3	@0.5	@0.7	R5@0.3	@0.5	@0.7
IRON* (Ours)	DistilBERT	**69.43**	**50.90**	**24.32**	97.43	**85.92**	**54.06**
IRON* (Ours)	Glove	69.11	50.17	23.94	**97.50**	85.39	53.95
IRON (Ours)	DistilBERT	**70.71**	**51.84**	**25.01**	**98.96**	**86.80**	**54.99**
IRON (Ours)	Glove	70.28	51.33	24.71	98.25	86.35	54.93
CNM [11]	Glove	60.39	35.43	15.45	-	-	-
CPL [12]	Glove	66.40	49.24	22.39	96.99	84.71	52.37

Table 3. **Comparison results (%) with DistilBERT [6] and Glove [5] language encoder on ActivityNet Captions dataset.** IRON* uses MIL for grounding and IRON follows the reconstruction strategy.

Method	Text Encoder	R1@0.1	@0.3	@0.5	R5@0.1	@0.3	@0.5
IRON* (Ours)	DistilBERT	**82.83**	**56.81**	**33.67**	**95.09**	**83.46**	**67.38**
IRON* (Ours)	Glove	82.25	56.56	33.43	94.78	83.11	67.30
IRON (Ours)	DistilBERT	**84.42**	**58.95**	**36.27**	**96.74**	**85.60**	**68.52**
IRON (Ours)	Glove	84.13	58.57	36.04	96.25	85.32	68.44
CNM [11]	Glove	78.13	55.68	33.33	-	-	-
CPL [12]	Glove	82.55	55.73	31.37	87.24	63.05	43.13

Table 4. **Concept set of Charades-STA dataset**, *i.e.*, the top-30 high frequent verbs, adjectives and nouns in the training set of Charades-STA dataset.

Rank	Word	Frequency	Rank	Word	Frequency
#1	`person`	12373	#2	`put`	1522
#3	`open`	1502	#4	`door`	1270
#5	`take`	1172	#6	`eat`	953
#7	`close`	819	#8	`sit`	776
#9	`light`	631	#10	`glass`	623
#11	`hold`	592	#12	`drink`	569
#13	`turn`	555	#14	`throw`	523
#15	`run`	507	#16	`book`	499
#17	`bag`	459	#18	`table`	456
#19	`shoe`	451	#20	`sandwich`	449
#21	`chair`	438	#22	`food`	434
#23	`start`	426	#24	`cabinet`	424
#25	`laptop`	408	#26	`box`	404
#27	`window`	398	#28	`begin`	384
#29	`cloth`	372	#30	`cup`	353

Table 5. **Concept set of ActivityNet Captions dataset**, *i.e.*, the top-30 high frequent verbs, adjectives and nouns in the training set of ActivityNet Captions dataset.

Rank	Word	Frequency	Rank	Word	Frequency
#1	`man`	9455	#2	`woman`	4108
#3	`people`	3879	#4	`camera`	3610
#5	`play`	2902	#6	`shown`	2829
#7	`stand`	2705	#8	`seen`	2334
#9	`person`	2174	#10	`continue`	2051
#11	`talk`	2046	#12	`walk`	2025
#13	`see`	1917	#14	`girl`	1915
#15	`hold`	1793	#16	`begin`	1731
#17	`ball`	1683	#18	`hand`	1673
#19	`sever`	1626	#20	`put`	1608
#21	`men`	1601	#22	`show`	1596
#23	`sit`	1535	#24	`water`	1518
#25	`jump`	1486	#26	`boy`	1478
#27	`screen`	1353	#28	`end`	1314
#29	`speak`	1281	#30	`move`	1267

actions ("`fix`", "`cook`", "`play`", and "`get`") in Figure 5. The ground truth is plotted by **green density distribution** while the prediction results are plotted by **red points**. We can observe that the proposed semantic distillation can effectively rectify the prediction results, especially for low frequency actions.

We explain this from two aspects. Firstly, the pre-trained VL models have shown great transfer potential in open-vocabulary detection [1, 3], few-shot learning [2, 10], and zero-shot learning [4, 9]. Therefore, distilling the knowledge from these pre-trained VL models can naturally benefit the long-tailed issue since it can be viewed as a *weaker* version of open-vocabulary detection. Secondly, the proposal-wise semantic distillation targets provide explicit and distinctive clues for proposal updates. This additional supervision information does not depend on the distribution of the overall dataset, thus alleviating the long-tailed performance.

References

[1] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of*

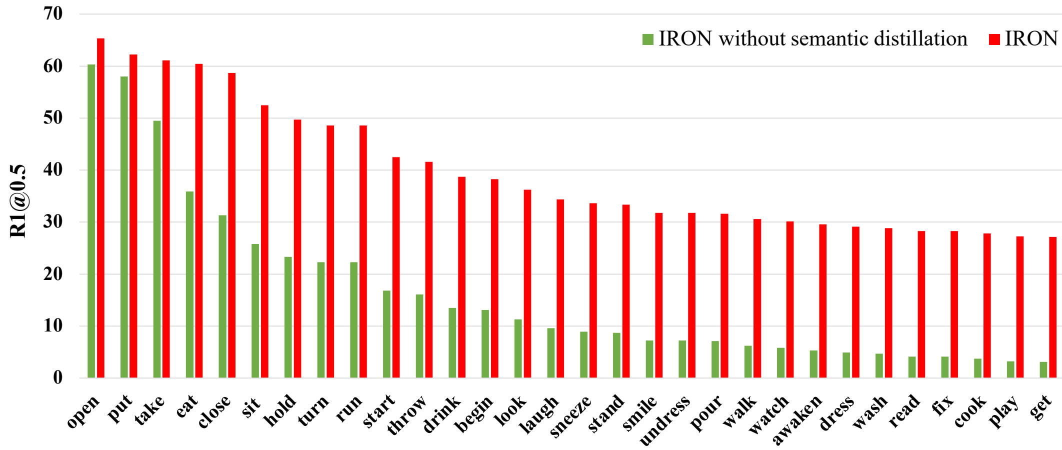

Figure 3. **The R1@0.5 performance of the top-30 high frequent actions (*i.e.*, verbs) on Charades-STA dataset.** IRON without semantic distillation shows a long-tailed distribution while our IRON alleviates this to some extent.

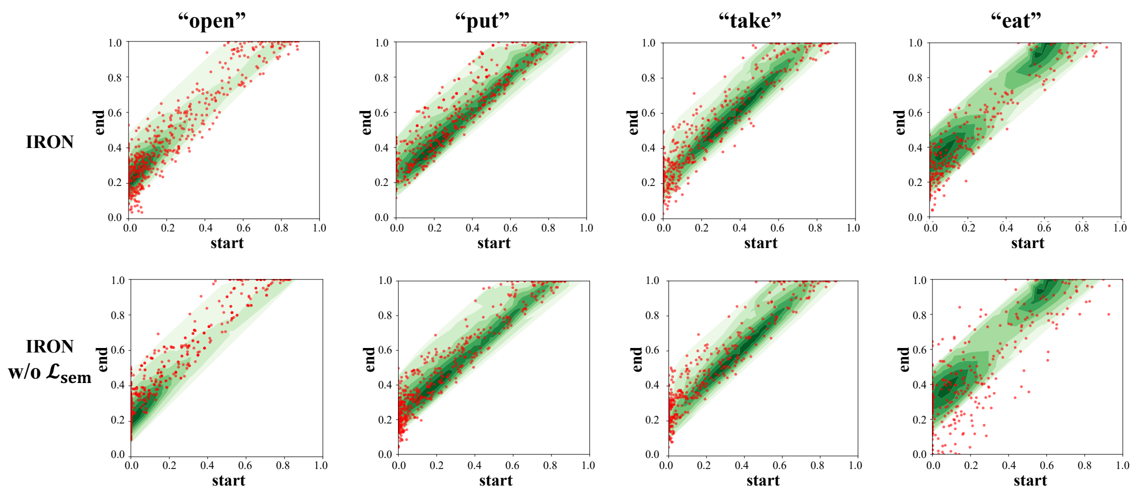

Figure 4. **Visualizations of the ground truth distribution and prediction results for high frequency actions** including "`open`", "`put`", "`take`", and "`eat`". We visualize the results for IRON and IRON without semantic distillation loss $\mathcal{L}_{\text{sem}}$, respectively.

the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3

[2] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, 2021. 3

[3] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[4] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 2, 3

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 2, 3

[7] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 2

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[9] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceed-*

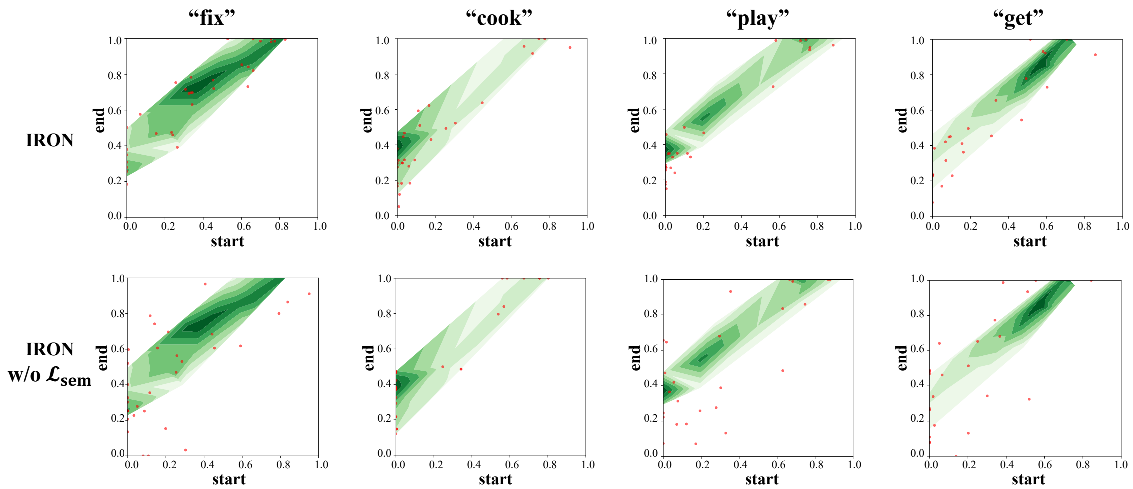

Figure 5. **Visualizations of the ground truth distribution and prediction results for low frequency actions** including “fix”, “cook”, “play”, and “get”. We visualize the results for IRON and IRON without semantic distillation loss $\mathcal{L}_{\text{sem}}$, respectively.

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3313–3322, 2022. 3

[10] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021. 3

[11] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, page 3, 2022. 1, 2, 3

[12] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15555–15564, 2022. 1, 2, 3