TableR 1: Performance comparisons about the proposal selection approaches and negative sample mining methods

Method	Charades-STA		ActivityCaption		RefCOCO		RefCOCO+		RefCOCOg
	R@1 IoU=0.5	R@1 IoU=0.7	R@1 IoU=0.3	R@1 IoU=0.5	TestA	TestB	TestA	TestB	Val
score-based selection	32.65	15.24	49.30	30.77	36.41	32.27	36.48	32.60	34.54
hard negative mining	31.19	13.38	46.92	29.41	34.74	28.77	34.69	30.72	32.66
direct proposal masking	31.54	14.15	47.29	29.33	35.59	30.22	35.60	31.95	33.16
full	**33.21**	**15.68**	**50.12**	**31.07**	**37.64**	**32.59**	**36.91**	**33.56**	**34.92**

1 To Reviewer 1

R1.1 The idea of counterfactual contrastive learning is similar to adversarial erasing [a,b,c]. In fact, the core idea of CCL is completely different from the adversarial erasing works [a,b,c]. The only similarity is the process of detecting the crucial parts by the CAM method [32]. After it, the adversarial erasing approaches hide the crucial parts to make the network further focus on other relevant parts. However, our CCL constructs the counterfactual transformations based on the detected proposals for contrastive training. So the connection between CCL and adversarial erasing is very weak. The key contributions of CCL are the usage of contrastive learning for WSVLG, the transformation strategies from three levels and the design of score-based and distribution-based losses, rather than the conventional gradient-based CAM for proposal selection. We will cite these works and illustrate the differences in revision.

R1.2 The "relation module" includes relational modeling and score inference. Thanks for your comments and we will separate the score inference from the relation module as an independent module.

R1.3 What are its advantages of gradient-based selection compared to the direct selection of the proposals with higher scores as the critical proposals. As shown in TableR 1, the model with score-based selection achieves a worse performance than the full model with gradient-based selection. By qualitative observation, we find the gradient-based method can select more disperse proposals but the score-based method often chooses proposals with large overlaps.

R1.4 Verifying the performance of the algorithm w.r.t simple framework seems unconvincing. Actually, as shown in Tables 1 and 4, our basic VGN and IGN achieve the performance close to SOTA, so they are not simple frameworks and our CCL further improves their accuracy. Moreover, we integrate CCL into the existing SOTA approach CTF [7] to prove its effectiveness, shown in TableR 2.

TableR 2: The CCL paradigm on other networks (R@1).

Method	Charades-STA		ActivityCaption	
	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5
CTF	27.30	12.90	44.30	23.60
CTF+CCL	**32.11**	**14.57**	**48.61**	**28.73**

R1.5 The authors should cite and compare with [d]. Thanks for your suggestion and we will add the missing related works in revision. But [d] is a supervised approach and we mainly compare with weakly-supervised methods.

2 To Reviewer 2

R2.1 Influences of memory bank size B and the memory update strategy are expected to be discussed. We validate the effect of memory size B and update hyper-parameter α. When α is set to 0.9 and B is set to [25, 50, 100, 125, 150], R@1 IoU=0.5 on Charades-STA is [32.86, 32.98, 33.21, 33.13, 33.01]. And when B is set to 100 and α is set to [0.6,0.7,0.8,0.9,1.0], R@1 IoU=0.5 on Charades-STA is [32.89, 33.05, 33.15, 33.21, 33.04]. So our CCL paradigm is relatively robust to B and α. More experiment data will be provided in the revision.

3 To Reviewer 3

R3.1 How can authors guarantee the positive results have higher alignment scores than the negative results. The DCT destroys crucial proposals to generate negative results and the RCT damages inessential ones to generate positive results. Because alignment scores rely on the crucial proposals, it is natural that positive results have higher scores.

R3.2 Could you provide some visualization or analysis about the distribution of original, counterfactual negative and positive results. By qualitative observation, we can find the distribution of positive results is close to the original results, but the negative results have an inconsistent distribution with the positive and original ones. Due to the page limitation, we will add the visualization of these distributions in the revision.

4 To Reviewer 4

R4.1 The feature and proposal details of the baseline VGN. We have presented the details of VGN and IGN in Section 1 of the supplementary material, including the model architecture and parameter settings.

R4.2 Stronger baselines from negative sample mining perspective. As shown in TableR 1, we further compare CCL with two baselines, where "hard negative mining" selects those unmatched vision-language pairs with high alignment scores during model inference as the negative samples for MIL-based training, and "direct proposal masking" synthesizes negative samples by directly masking important proposals. Our CCL still outperforms the two baselines.

R4.3 The image grounding part is not evaluated on the ReferitGame and Flickr30K Entities dataset. Due to the page limitation, we will add the experiments on ReferitGame and Flickr30k in the revision.