All reviewers. Thank you for the constructive comments and suggestions. Our work (COOT) aims to exploit long-range temporal context and leverage hierarchy of semantics to learn joint video-text embeddings. COOT consists of three new components: an Attention-aware Feature aggregation module (AF), a Contextual Transformer (CoT) and Cross-Modal Cycle consistency loss (CMC). The proposed method achieves state-of-the-art results on retrieval tasks (Table 2 and Table 3). We obtained larger improvements on paragraph-video retrieval compared to sentence-clip retrieval (Table 3). This indicates success of our model in capturing long-range semantics, which is the main theme of our paper. We believe that retrieval task is sufficient for measuring the alignment of semantic representations, but we agree that other tasks like video captioning may further support our claims. We report the results of video captioning in TabA 1-Left. Note that due to limited time, our method uses a simple caption generation method based on RNNs while the result of VideoBERT uses more sophisticated transformer based method. Hence, we expect to get even better improvement over VideoBERT if we use the same captioning method.

Video Captioning

Method	CIDEr	BLEU-3	BLEU-3	ROUGE-L
VideoBERT+ Transformer	0.49	6.80	4.04	27.50
COOT + RNN	**0.67**	**6.91**	**4.43**	**27.80**

	Paragraph $\Longrightarrow$ Video		Video $\Longrightarrow$ Paragraph	
Method	R@1	R@5	R@1	R@5
FSE	11.5	31.0	11.0	30.6
HSE	32.9	62.7	32.6	63.0
COOT	**48.5**	**78.9**	**48.9**	**79.5**

TabA 1: Video captioning on Youcook2 dataset (Left) and Retrieval on AcitvityNet-captions-val2 (Right).

Reviewer1. **Transformers and attention modules are frequently used.** We do not claim novelty for using transformers or attention for video-text representation learning. The novelty is in the way we handle long-range interactions between semantics. Our model improves the semantic representation by learning the interactions from long-range video-text data. Particularly the proposed contextual transformer and the cross-modal cycle consistency loss contribute to this, as shown by Tab.1 in the paper. **AF vs [CLS] and discussion**. The [CLS] token is one way of pooling the information of a sequence. Alternatively, one could average representations of all tokens with the same weight (Kim et al. [25]), which will integrate all the semantics equally. Table 6 of Reimers et al.[**A1**] shows that average pooling works better than the [CLS] token. We propose AF (Attention-aware Feature aggregation) to learn attention weights that emphasize the most relevant content of the input. In Tab.3 of the supplementary material we have a comparison to [CLS] and average pooling. It shows the benefit of AF. There have been ideas similar to our AF module in text-only or video-only domains, yet we are the first to apply this idea to the joint video-text domain, which is an additional but admittedly not groundbreaking contribution. We will discuss this a bit more in the final manuscript. **i) What means (1,1)**? Only negative pairs are considered, i.e. the D(x,y) parts of positive pairs in Eq.1 become 0. **ii) Fig2. Global context.** It's simply the concatenation of all local captions of the video, i.e., the paragraph.**iii) Line 212**. Thanks. We will follow your suggestion. **iv) HSE in Tab1 and Tab2.** HSE in Tab2 is copied the HSE original paper. HSE in Tab1 is our reproduced results of the same architecture to have fair ablations. **v) Replacing CoT.** We use a one-layer transformer and then average pooling, but we do not use the cross attention layer.

Reviewer2. **Why do modules strengthen each other?** We observed that networks with higher performance benefit more from our CMC loss. This is probably because with more meaningful features the cycle consistency is more informative. We also believe there is a mutual regularization effect. **Hyper-parameter optimization.** We designed the architecture from scratch. Thus, it is necessary to set proper hyper-parameter ranges. For instance, the learning rate range of [0.0005-0.001] results in accuracy range of [59.4%-61.6%] with Adam optimizer on Activitynet retrieval task. We will provide more results on effect of hyper-parameters range in the supplementary material.
Results on ActivityNet-val2. We do use cross-validation by holding out a part of training set for tuning the hyper-paramters. Moreover, we report results on ActivityNet-val2 in TabA 1-Right. Ours outperforms all baselines including HSE. Note that similar to HSE, our results on val1 is better than val2 which indicates the difficulty of this split. **Prior work.** Thanks. We will cite it.

Reviewer3. **1) AF module.** The purpose of the co-attentional transformer and the tangled transformer is different from our AF module. We designed AF for pooling, while VilBERT and ActBERT use the [CLS] token. We will add more discussions to the final paper (also see response to R1 in line 16-21). **2) Value of λ.** For Activitynet we used $1\,\mathrm{e}{-2}$ and for Youcook2 $1\,\mathrm{e}{-3}$. **3) Difference to [21].** Our architecture is based on transformers while HSE is based on GRUs. Additionally, we have new components (AF, CoT and CMC) in our design, which significantly improve over the base architecture (~ 9 points). **4) Improvement of COOT without AF, CMC, CoT over HSE.** We believe, this improvement comes from using transformers rather than GRUs. This baseline shows that the HSE idea of using a hierarchy is well compatible with transformers.

Reviewer4. **Attention-FA and [CLS] output.** Conventional attention does not involve any pooling. We either need average/max pooling or define something like a [CLS] token. Please see the response to R1 in line 16 -23. In Table 4, CLS method means adding a learnable CLS token to the inputs of all T-Transformers and taking its output value as the pooled value. We agree that it would be interesting to see how CLS works with more layers. However, it will increase the complexity of the model. Also, Reimers et al.[**A1**] showed that average pooling performed better than [CLS]. **Alternatives to contextual transformer?** This is good idea for future work. The second last row of Table 1 can be seen as an alternative where we replace CoT with a one-layer transformer and average pooling. **Other video-language tasks.** We report the results of video captioning in TabA 1-Left. **Single transformer layer?** We use one layer for T-transformer and two layers for CoT. We will clarify this in the paper.

[A1]- Nils Reimers and Iryna Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP2019