DeCo : Decomposition and Reconstruction for Compositional Temporal Grounding via Coarse-to-Fine Contrastive Ranking Supplymentary Material

1. Implementation details

We use PyTorch [5] for our implementation. For a fair comparison, we follow [4] to use C3D [6] features for the ActivityNet-CG dataset and I3D [1] features for the Charades-CG dataset. In detail, the features are extracted by first downsampling each video by a factor of 8, and we follow [7] by setting the maximum video feature length as 200. We use pre-trained GloVe word2vec for extracting the language features.

The output video feature dimensions from C3D/I3D video encoder are 500/1024, respectively. The output language feature dimension from text encoder is 300. The extracted video features and language features will further pass through different MLP layers, generating features with dimension $D = 256$. For model architecture, text transformer, vision transformer, cross-transformer for temporal boundary prediction and mask-conditioned transformer each with 3 layers and 4 attention heads. And the transformer decoder used in sentence recomposition is of 2 layers and 4 attention heads. We set $M = 2$ and $N = 4$ for both Charades-CG dataset and ActivityNet-CG dataset.

In all experiments, we use the Adam optimizer [2] with an initial learning rate of 4e-4 to train the model for 30 epochs. For loss balancing we empirically set the factors to $\lambda_{reg} = 10, \lambda_{rec} = 1$, and $\lambda_{rank} = 1$. In both datasets, we set $m_0 = 0.05, m_1 = 0.05, m_2 = 0.15, m3 = 0.25$.

2. Details of fixed hand-craft prompt template

In the main submission, we choose to use learnable context prompt in our method instead of hand-craft prompt template. We made comparison between using learnable context prompt and the hand-craft prompt in Section 4.4.2. Below shows the details of the hand-craft prompt templates we used in the comparison, where we use 3 prompt templates for each composition pair:

(Verb, Noun) prompts:

- "The person is {*Verb*} {*Noun*}."
- "The person is performing the action of {*Verb*} on {*Noun*}."
- "The action is {*Verb*} {*Noun*}."

(Adjective, Noun) prompts:

- "The video contains {*Adj*} {*Noun*}."
- "There is a/an {*Adj*} {*Noun*} in the video."
- "{*Adj*} {*Noun*} is now showing in the video."

(Preposition, Noun) prompts:

- "The person is performing an action {*Prep*} {*Noun*}."
- "The action is being played {*Prep*} {*Noun*}."
- "The person is doing something {*Prep*} {*Noun*}."

(Noun, Noun) prompts:

- "The video contains {*Noun*} {*Noun*}."
- "In the video, the {*Noun*} is for {*Noun*}.
- "{*Noun*} {*Noun*} appears in the video."

(Verb, Adverb) prompts:

- "The action of {*Verb*} is done {*Adv*} in the video"
- "The person is performing {*Verb*} {*Adv*}."
- "The video contains {*Verb*} {*Adv*}."

3. Object Detector as Auxiliary Knowledge

In this section, we compare our method with VISA [3] which takes object detector and action detector as auxiliary knowledge, and analyze the reason why we do not use the auxiliary knowledge in detail.

Why VISA [3] can benefit from auxiliary knowledge For compositional temporal grounding, the model needs to have 1) ***decompositional*** ability to understand individual components of seen compositions during training, and 2) ***recompositional*** ability to recompose atomic-level components to understand novel compositions.

Based on these requirements, we assume two reasons can explain *why VISA can benefit from object detection*:

	Verb-Noun	Prep-Noun	Adj-Noun	Noun-Noun
Detector-supplemented	76.7%	78.4%	38.8%	67.6%
Ours-supplemented	4.6%	0%	0%	5.9%

Table A1. Overlap ratio between the supplemented compositions during training and the *Novel-Composition* split, categorized by different composition types on the Charades-CG dataset.

	NC-R1@0.5	NC-R1@0.7	NW-R1@0.5	NW-R1@0.7
VISA	45.41	**22.71**	42.35	20.88
DeCo	**47.39**	21.06	**45.61**	**23.31**

Table A2. Results on the *Novel-Composition* split (NC) and *Novel-Word* split (NW) of the Charades-CG dataset.

firstly, the use of object detectors directly provides post-decomposition information, which reduces the reliance on decompositional power. Secondly, object detectors can provide privileged semantic information towards novel compositions during training (e.g. verbs from query sentences and nouns from detections), leading to stronger recompositional ability.

Why we do not use auxiliary knowledge In detail, VISA takes the GloVe features of the top-5 detections as additional input alongside the query sentence. Purely training with compositions from the *Training* split queries ensures a clear separation from unseen compositions in the *Novel-Composition* split, whereas training with detector-supplemented compositions does overlap significantly with the *Novel-Composition* split. To this end, we analyze the overlap ratio between the detector-supplemented compositions and the ones in the *Novel-Composition* split in Table A1. Evidently, this large overlap between train and test compositions increases the chance for detector-supplemented models to learn compositions in the *Novel-Composition* split which ***should not be seen during training***. Thus, while using detectors can increase the performance especially on the *Novel-Composition* split, it ***sidesteps the actual task challenge*** of compositional temporal grounding of unseen concepts.

As a result, to fully validate the ability of compositional temporal grounding, we choose to experiment in a regime where no knowledge from detectors is leveraged. As shown in Table A2, even without the help of semantic detection, we achieve almost comparable results with detector-supplemented VISA. Furthermore, we achieve better performance on the *Novel-Word* split. This suggests that our model has stronger decomposition ability, as it can extract word-wise knowledge from known words without being affected by the existence of novel words.

4. More Qualitative Results on ActivityNet-CG dataset

In the main submission, we show the qualitative examples of grounding results by our model with and without the proposed contrastive ranking loss in Section 4.5. Here we show additional qualitative results on the ActivityNet-CG dataset. From Figure A1, similarly with the main submission, we can also see the positive improvement brought by the contrastive ranking loss.

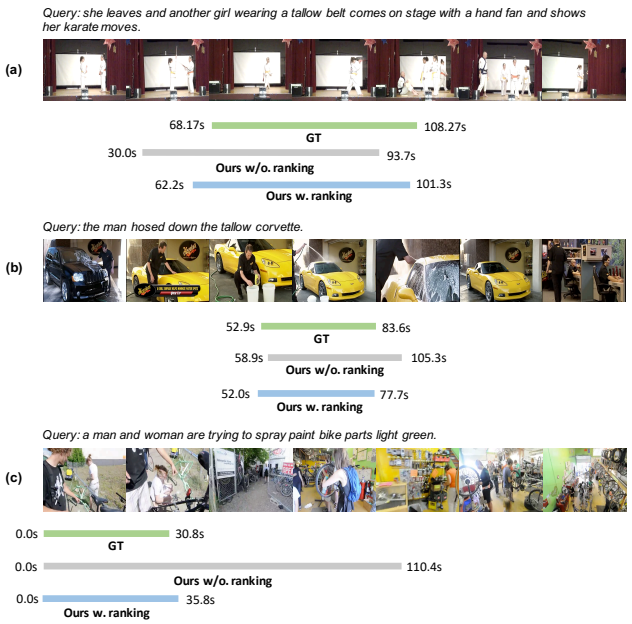

Figure A1. Qualitative examples of the ground truth (GT), Ours (without contrastive ranking), and Ours (with contrastive ranking). Examples are from the ActivityNet-CG dataset.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3032–3041, 2022.

[4] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020.

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[6] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[7] 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.