Exploring the Effect of Primitives for Compositional Generalization in Vision-and-Language - Supplementary Material

Chuanhao Li[1], Zhen Li[1], Chenchen Jing[3]*, Yunde Jia[2,1], Yuwei Wu[2,1]*
[1]Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing,
Shenzhen MSU-BIT University, China
[3]School of Computer Science, Zhejiang University, Hangzhou, China
{lichuanhao,li.zhen,jiayunde,wuyuwei}@bit.edu.cn
jingchenchen@zju.edu.cn

1. Overview

In this document, we provide supplementary experimental results including:
(1) experimental results compared to state-of-the-art temporal video grounding (TVG) methods on the ActivityNet Captions [8] and ActivityNet-CG [9] datasets;
(2) parameter analysis of the quantitative effect of words with different part-of-speech tags (α, β, and γ);
(3) more qualitative examples for TVG and visual question answering (VQA).

2. Experiments

We use 2D-TAN [22] and MS-2D-TAN [21] as baseline methods, and incorporate them into our framework, which are dubbed as 2D-TAN+Ours and MS-2D-TAN+Ours, respectively.

2.1. Experimental Results on ActivityNet Captions and ActivityNet-CG

Datasets. The ActivityNet Captions [8] dataset is a large-scale dataset with 19,209 videos taken from the real world. There are a train split, a validation split and a test split, which contain 37,421, 17,505 and 17,031 video-query pairs, respectively. The ActivityNet-CG [9] dataset is developed from ActivityNet Captions by re-splitting its samples. ActivityNet-CG provides four splits including: a train split with 36,724 video-query pairs for training, a Novel-Composition test split with 12,028 video-query pairs for testing compositional capability, a Test-Trivial test split with 3,944 video-query pairs for testing the generalization capability of seen words, and a Novel-Word test split with 15,712 video-query pairs for testing the generalization capability of unseen words.

Results. The results compared to state-of-the-art methods on ActivityNet-CG [9] are shown in Tab. 1. We can observe that our framework consistently improves 2D-TAN and MS-2D-TAN on all three test splits, and the performance gains are remarkable on the Novel-Composition test split (*e.g.*, 1.51% and 0.94% absolute performance gains in R1@0.5 for 2D-TAN and MS-2D-TAN, respectively). Compared to VISA [9], MS-2D-TAN+Ours achieves competitive performance (*e.g.*, 30.80% vs. 31.51% in R1@0.5) on the Novel-Composition test split. Our framework is compatible with VISA, and can further improve its performance by incorporating it into the framework.

The results on ActivityNet Captions [8] are listed in Tab. 2. Using our framework, both 2D-TAN and MS-2D-TAN achieve better performance with different improvements in different metrics. In addition, MS-2D-TAN+Ours achieves comparable performance compared to state-of-the-art methods (*e.g.*, 29.99% vs. MMN's 29.26% in R1@0.7, 79.36% vs. MMN's 79.50% in R5@0.5).

2.2. Parameter Analysis

We provide parameter analysis of the quantitative effect of words with different part-of-speech tags (α, β, and γ) on the Charades-STA [5] and Charades-CG [9] datasets. α, β, and γ denote the quantitative effect for nouns/verbs, adjectives/adverbs, and other words, respectively. The results of MS-2D-TAN+Ours with different setting of α, β and γ are listed in Tab. 3. We observe from the table that: (1) The performance of MS-2D-TAN+Ours fluctuates significantly with β. (2) MS-2D-TAN+Ours achieves the best overall performance on the three test splits under the setting $\alpha = 1$, $\beta = 0.6$, and $\gamma = 0$.

*Corresponding author: Chenchen Jing and Yuwei Wu

Table 1. Performance (%) of the state-of-the-art methods on the ActivityNet-CG [9] dataset. The best scores are bold and the second-best scores are underlined.

Type	Method	*Test-Trivial*			*Novel-Composition*			*Novel-Word*		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
Weakly-supervised	WSSL [4]	11.03	4.14	15.07	2.89	0.76	7.65	3.09	1.13	7.10
RL-based	TSP-PRL [16]	34.27	18.80	37.05	14.74	1.43	12.61	18.05	3.15	14.34
Proposal-free	VSLNet [20]	39.27	23.12	42.51	20.21	9.18	29.07	21.68	9.94	29.58
	LGI [12]	43.56	23.29	41.37	23.21	9.02	27.86	23.10	9.03	26.95
	VISA*[9]	47.13	29.64	44.02	**31.51**	**16.73**	**35.85**	<u>30.14</u>	**15.90**	**35.13**
Proposal-based	TMN [10]	16.82	7.01	17.13	8.74	4.39	10.08	9.93	5.12	11.38
	2D-TAN [22]	44.50	26.03	42.12	22.80	9.95	28.49	23.86	10.37	28.88
	2D-TAN*[22]	43.85	26.04	42.44	25.67	11.76	29.77	24.85	10.82	28.61
	2D-TAN + Ours	46.58	29.65	45.60	27.18	12.60	30.98	26.58	12.55	30.09
	MS-2D-TAN*[21]	<u>48.80</u>	<u>31.52</u>	<u>46.58</u>	29.86	14.40	31.80	28.90	13.83	31.01
	MS-2D-TAN + Ours	**49.63**	**31.73**	**47.22**	<u>30.80</u>	<u>15.39</u>	<u>33.18</u>	**30.15**	<u>14.97</u>	<u>32.14</u>

* indicates the results from our reimplementation using official released codes.
* indicates that the method can be incorporated into our framework for further improvements.

Table 2. Performance (%) of the state-of-the-art methods on the ActivityNet Captions [8] dataset. The best scores are bold and the second-best scores are underlined.

Type	Method	Feature	R1@0.5	R1@0.7	R5@0.5	R5@0.7	mIoU
RL-based	RWM [7]	C3D	36.90	-	-	-	-
	TSP-PRL [16]	C3D	38.82	-	-	-	-
	MABAN [14]	C3D	42.42	24.34	-	-	-
Proposal-free	LGI [12]	C3D	41.51	23.07	-	-	41.13
	IVG [13]	C3D	43.84	27.10	-	-	44.21
	DeNet†[23]	C3D	43.79	-	74.13	-	-
	DCM [19]	C3D	44.90	27.70	-	-	43.30
	HiSA [18]	C3D	45.36	27.68	-	-	<u>45.45</u>
	CBLN†[11]	C3D	<u>48.12</u>	27.60	79.32	63.41	-
Proposal-based	BPNet [17]	C3D	42.07	24.69	-	-	42.11
	FVMR [6]	C3D	45.00	26.85	77.42	61.04	-
	SSCS [3]	C3D	46.67	27.56	78.37	<u>63.78</u>	-
	MMN [15]	C3D	**48.59**	29.26	**79.50**	**64.76**	-
	2D-TAN [22]	C3D	44.05	27.38	76.65	62.26	-
	2D-TAN*[22]	C3D	44.72	26.89	76.38	61.18	43.31
	2D-TAN + Ours	C3D	45.46	28.01	77.01	62.11	43.62
	MS-2D-TAN [21]	I3D	45.50	28.28	79.36	61.70	-
	MS-2D-TAN [21]	C3D	46.16	29.21	78.80	60.85	-
	MS-2D-TAN*[21]	C3D	46.91	<u>29.79</u>	79.04	59.43	45.00
	MS-2D-TAN + Ours	C3D	47.57	**29.99**	<u>79.36</u>	62.19	**46.27**

† indicates that the method is a special proposal-free method, which can provide multiple predictions without using proposals.
* indicates the results from our reimplementation using official released codes.

2.3. Qualitative Examples

Temporal Video Grounding. Fig. 1 depicts several qualitative examples that show the effectiveness of our framework for temporal video grounding. The examples come from three test splits of Charades-CG [9] mentioned above, and we visualize four qualitative examples for each test split. These qualitative examples demonstrate that our

Table 3. Parameter analysis of the quantitative effect of words with different part-of-speech tags. Performance (%) on the Charades-CG [9] dataset of our framework with different α, β and γ settings on MS-2D-TAN. The best scores are bold and the second-best scores are underlined.

α	β	γ	*Test-Trivial*			*Novel-Composition*			*Novel-Word*		
			R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
1.0	1.0	0.0	58.33	36.79	50.76	42.30	22.60	38.18	45.32	25.90	40.41
1.0	0.8	0.0	58.17	37.53	50.66	43.06	22.31	37.79	46.04	26.91	41.37
1.0	0.6	0.0	58.14	37.98	50.58	**46.54**	**25.10**	**40.00**	**50.36**	**28.78**	**43.15**
1.0	0.4	0.0	59.27	**38.44**	**51.27**	44.68	23.65	39.56	49.07	26.76	41.90
1.0	0.2	0.0	**59.30**	38.02	51.15	44.28	23.68	39.42	48.20	26.04	41.38

framework is effective to improve the generalization capability of TVG methods from three aspects: compositional generalization, seen words generalization, and unseen words generalization.

Visual Question Answering. We visualize several qualitative examples from the test split of the CLOSURE [1] dataset in Fig. 2. The test samples in CLOSURE are divided into six categories including material, color, size, shape, yes/no (y/n), and number (num.), according to the question type of the sample. For each category, we provide two qualitative examples. For the first shown example in Fig. 2 (b), GLT makes a wrong prediction "red" even though the image contains no red objects, which suggests that GLT neglects the image when making predictions. By using our framework, GLT+Ours correctly answers "gray" for the example, which proves the framework is effective to establish the relationship between primitives and ground-truth, thereby improving the compositional generalization capability of GLT.

References

[1] Dzmitry Bahdanau, Harm de Vries, Timothy J O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. Closure: Assessing systematic generalization of clevr models. *arXiv preprint arXiv:1912.05783*, 2019. 3, 5

[2] Ben Bogin, Sanjay Subramanian, Matt Gardner, and Jonathan Berant. Latent compositional representations improve systematic generalization in grounded question answering. *Transactions of the Association for Computational Linguistics*, 9:195–210, 2021. 5

[3] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11573–11582, 2021. 2

[4] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3062–3072, 2018. 2

[5] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5267–5275, 2017. 1

[6] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1523–1532, 2021. 2

[7] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8393–8400, 2019. 2

[8] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 1, 2

[9] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3032–3041, 2022. 1, 2, 3, 4

[10] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–568, 2018. 2

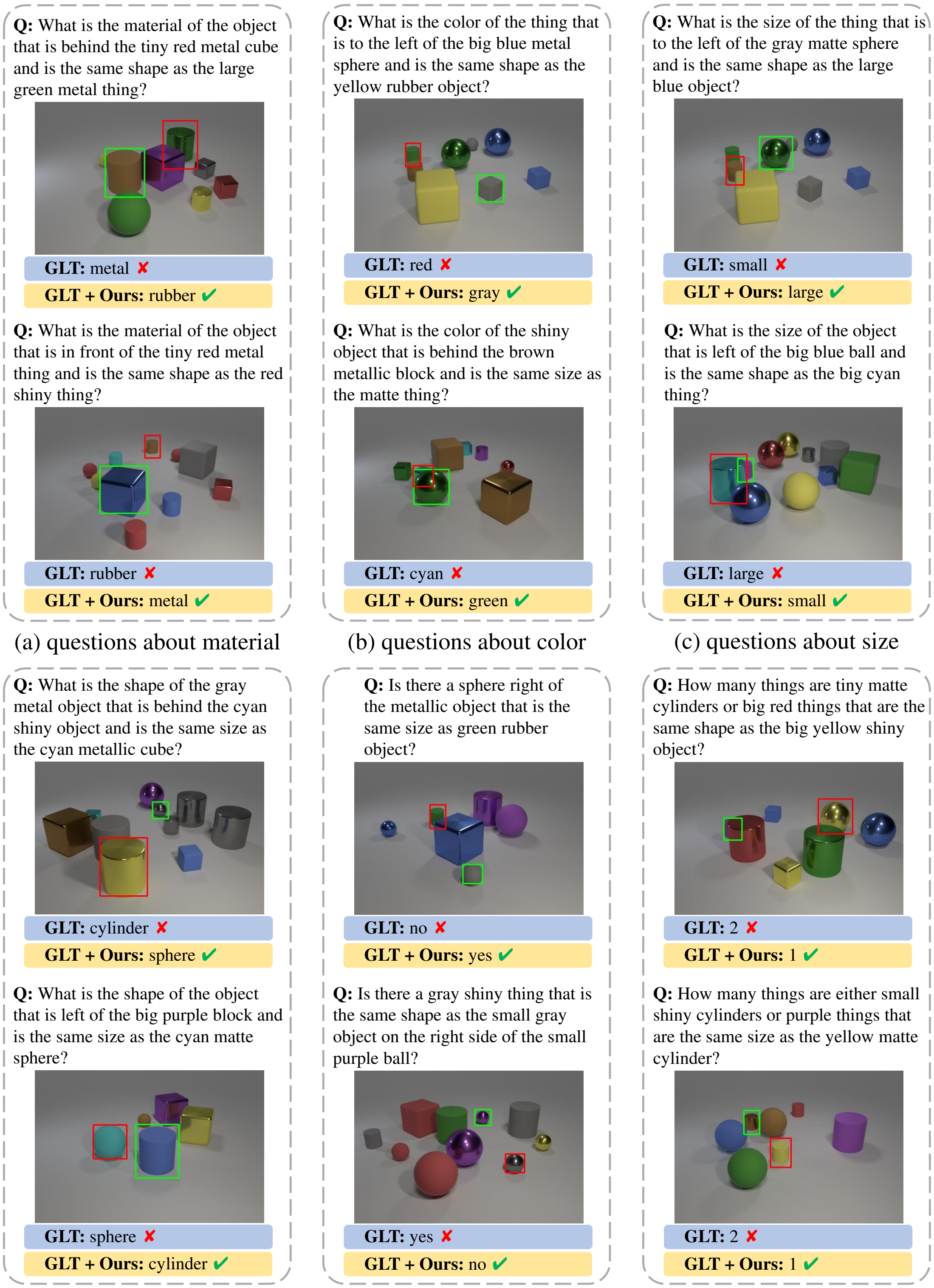
[11] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11235–11244, 2021. 2

[12] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10810–10819, 2020. 2

[13] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2765–2775, 2021. 2

[14] Xiaoyang Sun, Hanli Wang, and Bin He. Maban: Multi-agent boundary-aware network for natural language moment

Figure 1. Qualitative comparisons between MS-2D-TAN+Ours and MS-2D-TAN [21] on samples from different test splits of Charades-CG [9]. The words in red font in (b) and (c) denote novel compositions and novel words, respectively.



(a) questions about material (b) questions about color (c) questions about size

(d) questions about shape (e) questions about y/n (f) questions about num.

Figure 2. Qualitative comparisons between GLT+Ours and GLT [2] on questions with novel compositions from CLOSURE [1].The green and red boxes indicate the image regions with the high- est attention weights of GLT+Ours and GLT for object referring, respectively.

retrieval. *IEEE Transactions on Image Processing*, 30:5589–5599, 2021. 2

[15] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2613–2623, 2022. 2

[16] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12386–12393, 2020. 2

[17] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2986–2994, 2021. 2

[18] Zhe Xu, Da Chen, Kun Wei, Cheng Deng, and Hui Xue. Hisa: Hierarchically semantic associating for video temporal grounding. *IEEE Transactions on Image Processing*, 31:5178–5188, 2022. 2

[19] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the International Conference on Research on Development in Information Retrieval (SIGIR)*, pages 1–10, 2021. 2

[20] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6543–6554, 2020. 2

[21] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9073–9087, 2021. 1, 2, 4

[22] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12870–12877, 2020. 1, 2

[23] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8454, 2021. 2