# −Supplementary Material−
# LocVTP: Video-Text Pre-training for Temporal Localization

Meng Cao[1*], Tianyu Yang[2‡], Junwu Weng[2], Can Zhang[1], Jue Wang[2], and Yuexian Zou[1,3†]

[1] School of Electronic and Computer Engineering, Peking University
[2] Tencent AI Lab
[3] Peng Cheng Laboratory

The supplementary document is organized as follows. In Section 1, we provide the transfer results on temporal action localization; In Section 2, we present the performance of LocVTP on more video retrieval datasets; In Section 3, we evaluate our pre-trained features on more temporal grounding methods; In Section 4, more visualization results are given.

## 1 Transfer Results on Temporal Action Localization

| Models | Modal | mAP@0.5 | @0.75 | @0.95 | AVG |
|---|---|---|---|---|---|
| BSP [35] | R | 50.9 | 35.6 | 8.0 | 34.8 |
| LoFi-E2E [36] | R | 50.4 | 35.4 | 8.9 | 34.4 |
| TSP [1] | R | 51.3 | 37.1 | 9.3 | 35.8 |
| CoCLR [13] | R+F | 47.9 | 32.2 | 7.3 | 31.9 |
| XDC [2] | R+A | 48.4 | 32.6 | 7.6 | 32.3 |
| VideoMoCo [22] | R | 47.8 | 32.1 | 7.0 | 31.7 |
| RSPNet [9] | R | 47.1 | 31.2 | 7.1 | 30.9 |
| AoT [33] | R | 44.1 | 28.9 | 5.9 | 28.8 |
| SpeedNet [6] | R | 44.5 | 29.5 | 6.1 | 29.4 |
| VideoBERT [29] | R+L | 38.5 | 16.7 | 2.6 | 19.4 |
| UniVL [19] | R+L | 40.3 | 17.4 | 4.4 | 21.2 |
| SupportSet* [23] | R+L | 39.3 | 18.3 | 1.4 | 20.0 |
| LocVTP(Ours) | R+L | **51.6** | **38.4** | **9.8** | **36.5** |

Table 1: **Performance of Temporal Action Localization using different pre-trained features.** We take G-TAD [37] as the downstream method and evaluate on ActivityNet v1.3 dataset [7]. Methods highlighted in blue use fully-supervised pre-training. * denotes our implementation. R: RGB; F: optical flow; A: audio; L: language.

Temporal action localization (TAL) [26,27] is a fundamental video understanding task which requires to detect the start and end frames of action instances, as well as to predict their class labels. Here we evaluate the TAL performance using our LocVTP pre-trained features with G-TAD [37] as the downstream method. Specifically, we retrain G-TAD by only replacing the original features with our LocVTP features. The popular large-scale benchmark ActivityNet v1.3 [7] is taken as the experimental dataset.

| Method | R@1 | R@5 | R@10 | MdR |
|---|---|---|---|---|
| NE [3] | 13.7 | 35.7 | 47.7 | 12.0 |
| SupportSet [23] | 28.4 | 60.0 | 72.9 | 4.0 |
| Frozen [5] | 45.6 | 79.8 | 88.2 | 2.0 |
| BridgeFormer [12] | 52.0 | 82.8 | 90.0 | 1.0 |
| LocVTP(Ours) | **52.4** | **83.2** | **92.4** | **1.0** |

(a) MSVD

| Method | R@1 | R@5 | R@10 | MdR |
|---|---|---|---|---|
| NE [3] | 6.4 | 19.8 | 28.4 | 39.0 |
| CE [18] | 11.2 | 26.9 | 34.8 | 25.3 |
| Frozen [5] | 15.0 | 30.8 | 39.8 | 20.0 |
| BridgeFormer [12] | 17.9 | 35.4 | 44.5 | 15.0 |
| MMT [34] | 12.9 | 29.9 | 40.1 | 19.3 |
| LocVTP(Ours) [32] | **18.4** | **36.2** | **44.7** | **15.0** |

(b) LSMDC

| Method | R@1 | R@5 | R@10 | MdR |
|---|---|---|---|---|
| S2VT [31] | 11.9 | 33.6 | – | 13.0 |
| CE [18] | 16.1 | 41.1 | – | 8.3 |
| ClipBERT [16] | 20.4 | 48.0 | 60.8 | 6.0 |
| Frozen [5] | 31.0 | 59.8 | 72.4 | 3.0 |
| BridgeFormer [12] | 37.0 | 62.2 | 73.9 | 3.0 |
| OA-Trans [32] | 34.8 | 64.4 | **75.1** | 3.0 |
| LocVTP(Ours) | **37.5** | **64.7** | 74.2 | **3.0** |

(c) DiDeMo

Table 2: **Video retrieval performance.**

The comparison results to current state-of-the-art video pre-training and VL pre-training approaches are listed in Table. 1. The results show that our LocVTP pre-trained feature outperforms both video pre-training and VL pre-training methods at all four evaluation metrics. Notably, we even surpass the supervised pre-trained features (*e.g.*, BSP [35], LoFi-E2E [36], TSP [1]) which require classification labels. The excellent performance on the TAL task further demonstrates the superiority of our LocVTP on the localization tasks.

## 2 More Transfer Results on Video Retrieval

We provide the video retrieval performance on more datasets including MSVD [8], LSMDC [24], and DiDeMo [4]. We use HowTo100M as the pre-training dataset and the visual encoder is initialized with ImageNet-21k pre-trained weights (see "Settings of Pre-training" of the main paper for details). The results listed in Table. 2 show that our LocVTP also achieves state-of-the-art performance on these datasets.

## 3 Temporal Grounding Results on More Baselines

We provide temporal grounding results on more baselines with our pre-trained LocVTP features. We select two more recent state-of-the-art algorithms CSMGAN [17] and VLGNet [28]. The results are listed in Table. 3. We can observe that our LocVTP pre-trained feature consistently outperforms other VL pre-trained features using both advanced temporal grounding methods.

| Models | PT Data | $R_1^{0.5}$ | $R_1^{0.7}$ | $R_5^{0.5}$ | $R_5^{0.7}$ |
|---|---|---|---|---|---|
| Sep.Pre. [17] | Kinetics | 48.9 | 29.0 | 77.2 | 59.1 |
| VideoBERT*[29] | HT | 38.5 | 22.6 | 69.3 | 55.4 |
| MIL-NCE [21] | HT | 43.2 | 26.3 | 75.4 | 58.7 |
| UniVL [19] | HT | 43.1 | 27.0 | 76.9 | 58.3 |
| SupportSet* [23] | HT | 42.4 | 26.8 | 76.7 | 58.6 |
| **LocVTP (Ours)** | **HT** | **53.9** | **34.6** | **83.7** | **62.3** |
| Frozen [5] | CC,WV | 47.3 | 26.8 | 72.5 | 53.6 |
| OA-Trans*[32] | CC, WV | 48.0 | 27.3 | 73.0 | 53.9 |
| **LocVTP (Ours)** | **CC,WV** | **50.2** | **30.8** | **79.8** | **60.3** |
| December [30] | HT | 44.1 | 26.9 | 77.4 | 58.8 |
| ClipBERT [16] | CO,VG | 43.7 | 26.2 | 76.8 | 58.1 |

(a) Downstream method: CSMGAN [17].

| Models | PT Data | $R_1^{0.5}$ | $R_1^{0.7}$ | $R_5^{0.5}$ | $R_5^{0.7}$ |
|---|---|---|---|---|---|
| Sep.Pre. [17] | Kinetics | 46.5 | 30.1 | 77.4 | 64.3 |
| VideoBERT*[29] | HT | 36.9 | 22.4 | 69.1 | 57.7 |
| MIL-NCE [21] | HT | 40.3 | 25.6 | 74.8 | 60.6 |
| UniVL [19] | HT | 40.8 | 26.4 | 76.2 | 59.9 |
| SupportSet* [23] | HT | 39.3 | 26.3 | 76.0 | 59.2 |
| **LocVTP (Ours)** | **HT** | **54.9** | **34.2** | **83.1** | **64.4** |
| Frozen [5] | CC,WV | 45.6 | 26.6 | 72.1 | 56.0 |
| OA-Trans*[32] | CC, WV | 46.2 | 27.2 | 72.6 | 56.5 |
| **LocVTP (Ours)** | **CC,WV** | **51.2** | **31.6** | **80.7** | **61.2** |
| December [30] | HT | 41.6 | 26.4 | 76.4 | 59.2 |
| ClipBERT [16] | CO,VG | 40.3 | 26.0 | 76.1 | 58.5 |

(b) Downstream method: VLGNet [28].

Table 3: **Temporal grounding performance using different pre-trained features.** We retrain the the temporal grounding method (a) CSMGAN [17] and (b) VLGNet [28] using the pre-trained features on ActivityNet Caption dataset. `Sep.Pre.` means separately pre-training, *i.e.*, the video encoder supervisedly pre-trained on Kinetics [14] and text encoder taken from BERT [11]. HT: HowTo100M; CO: Coco Captions [10]; VG: Visual genome [15]; CC: Conceptual captions [25]; WV: WebVid-2M [5]. * denotes our implementation.

## 4 More Visualizations

**UMAP Visualizations.** We provide UMAP visualizations of *fused* multi-modal features, which are generated by multiplying the extracted video feature by the query feature. Formally, given the untrimmed video feature $\boldsymbol{v}$ and query feature $\boldsymbol{q}$ corresponding to time interval $[s, e]$, the fused features are computed by $\boldsymbol{f} = \boldsymbol{v}[s:e] \cdot \boldsymbol{q}$. Clips corresponding to ground-truth caption are marked with **green** while others are with **gray**.

Fig. 1 shows the UMAP visualization results for LocVTP (*w/* $\mathcal{L}_t$), LocVTP (*w/o* $\mathcal{L}_t$), UniVL [19], and MIL-NCE [20]. The results show that the temporal-aware contrastive loss $\mathcal{L}_t$ helps distinguish action-of-interest from background and thus benefits the localization task.

**Cross-modal Correspondence Visualizations.** Fig. 2 shows typical frames and their corresponding similarity scores with caption words. The top $K$ highest scored words are marked with red ($K = 3$). As shown, the most responsive words exactly reflect the frame content.


Fig. 1: **UMAP visualizations.** We mark clips corresponding to the ground-truth caption with **green** and others with **gray**.

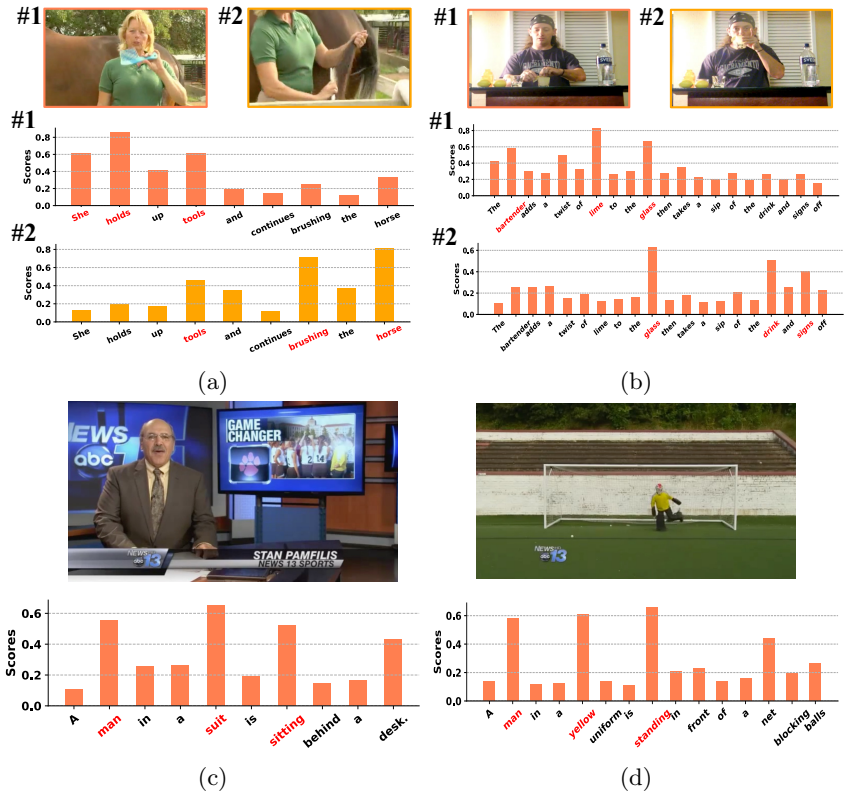


Fig. 2: **Cross-modal correspondence visualizations.** Top $K$ responsive words are marked with red. $K = 3$

## References


1. Alwassel, H., Giancola, S., Ghanem, B.: Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3173–3183 (2021) 1, 2
2. Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. Advances in Neural Information Processing Systems **33** (2020) 1
3. Amrani, E., Ben Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. arXiv preprint arXiv:2003.03186 (2020) 2
4. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: ICCV. pp. 5803–5812 (2017) 2
5. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. arXiv preprint arXiv:2104.00650 (2021) 2, 3
6. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: Learning the speediness in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9922–9931 (2020) 1
7. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015) 1
8. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011) 2
9. Chen, P., Huang, D., He, D., Long, X., Zeng, R., Wen, S., Tan, M., Gan, C.: Rspnet: Relative speed perception for unsupervised video representation learning. In: AAAI Conference on Artificial Intelligence. vol. 1 (2021) 1
10. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 3
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 3
12. Ge, Y., Ge, Y., Liu, X., Li, D., Shan, Y., Qie, X., Luo, P.: Bridgeformer: Bridging video-text retrieval with multiple choice questions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 2
13. Han, T., Xie, W., Zisserman, A.: Self-supervised co-training for video representation learning. Advances in Neural Information Processing Systems **33**, 5679–5690 (2020) 1
14. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv (2017) 3
15. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV pp. 32–73 (2017) 3

16. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7331–7341 (2021) 2, 3
17. Liu, D., Qu, X., Liu, X.Y., Dong, J., Zhou, P., Xu, Z.: Jointly cross-and self-modal graph attention network for query-based moment localization. In: ACM MM. pp. 4070–4078 (2020) 2, 3
18. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487 (2019) 2
19. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020) 1, 3, 4
20. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020) 4
21. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2630–2640 (2019) 3
22. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11205–11214 (2021) 1
23. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020) 1, 2, 3
24. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015) 2
25. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL. pp. 2556–2565 (2018) 3
26. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5734–5743 (2017) 1
27. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: CVPR. pp. 1049–1058 (2016) 1
28. Soldan, M., Xu, M., Qu, S., Tegner, J., Ghanem, B.: Vlg-net: Video-language graph matching network for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3224–3234 (2021) 2, 3
29. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7464–7473 (2019) 1, 3
30. Tang, Z., Lei, J., Bansal, M.: Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2415–2426 (2021) 3

31. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014) 2
32. Wang, A.J., Ge, Y., Cai, G., Yan, R., Lin, X., Shan, Y., Qie, X., Shou, M.Z.: Object-aware video-language pre-training for retrieval. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 2, 3
33. Wei, D., Lim, J.J., Zisserman, A., Freeman, W.T.: Learning and using the arrow of time. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8052–8060 (2018) 1
34. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: AAAI. pp. 9062–9069 (2019) 2
35. Xu, M., Pérez-Rúa, J.M., Escorcia, V., Martinez, B., Zhu, X., Zhang, L., Ghanem, B., Xiang, T.: Boundary-sensitive pre-training for temporal localization in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7220–7230 (2021) 1, 2
36. Xu, M., Perez Rua, J.M., Zhu, X., Ghanem, B., Martinez, B.: Low-fidelity video encoder optimization for temporal action localization. Advances in Neural Information Processing Systems **34** (2021) 1, 2
37. Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B.: G-tad: Sub-graph localization for temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10156–10165 (2020) 1