LLaViLo: Boosting Video Moment Retrieval via Adapter-Based Multimodal Modeling

Kaijing Ma[1,2*] Xianghao Zang[1*] Zerun Feng[1] Han Fang[1†] Chao Ban[1]
Yuhan Wei[1,3] Zhongjiang He[1] Yongxiang Li[1] Hao Sun[1†]
[1]China Telecom Corporation Ltd. Data&AI Technology Company [2]Xi'an Jiaotong University [3]Rice University

Abstract

*Recent studies have explored the potential of large language models (LLMs) for understanding the semantic information in images. However, the use of LLMs to understand videos, which contain continuous contextual information, remains limited. In this paper, we propose **LLaViLo** (**LLa**Ma-**Vi**deo-**Lo**calizer), a video moment retrieval pipeline powered by a large language model. LLaViLo has two key features: 1) In contrast to fine-tuning the entire LLM, we introduce and optimize only 1.7% of additional parameters in adapter modules, freezing the pre-trained LLM to enable efficient alignment of video and text. 2) A multi-objective optimization framework concurrently optimizes two objectives: a set prediction objective and a captioning objective. The joint training of these two objectives allows the proposed framework to produce high-quality time coordinates. Compared with other state-of-the-art methods, the proposed LLaViLo achieves significant performance improvement on QVHighlights and Charades-STA datasets.*

1. Introduction

Given a natural language query, moment retrieval (MR) aims to locate the most relevant segments from an untrimmed video, requiring effective modeling of the visual and textual semantics across continuous time. However, current models [6, 9, 31] still struggle with semantic reasoning and relevance matching. They rely on separate modeling of visual and textual features, lacking deep integration between these two modalities, as shown in Figure 1 (a).

Recent works on LLMs for visual tasks [1, 7, 32] suggest potential benefits, *i.e.,* LLMs have shown impressive capabilities in visual-textual semantic reasoning. However, LLMs have not been thoroughly explored in the video domain.

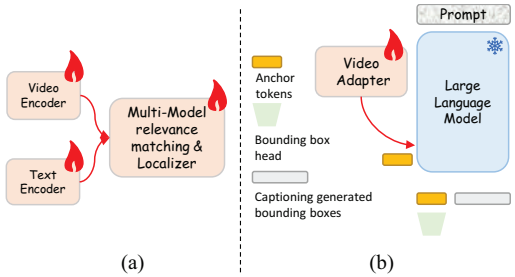

Figure 1. Comparison of two kinds of MR Methods. (a) Separate modeling of visual and textual features, (b) our proposed LLM-based video moment retrieval.

Fine-tuning LLMs confronts two key obstacles: 1) The immense computational complexity of fine-tuning billions of parameters results in prohibitive resource costs; 2) catastrophic forgetting happens for the pre-trained knowledge as models adapt to new tasks [12, 29].

We propose an efficient video moment retrieval framework integrating LLMs as shown in Figure 1 (b). Textual information is integrated into video representations to generate joint feature representation, which is injected into an LLM. Compared to full fine-tuning, we only optimize parameters in the adapter, avoiding catastrophic forgetting. We further adopt a multi-objective optimization approach with two complementary objectives: 1) a DETR-like [2] set prediction objective localizing relevant segments, and 2) a language modeling objective generating textual time coordinates. By concurrently optimizing set prediction and language modeling, a co-learning manner is achieved, improving the video-text alignment. Set prediction focuses on the cross-modal alignment between clips and query semantics, while language modeling facilitates textual understanding. The proposed video adapter and multi-objective optimization approach leverage the potential benefits of LLMs and achieve better video-text alignment.

In summary, the main contributions of this work are three-fold: 1) We propose a lightweight adapter module

*Both authors contributed equally to this work.
†Corresponding authors.

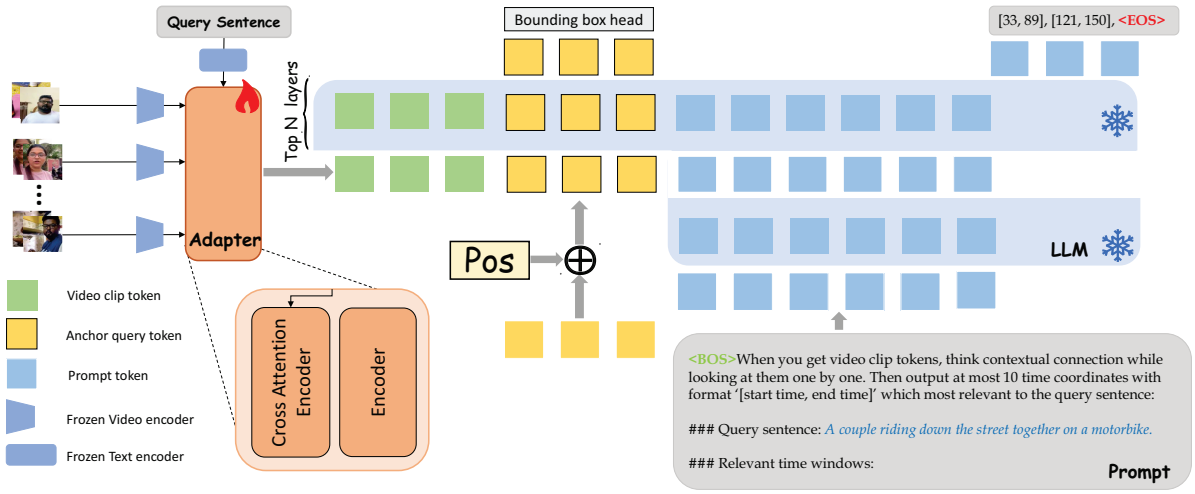

Figure 2. Overview of our proposed model architecture. It consists of a frozen video encoder, a frozen text encoder, an adapter module to integrate multimodal context, and two objective-specific heads for set prediction and language modeling.

to incorporate multimodal video-text representations into LLMs, which leverages the capability of LLMs for video understanding. 2) We develop a multi-objective learning framework including a set prediction objective and a language modeling objective, which are complementary and optimized concurrently. 3) Our model achieves state-of-the-art performance on two benchmarks for the moment retrieval task.

2. Related work

Moment Retrieval. Prior works on MR struggled to capture fine-grained semantics and achieve limited alignment between modalities. The existing methods mainly employ stronger video encoders [5, 28] and language encoders [3, 19] to capture representations independently. However, such a dual-stream network introduces significant gaps(the inherent difference between visual and textual representations) in the subsequent matching and fusion processes of the two modalities.

To tackle this issue, Liu *et al*. [18] propose a unified multi-modal transformer (UMT) inspired by DETR [2], which jointly optimizes MR and highlight detection [13], reducing the gap between two modalities effectively. However, the potential of LLMs, a naturally powerful textual transformer decoder, remains unexplored in the MR task.

LLM-based Visual Understanding. Recent studies [11, 15, 16, 26] have begun exploring language models for visual analysis, showing strong reasoning abilities. LLaMa-Adapter [7, 32] enables efficient fine-tuning of large pre-trained language models through adapters with minimal trainable parameters. Compared to full fine-tuning methods like Alpaca [27], it significantly reduces computational costs and storage requirements. However, LLaMa-Adapter is designed primarily for textual and visual instructions, with limited capability in handling video inputs. Extending LLaMa-Adapter to video domains remains an open research direction.

Moreover, the existing LLM-based video moment retrieval approach [30] operates on individual video frames rather than leveraging temporal context. Our work addresses this limitation by integrating continuous video encoding into LLM architectures.

3. Method

Given an untrimmed video V and textual query q, the goal of moment retrieval is to predict start and end times (t_s, t_e) to localize relevant moments in V that semantically match q.

As illustrated in Figure 2, the overall architecture integrates an LLM decoder with a multi-objective optimization framework. Video clip and text representations are fused by the adapter module and then inserted into the LLM. A set prediction head and a language modeling head are added to LLM to utilize the LLM's pre-trained representation capability and achieve efficient video-text alignment.

3.1. Video Semantic Modeling Adapter

Video Adapter. We utilize two pre-trained models to extract embeddings for video clips and text queries separately. The resulting video and text embeddings are then fused through a cross-transformer encoder with two layers. The cross-attention between video and text embeddings can be formulated as:

$$\text{CrossAtt}(\mathbf{Q_v}, \mathbf{K_t}, \mathbf{V_t}) = \text{softmax}\left(\frac{\mathbf{Q_v}{\mathbf{K_t}}^T}{\sqrt{d}}\right)\mathbf{V_t}, \quad (1)$$

where $\mathbf{Q}_v$ represents video query vectors, $\mathbf{K}_t, \mathbf{V}_t$ are text key and value vectors, d is the dimension of *query*, *key* and

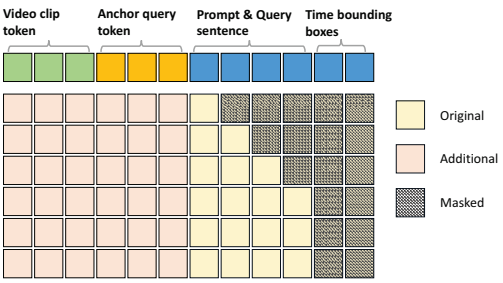

Figure 3. Adaption attention mask.

value. This allows the model to capture fine-grained relationships between the multimodal inputs.

Following cross-modal encoding, the fused video-text representations are further processed by a transformer encoder with two layers, which provides a deeper understanding of each clip-text pair. Compared with the original LLM, the proposed adapter with a two-layer cross-attention encoder and a two-layer transformer encoder only introduces 1.7% additional parameters, which leverages the understanding ability of LLMs for video-text content efficiently.

LLM-based Video Understanding. In this section, we insert encoded video-text representations into an LLM architecture, *i.e.,* LLama, and modify the LLaMa architecture to adapt the video domain.

Specifically, we take fused video-text tokens generated by the video adapter and project them to the same hidden dimension size C as LLaMa. For saving GPU memory, we only adjust the top N layers of LLaMa. The projected video tokens P are inserted into each one of the N layers. For the n^{th} layer, P_n are concatenated with anchor query tokens A_n (please refer to 3.2) and text tokens T_n, formulated as $[P_n; A_n; T_n] \in \mathbb{R}^{(V+J+K)\times C}$, where V, J, and K represent the lengths of video clip, anchor query token, and text prompt token, respectively.

We use zero-initialized attention to avoid disruptions from randomly initialized video adapter tokens during fine-tuning. As formulated in Equation 2, the attention scores S_n^V of video clip tokens are controlled by a gating factor g_l, which is initialized to zero:

$$S_n^g = \left[\text{softmax}(S_n^V) \cdot g_l \,; \text{softmax}(S_n^{J+K})\right], \tag{2}$$

where S_n^{J+K} denotes the attention scores of concatenated anchor query tokens and text tokens. In order to avoid destroying the original information in LLaMa at the beginning, we gradually increase g_l, to progressively incorporate the video semantics into the model.

3.2. Multi-Objective Optimization

Set Prediction Head. As mentioned above, the set prediction head follows a DETR-like design to localize relevant moments. Specifically, J additional anchor tokens are employed to represent learnable query embeddings, which are denoted as A. For the n^{th} layer at the top of LLaMa, the corresponding anchor token is represented by A_n.

Each anchor query token has a unique positional embedding and undergoes self-attention to predict the start and end time points in a video for the given query text. An alignment loss using the Hungarian algorithm is employed to match the predicted time coordinates and ground truth as follows,

$$\mathcal{C}_{\text{m}} = -c_i \neq \emptyset \hat{p}_{\sigma(i)}(c_i) + c_i \neq \emptyset \mathcal{L}_{\text{m}}, \tag{3}$$

where c_i is the ground truth label and m is a vector that defines the normalized center coordinate and duration.

Then, the loss of moment localization L_{m} is defined as follows:

$$\mathcal{L}_{\text{m}} = \lambda_1 ||m - \hat{m}|| + \lambda_2 L_{\text{gIoU}}(m, \hat{m}) + \lambda_3 L_{\text{CE}}, \tag{4}$$

where m and $\hat{m}$ are ground-truth and its corresponding prediction containing center coordinate and duration. Also, λ_1, λ_2 and λ_3 are hyperparameters for balancing the losses.

As shown in Figure 3, we propose a novel masking strategy to locate specific moments in a video-query context. The pink tiles represent the additional tokens for the top N layers of LLaMa. We also mask all text tokens containing the start and end time points to avoid information disclosure. This ensures anchor tokens don't rely only on timestamps to predict boxes in the training process.

Multi-Objective Optimization. Our model combines a set prediction objective and a time-window captioning objective, which are optimized concurrently.

For captioning, we provide a prompt to instruct the model to generate the time windows in a standard textual format (as shown in Figure 2) according to the text query. The captioning head is trained to predict the ground truth time stamp captions using a cross-entropy loss $\mathcal{L}_{\text{cap}}$:

$$\mathcal{L}_{\text{cap}} = -\sum_{t=1}^{T}\sum_{f=1}^{F} y_{t,f} \log(p_{t,f}) \tag{5}$$

where T is caption length, F is vocabulary size, $y_{t,f}$ is 1 if word f is the ground truth at position t, and $p_{t,f}$ is the predicted probability.

By optimizing the captioning and set prediction objectives during training, we improve the alignment between video as follows,

$$\mathcal{L} = \alpha \mathcal{L}_{\text{cap}} + \beta \mathcal{L}_{\text{m}}, \tag{6}$$

where α and β are trade-off parameters weighing the importance of each objective.

4. Experiments

4.1. Datasets & Implementation details

We experiment with our LLaViLo on below two datasets, following existing data splits from the existing works [13, 18].

QVHighlights [13] contains over 10,000 YouTube videos with diverse topics and both first-person and third-person perspectives. **Charades-STA** [6] consists of 6,768 indoor activity videos with over 16,000 textual queries labeled with relevant moments.

A SlowFast [5] network and a text encoder in CLIP [10, 23, 25] are used as frozen encoders to extract representations. Following [21], the clip length is set to 2 for QVHighlights and 1 for Charades-STA. We use a 3-layer MLP to match the hidden dimension of LLaMa's architecture. The adapter module has a hidden dimension of 256, which we project to 4096 before injecting tokens into LLaMa.

For evaluation, we report four commonly utilized metrics on the test sets: Recall at rank 1 with intersection over union (IoU [8, 17, 24]) thresholds of 0.5 and 0.7 (R1@0.5 and R1@0.7), along with mean Average Precision at [13].

4.2. Compare to the state-of-the-art

We conduct comparative experiments against both conventional moment retrieval methods and LLM-based approaches. As shown in Table 1, our proposed LLaViLo model achieves superior performance over the current state-of-the-art methods on all evaluation metrics. Specifically, we obtain significant improvements of over 4% on four metrics compared to the previous best method. This validates that through our video semantic modeling adapters and multi-objectives, our model is able to gain a more accurate understanding of complex video-text queries. We also conduct experiments on the Charades-STA dataset as shown in Table 2, and LLaViLo shows better performances compared with other state-of-the-art models.

4.3. Ablation results

We conducted ablation studies to validate the effectiveness of the proposed modules, as shown in Table 3. Firstly, removing the caption loss results in significant performance drops across all metrics, demonstrating the importance of joint optimization with the language modeling objective. Secondly, we vary the number N of top layers in LLaMa, from 2 to 8. The steady performance gains are observed as more top layers were injected with fused video-text information. These performance improvements demonstrate that more injected layers introduce better video-text content understanding for the LLaMa model.

Besides, we also evaluate the effectiveness of instruction in natural language prompt, *i.e.,* the sentence started with <BOS> in Figure 2. As illustrated in Table 4, after using the task-guided instruction, the model performance achieves significant improvements. These comparisons demonstrate that task-guided instructions can help to explore more potentials of LLMs for video understanding tasks.

Method	R@0.5	R@0.7	mAP@0.5	avg mAP
MCN [9]	11.41	2.72	24.94	10.67
CAL [4]	25.49	11.54	23.40	9.89
CLIP [22]	16.88	5.19	18.11	7.67
XML [14]	41.83	30.35	44.63	32.14
XML+	46.69	33.46	47.89	34.90
Moment-DETR [13]	52.89	33.02	54.82	30.73
UMT [18]	-	-	-	36.12
LLM-based Method	R@0.5	R@0.7	mAP@0.5	avg mAP
SeViLA [30]	54.5	36.5	-	32.3
LLaViLo(ours)	**59.23**	**41.42**	**59.72**	**36.94**

Table 1. Performance comparison on QVHighlights

Method	R@0.5	R@0.7
CTRL [6]	23.63	8.89
2D-TAN [33]	39.81	23.31
SimVTP [20]	44.7	26.3
UMT	49.35	26.16
Moment-DETR	53.63	31.37
LLaViLo(ours)	**55.72**	**33.43**

Table 2. Performance comparison on Charades-STA

N	w/o $\mathcal{L}_{\text{cap}}$		w/ $\mathcal{L}_{\text{cap}}$	
	R@0.5	R@0.7	R@0.5	R@0.7
2	53.81	35.06	56.97	37.10
4	55.52	38.48	57.34	39.74
6	55.65	37.61	58.32	40.06
8	56.42	39.74	**59.23**	**41.42**

Table 3. **Ablation study on Caption Loss.** N refers to the top N layers of LLM. The evaluation is conducted on the QVHighlights dataset

	R@0.5	R@0.7	mAP@0.5	avg mAP
w/o Instruction	53.12	34.02	56.52	34.45
w/ Instruction	**59.23**	**41.42**	**59.72**	**36.94**

Table 4. **Ablation study on Natural Language Prompt.** The evaluation is conducted on the QVHighlights dataset.

5. Conclusion

This work presents an efficient video-text modeling approach integrating a video semantic modeling adapter and a language model. The lightweight adapters enable incorporating multimodal semantics. A multi-objective learning framework optimizes complementary moment localization and language modeling jointly. Comprehensive experiments validate state-of-the-art video retrieval performance. This demonstrates the efficacy and efficiency of our proposed techniques for advancing language models on multimodal understanding tasks.

References

[1] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, Mar. 2023.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[4] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[6] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[7] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[9] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017.

[10] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[11] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 18–24 Jul 2021.

[12] Tomasz Korbak, Hady Elsahar, German Kruszewski, and Marc Dymetman. Controlling conditional language models without catastrophic forgetting. In *International Conference on Machine Learning*, pages 11499–11528. PMLR, 2022.

[13] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

[14] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[16] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.

[18] Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, 2022.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[20] Yue Ma, Tianyu Yang, Yin Shan, and Xiu Li. Simvtp: Simple video text pre-training with masked autoencoders. *arXiv preprint arXiv:2212.03490*, 2022.

[21] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv*, 2021.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[26] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[27] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[28] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[29] Matthew Wallingford, Hao Li, Alessandro Achille, Avinash Ravichandran, Charless Fowlkes, Rahul Bhotika, and Stefano Soatto. Task adaptive parameter sharing for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7561–7570, 2022.

[30] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023.

[31] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, Online, July 2020. Association for Computational Linguistics.

[32] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[33] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020.