Alignment and Generation Adapter for Efficient Video-text Understanding

Han Fang[1*] Zhifei Yang[1,2*] Yuhan Wei[1,3*] Xianghao Zang[1] Chao Ban[1]
Zerun Feng[1] Zhongjiang He [1] Yongxiang Li [1] Hao Sun[1†]

[1]China Telecom Corporation Ltd. Data&AI Technology Company
[2]Beijing University of Posts and Telecommunications [3]Rice University

Abstract

*Pre-trained models have demonstrated considerable performance, especially in enhancing cross-modal understanding between videos and text. However, fine-tuning them at scale becomes costly and poses challenges for adapting to various downstream tasks. To tackle these challenges, we propose the Alignment-generation Adapter (**AGAdapter**), establishing semantic coherence between alignment and generation models for efficient video-text adaptation across multiple tasks simultaneously. We propose an alignment adapter with knowledge-sharing to adapt the frozen CLIP model for fine-grained video-language interaction. Additionally, we introduce the generation adapter with prompt tuning to leverage the large language model for captioning. Furthermore, we introduce instruction joint tuning, combining textual and cross-modal instructions, to capture detailed descriptions. Our AGAdapter achieves state-of-the-art performance on video-text retrieval and video captioning tasks, including two benchmarks, MSR-VTT and ActivityNet.*

1. Introduction

Video-text understanding [5, 23, 3, 45, 46], encompassing video-text retrieval [19, 44, 7, 17] and video captioning [34, 26, 8], represents a fundamental task that revolves learning semantic coherence. Video-text retrieval [3, 4, 18] refers to the process of searching for videos or captions using a cross-modal query. In contrast, video captioning [35, 33, 51] aims to generate descriptions for a video.

Advancements in image-text pre-trained models have demonstrated remarkable generalization, inspiring various video-text methods [23, 5, 17] to leverage the knowledge of pre-trained models [31]. In video-text retrieval, several works focus on designing temporal information [19, 17] to align image representations at the video level. However, these methods still train the model in an end-to-end manner [23], resulting in significant computational overhead. Additionally, the captioning models [9] emphasize reasoning about sophisticated relations and objects. Decoder networks such as GPT-2 [32] are employed to transform video representations. Recently, large language models have proven potential to handle visual inputs [49, 16] for image captioning. However, efficiently capturing spatial and temporal relations at the video level remains challenging.

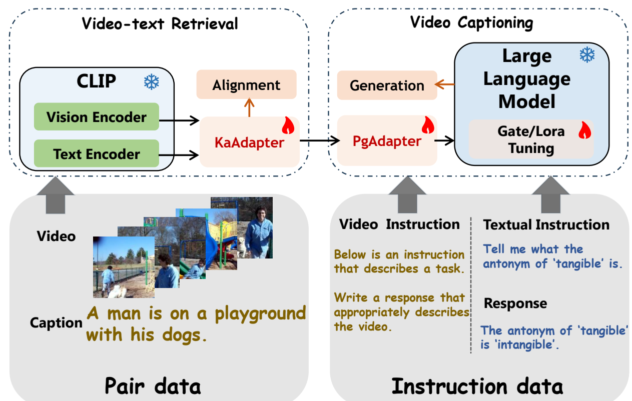

Figure 1. Training pipeline of **AGAadapter.** Our parameter-efficient adaptation method incorporates the pre-trained CLIP model with a large language model, using cross-modal instructions for video-text retrieval and video captioning tasks.

To address these limitations, we present the alignment and generation adapter (**AGAdpter**), which unifies the pre-trained aligned model and large language model. To facilitate the adaptation of pre-trained aligned models in video-text retrieval, we propose a knowledge-sharing alignment adapter (**KaAdapter**). By incorporating textual and video queries to indicate modality-specific knowledge and applying parameter-sharing modeling, cross-modal representations can be fully aligned. On the other hand, we present the prompt-following generation adapter (**PgAdapter**) to leverage the reasoning power of large language models for video captioning tasks. The proposed PgAdapter learns to transform aligned video representations into adaptation prompts. These prompts serve as video content, being injected into each layer of the large language model, pro-

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

gressively enhancing its video reasoning ability to generate captions. Additionally, we introduce an **instruction joint tuning strategy** that combines video-text pairs with instruction-following data. This strategy enhances the extraction of specific video information, enabling the model to capture finer details in the video content.

2. Related Work

Video-text Retrieval. Early methods [3, 4, 18] leverage multiple representations fusion for cross-modal alignment [48, 41]. Recent studies [7, 39] have adopted an end-to-end manner to train models. CLIP4Clip [23] and CLIP2Video [5] propose temporal modeling to transfer prior knowledge from CLIP [31]. However, these methods, which employ full-parameter training, are unable to utilize larger CLIP models due to high computational costs.

Video Captioning. The encoder-decoder frameworks [42, 34] are adopted for video captioning in which traditional methods focus on graph modeling [50] or mutual knowledge distillation [29]. Building upon these studies, recent researches [27, 6, 35] have utilized pre-trained models to extract aligned features for cross-modal decoding. Furthermore, we utilize large language models with the proposed adapters to generate detailed descriptions.

LLMs for Vision-Language Tasks. The adaption of large language models (LLMs) has increased in vision-language tasks [28, 25, 20, 40]. MiniGPT-4 [52] aligns visual information with Vicuna [2] without external visual reasoning modules. To bridge the gap between LLaMA [37] and visual instructions, LaVIN [21] introduces adaptation modules. In this work, we propose an adapter-based method to address video-related language tasks using LLMs.

3. Method

3.1. Knowledge-sharing Alignment Adapter

As illustrated in Fig. 2, to achieve video-text retrieval, we first apply frozen CLIP to extract frame and word representations. The M frames are sampled and fed into the vision encoder to generate frame tokens as $e_f = \{f_0, f_1, \cdots, f_{M-1}\}$. Besides, the caption is appended with two special tokens and input into the text encoder to generate word tokens as $e_w = \{w_{\mathrm{SOS}}, w_1, \cdots, w_{N-2}, w_{\mathrm{EOS}}\}$. N represents the number of text tokens.

To model the frame and word representations, which are extracted from frozen CLIP, into the joint space, we propose the weight-sharing adapter (**KaAdapter**). The KaAdapter employs unified attention interaction with shared weights to encode both textual and video information in a more parameter-efficient manner. Additionally, we introduce video and textual queries as indicators to model modality-specific knowledge. The learnable video and textual queries are denoted as q_v^A and q_t^A ($R^{n_a \times d_a}$), where d_a is the dimension of token embedding, and n_a is the number of queries. Therefore, the aligned video representations are captured through unified attention interaction as follows:

$$e_v = \mathit{Softmax}(Q_u(e_f)K_u(q_v^A)/\sqrt{d_a})^T \cdot V_u(q_v^A), \quad (1)$$

where the cross-attention transformation is represented by Q_u, K_u, and V_u. The output video representation is denoted by e_v and shares the same dimension ($R^{M \times d_a}$) as the frame representation e_v. The aligned text representation can be modeled in the same manner by applying e_w and q_t^A to replace e_f and q_v^A. The training objective of token-based contrastive loss based on WTI [39] can be formulated as:

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp\left(\mathrm{WTI}\left(\mathbf{e}_{v,i}, \mathbf{e}_{t,i}\right)/\tau\right)}{\sum_j^B \exp\left(\mathrm{WTI}\left(\mathbf{e}_{v,i}, \mathbf{e}_{t,j}\right)/\tau\right)}, \quad (2)$$

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp\left(\mathrm{WTI}\left(\mathbf{e}_{t,i}, \mathbf{e}_{v,i}\right)/\tau\right)}{\sum_j^B \exp\left(\mathrm{WTI}\left(\mathbf{e}_{t,i}, \mathbf{e}_{v,j}\right)/\tau\right)}, \quad (3)$$

$$\mathcal{L}_{vt} = \frac{1}{2}(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}). \quad (4)$$

where B represents batch size, and τ is the temperature and pair-wise token correlations are fully exploited to maximize the similarity between positive pairs based on all tokens.

3.2. Prompt-following Generation Adapter

To transform video representation into the video prompt, which can be integrated with LLMs, we propose the prompt-following generation adapter (**PgAdapter**). As CLIP and LLMs have different distributions, PgAdapter maps the video representations $\mathbf{e}_v$ to prompt embedding as:

$$p_v = \sum_{i=1}^{K} \mathit{Softmax}(Q_g^i(q_v^G)K_g^i(e_v)^T/\sqrt{d_g}) \cdot V_g^i(e_v), \quad (5)$$

where Q_g^i, K_g^i, and V_g^i are the i-th mapping network for cross-attention transformation. q_v^G is the prompt token, which is the N_g learnable embedding with the same dimension d_g as the internal mappings as LLMs. To capture and preserve more temporal and spatial information, we utilize multiple mapping mechanisms and sum their outputs to form the final prompt, denoted as $p_v \in R^{N_g \times d_g}$.

Inspired by LLaMA-adapter [49], we insert video prompts into each layer of the LLMs with zero-init gating attention [49], progressively incorporating video information with the reasoning power. To achieve this, the video prompts are reshaped into the dimension as $R^{(N_g//l) \times l \times d_g}$, where l is the number of layers in LLMs. Therefore, the group of video prompts can be obtained as $p_v^{LLM} = \{p_v^1, p_v^2, \cdots, p_v^l\}, p_v^l \in R^{(N_g//l) \times d_g}$. The video prompts are injected by zero-init gating attention as:

$$S^{V-T} = [S_l^V(p_v^l) \cdot \alpha^l, S_l^T], \quad (6)$$

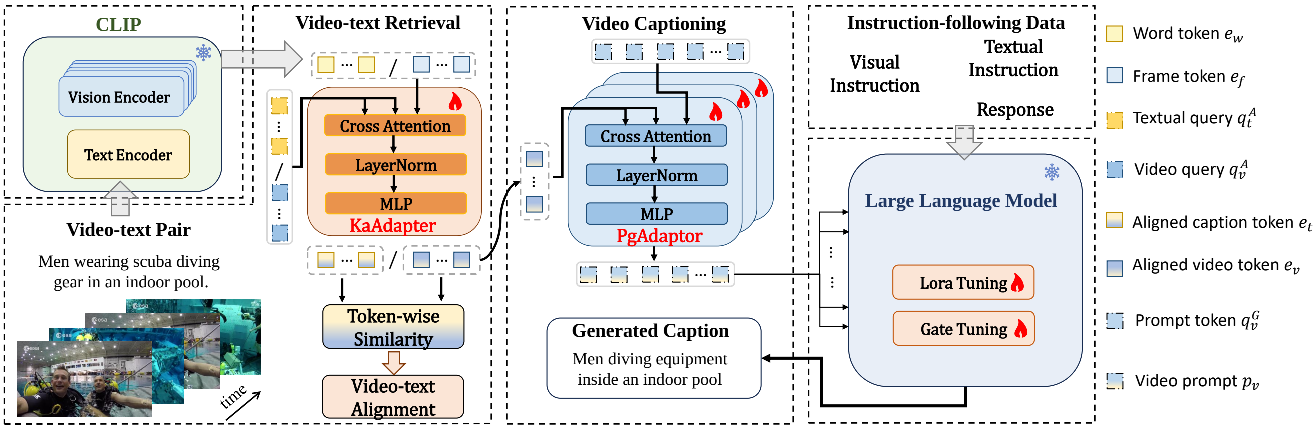

Figure 2. The overall framework of Alignment-generation Adapter. Given the video and captions, we first adopt CLIP to extract image and text representations. Then, we employ **KaAdapter** for fine-grained video-text alignment. The aligned video representations are transformed into video prompts by **PgAdapter** and incorporated with a large language model to generate video captions.

where $S_l^V(p_v^l)$ is the attention score, which contains the video content. By applying video instructions such as "describe the video" to instruct LLMs, the attention scores S_l^T between instructions are also measured. By applying α^l as the learnable weight of video content, the multi-modal alignment is progressively achieved. Moreover, we utilize Lora [10] tuning to optimize the adaptation of the large language model, enhancing cross-modal reasoning ability. Therefore, our training objective is to predict the caption tokens conditioned on the video prompts, where the cross-entropy loss function is optimized:

$$\mathcal{L}_{cap} = -\sum_{j=1}^{J} \log p(w_j | w_{<j}, Instruction), \qquad (7)$$

where J is the maximal length of the predicted word tokens, and w_j is the j-th predicted word token. By incorporating the video prompts, LLMs are able to generate textual descriptions in the context of relevant video concepts, facilitating more contextually meaningful captions.

3.3. Instruction Joint Tuning

To enhance multi-modal understanding, we present an instruction joint tuning that effectively combines video-text pairs with textual instruction. Specifically, we adopt Alphaca-52k [36] to instruct large language models by Lora tuning, thus adapting LLMs for knowledge reasoning. The Lora mappings within LLMs, trained by instructions, also enhance the understanding of sophisticated relations in video content, leading to more detailed descriptions. Overall, the total loss is obtained as:

$$\mathcal{L}_{all} = \mathcal{L}_{vt} + \beta \mathcal{L}_{cap} + \gamma \mathcal{L}_{cap}^{I}(\theta_{Lora}), \qquad (8)$$

where $\mathcal{L}_{cap}^{I}(\theta_{Lora})$ is the loss function that accepts the textual instruction as input and only fine-tunes the Lora mapping. β and γ are the weight to control trade-off.

4. Experiments

4.1. Experimental Setting

Datasets. We evaluate video-text retrieval on MSR-VTT [43] and ActivityNet [12], where 9k protocol in MSR-VTT including 9k and 1k videos for train and testing, and video-paragraph retrieval settings in ActivityNet [12] are utilized. For video captioning, we use the full protocol [3] of MSR-VTT. We also incorporate WebVid-2M [1] for pre-training.

Evaluation. Following the existing retrieval task, Recall at rank K (R@K) and mean rank (MnR) are reported, where lower MnR and higher R@K indicate better performance. For video captioning, we report metrics, such as BLEU [30], ROUGE [14], and CIDEr [38].

Implementation Details. We employ the frozen pre-trained CLIP-bigG/14 [11] to encode both frame and word tokens. The KaAdapter serves as a 1-layer transformer following the adopted CLIP model. The dimension of both video and textual queries is 3×1280. As for the PgAdapter, we stack three 1-layer transformers and sum their outputs to create video prompts of dimension $R^{320 \times 1280}$. In order to adapt to the LLaMA-7B [37], we split the prompt into 32 tokens of dimension $R^{10 \times 1280}$ and insert them into each layer. For the MSR-VTT dataset, we set the video and caption lengths to 12 and 32, respectively. For the ActivityNet dataset, both video and caption lengths are set to 64. The model is trained for 5 and 10 epochs for the MSR-VTT and the ActivityNet dataset, with a batch size of 32. To pre-train on WebVid-2M, we use the same settings as for MSR-VTT, with the exception of training for 2 epochs. Additionally, we set the values of β and γ to 0.5 and 0.1, respectively.

Method	K	Text2Video R@1	R@5	R@10	MnR	Video2Text R@1	R@5	R@10	MnR	Video Captioning BLEU@4	ROUGE	CIDEr	Training Time
CLIP-finetune	-	46.6	73.4	83.5	13.0	45.4	73.4	81.9	9.1	-	-	-	1.8h
$\mathcal{L}_{vt}$	-	48.8	74.0	83.6	12.3	48.3	74.5	84.1	8.6	-	-	-	0.12h
$\mathcal{L}_{vt} + \mathcal{L}_{cap}$	1	49.5	74.3	83.8	11.9	49.2	75.1	84.6	8.2	46.7	64.4	59.8	0.25h
$\mathcal{L}_{vt} + \mathcal{L}_{cap}$	3	50.4	74.9	84.1	11.1	50.0	75.7	85.1	8.1	47.5	64.5	60.9	0.33h
$\mathcal{L}_{vt} + \mathcal{L}_{cap} + L^{I}_{cap}$	3	51.2	75.6	84.8	10.8	50.8	76.2	86.2	7.8	48.0	64.7	62.1	0.5h

Table 1. Ablation results on the different settings of the proposed method. All the results are evaluated on the MSR-VTT dataset. K refers to the number of stacked layers in the PgAdapter. Training time represents the time to train for 1 epoch on V100 $\times$ 8 GPUs.

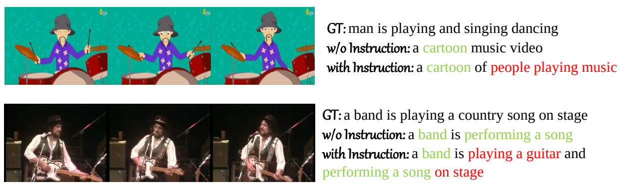

Figure 3. Visualizations of the generated captions with and without the instruction joint tuning strategy on the MSR-VTT dataset.

4.2. Main Results

Ablation Study. We thoroughly investigate various settings for our proposed AGAdapter and present comprehensive comparisons in Tab. 1. In the experiments, we use CLIP-finetune as the baseline, which employs the learned ViT-B/32 as the backbone and WTI [39] for interaction. As observed, applying only KaAdapter for $\mathcal{L}_{vt}$ can effectively transfer the knowledge of frozen CLIP for parameter-efficient video-text adaptation. By utilizing PgAdapter to incorporate LLaMA [37] for multi-task learning, the performance of video-text retrieval is further improved. Moreover, increasing the value of K to transform video representations as prompts allows more video content to be modeled and preserved, and the performance of both two tasks is improved. Additionally, introducing textual instruction to train Lora mapping enhances the cross-modal understanding, which leads to the best performance.

Comparisons with State-of-the-art Models. In video-text retrieval, we compare the performance of our AGAdapter with other state-of-the-art methods on two datasets: MSR-VTT [43] and ActivityNet [12]. The results for both video-to-text (V2T) and text-to-video (T2V) retrieval are presented in Tab. 2. Our method demonstrates superior performance, achieving a significant margin of improvement while still maintaining limited computational costs, even without any pre-training. Moreover, when we apply WebVid-2M [1] for pre-training, our performance is further enhanced. We also evaluate the performance of video captioning, as shown in Tab. 3. Remarkably, our method outperforms the other results while requiring limited datasets for pre-training.

Qualitative Results. We illustrate the visualization of generated captions under different settings in Fig. 3. It is evident that AGAdapter without instruction joint tuning tends to produce more generic descriptions. However, when the instruction joint tuning strategy is applied, AGAdapter generates video captions with finer details. This outcome the effectiveness of the instruction joint tuning method.

Method	MSR-VTT T2V R@1	R@5	R@10	MnR	V2T R@1	R@5	R@10	MnR
CLIP4Clip [23]	44.5	71.4	81.6	15.3	42.7	70.9	80.6	11.6
CLIP2Video [5]	45.6	72.6	81.7	14.6	43.5	72.3	82.1	10.2
X-CLIP [24]	46.1	73.0	83.1	13.2	46.8	73.3	84.0	8.1
TS2Net [19]	47.0	74.5	83.8	13.0	45.3	74.1	83.7	8.9
AGAdapter	**51.2**	**75.6**	**84.8**	**10.8**	**50.8**	**76.2**	**86.2**	**7.8**
AGAdapter*	**51.8**	**76.5**	**86.9**	**10.5**	**51.5**	**77.3**	**86.9**	**7.2**
	ActivityNet							
CLIP4Clip [23]	40.5	72.4	-	7.5	41.4	73.7	-	6.7
TS2-Net [19]	41.0	73.6	84.5	8.4	40.5	73.4	-	-
AGAdapter	**49.0**	**78.1**	**88.6**	**5.2**	**45.6**	**76.2**	**86.9**	**6.3**
AGAdapter*	**50.1**	**79.5**	**89.1**	**5.1**	**46.4**	**78.3**	**87.3**	**5.9**

Table 2. Performance comparisons on video-text retrieval. * means adopting WebVid-2M [1] for pre-training.

Method	#PT Data	BLEU@4	ROUGE	CIDEr
UniVL[22]	136M	42.2	61.2	49.9
SwinBERT[15]	-	41.9	62.1	53.8
CLIP4Caption[35]	400M	46.1	63.7	57.7
MV-GPT[34]	53M	48.9	64.0	60.0
LAVENDER[13]	30M	-	-	60.1
Vid2Seq[45]	314M	-	-	61.5
HiTeA[47]	5M	-	-	62.5
AGAdapter	**-**	**48.0**	**64.7**	**62.1**
AGAdapter*	**2M**	**48.2**	**65.0**	**63.7**

Table 3. Comparisons on video captioning on MSR-VTT [3]. * means adopting WebVid-2M [1] for pre-training. #PT Data is the number of video-text pairs for pre-training.

5. Conclusion

This paper addresses the challenges posed by the costly fine-tuning of pre-trained models for efficient video-text adaptation across multiple tasks. The Alignment-generation Adapter including knowledge-sharing alignment adapter and prompt-following generation adapter are adopted to incorporate the CLIP model and large language model, leveraging for video-text retrieval and video captioning. We introduce instruction joint tuning, combining text and cross-modal instructions, to enhance video-text understanding.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. 2021.

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[3] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. MDMMT: multidomain multimodal transformer for video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 3354–3363, 2021.

[4] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[5] Han Fang, Pengfei Xiong, Luhui Xu, and Wenhan Luo. Transferring image-clip to video-text retrieval via temporal relations. *IEEE Transactions on Multimedia*, 2022.

[6] Federico Galatolo., Mario Cimino., and Gigliola Vaglini. Generating images from caption and vice versa via clip-guided generative latent space search. *Proceedings of the International Conference on Image Processing and Vision Engineering*, 2021.

[7] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5006–5015, 2022.

[8] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18941–18951, 2023.

[9] Xingjian He, Sihan Chen, Fan Ma, Zhicheng Huang, Xiaojie Jin, Zikang Liu, Dongmei Fu, Yi Yang, Jing Liu, and Jiashi Feng. Vlab: Enhancing video language pre-training by feature adapting and blending. *arXiv preprint arXiv:2305.13167*, 2023.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[11] 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.

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[13] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129, 2023.

[14] Chin Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003.

[15] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022.

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[17] Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H Li. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. *arXiv preprint arXiv:2301.11116*, 2023.

[18] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019.

[19] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European Conference on Computer Vision*, pages 319–335. Springer, 2022.

[20] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023.

[21] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint arXiv:2305.15023*, 2023.

[22] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[23] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[24] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[25] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[26] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5073–5084, 2023.

[27] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[28] OpenAI. Gpt-4 technical report, 2023.

[29] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Bleu: a method for automatic evaluation of machine translation. 2002.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[33] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6914–6924, 2023.

[34] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022.

[35] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021.

[36] 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, 2023.

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[38] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[39] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022.

[40] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023.

[41] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2021.

[42] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212, 2016.

[43] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016.

[44] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[45] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023.

[46] Bang Yang, Tong Zhang, and Yuexian Zou. Clip meets video captioning: Concept-aware representation learning does matter. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 368–381. Springer, 2022.

[47] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. *arXiv preprint arXiv:2212.14546*, 2022.

[48] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018.

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

[50] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020.

[51] Xian Zhong, Zipeng Li, Shuqin Chen, Kui Jiang, Chen Chen, and Mang Ye. Refined semantic enhancement towards frequency diffusion for video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3724–3732, 2023.

[52] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language

understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.