BT-Adapter: Video Conversation is Feasible Without Video Instruction Tuning

Supplementary Material

6. Datasets

Test-to-Video Retrieval. The settings of the four zero-shot retrieval benchmarks are presented as follows: (1) **MSRVTT** [39], a widely used video-text retrieval benchmark, comprises 10,000 YouTube videos, each accompanied by 20 captions. Our reported results are based on the 1K-A split, which consists of 9,000 training videos and 1,000 testing videos. For MSRVTT, we sample 12 frames for each video and set max token length as 12. (2) **DiDemo** [1] comprises 10,611 videos gathered from Flickr, along with 40,000 sentences. To form queries, we concatenate all captions associated with a video. We use a frame number of 64 and a mask token length of 64, consistent with prior research. (3) **LSMDC** [30] consists of 118,081 videos extracted from 202 movies. We configure it with a frame number of 12 and a maximum token length of 32. (4) **ActivityNet** [4] comprises 20,000 YouTube videos. To create queries, we concatenate all video descriptions into paragraphs. Our evaluation focuses on video-paragraph retrieval using the 'val1' split. We set the frame number and maximum token length to 64.

Action Recognition. In all video recognition datasets, we refrain from utilizing templates such as “a video of **” and instead employ the tag itself as the textual query. The parameters for frame number and maximum token length are consistently configured at 16 and 12 respectively. The statistics pertaining to the three zero-shot action recognition benchmarks are provided below: (1) **Kinetics-400** [5] is a widely recognized dataset for video action recognition. It comprises a substantial collection of 260,000 videos, each with an average duration of approximately 300 frames. The dataset encompasses a diverse set of 400 action classes. (2) **HMDB-51** [13] includes a total of 5,000 videos spanning 51 distinct action categories. The dataset is partitioned into training and test sets, with 3,500 videos allocated for training and 1,500 videos for testing. (3) **UCF-101** [31] comprises a comprehensive collection of 13,000 videos, representing 101 unique action categories. Within this dataset, the training set consists of 9,500 videos, while the test set contains 3,500 videos.

Video-Text pretraining. We adopt the **WebVid2M** [2] for pretraining, laying the foundation for the BT-Adapter's video encoding capabilities. WebVid2M is a substantial video-text pretraining dataset composed of short videos paired with textual descriptions, sourced from stock footage sites. This dataset is characterized by its vast scale, encompassing approximately 2.5 million video-caption pairs and totaling 12,000 video hours. The videos within WebVid2M exhibit a rich diversity of content. During the pretraining, we configured the frame number and maximum token length to be 8 and 32 respectively.

Video Conversation. VideoChatGPT benchmark [23] s the first benchmark designed for the quantitative evaluation of video conversation models. It was collaboratively annotated by ChatGPT and human annotators using the ActivityNet dataset, resulting in a dataset containing 100k video-text instruction pairs. For the video-based text generation benchmark, a test set was curated based on ActivityNet, which included captions and associated question-answer pairs obtained from human annotations. The evaluation pipeline used the GPT-3.5 model and assessed the model's performance in various aspects, including Correctness of Information, Detail Orientation, Contextual Understanding, Temporal Understanding, and Consistency. The pipeline assigns a relative score to the generated predictions on a scale of 1 to 5 for each of these aspects. For zero-shot question-answer evaluation, three open-source video QA datasets were employed: MSRVTT-QA, MSVD-QA, and ActivityNet-QA. Also, GPT was used as the zero-shot evaluation assistor to assign relative scores on a scale of 1 to 5 for generated answers.

7. Implementation Details

All experiments were conducted using PyTorch [27]. The pretraining and zero-shot inference processes were implemented based on mmaction2.0 [6]. Our configuration settings are detailed in Table 8, with the exception of specific cases where alternate configurations were used. It is noteworthy that our data augmentation techniques are notably simpler in comparison to those employed by other methods.

8. Zero-Shot Results on Action Recognition

The results of zero-shot video recognition are reported in Table 9. Despite being pretrained solely on video-language datasets, BT-Adapter consistently contributes to the video-only task, achieving state-of-the-art zero-shot results among CLIP-based methods. Notably, even when compared to InternVideo, which employed self-supervised reconstruction during pretraining (proven to be more effective on single-modality tasks than contrastive learning), BT-Adapter still outperforms it, underscoring the effectiveness of BT-Adapter in video encoding and spatial-temporal modeling.

Table 8. Default implementation details for pretraining and instruction tuning.

Task	Video-Text Pretraining	Video Instruction Tuning
num. BT-Adapter layers	4	3
num. CLIP layers	24	23
optimizer	AdamW, $\beta = (0.9, 0.98)$	AdamW, $\beta = (0.9, 0.998)$
weight decay	0.05	0.1
learning rate	2e-6 (for BT-Adapter)	2e-5 (for BT-Adapter and linear projection)
fp16	✗	✓
batch size	640	4
augmentation	RandomResizedCrop	CenterCrop
training source	8 V100-32G	4 A100-40G

Table 9. The zero-shot results of video recognition on HMDB, UCF, and K400.

Method	HMDB-51		UCF-101		K400	
	A@1	A@5	A@1	A@5	A@1	A@5
JigsawNet [28]	38.7	-	56.0	-	45.9	-
CLIP [29]	45.0	74.4	73.5	92.7	59.1	82.8
X-Florence [24]	48.4	-	73.2	-	-	-
InternVideo [38]	-	-	-	-	64.2	-
TVTSv2 [46]	52.1	-	78.0	-	59.6	-
BT-Adapter	**54.6**	**79.7**	**79.1**	**96.2**	**64.3**	**86.7**

9. Experimental Comparison With Similar Methods

In this section, we conduct an experimental comparison between two closely related works, ST-Adapter [26] and STAN [20]. ST-Adapter is also notably recognized as a parameter-efficient method for temporal modeling, while STAN also employs the branching temporal modeling strategy. We pretrain the three methods on MSRVTT for one epoch first, and the results of zero-shot performance on MSRVTT retrieval and video conversation are presented in Table 10. Initially, it is evident that ST-Adapter exhibits suboptimal results across all metrics. This outcome may be attributed to the fact that ST-Adapter is a single-modality temporal adapter, where the insertion of 3-D convolutions between transformer layers may lead to the rapid degradation of the pretrained multimodal knowledge. Next, we assess STAN under two conditions: with frozen CLIP and without. The results reveal that STAN, when used with an open CLIP, performs admirably in zero-shot retrieval tasks. However, it exhibits poorer outcomes in video conversation tasks, and it requires significantly longer pretraining hours. Conversely, when STAN is employed with a frozen CLIP, it shows improvements across all metrics, although it still falls short of BT-Adapter in all aspects. In contrast, BT-Adapter achieves both efficiency and effectiveness simultaneously, underscoring the superiority of our design over ST-Adapter and STAN in the context of zero-shot video encoding and video conversation.

10. More Ablation Results

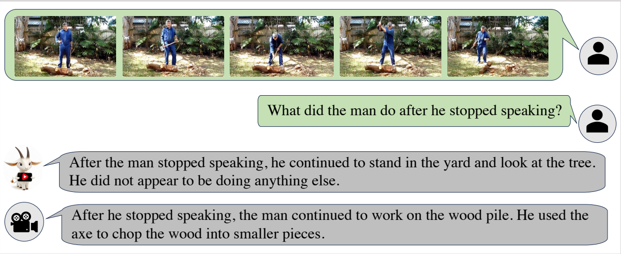
Temporal Projection and Initialization. We examine the appropriate way for instantiating the temporal projection. As demonstrated in Table 11(above), random initialization for the projection yields performance results similar to those obtained without projection. In contrast, zero initialization outperforms them by a significant margin. This suggests that building temporal reasoning capability from scratch, as opposed to random initialization, mitigates adverse effects on the well-established spatial prior. Consequently, zero initialization is better suited for knowledge transfer from images to videos. **Backbone-Branch Combination.** We further perform an ablation study to explore the most effective method for combining the output from the backbone and the branch, considering three approaches: direct addition, weighted selection, and concatenation with subsequent linear projection. As illustrated in Table 11(below), weighted selection yields the most favorable results. This observation suggests that different layers and samples require distinct degrees of information from the backbone and the branch.

Table 10. The experimental comparison with closely related works.

Method	MSR-VTT R@1	Correctness	Temporal	GPU Hours
CLIP(baseline)	35.4	2.06	1.78	-
ST-Adapter	33.6	1.52	1.71	2.5
STAN(open)	40.5	1.84	1.77	22
STAN(frozen)	38.1	2.07	1.92	3
BT-Adapter	40.9	2.16	2.13	3

Table 11. Ablation Studies on Temporal Projection and Backbone-Branch Combination.

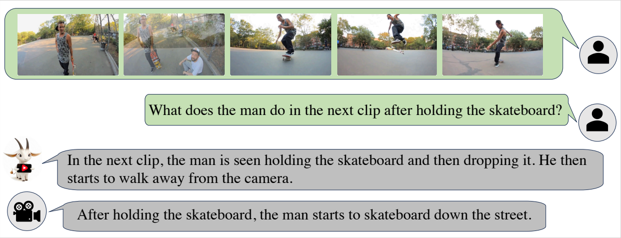
Method	MSR-VTT R@1	DiDemo R@1
None Projection	39.1	33.5
Random Initialization	39.3	34.0
Zero Initialization	**40.9**	**35.6**
Addition	40.0	34.4
Weighted Selection	**40.9**	**35.6**
Concatenation	39.8	34.5


What did the man do after he stopped speaking?

After the man stopped speaking, he continued to stand in the yard and look at the tree. He did not appear to be doing anything else.

After he stopped speaking, the man continued to work on the wood pile. He used the axe to chop the wood into smaller pieces.

Figure 6. Qualitative results of video conversation in terms of the sequence of actions in the video. We present the answers from VideoChatGPT (upper) and our BT-Adapter-LLaVA (down).


What does the man do in the next clip after holding the skateboard?

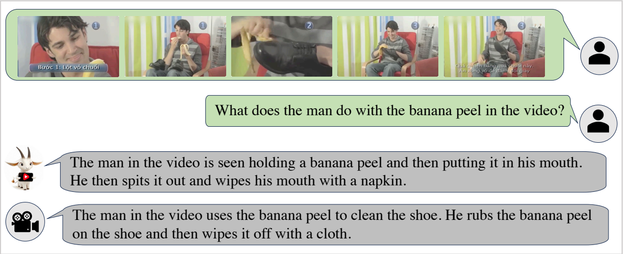
In the next clip, the man is seen holding the skateboard and then dropping it. He then starts to walk away from the camera.

After holding the skateboard, the man starts to skateboard down the street.

Figure 7. Qualitative results of video conversation in terms of actions in a specific frame of the video. We present the answers from VideoChatGPT (upper) and our BT-Adapter-LLaVA (down).

11. More Qualitative Results

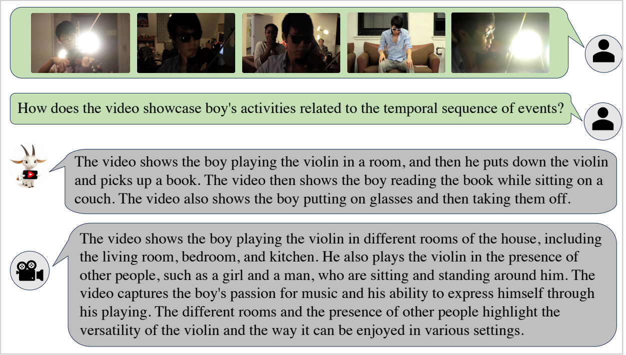
In Figures 6, 7, 8, and 9, we present a comprehensive overview of the qualitative results obtained in video dialogues, encompassing diverse aspects. These visualizations vividly illustrate the capacity of our BT-Adapter to provide contextually appropriate responses in a variety of scenarios where temporal sensitivity is paramount. These results serve to underscore the efficacy of the BT-Adapter in video understanding. In Figure 10, we present a notable outlier case in which our method encounters challenges, where the BT-Adapter struggles to recognize the text content within the frames. This particular instance sheds light on the fact that, while the BT-Adapter diligently strives to preserve pretraining information to the greatest extent possible, it may still introduce some disruption to the pretraining knowledge compared to the fully concatenation-based modeling of VideoChatGPT.


What does the man do with the banana peel in the video?

The man in the video is seen holding a banana peel and then putting it in his mouth. He then spits it out and wipes his mouth with a napkin.

The man in the video uses the banana peel to clean the shoe. He rubs the banana peel on the shoe and then wipes it off with a cloth.

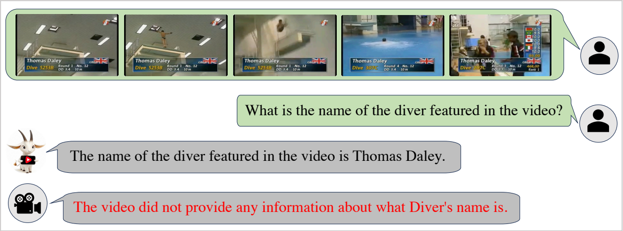
Figure 8. Qualitative results of video conversation in terms of unusual actions in the video. We present the answers from VideoChatGPT (upper) and our BT-Adapter-LLaVA (down).


How does the video showcase boy's activities related to the temporal sequence of events?

The video shows the boy playing the violin in a room, and then he puts down the violin and picks up a book. The video then shows the boy reading the book while sitting on a couch. The video also shows the boy putting on glasses and then taking them off.

The video shows the boy playing the violin in different rooms of the house, including the living room, bedroom, and kitchen. He also plays the violin in the presence of other people, such as a girl and a man, who are sitting and standing around him. The video captures the boy's passion for music and his ability to express himself through his playing. The different rooms and the presence of other people highlight the versatility of the violin and the way it can be enjoyed in various settings.

Figure 9. Qualitative results of video conversation in terms of complex actions and scenes in a long video (3 min). We present the answers from VideoChatGPT (upper) and our BT-Adapter-LLaVA (down).



What is the name of the diver featured in the video?

The name of the diver featured in the video is Thomas Daley.

The video did not provide any information about what Diver's name is.

Figure 10. The bad case of our method in video conversation. We present the answers from VideoChatGPT (upper) and our BT-Adapter-LLaVA (down).

12. Broader Impact

The research presented in this paper, which leverages Large Language Models (LLMs), comes with several important considerations for its broader impact. The use of LLMs to generate content, while powerful, can inherit biases from the data used for training, potentially resulting in content that reflects these biases, some of which may have negative societal implications. Moreover, the model may generate inaccurate or non-factual content, which can undermine trust in online information sources.