# MotionBoost: Bootstrapping Image-Language Models with Motion Awareness for Efficient Video Understanding

Anonymous ACL submission

## Abstract

We present a novel fine-tuning framework that improves the motion sensitivity and length adaptability of Vision-Language Pretraining Models (VLPs), which are currently constrained by their dependence on static images or fixed-length video segments due to data and computational limits. Our framework introduces two main components: the Temporal Prompt Sampler (TPS), which uses optical flow to selectively sample video content based on motion, and the Spatial Prompt Solver (SPS), which accurately captures the complex spatial interplay between visual and textual elements. We further propose a self-boost training approach to harmonize TPS and SPS. Our framework's effectiveness is validated through rigorous testing on various advanced videoQA tasks and a temporal question grounding task, showing marked improvements in performance, efficiency, and generality across various VLPs and large language models (LLMs).

## 1 Introduction

Existing methods in video-langauge modeling have been greatly improved by the pertaining technicals and LLMs (Maaz et al., 2023a; Li et al., 2023c; Zhang et al., 2023a; Lin et al., 2023). However, understanding videos with task-oriented linguistic queries still suffers from the significant computational overhead (Buch et al., 2022; Gao et al., 2023a; Yu et al., 2023; Song et al., 2023) imposed by high-dimensional video data and the disparity between language and spatial-temporal visual cues (Lei et al., 2022; Xiao et al., 2023a). To address the computational burden of video processing, research has focused on sampling methods that select only relevant frames to reduce input size (Lei et al., 2021; Wang et al., 2023; Bain et al., 2021; Buch et al., 2022; Gao et al., 2023a). Despite this, these approaches are hindered by low efficiency and slow speeds due to extensive parameters. Achieving a balance between effective spatial-temporal video-language extraction and computational efficiency continues to be a significant challenge, especially for advanced and long videos.

Drawing upon the insights, we introduce MotionBoost, a general and efficient finetuning framework capable of integrating temporal priors into LLMs for a range of Video-language understanding tasks. As illustrated in fig. 1, our framework comprises a TPS to bootstrap information from temporal priors, and a SPS to grasp spatial visual-text cues. The primary advantages that differentiate MotionBoost from prior arts can be outlined as follows:

**Computationally efficient and effective** Our lightweight TPS effectively extracts keyframes from video using language queries without extra pre-trained models, optimizing both efficacy and efficiency in video-language understanding.

**Temporally extrapolated** We enhance the TPS's flexibility and scalability by incorporating RoPE (Su et al., 2021), which encodes absolute positions and relative dependencies in cross-attention. Our adaptation applies RoPE to both visual and language embeddings, enabling our sampler to handle long videos efficiently.

**Collaborative Spatial-Temporal Self-Boost** In MotionBoost, TPS and SPS mutually enhance performance. TPS selects keyframes for SPS, which uses advanced tools for spatial-textual analysis. A self-boost loop connects them, and Gumbel-Softmax bridges the gap for joint fine-tuning, synergizing LLM, SPS, and TPS effectively and efficiently without additional annotation.

## 2 The MotionBoost Framework

The open-ended video-language understanding task involves analyzing a video, represented as a sequence of frames $V = \{fr_1, fr_2, \cdots, fr_T\}$, and a language prompt $L$ consisting of $N$ tokens, to identify keyframes relevant to the prompt and gen-

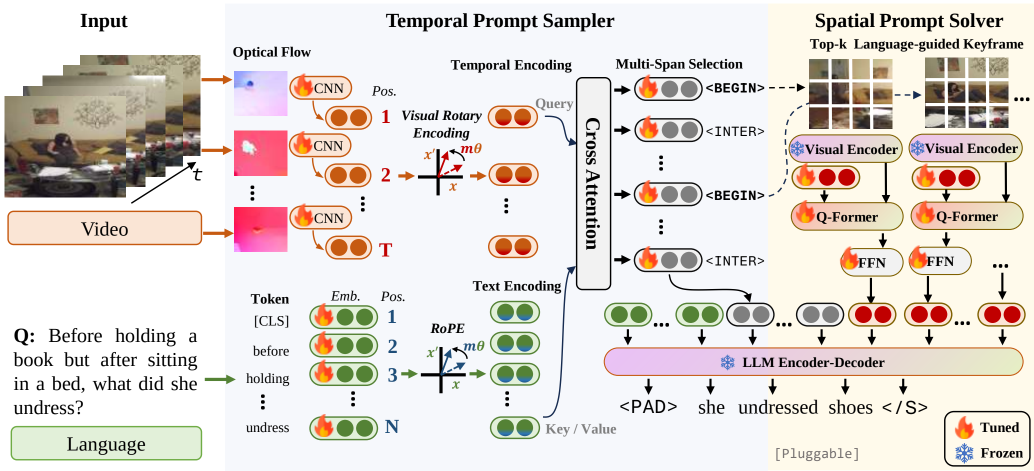



Figure 1: **Overview of MotionBoost framework.** The TPS is designed to capture temporal priors and specific moments. The SPS bridges the gap between the sampled frames and language. A collaborative spatial-temporal self-boost algorithm is devised to incorporate spatial-temporal-language alignment.

erate a natural language response $y$. Trainable parameters or neural networks are denoted by $f(\cdot)$, while $f(\cdot)$ represents frozen pre-trained models.

**Temporal Prompt Sampler** We introduce a TPS that encodes video-text temporal features more effectively using optical flows (OFs) than traditional offline encoders. Optical flows capture frame-to-frame motion and are processed by a compact CNN and an MLP for visual data, while language inputs are handled by a trainable embedding layer, denoted as $E_{of} = \mathrm{MLP}(\mathrm{CNN}(of))$. To manage long inputs in Transformer models, we use RoPE (Su et al., 2021) for positional encoding of both OF and language tokens, represented as $E^R_{of} = RoPE(W_{of} RoPE(W_{of} E_{of}, Pos_{of})$, where $W_{of}, W_l$ are transformation matrices and $Pos_{of}, Pos_l$ are position indices. Cross-attention is applied to these features to create language-informed temporal features. We formulate temporal question grounding as a multi-span reading comprehension task, employing an RC head to pinpoint keyframe spans and optimizing with cross-entropy, as explained in Appendix D.1. Our approach allows for the extraction of multiple video segments efficiently during inference, with low time and space complexity.

**Spatial Prompt Solver** For each keyframe $fr_k$, we capture spatial information using a pre-trained visual encoder: $E_{fr} = Enc_v(fr_k)$. We then adapt these features with a pre-trained Q-former (Li et al., 2022a) to generate query representations $\tilde{E}_q = Enc_q(E_q, E_{fr})$, where $E_q$ is a learnable query and $\tilde{E}_q$ is the output of the SPS. The final output $y$ is obtained by inputting spatial-temporal-language information into a frozen LLM: $y = \mathrm{LLM}(E_r, \tilde{E}_q, E_l)$. The SPS is pluggable and could be replaced with any VLPs.

**Collaborative Spatial-Temporal Self-Boost Algorithm** We create a self-boost algorithm to boost TPS performance using the capabilities of the SPS due to the lack of temporally annotated video-language datasets and the expensive nature of human labeling. Our algorithm caters to both close-ended and open-ended video-language understanding tasks. For close-ended tasks, we use an iterative SPS-based evaluation of video frames, labeling frames with correct SPS predictions as positive and incorrect ones as negative. For open-ended tasks, we analyze SPS results of sampled frames, comparing them with ground truth using sentence semantic similarity score, and employing a monotonic stack algorithm to find the span with the highest similarity for pseudo labeling. More details are available in Appendix A. Furthermore, The lightweight TPS's ability in localizing keyframes is improved by proposing a joint optimization technique using Gumbel-Softmax, which samples key spans and connects temporal samplers with spatial solvers. This approach enhances spatial-temporal grounding by combining large language models, visual feature extraction, and optical flow insights.

## 3 Experiments

In this section, we utilize the MotionBoost on a variety of VLPs and advanced VidL tasks. You can find all the experiment setups, baselines, implementation details in Appendix D.

| Model | Object-relation | Relation-action | Object-action | Superlative | Sequencing | Exists | Duration comparison | Action recognition | Overall |
|---|---|---|---|---|---|---|---|---|---|
| *Retrieval-based Video-Language Models* | | | | | | | | | |
| HME (Fan et al., 2019) | 37.42 | 49.90 | 49.97 | 33.21 | 49.77 | 49.96 | 47.03 | 5.43 | 39.89 |
| PSAC (Li et al., 2019) | 37.84 | 49.95 | 50.00 | 33.20 | 49.78 | 49.94 | 45.21 | 4.14 | 40.18 |
| HCRN (Le et al., 2020) | 40.33 | 49.86 | 49.85 | 33.55 | 49.70 | 50.01 | 43.84 | 5.52 | 42.11 |
| AIO (Wang et al., 2023) | 48.34 | 48.99 | 49.66 | 37.53 | 49.61 | 50.81 | 45.36 | 18.97 | 48.59 |
| ATP (Buch et al., 2022) | 50.15 | 49.76 | 46.25 | 39.78 | 48.25 | 51.79 | 49.59 | 18.96 | 49.79 |
| ALBEF (Li et al., 2021) | 50.53 | 49.39 | 49.97 | 38.22 | 49.79 | 54.11 | 48.01 | 10.40 | 50.68 |
| SINGULARITY (Lei et al., 2022) | 50.87 | 50.67 | 49.70 | 40.47 | 40.79 | 55.34 | 48.20 | 11.59 | 51.11 |
| VIOLET (Fu et al., 2021) | 50.89 | 50.24 | 50.93 | 40.76 | 50.51 | 58.07 | 38.97 | 6.53 | 51.03 |
| MIST-AIO (Gao et al., 2023a) | 51.43 | 54.67 | 55.37 | 41.34 | 53.14 | 53.49 | 47.48 | **20.18** | 50.96 |
| MIST-CLIP (Gao et al., 2023a) | 51.68 | **67.18** | **68.99** | 42.05 | **67.24** | 60.33 | **54.62** | 19.69 | 54.39 |
| *Open-ended Video-Language Models* | | | | | | | | | |
| SeViLA* (Yu et al., 2023) | 51.15 | 48.93 | 62.08 | 42.24 | 55.96 | 53.02 | 38.91 | 0.00 | 51.70 |
| BLIP2 (Li et al., 2023b) | 53.72 | 48.64 | 62.1 | 43.84 | 55.94 | 55.14 | 40.39 | 0.28 | 54.00 |
| TPS + ALBEF (Li et al., 2021) | 51.05 | 51.11 | 51.66 | 38.36 | 51.33 | 58.10 | 49.20 | 11.78 | 51.73 |
| TPS + VIOLET (Fu et al., 2021) | 51.59 | 54.54 | 56.96 | 40.94 | 55.61 | 59.12 | 42.81 | 9.02 | 52.59 |
| TPS + SINGULARITY (Lei et al., 2022) | 52.33 | 54.12 | 55.07 | 40.71 | 54.49 | 57.88 | 48.35 | 12.24 | 53.13 |
| MotionBoost (Ours, BLIP2-based) | **62.27** | 51.74 | 66.09 | **53.67** | 60.11 | **60.85** | 36.99 | 0.00 | **61.45** |

* Re-implementation result. We removed prior information from QVHighlights (Lei et al.) used in (Yu et al., 2023) for fair comparison.

Table 1: **Comparison accuracy of different sampling-based SOTA models on AGQA 2.0**.

| Model | Temporal | Causal | Description | All |
|---|---|---|---|---|
| CLIP (Radford et al., 2021a) | 46.3 | 39.0 | 53.1 | 43.7 |
| HGA (Jiang and Han, 2020) | 44.2 | 52.5 | 44.1 | 49.7 |
| AIO (Wang et al., 2023) | 48.0 | 48.6 | 63.2 | 50.6 |
| VQA-T (Yang et al., 2021) | 49.6 | 51.5 | 63.2 | 52.3 |
| MIST-AIO (Gao et al., 2023a) | 51.6 | 51.5 | 64.2 | 53.5 |
| ATP (Buch et al., 2022) | 50.2 | 53.1 | 66.8 | 54.3 |
| VGT (Xiao et al., 2022) | 52.3 | 55.1 | 64.1 | 55.0 |
| MIST-CLIP (Gao et al., 2023a) | 56.6 | 54.6 | 66.9 | 57.1 |
| BLIP2 (Li et al., 2023b) | 64.9 | 69.7 | 79.4 | 69.6 |
| SeViLA* (Yu et al., 2023) | 66.4 | 71.9 | 80.8 | 71.5 |
| MotionBoost (Ours, BLIP2-based) | **66.5** | **72.8** | **81.2** | **72.1** |

* Re-implementation result. We removed prior information from QVHighlights (Lei et al.) used in (Yu et al., 2023) for fair comparison.

Table 2: **Comparison accuracy of long-form video QA on NExT-QA** (Xiao et al., 2021).

### 3.1 Complicated Video Question Answering

**Results on AGQA 2.0 (Grunde-McLaughlin et al., 2021)** The MotionBoost framework marginally improves BLIP2's performance in video-language tasks, but it still falls short of MIST-CLIP. Enhancements from MotionBoost increase BLIP2's accuracy by 7.45 points, indicating better spatial-temporal feature learning. However, BLIP2 struggles with certain question types, such as"Activity Recognition," This difficulty arises from the reliance on an unsuitable evaluation method, namely, the requirement for exact matches between the generative model's outputs and a predefined set of answer vocabulary.

**Results on NExTQA (Xiao et al., 2021)** Table 2 presents the results on the NExTQA dataset. Our method surpasses various baseline models, including the recent SeViLA model that utilizes LLM for keyframe selection. The lesser performance gain on NExTQA over AGQA is attributed to its focus on causality and the inherent "static appearance bias" (Lei et al., 2022) in its source videos from the VidOR dataset (Shang et al., 2019).

**Analysis** Our study evaluated the impact of TPS on various VLPs by comparing them with different frame sampling methods, excluding optical flow features. For VLPs that use a single image, we combined multiple images through early fusion. Results on the AGQA 2.0 dataset showed that TPS significantly improves VLPs' performance on temporal questions, such as "`Relation-action`," "`Sequencing`, and "`Exists` ", over uniform sampling. However, the lack of temporal priors limits ensemble methods' effectiveness, with SINGULARITY outperforming ALBEF due to its video corpus pre-training. While TPS-augmented models show limited improvement on "`Superlative` " questions, integrating optical flow into our BLIP2-based framework resulted in a 22.42% performance increase, demonstrating that optical flow can mitigate the temporal information loss from frame sampling. In addition, We replaced BLIP2-based SPS with different types of VLPs, excluding optical flow input, and tested on AGQA 2.0. Results show a 3.68% accuracy increase using keyframes over uniform frames, proving our model's effectiveness with various VLPs. For the effectiveness of our components, refer to Appendix C.1.

### 3.2 Temporal Question Grounding on Video

The results on NExTGQA (Xiao et al., 2023a) are shown in table 3, our method outperforms baselines using additional feature extractors (Ren et al., 2015; Liu et al., 2021c,b; Radford et al., 2021a). Our TPS with OF improves temporal learning for video-language tasks, reducing the irrelevant visual noise from discrete frames. Current methods show weak temporal grounding (mIoU < 0.20), but our TPS's

| Method | Vision Encoder | mIoU | IoU@0.3 | IoU@0.5 |
|---|---|---|---|---|
| VGT | RCNN | 3.0 | 4.2 | 1.4 |
| VIOLETv2 | VSWT | 3.1 | 4.3 | 1.3 |
| Temp[Swin] | SWT | 4.9 | 6.6 | 2.3 |
| Temp[CLIP] | ViT-B | 6.1 | 8.3 | 3.7 |
| Temp[BLIP] | ViT-B | 6.9 | 10.0 | 4.5 |
| FrozenBiLM | ViT-L | 7.1 | 10.0 | 4.4 |
| IGV | ResNet | 14.0 | 19.8 | 9.6 |
| SeViLA* | ViT-G | 21.7 | 29.2 | 13.8 |
| MotionBoost (BLIP2-based) | OF+CNN | 19.9 | 23.3 | 11.2 |

* pre-trained on QVHighlights (Lei et al.).

Table 3: **Comparison results of Temporal Question Grounding task on NExT-GQA (Xiao et al., 2023b).**

| Methods | LLM size | MSVD-QA Accuracy | MSVD-QA Score | MSRVTT-QA Accuracy | MSRVTT-QA Score | ActivityNet-QA Accuracy | ActivityNet-QA Score |
|---|---|---|---|---|---|---|---|
| FrozenBiLM | 1B | 32.2 | - | 16.8 | - | 24.7 | - |
| VideoChat | 7B | 56.3 | 2.8 | 45.0 | 2.5 | - | 2.2 |
| LLaMA-Adapter | 7B | 54.9 | 3.1 | 43.8 | 2.7 | 34.2 | 2.7 |
| Video-LLaMA | 7B | 51.6 | 2.5 | 29.6 | 1.8 | 12.4 | 1.1 |
| Video-ChatGPT | 7B | 64.9 | 3.3 | 49.3 | 2.8 | 35.2 | 2.7 |
| Video-LLaVA | 7B | 70.7 | 3.9 | 59.2 | 3.5 | 45.3 | 3.3 |
| MotionBoost (Vicuna-7b-based) | 7B | 71.4 | 3.9 | 57.3 | 3.3 | 43.9 | 3.3 |

Table 4: **Zero-shot Open Domain Video QA.**

| Methods | Base Model | # of Frames | Accuracy |
|---|---|---|---|
| Video-LLaVA | LLaVA-7b | 8 | 36.8 |
| Sevila | BLIP2 | 32 | 25.7 |
| MotionBoost (BLIP2) | BLIP2 | 4 | 41.2 |
| MotionBoost (BLIP2) | BLIP2 | 8 | 41.4 |
| MotionBoost (BLIP2) | BLIP2 | 32 | 42.8 |

Table 5: **Zero-shot Result on subset of EgoSchema**

features could close this gap in spatial-temporal research. For qualitative results, refer to Appendix E.

### 3.3 Generality of MotionBoost

To illustrate the generality of our approach, we implemented our model on visual instruction datasets, namely VideoChatGPT (Maaz et al., 2023a) and LLava-1.5K (Liu et al., 2023a). Additionally, we change the LLM to the Vicuna-7b (Chiang et al., 2023) for an equitable comparison with the latest SOTA techniques. Table 4 displays our model's performance on the videoQA dataset in a zero-shot scenario. In contrast to VideoLLaVA, our model was solely fine-tuned on these visual instruction datasets, without any pretraining on extra datasets. The outcomes affirm that our method rivals the performance of the most recent SOTA MLLMs, despite our model's LLM being static and not pretrained on video-specific corpora. This underscores the significant potential and broad applicability of our framework within this field.

### 3.4 Length Extrapolation of MotionBoost

In this section, we will assess MotionBoost's capabilities in long video language understanding

| Model | FLOPs (GFLOPs) ↓ | MACs (GMACs) ↓ | Acc. ↑ |
|---|---|---|---|
| BLIP2 (ViT-G) | 2,705 | 1,350 | 69.6 |
| Sevila (ViT-G) | 13,720 | 14,357 | 71.5 |
| MotionBoost (ViT-G, BLIP2-based) | 19,620 | 9,840 | 72.3 |
| MotionBoost (OFs, BLIP2-based) | 2,950 | 1,474 | 72.1 |

Table 6: **Computational Efficiency of MotionBoost.**

tasks. We evaluate the model's performance on EgoSchema(Mangalam et al., 2023), which is one of the longest videoQA datasets available. As depicted in table 5, MotionBoost exhibits a robust understanding of long videos. Moreover, although MotionBoost is trained on sequences of 4 frames, it is evaluated on varying lengths during the testing phase. The consistently improved results suggest that our method possesses a strong capacity for length extrapolation.

### 3.5 Time Efficiency

We evaluated the average inference time efficiency of our method against BLIP2 using calflops (xiaoju ye, 2023) on the NExT-QA dataset, as shown in Table 6. Our method outperformed the current SOTA model SeViLa, which uses the LLM to select keyframes, both in performance and efficiency. While replacing the OFs with features from ViT-G (Zhai et al., 2021) resulted in minor improvements, it significantly increased computation costs due to the offline feature extractor. Compared to BLIP2, our method required minimal additional computation. The major computation costs were associated with the LLMs from BLIP2 and the offline feature extractor. We believe our method strikes a balance between being effective and efficient. Further details on the composition of inference time of MotionBoost are provided in *SM*. In addition, we investigate the composition of inference time of MotionBoost and offline demo in Appendix B.

## 4 Conclusion

In this work, we propose an efficient pluggable framework MotionBoost for advanced video-language understanding tasks, which comprises a temporal prompt sampler and a spatial prompt solver to combine spatial-temporal-language alignment and temporal grounding. Experiments on advanced video question answering and temporal question grounding on video demonstrate a consistent improvement over various types of VLPs. Comprehensive analysis verifies the effectiveness, efficiency, and generality of our framework.

## 5 Limitations

Our study has one primary limitation: *i.e.* **Limited Temporal Grounding Capability** As shown in section 3.2, our method outperforms existing approaches but still has restricted temporal grounding capabilities, a common issue in current research. We suspect that this limitation may be due to the constraints of the lightweight 6-layer transformer-based TPS. In future work, we aim to enhance this aspect of our method without sacrificing efficiency.

## References


Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. *International Conference on Computer Vision (ICCV)*, pages 1708–1718.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. 2022. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

S. Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the "video" in video-language understanding. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2907–2917.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *ArXiv*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1999–2007.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. Violet : End-to-end video-language transformers with masked visual-token modeling. *ArXiv*, abs/2111.12681.

Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. 2021a. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. *International Conference on Computer Vision (ICCV)*, pages 1655–1665.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023a. MIST : Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14773–14783. IEEE.

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. 2023b. Llama-adapter v2: Parameter-efficient visual instruction model.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021b. Making pre-trained language models better few-shot learners. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3816–3830. Association for Computational Linguistics.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and

Furu Wei. 2023a. Language is not all you need: Aligning perception with language models.

Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. 2023b. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6565–6574. IEEE.

Y. Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1359–1367.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, volume 13693 of *Lecture Notes in Computer Science*, pages 709–727. Springer.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 11109–11116. AAAI Press.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. 2023. Imagenhub: Standardizing the evaluation of conditional image generation models. *ArXiv*.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9969–9978. Computer Vision Foundation / IEEE.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2022. Revealing single frame bias for video-and-language learning. *ArXiv*, abs/2206.03428.

Jie Lei, Tamara Lee Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7327–7337.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3045–3059. Association for Computational Linguistics.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *ArXiv*, abs/2305.06355.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023d. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023e. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4582–4597. Association for Computational Linguistics.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 8658–8665. AAAI Press.

Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. 2023f. Efficient multimodal fusion via interactive prompting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2604–2613. IEEE.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022b. Invariant grounding for video question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2927. IEEE.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *ArXiv*, abs/2311.10122.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *ArXiv*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021c. Video swin transformer. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:2507–2521.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023a. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *CoRR*, abs/2306.09093.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023b. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023a. Video-chatgpt: Towards detailed video understanding via large vision and language models.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. 2023b. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2463–2473. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. 39:1137–1149.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2023. Moviechat: From dense token to sparse memory for long video understanding. *CoRR*, abs/2307.16449.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 20841–20855. PMLR.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning.

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. *Lecture Notes in Computer Science*, page 402–419.

Andrés Villa, Juan León Alcázar, Motasem Alfarra, Kumail Alhamoud, Julio Hurtado, Fabian Caba Heilbron, Alvaro Soto, and Bernard Ghanem. 2023. PIVOT: prompting for video continual learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24214–24223. IEEE.

Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. All in one:

Exploring unified video-language pre-training. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 139–149. IEEE.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 2021. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS Datasets and Benchmarks*.

Chen Henry Wu, Saman Motamed, Shaunak Srivastava, and Fernando De la Torre. 2022. Generative visual prompt: Unifying distributional control of pre-trained generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9772–9781.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2023a. Can I trust your answer? visually grounded video question answering. *CoRR*, abs/2309.01327.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2023b. Can i trust your answer? visually grounded video question answering. *ArXiv*.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022. Video graph transformer for video question answering. In *European Conference on Computer Vision (ECCV)*, volume 13696 of *Lecture Notes in Computer Science*, pages 39–58. Springer.

xiaoju ye. 2023. calflops: a flops and params calculate tool for neural networks in pytorch framework.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Liqi Yan, Cheng Han, Zenglin Xu, Dongfang Liu, and Qifan Wang. 2023. Prompt learns prompt: Exploring knowledge-aware generative prompt collaboration for video captioning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1622–1630. ijcai.org.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1686–1697.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering.

Zhou Yu, D. Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. *AAAI Conference on Artificial Intelligence (AAAI)*.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2021. Scaling vision transformers. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1204–1213.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2023b. Temporal sentence grounding in videos: A survey and future directions. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, page 1–20.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023c. Magicbrush: A manually annotated dataset for instruction-guided image editing. *ArXiv*.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023. Chatbridge: Bridging modalities with large language model as a language catalyst. *CoRR*, abs/2305.16103.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

# Appendices

**We provide supplementary materials as follows, in addition, we provide the demo and anonymous code in the uploaded zip files.**

## Table of Contents



## A Self-Boost Algorithm

algorithm 1 shows our self-boost algorithm of automatically generating pseudo labels under open-ended settings by the SPS, which is used to optimize the TPS.

**Algorithm 1:** Pseudo Label Algorithm

**Input:** frames ($V = \{fr_1, fr_2, \cdots, fr_T\}$), query ($q$), answer ($a$)
**Output:** temporal grounded span
$score_{best} \leftarrow 0$
$start \leftarrow 0$
$end \leftarrow T - 1$
$stack \leftarrow empty\ list$
$scores \leftarrow empty\ list$
**for** $fr$ *in* $V$ **do**
    $prediction = LLM_{SPS}(fr, q)$
    $scores.add(SIM(prediction, a))$
**end**
**for** $i$ *in* $scores.length$ **do**
    **while** $stack$ *is not empty and* $stack.get(score.top) > score.get(i)$ **do**
        $tmp = stack.pop()$
        $score_{tmp} = (i - stack.top - 1) \times score.get(tmp)$
        **if** $score_{tmp} > score_{best}$ **then**
            $score_{best} = score_{tmp}$
            $start = 0$
            $end = i - 2$
        **else**
        **end**
    **end**
    $stack.push(i)$
**end**

## B Inference Time Analysis

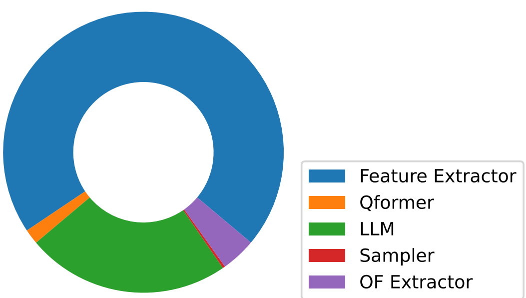



Figure 2: Inference time Analysis

We further investigate the composition of inference time of MotionBoost on the NExT-QA dataset. We find most computation costs come from LLM and the offline feature extractor. Compared with other components, the computation cost is trivial, indicating the strong efficiency of our method. The offline demo is presented in the supplementary material.

| Model | Object-relation | Relation-action | Object-action | Others | All |
|---|---|---|---|---|---|
| MotionBoost | 62.27 | 51.74 | 66.09 | 57.04 | 61.45 |
| w/o optical flow | 59.13 | 15.06 | 50.79 | 51.29 | 55.00 |
| w/ fixed sampler | 62.28 | 47.84 | 50.68 | 53.47 | 59.88 |
| w/ uniform sampler | 53.72 | 48.64 | 62.10 | 50.68 | 54.00 |
| w/ zero-shot | 23.60 | 17.09 | 29.37 | 40.72 | 25.54 |

Table 7: **Ablation study of our method on reasoning questions from AGQA 2.0.** We list the major outputs of complicated relationships and summarize the rest; see *SM* for complete results.

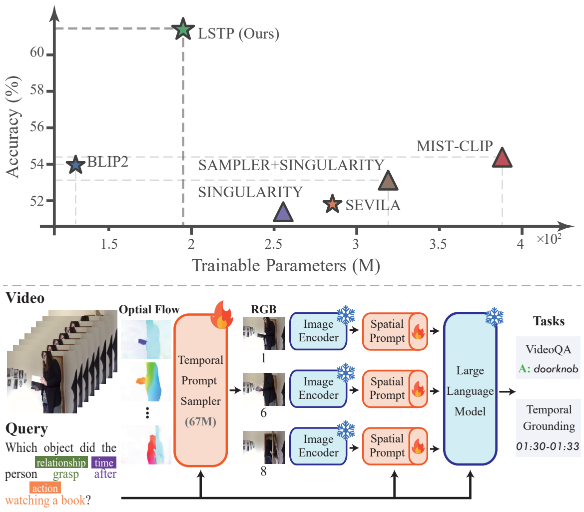



Figure 3: Efficiency Illustration and Task Definition.

# C More Analysis Experiments

## C.1 Ablation Study

We apply ablation study on MotionBoost to investigate the effects of our joint training framework. All the experiments are performed on AGQA 2.0 (Grunde-McLaughlin et al., 2021). As shown in Table 7, the framework incorporating motion feature significantly improved performance by 11.72%, underscoring its effectiveness in tackling spatial-temporal problems. We also found that fixing the pre-trained sampler during training notably affected performance on temporal questions like "`Relation-action` ", suggesting that joint training can further optimize the sampler. Lastly, comparing with zero-shot and fine-tuned BLIP2 (Li et al., 2023b) with uniformly-sampled frames, our method showes significant improvements, demonstrating its overall effectiveness. In Appendix C.2, we provide detailed ablation study about the TPS-augmented models.

## C.2 Ablated TSP-augmented models

| Sampler | Solver | # of frames (Train) | # of frames (Infer.) | Acc. |
|---|---|---|---|---|
| OF | SING-17M | 1 | 6 | 53.13 |
| OF | SING-17M | 1 | 1 | 51.36 |
| OF | SING-17M | 6 | 6 | 53.85 |
| OF | SING-5M | 1 | 6 | 51.10 |
| Swin. | SING-17M | 1 | 6 | 53.76 |

Table 8: Detailed Analysis on the Sampler.

In table 8, we analyzed TSP+SINGULARITY to evaluate the TSP-augmented paradigm. Our study revealed that increasing the number of frames during inference improved performance by 3.4%, but further increases did not proportionally enhance results. We also found that VLP benefits more from the sampling strategy when adequately pretrained (*i.e.*, 17M denotes the model is pretrained on 17M video corpora). Additionally, we proposed two sampler variants, replacing optical flow with features extracted by the video SwinTransformer (Liu et al., 2021c) for pre-training. The comparable results suggest that our TSP can effectively reason over time without any prior perception information.

## C.3 Influence of the number of frames on solver

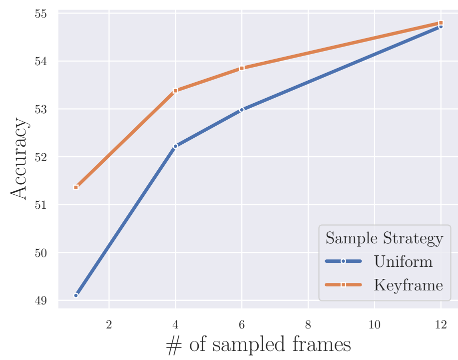



Figure 4: Further study on the number of sampled frames.

We trained the solver with different numbers of sampled frames. Results are shown in Figure 4. The fewer sampled frames the better performance of the keyframe strategy, and after a certain point, the uniform strategy performs close to the keyframe strategy. This is because the average duration of videos in AGQA is around 30 seconds, 12 frames are close to dense sampling which covers almost all visual cues. In other words, video-language tasks require bountiful frame inputs that have high computational complexity, but our method efficiently learns near-complete video information.

### C.4 Detailed Ablation Study Results

| | MotionBoost | w/o Optical Flow | fixed Sampler | Uniform Sample | Zero-Shot |
|---|---|---|---|---|---|
| **Obj-rel** | 62.27 | 59.13 | 62.28 | 53.72 | 23.60 |
| **Rel-act** | 51.74 | 15.06 | 47.84 | 48.64 | 17.09 |
| **Obj-act** | 66.09 | 50.79 | 50.68 | 62.10 | 29.37 |
| **Superlative** | 53.67 | 59.79 | 52.12 | 43.84 | 28.39 |
| **Sequencing** | 60.11 | 35.04 | 49.43 | 55.94 | 48.79 |
| **Exists** | 60.85 | 60.92 | 60.96 | 55.14 | 48.79 |
| **Duration** | 36.99 | 26.48 | 40.18 | 40.39 | 26.99 |
| **Action** | 0.00 | 0.00 | 0.00 | 0.28 | 0.28 |
| **All** | 61.45 | 55.00 | 59.88 | 54.00 | 25.54 |

Table 9: Ablation study of our method on reasoning questions from AGQA 2.0 (Grunde-McLaughlin et al., 2021).

In table 9, we demonstrate the details of the ablation study of MotionBoost on AGQA 2.0. Specifically, we demonstrates the ablation study results of different question types.

## D Implementation Details

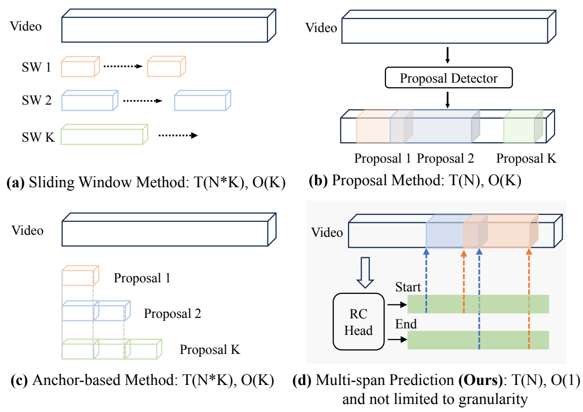
### D.1 Details of Multi-span Prediction

Based on the flow-language encoding, we formulate the temporal question grounding video task as multi-span reading comprehension (RC) problem, where an RC head is to predict the label of fused encoding $\{e_{R1}, e_{R2}, \dots, e_{RT}\}$ as one of {"<BEGIN>", "<END>", "<NONE>"} of the grounded video spans. The selection can be formulated as:

$$h = \mathcal{F}_\theta(e_{R1}, e_{R2}, \dots, e_{RT}), \quad (1)$$
$$index = \arg\max(\text{Softmax}(h)),$$

where $\mathcal{F}_\theta$ denotes the RC head for span selection, $index$ is the prediction of the start or end index. The objective is computed as the cross-entropy between the prediction and pseudo labels.

During Inference, we can obtain an arbitrary number of $K$ segments of grounded video by predicting $K$ <BEGIN> s and $K$ <END> s with the RC Head. Finally, we union these segments to eliminate the overlap between these extracted spans. Appendix D.1 demonstrates commonly used methods for temporal sentence grounding on video tasks (TSGV) (Zhang et al., 2023b). Compared with other span-fixed methods, our method could obtain multiple grounded video spans with the least time complexity and space complexity.



(a) Sliding Window Method: T(N*K), O(K) (b) Proposal Method: T(N), O(K)
(c) Anchor-based Method: T(N*K), O(K) (d) Multi-span Prediction (**Ours**): T(N), O(1) and not limited to granularity

Figure 5: **Comparison of multi-span RC prediction (d)** and other methods (a-c) in terms of time and space complexity.

In fig. 5, we compare our proposed multi-span reading comprehension prediction algorithm and other commonly used methods for temporal sentence grounding on video tasks, including the sliding window method, proposal method, and anchor-based method.

### D.2 Baselines and Setups

**Advanced VideoQA** We take two advanced video question answering (VideoQA) benchmarks AGQA (Grunde-McLaughlin et al., 2021) and NExTQA (Xiao et al., 2021) for evaluation. AGQA is specially designed for compositional spatial-temporal reasoning[1] including 1,455,610/669,207 question answering for train/test splits. NExTQA is a multiple choice VideoQA benchmark for causal, temporal, and descriptive reasoning, including 52K questions. We use two types of baselines: retrieval-based models and open-ended models focusing on recent SOTA temporal priors learning models for comparative analysis. For the retrieval-based models, in addition to traditional methods (Fan et al., 2019; Li et al., 2019; Le et al., 2020; Wang et al., 2023; Li et al., 2021; Lei et al., 2022; Fu et al., 2021), we use recent SOTA temporal learning models, specifically ATP (Buch et al., 2022) and MIST (Gao et al., 2023a). For the open-ended models, we use BLIP2 (Li et al., 2023b) and SEVILA (Yu et al., 2023). For the number of keyframes, we sample 4 frames for MotionBoost and 6 frames for TPS-augmented methods in all experiments. For more implementation details, please refer to Appendix D.3.

**Temporal Question Grounding on Video** We use the Temporal Question Grounding on Video

[1] We use AGQA 2.0 which has more balanced distributions.

(TQGV) dataset NExT-GQA (Xiao et al., 2023a) to evaluate the efficacy of our temporal prompt sampler. NExT-GQA is an extension of NExT-QA (Xiao et al., 2021) with $10.5K$ temporal grounding labels tied to questions, which contains 3,358/5,553 questions for val/test splits. We report mean Intersection over Union (mIoU), IoU@0.3, and IoU@0.5 as metrics following (Xiao et al., 2023a). We select a wide range of VLPs as baselines: VGT (Xiao et al., 2022), Temp (Buch et al., 2022; Xiao et al., 2023b), FrozenBiLM (Yang et al., 2022), IGV (Li et al., 2022b), and SeViLA (Yu et al., 2023). These baseline models encompass a variety of architectures, text encoders, and vision encoders. In contrast, our method does not depend on heavy offline vision feature extractors. We obtain the optical flow using a fixed RAFT (Teed and Deng, 2020), a model with only 5.26 million parameters. This comparison highlights the efficiency and simplicity of our approach.

**Long VideoQA** We take the long videoQA dataset EgoSchema (Mangalam et al., 2023) to evaluate MotionBoost's ability over long video understanding. EgoSchema consists of over 5000 human curated multiple choice question answer pairs with an average video length of 3 minutes. The EgoSchema subset, including 500 question-answer pairs are publicly available. Our experiments are applied on the subset.

### D.3 Implementation Details of MotionBoost on Downstream Tasks

The sampler is a 6-layer transformer with RoPE (Su et al., 2021). For MotionBoost, We use BLIP2-flant5-xl (Li et al., 2023b) as TPS. For the TPS-augmented framework, we take three vison-language pretraining models as the solver: ALBEF (Li et al., 2021), SINGULARITY (Lei et al., 2022), and VIOLET (Fu et al., 2021) For the number of keyframes, we sample 4 frames for MotionBoost and 6 frames for TPS-augmented methods to keep consistent with baselines. We take $K = 2$ for Gumbel-Softmax tricks in practice. We extract the dense optical flow from the video by RAFT (Teed and Deng, 2020). For the BLIP2-based model, the total trainable parameters are 195M, thus our framework is lightweight and can be easily adapted to any LLM. All the experiments are performed on NVIDIA A100 80G GPU. Furthermore, all models on zero-shot setting, including section 3.3 and section 3.4 are fine-tuned on VideoLLaVA(Lin et al., 2023) fine-tuning dataset without any pretraing.

### D.4 Prompt for Multiple-choice Task on BLIP2

Following (Yu et al., 2023), we construct additional prompts to adapt the generative model to the multiple-choice task.

Question: why did the boy pick up one present from the group of them and move to the sofa ?
Option A: share with the girl
Option B: approach lady sitting there
Option C: unwrap it
Option D: playing with toy train
Option E: gesture something
Considering the information presented in the frame, select the correct answer from the options.

Figure 6: Additional prompt for NExT-MC task

## E Qualitative Studies on NExTGQA

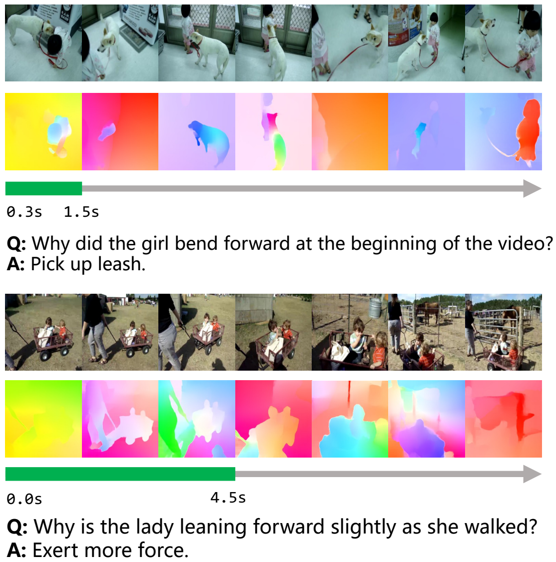


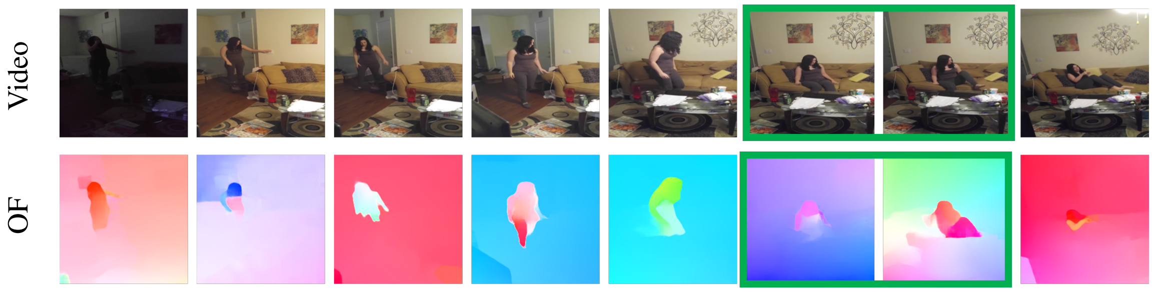

Figure 7: Qualitative results on temporal grounding

fig. 7 presents two random outputs from MotionBoost on the TQGV task. The first example demonstrates how our method can ground video using the semantic information from the question, specifically, the phrase "`at the beginning` ". The second example demonstrates the efficacy of our method in temporal reasoning, as evidenced by the phrase "`as she walked` ".

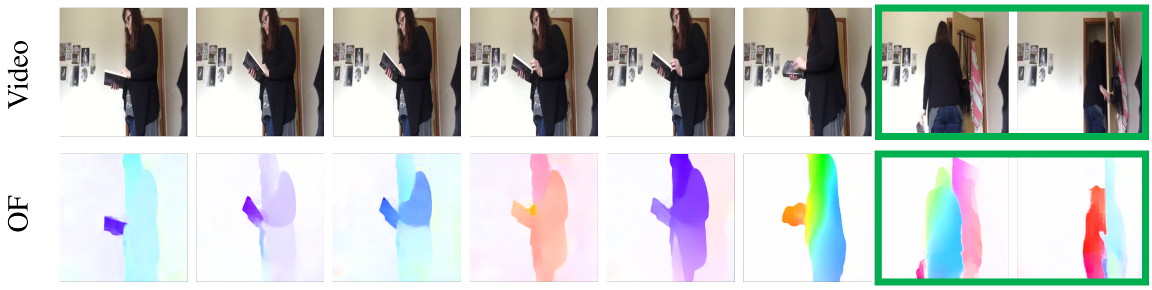
## F Qualitative Studies on AGQA 2.0

## G Related Work

**Long-form Video Question Answering** In the realm of Video Question Answering




**Question:** Before holding a book but after sitting in a bed, what did they undress?
**Ground Truth:** shoe **MotionBoost:** shoe **BLIP2:** dish **SEVILA:** clothes




**Question:** Which object did the person grasp after watching a book?
**Ground Truth:** doorknob **MotionBoost:** doorknob **BLIP2:** NA **SEVILA:** doorway

Figure 8: Case Studies. OF: Optical Flow. Green and red boxes indicate correct and wrong keyframe predictions, respectively. In these cases, our method could correctly localize the keyframes and predict the right answer. “NA" indicates the BLIP2 can’t generate an answer hitting the answer vocabulary.


**Question:** Between putting a book somewhere and tidying something on the floor, which object were they undressing?
**Prediction:** shoe **Ground Truth:** clothes


**Question:** What was the person taking between putting a cup somewhere and holding a book?
**Prediction:** box **Ground Truth:** food

Figure 9: Filure Cases. OF: Optical Flow. Green and red boxes indicate correct and wrong keyframe predictions, respectively. For complicated situations involving more than one event, *e.g.*, “between putting a cup and holding a book", our method could fail to localize the keyframes and thus print the wrong answer.

(VideoQA), traditional datasets such as TGIF-QA (Jang et al., 2017), MSRVTT-QA (?), and ActivityNetQA (Yu et al., 2019) consist of short videos about daily human activities. Notably, Buch et al. (2022); Lei et al. (2022) reveal limitations in common VideoQA benchmarks, failing to mitigate static appearance bias, hindering performance gains from temporal cues. Recent strides introduce intricate spatio-temporal reasoning datasets (Gao et al., 2021a; Grunde-McLaughlin et al., 2021; Wu et al., 2021; Xiao et al., 2021), catalyzing a surge in associated research.

**Visual Prompt Learning** Prompt learning, a label-free approach utilizing language models for text prediction, has shown promise in few-shot and zero-shot learning for NLP tasks (Petroni et al., 2019; Brown et al., 2020; Gao et al., 2021b; Sun et al., 2022). Evolving into prompt tuning, which combines continuous prompts with supervised learning for efficient training (Lester et al., 2021; Li and Liang, 2021; Liu et al., 2021a), this method has extended to image prompts for computer vision (Jia et al., 2022; Wang et al., 2022; Wu et al., 2022; Bar et al., 2022). The integration of vision and language prompts enables low-cost cross-modal alignment, as evidenced by recent studies (Radford et al., 2021b; Zhou et al., 2022; Li et al., 2023f; Huang et al., 2023b). This concept has further expanded to video-language prompts (Villa et al., 2023; Yan et al., 2023), with research integrating LLMs with video data to improve visual tasks like video captioning and question answering, demonstrating the potential of visual prompts in language models for diverse applications (Villa et al., 2023; Li et al., 2023d; Zhao et al., 2023; Maaz et al., 2023b; Lyu et al., 2023a).

**Bootstrapping Large Language Models for Visual Tasks** Capitalizing on the success of LLMs in NLP, there is a growing trend of applying them to computer vision tasks, such as VQA (Lu et al., 2022; Chen et al., 2023; Fu et al., 2023; Liu et al., 2023b; Li et al., 2023a), image generation (Ku et al., 2023; Zhang et al., 2023c), and visual instruction following (Xu et al., 2022; Li et al., 2023e). The research mainly progresses along three avenues: (i) leveraging LLMs' reasoning for visual tasks (Huang et al., 2023a; Wu et al., 2023; Driess et al., 2023; Surís et al., 2023); (ii) adapting Transformer or linear networks to equip LLMs with visual perception (Li et al., 2023b; Dai et al., 2023; Zhu et al., 2023; Xu et al., 2023; Gao et al., 2023b; Liu et al., 2023a); (iii) merging LLMs with video and audio inputs (Zhang et al., 2023a; Maaz et al., 2023a; Lyu et al., 2023b). Recently, Sevila's (Yu et al., 2023) self-chained VideoQA framework uses a two-step approach: selecting keyframes with a tailored prompt and applying them to tasks. However, it faces three issues: time-consuming keyframe localization, static frames missing motion details, and incomplete video representation by sampled frames. Addressing these, we introduce a sampler-solver framework that incorporates both static and dynamic features for video-language understanding.