Efficient Temporal Extrapolation of Multimodal Large Language Models with Temporal Grounding Bridge

Anonymous ACL submission

Abstract

Despite progress in multimodal large language models (MLLMs), the challenge of interpreting long-form videos in response to linguistic queries persists, largely due to the inefficiency in temporal grounding and limited pre-trained context window size. In this work, we introduce Temporal Grounding Bridge (TGB), a novel framework that bootstraps MLLMs with advanced temporal grounding capabilities and broadens their contextual scope. Our framework significantly enhances the temporal capabilities of current MLLMs through three key innovations: an efficient multi-span temporal grounding algorithm applied to low-dimension temporal features projected from flow; a multimodal length extrapolation training paradigm that utilizes low-dimension temporal features to extend the training context window size; and a bootstrapping framework that bridges our model with pluggable MLLMs without requiring annotation. We validate TGB across seven video benchmarks and demonstrate substantial performance improvements compared with prior MLLMs. Notably, our model, initially trained on sequences of four frames, effectively handles sequences up to 16× longer without sacrificing performance, highlighting its scalability and effectiveness in real-world applications. Our code is publicly available.

1 Introduction

A fundamental aspect of human intelligence is to effortlessly perceive, memorize, and comprehend daily multi-modal information such as events, observations, and videos that span hours and days. Such capacity of long-form multi-modal understanding, seamlessly integrating prolonged visual dynamics with textual cues, poses considerable challenges for contemporary machine perceptual systems. A wide range of research works in computer vision and multi-modal tasks has extensively delved into real-life videos, including video question answering (VideoQA) (Yu et al., 2018, 2019), text-to-video retrieval (Hendricks et al., 2017), video captioning (Xu et al., 2016; Krishna et al., 2017), *etc*. Despite the prominent advancements in many video-language benchmarks (Yu et al., 2018, 2019; Hendricks et al., 2017; Xu et al., 2016; Krishna et al., 2017), understanding long-form 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; He et al., 2024) imposed by high-dimensional video data and the disparity between language and spatial-temporal visual cues (Lei et al., 2022; Xiao et al., 2023a).

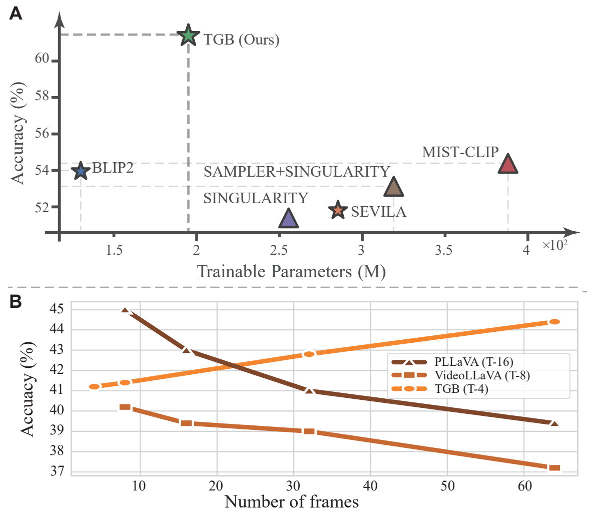


Figure 1: Training Efficiency and Length Extrapolation of TGB. **A.** Our method demonstrates the best performance with less trainable parameters. **B.** Results of frame extrapolation on EgoSchema (Mangalam et al., 2023) under zero-shot setting. T-num indicates the number of training context window size. By training with four-frame videos, our model shows consistent performance on extended video length.

Some researchers have proposed scaling up the amount of vision data fed into larger models(Bai et al., 2023; Liu et al., 2024a), following the scaling law observed in LLMs. However, the scarcity of high-quality, long video-language datasets makes this approach difficult. Others have explored sampling-based methods to reduce input overhead by selecting relevant frames at either the frame level(Lei et al., 2021; Wang et al., 2023; Bain et al., 2021; Buch et al., 2022) or token level (Gao et al., 2023a). These methods have three main limitations: first, they are computationally inefficient with slow training and inference speeds due to the large number of tunable parameters; second, the sampling strategy may miss important motion features, especially when there's misalignment between the video segment and the language query; and third, the complexity of the specialized vision encoder complicates the adaptation to long-video understanding.

To address these challenges, we present a novel framework, Temporal Grounding Bridge, which enriches image-language models with temporal priors, significantly improving the understanding of long videos. TGB distinguishes itself in the following key aspects:

Efficient and Adaptable Video Compression: TGB features a Bridge that is both lightweight and adaptable. To achieve this, we introduce a learnable multi-span algorithm capable of simultaneously extracting multiple relevant segments from low-dimension motion features. Subsequently, we can compress the entire video into several keyframes. This method efficiently balances performance and resource consumption when processing long-form videos, as demonstrated by our results on the AGQA (Grunde-McLaughlin et al., 2021), with a relatively low parameter count (see Fig. 1A).

Temporal Extrapolation Preserving Motion Features: A significant advantage of the TGB lies in its ability to preserve the continuity of video content, thereby maintaining the temporal dynamics discarded by previously extracted keyframes. To achieve this, we retain the low-dimensional motion features extracted by the TGB to supplement these keyframes. Additionally, we utilize extrapolative position encoding to ensure that these features remain extendable. This approach allows our method to extrapolate to longer sequences in a zero-shot setting (see Fig. 1B).

Bootstrapping Framework without Annotation: Due to the high cost of manual annotations and the limited availability of video data compared to image data, we developed a framework that leverages MLLMs without requiring them to be pretrained on videos. Our approach employs a bootstrapping strategy to refine TGB using MLLMs, eliminating the need for explicit temporal grounding annotations. This strategy also allows for joint training with MLLMs by incorporating the Gumbel-Softmax trick. Additionally, our bootstrapping framework, when integrated with the aforementioned mechanism, can be trained on standard video data and still achieve strong performance on much longer sequences (see Fig. 1B).

To validate the effectiveness of TGB, we conducted experiments on long-form video question answering with seven datasets: AGQA 2.0 (Grunde-McLaughlin et al., 2021), NExT-QA (Xiao et al., 2021), Egoschema (Mangalam et al., 2023), MSVD (Xu et al., 2017), MSRVTT (Xu et al., 2016), and ActivityNet (Yu et al., 2019). Additionally, we tested temporal question grounding on video using the NExT-GQA dataset (Xiao et al., 2023a). Consistent improvements across these datasets confirm the efficacy of our approach. TGB has shown strong generalization capabilities across five MLLMs (across encoder, encoder-decoder, and decoder-only) and two LLMs. Further enhancements include the incorporation of a general multimodal instruction-tuning dataset, which shows promise for video chat agent applications. In comparison to other leading-edge methods, TGB provides substantial efficiency and efficacy benefits.

2 Related Work

Long-form Video Understanding The computational demands of processing long-form videos have led to research exploring various methods to address the challenge. A common approach involves sampling-based techniques that aim to reduce the computational load by selectively choosing relevant frames. Research (Lei et al., 2021; Wang et al., 2023; Bain et al., 2021) integrate sparse sampling within the framework of video-language pretraining. (Buch et al., 2022) introduce an atemporal probe (ATP) model that seeks to distill a single image representation from a video clip for more details. Despite these advancements, there's a risk that sparse sampling may lead to an insufficient representation of visual information, which may not be relevant to corresponding language queries. MIST (Gao et al., 2023a) attempts to address this by leveraging the inherent structure of videos to iteratively select and sample spatial-temporal information within a Transformer architecture. Nonetheless, these methods often suffer from reduced computational efficiency and prolonged training and inference times due to the extensive tunable parameters required for processing either spatial or temporal dimensions. More recent studies are exploring the utilization of LLMs for enhancing long-form video understanding. These approaches include a range of techniques such as incorporating temporal embeddings (Qian et al., 2024), applying prompt-based strategies (Yu et al., 2023; Ren et al., 2023), condensing video frames through a similarity metric (Song et al., 2023), compressing visual tokens with resampling methods (Korbar et al., 2023; Ma et al., 2023; Liu et al., 2024b), and employing retrieval-based methods that integrate visual features (He et al., 2024). To overcome the constraints of current methods, which usually depend on human-provided annotations for time alignment or require intricate encoding of context, our proposed approach employs a novel bootstrapping framework. This framework enhances a temporal grounding bridge, using MLLMs. This bridge is designed to simultaneously capture multiple granular pieces of key information by leveraging multi-span sampling, which it then integrates with low-dimensional motion features for a more efficient and effective representation.

Bootstrapping Large Language Models for Visual Tasks Capitalizing on the success of large language models (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., 2023b), and visual instruction following (Xu et al., 2022; Li et al., 2023c). The research mainly progresses along three avenues: (i) leveraging LLMs' reasoning for visual tasks (Huang et al., 2023; 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., 2023; Lyu et al., 2023). 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 TGB that incorporates both static and dynamic features for video-language understanding.

3 Methodology

In the subsequent sections, we begin with a detailed formulation of the video-language understanding problem in Section §3.1. Next, in Section §3.2, we outline the core components for efficient length extrapolation of our TGB. Section §3.3 explains the process of jointly tuning TGB with pluggable MLLMs on new video-language datasets within our Bootstrapping framework. The overall architecture of TGB is illustrated in Figure 2.

3.1 Problem Definition

We formalize the open-ended video-text understanding problem. The input video V is denoted as a sequence of image frames $V = \{fr_1, fr_2, \cdots, fr_T\}$, where T is the total number of frames. The input language L, denoted as a sequence of N tokens starting with `[CLS]`, is a task-relevant prompt or question related to interactions among objects, relationships, or events that occur within a few frames of the video. Our goal is to identify the keyframes that relate to the query as grounded moments and generate an open-ended answer in the form of natural language response y, incorporating time priors. In the following sections, we use $f(\cdot)$ to indicate trainable parameters or neural networks and $f(\cdot)$ to indicate frozen pre-trained models.

3.2 Temporal Grounding Bridge

Previous Video-Language Understanding models commonly extract temporal features from video-text data using offline video encoders or image encoders (Carreira and Zisserman, 2017; Jiang et al., 2017; Xie et al., 2017; Feichtenhofer et al., 2019; Liu et al., 2021a; Tong et al., 2022), causing the model to be time-consuming and lack generality. To address these limitations, we propose a novel mechanism that combines high-dimension key visual cues with low-dimension motion features, ensuring efficiency without compromising visual information. We further contend that temporal grounding does not necessitate dense frame-level features. To support this claim, we introduce a Temporal Grounding Bridge that incorporates optical flows (Jiang et al., 2019; Feichtenhofer et al., 2019; Pfister et al., 2015; Feng et al., 2023; Zhang et al., 2018) during the temporal grounding

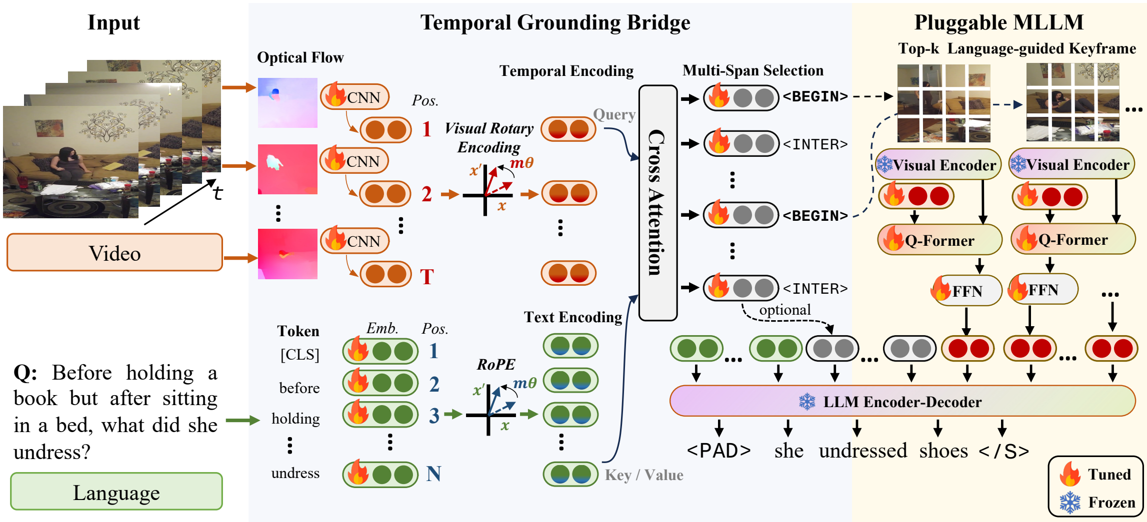


Figure 2: **Overview of TGB framework (BLIP-based).** The Temporal Grounding Bridge (§3.2) is designed to capture temporal priors as well as the specific moments in a video that are grounded by language. We further develop a pluggable bootstraping framework (§3.3) that incorporates TGB-MLLM alignment, utilizing a joint optimization strategy.

stage through a dimensionality reduction. By injecting language queries, this approach generates parameter-efficient, language-guided temporal features. **A key distinction of our work is that we do not use optical flow merely as supplementary information to enhance frame-based performance. Instead, our framework employs flow as a low-dimensional bridge, which can be directly or indirectly applied to infuse motion details into MLLMs. Importantly, the flow feature can be substituted with other types of features if needed**.

Feature Extraction We denote the optical flow for each pair of video frames as $OF = \{of_1, of_2, \cdots, of_T\}$ The low-dimension visual encoding is then computed over these extracted optical flows with a simple convolutional layer followed by a multi-layer perceptron (MLP) $E_{of} = \mathrm{MLP}(\mathrm{CNN}(of))$. For language queries, we use a trainable embedding layer to represent the soft query prompt, *i.e.*, $E_l = Embedding(Q)$, where Q is the language query.

Temporal Feature Length Extrapolation Despite the impressive efficacy of Transformer-based models within the sphere of deep learning, their operational capacity is inherently constrained by the length of the input. In the context of our research, the bridge is meticulously devised to identify the most salient portions of an entire video, the duration of which may potentially exceed the predetermined limit and differ significantly across various instances. Current literature employs a sampling strategy to condense the video, a process that unfortunately results in the loss of substantial temporal information inherent in the video. To mitigate this challenge, inspired by rotary position embedding (RoPE) (Su et al., 2021), we add multimodal extrapolative position encoding to our TGB(Fig. 2). Specifically, we compute the position-encoded features using RoPE mechanism for each optical flow and language token, respectively. Formally, the position-encoded features can be denoted as

$$\begin{aligned} E_{of}^R &= RoPE(W_{of}E_{of}, Pos_{of}), \\ E_l^R &= RoPE(W_l E_l, Pos_l), \end{aligned} \tag{1}$$

where W_{of}, W_l are transformation matrices, Pos_{of}, Pos_l are corresponding position indices of OF and L.

Given the temporal features, we adopt the cross-attention (Vaswani et al., 2017) mechanism, which is computed using optical flow rotary encoding as query $\mathbf{Q}_R = E_{of}^R$, rotary language embedding as key $\mathbf{K}_R = E_l^R$, and language embedding as value $\mathbf{V} = W_V E_l$.

The final language-guided temporal feature E_R is calculated by the standard cross-attention mechanism, *i.e.*, $E_R = \mathrm{Softmax}(\frac{\mathbf{Q}_R\mathbf{K}_R^{\mathrm{T}}}{\sqrt{d_k}})\mathbf{V}$.

Multi-Span Keyframe Selection 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:

$$h = \mathcal{F}_\theta(e_{R1}, e_{R2}, \dots, e_{RT}), \quad (2)$$
$$index = \arg\max(\text{Softmax}(h)),$$

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** demonstrates commonly used methods for temporal sentence grounding on video tasks. Compared with other span-fixed methods, our method could obtain multiple grounded video spans with the least time complexity and space complexity.

Bridge with MLLMs For each selected keyframe fr_k, we utilize a frozen pre-trained visual encoder to capture its spatial information, *i.e.*, $E_{fr} = Enc_v(fr_k)$. In line with contemporary research, we adapt the visual feature via a pre-trained Q-former and obtain q query representations. $\tilde{E}_q = Enc_q(E_q, E_{fr})$, where E_q represents the learnable query, $\tilde{E}_q = \{e_q\}$ is the spatial visual feature output of the MLLM. The final output is produced by feeding obtained spatial-temporal-language information in to a forzen LLM, *i.e.*, $y = \text{LLM}(E_r, \tilde{E}_q, E_l)$.

3.3 Joint Training Bootstrapping Framework

Bootstraping Algorithm Due to the scarcity of video-language datasets with temporally grounded annotations and the high cost of acquiring human labeling, we have developed a self-improvement algorithm to enhance TGB using the capabilities of MLLM. There are two primary types of video-language understanding tasks: close-ended and open-ended. We have tailored algorithms to address both types. For close-ended tasks, we employ an iterative method in which each video frame is evaluated using the MLLM. Frames that lead to correct MLLM predictions are marked with positive labels, while those with incorrect predictions receive negative labels. For open-ended tasks, which often lack temporal labels, we introduce an innovative approach to generate pseudo labels for open-ended datasets. We analyze the MLLM-generated results of uniformly sampled frames and compute the sentence similarity between these results and the ground truth. We then apply a monotonic stack algorithm to identify the span with the highest similarity scores. These pseudo labels are used to optimize the TGB. Detailed information about the this algorithm can be found in the **Appendix A**.

Joint Optimization Despite the utilization of pseudo labels in the training process, in many videos, there is implicit alignment between query and videos. In addition, the fixation of the pre-trained bridge within the bootstrapping framework inevitably leads to the introduction of exposure bias. To mitigate this we suggest a joint training approach that extends the Gumbel-Softmax technique. We implement Gumbel-Softmax sampling K times to sample K spans:

$$\text{GumbelSoftmax}(\mathcal{F}_\theta(e_{R1}, \dots, e_{RT}), \tau), \quad (3)$$

where τ is the scaling term for reparameterizing. Consequently, our methodology is employed to facilitate a connection between TGB and MLLMs, thereby enabling our framework to be jointly optimized on domain-specific datasets.

4 Experiments

In this section, we utilize the TGB on 5 MLLMs, across encoder, encoder-decoder, and decoder-only three types of architectures. We demonstrate the effectiveness of our approach on three tasks: long-form videoQA and zero-shot open-domain videoQA (Section 4.1), temporal question grounding on video (Section 4.2). Furthermore, We provide a detailed analysis to showcase the effectiveness of our framework in length extrapolation (Fig. 1B), the effectivness of different components (Section 4.3), and compare its computational efficiency with other state-of-the-art models on a similar scale (Section 4.4).

4.1 Long-form Video Question Answering

Setups We take three long-form VideoQA benchmarks AGQA (Grunde-McLaughlin et al., 2021), NExTQA (Xiao et al., 2021), and EgoSchema (Mangalam et al., 2023) for evaluation. 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

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
MIST-AIO (Gao et al., 2023a)	51.43	54.67	55.37	41.34	53.14	53.49	47.48	20.18	50.96
ALBEF	50.53	49.39	49.97	38.22	49.79	54.11	48.01	10.40	50.68
ALBEF + TGB (Ours)	51.05	51.11	51.66	38.36	51.33	58.10	49.20	11.78	51.73
SINGULARITY (Lei et al., 2022)	50.87	50.67	49.70	40.47	40.79	55.34	48.20	11.59	51.11
SINGULARITY + TGB (Ours)	52.33	54.12	55.07	40.71	54.49	57.88	48.35	12.24	53.13
VIOLET (Fu et al., 2021)	50.89	50.24	50.93	40.76	50.51	58.07	38.97	6.53	51.03
VIOLET + TGB (Ours)	51.59	54.54	56.96	40.94	55.61	59.12	42.81	9.02	52.59
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
TGB-BLIP2 (Ours)	**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 SeViLA for fair comparison.

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

Model	Temporal	Causal	Description	All
Retrieval-based Video-Language Models				
CLIP (Radford et al., 2021)	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
Open-ended Video-Language Models				
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
TGB-BLIP2 (Ours)	**66.5**	**72.8**	**81.2**	**72.1**

* We removed prior information from QVHighlights used in SeViLA for fair comparison.

Table 2: **Comparison accuracy of long-form video QA on NExT-QA**.

Methods	Base Model	# of Frames	Accuracy
Sevila	BLIP2	32	25.7
mPLUG-Owl	LLaMA-7b	5	33.8
Video-LLaVA	LLaVA-7b	8	40.2
TGB-BLIP2	BLIP2	4	**41.2**

Table 3: **Zero-shot Result on subset of EgoSchema**

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 TGB and 6 frames for TGB-augmented methods (where we don't incorporate the motion feature to the input directly) in all experiments. For more implementation details, please refer to **Appendix F.1**.

Results on AGQA 2.0 Our TGB framework, compared with prior works that integrate keyframe localization into video-language tasks, shows that BLIP2, despite its 4.1B parameters pre-trained on 129M images, offers only a slight improvement over smaller models, as demonstrated in AGQA 2.0 results. BLIP2 even falls short of the state-of-the-art MIST-CLIP, which has a parameter count comparable to BERT (Devlin et al., 2019). This indicates that simply adapting videos for LLMs is inadequate for complex video question-answering tasks. However, when enhanced with our TGB framework, BLIP2's accuracy increases by 7.45 points, underscoring the framework's ability to learn spatial-temporal video features effectively. We believe this is due to our framework's superior temporal information capture, which other methods miss. Nonetheless, it still lags behind MIST-CLIP on certain question types, stemming from the inherent differences in how retrieval-based and open-ended models produce answers. For example, open-ended models struggle with "Duration comparison" questions because they are limited to generating answers from a specific set of 171 words or phrases, which are infrequently found in generative models' pre-training data, posing a challenge for exact match generation.

Results on NExTQA Table 2 presents the results on the NExTQA dataset. Generally, our method outperforms a variety of baselines, particularly SeViLA, a recent model using LLM for keyframe selection. However, the performance im-

Methods	LLM size	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	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
TGB (BLIP2)	3B	66.0	3.6	53.5	3.1	41.3	3.1
TGB (Vicuna7B)	7B	**71.4**	**3.9**	**57.3**	**3.3**	**43.9**	**3.3**

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

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
TGB	OF+CNN	**19.9**	**23.3**	**11.2**

Table 5: **Comparison results of Temporal Question Grounding task on NExT-GQA (Xiao et al., 2023b).**

provement of our framework on NExTQA is not as significant as on AGQA. This is because NExTQA places more emphasis on causality, and videos in NExTQA, sourced from VidOR (Shang et al., 2019; Thomee et al., 2016), a dataset focused on video objects and relation recognition, exhibit more "static appearance bias" (Lei et al., 2022) than AGQA.

Results on EgoSchema We evaluated our model's performance on the EgoSchema (Mangalam et al., 2023), one of the longest videoQA datasets available. We apply this experiment under the zero-shot setting, thereby trained on video instruction dataset from VideoLLaVA (Lin et al., 2023). As shown in Table 3, our model outperforms other models that use similar pretraining data. This superior performance is particularly notable given that our base model is smaller and processes fewer input instances compared to the others. We believe our approach is highly effective for understanding long-form video content.

Impact of TGB-grounded frames We assessed the influence of TGB on different MLLMs by testing them with alternative MLLMs and TGB-grounded frames, excluding optical flow features. For MLLMs using single-image input, we merged multiple images using an early fusion approach. Our experiments on the AGQA 2.0 dataset in Table 1 revealed: ❶ *TGB matters in temporal learning over different MLLMs.* TGB-augmented methods significantly enhances MLLMs' ability in solving temporal question (*i.e.*, "`Relation-action` ", "`Sequencing` ", "`Exists` ") compared to the uniform sampling strategy. ❷ *Absence in temporal priors hinders the performance of ensemble methods.* The improvement gained on SINGULARITY is better than ALBEF, despite they have similar objectives but SINGULARITY is pre-trained with video corpora. ❸ *Temporal features of optical flow can compensate for the information loss caused by frame sampling.* The marginal improvement of our TGB-augmented models on "`Superlative` " suggests that the sampling strategy cannot enhance the model's overall video understanding ability. In contrast, our BLIP2-based framework with optical flow improves from 43.84 to 53.67 (a relative increase of 22.42%). indicating that optical flow features can reduce the temporal information loss caused by the sampling strategy.

Analysis of Pluggable MLLMs We substitute the BLIP2 with three popular types of MLLMs, mainly encoder-based models, *i.e.*, VIOLET (Fu et al., 2021) as a representative of video-language models, ALBEF (Li et al., 2021) as an image-language model, SINGULARITY (Lei et al., 2022) as a pre-trained model on a single frame of video and image corpus. It's noteworthy that we did not incorporate the learned optical flow feature into these MLLMs' input. In this part, we also apply all the experiments on AGQA 2.0 dataset. Table 1 (ALBEF + TGB, VIOLET + TGB, SIGULARITY + TGB) validates the efficacy of our TGB and the versatility of our framework. On average, the solver achieves a 3.68% accuracy improvement after replacing the uniform sampled frames with keyframes extracted by the TGB. These results consistently demonstrate the effectiveness of our TGB framework across various MLLMs.

Generality of TGB To demonstrate the generality of our approach, we applied our model to visual instruction datasets (Lin et al., 2023). We also adapted the LLM using LoRA (Hu et al., 2022) to ensure a fair comparison with current SOTA methods. As shown in Table 4, our method's performance on the videoQA dataset in a zero-shot setting is presented. Unlike VideoLLaVA, our method was not pretrained on additional datasets; it was only fine-tuned on the same visual instruction datasets. The results demonstrate that our method can match the performance of the latest state-of-the-art (SOTA) MLLMs, even though the LLM of our model is less than half their size. This highlights the considerable promise of our framework in this domain.

Model	Object-relation	Relation-action	Object-action	Others	All
TGB	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 bridge	62.28	47.84	50.68	53.47	59.88
w/ uniform sampling	53.72	48.64	62.10	50.68	54.00
w/ zero-shot	23.60	17.09	29.37	40.72	25.54

Table 6: **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.

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

Table 7: **Computational Efficiency of TGB.**

4.2 Temporal Question Grounding on Video

Setup We use the Temporal Question Grounding on Video (TQGV) dataset NExT-GQA (Xiao et al., 2023a) to evaluate the efficacy of our TGB. We select a wide range of MLLMs 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., 2022), 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.

Main Results and Analysis As shown in Table 5, our method outperforms baselines using additional feature extractors (Ren et al., 2015; Liu et al., 2021b,a; Radford et al., 2021). Our TGB with optical flow effectively learns temporal priors for video-language tasks. We suggest that discrete frames may introduce irrelevant visual cues, increasing the computational load for temporal learning. Despite this, all methods struggle with temporal grounding, with most mIoU values under 0.20, indicating a significant gap in current temporal modeling. Conversely, our TGB's temporal features could mitigate these issues. We posit that our approach could significantly advance spatial-temporal research for extended videos. Qualitative results are presented in Appendix G.

4.3 Ablation Study

We apply ablation study on TGB 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 6, 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 TGB during training notably affected performance on temporal questions like "`Relation-action` ", suggesting that joint training can further optimize the bridge. 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.1, we provide detailed ablation study about the TGB-augmented models.

4.4 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 7. Our method outperformed the current SOTA model SeViLa, which uses the LLM to select keyframes, both in terms of 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 computationally efficient. Further details on the composition of the inference time of TGB are provided in *SM*. In addition, we investigate the composition of inference time of TGB and offline demo in Appendix B.

5 Conclusion

In this work, we propose a pluggable framework TGB for long Video-Language Understanding tasks, which comprises a TGB and a spatial prompt solver to combine spatial-temporal-language alignment and temporal grounding. Experiments on long-form video question answering and temporal question grounding on video demonstrate a consistent improvement over various types of MLLMs. Comprehensive analysis verifies the effectiveness, efficiency, and generality of our framework.

Limitations

Our study has one primary limitation: *i.e*. **Limited Temporal Grounding Capability** As shown in Section 4.2, 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 TGB. In future work, we aim to enhance this aspect of our method without sacrificing efficiency.

Ethics Statement and Broad Impact

References

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L. Yuille, Trevor Darrell, Jitendra Malik, and Alexei A. Efros. 2023. Sequential modeling enables scalable learning for large vision models. *CoRR*, abs/2312.00785.

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.

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.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE Computer Society.

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*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

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.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *International Conference on Computer Vision (ICCV)*, pages 6201–6210. IEEE.

Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. 2023. Mutual information-based temporal difference learning for human pose estimation in video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17131–17141.

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, 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.

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)*.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: memory-augmented large multimodal model for long-term video understanding. *CoRR*, abs/2404.05726.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing moments in video with natural language. *International Conference on Computer Vision (ICCV)*, pages 5804–5813.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

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

Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. 2019. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2000–2009.

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.

Zhuolin Jiang, Viktor Rozgic, and Sancar Adali. 2017. Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 309–317. IEEE Computer Society.

Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. 2023. Text-conditioned resampler for long form video understanding. *CoRR*, abs/2312.11897.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. *International Conference on Computer Vision (ICCV)*.

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.

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 pretraining with frozen image encoders and large language models.

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)*.

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. 2023c. M3it: A large-scale dataset towards multimodal multilingual instruction tuning. *ArXiv*.

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.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. 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.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with blockwise ringattention. *CoRR*, abs/2402.08268.

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

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2024b. ST-LLM: large language models are effective temporal learners. *CoRR*, abs/2404.00308.

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. 2021a. 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. 2021b. 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. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2023. Vista-llama: Reliable video narrator via equal distance to visual tokens. *CoRR*, abs/2312.08870.

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

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*.

Tomas Pfister, James Charles, and Andrew Zisserman. 2015. Flowing convnets for human pose estimation in videos. In *International Conference on Computer Vision (ICCV)*, pages 1913–1921.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning. *CoRR*, abs/2402.11435.

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. 2021. 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.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023. Timechat: A time-sensitive multimodal large language model for long video understanding. *CoRR*, abs/2312.02051.

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.

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.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6000–6010.

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)*.

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.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995. IEEE Computer Society.

D. Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. *Association for Computing Machinery's Annual Conference on Multimedia (ACM MM)*.

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.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

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

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.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. *European Conference on Computer Vision (ECCV)*.

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.

Dingwen Zhang, Guangyu Guo, Dong Huang, and Junwei Han. 2018. Poseflow: A deep motion representation for understanding human behaviors in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6762–6770.

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

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

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

Appendices

Table of Contents

A Self-Improvement Algorithm

Algorithm 1 shows our self-improvement algorithm of automatically generating pseudo labels by the MLLM, which is used to optimize the TGB.

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**
  $prediction = LLM_{MLLM}(fr, q)$
  $scores.add(SIM(prediction, a))$
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

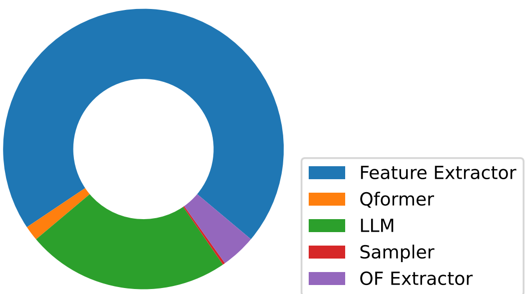


Figure 3: Inference time Analysis

We further investigate the composition of inference time of TGB 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.

C More Analysis Experiments

C.1 Ablated TSP-augmented models

TGB	MLLM	# 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 TGB.

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 MLLM benefits more from the sampling strategy when adequately pre-trained (*i.e.*, 17M denotes the model is pretrained on 17M video corpora). Additionally, we proposed two TGB variants, replacing optical flow with features extracted by the video SwinTransformer (Liu et al., 2021b) for pre-training. The comparable results suggest that our TSP can effectively reason over time without any prior perception information.

C.2 Influence of the number of frames on solver

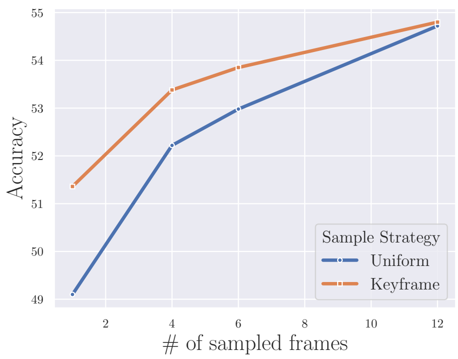


Figure 4: Further study on the number of sampled frames.

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.3 Detailed Ablation Study Results

	TGB	w/o Optical Flow	fixed TGB	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 TGB on AGQA 2.0. Specifically, we demonstrates the ablation study results of different question types.

D Details of Multi-span Prediction

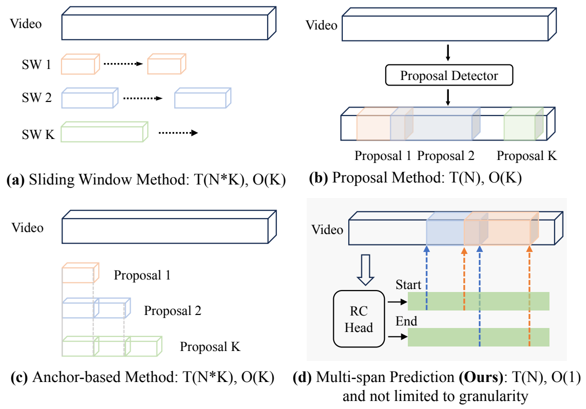


Figure 5: **Comparison of multi-span RC prediction (d)** and other methods (a-c) in terms of time and space complexity.

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.

E Implementation Details

F Details of Datasets

Long-form VideoQA. 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,

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

and descriptive reasoning, including 52K questions.

Temporal Question Grounding on Video. 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).

F.1 Implementation Details of TGB on Downstream Tasks

The TGB is a 6-layer transformer with RoPE (Su et al., 2021). For TGB, We use BLIP2-flant5-xl (Li et al., 2023b) as TGB. For the TGB-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 TGB and 6 frames for TGB-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.

F.2 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

G Qualitative Studies on NExTGQA

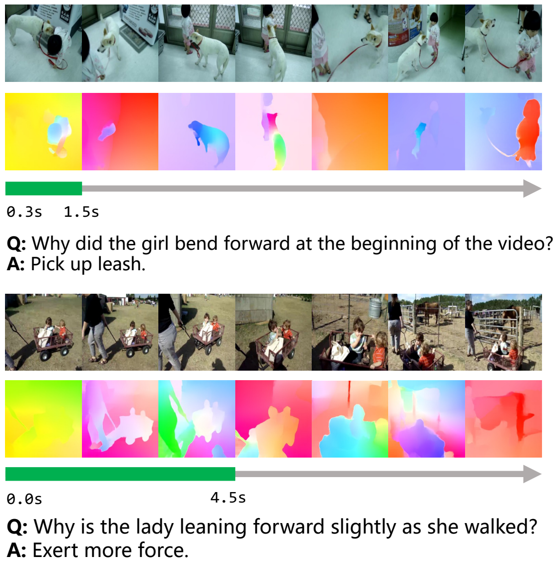
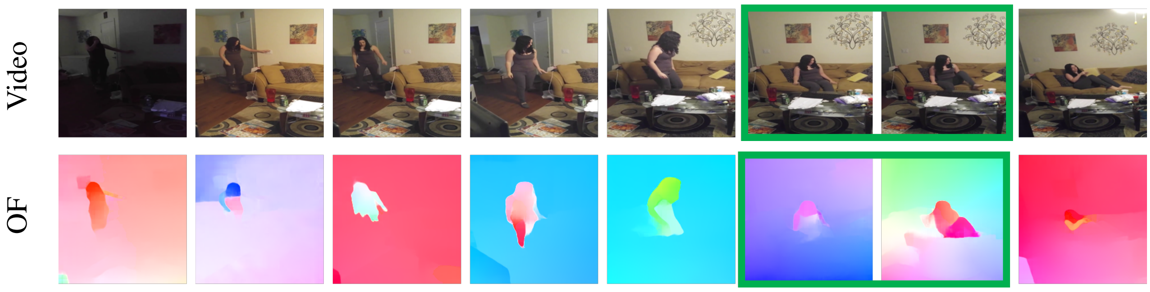


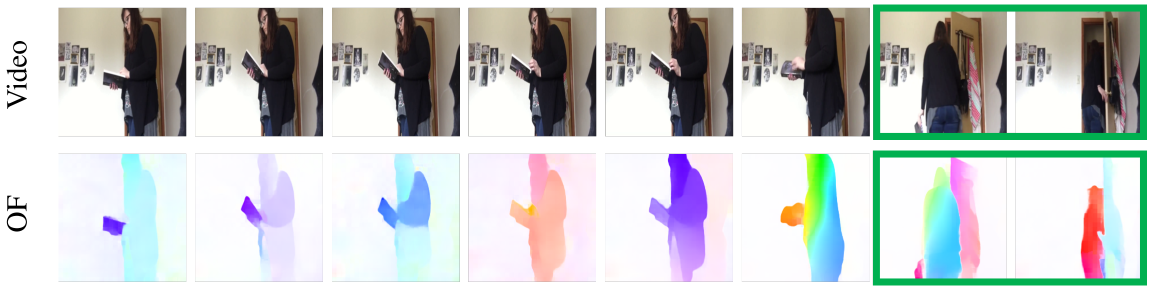
Figure 7: Qualitative results on temporal grounding

Fig. 7 presents two random outputs from TGB 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`".

H Qualitative Studies on AGQA 2.0



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



Question: Which object did the person grasp after watching a book?
Ground Truth: doorknob **TGB:** **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.