This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VISTA-LLAMA: Reducing Hallucination in Video Language Models via Equal Distance to Visual Tokens

Fan Ma[1], Xiaojie Jin[2*], Heng Wang[2], Yuchen Xian[1], Jiashi Feng[2], Yi Yang[1*]
[1]Zhejiang University [2]ByteDance Inc.

Abstract

Recent advances in large video-language models have displayed promising outcomes in video comprehension. Current approaches straightforwardly convert video into language tokens and employ large language models for multi-modal tasks. However, this method often leads to the generation of irrelevant content, commonly known as "hallucination", as the length of the text increases and the impact of the video diminishes. To address this problem, we propose VISTA-LLAMA, *a novel framework that maintains the consistent distance between all visual tokens and any language tokens, irrespective of the generated text length.* VISTA-LLAMA *omits relative position encoding when determining attention weights between visual and text tokens, retaining the position encoding for text and text tokens. This amplifies the effect of visual tokens on text generation, especially when the relative distance is longer between visual and text tokens. The proposed attention mechanism significantly reduces the chance of producing irrelevant text related to the video content. Furthermore, we present a sequential visual projector that projects the current video frame into tokens of language space with the assistance of the previous frame. This approach not only captures the temporal relationship within the video, but also allows less visual tokens to encompass the entire video. Our approach significantly outperforms various previous methods* (e.g., *Video-ChatGPT, MovieChat) on four challenging open-ended video question answering benchmarks. We reach an accuracy of 60.7 on the zero-shot NExT-QA and 60.5 on the zero-shot MSRVTT-QA, setting a new state-of-the-art performance. This project is available at https://jinxxian.github.io/Vista-LLaMA.*

1. Introduction

The surge in multi-modal vision-and-language models [1, 12, 14], capable of comprehending both visual (*e.g.*, image/video) and language data, can be attributed to the recent achievements of large language models (LLMs) such as GPT [8], GLM [38], and LLaMA [40]. Video-language models (video-LMs) pose greater challenges in scaling due to increased computational and annotation costs compared to image-language models (image-LMs). Recent research has therefore focused on effectively training video-LMs by utilizing pre-trained image-LMs [22, 40].

Video-LMs benefit from this warm-start approach to enhance visual representation learning by projecting video frames into language space and treating videos as several prompt language tokens [18, 34]. However, this diminishes the visual impact of videos on text generation and lacks explicit temporal modeling in videos. The generated text is often not related to the video content, as depicted in Figure 1, a phenomenon known as hallucination in language processing [6]. The distance between the generated text token and the visual tokens in large language models may

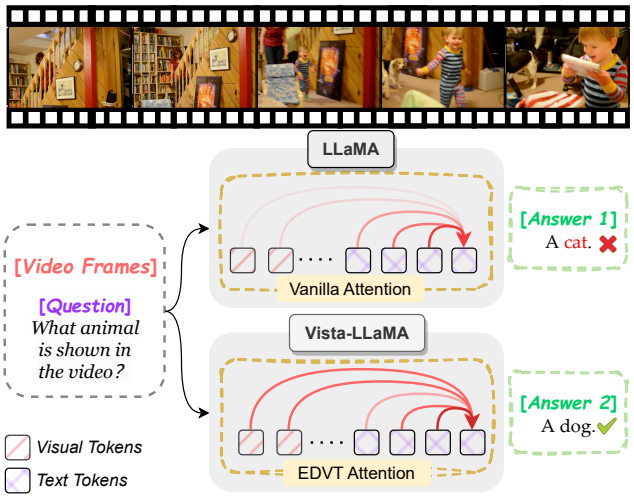


Figure 1. **Video language processing with LLaMA [26] and our VISTA-LLAMA.** The vanilla LLaMA treats visual tokens (▨) the same as other language tokens (▩), weakening the impact for tokens in long distance. Our model retains the same mechanism for language tokens and strengthens the impact of the visual tokens. The intensity of the impact of each token is conveyed through the depth of the line color (→). Our model provides the accurate response for the presented scenario.

* Corresponding author

be a contributing factor, especially when the visual tokens are distant from the generated text. Additionally, handling longer videos poses a challenge because of the constraints on context length in large language models, where numerous visual tokens take up a substantial portion of the context.

We present VISTA-LLAMA, an innovative video language framework to tackle above issues. The primary concept is to preserve equal distance between all visual tokens and any language tokens, while also retaining the relative distance between any two language tokens, as depicted in Figure 1. The rotary position embedding is only applied to language tokens to capture relative distance when calculating similarity between language tokens. When computing similarity between visual and text tokens, the rotary position embedding is removed to reduce the impact of relative distance. With the equal distance to visual tokens (EDVT) attention, the impact of visual cues on text production is amplified without compromising the text production capability. The experiments also show that our design produces more precise text description for input videos and the phenomenon of visual hallucinations occurring in language models has been greatly reduced.

To further improve the temporal modeling of video, we introduce a sequential visual projector. Instead of mapping each frame of the video independently into fixed-length visual tokens, which ignores the temporal relationship between frames, we generate visual tokens for each frame using the previous projected visual tokens. It not only incorporates the temporal relationship between frames into the language model to improve video comprehension without any extra parameters, but also enables the language model to encode longer videos with fewer visual tokens by sampling fewer projected frames.

We demonstrate the effectiveness of VISTA-LLAMA on four challenging video question answering (QA) benchmarks, where our method outperforms several previous works, and achieves the state-of-the-art in zero-shot NExT-QA [30], and MSRVTT [32]. We also show that our attention design and the temporal modeling mechanism improves capacity of large language model on video-question answering by a large margin. Comprehensive experiments are conducted to demonstrate the effectiveness of our designs. We summarize our contributions as follows:

- We introduce a novel video-language model, dubbed as VISTA-LLAMA, to enhance the video understanding and facilitate temporal modeling within the language model.
- A novel multi-modal attention is proposed to enable reliable video text generation by maintaining equal distance to visual tokens. Additionally, temporal modeling is facilitated by employing a sequential visual projector.
- Our method exhibits superior empirical performance, establishing new benchmarks for zero-shot open-ended video question answering tasks.
- A detailed analysis further explains the design choices inherent in our proposed framework, contributing significantly to a better comprehension of multi-modal dynamics in large language models.

2. Related Work

2.1. Large Language Models

Large language models (LLMs) [2, 19, 27, 28, 42] have demonstrated exceptional proficiency in understanding language and reasoning, resulting in the generation of high-quality natural language text in diverse domains. LLMs have already initiated a technological revolution and have found extensive application in various domains. Additionally, a series of open source large models, including LLaMA [26], and OPT [41], have significantly contributed to technological advancements and made remarkable contributions to the NLP community. Leveraging the foundation established by these impressive LLMs, researchers have further expanded their capabilities and developed exceptional models for diverse NLP tasks (*e.g.* Vicuna [4]). Our work is also built upon these remarkable LLMs, and we equip the language models with a novel attention mechanism and a sequential visual projector that enhances their abilities to comprehend visual content in videos.

2.2. Multi-modal Large Language Models

Researchers have been diligently exploring the application of LLMs for processing multi-modal problems [9, 13, 15, 17]. The existing approaches can be classified into two primary categories. In the first category, LLMs are treated as controllers while the multi-modal models as tools [10]. When presented with the specific task, the LLMs first interact with user instructions and decide which tools to employ. Subsequently, it generates comprehensive responses by amalgamating the outcomes derived from these readily available multi-modal models. These approaches, such as Visual ChatGPT [29], HuggingGPT [21], and DoraemonGPT [35], have shown impressive results on various multi-modal tasks without training models. For the second category, large-scale multi-modal models are trained on the multi-modal data. The principle behind this line of researches is to align other modal pre-trained models with textual LLMs. Flamingo [1] incorporates a perceiver resampler and a gated cross-attention layer to connect a frozen image encoder with the LLM. BLIP2 [12] introduces a Q-Former to map each image into fixed-length tokens in the language embedding space via the learned queries. LLaVA [14], mPLUG-owl [36], and MiniGPT4 [42] develop the image-LLMs utilizing image-instruction training pairs. Video Chat [13] extends image encoders to enable large models to comprehend visual content in videos. Video-ChatGPT [18] is trained on video instructional data to give appropriate answers for multi-modal

inputs. In this work, we improve the understanding of videos in a video-language model by incorporating temporal modeling and a novel multi-modal attention mechanism.

2.3. Video Question Answering

Video Question Answering (VideoQA) is a task that involves answering language questions that are based on a given video [32]. It requires the ability to understand and reason across different semantic levels, which in turn demands a capacity for multi-modal understanding. Previous VideoQA benchmarks primarily concentrated on short videos, asking questions according to the specified visual facts (*e.g.*, location and objects) [16, 33]. Recently, several new benchmarks have been proposed to focus on resolving temporal and causal questions in longer video clips [25]. NExT-QA [30] is an example of such a benchmark, aiming to uncover the causalities or intentions of particular events, and infer subsequent actions within the video. In this work, we investigate the video-language model on zero-shot open-ended VideoQA where no training question-answer pairs are provided in the training stage.

3. Method

3.1. Overview

VISTA-LLAMA comprises three fundamental components: a visual encoder, a visual projector, and a pre-trained large language model. Figure 2 shows an overview of VISTA-LLAMA. The components' design and implementation details are provided below:

- ***Visual Encoder***. We utilize pre-trained EVA-CLIP-g [24] and ViT-L [20] as visual encoder. The last layer of ViT encoder is removed because it specializes in aggregating the features of the first token for contrastive learning.
- ***Visual Projector***. The visual projector maps the output of the visual encoder into tokens that occupy the same space as the text features from word embedding. Various visual projectors are considered in Sec. 3.3. The visual projector is to be trained in our work.
- ***Pre-trained Large Language Model***. In this study, we utilize LLaVa [14], which is fine-tuned on Vicuna-7B [4] using instructional image-text pairs, for processing videos and texts. The causal mask is employed in all attention processes, incorporating the attention interplay between visual and text tokens. The pre-trained language model is frozen in the present study, and a new attention mechanism is developed to maintain a consistent distance to visual tokens for all textual tokens.

In our framework, the video is first encoded into visual tokens using the frozen visual encoder and the trainable visual projector. These visual tokens are then combined with text tokens, which are projected with word embedding from the prompt and question sentences. The combined visual and text tokens are then input into the language model to generate answers.

Algorithm 1 Pseudocode of EDVT attention in a PyTorch-like style.

```
# x: hidden input in each attention laye
# v_mask: indcitate which input are from visual tokens

def edvt_attention_layer(x, v_mask):
    # query, answer, value projection
    q, k, v = qkv_proj(x)

    # apply RoPE for query and key inputs
    r_q, r_k = rope(q, k)

    # attention weights without RoPE
    attention = bmm(q.T, k)

    # attention weights with RoPE
    r_attention = bmm(r_q, r_k)

    # Merge attention weights based on visual token
    attention = v_mask * attention + (1 - v_mask) *
        r_attention
    attention = Softmax(attention, dim=-1)

    # Update representation based on attention weights
    v = bmm(attention, v)
    out = linear_proj(v)
    return out
```

`qkv_proj`: linear projection layer; `bmm`: batch matrix multiplication; `rope`: apply rotary position embedding.

3.2. Equal Distance to Visual Tokens

Preliminary. For the concatenate visual-text input, three linear projection layers are applied in every attention layer to produce the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$. Let $\mathbf{q}_j$ be j^{th} query in $\mathbf{Q}$. The conventional attention mechanism updates the input by initially determining the similarity between the query and key, then applying the attention weights to the value state. Specifically, for the j^{th} position, the update procedure can be articulated as follows:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})_j = \frac{\sum_{i=1}^{j} sim(\mathbf{q}_j,\mathbf{k}_i)\mathbf{v}_i}{\sum_{i=1}^{j} sim(\mathbf{q}_j,\mathbf{k}_i)}, \tag{1}$$

where $sim(\mathbf{q}_j,\mathbf{k}_i) = exp(\mathbf{q}_j^T\mathbf{k}_i/\sqrt{d})$. Here, a causal mask is utilized so that the j^{th} query can only attend to the key positioned less than j.

EDVT-Attention. The vanilla attention model lacks positional awareness, with no encoded relative distance for natural language processing. In contrast, Rotary Positional Embeddings (RoPE) [23] encodes the position data of tokens using a rotation matrix, which inherently includes an explicit relative position dependency. Within each attention layer, RoPE is implemented across all projected query and key inputs in order to compute the attention weights via leveraging relative distance between tokens. The query situated at the j^{th} position incorporates rotary position embedding through $\tilde{\mathbf{q}}_j = \mathbf{R}_j\mathbf{q}_j$, wherein $\mathbf{R}_j \in \mathbb{R}^{d\times d}$ represents the rotary matrix for the j^{th} position. Consequently, the attention involving

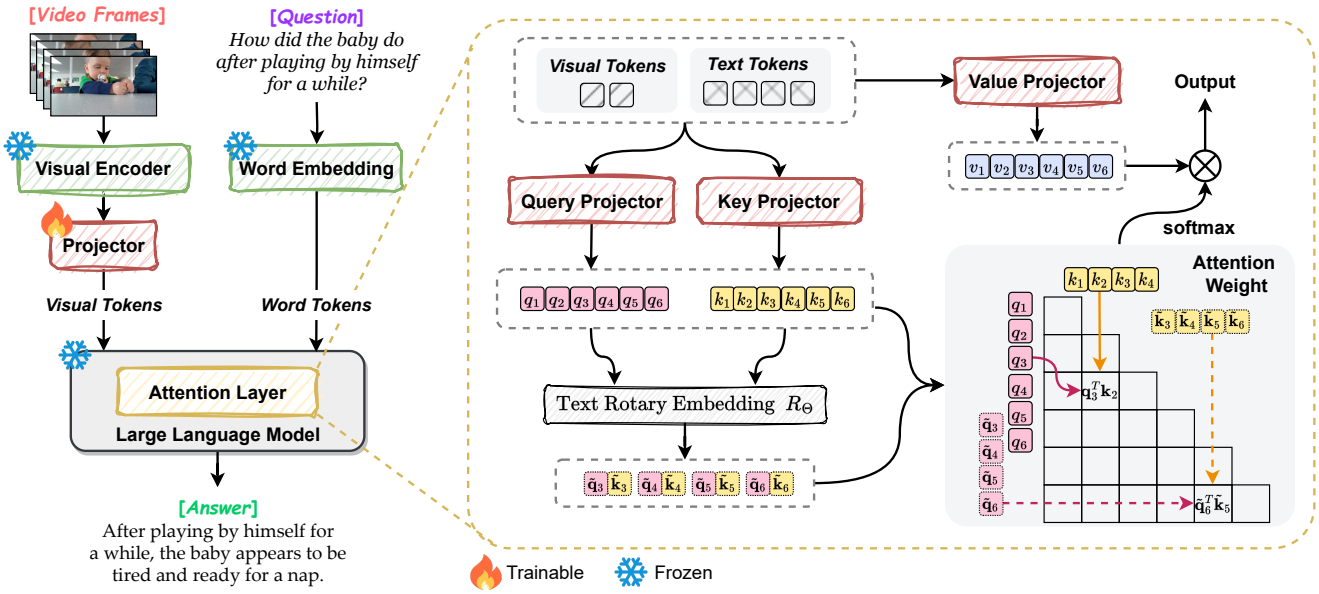


Figure 2. **The framework of VISTA-LLAMA**. The visual encoder and large language model are both frozen (❄) during training, while the projector is trainable (🔥) to map video into the language's space. The attention operation in each layer is present on the right part. Only the text tokens are applied with rotary position embedding to include relative distance information. The attention weights between visual and language tokens are calculated without the rotary position embedding. The casual mask is applied to the bottom-right attention weights.

relative position embedding is expressed as follows:

$$\text{Attention}_{rope}(\mathbf{Q},\mathbf{K},\mathbf{V})_j = \frac{\sum_{i=1}^{j} sim(\mathbf{R}_j\mathbf{q}_j, \mathbf{R}_i\mathbf{k}_i)\mathbf{v}_i}{\sum_{i=1}^{j} sim(\mathbf{R}_j\mathbf{q}_j, \mathbf{R}_i\mathbf{k}_i)}. \quad (2)$$

RoPE inherently integrates relative position data via the multiplication of rotation matrices rather than appending it to the input as a positional embedding. In natural language processing, this relative proximity between two words is vital, given that remote words should have less influence than adjacent words when generating the current word. However, using the same attention mechanism for visual and text tokens may result in unintentional text generation, a phenomenon often referred to as hallucination in LLMs. With multi-modal input, the generated text should depend on the visual content, disregarding the influence of relative distance.

To alleviate this issue and enhance the video understanding with LLMs, we introduce the EDVT-Attention where the equal distance to visual tokens is maintained while the relative distance between text tokens is attained. As shown in Figure 2, the rotary position embedding is only applied on text tokens. Let $\mathcal{V}$ and $\mathcal{T}$ be the set of visual and text tokens. The EDVT-Attention is formulated as:

$$\text{Attention}_{edvt}(\mathbf{Q},\mathbf{K},\mathbf{V})_j = \frac{\sum_{k_i\in\mathcal{T}} sim(\tilde{\mathbf{q}}_j, \tilde{\mathbf{k}}_i)\mathbf{v}_i + \sum_{k_i\in\mathcal{V}} sim(\mathbf{q}_j, \mathbf{k}_i)\mathbf{v}_i}{\sum_{k_i\in\mathcal{T}} sim(\tilde{\mathbf{q}}_j, \tilde{\mathbf{k}}_i) + \sum_{k_i\in\mathcal{V}} sim(\mathbf{q}_j, \mathbf{k}_i)}, \quad (3)$$

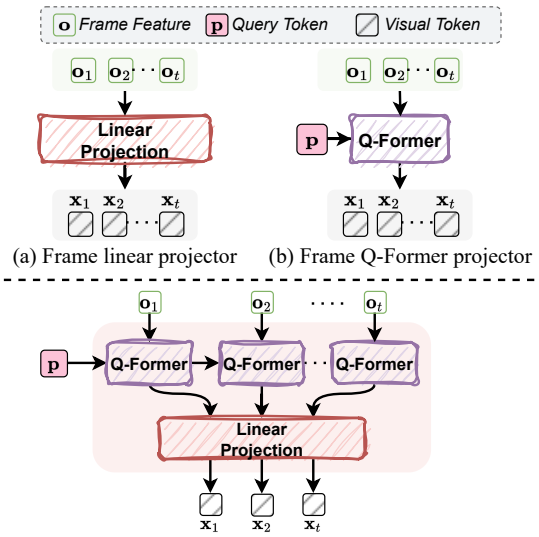
where $\tilde{\mathbf{q}}_j$ and $\tilde{\mathbf{k}}_j$ denote the query and key applied with rotary embedding, separately. The distance between language tokens is determined using the rotary matrix, while the correlation between visual and textual inputs remains unaffected by the rotary embedding. This improves the influence of visual information on long-term text generation and decreases the incidence of hallucinations wherein the fabricated content is absent in the videos.

For ease comprehension of how EDVT-Attention works, Algorithm 1 exhibits the pseudo-code in the decoder layer of LLMs. It involves the combinations of attention weights prior to and subsequent to the implementation of rotary embedding with the visual mask.

3.3. Sequential Visual Projector

The visual projector aims to map video features into the language embedding space, allowing for the fusion and processing of visual and textual inputs by the substantial language model. As shown in Figure 3, earlier visual projectors either employ the linear layer or the query transformer (Q-Former) [12] to directly convert frame features into language tokens. However, the lack of temporal relationship in these methods impedes thorough video understanding in LLMs. We introduce the sequential visual projector in Figure 3 to encode the temporal context into the visual tokens.

Let $\mathbf{o} \in \mathbb{R}^{t\times l\times d}$ be the extracted video feature of length t where l denotes the length of visual tokens. We use a Q-Former f_Q to map each frame into the fixed length of k representations $\mathbf{x}_i = f_Q(\mathbf{o}_i, \mathbf{p}) \in \mathbb{R}^{k\times d}$ where $\mathbf{p} \in \mathbb{R}^{k\times d}$ is the learnable query embedding. In the prior approach, all projected frame features are merely combined with word tokens to serve as mixed input for LLMs. For encoding temporal



(a) Frame linear projector (b) Frame Q-Former projector

(c) Temporal Q-Former projector

Figure 3. **Comparison of three visual projectors.** (a): Each frame feature is projected into the visual tokens independently with the linear projection. (b): Q-Former uses shared learnable query tokens to separately map each frame into fixed-length tokens. (c): The sequential Q-Former with linear projection layer to enable temporal modeling.

context, we utilize the previously projected frame feature as the query to attend to the current frame feature. The current projected frame feature is then updated accordingly:

$$\mathbf{x}_t = f_Q(\mathbf{o}_i, \mathbf{x}_{t-1}), \tag{4}$$

where $\mathbf{x}_{t-1} \in \mathbb{R}^{k \times d}$ represents the previous projected frame feature. This allows the visual tokens to encode the temporal relationship, as the current frame's visual token is generated using the previous feature.

Moreover, we can represent the entire video with fewer tokens by merely sampling a small number of visual tokens. This method tackles the challenge of encoding lengthy videos by employing sequential encoding. By integrating prior visual context into subsequent visual tokens, we can accomplish an adequate representation of the entire video through sparse sampling of projected frame features. Besides, this technique permits larger language models to manage much longer videos without length limitation.

3.4. Implementation Details

We fine-tune the model based on LLaVA [14] and use $100K$ video instruction pairs provided by [18]. We add Q-Former initialized from BLIP2 [12] to project frame features into fixed-length tokens. We test our model with both ViT-L-14 [20] and EVA-CLIP-g [24] visual encoders in experiments. We only update the visual projector, which contains the Q-Former and linear projection layer to project the video features to the LLMs' input space. The visual backbone and the language model are frozen during training. The VISTA-LLAMA is fine-tuned for 3 epochs on 8 A100 80GB GPUs with a learning rate of $2e^{-5}$ and an overall batch size of 32. We run all the inference experiments with FP16 to save memory and faster testing.

4. Experiments

4.1. Experimental Setup

Datasets. Our method is evaluated on four datasets:

- **NExT-QA** [30] is designed to advance video understanding from descriing to explaining the temporal actions. It comprises 5,440 videos and approximately $52K$ manually annotated QA pairs, which are categorized into *temporal* (Tem.), *causal* (Cau.), and *descriptive* (Des.) questions.
- **MSVD-QA** [32] is a dataset built upon Microsoft Research Video Description Corpus [3], commonly used in video caption tasks. The MSVD-QA dataset comprises a total of 1,970 video clips with 50,505 QA pairs.
- **MSRVTT-QA** [32] is based on MSR-VTT dataset [33], which includes $10K$ videos and $243K$ QA pairs with larger and has more complex scenes.
- **ActiviytNet-QA** [37] is a fully annotated and large-scale videoQA dataset. It contains $58K$ QA pairs derived from 5,800 complex web videos derived from the popular ActivityNet dataset [11].

Evaluation. We adopt two metrics to evaluate the performance of video-language models.

- **Open-Ended Zero-Shot Question-Answer Evaluation.** As video-language models generate responses of varying lengths to open-ended questions, it is challenge to evaluate models with traditional word matching strategy. We employ LLM-Assisted evaluation, in line with [18], for fair comparison. Given the question, correct answer, and predicted answer, GPT3.5-turbo-0613 is used to return *True* or *False* judgement and relative score (0 to 5).
- **Video-Based Text Generation Benchmarking.** We evaluate the text generation performance following [18] from five aspects: *Correctness of Information*, *Consistency*, *Detail Orientation*, *Contextual Understanding*, and *Temporal Understanding*. The test set for this evaluation is based on the ActivityNet-200 [7], featuring videos with rich, dense descriptive captions and associated question-answer pairs from human annotations. The evaluation pipeline is also built with the GPT-3.5 model and relative score (0 to 5) is generated.

4.2. Comparison to State-of-the-Arts

For the zero-shot open-ended video question answering tasks, we compare our model with FrozenBiLM [34], Video Chat [13], LLaMA Adapter [40], VideoLLaMA [39], Video-ChatGPT [18], and MovieChat [22] in Tab. 1. FrozenBiLM

Method	NExT-QA [30]		MSVD-QA [32]		MSRVTT-QA [32]		ActivityNet-QA [37]	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM [34]	-	-	32.2	-	16.8	-	24.7	-
Video Chat [13]	56.2	3.2	56.3	2.8	45.0	2.5	26.5	2.2
LLaMA Adapter [40]	-	-	54.9	3.1	43.8	2.7	34.2	2.7
Video LLaMA [39]	-	-	51.6	2.5	29.6	1.8	12.4	1.1
MovieChat [22]	49.9	2.7	61.0	2.9	49.7	2.8	**51.5**	3.1
Video-ChatGPT [18]	54.6	3.2	64.9	3.3	49.3	2.8	35.2	2.7
VISTA-LLAMA (Ours)	**60.7**	**3.4**	**65.3**	**3.6**	**60.5**	**3.3**	48.3	**3.3**

Table 1. **Comparison with SoTA methods** on zero-shot VideoQA. See §4.2 for more details.

Method	Cr.	Cs.	De.	Ct.	Te.
Video Chat [13]	2.23	2.24	2.50	2.53	1.94
LLaMA Adapter [40]	2.03	2.15	2.32	2.30	1.98
Video LLaMA [39]	1.96	1.79	2.18	2.16	1.82
Video-ChatGPT [18]	2.40	**2.37**	2.52	2.62	1.98
VISTA-LLAMA (Ours)	**2.44**	2.31	**2.64**	**3.18**	**2.26**

Table 2. **Quantitative results** on video-based text generation with different video-language methods (§4.2). For clarity, five scores are reported ("Cr.": *Correctness of Information*, "Cs.": *Consistency*, "De.": *Detail Orientation*, "Ct": *Contextual Understanding*, "Te.": *Temporal Understanding*).

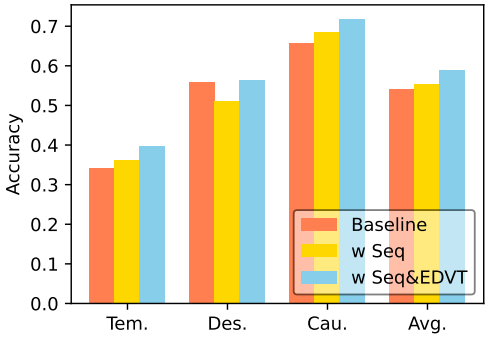


Figure 4. **Comparison of different design choices** on NExT-QA [30](§4.3). For clarity, accuracy of base model and two variants are given ("Baseline": the frozen LLaVA [14] with trainable Q-Former [12], "*w Seq*": base model with sequential visual projector, "*w Seq&EDVT*": base model with both sequential visual projector and EDVT-Attention).

adapts frozen the bidirectional language model, showing promising results in zero-shot VideoQA settings. Other compared models are built on recent large auto-regressive language models. Despite pre-existing models have produced substantial results, VISTA-LLAMA consistently outperforms them, achieving state-of-the-art (SoTA) performance across three datasets: NExT-QA [30], MSVD-QA [32], and MSVTT-QA [32]. Our method obtains the highest results, with 60.5% accuracy on MSRVTT, elevating the performance of the second-best model by nearly 10%. Additionally, our method attains 60.7% accuracy on NExT-QA, markedly superior to Video-ChatGPT [18]. These results demonstrate VISTA-LLAMA's capability to comprehend video content and produce precise answers.

We present the results of the evaluation of video-based text generation in Tab. 2. The results reveal its competent performance across all aspects when compared with the recent video-language models, Video Chat[13], VideoLLaMA [39], and Video-ChatGPT [18]. Despite being trained with identical datasets, our model outperforms Video-ChatGPT in four aspects. Our mothod offers a more comprehensive interpretation, and its responses are more in tune with the overarching context of the video content than comparable approaches. By utilizing the EDVT-Attention and sequential temporal modeling techniques, our model demonstrates a strong ability to generate text that is contextually appropriate, detailed, and includes precise timing for video inputs.

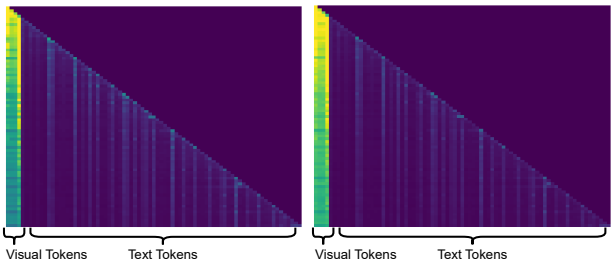
4.3. Abalation Study

Effect of Design Choices. We investigate the effect of our two designs, including the equal-distance to visual-tokens (EDVT) attention and the sequential visual projector. As shown in Figure 4, our model is tested on NExT-QA [30], which comprises three different types of questions. The baseline model employs LLaVA [14] as the language model, and Q-Former in BLIP-2 [12] as the visual projector. We first incorporate temporal modeling into the baseline, utilizing the projected tokens from the previous frame as the query tokens to generate visual tokens for the current frame, a setup denoted as "*w Seq*" in Figure 4. Further enhancing this version with EDVT-Attention, we introduce "*w Seq&EDVT*", ultimately forming our final VISTA-LLAMA. The results affirm the efficacy of our design across all question types.

EDVT-Attention. We here validate the effectiveness of our core EDVT-Attention. Table 3 reports the comparison results of different models in combination with our EDVT-Attention on NExT-QA [30]. Initially, we integrate the EDVT-Attention into Video-ChatGPT [18], where only the attention component is substituted with our EDVT-Attention

Method	NExT-QA [30]			
	Tem.	Cau.	Des.	Avg.
Video-ChatGPT [18]	37.6	65.1	54.9	54.6
+ **EDVT-Attention**	39.5 (+1.9)	72.8 (+7.7)	54.8	59.3 (+4.7)
VISTA-LLAMA (ViT-L-14)	34.0	69.1	42.2	53.6
+ **EDVT-Attention**	36.8 (+2.8)	72.2 (+3.1)	47.7 (+5.5)	56.5 (+2.9)
VISTA-LLAMA (EVA-CLIP-g)	34.3	65.8	55.9	54.1
+ **EDVT-Attention**	40.7 (+6.4)	72.3 (+6.5)	57.0 (+1.1)	59.7 (+5.6)

Table 3. **Comparison of EDVT-Attention design** with diffent visual encoders on NExT-QA [30] (§4.3). For clarity, accuracy of all questions and three types of questions are reported ("Tem.": *temporal*, "Cau.": *causal*, "Des.": *descriptive*, "Avg.": average).



(a) Vanilla attention (b) EDVT-Attention (Ours)

Figure 5. **Comparison of attention weights** for varing context lengths. Lighter colors represent higher weights. To improve clarity, we have combined visual token weights into the first four tokens. We recommend zooming in for optimal viewing.

during training. Further deployment of the EDVT-Attention into Video-ChatGPT [18] yields significant performance gains (*e.g.,* 65.1% escalates to 72.8% for causal questions). The accuracy is enlarged across all three setting types, indicating that our EDVT-Attention considerably enhances the multi-modal understanding in large language models.

We further illustrate the attention weights of both EDVT-Attention and conventional attention in Figure 5. For this experiment, we employ 128 visual tokens and integrate the attention weights into the first four tokens to enhance visualization. The attention weights assigned to visual tokens are markedly greater in our EDVT-Attention compared to those in conventional attention. This indicates that the impact of visual tokens on language tokens is considerably more substantial under the proposed EDVT-Attention, shedding light on why the performance notably improves with our design.

Visual Projector. The visual projector is trainable within the model. Therefore, we have examined the impacts of different visual projector designs, as shown in Table 4. We evaluated three visual projector variants. The Q-Former, initialized using BLIP-2 [12], considerably outperforms the Q-Former that was initialized with BERT [5]. This indicates that visual projectors, when pre-trained on image-text pairs, can significantly enhance video comprehension. The model's performance peaks when integrating the sequential design,

Visual projector	NExT-QA [30]			
	Tem.	Cau.	Des.	Avg.
Linear Projector	37.6	65.1	54.9	54.6
Q-Former (*BERT init.*)	35.2	62.7	49.2	51.8
Q-Former (*BLIP-2 init.*)	34.3	65.8	55.9	54.1
SeqQ-Former (*BLIP-2 init.*)	36.2	68.5	51.1	55.4

Table 4. **Comparison of different visual projectors** on NExT-QA [30]. The linear projector is initialized with pre-trained weights in LLaVa [14]. *BERT init.* and *BLIP-2 init.* indicate that the visual projector is initialized with weights from BERT [5] and BLIP-2 [12]. SeqQ-Former is the proposed sequential visual projector. See §4.3 for more details.

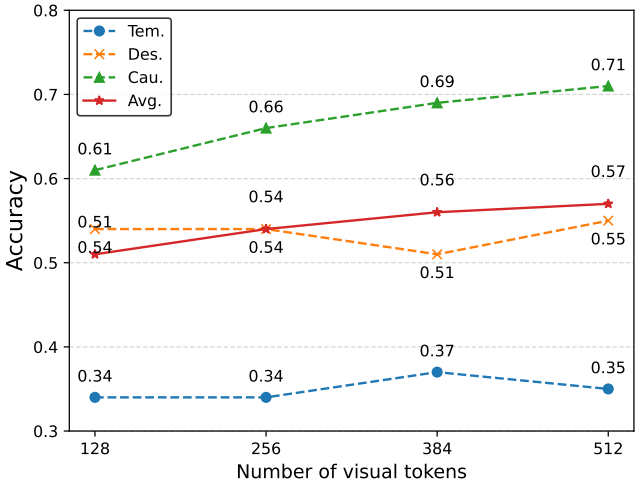


Figure 6. **The effect of training visual tokens** on NExT-QA [30]. Accuracy of all questions and three types of questions, including *temporal* (Tem.), *descriptive* (Des.), and *causal* (Cau.), are presented with different colors.

referred to as "SeqQ-Former" in the figure, outperforming all other visual projectors in terms of accuracy.

Number of Training Visual Tokens. We assess the impact of visual tokens on NExT-QA as illustrated in Figure 6. In this study, a frame is converted into 32 visual tokens using Q-Former. Frames were sampled at various timestamps throughout the training process. The overall accuracy improves as more visual tokens are incorporated into the language model. However, for temporal queries, precision declined with an increased number of visual tokens. The accuracy of temporal reasoning questions is notably lower than that of other question types. Temporal reasoning presents more difficulties, and the language model may not excel at the temporal modeling of visual tokens. In terms of descriptive question types, the accuracy steadily increases as the model gains more visual information.

Quantitative Results. We visualize the generated responses of Video-ChatGPT [18] and VISTA-LLAMA for different videos in Figure 7. The frames sampled at different times-

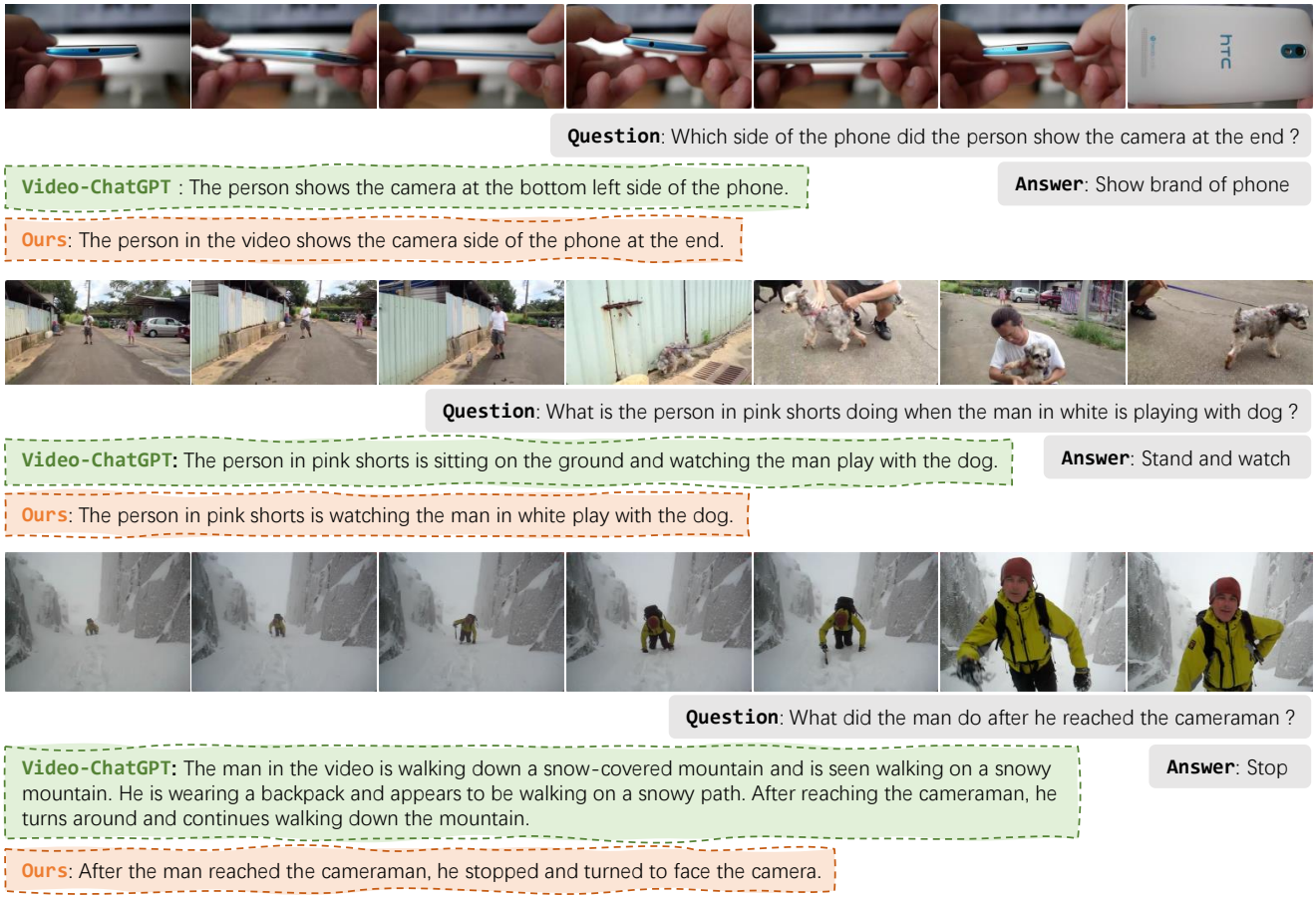


Figure 7. **Visualization results** on different video questions. The questions and annotated answers are located on the left side. The generated text from Video-ChatGPT [31] and our model is presented in the green and orange boxes, respectively. See §4.3 for more details.

tamps is presented, with questions listed below the images. In the first video, the mobile phone is turned, sequentially exposing each side over time. The question asks for the side of the mobile phone at the end of the video. Video-ChatGPT gives the wrong response, as the camera is never present at the bottom left side in the video. The hallucination occurs in Video-ChatGPT, whereas our model predicts the answer correctly. The responses in three cases demonstrate that Video-ChatGPT provides unrelated answers that do not correspond to the video content. In the second case, the person in pink is always standing while the man is sitting on the ground, but Video-ChatGPT incorrectly states that the person in pink is also sitting on the ground. In the third video, the man is consistently walking up the mountain and never walking down. However, Video-ChatGPT falsely outputs that the man continues walking down the mountain. In contrast, our model provides the correct and reliable responses. This demonstrates that our model significantly reduces hallucination in video understanding and delivers more accurate responses.

5. Conclusion

In this work, we present VISTA-LLAMA to improve the video understanding in large language model. A new vision-aware attention is introduced to maintain same relative distance between all visual tokens and language tokens. In addition, we propose a sequential visual projector to map video into the language space to enable temporal modeling. Experiments on several video question answering task demonstrate that our designs significant improves the current SoTA. Current work is built on the pre-trained image-text architecture, the power of our designs could be enlarged when applied in pre-training stage. Furthermore, the performance of our sequential encoder on longer videos can also be assessed. We would consider evaluating the model in more tasks and applied the attention mechanism to more vision language models.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (62293554, U2336212), and China Postdoctoral Science Foundation (524000-X92302).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, pages 23716–23736, 2022. 1, 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901, 2020. 2

[3] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics*, 2011. 5

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 2, 3

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7

[6] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023. 1

[7] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 5

[8] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694, 2020. 1

[9] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yezhou Yang, and Mike Zheng Shou. Mist : Multi-modal iterative spatial-temporal transformer for long-form video question answering. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14773–14783, 2022. 2

[10] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 2

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 5

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 1, 2, 4, 5, 6, 7

[13] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 5, 6

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. 1, 2, 3, 5, 6, 7

[15] Yu Lu, Ruijie Quan, Linchao Zhu, and Yi Yang. Zero-shot video grounding with pseudo query lookup and verification. *IEEE Transactions on Image Processing*, 33:1643–1654, 2024. 2

[16] Fan Ma, Linchao Zhu, and Yi Yang. Weakly supervised moment localization with decoupled consistent concept prediction. *International Journal of Computer Vision*, 130(5): 1244–1258, 2022. 3

[17] Fan Ma, Xiaojie Jin, Heng Wang, Jingjia Huang, Linchao Zhu, Jiashi Feng, and Yi Yang. Temporal perceiving video-language pre-training. 2023. 2

[18] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 2, 5, 6, 7

[19] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5

[21] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 2

[22] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 1, 5, 6

[23] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. 3

[24] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3, 5

[25] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2015. 3

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. 1, 2

[27] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, Felix Hill, and Zacharias Janssen. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021. 2

[28] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Luccioni, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 2

[29] Chenfei Wu, Sheng-Kai Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *ArXiv*, abs/2303.04671, 2023. 2

[30] Junbin Xiao, Xindi Shang, Angela Yao, and Tat seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9772–9781, 2021. 2, 3, 5, 6, 7

[31] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *ArXiv*, abs/2304.01196, 2023. 8

[32] D. Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. *Proceedings of the 25th ACM international conference on Multimedia*, 2017. 2, 3, 5, 6

[33] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 3, 5

[34] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022. 1, 5, 6

[35] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models, 2024. 2

[36] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178, 2023. 2

[37] Zhou Yu, D. Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019. 5, 6

[38] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 1

[39] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 5, 6

[40] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 1, 5, 6

[41] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022. 2

[42] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592, 2023. 2