VaQuitA: Enhancing Alignment in LLM-Assisted Zero-Shot Video Understanding

Anonymous authors
Paper under double-blind review

Abstract

Recent advancements in language-model-based video understanding have been progressing at a remarkable pace, spurred by the introduction of Large Language Models (LLMs). However, the focus of prior research has been predominantly on devising a projection layer that maps video features to tokens, an approach that is both rudimentary and inefficient. In our study, we introduce a cutting-edge framework, `VaQuitA`, designed to refine the synergy between video and textual information. At the data level, instead of sampling frames uniformly, we implement a sampling method guided by CLIP (Radford et al., 2021)-score rankings, which enables a more aligned selection of frames with the given question. At the feature level, we integrate a trainable Video Perceiver alongside a Visual-Query Transformer (abbreviated as VQ-Former), which bolsters the interplay between the input question and the video features. We also discover that incorporating a simple prompt, "Please be critical.", into the LLM input can substantially enhance its video comprehension capabilities. Our experiments show that `VaQuitA` consistently sets a new benchmark for zero-shot video question-answering tasks and is adept at producing high-quality, multi-turn video dialogues with users. The code will be released.

1 Introduction

The rise of deep learning tools for video interpretation has ushered in significant progress in video-centric tasks (Xu et al., 2021; Wang et al., 2022b; 2023). Yet, current models for video comprehension often falter when engaging in spontaneous discussions about video content (Zhong et al., 2022). A dialogue system rooted in video content can transform video searches, enhance monitoring techniques, and assist in summarizing pivotal events. Importantly, it offers a unified, accessible interface for video tasks, including action recognition, location identification, detection, retrieval, and tracking (Mu et al., 2023). This proficiency is especially noteworthy, highlighting the model's ability to understand temporal and spatial indications, grasp context, and perceive extended relationships (Liu et al., 2023d).

Existing research in Large Video Language Models (Yang et al., 2022; Zhang et al., 2023; Gao et al., 2023; Li et al., 2023b; Maaz et al., 2023; Liu et al., 2023c) predominantly adopts a uniform sampling strategy for frame selection. These models typically use a single projection layer to transfer and align video semantic content into the token space. The resulting tokenized video embeddings are then concatenated with query embeddings and fed into Large Language Models for response generation. However, this straightforward approach fails to adequately guide the projection of video features into specific text representations or sufficiently highlight which spatial or temporal aspects of the video should be emphasized. Given the constraints of limited training data, this methodology often leads to suboptimal performance in out-of-distribution video understanding tests (Maaz et al., 2023). In real-world scenarios, this can lead to perplexing errors in video conversation systems (Liu et al., 2023c).

To mitigate the above problems, we introduce `VaQuitA`, an innovative framework that redefines the approach to video and textual information integration. `VaQuitA` diverges from traditional methodologies by implementing a CLIP (Radford et al., 2021)-score guided frame sampling method. This innovation allows for the selection of frames that exhibit a higher relevance to the input question, thereby addressing the limitations of uniform frame sampling. The framework further advances the interaction between video content and textual queries through the integration of a trainable Video Perceiver. This component enhances the processing of video features, ensuring a more nuanced understanding of the visual content. Complementing this is our Visual-Query Transformer (VQ-Former), which acts as a pivotal element in aligning the video features with the textual query, facilitating a more coherent and context-aware interplay. Furthermore, `VaQuitA` incorporates a novel approach in its interaction with LLMs. By introducing a simple, yet effective prompt — "Please be critical." — into the LLM input during testing, we notice a marked enhancement in the model's capability to interpret video material. This refinement leads to a more critical and discerning analysis by the LLM, enhancing its performance in complex video understanding tasks.

In summary, the main contributions of the paper are:

- We propose `VaQuitA`, a novel video understanding model that strengthens the alignment of text features and video features. The alignment lies in both the raw data level and the feature level, which enhances the fusion of question and video information, leading to stronger reasoning ability of the video question answering model.
- We uncover the fact that adding an additional prompt, "Please be critical.", before the question can improve the understanding ability of `VaQuitA`.
- Our proposed `VaQuitA` achieved state-of-the-art performance on the Zero-shot Video Question Answering task. It can also conduct top-notch multi-turn conversations.

2 Related Works

We briefly summarize existing works in the related areas of video conversation, vision large lanugage models, and visual-text alignment.

Video Conversation. With the rapid development of LLMs, researchers begin to transfer their extraordinary reasoning abilities to the video conversation area (Song et al., 2023; Kim et al., 2024; Maaz et al., 2024; Liu et al., 2024). The SeViLA framework (Yu et al., 2023) leverages a single image-language model (BLIP-2 (Li et al., 2023a)) for both temporal keyframe localization and question answering in videos, with a novel method of chaining modules for cascaded inference and self-refinement without the need for expensive annotations. VideoChat (Li et al., 2023b) integrates foundational video models and LLM using a learnable neural interface, comprising two branches: VideoChat-Text which textualizes videos in real-time, and VideoChat-Embed which encodes video into embeddings using Video Foundation Models and Token Projection; the processed video content and questions are then passed to the LLM for generating answers. Video-LLaMA (Zhang et al., 2023) employs a multi-branch cross-modal pre-training approach, effectively achieving alignment between vision-language and audio-language. Nevertheless, both these two approaches have limited ability to handle long videos and have no quantitative results. Different from these, Video-ChatGPT (Maaz et al., 2023) develops a multimodal model that merges a video-adapted visual encoder with a large language model, capable of understanding and generating detailed conversations about videos, supported by a novel dataset of $100,000$ video-instruction dataset for video-based dialogue. More recently, BT-Adapter (Liu et al., 2023c) method extends image-language pretrained models into the video domain by acting as a plug-and-use temporal modeling branch next to the pretrained visual encoder, which is fine-tuned with the main backbone remaining unchanged. Despite the progress, the current video conversation capability is still limited due to the insufficient exploitation of question and video interplay.

Vision Large Language Models. Recent progress in computer vision has been propelled by the emergence of groundbreaking vision-language models. These models mark a considerable step forward in developing versatile vision models that can handle multiple tasks at once (Gupta et al., 2022; Maaz et al., 2022). A standout model in this realm is CLIP (Radford et al., 2021), trained on 400 million image-text pairs, showcasing exceptional zero-shot capabilities across many benchmarks. In more recent times, Flamingo (Alayrac et al., 2022) is a new family of Visual Language Models designed to rapidly adjust to novel tasks using a minimal number of annotated examples. It proposes perceiver resampler and gated cross-attention architectures, enabling its superior few-shot learning capabilities by training on large-scale multimodal web datasets with mixed text and

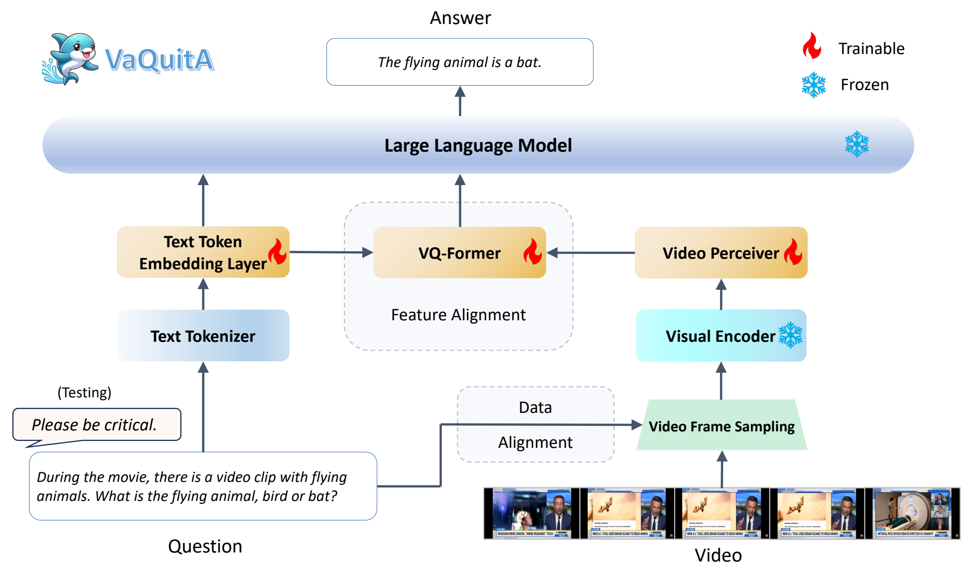


Figure 1: **Framework overview.** In response to a specific question, our framework begins by processing the input video with a sampling module that identifies key frames based on their relevance to the question's context. These frames are then processed by a pre-trained visual encoder to obtain spatio-temporal features. These features are subsequently refined into condensed embeddings by our newly developed Video Perceiver. In parallel, the question undergoes tokenization. Both the video and text embeddings are then synergized using our Visual-Query Transformer, which aligns the multimodal information more effectively. The resulting text-influenced video features are concatenated with the text embeddings and fed into the Large Language Model to generate the answer. During the testing phase, we propose to add an additional prompt, "Please be critical.", before the question for performance enhancement. The whole framework supports end-to-end training.

images. BLIP-2 (Li et al., 2023a) represents an effective method for pre-training that leverages existing image encoders and language models that have undergone pre-training, connecting them using a lightweight Querying Transformer in two stages: the vision-language representation and vision-to-language generative training. While some of these models are compatible with both images and videos, videos that are minute-level are yet challenging to process with input questions for accurate answers, and there is a growing demand for a robust large video-language model.

Visual-Text Alignment. Contemporary progress in aligning visual-text features primarily revolves around the concept of harmonizing multimodal features originating from various representational spaces. The foundational work (Duan et al., 2022) highlighted the challenges of aligning evolving features during training. Progressing from this, the Multi-Modality Cross Attention Network (Wei et al., 2020) and the "MVPTR" framework (Li et al., 2022b) both emphasized the significance of fine-grained feature alignment and cross-modal interactions, illustrating a shift towards more sophisticated semantic alignment tasks. Further innovations in multimodal fusion were proposed through CentralNet (Vielzeuf et al., 2018) presenting a multilayered integration approach, and ADAPT (Lin et al., 2022) which introduced dynamic action-based context alignment for Vision-Language Navigation, showcasing the practical application of alignment in autonomous systems. These developments culminated in the Multimodality-guided Visual Pre-training (MVP) approach (Wei et al.), which leveraged large-scale image-text datasets to refine the alignment process, marking a significant step forward in pre-training methodologies. Contrasting with existing approaches, our method enforces the alignment of video and text embeddings through a novel video feature resampling network and a bespoke cross-attention module tailored for the LLM input space. The resampling module enhances the alignment in the input raw data level and the cross-attention module strengthens the alignment in the feature learning level. Our approach represents an innovative direction in the field of large vision language modeling.

3 VaQuitA Framework

Our proposed VaQuitA framework consists of three novel components: Data Alignment module (Sec. 3.1), Feature Alignment module (Sec. 3.2), and test-time Prompt Engineering (Sec. 3.4). The entire pipeline is illustrated in Fig. 1.

3.1 Data Alignment

Existing methodologies typically employ a uniform sampling approach to extract frames for video conversation (Maaz et al., 2023; Bhattacharya et al., 2023; Yang et al., 2023a) or video understanding in general (Lin et al., 2019; Li et al., 2022a; Wang et al., 2022a). Such uniform sampling method, while straightforward, often results in the loss of critical information contained in the frames that are not selected, affecting the model's ability to understand videos effectively. To address this limitation, we present a new method in our VaQuitA that leverages the semantic similarity between the video frames and the question prompt for frame selection. This technique ensures a more congruent alignment between the features of the question and those of the frames at the raw data level. We refer to this as the "Data Alignment" module.

CLIP Feature Similarity-based Frame Selection for Training. Given the input video of L frames in total, instead of getting a certain number of frames with only uniform sampling, we also select frames based on the similarity between the frame features and the input query. Suppose we sample T frames in total, we propose to select $\frac{T}{2}$ frames uniformly over the temporal dimension and another $\frac{T}{2}$ frames using the similarity-based approach. Specifically, we extract the text feature of the query using CLIP model, denoted as f_{query}, and the visual features of the remaining frames that are not selected as $\{f_{\text{video}}^1, f_{\text{video}}^2, \cdots, f_{\text{video}}^{L-\frac{T}{2}}\}$, the similarity is calculated as

$$\text{Cosine-Similarity}(f_{\text{query}}, f_{\text{video}}^i) = \frac{f_{\text{query}} \cdot f_{\text{video}}^i}{\|f_{\text{query}}\|_2 \times \|f_{\text{video}}^i\|_2}, \tag{1}$$

and we select the indices of the top $\frac{T}{2}$ values. The motivation is that uniform sampling will lead to information loss due to its non-adaptivity, and by employing the proposed similarity-based approach, frames that are most related to the question will be selected, improving representation learning ability. Our proposed sampling strategy is different from some more recent works, *e.g.*, using motion importance (Zhi et al., 2021), and dynamic sampling (Zheng et al., 2020), which are based on the inherent properties of the video frame statics. Our sampling approach makes the first attempt to employ the prompt-frame similarity as guidance for frame sampling, which bridges the two modalities for effective modeling.

Uniform Sampling for Testing. Our proposed sampling method is implemented during the training phase of our model. For the testing or inference stage, we revert to uniform sampling due to efficiency considerations and the need for speed in real-world applications. This approach is enough to demonstrate satisfactory performance in our experimental evaluations. We supplement an example in the supplementary showing that our sampling approach can improve testing performance as well compared with uniform sampling, despite being slower.

3.2 Feature Alignment

Visual data are regarded as the reflection and capture of the physical world while text data can be seen as the abstract of the understanding of the world and the fundamental logic (LeCun, 2022). The successful alignment of visual and textual information is significant for an intelligent system to work appropriately. Instead of directly concatenating the tokenized text and visual features to put into LLM (Liu et al., 2023b; Maaz et al., 2023; Zhang et al., 2023; Chen et al., 2023; Liu et al., 2023c), we propose a novel Visual-Query Transformer, abbreviated as VQ-Former, to produce text-guided video embeddings before concatenation with the text embeddings. The inspiration comes from recent work on visual-text pretraining (Li et al., 2023a; Alayrac et al., 2022), and the illustration of the architecture is provided in Fig. 2. Notice that although the self-attention mechanism in LLM already interacts text tokens with visual tokens to some extent by treating visual tokens as language tokens, our proposed feature alignment module treat text and video features as different domain

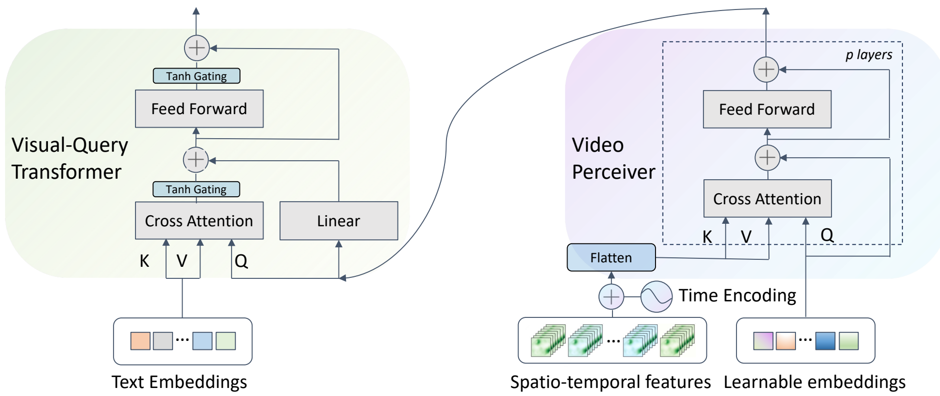


Figure 2: **Feature alignment**. The extracted spatio-temporal features of the video clip first go through Video Perceiver for representative embedding extraction, and are afterwards sent to Visual-Query Transformer for interleaving with text embeddings.

features through cross-attention operations. This practice enhance the input visual token qualities by computing attention and attending over input query tokens.

3.2.1 Video Perceiver

Given a sampled video snapshot, we first apply a pretrained CLIP model to extract semantic features for each frame. Suppose the extracted spatio-temporal feature embeddings are $F \in \mathbb{R}^{T\times n\times d}$, where T is the sampled frame number, n is the number of features for each frame, *i.e.*, the patch number for CLIP model, and d denotes the dimension of feature. To facilitate the alignment with text embeddings and input into the LLM, we need to resample and reduce the number of video features for computation feasibility. Inspired by Perceiver Resampler (Alayrac et al., 2022), we put forward Video Perceiver which transforms the spatio-temporal visual attributes into a number of learned output tokens. The spatio-temporal features are first added by Time Encodings of shape $T \times 1 \times d$ to store the sequence order information and then flattened to shape $Tn \times d$ for the cross-attention module dimension match. This cross-attention module employs a collection of learned latent vectors to query (Q), while the keys (K) and values (V) combine the flattened spatio-temporal visual attributes with these learned latent vectors. The shape of the learned latent embeddings is $m \times d$, where m denotes the number of latent embeddings. The weights of the learned latent embeddings are randomly initialized. Following transformer (Vaswani et al., 2017), feed-forward networks and residual connections are added for efficient modeling. The output embedding shape remains the same as the input learnable embeddings, *i.e.*, $m \times d$. We use p to denote the number of layers of the Video Perceiver.

3.2.2 Visual-Query Transformer

The input question goes through a text tokenizer and a text token embedding layer and turns into query embeddings, which, together with the learnable embeddings output of the video perceiver, are sent into VQ-Former. The layers derive their queries from vision features, whereas the keys and values originate from the language inputs. Visual-Query Cross-Attention layer is applied for the query feature (denoted as $X \in \mathbb{R}^{l\times d_{text}}$) and video feature (denoted as $M \in \mathbb{R}^{m\times d}$) interleaving, where l is the length and d_{text} is the text embedding dimension.

Visual-Query Cross Attention. In the Visual-Query Cross Attention layer, we adopt a multi-head mechanism as in Transformer (Vaswani et al., 2017). We denote the head index as h and the inner feature dimension of each head as d_h. Given the input learned video feature, we first apply Layer Normalization (Ba et al., 2016) and denote the normalized one as M. Note that although M is learned, the existence of the Time Encodings guarantees that the temporal information of input frames is kept. Then we have the Q, K, Vs for each head calculated as

$$Q^{(h)} = MW_Q^{(h)}/s_q, K^{(h)} = XW_K^{(h)}, V^{(h)} = XW_V^{(h)}, \tag{2}$$

and

$$O_a^{(h)} = \texttt{Softmax}(Q^{(h)}{K^{(h)}}^\top)V^{(h)}W_O^{(h)}, \tag{3}$$

where $W_Q^{(h)} \in \mathbb{R}^{d \times d_h}$, $W_K^{(h)} \in \mathbb{R}^{d_{text} \times d_h}$, $W_V^{(h)} \in \mathbb{R}^{d_{text} \times d_h}$ and $W_O^{(h)} \in \mathbb{R}^{d_h \times d_{text}}$ are learnable weight parameters of head h. s_q is a scaler representing the scale parameter, and h represents the head index. Denoting the Visual-Query Cross Attention layer output as O_a, we have that O_a is the multi-head concatenation of each head output:

$$O_a = \text{Concat}(O_a^{(1)}, \cdots, O_a^{(H)}), \tag{4}$$

where H denotes the head number. The dot product attention computation aligns the semantics of the video embedding M and query embedding X, contributing to the selection and learning of the visual features more relevant with the question. The multi-head design enables the exploration of the weight parameters in more feature subspaces for superior representation learning (Vaswani et al., 2017).

VQ-Former Overview. We use `Cross_Attn` to denote the Visual-Query Cross Attention, and the entire procedure of our VQ-Former can be written as:

$$O_a = \texttt{Cross\_Attn}(M, X), M' = O_a \cdot \tanh(g_{\text{attn}}) + MW_M, \tag{5}$$

$$O_f = \texttt{Feed\_Forward}(M'), M'' = O_f \cdot \tanh(g_{\text{ff}}) + M'. \tag{6}$$

Here the learnable parameter $W_M \in \mathbb{R}^{d \times d_{text}}$ is applied to transform the dimension of the video representative features into token feature dimension for the residual architecture. Here $\tanh$ denotes Hyperbolic Tangent function and `Feed_Forward` denotes a Feed Forward net block containing 2 linear layers with Layer Normalization and GELU (Hendrycks & Gimpel, 2016) activation layer. g_{attn} is Attention Gate and g_{ff} is FeedForward Gate, which are both learnable scalar parameters borrowed from Flamingo (Alayrac et al., 2022) for improved stability and performance. Eventually, the output question-interacted video features M'' are input to the LLM together with the input question embeddings. Different from the existing visual-text interleaving architectures, *e.g.*, Q-Former (Li et al., 2023a) or Gated Cross-Attention layer (Alayrac et al., 2022), our VQ-Former converts visual features to Queries and text features to Keys and Values for attention value computation. The underlying rationale of our approach is to utilize the information from the query as a directive to enhance the learning of pivotal visual embeddings. **This is significantly different from existing literature where visual features are converted to Keys and Values and text features are converted to Queries (Alayrac et al., 2022; Li et al., 2023a).** Also, the output of the VQ-Former is concatenated to the question, which is also different from existing works where the output is directly sent to the Language Models.

3.3 End-to-end Training

Our `VaQuitA` supports end-to-end training: the trainable parameters include the Text Token Embedding Layer, the VQ-Former, and the Video Perceiver. The visual encoder (CLIP) and the Large Language model weights are derived from pretrained weights and are frozen during our training. The CLIP model employed to extract f_{query} and f_{video} is also frozen during the training. We employ the standard smoothed Negative Log-Likelihood Loss in NLP literature.

3.4 Prompt Engineering

Prompt engineering (Wei et al., 2022; Zhou et al., 2022; Gu et al., 2023) refers to the systematic design and modification of input prompts to guide machine learning models, particularly pretrained LLMs, to produce desired or more accurate outputs. The essence of this technique is rooted in the understanding that the input provided to a model doesn't merely serve as a query but also as a form of soft guidance, potentially shaping the model's behavior and outputs. In our experiments, we are excited to discover that in the testing phase, if we add a prompt "Please be critical." before the question, zero-shot question answering performance can be significantly and consistently improved. This might imply an intriguing principle that, unlike in question answering in NLP the models are demanded to be calmer or more organized (Kojima et al., 2022; Yang et al., 2023b), the model needs to be more critical or judgmental for video question answering tasks. An ablation study on the prompts is carried out in Sec. 4.3.2 which verifies the implication.

Table 1: Zero-Shot question-answering performance comparison of VaQuitA with other models. Our VaQuitA demonstrates SOTA performance across all examined datasets.* denotes the results reported in Maaz et al. (2023) and † denotes the results reported in Liu et al. (2023c). The best performance in **bold** and the second best underlined.

Model	MSVD-QA		MSRVTT-QA		Activity Net-QA	
	Acc. (↑)	Score (↑)	Acc.(↑)	Score(↑)	Acc.(↑)	Score (↑)
FrozenBiLM* (Yang et al., 2022)	32.2	–	16.8	–	24.7	–
VideoLLaMA† (Zhang et al., 2023)	51.6	2.5	29.6	1.8	12.4	1.1
LLaMA-Adapter† (Gao et al., 2023)	54.9	3.1	43.8	2.7	34.2	2.7
Video Chat* (Li et al., 2023b)	56.3	2.8	45.0	2.5	26.5	2.2
Video-ChatGPT* (Maaz et al., 2023)	64.9	3.3	49.3	2.8	35.2	2.7
BT-Adapter† (Liu et al., 2023c)	67.0	3.6	51.2	2.9	46.1	3.2
LLaMA-VID (Li et al., 2023c)	70.0	3.7	58.9	3.3	47.5	**3.3**
Vista-LLaMA (Ma et al., 2023)	65.3	3.6	60.5	3.3	48.3	**3.3**
Video Chat 2 (Li et al., 2024)	70.0	**3.9**	54.1	3.3	49.1	**3.3**
Video-LaVIT (Jin et al., 2024)	73.2	**3.9**	59.3	3.3	**50.1**	**3.3**
Video-LLaVA (Lin et al., 2024)	70.7	**3.9**	59.2	**3.5**	45.3	**3.3**
VaQuitA (Ours)	**74.6**	3.7	**68.6**	3.3	48.8	**3.3**

4 EXPERIMENTS

In the experimental implementation, we employ Llama 2 (7B) (Touvron et al., 2023b) as the foundational LLM backbone and initialize its weight using the weights of LLaVA-1.5 (Liu et al., 2023a). We fine-tune the trainable parameters in VaQuitA using the video instruction dataset VideoInstruct-100K[1] (Maaz et al., 2023), comprising roughly 100,000 pairs of video instructions. The fine-tuning phase spans three epochs, utilizing a step size of value $2e-5$ and a total batch size of value 32. For fair comparison, we keep the data-level hyperparameters as the same in literature: $T = 100, d = 1024, d_{text} = 4096$,. We employ the "clip-vit-large-patch14" CLIP version for video feature extraction. Specifically, for the sampling-period features f_{query} and f_{video}, we use the last layer of the CLIP model with dimension 768. For the video feature extraction before the video perceiver, we utilize the last but one layer of CLIP with patch number $n = 256$ and feature dimension $d = 1024$. We chose $m = 356$ in Video Perceiver, which is the same as the dimension after spatio-temporal pooling in Video-ChatGPT (Maaz et al., 2023) for a fair comparison. The perceiver depth is set as $p = 1$. For all the attention blocks in both Video Perceiver and VQ-Former, we set $d_h = 64, H = 8$ and scale parameter $s_q = 8$. All the training experiments are conducted on eight A100 80GB GPUs. For testing, one GPU with 15 GB GPU memory is sufficient.

4.1 ZERO-SHOT VIDEO QUESTION ANSWERING

We carry out an exhaustive quantitative assessment using several prevalent open-ended video question-answer datasets, encompassing MSRVTT-QA (Xu et al., 2017), MSVD-QA (Xu et al., 2017), and Activity Net-QA (Yu et al., 2019). Following Maaz et al. (2023), the assessments are performed in a zero-shot setting, utilizing GPT-guided evaluation to gauge the model's proficiency. This assessment method calculates the precision of the model's predicted outputs (accuracy) and ranks them on a 1-5 scale (score). To ensure a fair comparison with the baselines, we employ Azure GPT-3.5-turbo API for evaluation, which is consistent with Maaz et al. (2023). Our VaQuitA's efficacy is juxtaposed with other notable models, namely FrozenBiLM (Yang et al., 2022), VideoLLaMA (Zhang et al., 2023), LLaMA-Adapter (Gao et al., 2023), Video Chat (Li et al., 2023b), Video-ChatGPT (Maaz et al., 2023), BT-Adapter (Liu et al., 2023c), LLaMA-VID (Li et al., 2023c), Vista-LLaMA (Ma et al., 2023), Video Chat 2 (Li et al., 2024), Video-LaVIT (Jin et al., 2024) and Video-LLaVA (Lin et al., 2024). From Table 1, we can draw the conclusion that VaQuitA achieves state-of-the-art performance in both accuracy and score across the three benchmark datasets.

4.2 MULTI-ROUND CONVERSATION

The experiments conducted predominantly address scenarios involving a singular question and answer. However, in practical applications such as Copilot or assistants for industrial products, the capacity for multi-round conversations is crucial for user experience. To evaluate this aspect, we

[1] https://huggingface.co/datasets/MBZUAI/VideoInstruct-100K

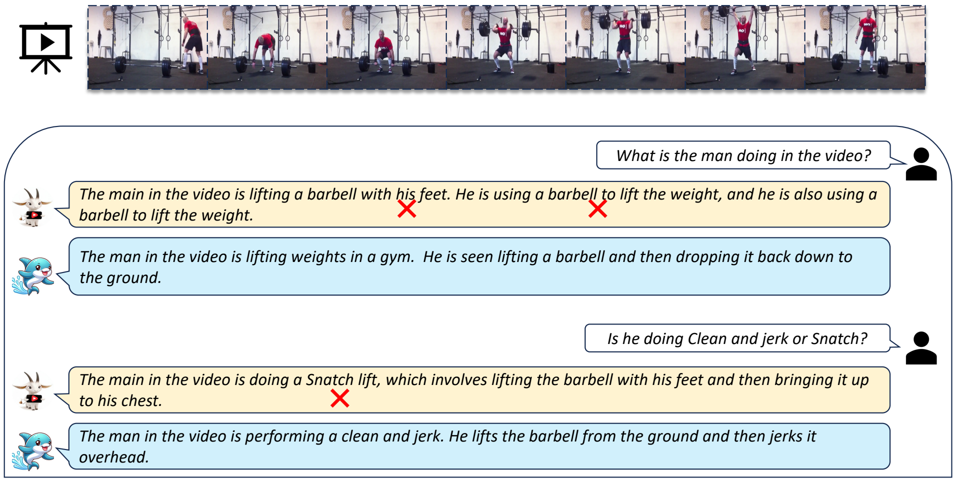


Figure 3: Given a video clip of a man lifting weights, we ask questions on what the man is doing and whether he is doing Clean and Jerk or Snatch. Our `VaQuitA` answers both the questions correctly. While the baseline Video-ChatGPT (Maaz et al., 2023) generates a repetitive answer to the first question, seeming somewhat chaotic, and fails to discriminate that the man is doing a Clean and Jerk, rather than Snatch.

Table 2: Ablation of the components of `VaQuitA`. FA, DA, and PE signify Feature Alignment, Data Alignment, and Prompt Engineering. A. denotes accuracy and S. demotes score.

FA	DA	PE	MSVD		MSRVTT		Activity	
			A.	S.	A.	S.	A.	S.
✗	✗	✗	65.1	3.3	49.9	2.8	42.5	3.0
✗	✗	✓	65.8	3.3	50.5	2.9	43.9	3.1
✗	✓	✗	64.5	3.2	50.8	2.9	44.9	3.1
✗	✓	✓	65.9	3.3	52.8	3.0	45.7	3.1
✓	✗	✗	70.8	3.5	59.7	3.1	47.4	3.1
✓	✗	✓	71.0	3.5	60.3	3.1	47.8	3.2
✓	✓	✗	74.4	3.7	68.5	3.3	47.7	3.3
✓	✓	✓	**74.6**	**3.7**	**68.6**	**3.3**	**48.8**	**3.3**

Table 3: Comparison of our cross-attention computation approach and the traditional approach. Here t-Q-v-KV denotes that text features serve as queries and video features serve as keys and values; v-Q-t-KV denotes that video features serve as queries and text features serve as keys and values. A. denotes accuracy and S. denotes score.

t-Q-v-KV	v-Q-t-KV	DA	PE	MSVD		MSRVTT		Activity	
				A.	S.	A.	S.	A.	S.
✓		✗	✗	66.8	3.3	52.7	2.9	44.3	2.9
	✓	✗	✗	70.8 (+4.0)	3.5 (+0.2)	59.7 (+7.0)	3.1 (+0.2)	47.4 (+3.1)	3.1 (+0.2)
✓		✗	✓	67.4	3.3	54.1	3.0	45.2	3.0
	✓	✗	✓	71.0 (+3.6)	3.5 (+0.2)	60.3 (+6.2)	3.1 (+0.1)	47.8 (+2.6)	3.2 (+0.2)
✓		✓	✗	67.9	3.4	56.2	3.0	45.3	3.0
	✓	✓	✗	74.4 (+6.5)	3.7 (+0.3)	68.5 (+12.3)	3.3 (+0.3)	47.7 (+2.4)	3.3 (+0.3)
✓		✓	✓	68.2	3.4	56.5	3.0	45.8	3.1
	✓	✓	✓	**74.6** (+6.4)	**3.7** (+0.3)	**68.6** (+12.1)	**3.3** (+0.3)	**48.8** (+3.0)	**3.3** (+0.2)

compare the multi-round conversation capabilities of `VaQuitA` with one of the baselines Video-ChatGPT (Maaz et al., 2023). As depicted in Fig. 3, `VaQuitA` demonstrates consistently more accurate and comprehensive conversational abilities compared to Video-ChatGPT. This highlights `VaQuitA`'s potential for industrial applications. More video dialogue examples are provided in the supplementary.

4.3 Ablation Studies

4.3.1 Ablation of the components of VaQuitA

We perform ablation studies w.r.t. the components of `VaQuitA` including the Data Alignment, Feature Alignment, and Prompt Engineering. We conduct experiments on all three Video Question Answering datasets. We also list the baseline results using Llama 2 as the LLM backbone without all three modules, in which one MLP layer is used to project the visual embeddings to the token space as Maaz et al. (2023). Tab. 2 shows that Data Alignment, Feature Alignment, and Prompt Engineering all contribute to zero-shot video QA perfor-

Table 4: Ablation of Video Perceiver and VQ-Former in FA.

Dataset	VP	VQ-Former	Acc. (↑) Score (↑)
MSVD/MSRVTT/Activity	✗	✗	65.3/51.0/45.73.2/2.9/3.1
	✓	✗	68.5/59.4/46.93.4/3.1/3.2
	✗	✓	70.9/62.3/47.43.5/3.1/3.2
	✓	✓	**74.6/68.6/48.83.7/3.3/3.3**

mance. On one hand, the performance of merely adopting Feature Alignment performs better than merely adopting Data Alignment without Prompt Engineering, implying that feature-level learning is comparatively more significant than input data selection for our task. On the other hand, with Prompt Engineering, the model will degrade a lot without Data Alignment. In addition, Prompt Engineering improves the performance in all cases. We also conduct experiments ablating the Video Perceiver (VP) and VQ-Former respectively. When ablating VP, we directly pool the CLIP features along spatial dimension and temporal dimension and then concatenate the features as Maaz et al. (2023) does. As is shown in Tab. 4, both VP and VQ-Former contribute much to the performance and the combination of them leads to the best results. We further compare our strategy of converting video features to Quries and prompt features to Keys and Values with the strategy of converting prompt features to Queries and video features to Keys and Values. The architecture is kept the same and the results in Tab. 3 indicate that our approach manifests obvious superiority whether DA or PE is used consistently. **This largely arises from the fact that because the output of the cross-attention layer is concatenated with the prompt tokens to be sent to LLM, video is the primary context for which we want to enhance or refine representations using information from textual modality.** Therefore video features should serve as Queries while text features serve as Keys and Values.

4.3.2 Ablation of Hyperparameters

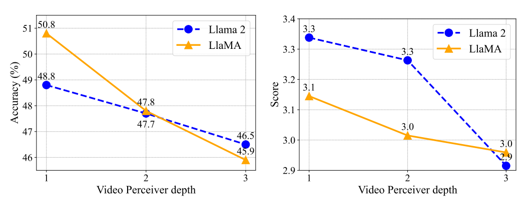
We further study the effects of changing the hyperparameter values in our `VaQuitA` framework. We conduct the ablation studies on the Activity Net-QA testing dataset.

Similarity-based Sampling Frame Number. We employed a mixed strategy of sampling to focus on question-related frames while looking broadly for performance stability. CLIP features are not perfect and it is likely that the CLIP-similarity selected frames are not the places of interest. We provide additional experiments by changing the similarity-based sampling frame number. As shown in Tab. 5, sampling completely uniformly or completely based on CLIP feature similarities gives the inferior performance. **This means that the model should both look broadly and focus on certain frames of interest across the temporal dimension to reach the best performance in the data input phase.**

Table 5: Ablation of similarity-based sampling frame #.

Length	0	20	40	60	80	100
Acc. (↑)	47.8	48.2	48.5	**49.0**	48.1	47.5
Score (↑)	3.2	3.3	3.3	**3.3**	3.3	3.2

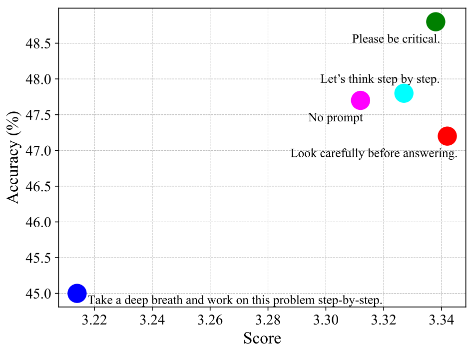
Video Perceiver Depth & Pretrained Model. We try using multiple layers in Video Perceivers and using the LLaMA (Touvron et al., 2023a) model with weight initialization from LLaVA (Liu et al., 2023b). As illustrated in Fig. 4, the accuracy of `VaQuitA` drops when the layer number p of the Video Perceiver increases for both LLaMA and Llama 2 backbone. This might largely result from the small training epoch we use and the limited size of training data. For the LLM weights initialized from LLaVA and LLaVA-1.5, we find that the performance gap is not as large as expected, and using LLaMA (LLaVA-1.5) pretrained weights with one layer in Video Perceiver even achieve 50.8 accuracy on Activity Net-QA dataset. On the other hand, models initialized using LLama 2 are obviously more robust to the perceiver depth and are significantly better in relative score evaluation.



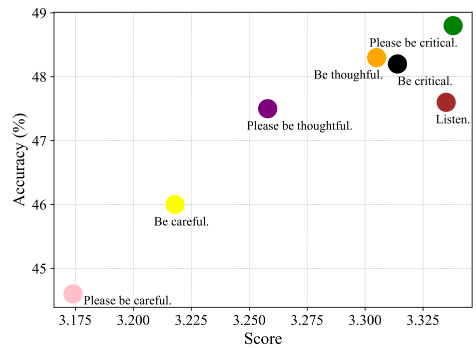
(a) Accuracy w.r.t. p. (b) Score w.r.t.p.

Figure 4: Performance on Activity Net-QA (Yu et al., 2019) using pretrained LLama 2 (Touvron et al., 2023b) and LLaMA (Touvron et al., 2023a).

Prompt Engineering Design. We ablate the prompt added before the question. We compare our designed prompt with two popular instruction prompts in the NLP filed: "Take a deep breath and work on this problem step-by-step." (Yang et al., 2023b) and "Let's think step by step." (Kojima et al., 2022). We also compare with another prompt "Look carefully before answering." and indicate the performance when not adding a prompt. From the accuracy and score results shown in Fig. 5a, we can draw the conclusion that our designed prompt, "Please be critical.", performs the best with both the highest accuracy and the highest score. "Let's think step by step." (Kojima et al., 2022)



(a) Comparison with the classic prompts in NLP.



(b) Comparison with the prompts with $\leq$ 3 words.

Figure 5: Accuracy and score results on Activity Net-QA (Yu et al., 2019) dataset of different prompt designs.

improves the performance slightly while "Take a deep breath and work on this problem step-by-step." (Yang et al., 2023b) degrades the performance. We conduct additional experiments with different prompts that have less than or equal to 3 words: "Please be careful." (pbc), "Please be thoughtful." (pbt), "Be critical." (bcr), "Be thoughtful." (bt), "Be careful." (bca), "Listen." (l). As shown in Tab. 5b, our prompt "Please be critical." exhibits the best performance. Also, "Be critical." also exhibits nearly excellent performance, which implies that the word "critical" is significant.

5 Conclusion and Limitation

Our proposed VaQuitA represents a significant stride in video understanding. By moving away from traditional frame sampling methods and adopting a CLIP-score guided technique, we have achieved a more nuanced and effective integration of video frame and text data. The innovative combination of a trainable video perceiver with a visual-query transformer mechanism allows for a dynamic interplay between video features and input questions, further augmented by the strategic use of prompts. The results clearly demonstrate that VaQuitA not only excels in zero-shot video question-answering tasks but also in generating coherent and contextually rich multi-turn video dialogues. Our VaQuitA therefore sets a new standard for LLM-based video understanding.

One limitation of our work is the reliance of the model on pre-trained models like CLIP (Radford et al., 2021) for feature extraction, which might limit the framework's adaptability to other domains or tasks where such pre-trained models are not available or effective. There are two solutions: one is to train the vision encoder weights as well in the instruction tuning, and the other is to pre-train the vision encoder using large-scale video/image text pairs related to the target domains or tasks. We leave the adaption to other tasks of our model without using fixed CLIP encoder for future work.

REFERENCES

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019.

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. *NeurIPS*, 35:23716–23736, 2022.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Aanisha Bhattacharya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot. *arXiv preprint arXiv:2305.09758*, 2023.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, 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*, pp. 961–970, 2015.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023.

Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15651–15660, June 2022.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16399–16409, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *arXiv preprint arXiv:2403.18406*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Yann LeCun. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. *https://openreview.net/pdf?id=BZ5a1r-kVsf*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022a.

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*, 2023b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023c.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4641–4650, 2016.

Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4395–4405, 2022b.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *EMNLP*, 2024.

Bingqian Lin, Yi Zhu, Zicong Chen, Xiwen Liang, Jianzhuang Liu, and Xiaodan Liang. Adapt: Vision-language navigation with modality-aligned action prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15396–15406, 2022.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7083–7093, 2019.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

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

Ruyang Liu, Chen Li, Yixiao Ge, Ying Shan, Thomas H Li, and Ge Li. One for all: Video conversation is feasible without video instruction tuning. *arXiv preprint arXiv:2309.15785*, 2023c.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. *arXiv preprint arXiv:2404.00308*, 2024.

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023d.

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

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European Conference on Computer Vision*, pp. 512–531. Springer, 2022.

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.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

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. pp. 8748–8763. PMLR, 2021.

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.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022a.

Yizhou Wang, Can Qin, Yue Bai, Yi Xu, Xu Ma, and Yun Fu. Making reconstruction-based method great again for video anomaly detection. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 1215–1220. IEEE, 2022b.

Yizhou Wang, Dongliang Guo, Sheng Li, and Yun Fu. Towards explainable visual anomaly detection. *arXiv preprint arXiv:2302.06670*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

L Wei, L Xie, W Zhou, H Li, and Q Tian. Mvp: Multimodality-guided visual pre-training." arxiv, mar. 10, 2022. doi: 10.48550. *arXiv preprint arXiv.2203.05175*.

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Dejing 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. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

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.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, pp. 10714–10726, 2023a.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023b.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023.

Zhou Yu, Dejing 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

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.

Yin-Dong Zheng, Zhaoyang Liu, Tong Lu, and Limin Wang. Dynamic sampling networks for efficient action recognition in videos. *IEEE transactions on image processing*, 29:7970–7983, 2020.

Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pp. 1513–1522, 2021.

Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

VaQuitA: Enhancing Alignment in LLM-Assisted Zero-shot Video Understanding

Supplementary Material

6 RAW VIDEOS OF SEC. 4.2

We supplement the raw videos of the two examples in Sec. 4.2 of the main paper, namely "multi_round_example_1.mp4" and "multi_round_example_2.mp4". They are chosen from the test set of ActivityNet-200 (Caba Heilbron et al., 2015) dataset.

7 VAQUITA ASSISTANT DEMO

We also provide a video demo recording of our VaQuitA Assistant on Gradio (Abid et al., 2019) ("VaQuitA_demo.mp4"). The three example videos are chosen from the test set of TGIF (Li et al., 2016), Social-IQ 2.0 and ActivityNet-200 (Caba Heilbron et al., 2015) datasets. The videos are about a boy falling down the skateboard on a ramp, a doctor and patient talking to each other in the hospital and a man shaving himself in the bathroom, respectively. We show in our demo recording that the VaQuitA Assistant is able to generate high-quality multi-round conversations at a high responding speed. It is able to precisely summarize the content of a video, identify the relationships between characters and events, and pinpoint locations.

8 TEST-TIME DATA ALIGNMENT

The illustration of our proposed frame sampling approach is shown in Fig. 6. We conduct an additional experiment using our proposed sampling approach in Sec. 3.1 during the inference stage. We use the Video-ChatGPT (Maaz et al., 2023) trained model and only change the sampling way in inference. The baseline is uniform sampling. Given a video clip of MSNBC news report (video given at "test_time_da_example.mp4"), we ask a video question: "During the movie, there is a video clip with flying animals. What is the flying animal, bird or bat?" for 3 independent times. The correct answer is "bat", which corresponds to 2:16-2:22 time stamp of the video. For uniform sampling, the model answers: "The flying animal in the video is a bird." for 3 times, which is wrong; for our proposed sampling method, the model answers: "The flying animal in the video is a bat" for 3 times, which is correct.

This superiority of our Data Alignment module mainly results from the CLIP Feature Similarity-based Frame Selection component, which is verified by checking the selected frames. We supplement the directories of the sampled frame of uniform sampling and our data alignment sampling method. The sampled frames using uniform sampling are stored under directory "uniform_sampled_frames" and the sampled frames using our proposed sampling method are under directory "ours_sampled_frames". We can see that the uniform sampling only samples one frame ("frame_4223.jpg") related to the question, while our proposed sampling method samples 13 related frames ("frame4197.jpg", "frame4198.jpg", "frame4201.jpg", "frame4206.jpg", "frame4207.jpg", "frame4247.jpg",
"frame4280.jpg", "frame4281.jpg", "frame4282.jpg", "frame4287.jpg",
"frame4197.jpg", "frame4288.jpg", "frame4289.jpg", "frame4304.jpg"). Since our sampling method samples more frames corresponding to the question, the model can answer more correctly, which reflects the effect of Data Alignment in the inference phase.

9 MORE MULTI-ROUND CONVERSATION EXAMPLES COMPARED WITH VIDEO-CHATGPT (MAAZ ET AL., 2023)

From Fig. 7 to Fig. 13, we supplement more multi-round conversation examples in addition to Sec. 4.2 with their raw videos, namely

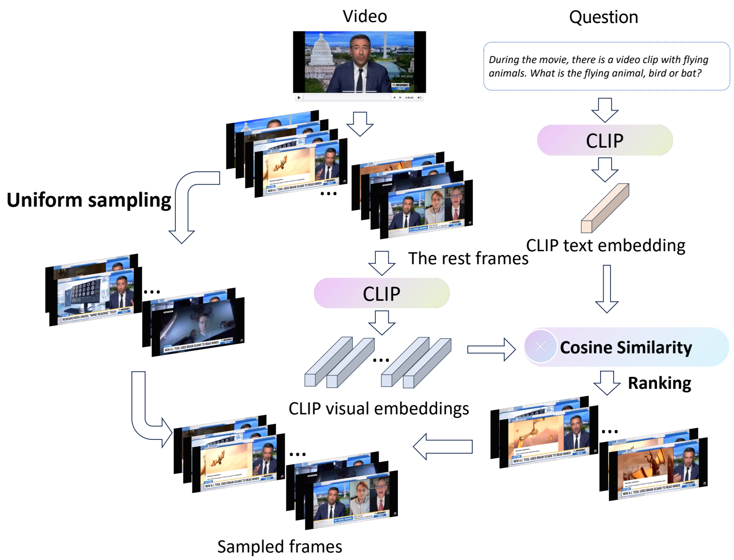


Figure 6: **Data alignment.** Our proposed sampling module consists of both uniform sampling and similarity-based sampling for the training process. Best viewed in color.

"multi_round_example_3.mp4", "multi_round_example_4.mp4", "multi_round_example_5.mp4", "multi_round_example_6.mp4", "multi_round_example_7.mp4", and "multi_round_example_8.mp4". They are chosen from the test set of ActivityNet-200 (Caba Heilbron et al., 2015) dataset. It is shown that `VaQuitA` generates consistently better answers and exhibits significantly stronger video understanding ability compared to Video-ChatGPT (Maaz et al., 2023).

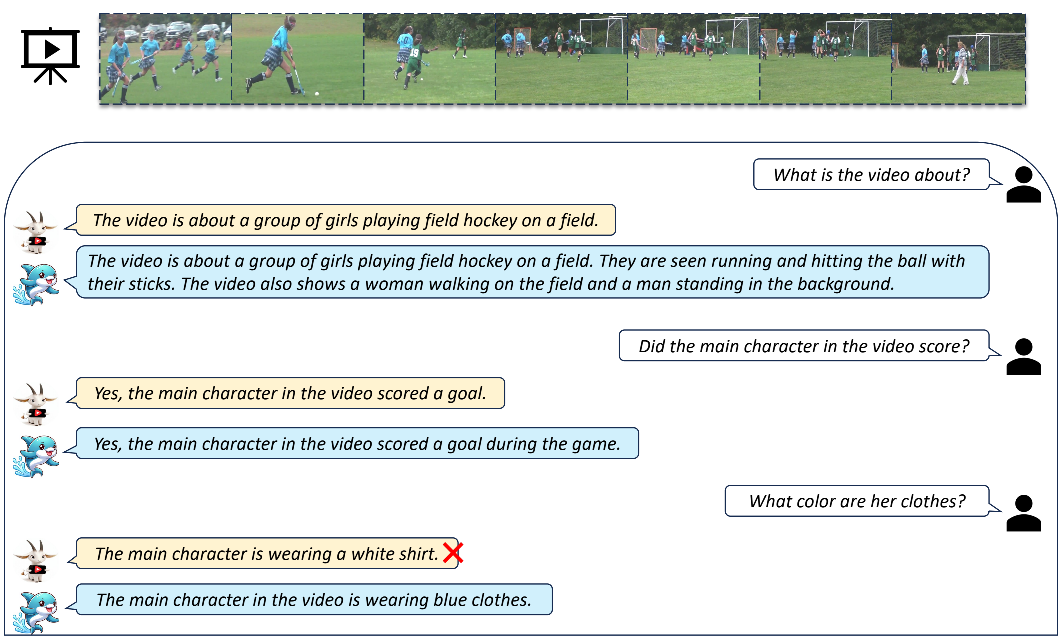


Figure 7: Given a video clip on a group of girls playing field hockey, we ask questions on the content of the video, whether the main character scores, and the color of the main character's clothes. Our `VaQuitA` can answer all the questions correctly while the baseline Video-ChatGPT (Maaz et al., 2023) fails to tell the correct color of the clothes the girl is wearing (marked by red cross). In addition, the generated answers of `VaQuitA` are more detailed and specific like a human chatting with the user, while the responses of Video-ChatGPT are short and like being forced to complete a task. Best viewed in color.

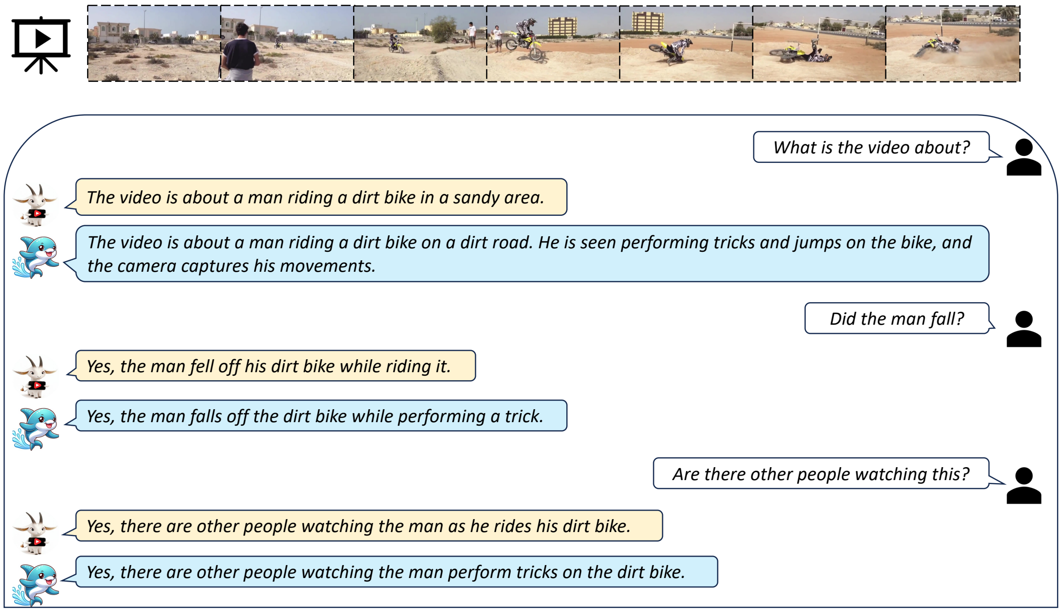


Figure 8: Given a video clip on a man riding a dirt bike, we ask questions on the content of the video, whether the man fell, and whether there are other people watching this. Our VaQuitA can answer all the questions correctly and can identify that the person was trying to do a trick when riding, which leads to his falling. In contrast, the baseline Video-ChatGPT (Maaz et al., 2023) fails to tell that the man fell when doing a trick when riding. In addition, the generated answers of VaQuitA are more detailed and specific. Best viewed in color.

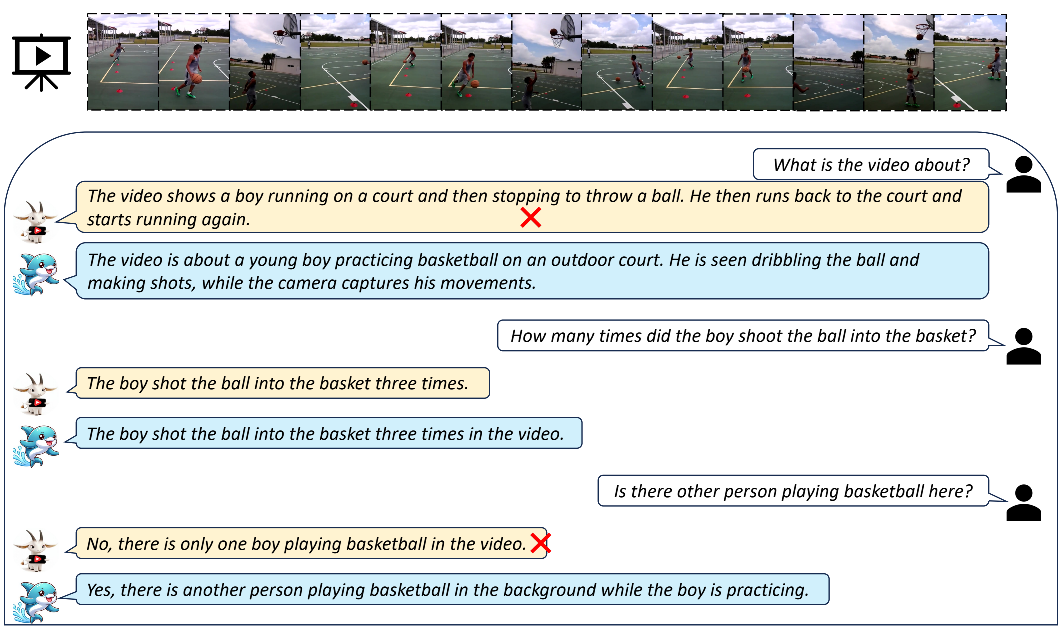


Figure 9: Given a video clip on a boy playing basketball on an outside court, we ask questions on the content of the video, how many times the boy shot the ball, and whether there is another person playing basketball there. Our VaQuitA can identify that the boy was practicing basketball with dribbling and shooting and tell that there is another man playing basketball behind the boy. In contrast, the baseline Video-ChatGPT (Maaz et al., 2023) fails to identify the movements of the boy well and gives the wrong answer on the last question. In addition, the generated answers of VaQuitA are more detailed and specific. Best viewed in color.

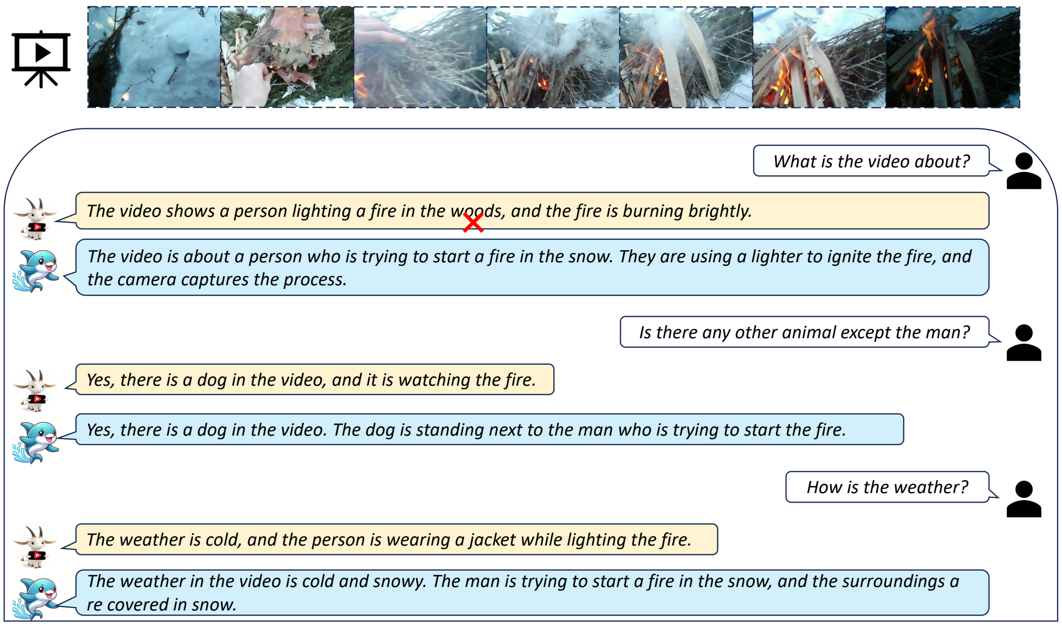


Figure 10: Given a video clip of a person making a fire in the snow, we ask questions on the content of the video, whether there is another animal, and how the weather is. Our VaQuitA can answer all the questions correctly and can identify that it is snowy and the fire is started in the snow. In contrast, the baseline Video-ChatGPT (Maaz et al., 2023) fails to tell the snowy weather and mistakenly identifies that the fire is lit in the woods. In addition, the generated answers of VaQuitA are more detailed and specific.

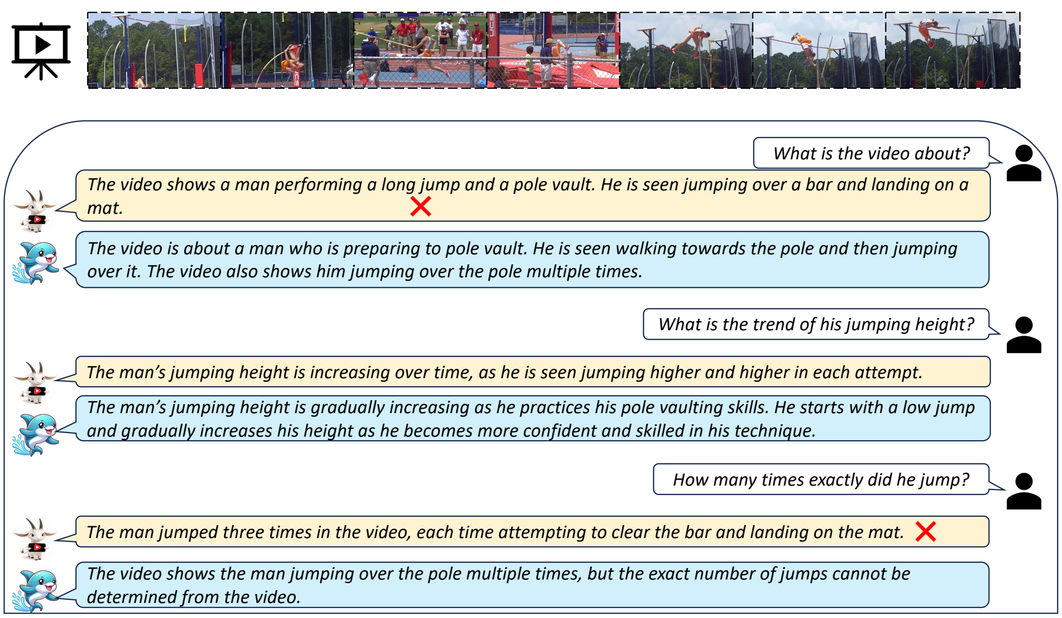


Figure 11: Given a video clip of a man doing pole vault, we ask questions on the content of the video, the trend of his jumping height, and the times of his jumping. Our VaQuitA can answer most of the questions correctly, except the third one as the man jumped four times in total. The baseline Video-ChatGPT (Maaz et al., 2023) gives the wrong answer to the third question, either. The advantage of VaQuitA is that for the first question on the content of the video, it can identify that the man jumped over the pole multiple times, which is impressive. Note that it is bearable to fail to remember how many times in total the man jumped, which is also challenging even for a human being. In addition, the generated answers of VaQuitA are more detailed and specific.

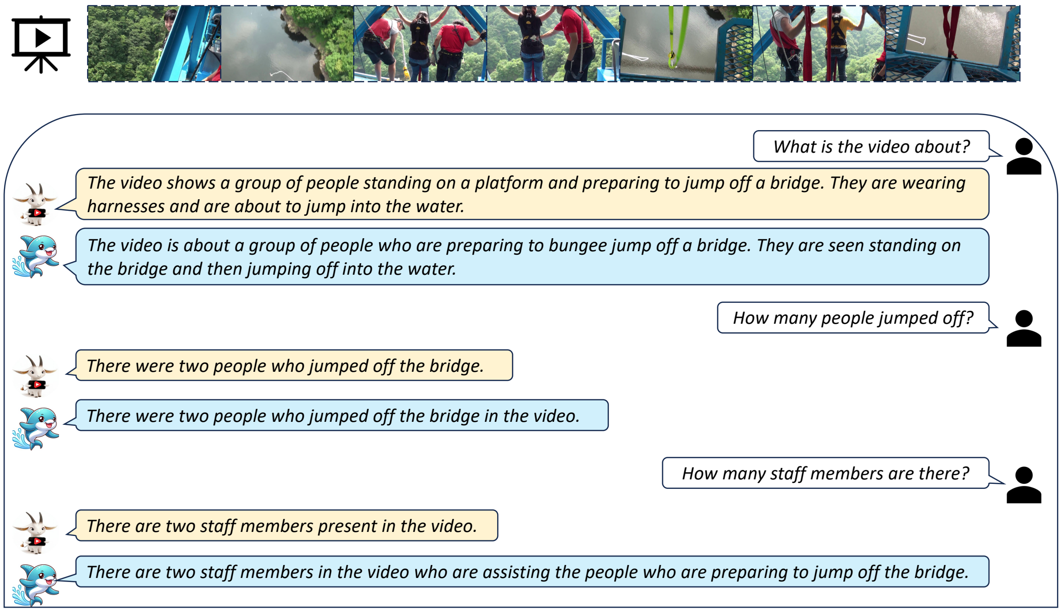


Figure 12: Given a video clip of a group of people doing bungee jump off a bridge, we ask questions on the content of the video, the number of people jumped, and the number of staff members. Our `VaQuitA` can answer all the questions correctly and can identify that the sports is bungee jump. In contrast, the baseline Video-ChatGPT (Maaz et al., 2023) fails to tell the name of the sports the people are doing. In addition, the generated answers of `VaQuitA` are more detailed and specific.

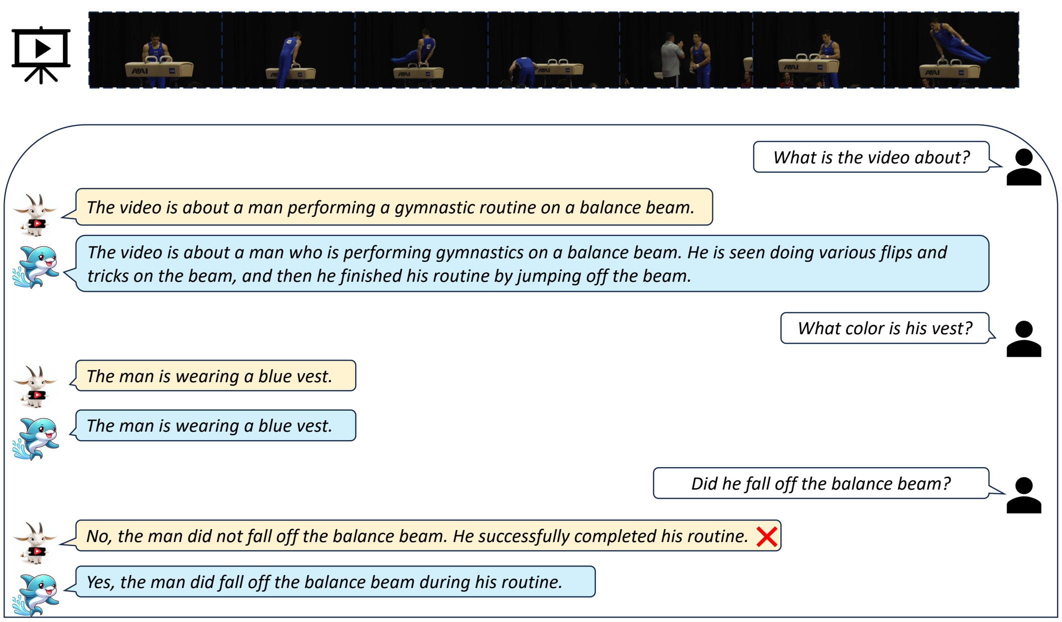


Figure 13: Given a video clip of a man performing gymnastics on a balance beam, we ask questions on the content of the video, the color of his vest, and whether he fell off the balance beam. Our `VaQuitA` can answer all the questions correctly. In contrast, the baseline Video-ChatGPT (Maaz et al., 2023) fails to find that the man actually fell off the balance beam once. In addition, the generated answers of `VaQuitA` are more detailed and specific in describing the movements of the gymnastic, especially for the first question.