Video-Context Aligned Transformer for Video Question Answering

Linlin Zong[1], Jiahui Wan[1], Xianchao Zhang[1], Xinyue Liu[1]*, Wenxin Liang[1], Bo Xu[2]

[1] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province,
School of Software, Dalian University of Technology, Dalian 116620, China
[2]School of Computer Science and Technology, Dalian University of Technology
{llzong, xczhang, xyliu, wxliang, xubo}@dlut.edu.cn, wanjiahui1011@gmail.com

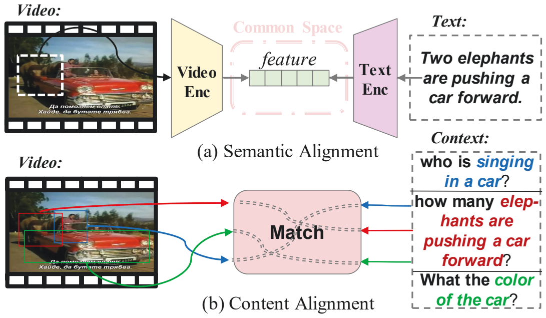
Abstract

Video question answering involves understanding video content to generate accurate answers to questions. Recent studies have successfully modeled video features and achieved diverse multimodal interaction, yielding impressive outcomes. However, they have overlooked the fact that the video contains richer instances and events beyond the scope of the stated question. Extremely imbalanced alignment of information from both sides leads to significant instability in reasoning. To address this concern, we propose the Video-Context Aligned Transformer (V-CAT), which leverages the context to achieve semantic and content alignment between video and question. Specifically, the video and text are encoded into a shared semantic space initially. We apply contrastive learning to global video token and context token to enhance the semantic alignment. Then, the pooled context feature is utilized to obtain corresponding visual content. Finally, the answer is decoded by integrating the refined video and question features. We evaluate the effectiveness of V-CAT on MSVD-QA and MSRVTT-QA dataset, both achieving state-of-the-art performance. Extended experiments further analyze and demonstrate the effectiveness of each proposed module.

Introduction

Video Question Answering (VideoQA) is a challenging task within the multimodal learning domain, aiming to understand videos and answer questions (Zhong et al. 2022). VideoQA not only requires precise semantic understanding of both the video and the question, but also need effectively interactions to locate the most critical feature within the video with rich spatiotemporal information.

Currently, the mainstream paradigm for video question answering is to encode the video and question using separate pre-trained models (He et al. 2016; Hara, Kataoka, and Satoh 2018; Devlin et al. 2018), and then fuse the visual and textual features through a complex interaction model for classifying the final answer. Many existing works have exhibited powerful video content modeling abilities and achieved high performance on multi-modal information interaction. They input extracted features into spatio-temporal (Jin et al. 2021; Jiang et al. 2020; Dang et al. 2021; Gao et al. 2023), hierarchical (Le et al. 2020; Liu et al. 2021; Peng et al. 2022; Xiao et al. 2022; Dang et al. 2021), multiscale (Peng et al. 2022; Guo et al. 2021), or multi-granularity (Xiao et al. 2022) structured models, and generate useful answers through complex and precise interactions. Furthermore, some researchers have focused on causal analysis (Li et al. 2022b,a; Yu et al. 2023; Zang et al. 2023), answering questions by finding relevant information in the video. Although achieving promising performance improvement, there exist two main challenges that hinder precisely matching the exact answers. The first challenge lies in the semantic unalignment of visual contents in videos and textual contents in questions. To obtain accurately matched answers, semantics in video and question need to be precisely aligned into the same common space, and then generate the exact answer for the question. As shown in Figure 1(a), the text 'two elephants are pushing a car forward' and its corresponding video event should be semantically aligned into the same feature space for effectively question answering.



(a) Semantic Alignment

(b) Content Alignment

Figure 1: The two challenges of video question answering are semantic alignment, which refers to ensuring that the semantics of the video and related text are consistent in the encoded features, and content alignment, which entails matching the questions and related content in the video.

The second challenge lies in the content unalignment within videos that contain much richer information than a single question sentence. More complete coverage of video information provides a deeper understanding than a single question sentence that only aligns with a specific event within the video. To improve the content alignment, multiple sentences can be used to offer a more comprehensive cover-

*Corresponding author.

age of video contents, which provides a large space for answer acquisition. Existing methods directly align the video and question, requiring the model to correspond vastly informative videos with information-scarce question-answer pairs. This extremely imbalanced alignment of information from both sides leads to significant instability in reasoning. Therefore, we propose to first expand the information volume on the textual side to align it with the video, and then further obtain accurate answers through single-question interaction with the video. This gradually progressive video refinement approach is smoother, resulting in more stable alignment between the two modalities, which is shown in Figure 1(b).

To address the above mentioned challenges, we treat the VideoQA task from a perspective of reading comprehension (Sun et al. 2022). Intuitively, To accelerate human reading comprehension, one effective approach is to quickly skim through all the questions, and then read the article while bearing in mind an understanding of these questions. This helps pinpoint the most important areas in the article that require attention. Afterwards, a more thorough reading of the current question allows for a speedy and correct answer. Thus, drawing on this thought process, we propose a Video Context Alignment Transformer (V-CAT), utilizing all the questions related to videos as the context. Firstly, the pre-trained models and a trainable encoder are used to extract each single modality features. Then, contrastive learning are used to semantically constrain global features of both the video and context, achieving semantic alignment for both modalities. Afterward, aligning the contents of the video with the context, we preliminarily extract the vital information that requires attention. Finally, interact granularly with the video by the answer decoder, and use the global token of question to classify answer.

Our work makes the following three contributions:

- We revise the existing traditional paradigm and proposed V-CAT, the Video-Context Alignment Transformer method. By introducing context, we balance the informational levels of both modalities and achieve preliminary semantic and content alignment between the video and context. We then proceed to interact the current granular question and refined video content to predict answer.

- We propose a semantic alignment method based on contrastive learning. By utilizing contrastive loss on global features of videos and contexts within the same batch, the semantic matching features are pulled closer while the non-matching features are pushed apart, effectively enhancing the effectiveness of semantic alignment.

- We conduct experiments on the traditional datasets MSVD-QA (Xu et al. 2017) and MSRVTT-QA (Xu et al. 2016), and obtained state-of-the-art performance, demonstrating the effectiveness of our proposed method. In addition, we conducted further analysis on each module through extended experiments, confirming the remarkable ability enhancement brought about by each module.

Related Work

In recent years, the VideoQA paradigm has mainly followed a three-step process (Zhong et al. 2022): firstly, extracting features using pre-trained models; secondly, performing feature interaction between videos and questions; and finally, classifying answers in open-ended manner. Typically, pre-trained models in the computer vision field, such as ResNet (He et al. 2016) and ResNeXt (Hara, Kataoka, and Satoh 2018), are used to extract video features, while word embedding vector like Glove (Pennington, Socher, and Manning 2014) or pre-trained models such as Bert (Devlin et al. 2018) are used to extract question features in natural language processing. In recent works, various attempts have been made to improve the most critical interaction process of the model. Due to the time and space dimensions in videos, some works have attempted to model them spatiotemporally(Jin et al. 2021; Jiang et al. 2020; Dang et al. 2021; Gao et al. 2023), focusing on the relationship between video clips and regions. To further improve the model's ability to learn high-order semantic information, some works have modeled the structure hierarchically(Le et al. 2020; Liu et al. 2021; Peng et al. 2022; Xiao et al. 2022; Dang et al. 2021). In these structures, the model can learn different semantic characteristics at different levels, making the learning process more enriched. Similarly, some works have extracted multi-scale(Peng et al. 2022; Guo et al. 2021) or multi-granularity(Xiao et al. 2022) features from videos for interaction with questions. Previous studies have concluded that videos contain more diverse and informative instances and events than a sentence (Lin et al. 2022). Although these sophisticated models have strengthened the representation ability of video features, excessive redundant video information may cause instability during interaction with questions. Therefore, some works(Li et al. 2022b,a; Yu et al. 2023; Zang et al. 2023) have attempted to locate key video clips using causal analysis to filter out irrelevant information and obtained satisfactory results. With the rise of pre-trained models, some works (Xue et al. 2022; Zellers et al. 2021; Seo, Nagrani, and Schmid 2021) hope to enhance the model's semantic alignment and generalization ability by pre-training on large-scale datasets, followed by fine-tuning on VideoQA subtasks. However, these methods require a considerable amount of dataset and training resources, and lack interpretability. Additionally, due to the extremely unequal distribution of semantic information between videos and question answers, abrupt alignment methods are highly unstable and prone to failure.

Our approach differs from previous work in that we attempt to address the lack of information on the textual side to achieve balance in aligning features of the two modalities as much as possible. We introduce all video-related questions as context, initially aligning them with semantic and content of the video. We then progressively interact the current question and refined video to obtain the final answer. Despite using simple modules compared to previous work, our method still achieves decent performance because of stable and smooth semantic and content alignment between the video and text.

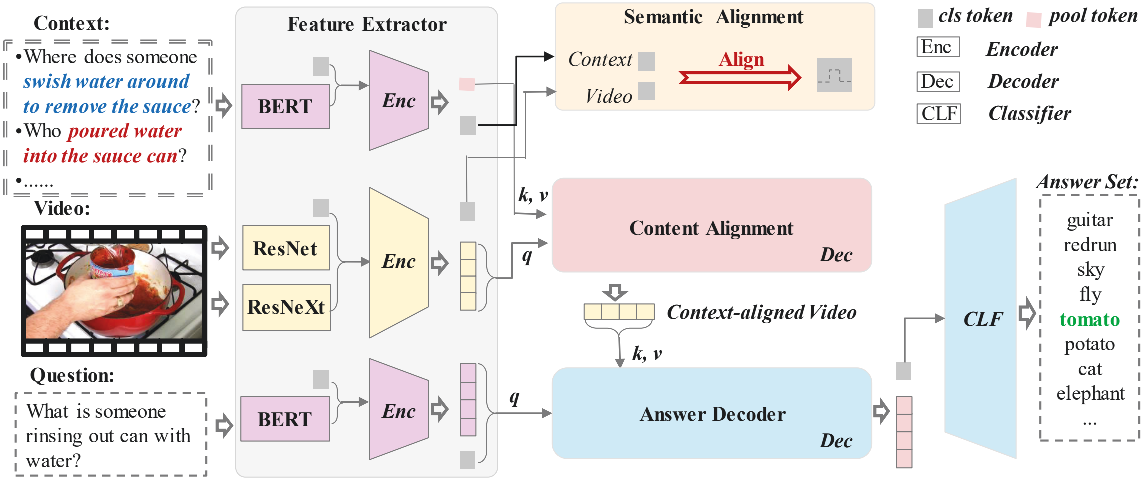


Figure 2: The V-CAT model. All the questions related to videos are utilized as the context. Firstly, the pre-trained model is used to extract the text and video features. Then, a trainable encoder and contrastive learning is used to semantically constrain global features of both the video and context, achieving semantic alignment for both modalities by the encoder. Afterward, aligning the content of the video with the context, we preliminarily extract the vital information that requires attention. Finally, the video and the question interact at a fine-grained level by a decoder, after that the global token is utilized to classify answers.

Method

As shown in Figure2, V-CAT consists of four main modules: a feature extractor module, which extracts video and question features separately; a semantic alignment module, which encodes the videos and questions and applies contrastive learning to constrain video and question features, thereby enhancing the semantic alignment of the encoder; a content alignment module, which enables the capture of key video features in the given context and ensures the stability of subsequent training; and a decoding module, which generates more precise answers using question features.

Feature Extractor

Pre-trained model In this module, the conventional approach involves using pre-trained models to extract video and text features separately. First, we sample an equal number of frames from the video, following the previous work and adhering to a simple and efficient principle. We choose the traditional ResNet(He et al. 2016) and ResNeXt(Hara, Kataoka, and Satoh 2018) models to extract appearance and motion features from each frame of the video, denoted as $v_a \in R^{f \times d_v}$ and $v_m \in R^{f \times d_v}$ respectively. The variables f and d represent the number of frames sampled per video and the dimensionality of the embedded features, respectively. Next, we concatenate the two types of features at the frame level to obtain the video feature $v \in R^{2f \times d_v}$. For extracting text features, we utilize a pre-trained BERT model(Devlin et al. 2018). We perform pooling on the extracted context features in order to capture the global contextual information. This feature is represented as $c \in R^{1 \times d_q}$. And the question feature is denoted as $q \in R^{l \times d_q}$. Here, l represents the maximum length of a single sentence, and d_q represents the feature dimension. It is worth mentioning that these three feature extractors are not involved in the subsequent training process.

Encoder The pre-trained models are trained in their respective modal domains, resulting in significant semantic discrepancies in the learned feature representation spaces. To ensure that the encoded semantic features of both modalities are more consistent, we introduce a learnable semantic alignment encoder after the pre-trained models. Through this encoder, we effectively project the video and question features to a common semantic space, achieving alignment between the visual and textual modalities.

We design the encoder based on the encoder of the Transformer model (Vaswani et al. 2017). For visual features, we first use a linear layer to project the features to the model dimension. Additionally, before inputting them into the encoder, we introduce a learnable token v_g to capture global information for semantic alignment. Then, we incorporate positional encoding to capture the relative temporal position information of each frame. After normalization, we obtain visually features that are more uniformly standardized. These operations can be represented by

$$v = LN([vW, v_g] + pos) \tag{1}$$

where pos represents the positional encoding, LN denotes the normalization operation, and $W \in R^{d_v \times d}$ is trainable parameter, d denotes the hidden size of model.

Next, through self-attention and feed-forward networks, we obtain visual features v with contextual information. The formulas for the attention mechanism and feed-forward network are as

$$v_{i+1} = FFN(MHA(v_i, v_i, v_i)) \tag{2}$$

where FFN is feed-forward network and MHA represents multi-head attention. FFN formulation are denoted as

$$FFN(x) = LN(ReLU(xW_1)W_2 + x) \tag{3}$$

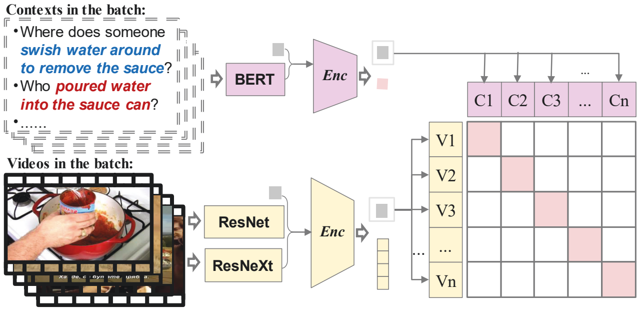

Contexts in the batch:
•Where does someone swish water around to remove the sauce?
•Who poured water into the sauce can?
•......
BERT
Enc
C1
C2
C3
...
Cn
Videos in the batch:
ResNet
ResNeXt
Enc
V1
V2
V3
...
Vn

Figure 3: An illustration of contrastive alignment. A higher similarity is expected in the alignment tokens of the same sample's video and context within a batch, while a lower similarity is preferred for different samples.

where $W_i \in R^{d \times d}$ is trainable parameter, $ReLU$ represents the activation function. Also, MHA are denoted as

$$MHA(Q, K, V) = Softmax(\frac{(QW_1)(KW_2)^T}{\sqrt{t}})(VW_3) \tag{4}$$

where $W_i \in R^{d \times d}$ is trainable parameter, $Softmax$ is a activation function, t denotes as the temperature of attention. In this process, Q, K, and V are all video features v.

For context and question features, we similarly use linear layers to transform the features to the model dimension. Then, we introduce learnable tokens q_g and c_g to capture global information for both context and question. And we obtain the textural features that are more uniformly standardized, which can be denoted as

$$q \quad = LN([qW_q, q_g] + pos), \tag{5}$$

$$c \quad = LN([cW_c, c_g] + pos) \tag{6}$$

where W_q and W_c are learnable parameters. Afterward, similar to visual features, they are input into the encoder for processing. The formulas for this process can be represented as

$$q_{i+1} \quad = FFN(MHA(q_i, q_i, q_i)), \tag{7}$$

$$c_{i+1} \quad = FFN(MHA(c_i, c_i, c_i)). \tag{8}$$

Semantic Alignment Module

In the encoder module, we have connected an encoder after the pre-trained models to encode the visual and textual modality features separately, ensuring that their feature representations have consistent semantics. However, since the information contained in a single question text is much less than that in a video, the model may not always learn semantic-aligned features as desired. To better align the two modalities and draw inspiration from the work of multi-modal alignment (Radford et al. 2021a; Tsimpoukelli et al. 2021; Hou et al. 2022; Li et al. 2023), we utilize context with richer textual information for semantic aligning with the video, and introduce contrastive learning to supervise the learning process of the encoder. This encourages semantically similar multimodal features to be closer in the feature space, while pushing semantically dissimilar features further apart. Specifically, we adopt a contrastive learning approach as shown in the Figure 3, where the global alignment token of the video and the global alignment token of the context are normalized and then used to compute a large similarity matrix. The formula for this process can be represented as

$$v_g \quad = \quad v[0], \tag{9}$$

$$c_g \quad = \quad c[0], \tag{10}$$

$$sim_{ij} \quad = \quad v_g^i \cdot (c_g^j)^T \tag{11}$$

where symbol $[i]$ represents the i-th token in the sequence, i and j denotes the number of v and c in a batch.

It is evident that the video and the context from the same sample are more related, while different samples within a batch are unrelated. Therefore, we naturally desire the diagonal of this similarity matrix to be 1, while the other positions are 0. To achieve this, we adopt a loss calculation method inspired by (Radford et al. 2021a) and construct sample labels as

$$z' = I(bsz) \tag{12}$$

where $I(x)$ denotes a Identity matrix with the size of x, bsz denotes the size of batch. Then, we use cross-entropy to achieve the goal of semantic similarity differentiation, allowing the features of the two modalities to be embedded into a common feature space, which represents as

$$L_{cl} = -\Sigma_{i=1}^{bsz}(z_i')^T Sim_i. \tag{13}$$

Unlike simple embedding features in the past, we further extract video and context global features as interface, to achieve video-context semantic alignment by utilizing the contrastive learning. Consistent features in a shared feature space facilitate subsequent interactions between two modalities.

Content Alignment Module

After obtaining semantically aligned visual and textual information, previous work (Lin et al. 2022) has shown that videos often contain richer instance and event information compared to individual sentences. However, for video question answering tasks, redundant information can often lead to interference when answering questions, resulting in instability during modality interaction. For answering the current question, we only need video information relevant to the question. Refined visual features can enable faster and more accurate modality interaction.

Here, to better align with human thinking, we first roughly examine the context, which allows us to identify the video information of interest. Therefore, we use the query token c_q of context as the key and value in the cross-attention module, while the video features serve as the query. This enables us to use the contextual semantic information to search for and aggregate video features while reducing the attention on redundant and irrelevant video features. It also avoid the instability of alignment between video and single sentence. This way, we can obtain more refined and effective visual features. Specifically, we interact the localized video features with the contextual content alignment features through a decoder. The decoder consists of cross-attention, and feed-forward networks. The formulas for this process can be represented as

$$c_q = c[1], \tag{14}$$

$$v_l = v[1 :], \tag{15}$$

$$v_l^{i+1} = FFN(MHA(v_l^i, c_q, c_q)). \tag{16}$$

Answer Decoder

After obtaining the refined video information that can be used to answer the question, we can proceed to answer the question using this content. Inspired by previous work (Sun et al. 2022), video question answering is similar to reading comprehension. After a rough examination of the context to identify key content, we re-examine the question to provide an answer. By using question features, we can facilitate more comprehensive modality interaction and obtain more accurate answers.

Therefore, here we use a decoder module consisting of cross-attention, and feed-forward networks to interact between the visual and textual modalities and decode the answer. We use word-level question features, including learnable global features, as the query, and the video features as the key and value, which are input into the decoder. The formula for this process can be represented as follow:

$$q_{i+1} = FFN(MHA(q_i, v_l, v_l)). \tag{17}$$

Next, we use the global token of the question, which encapsulates rich multimodal interactions, as input to the subsequent answer generation module. It can be represented as follows:

$$q_g = q[0], \tag{18}$$

$$p = ELU(q_g W_1)W_2 \tag{19}$$

where $W_1 \in R^{d\times d}$ and $W_2 \in R^{d\times m}$ denotes the trainable parameters, m denotes the size of answer set, ELU represents the activation function. The obtained p can be used for predicting various answers. In the training process, cross-entropy loss is used:

$$L_{ce} = -\Sigma_{i=1}^{m} z_i ln(p_i) \tag{20}$$

where z_i = 1 if the answer index corresponds to the ith sample's ground-truth answer and 0 otherwise. Finally, our construct overall loss:

$$L = L_{ce} + \alpha L_{cl} \tag{21}$$

where α denotes the weight of loss.

Experiment

Datasets

We experiment on the traditional and widely used datasets in the video question answering domain, MSVD-QA(Xu et al. 2017) and MSRVTT-QA(Xu et al. 2016). They are both open VideoQA datasets constructed using videos and descriptions. The videos mainly consist of short videos, and MSVD-QA contains 1,970 short videos with 50,505 open-ended Q&A pairs. MSRVTT-QA includes 10,000 videos with 243,000 Q&A pairs. These two datasets are both open-ended, which are more challenging compared to the current multiple-choice question answering datasets. They require models to have a strong ability to understand both visual and textual modalities and generate answers from a large condstructed answer set accordingly.

Implementation Details

In our experiments, each video is uniformly sampled into segments of 16 frames. For videos with insufficient frames, we pad using either the initial or terminal frame. In order to balance the information of video and text to better alignment, we introduce the context, which possess richer information than one sentence. Video context can be constructed by many method such as video description and video comment. In our work, we obtain context by concatenating extracted features from all questions related to current video and utilizing pooling operation. It is worth mentioning that, there are no same video in both train set and test set, which avoid the leakage of test set problems. The extracted visual and textual features possess dimensions of 2048 and 768, respectively, while the model's dimension stands at 1024. The model's context encoder, video encoder, question encoder, content decoder, and answer decoder are all composed of stackable transformer layers, facilitating adaptability to diverse datasets. When employing the MSVD-QA dataset, the numbers of layer for each module are set at 8, 1, 1, 7 and 4, respectively, which are searched from 1 to 8. For MSRVTT-QA, the numbers are 1, 1, 2, 2, and 1. Concerning the loss weight α, they are designated at 1e-5 for MSVD-QA and 1e-6 for MSRVTT-QA, which are searched from 1e-6 to 1 increasing by multiples of 10 each time. Throughout the training process, the model underwent 30 epochs of iterative training with a batch size of 128 and a learning rate of 1e-4.

Comparison with State-of-the-arts

Table 1 presents the evaluation results of our approach and the performance of state-of-the-art models in the VideoQA domain, where accuracy is utilized as metric to evaluate the performance of the models. To ensure a comprehensive comparison, we also showcase the visual and textual feature extractors employed by each method, as well as whether additional datasets were used for pretraining.

Our approach V-CAT leverages conventional computer vision techniques, namely ResNet(He et al. 2016) and ResNeXt(Hara, Kataoka, and Satoh 2018), to extract video features. And we employ the classical BERT(Devlin et al. 2018) model to extract textual features. Similarly, HCRN(Le et al. 2020), B2A(Park, Lee, and Sohn 2021), HAIR(Liu et al. 2021), MHN(Peng et al. 2022), HQGA(Xiao et al. 2022) and EIGV(Li et al. 2022a) also utilize traditional image-based models such as Faster R-CNN(Anderson et al. 2018). However, their performance falls significantly short of our method. Our accuracy surpasses that of the highest-performing EIGV model by 2.6 on MSVD-QA and 4 on MSRVTT-QA, respectively.

Furthermore, even when compared to CLIP-QA(Radford et al. 2021b), which incorporates the powerful multi-modal CLIP (Radford et al. 2021a) model, and PMT(Peng

Method	Video Ex.	Text Ex.	PT	MSVD↑	MSRVTT↑
HCRN(Le et al. 2020)	ResNet, ResNeXt	Glove	-	36.1	35.6
B2A(Park, Lee, and Sohn 2021)	ResNet, ResNeXt	Glove	-	37.2	36.9
HAIR(Liu et al. 2021)	ResNet, Faster R-CNN	Glove	-	37.5	36.9
CLIP-QA(Radford et al. 2021b)	CLIP	Bert	-	38.5	39
MHN(Peng et al. 2022)	ResNet, ResNeXt	Glove	-	40.4	38.6
HQGA(Xiao et al. 2022)	ResNet, ResNeXt, Faster R-CNN	Bert	-	41.2	38.6
EIGV(Li et al. 2022a)	ResNet, ResNeXt	Bert	-	42.6	39.3
CoVGT(Xiao et al. 2023)	ResNet, Faster R-CNN	RoBERTa	-	-	40.0
PMT (Peng et al. 2023)	X3D-M	Glove	-	41.8	40.3
HD-VILA(Xue et al. 2022)	ResNet, TimeSformer	Bert	100M	41.8	40.3
MERLOT(Zellers et al. 2021)	ViT	RoBERTa	180M	-	43.1
CoMVT (Seo, Nagrani, and Schmid 2021)	S3D	Bert	100M	42.6	39.5
V-CAT(ours)	ResNet, ResNeXt	Bert	-	45.2	43.3

Table 1: Comparison with state-of-the-art methods on VideoQA datasets. Video Ex. and Text Ex. denote the video and text feature extractor, respectively. PT denotes the a mount of dataset used for pre-training.

Method	MSVD↑	MSRVTT↑
V-CAT w/o context	35.7	34.2
V-CAT w/ question	31.3	31.5
V-CAT w/o SA	45.1	43.0
V-CAT w/o CA	37.2	36.6
V-CAT	45.2	43.3

Table 2: Ablation study of the alignment module.

et al. 2023), which is based on a video-based model X3D(Feichtenhofer 2020), and CoVGT(Xiao et al. 2023), which utilizes the robust RoBERTa(Liu et al. 2019) model for textual feature extracting, our approach still exhibits a notable advantage. On the MSVD-QA and MSRVTT-QA datasets, our accuracy exceeds that of the leading PMT model by 3.4 and 3, respectively. Evidently, although our pretrained model is straightforward, the feature processing pipeline remains stable, and the extracted features are effectively utilized.

Meanwhile, HD-VILA(Xue et al. 2022), MERLOT(Zellers et al. 2021) and CoMVT(Seo, Nagrani, and Schmid 2021) aim to enhance the model's modality alignment and generalization capabilities through large-scale dataset pretraining, with the goal of improving accuracy. Despite not employing any data pretraining, our model achieves accuracy surpassing that of the highest-performing CoMVT model by 2.6 on MSVD-QA and MERLOT model by 0.2 on MSRVTT-QA. This demonstrates that our balanced approach to video and context ensures a more stable and efficient alignment process.

Ablation Analysis

Method	MSVD↑	MSRVTT↑
V-CAT w/o CL	45.1	43.0
V-CAT w/ L1	43.6	42.8
V-CAT w/ L2	44.1	42.2
V-CAT w/ KL	42.2	41.5
V-CAT	45.2	43.3

Table 3: Variants of our model specifically in loss function.

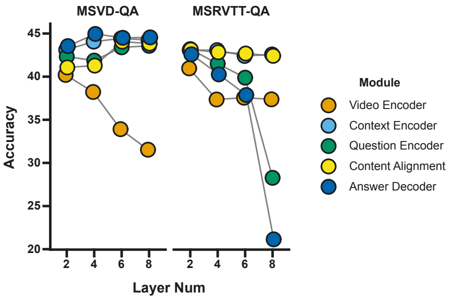


Figure 4: The figure showcases the impact of stacking layers in different modules of the model on its performance on the MSVD-QA and MSRVTT-QA datasets. The horizontal axis represents the number of layers in the encoder or decoder, while the vertical axis represents the model's accuracy. Each point of a specific color corresponds to a respective module.

Alignment We conducted an ablation analysis on the strategy of using context for alignment, which result is shown in Table 2. The 'w/o context' indicates a model that does not use any context, while 'w/ question' represents the model using the current question as context. It was observed that the video features without contextual alignment struggle to align effectively with the semantic and content of the text, resulting in a significant impact on the model. Meanwhile, using only the question as context yields unsatisfactory results. The primary reason is that the amount of information contained in a single question is much less than that in the video (Lin et al. 2022), leading to an extremely imbalanced of two semantic information during alignment, which destabilizes the model. Additionally, aligning with a single sentence with minimal information reduces the richness of video information, resulting in a loss of smoothness and potentially causing the video features to lose crucial information for reasoning. Furthermore, to further analyze the roles of semantic and content alignment in the model, we evaluated 'w/o SA' and 'w/o CA' separately. It was found that both cases led to a decrease in model accuracy, with a greater loss observed when content alignment was removed.

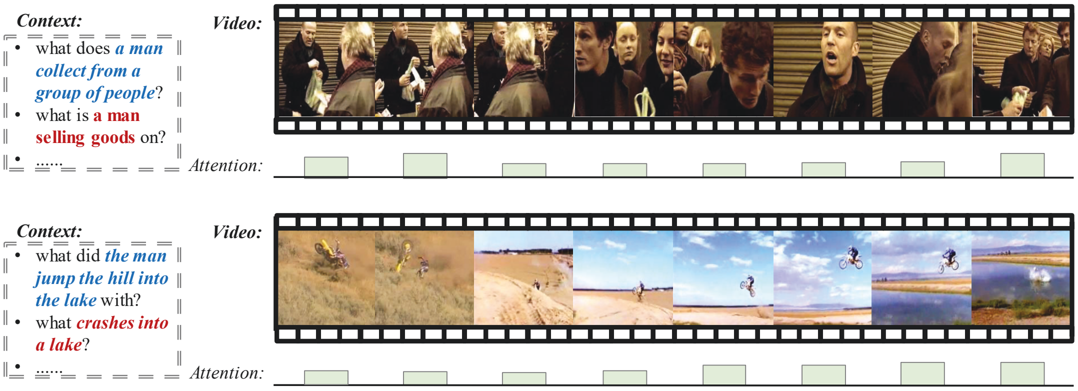

Context:
what does a man collect from a group of people?
what is a man selling goods on?
......
Video:
Attention:
Context:
what did the man jump the hill into the lake with?
what crashes into a lake?
......
Video:
Attention:

Figure 5: Two case of the attention of different video contents under the context. The context is constructed by collecting video-related questions. It is evident that, within the context alignment, the model is more adept at capturing the primary events in the video, while disregarding some peripheral and unrelated scenes.

Contrastive Learning Table 3 shows the analysis impact of contrastive learning loss. Firstly, we present the model's performance removing the loss constraint. The accuracy of the model decreases, which shows the evidence that the contrastive loss imposes constraints on video and contextual features, enabling the model to align video and text semantically more effectively, thereby enhancing the model's performance. Also, we replaced the contrastive loss function to obtain multiple variants of the model and analyze the effectiveness. It was observed that regardless of whether L1 loss, L2 loss or KL(Kullback-Leibler) divergence loss was used as the loss function, they all led to a decrease in model performance, even lower than without incorporating the loss constraint. Although these three types of loss functions aim to encourage similarity between the feature distributions of two global tokens, the contrastive loss used in addition to this is able to further push apart semantically dissimilar pairs. This makes the model more stable and reliable in semantic alignment.

Encoders and Decoders For each module, we employed transformer-based encoders and decoders. Such stackable modules enhance the model's scalability and can adapt to different datasets by altering the number of layers. To explore the impact of layer numbers on the model's performance on the MSVD-QA and MSRVTT-QA, we present the results for different parameters, as shown in the Figure 4. Overall, both datasets achieve satisfactory results with fewer layers, and an excessive number of stacked layers actually leads to a decline in model performance. Regarding the MSVD-QA dataset, apart from the layers in the Video Encoder, an appropriate increase in other layers contributes to model improvement. However, for the MSRVTT dataset, increasing the number of layers generally results in a decrease in model accuracy, particularly for the Question Encoder and Answer Decoder, where the accuracy drops below 30 when the number of layers reaches 8. It is evident that by adjusting the layer numbers of each module, we can adapt the model to different datasets.

Content Attention Visualization

We conducted a visual analysis of the attention mechanism during the content alignment process to explore the selection and filtering of actual video contents, which is shown in Figure 5. In the first example, the event in the video where the man sells a product to earn money is assigned greater attention, while certain shots of the crowd are given lower weight. In contrast, in the second example, the entire video emphasizes the scene where the man leaps on a motorcycle, soaring through the air before plunging into the water, garnering higher attention, while some non-essential shots receive lower attention. It is evident that, within the context alignment, the model is more adept at capturing the primary events in the video, while disregarding some peripheral and unrelated scenes. These primary events are precisely the subjects that tend to be inquired about. Furthermore, the model effectively assigns higher weight to these crucial events, which aids in subsequent answer decoding, while appropriately disregarding environmental scenarios.

Conclusion

We proposes V-CAT, which strengthens the information on the textual side by introducing context to align with the video. Specifically, the feature extractor and encoder embed visual and textual information into a shared feature space, and innovatively introduce contrastive learning to align global visual and contextual semantics, ensuring the consistency of multimodal features. Subsequently, more relevant video information is extracted through contextual information refinement to stabilize the subsequent answering process. Finally, the interaction between the current question features and video features yields the final answer. Our model structure is simple yet remarkably effective. Evaluation results on the MSVD-QA and MSRVTT-QA datasets demonstrate that our approach outperforms existing models. In future, we will consider replacing the simplistic encoder module during answer generation and introduce more refined interaction methods. Additionally, alternative approaches can be explored for modeling the context of videos.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62006034; in part by the Ministry of Education Humanities and Social Science Project under Grant 22YJC740110; in part by the Dalian High-level Talent Innovation Support Plan under Grant 2021RQ056; and in part by the Fundamental Research Funds for the Central Universities under Grant DUT23YG136.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Dang, L. H.; Le, T. M.; Le, V.; and Tran, T. 2021. Hierarchical object-oriented spatio-temporal reasoning for video question answering. *arXiv preprint arXiv:2106.13432*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feichtenhofer, C. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 203–213.

Gao, D.; Zhou, L.; Ji, L.; Zhu, L.; Yang, Y.; and Shou, M. Z. 2023. MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14773–14783.

Guo, Z.; Zhao, J.; Jiao, L.; Liu, X.; and Li, L. 2021. Multi-scale progressive attention network for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 973–978.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6546–6555.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, Z.; Zhong, W.; Ji, L.; Gao, D.; Yan, K.; Chan, W.-K.; Ngo, C.-W.; Shou, Z.; and Duan, N. 2022. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. *arXiv preprint arXiv:2209.10918*.

Jiang, J.; Chen, Z.; Lin, H.; Zhao, X.; and Gao, Y. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11101–11108.

Jin, W.; Zhao, Z.; Cao, X.; Zhu, J.; He, X.; and Zhuang, Y. 2021. Adaptive spatio-temporal graph enhanced vision-language representation for video qa. *IEEE Transactions on Image Processing*, 30: 5477–5489.

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9972–9981.

Li, M.; Shi, X.; Leng, H.; Zhou, W.; Zheng, H.-T.; and Zhang, K. 2023. Learning Semantic Alignment with Global Modality Reconstruction for Video-Language Pre-training towards Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1377–1385.

Li, Y.; Wang, X.; Xiao, J.; and Chua, T.-S. 2022a. Equivariant and invariant grounding for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4714–4722.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2022b. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2928–2937.

Lin, C.; Wu, A.; Liang, J.; Zhang, J.; Ge, W.; Zheng, W.-S.; and Shen, C. 2022. Text-adaptive multiple visual prototype matching for video-text retrieval. *Advances in Neural Information Processing Systems*, 35: 38655–38666.

Liu, F.; Liu, J.; Wang, W.; and Lu, H. 2021. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1698–1707.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Park, J.; Lee, J.; and Sohn, K. 2021. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15526–15535.

Peng, M.; Wang, C.; Gao, Y.; Shi, Y.; and Zhou, X.-D. 2022. Multilevel hierarchical network with multiscale sampling for video question answering. *arXiv preprint arXiv:2205.04061*.

Peng, M.; Wang, C.; Shi, Y.; and Zhou, X.-D. 2023. Efficient End-to-End Video Question Answering with Pyramidal Multimodal Transformer. *arXiv preprint arXiv:2302.02136*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Seo, P. H.; Nagrani, A.; and Schmid, C. 2021. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16877–16887.

Sun, X.; Wang, X.; Gao, J.; Liu, Q.; and Zhou, X. 2022. You need to read again: Multi-granularity perception network for moment retrieval in videos. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1022–1032.

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiao, J.; Yao, A.; Liu, Z.; Li, Y.; Ji, W.; and Chua, T.-S. 2022. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2804–2812.

Xiao, J.; Zhou, P.; Yao, A.; Li, Y.; Hong, R.; Yan, S.; and Chua, T.-S. 2023. Contrastive Video Question Answering via Video Graph Transformer. *arXiv preprint arXiv:2302.13668*.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, 1645–1653.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5036–5045.

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2023. Self-Chained Image-Language Model for Video Localization and Question Answering. *arXiv preprint arXiv:2305.06988*.

Zang, C.; Wang, H.; Pei, M.; and Liang, W. 2023. Discovering the Real Association: Multimodal Causal Reasoning in Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19027–19036.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34: 23634–23651.

Zhong, Y.; Xiao, J.; Ji, W.; Li, Y.; Deng, W.; and Chua, T.-S. 2022. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*.