Question-Instructed Visual Descriptions for Zero-Shot Video Question Answering

David Romero and **Thamar Solorio**
MBZUAI
{david.mogrovejo,thamar.solorio}@mbzuai.ac.ae

Abstract

We present Q-ViD, a simple approach for video question answering (video QA), that unlike prior methods, which are based on complex architectures, computationally expensive pipelines or use closed models like GPTs, Q-ViD relies on a single instruction-aware open vision-language model (InstructBLIP) to tackle video QA using frame descriptions. Specifically, we create captioning instruction prompts that rely on the target questions about the videos and leverage InstructBLIP to obtain video frame captions that are useful to the task at hand. Subsequently, we form descriptions of the whole video using the question-dependent frame captions, and feed that information, along with a question-answering prompt, to a large language model (LLM). The LLM is our reasoning module, and performs the final step of multiple-choice QA. Our simple Q-ViD framework achieves competitive or even higher performances than current state of the art models on a diverse range of video QA benchmarks, including NExT-QA, STAR, How2QA, TVQA and IntentQA. Our code is publicly available at: `https://github.com/Daromog/Q-ViD`

1 Introduction

Recently, vision-language models have shown remarkable performances in image question-answering tasks (Goyal et al., 2017; Marino et al., 2019; Schwenk et al., 2022), with models such as Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), InstructBlip (Dai et al., 2023) and mPLUG-Owl (Ye et al., 2023) showing strong reasoning capabilities in the vision-language space. Image captioning (Vinyals et al., 2015; Ghandi et al., 2023) is one of the tasks in which these models truly excel, as they can generate detailed linguistic descriptions from images. Different works have leveraged this capability in many ways for zero-shot image-question answering, such as giving linguistic context to images (Hu et al., 2022; Ghosal et al., 2023), addressing underspecification problems in questions (Prasad et al., 2023), coordination of multiple image captions to complement information (Chen et al., 2023b), or by combining captions with other type of linguistic information from the image (Berrios et al., 2023). In this manner, the reasoning capabilities of LLMs can be directly used to reason about the linguistic image descriptions and generate an answer for the given visual question.

This approach has been successful for images, but in the case of video-question answering tasks (Lei et al., 2018; Li et al., 2020; Xiao et al., 2021; Wu et al., 2021; Li et al., 2023a) this is more challenging. Video possesses multiple image frames that have relationships between each other and involve the recognition of objects, actions, as well as the inference about semantic, temporal, causal reasoning and much more (Zhong et al., 2022). Thus, some works (Chen et al., 2023a; Wang et al., 2023) have focused on using ChatGPT to either ask visual questions to image-language models like BLIP-2 or to respond and retrieve useful information from large datasets with detailed linguistic information from the video. Similarly, Zhang et al. (2023a) have leveraged the reasoning capabilities of GPT-3.5 to create textual summaries from the video, and later perform video QA using only textual information. While others (Wang et al., 2022b; Zeng et al., 2022) combine linguistic information from multiple sources such as captions, visual tokenization or even subtitles of input speech. In summary, current methods for video QA rely on any combination of closed LLMs, expensive training regimes, and complex architectures with multiple modules (Yang et al., 2022; Ko et al., 2023; Yu et al., 2023; Momeni et al., 2023; Li et al., 2023c; Zhang et al., 2023a). In contrast, we introduce Q-ViD a simple **Q**uestion-Instructed **Vi**sual **D**escriptions for video QA approach that relies on an instruction-aware vision-language model, InstructBLIP (Dai

Findings of the Association for Computational Linguistics: ACL 2024, pages 9329–9339
August 11-16, 2024 ©2024 Association for Computational Linguistics

et al., 2023), to automatically generate rich specific captions from video frames. In this manner, we effectively turn the video QA task into a text QA task. More specifically, given an input video V we sample n number of frames, then, we generate question-specific instructions to prompt the multimodal instruction tuned model to generate captions for each frame. Afterwards, we form a video description by concatenating all the generated question-dependent captions from InstructBLIP, and use it along with the question, options and a question-answering instruction prompt as input to the LLM-based reasoning module that generates an answer to the multiple-choice question about the video. We demonstrate the effectiveness of Q-ViD on five challenging multiple choice video question answering tasks (NExT-QA, STAR, How2QA, TVQA, IntentQA), showing that this simple framework can achieve strong performances comparable with more complex pipelines. Our contributions are summarized as follows:

- We propose Q-ViD, a simple gradient-free approach for zero-shot video QA that relies on an open instruction-tuned multimodal model to extract question-specific descriptions of frames to transform the video QA task into a text QA one.

- Our approach achieves strong zero-shot performance that is competitive or even superior to more complex architectures such as SeViLa, Internvideo, and Flamingo. It even compares favorably with recent solutions that include GPT APIs, like LloVi and ViperGPT.

2 Related Work

2.1 Multimodal Pretraining for Video QA

The strong reasoning capabilities of LLMs (Chung et al., 2022; Touvron et al., 2023; Brown et al., 2020; Hoffmann et al., 2022) in natural language processing tasks has motivated to apply these models for visual understanding. Currently, LLMs have been successfully adapted to understand images (Li et al., 2023b; Ye et al., 2023; Chen et al., 2023c), but applying the same principles for video is more challenging. Approaches for VideoQA rely on image-language models, and adapt those to process video by using fixed amounts of video frames as input (Alayrac et al., 2022; Yu et al., 2023; Yang et al., 2022), or by selecting key-frames from the initial sequence (Yu et al., 2023; Li et al., 2023c).

Commonly, these works use frozen visual and language models and focus only on modality alignment. Models like Flamingo (Alayrac et al., 2022) uses a fixed amount of video frames as input and bridges modalities by training a perceiver resampler and gated attention layers in the Chinchilla LLM (Hoffmann et al., 2022). While others, like SeViLa (Yu et al., 2023) relies on BLIP-2 (Li et al., 2023b) for modality alignment, using an intermediate pretrained module called Q-former. SeViLa, first perform key-frame localization and then video QA with Flan-T5 LLMs (Chung et al., 2022). On the other hand, other works apart from using frozen vision models, adapt the LLM to visual inputs using adapter tokens (Zhang et al., 2023b) or intermediate trainable modules (Houlsby et al., 2019). Models like Flipped-VQA (Ko et al., 2023) focuses on adapting LLaMa (Touvron et al., 2023) to video QA by using adapter tokens along with different training objectives to leverage the temporal and causal reasoning abilities of LLMs. Similarly, FrozenBilm (Yang et al., 2022) exploit the strong zero-shot performance of BILM, a frozen bidirectional language model that is adapted to video QA by using lightweight trainable modules. Despite the success of all these models, they require complex architectures and training regimes, unlike these works we build a simple, gradient-free, approach for zero-shot video QA.

2.2 Image Captions for Video Understanding

One of the core strengths of image-language models (Alayrac et al., 2022; Li et al., 2023b; Dai et al., 2023) is the generation of image captions, thus due to the current strong zero-shot capabilities of LLMs, captions can be directly use to reason about visual content. This has been successfully leveraged in the image-language space for image question-answering with approaches such as Lens (Berrios et al., 2023), Img2LLM (Guo et al., 2023) and PromptCat (Hu et al., 2022) that gather image captions and other type of linguistic information to answer a visual question. While similar approaches have been taken for videos, the use of large models like GPTs is very common, with models such as ChatCaptioner (Chen et al., 2023a), ViperGPT (Surís et al., 2023) , ChatVideo (Wang et al., 2023), VidIL (Wang et al., 2022b), Socratic Models (Zeng et al., 2022), and LLoVi (Zhang et al., 2023a) have been applied for video-language tasks, common methods use GPTs to either interact with image-language models to get visual descriptions, or to

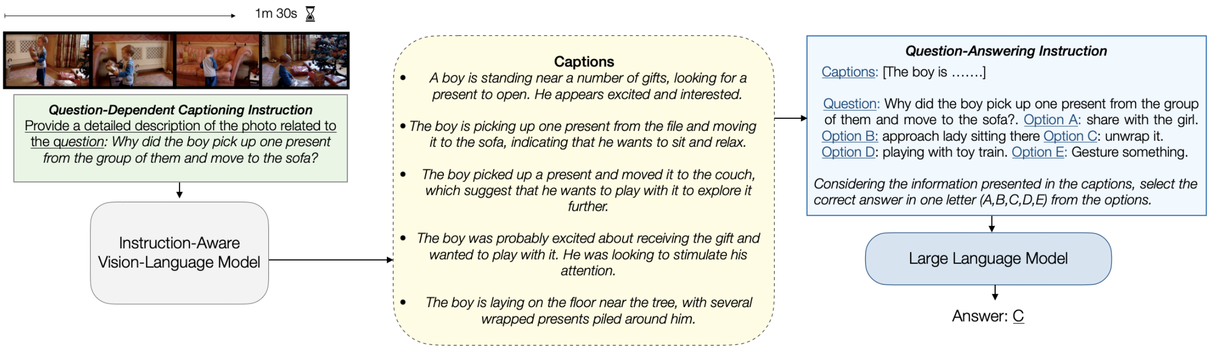

Figure 1: **Overview of Q-ViD**. We propose relying on a instructed-tuned multimodal model to generate question-dependent frame captions to perform video QA using text. This simple approach achieves competitive results with more complex architectures or GPT-based methods.

make summaries from captions and other type of information such as visual tokenization, subtitles of speech and more. Unlike these approaches, we do not use GPTs or multiple computationally expensive modules in any part of our pipeline to achieve strong zero-shot performance on video QA.

3 Method

Recently, vision-language models trained with instruction tuning (Dai et al., 2023; Zhu et al., 2023; Liu et al., 2023) have shown impressive capabilities to faithfully follow instructions and extract visual representations adapted to the task at hand. Thus, with Q-ViD (Figure 1),we propose to leverage these capabilities for multiple-choice video QA, and turn this task into textual QA using InstructBLIP (Dai et al., 2023). We use a question-dependent captioning prompt as the input instruction, to guide InstructBLIP to generate video frame descriptions that are more relevant for the given question. Afterwards, we reuse the LLM from InstructBLIP and use it as our reasoning module. This LLM (Flan-T5) takes a question-answering prompt as input, that consists of a video description formed by the concatenation of all the question-dependent frame captions, the question, options and a task instruction. Considering that Flan-T5 is also originally trained with instructions, we aim to leverage its reasoning capabilities to correctly answer the question given only the text we just described as input. Our simple approach does not rely on complex pipelines or closed GPT models, which makes it easy, cheaper and straight forward to use for zero-shot video QA. On the other hand, Q-ViD is flexible and model agnostic, which means we can use any multimodal models available. This section presents our approach in detail. First, we introduce some preliminary information on InstructBLIP, which serves as the foundation of our work, and then we provide a detailed overview for all components from our Q-ViD framework.

3.1 Preliminaries: InstructBLIP

We rely on InstructBLIP (Dai et al., 2023) as the foundational architecture of Q-ViD. InstructBLIP is a vision-language instruction tuning framework based on a Query Transformer (Q-former) and frozen vision and language models. Unlike BLIP-2 (Li et al., 2023b), which is based on an instruction-agnostic approach, InstructBLIP can obtain visual features depending on specific instructions of the task at hand using an instruction-aware Q-former, which in addition to query embeddings, uses instruction tokens to guide the Q-former in extracting specific image features. Subsequently, a LLM (Flan-T5) uses these features to generate visual descriptions depending on the input instructions. In our approach we adapt this model to video, we adopt it to obtain video frame captions that are dependant on the questions of the video QA task, thus, we aim to gather the most important information from each part of the video and use it as input for our reasoning module to answer the given question. Because of our Q-ViD framework is a zero-shot approach, we do not train any part of InstructBLIP, and keep all of its parts frozen.

3.2 Q-ViD: Generating Frame Descriptions for Video QA

We focus on automatically generating meaningful captions that can provide enough information about what is happening in the video to the LLM. We assume that if captions for the frames contain relevant information related to the question

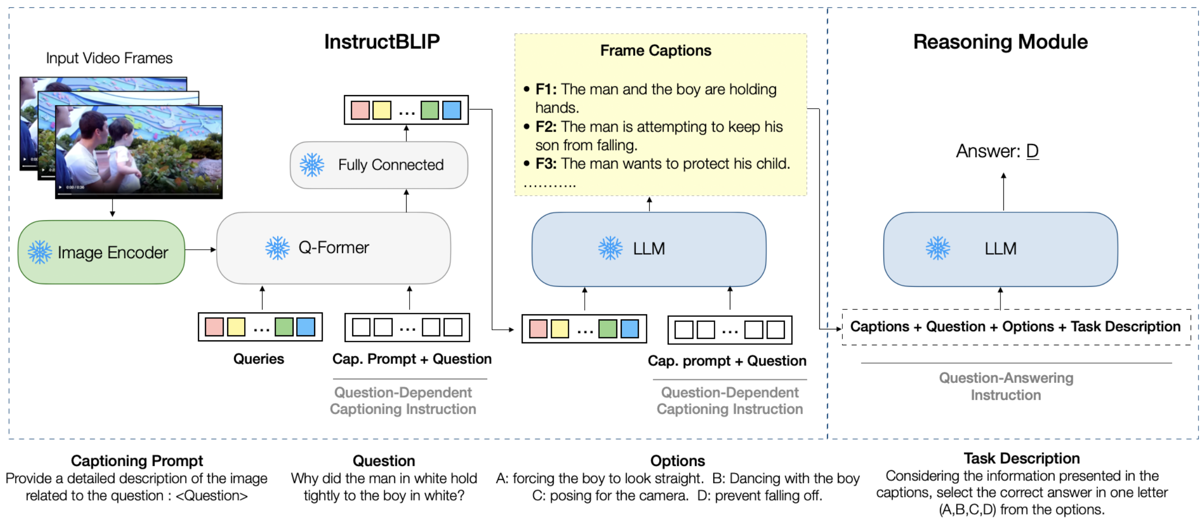

Figure 2: **Our pipeline for Zero-shot Video QA.** Q-ViD prompts InstructBLIP, to obtain video frame descriptions that are tailored to the question needing answer.

needing answer, then an LLM should be able to answer the question correctly without additional need for frame/video input. As shown in Figure 2, given an input video V, we use a uniform sampling strategy and extract a set of n video frames $\{f_1, f_2, ..., f_n\}$. We then use InstructBLIP, refered as I_b, to obtain instruction-aware visual captions c_i for each frame f_i, as follows $c_i = I_b(f_i, E)$, where E represents the question-dependent captioning instruction. Q-ViD generates E, by concatenating a captioning prompt, referred as B (e.g *"Provide a detailed description of the image related to the question:"*) and a question, referred as Q (e.g *"Why did the man in white held tightly to the boy in white?"*), represented as follows $E = concat(B, Q)$. Specifically, E is used as input to the Q-former and the LLM of InstructBLIP to obtain specific visual representations and frame descriptions respectively. Thus, we represent the input video V as a set of question-dependent frame captions $c = (c_1, c_2, ..., c_n)$, where each caption is conformed by a sequence of w_m words $c_i = (w_1, w_2, ..., w_m)$. In this way, we extract specific textual information from the frames of V, that is going to be useful for the question answering task. Next, we describe the reasoning module of Q-ViD and how these question-dependent captions are used to perform video QA.

3.3 Q-ViD: Reasoning Module

We reuse the frozen LLM (Flan-T5) from InstructBLIP and implement it as the reasoning module of Q-ViD. In order to perform video QA using language, we first concatenate the question-dependent frame captions $C = [c_1, c_2, ..., c_n]$ in the same order they appear in the video. Then, we create a question-answering instruction L as follows: $L = concat(C, Q, A, T)$. In other words, we concatenate in L the list of captions C, question Q, possible answers A and a task description T (e.g *"Considering the information presented in the captions, select the correct answer in one letter (A,B,C) from the options."*). Our goal is to leverage the LLM reasoning linguistic capabilities by providing a set of captions that were tailored to be relevant for the specific question Q. Our experiments in Section 4, show that this simple approach works surprisingly well, showing to be competitive, and even superior in some cases, in comparison with more complex pipelines. Next, we describe in more detail the prompts used for question-dependent captioning and video QA.

3.4 Q-ViD: Prompt Design

First, to get question-dependent captions for each frame, given the question Q we prompt InstructBLIP with a question-dependent captioning instruction: *"Provide a detailed description of the image related to the question: {Q}"*. This instruction is used along with queries as input to the frozen Q-Former and LLM modules of InstructBLIP to extract specific visual features and generate question-dependent descriptions. Afterwards, to perform QA with the reasoning module, given the list of captions C and the list of possible answers $A = [a_1, ..., a_m]$ with m being the number of options provided in each dataset, we prompt the language model as follows: *"Captions: {C} Question: {Q}. Option A: a_1. Option B: a_2. Option C: a_3. Considering the information presented in the captions, select the correct answer in one letter from the options (A,B,C)"*. In this prompt, in addition to the list of captions, the question and the list of possible answers, we added a small instruction at the end to specify in detail that a single letter is needed as output.

4 Experiments

In this section, we present our experiments for zero-shot video QA. First, we describe the datasets we used and the implementation details. Then, we evaluate our approach, compare Q-ViD with other state of the art models for video QA and provide a comprehensive analysis of the model's performance. Lastly, we conduct some ablation studies of Q-ViD regarding the instructions prompt design.

4.1 Datasets

To test our approach, we conduct experiments on the following multiple-choice video QA benchmarks. To make comparisons with prior work we use the validation set in NExT-QA, STAR, How2QA and TVQA, meanwhile in IntentQA we use the test set. More details are shown below:

- **NExT-QA** (Xiao et al., 2021): A benchmark focused on Temporal, Causal and Descriptive reasoning type of questions. Contains 5,440 videos and 48K multiple-choice questions in total. We perform our experiments using the validation set that is conformed by 570 videos and 5K multi-choice questions.
- **STAR** (Wu et al., 2021): A benchmark that evaluates situated reasoning in real-world videos, is focused on interaction, sequence, prediction and feasibility type of questions. It contains 22K situation video clips and 60K questions. We perform evaluations on the validation set with 7K multiple-choice questions.
- **HOW2QA** (Li et al., 2020): A dataset that consists on 44K question-answering pairs for 22 thousand 60-second clips selected from 9035 videos. We perform experiments on the validation set with 2.8K questions.
- **TVQA** (Lei et al., 2018): A large scale video QA dataset based on six popular TV shows. It has 152K multiple-choice questions and 21K video clips. For our zero-shot evaluations we use the validation set with 15K video-question pairs.
- **IntentQA** (Li et al., 2023a): A dataset focused on video intent reasoning. It contains 4K videos and 16K multiple-choice question-answer samples. In this case, we use the test set for our zero-shot evaluations which contains 2K video-question answering samples.

4.2 Implementation Details

For Q-ViD we adopt InstructBLIP-Flan-T5$_{\text{XXL}}$ with 12.1B parameters, as a default vision encoder it uses VIT-g/14 (Fang et al., 2023), and as language model FlanT5$_{\text{XXL}}$ (Chung et al., 2022). We extract 64 frames per video, as in preliminary experiments this number worked well. For frame captioning, we use a maximum number of 30 tokens per description and top-p sampling with $top_p = 0.7$ to get varied captions. Regarding our reasoning module, we reuse and adopt the corresponding Flan-T5 language model from InstructBLIP. In this case we do not use top-p sampling. Our experiments were conducted using 4 NVIDIA A100 (40GB) GPUs using the Language-Vision Intelligence library LAVIS (Li et al., 2022) and the released code from SeViLa (Yu et al., 2023) .

4.3 Overall Performance

Table 1 provides a detailed overview on the performance of Q-ViD on the validation set of NExT-QA, STAR, HOW2QA and TVQA. We compare our approach with current state of the art methods such as SeViLa (Yu et al., 2023), FrozenBILM (Yang et al., 2022) and VideoChat2 (Li et al., 2024), as well as, with GPT-based models like ViperGPT (Surís et al., 2023) and LloVi (Zhang et al., 2023a). The results obtained from our experiments demonstrate the surprisingly competitive nature of Q-ViD, outperforming or being competitive with previous methods with more complex architectural pipelines such as SeViLa, VideoChat2 and LLoVi. For fair comparisons, we gray out methods that use GPTs.

Specifically, on **NExT-QA**, Q-ViD outperforms SeViLa by **2.7%** of average accuracy, and achieves almost the same state of the art results of Llovi, a framework based of GPT-3.5. Notably, Q-ViD is the best-performing model on causal questions, temporal questions, and overall average performance among methods that are not based on GPTs, showing the ability of this approach to perform action reasoning, which is the target of NExT-QA. With **STAR**, Q-ViD achieves the second

Models	NExT-QA				STAR					How2QA	TVQA
	Tem.	Cau.	Des.	Avg.	Int.	Seq.	Pre.	Fea.	Avg.		
GPT-Based Models											
ViperGPT (Surís et al., 2023)	-	-	-	60.0	-	-	-	-	-	-	-
LLoVi (Zhang et al., 2023a)	61.0	69.5	75.6	67.7	-	-	-	-	-		
Flamingo-9B (Alayrac et al., 2022)	-	-	-	-	-	-	-	-	41.8	-	-
Flamingo-80B (Alayrac et al., 2022)	-	-	-	-	-	-	-	-	39.7	-	-
FrozenBILM (Yang et al., 2022)	-	-	-	-	-	-	-	-	-	41.9	29.7
VFC (Momeni et al., 2023)	51.6	45.4	64.1	51.6	-	-	-	-	-	-	-
InternVideo (Wang et al., 2022a)	48.0	43.4	65.1	49.1	43.8	43.2	42.3	37.4	41.6	62.2	35.9
BLIP-2$^{\text{voting}}$ (Yu et al., 2023)	59.1	61.3	<u>74.9</u>	62.7	41.8	39.7	40.2	39.5	40.3	69.8	35.7
BLIP-2$^{\text{concat}}$ (Yu et al., 2023)	59.7	60.8	73.8	62.4	45.4	41.8	41.8	40.0	42.2	70.8	36.6
SeViLa (Yu et al., 2023)	<u>61.3</u>	61.5	**75.6**	<u>63.6</u>	<u>48.3</u>	45.0	<u>44.4</u>	40.8	44.6	**72.3**	38.2
VideoChat2 (Li et al., 2024)	57.4	<u>61.9</u>	69.9	61.7	**58.4**	**60.9**	**55.3**	**53.1**	**59.0**	-	<u>40.6</u>
Q-ViD (Ours)	**61.6**	**67.6**	72.2	**66.3**	48.2	<u>47.2</u>	43.9	<u>43.4</u>	<u>45.7</u>	<u>71.4</u>	**41.0**

Table 1: **Zero-shot results on video question answering.** For fair comparison we gray out methods that rely on closed GPTs. We **bold** the best results, and <u>underline</u> the second-best results. Q-ViD shows to be competitive and even outperform some more complex frameworks for zero-shot video QA.

best average accuracy behind VideoChat2, outperforming all other methods like SeViLa by **1.1%**, BLIP-2$^{\text{concat}}$ by **3.5%**, InternVideo by **4.1%** and Flamingo-80B by **6%**. Also note that Q-ViD achieves the second best performances on sequence and feasibility type of question of STAR. Lastly, on **How2QA** we achieve the second best performance behind SeViLa, and achieves the best overall performance for **TVQA** with an improvement of **0.4%** to the previous best-performing method VideoChat2.

On the other hand, in Table 2 we evaluate our approach on **IntentQA**, we use the test set of this benchmark in order to compare with prior works. We take the same comparison made from (Zhang et al., 2023a), and divide the table in two categories, Supervised and Zero-shot approaches. Q-ViD continues showing strong results, greatly outperforming all supervised methods and the SeViLa zero-shot performance by **2.7%**. Interestingly, Q-ViD almost achieves the best overall performance from the GPT-based method Llovi. These results demonstrate that our approach can be used among different video QA tasks and be able to achieve strong zero-shot performances.

Models	Acc.(%)
Supervised	
HQGA (Surís et al., 2023)	47.7
VGT (Alayrac et al., 2022)	51.3
BlindGPT (Alayrac et al., 2022)	51.6
CaVIR (Alayrac et al., 2022)	57.6
Zero-shot	
SeViLA (Yu et al., 2023)	<u>60.9</u>
LLoVi (Zhang et al., 2023a)	64.0
Q-ViD (Ours)	**63.6**

Table 2: **Performance on IntentQA.** Q-ViD shows to outperform supervised approaches, strong zero-shot baselines like SeViLa and obtain almost the same performance from the GPT-based model LLoVi.

5 Ablation Studies

In this section, we perform some ablation studies related to the instruction prompt selection for Q-ViD. For these experiments, we chose NExT-QA and STAR as our benchmarks, and report results on the validation sets on each dataset. Specifically, we test two model variations, using InstructBLIP-FlanT5$_{\text{XL}}$ (Q-ViD$_{\text{XL}}$) and the one used to report our main results, InstructBLIP-FlanT5$_{\text{XXL}}$ (Q-ViD$_{\text{XXL}}$), we test different prompts to analyze and compare the use of question-dependent and general descriptive captions. Additionally, we also make some ablation experiments for the question-answering instruction prompt that is used by the reasoning module to perform multi-choice QA. We discuss our findings in detail below.

5.1 Prompt Analysis

We focus on analyzing the impact on performance of the Captioning and QA instruction templates in Q-ViD. First, for the captioning instruction template (Figure 3), we compare two type of variants: (1) General prompts and (2) Question-dependent prompts. With general prompts we focus on obtaining general visual descriptions, and with question-dependent prompts on visual information related to the question of the task at hand. In order to test the impact of these captioning prompts, in both cases, we use a Base QA instruction template used as

Captioning Instruction Templates

General Prompts

1. Provide a description of what is presented in the photo.
2. Could you use a few words to describe what you perceive in the photo?

Question-dependent Prompts

1. Given the image, generate a description related with the Question: Q
2. Provide a detailed description of the image related to the Question: Q

Base - Question-Answering Instruction Template

- Captions: C. Question: Q. Option A: Option-1. Option B: Option-2. Option C: Option-3. Considering the information presented in the captions, select the correct answer in one letter from the options (A,B,C).

Figure 3: **Variation of captioning templates.** We focus on comparing general and question-dependent captioning prompts (Top). For both cases we use the same Base QA instruction template (Bottom).

Captioning Instruction Templates

General Prompt

- Could you use a few words to describe what you perceive in the photo?

Question-dependent Prompt

- Provide a detailed description of the image related to the Question: Q

Question-Answering Instruction Templates

1. Given the captions, answer the following question, select the correct answer in one letter from the options (A,B,C). Captions: C. Question: Q. Options. A: Option-1. B: Option-2. C: Option-3.
2. Using the following video description, answer the question by selecting the correct answer with a single letter (A,B,C). Video Description: C. Question: Q. Options. A: Option-1. B: Option-2. C: Option-3.

Figure 4: **Variation of QA prompt templates.** We focus on exploring two more complex and detailed variations for the QA instruction prompt (Bottom). We use the best captioning templates (Top) for Q-ViD$_{\text{XL}}$ (General) and Q-ViD$_{\text{XXL}}$ (Dependent).

input by the reasoning module (LLM) to perform multiple-choice QA. To leverage as much as possible the instruction-based capabilities of InstructBLIP, we create these prompts based on similar templates used by this model in its training setup.

Table 3 compares the performance of Q-ViD$_{\text{XL}}$ and Q-ViD$_{\text{XXL}}$ using the general, and question-dependent captioning prompts. It can be seen that performance varies between both models. First, Q-ViD$_{\text{XL}}$ achieves better performances with general captioning prompts, when comparing the best variants of this model, using the (2) General and (1) Dependent prompts, the former further increases the average accuracy by +1.4% on NExT-QA and +3.1% on STAR. On the other hand, the same behaviour is not shown using a bigger model, Q-ViD$_{\text{XXL}}$ achieves significant improvements in average performance by using question-dependent prompts, when comparing its best variants using the (2) General and (2) Dependent prompts, the latter obtains improvements of +3.5% on NExT-QA and +4.2% on STAR. Unsurprisingly, Q-ViD$_{\text{XXL}}$ provides significant performance boosts when compared to its smaller version Q-ViD$_{\text{XL}}$ achieving better performances on all type of questions in both datasets, showing a better capability to follow instructions, however, *this also demonstrates that using question-dependent prompts to obtain specific information for the task at hand, performs better for zero-shot Video QA than using captioning prompts that obtains general visual descriptions.*

Next, in Table 4 we investigate the impact on performance of the QA Instruction template. We propose two variations that are shown at the bottom of Figure 4 , in addition to the Base QA template (Figure 3). With these new variants we aim to test giving more details to our reasoning module based on Flan-T5, because of this LLM is also a model trained with instructions, we explore if using more complex and detailed QA prompts we can achieve better performances. For this comparison we take the best variants (Table 3) of Q-ViD$_{\text{XL}}$ and Q-ViD$_{\text{XXL}}$ using the (2) General and (2) Dependent captioning prompts respectively for each model, and explore their performances with different QA instruction templates. As shown in Table 4, using more complex variants of the initial Base QA Instruction prompt does not have a big impact on performance in any of the models, it even slightly affects the performance in some cases, showing that the simplest base prompt was enough for the LLM to understand the task. With this ablation study we can highlight the fact that the input instruction used to obtain dedicated frame descriptions is far more important than elaborated question-answering instruction prompts for zero-shot video QA.

6 Conclusion

In this paper, we introduce Q-ViD, a simple, gradient-free approach for zero-shot video QA. Q-ViD turns video QA into textual QA using frame captions. To do so, Q-VID relies on an instruction-aware visual language model and uses question-dependent captioning instructions to obtain specific frame descriptions useful for the task at hand.

Method	NExT-QA				STAR				
	Tem.	Cau.	Des.	Avg.	Int.	Seq.	Pre.	Fea.	Avg.
Q-ViD$_{XL}$									
(1) General	57.3	**60.3**	**62.0**	**60.5**	47.0	**45.2**	42.7	42.2	44.3
(2) General	**57.8**	60.1	60.8	60.1	**47.4**	44.8	**44.7**	**42.8**	**44.9**
(1) Dependent	55.9	59.8	57.5	59.1	45.0	41.7	40.5	40.2	41.8
(2) Dependent	56.6	58.8	61.1	59.0	45.8	40.6	40.2	39.5	41.5
Q-ViD$_{XXL}$									
(1) General	57.5	64.6	67.4	62.7	44.7	39.5	42.6	36.3	40.8
(2) General	57.1	64.8	68.0	62.8	44.6	39.5	43.1	38.7	41.5
(1) Dependent	**62.0**	66.5	71.2	65.8	47.8	44.2	42.1	41.8	44.0
(2) Dependent	61.6	**67.6**	**72.2**	**66.3**	**48.2**	**47.2**	**43.9**	**43.4**	**45.7**

Table 3: **Comparing the impact on performance using different Captioning Instruction templates.** We test two variants, General prompts and Question-Dependent prompts. All experiments use the Base QA instruction template.

Model	Templates		NExT-QA				STAR				
	Captioning	QA	Tem.	Cau.	Des.	Avg.	Int.	Seq.	Pre.	Fea.	Avg.
Q-ViD$_{XL}$	(2)General	Base	**57.8**	60.1	**60.8**	60.1	47.4	44.8	**44.7**	**42.8**	**44.9**
		(1)QA	56.4	60.6	58.4	**60.2**	**47.7**	**44.9**	43.5	41.0	44.3
		(2)QA	56.8	**60.9**	57.5	60.1	47.0	44.1	43.1	40.6	43.7
Q-ViD$_{XXL}$	(2)Dependent	Base	61.6	**67.6**	72.2	**66.3**	48.2	**47.2**	**43.9**	43.4	45.7
		(1)QA	**61.7**	65.8	73.7	65.5	48.9	46.8	43.5	**43.8**	**45.8**
		(2)QA	61.5	65.6	**73.9**	65.5	**49.1**	45.9	42.9	42.6	45.1

Table 4: **Performance using different variants for the QA Instruction template.** Base: Refer to the base QA instruction template. For the captioning prompts all models use their best variants, Q-ViD$_{XL}$ with (2)General and Q-ViD$_{XXL}$ with (2)Dependent. These results suggest that there is no improvements using more complex QA instruction prompts for the reasoning module.

This information is later used by a reasoning module with a question-answering instruction prompt to perform multiple-choice video QA. Our simple approach achieves competitive or even higher performances than more complex architectures and methods that rely on closed models like the GPT family. In our ablation studies we show that using dedicated instructions to get question-dependent captions works better than common prompts to get general descriptions from frames to perform video QA using captions.

Limitations

Even though, Q-ViD has shown to achieve strong performances for zero-shot video question answering, our approach suffers from some limitations. While the adopted instruction-aware multimodal model, InstructBlip, shows to successfully follow instructions from the question and extract meaningful information that can help the reasoning module to come up with the right answer, we have seen that in some cases the model tends to show hallucinations in the captions, or generate direct short one-word answers instead of a detailed and question-specific description of the image. On the other hand, even though experiments with really long videos are not within the scope of this paper, our approach would no be recommended in those cases, due to the high memory usage that comes with saving detailed frame captions to create an entire video description, which would also affect the LLM-based reasoning module because of the limited amount of tokens allowed as input or due to memory constrains to process the entire video description.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the lens of natural language. *Preprint*, arXiv:2306.16410.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. 2023a. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*.

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2023b. Language models are visual reasoning coordinators. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023c. Pali: A jointly-scaled multilingual language-image model. *Preprint*, arXiv:2209.06794.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39.

Deepanway Ghosal, Navonil Majumder, Roy Lee, Rada Mihalcea, and Soujanya Poria. 2023. Language guided visual question answering: Elevate your multimodal language model using knowledge-enriched prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12096–12102, Singapore. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10867–10877.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Yushi* Hu, Hang* Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *EMNLP*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. In *EMNLP*.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. 2022. Lavis: A library for language-vision intelligence. *Preprint*, arXiv:2209.09019.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023a. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11963–11974.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multimodal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pretraining. In *EMNLP*.

Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023c. Discovering spatio-temporal rationales for video question answering. *Preprint*, arXiv:2307.12058.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15579–15591.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2023. Rephrase, augment, reason: Visual grounding of questions for vision-language models. *Preprint*, arXiv:2310.05861.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *Preprint*, arXiv:2303.08128.

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

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *Preprint*, arXiv:2304.14407.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022a. Internvideo: General video foundation models via generative and discriminative learning. *Preprint*, arXiv:2212.03191.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal, and Heng Ji. 2022b. Language models with image descriptors are strong few-shot video-language learners. *Preprint*, arXiv:2205.10747.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 2021. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

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

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *Preprint*, arXiv:2204.00598.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023a. A simple llm framework for long-range video question-answering. *Preprint*, arXiv:2312.17235.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *Preprint*, arXiv:2303.16199.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Licences

We use standard licenses from the community for the datasets, codes, and models that we used in this paper:

- **NExT-QA** (Xiao et al., 2021): MIT
- **STAR** (Wu et al., 2021): Apache
- **How2QA** (Li et al., 2020): MIT
- **TVQA** (Lei et al., 2018): MIT
- **IntentQA** (Li et al., 2023a): N/A
- **SeViLa** (Yu et al., 2023): BSD 3 - Clause
- **LAVIS** (Li et al., 2022): BSD 3-Clause
- **Pytorch** (Paszke et al., 2019): BSD Style
- **Q-ViD** (Ours): BSD 3-Clause

B Use of Artifacts

In this work we adopt a open multimodal model, InstructBLIP (Dai et al., 2023), its application in our approach is consistent with its original intended use. For Q-ViD we release our code and we hope it will be useful for future works.