AssistSR: Affordance-centric Question-driven Video Segment Retrieval

Stan Weixian Lei, Yuxuan Wang, Dongxing Mao, Difei Gao, Mike Zheng Shou
Show Lab, National University of Singapore
weixian lei@u.nus.edu

Abstract

It is still a pipe dream that AI assistants on phone and AR glasses can assist our daily life in addressing our questions like “how to adjust the date for this watch?” and “how to set its heating duration? (while pointing at an oven)”. The queries used in conventional tasks (i.e. Video Question Answering, Video Retrieval, Moment Localization) are often factoid and based on pure text. In contrast, we present a new task called Affordance-centric Question-driven Video Segment Retrieval (AQVSR). Each of our questions is an image-box-text query that focuses on affordance of items in our daily life and expects relevant answer segments to be retrieved from a corpus of instructional video-transcript segments. To support the study of this AQVSR task, we construct a new dataset called AssistSR. We design novel guidelines to create high-quality samples. This dataset contains 1.4k multimodal questions on 1k video segments from instructional videos on diverse daily-used items. To address AQVSR, we develop a straightforward yet effective model called Dual Multimodal Encoders (DME) that significantly outperforms several baseline methods while still having large room for improvement in the future. Moreover, we present detailed ablation analyses. Our codes and data are available at https://github.com/StanLei52/AQVSR.

1. Introduction

As shown in Fig. 1a, a user is looking at his watch and wondering “how to adjust the date for this watch”. It would be great if our phone or glasses can be powered by an intelligent agent, i.e. AI assistant, which sees exactly what the user sees and can find out the answer in the instructional video of this watch, either provided by the watch’s seller or posted on YouTube. However, such a video is usually quite long and thus time-consuming for users to watch, let alone there are a lot of online tutorial videos for various items. To save a user’s time, we aim to retrieve only the relevant short segments in a long video for that specific watch.

This problem is related with several existing tasks including Video Question Answering (VideoQA), Video Retrieval and Moment Localization. Yet, to unlock the aforementioned application in Fig. 1a, we need a new task and a new benchmark, which have two major differences compared to the existing tasks when designing questions:

• Factoid vs. Affordance-centric. Existing tasks and datasets mainly focus on factoid questions or queries. For example in Fig. 2, Question 1 in ActivityNetQA [44], Query 2 in DiDeMo dataset [3], and Query 3 in MSR-VTT [43] are all about simple facts, e.g. basic attributes, relationships between common objects. In contrast, we expect the AI assistant to go beyond simple facts and tackle affordance-centric questions. As shown in Fig. 1b, instead of asking “what is the man holding”, we focus on questions like “how to clean up <this part> of the vacuum”. Such questions regarding the affordance [8] of objects or devices are often asked by people in daily life.

• Pure text vs. Multimodal. Questions or queries in ex-
Affordance-centric Question-driven Video Segment Retrieval

Figure 2. Illustration for AQVSR and compare AQVSR with other video-related tasks. AQVSR focuses on affordance-centric questions with multimodalities: an image or referring regions as visual part and a natural language query as textual part. The task is set-based: given a question, the system is to retrieve from the video corpus the relevant segments.

Existing datasets are mostly pure text. However, psychology literature shows that pointing to the interesting objects is one of the initial manners that a baby conveys intention [28, 29, 32]. It is unnatural for human to ask the AI assistant with phrases like Q2 in Fig. 1b or “on the top-right of the vacuum”. It is more straightforward to directly point to that part and ask: “How to clean up <this part> of the vacuum?”. This is a multimodal question consisting of a textual question, the current image seen by the AI assistant, and the visual region pointed out by the user’s finger.

**Task.** Given a multimodal, affordance-centric question, we propose a new task called Affordance-centric Question-driven Video Segment Retrieval (AQVSR), which expects the model to retrieve the relevant video segments that can address the user’s question. As shown in Fig. 2, (1) unlike video-text retrieval which focuses on returning a whole video, AQVSR takes a step further to locate the relevant segment within a long instructional video. (2) Unlike moment localization whose query directly describes the segment’s content, we find it is often difficult for different annotators to agree on where the exact start and end times are for affordance-centric question. To provide a fair comparisons, AQVSR first asks annotators to chunk a long video into short segments and then models can only focus on whether retrieve a segment or not. Details in Sec. 3.

**Dataset.** To support studying AQVSR, we create a new dataset called AssistSR. Throughout multiple iterations, we form a novel annotation pipeline and guideline, which can effectively address annotation challenges including (1) source high-quality videos that support asking affordance-centric questions; (2) annotators tend to ask factoid and simple questions; (3) how to alleviate ambiguity for answer segment annotation. In total, we have collected 1,431 high-quality questions on 1,065 segments from 130 videos.

**Model.** To develop model for AQVSR, we propose Dual Multimodal Encoders (DME), a straightforward yet effective approach that has two transformer-based multi-modal encoders for the user question and a video segment respectively. We further show DME outperforms previous methods and each modality in question and video matters.

In summary, our contributions are three-fold:

- We introduce a new AQVSR task to retrieve video segments whose contents can address a given question.
- We construct AssistSR dataset for AQVSR, which is annotated with high-quality affordance-centric questions and covers a wide range of common scenarios.
- We benchmark a number of video-language methods on AssistSR, and analyze performance and challenges under the task of AQVSR.

2. Related Work

**Video Question Answering.** This task requires the intelligent system to automatically answer a natural language question according to the content of a given video. In recent years, multiple VideoQA datasets and tasks have been proposed to facilitate research towards this goal, several video-based QA datasets have also been proposed, e.g. TGIF-QA [14], MovieFIB [27], VideoQA [48], LSMDC [35], TRECVID [33], and MarioQA [31]. Additionally, Video Question Answering [16, 22, 40], with naturally occurring subtitles are particularly interesting, as it combines both visual and textual information for question answering. Different from VideoQA, where a system is required to generate an answer in a video-instance based setting, the task of AQVSR is to ground the relevant answer segments from the video corpus for a given question.

**Video Retrieval.** The objective of Video Retrieval is that given a text query and a pool of candidate videos, select the video which corresponds to the text query. A good amount of work has been done in the area of natural language query based video search for complete videos, such
as MSR-VTT [43], DiDeMo [3], ActivityNet Captions [18], CharadesSTA [6], and TACoS [34]. However, returning the whole video is not always desirable, since sometimes they can be quite long (e.g. from a few minutes to hours). What’s more, most of the queries for Video Retrieval are based on declarative sentences, but not question-based, thus leading to the gap to real world application. However, the setting of video retrieval is far from practical application due to the factoid query. In AQVSR, we focus more on the affordance-centric user question expressed in a multi-modal manner and adopt the video segment level retrieval, whereby the returned video segments within proper time-span are able to address the user’s question.

**Moment Localization.** The task of Moment Localization is to localize moments from a video given a natural language query. Various datasets [3, 6, 18, 24, 34] have been proposed or repurposed for the task. While these datasets for moment localization collect only a single moment for each query-video pair, we annotate one or more video segments for each query in our dataset, which is more flexible. Meanwhile, most existing datasets [7, 12, 37, 39] for moment localization are query-agnostic, which do not provide customized moments for a specific user query. Recently [20] introduced a dataset for query-based video moment retrieval and highlight detection, which also lies in the category of moment localization. Unlike moment localization, AQVSR is question-driven and aims to retrieve all relevant segments. Sec. 3 analyses the challenges in adapting moment localization for AQVSR and our rationales behind formulating it as a segment retrieval problem.

### 3. Task Formulation for AQVSR

**Challenges.** Given a user question, naturally we would think of annotating the start time and end time for the answer span, as moment localization does. However, this leads to some problems. (1) **From the perspective of application,** different users have different preference for answer clips: some prefer to longer clips to learn more contexts, while others just tend to watch the key parts. This leads to the difficulty in the definition of correct answer span. (2) **From the perspective of metric evaluation,** in the task of moment localization, models are forced to generate a prediction with high tIoU to the ground truth annotation. However, tIoU might not measure the correctness of the predicted answer span well. For example, a predicted answer span with higher tIoU might miss some vital information for addressing the user’s question, while another candidate answer span with lower tIoU to ground truth might work better qualitatively. What’s more, a prediction span generated by localization might not be semantically intact, which is not user friendly.

**Our proposed solution of evaluating on segments.** To avoid such drawbacks and remove ambiguities in answer annotation, we propose to chunk a video into segments with well-designed rules and then annotate the answer segments among the candidate video segments. With predefined segments, the model no longer needs to predict the timestamp, but only needs to predict whether a segment contains the content for answering the user’s question or not, which significantly alleviates ambiguity during evaluation.

Notably, throughout multiple iterations of improving annotation guidelines, we arrived at the following design principles of how to segment a video: (1) **Keep consistent level of semantic granularity within one video segment.** (2) **Soft threshold for video segment duration:** 30 seconds as minimum and 2 minutes as maximum. (3) **Adjust segmentation according to the list of qualified questions.** More details and examples could be found in Supp.

**Task definition of AQVSR.** The task of AQVSR addresses an affordance-centric user question by retrieving from the video corpus the relevant video segments. Taking a question \( Q \) with multiple modalities as input, from the corpus of video segments \( V \), the task is to retrieve a set of relevant segments which can provide answer contents for the given question: \( V_{ans} = \{ v_i^q \mid q = Q, i \in |V|, v_i^q \in V \} \).

### 4. A New Benchmark: AssistSR

**Overview.** Our pipeline of data collection and annotation procedure is summarized in Fig. 3. In short, we conduct video collection, question collection, video segmentation and answer segment annotation for data collection. Quality control is carefully designed to ensure the data and annotation quality. We will explain them one by one.

**Challenges.** Here we list the challenging points in data collection and annotation, and provide our solutions and rationales behind. (1) **Source diverse videos of high-quality.** We show our principles in Sec. 4.1. (2) **How to ensure the quality of questions:** ask the ones people would naturally ask and balance the question types. We provide a verification-involved process in Sec. 4.2.

**Summary of our AssistSR dataset.** In total, we collected 1,431 questions associated with 1,065 segments in 130 videos for the AssistSR dataset. We provides some ex-

![Figure 3. Data collection procedure for our AssistSR benchmark.](image-url)
amples of AssistSR along with qualitative results in Fig. 8.

4.1. Video Collection

We collect a set of videos which cover a wide range of scenarios and contain interesting and diverse contents. Here, each scenario refers to a category of commonly used item in daily life, such as vacuums, digital watches and so on.

To source videos, we firstly maintain a scenario list which covers the majority of the highest level in HowTo100M [30] to ensure diversity. To obtain high-quality videos, we further set some criteria:

- To utilize auto-generated ASR captions, we exclude videos without voice-over or with only scene-text.
- Videos should contain actions or salient motion for showcasing the feature of the subject item;
- Video contents should be rich enough to support addressing user’s question.

Overall, these videos are captured via different devices (e.g., mobile phone or GoPro) with different view angles (e.g., first-person view or third-person view), posing important challenges to computer vision systems. All the transcripts generated by ASR are in English.

In total, we collected 130 videos with an average duration of 574 seconds. All collected videos are mainly from the domain of commonly used items, including sub-categories ranging from home appliances, digital gadgets to smart devices. Fig. 4 shows some scenarios with high frequency in our dataset.

![High frequency scenarios in our dataset.]

Figure 4. High frequency scenarios in our dataset.

4.2. Question Collection

Procedure for annotating questions. To collect high-quality questions, the annotators were asked to:

1. Go through and understand the whole video.
2. Come up with a list of questions which can be clearly tackled after watching the video. We exclude the questions that a user would not be interested in. For example, “What color is this juicer?” which is superficial.
3. Conduct self-verification and cross-verification on the question list. Each annotator checks its question list and filters out the unqualified questions while for cross-verification, another annotator who has not watched the same video checks the questions from the view of a real user.

4. Collect and annotate the qualified questions. After harvesting the qualified textual questions, the annotators are required to collect and annotate the query images.

Next, we provide details for our question collection and annotation.

Multimodal query. Each question annotated in AssistSR is multimodal, which means that it is composed of a textual part and a visual part. The textual part is a free-form query written by our annotators. The visual part is an image for the object the annotator poses a question on. Inspired by [29, 45], we allow annotators to pose a question on specific parts of an object and locate its position with a bounding box. By incorporating the visual information, annotators are able to write more natural questions. For example, they can ask “How to clean up <this part>?” instead of “How to clean up the plastic dust bucket of my Roidme X20S vacuum?”, with a bounding box tag referring to the specific part of the vacuum cleaner.

Query type. In practice, videos are associated with multi-modalities such as audio and text, e.g. subtitles for TV shows, transcripts for instructional videos or audience discourse accompanying live stream, which could be also important sources for retrieving relevant segments. Following [22, 24], we encourage annotators to write questions that are related to different modalities, aiming at enabling the model to learn the knowledge from the video in a systematic way. To test the model’s ability to retrieve the relevant segment for a given question, we categorize all the questions into 3 types:

1. t type: the visual part of the annotated question only helps to match the subject in the video, one can retrieve the relevant clips by textual clues from videos besides such matching.
2. v type: one can only locate the relevant clips by visual clues from videos.
3. v+t type: one should locate the relevant segments by leveraging both visual and textual clues from videos.

In this case, the visual part of the question not only helps to match the subject, but also provides vital information for segment retrieval.

Similar to [22, 24], in our pilot test, we observed that our annotators preferred to write t type questions, of which the answers segment can be easily retrieved by reading the transcripts. To ensure that we collect a balance of queries requiring one or both modalities, we set up awarding mechanism for asking v type and t+v type type questions. More examples of query types could be found in Supp.
Overview of the collected questions. We collected 1,431 questions in total. On average, our question contains 8.49 words. Each question is associated with one user image and with 0.54 bounding boxes annotated on average. Fig. 6b shows the distribution of question length, showing that most of the queries in our dataset have more than 7 words.

In Fig. 5, by visualizing the Alluvial diagram of the most frequent first three words in the questions of our dataset, it demonstrates that the text query pattern in our dataset is not limited to “how to” type questions. In real jobs, annotators can also ask questions with “what”, “when”, “where” and “why”. Also, they are allowed to state the goal (e.g. “I want to ...”) with a declarative sentence first and then ask a question. We can observe from Fig. 5 that “how to” type questions is the most dominant in our dataset, followed by “what” and “where”. The third word in “how to” type questions are mainly verbs, showcasing the diversity and complexity of our questions, which is affordance-centric.

4.3. Video Segmentation

Following Sec. 3, we ask annotators to chunk a video into predefined segments and detailed analysis can be found in Sec. 3 and Supp.

Finally we collected 1,065 segments upon 130 videos. The average duration for all chunked video segments is 70 seconds. Fig. 6a shows the distribution of the length of the chunked video segments. Most of the video segments in our dataset have a duration longer than 40 seconds.

4.4. Answer Annotation

With predefined segments, annotators are required to select the answer segments for each qualified question. Each query with contents that can address the given question should be annotated. In AssistSR, there could be multiple disjoint answer segments paired with a single query (on average 1.041 answer segments per question), while all the moment localization or video retrieval datasets can only have one single moment or one single video. This is a more realistic setup as relevant content to a query in a video might be separated by irrelevant content.

4.5. Quality Control

To ensure data and annotation quality, we’ve designed a detailed quality assurance mechanism, which is presented in Supp. Briefly, annotators were required to take the training curriculum and pass the qualification test before working on real jobs. Auditors are to evaluate the workers performance on the qualification test and check the quality of sampled annotations from the initial data pool. Typical issues in early training include asking unnatural questions, misunderstanding the rules for chunking a video, etc. Evaluation scores are rated on a scale of 1 (bad), 2 (minor errors) and 3 (good). Workers should retake the training curriculum when the rating score is deemed insufficient. In practice, the performance of an annotator is satisfying and acceptable if its average rating is above 2.5.

5. Method

We introduce Dual Multimodal Encoders (DME) for AQVSR. Fig. 7 gives an overview of our model. Formally, we define the inputs into the model as: a candidate video segment associated with transcripts and a user question composed of an image with or without referring regions, and a textual query.

5.1. Model architecture

DME is composed of two multimodal encoders based on the transformer network [41]. In AQVSR, both user question and a context video segment are multimodal, i.e. both of them are composed of vision input and language input, requiring a model for learning joint contextualized representations. Inspired by the recent success of the vision and language model [25, 26, 38], we follow [25] to use the self-attention mechanism within the Transformer to implicitly align elements of the input text and regions in the question and model the contextual information of the input video segment.

5.2. Input Representation

We extract visual features for video appearance, query image and region with ResNet-50 backbone [13] with
weights from grid-feat [15], which is trained on Visual Genome [19]. We extract contextualized text features using a 6-layer pretrained DistilBert [36] from [42]. Both visual features and textual features are projected into the same dimension, and added with token type embedding and positional embedding. We concatenate the text feature and visual feature to create a single sequence embedding and a learnable \([\text{cls}]\) token [5] is concatenated to the beginning of the input feature, which is used to produce the final output representation of each transformer-based encoder.

### 5.3. Training and Inference

During training, one relevant segment is randomly sampled to be paired with a question. For training loss, we employ [46] for segment retrieval setting, where matching question-segment pairs in a batch are treated as positives, and all other pairwise combinations in the same batch are treated as negatives. We maximize the score between positive pairs and minimize the score between negative pairs. At inference time, the DME model requires only the dot product between the multimodal question embedding and candidate video segment embeddings. More implementation details could be found in Supp.

### 6. Experiment

#### 6.1. Baseline

To explore the performance of other video-related tasks on AQVSR, we slightly modified the following #2-#4 methods from their original implementations for the AQVSR.

**#1 Random guess.** For each question, we randomly shuffle the list of all video segments as the ranked results.

**#2 TVQA** [22] proposed a multi-stream end-to-end trainable neural network for Multi-Modal Video Question Answering. In this model, the question-answer pairs are used to fuse with visual features and text features in the video separately. We modify the input module to fit in our multimodal question, and fuse the question with different modalities from the paired video segment separately.

**#3 XML** [24] is a late fusion approach for Video Corpus Moment Retrieval (VCMR). In XML, separated self-cross-encoders are used to encode visual and textual features of a video. The query is fed into a self-attention module followed by a FC-layer to generate modularized query representations. Late fusion is then applied for Video Retrieval and moment retrieval. We add a module image query in AQVSR, which is the same to the original query branch in XML. We use element-wise adding for image query and text query, to generate modularized query. XML(VR) views each video segment in AQVSR as a video, and just compiles the Video Retrieval part in XML. XML(VR+ML) uses the rough start-end time in AssistSR annotation for VCMR setting. At inference time, we pick the highest score of the moments in video segments as the score for that video segment; for each query, we rank the scores of all candidate video segment.

**#4 ClipBERT** [21] is a generic framework for video-language tasks. We modify ClipBERT by adding ResNet-50 [13] C5 max-pooled feature of the image query to text query feature. We also add corresponding transcript features for sampled frames to fit in the AssistSR input.

More implementation details could be found in Supp.

#### 6.2. Experimental Setting

**Dataset.** We split our AssistSR into 70% train, 15% val and 15% test such that videos and their associated queries appear in only one split. We include scenarios which are unseen in the train set into the validation set and test set. The distribution of query patterns and query types are kept aligned among these splits.

**Evaluation metrics.** To evaluate the performance of AQVSR, we use mean average precision (mAP) as in the information retrieval system. We also report standard metric Recall@1 (R@1), Recall@5 (R@5), Recall@10 (R@10) and Recall@50 (R@50) used in text-to-video retrieval and single moment retrieval. For AQVSR, we use mAP as the main metric.

#### 6.3. Comparisons with State-of-the-art Methods

**Performances on all scenarios.** AssistSR val set results are shown the Tab. 1. (1) All the methods for video-related tasks are clearly better than random guess. DME clearly outperforms other methods for video-related tasks, indicating that models designed for other tasks can not generalize...
Table 1. Experimental results on various methods for AQVSR.

| Method          | mAP  | R@1  | R@5  |
|-----------------|------|------|------|
|                 | All  | Unseen | All  | Unseen | All  | Unseen | All  | Unseen | All  | Unseen |
| Random guess    | 3.66 | 3.65 | 0.66 | 0.64 | 3.24 | 3.24 | 6.45 | 6.44 | 32.26 | 32.34 |
| TVQA [22]       | 10.13 | 10.32 | 2.97 | 3.51 | 12.13 | 12.28 | 27.97 | 28.07 | 61.88 | 55.26 |
| ClipBERT [21]   | 14.36 | 7.28 | 5.20 | 1.75 | 19.80 | 5.26 | 32.43 | 17.54 | 78.71 | 63.16 |
| XML(VR) [24]    | 12.56 | 8.89 | 3.47 | 0.00 | 19.31 | 12.28 | 30.20 | 19.30 | 78.47 | 68.42 |
| XML(VR + ML) [24] | 14.56 | 10.13 | 3.71 | 1.75 | 24.01 | 14.91 | 34.41 | 25.44 | 79.21 | 62.28 |
| DME(ours)       | 20.35 | 15.43 | 10.89 | 9.65 | 28.47 | 18.42 | 40.84 | 28.95 | 72.52 | 57.89 |

Table 2. Ablation experiment results on various modalities for AQVSR.

| Question | Text | Image | Trans. | Video | mAP  | R@1  | R@5  |
|----------|------|-------|--------|-------|------|------|------|
|          | All  | T     | V      | T + V | All  | T    | V    |
| #1       | ✓    | ✓     | ✓     | ✓     | 11.66 | 7.06 | 12.04 | 14.65 | 2.48 | 1.33 | 7.69 |
| #2       | x    | ✓     | ✓     | ✓     | 12.96 | 11.67 | 5.38 | 14.67 | 3.47 | 4.00 | 0.00 |
| #3       | x    | x     | ✓     | ✓     | 10.82 | 8.46 | 7.01 | 12.80 | 1.49 | 2.67 | 0.00 |
| #4       | ✓    | ✓     | ✓     | x     | 16.67 | 15.13 | 11.17 | 18.32 | 8.17 | 5.33 | 7.69 |
| #5       | ✓    | x     | ✓     | ✓     | 13.09 | 14.10 | 2.78 | 13.60 | 5.45 | 5.33 | 0.00 |
| #6       | ✓    | x     | x     | ✓     | 15.38 | 15.18 | 5.32 | 16.88 | 5.20 | 4.00 | 0.00 |
| #7       | ✓    | ✓     | ✓     | ✓     | 20.35 | 20.56 | 13.23 | 21.03 | 10.89 | 12.00 | 7.69 |

6.5. Ablation study

6.5.1 Ablation for different modalities.

In Tab. 2, we compare to variants that mask some of the text and visual features for the encoder in order to further verify the contribution of different modalities in our task. We first mask textual modalities(#1-#3) and then visual modalities(#4-#6) for the encoders, and analyze the effect on our model respectively. The result in #7 shows that the model with all modalities has significantly higher performance than those without some of the modalities, suggesting that all of the modalities have non-negligible contribution.

Modality removal from video source. For the video encoder, we respectively mask the transcript (shown as Trans.) and the video appearance (shown as Visual). Comparing #1 and #4 in Tab. 2, the mAP of all questions drops significantly when either the transcript or video appearance is masked, while masking transcript causes a more drastic drop. For t type and t+v type questions, mAP drops more in the case when masking transcript than masking video appearance. In contrast, mAP corresponding to v type questions drops more when the video appearance is masked. These results indicate that the transcript weighs more than video appearance, which could be attributed to two possible reasons: (1) Instead of using raw images and video frames as input, we use extracted features from pre-trained gridfeat [15] without fine-tuning. The adopted extractor might not match well with our task. (2) The existing method over-
relies on the textual information and lacks the alignment between text and vision, thus undermining the contribution of visual information. This can be observed from #7: with all modalities, the mAP of v type questions is lower than others.

What if removing modality from the user’s question?
Practically, users may bring about more simplified questions with less modalities considering the simplicity of using. Thus we mask the textual and visual part of the query (shown as Text and Image in Tab. 2) respectively for the encoder and analyze the effect on the model performance. Comparing #2 and #3, the mAP of all questions declines to almost the same when the visual part and when the visual part is masked. For v type and t+v type questions, the mAP declines less when masking the textual part than masking the visual part. In contrast, the mAP of t type queries declines more when the text query is masked. The results of comparison also indicate that textual information is better used than visual. Considering the complexity of our task, masking any modality could severely affect the model’s performance. Therefore, using multi-modalities is a direct way to improve the interaction quality.

6.5.2 Ablation for input fusion type.
We study the effect of different fusion types for the input of DME model. Here we conduct experiments on two fusion types for the multimodal encoder. (1) Adding fusion. For each multimodal encoder in DME, fuse visual and textual features of a video with positional alignment via element-wise adding as in [38]; token type embedding is not adopted in this case. (2) Concatenating fusion. For each multimodal encoder in DME, we concatenate the textual feature and visual feature into a single sequence embedding. This is used for our DME baseline. Experiment results in Tab. 3 show that applying concatenating fusion yields better performance over adding fusion. This suggests that to give full play to the advantages of Transformer in multimodal reasoning, it is necessary to decouple multi-modal information into separate sequences.

| Method                | mAP  | R@1  | R@5  | R@10 |
|-----------------------|------|------|------|------|
| Adding Fusion         | 16.32| 6.44 | 25.00| 39.36|
| Concatenating Fusion  | 20.35| 10.89| 28.47| 40.84|

Table 3. Results for different input fusion types for DME.

7. Conclusion and Future Work
This paper presents a new task, Affordance-centric Question-driven Video Segment Retrieval (AQVSR), requiring intelligent systems to address the given affordance-centric question by retrieving relevant segments from a video corpus. To support the evaluation of such a task, we construct a dataset, AssistSR, which contains 1.4k questions on 1k video segments from instructional videos on diverse daily-used items. Further, we provide analyses of this new dataset as well as several baselines and a multi-stream end-to-end trainable neural network framework for AQVSR. The experimental results show that AQVSR is still a formidable challenge for all current methods. We hope this benchmark can help drive the vision and language models to be deployed in real-world applications and boost the development of AI Assistants.

Limitations and societal impact. For the limitations of this work, we realize that the size of our dataset is limited due to the difficulties in the time consuming data collection and annotation process. We will make further efforts to create large-scale datasets in this topic, including widening for more scenarios and designing strategies to scale up data collection in a more efficient manner. For societal impact, we believe the proposed AQVSR is an important stepping stone towards building an intelligent agent. But to prevent the abuse of this technology, such as assisting unauthorized users to use dangerous items, we will take this into account and prohibit misuse cases explicitly in our license when releasing the developed codes and models.
8. Acknowledgement

This project is supported by Mike Shou’s SUG grant partially under Singapore Ministry of Education Academic Research Fund Tier 1.

9. Supp: More details of AssistSR Dataset

9.1. Comparison with other datasets.

In Tab. 4, we compare AssistSR to previous VideoQA, Video Retrieval, Video Moment Localization datasets. In summary, our dataset is possessed with multimodal questions. In addition, we provide affordance-centric questions, aiming to drive the models to learn the knowledge from multimodal instructional videos to address the user problem beyond simple facts. Besides, AssistSR can have multiple disjoint answer segments paired with a single query (on average 1.041 answer segments per question), while all the moment localization or video retrieval datasets can only have one single moment or one single video. This is a more realistic setup as relevant content to a query in a video might be separated by irrelevant content.

9.2. Visualization Examples

In Fig. 9, we show some visualization examples in our datasets. In AssistSR, each multimodal question is associated with an image (with or without bounding boxes for referring to a specific region) and a human-written query. GT video segments contain the content which can address the given question.

9.3. Video Segmentation

Principles and rationales for video segmentation. We provide principles for video segmentation in Sec. 3 in our paper. Here we provide more details: give a video, we ask the annotator to segment the videos following these principles:

1. Keep consistent level of semantic granularity within one video segment. The idea is that we want the content within one video segment to be semantically consistent. For example, a qualified segment could be the one introducing different aspects of a subject or introducing partial contents of an aspect (i.e. we can divide the content of this aspect into several segments).

2. Soft threshold for video segment duration: 30 seconds as minimum and 2 minutes as maximum. User study [1, 2], found that video clips should fully engage the audience within the first 30 seconds to attract their attention. Also, it is observed that keeping videos shorter than 2 minutes achieves the most viewer engagement while engagement drops off sharply after 2 minutes. Therefore, in video segmentation, we restrict the duration of each segment by setting a soft threshold.

3. Adjust segmentation according to the list of qualified questions. Since we adopt a flexible manner for segmenting a video, the result segmentation is not unique and thus adjustable. In order to test the models’ ability on retrieving the correct answer segment, we further ask our annotators to adjust the segmentation based on the
raised question. For example, if one video segment contains both contents of correct answer and distraction of a question, the annotator should separate this into a positive segment and a hard negative segment.

10.4. Quality Control

Here we provide the detailed quality assurance guideline for the auditor’s evaluation. As shown in our main paper, auditors are to evaluate the workers performance on the qualification test and check the quality of the sampled annotations from the initial data pool. To ensure the annota-
Table 4. Comparison of AssistSR to various existing VideoQA/Video Retrieval/Moment Localization datasets.

| Dataset           | Video source | Query source | Query modal | Affordance-centric |
|-------------------|--------------|--------------|-------------|--------------------|
| Movie-QA [40]     | movie        | ✓            | ✓           | -                  |
| TVQA [22]         | TV show      | ✓            | ✓ ✓         | -                  |
| TACoS [34]        | cooking      | ✓ ✓          | ✓ ✓         | -                  |
| DiDeMo [3]        | Flickr       | ✓ ✓          | ✓ ✓         | -                  |
| ActivityNet Cap [18] | Activities | ✓ ✓          | ✓ ✓         | -                  |
| TVR [24]          | TV show      | ✓ ✓ ✓        | ✓ ✓         | -                  |
| ScreenCast QA [47] | Software tutorial | ✓ ✓ ✓        | ✓ ✓         | ✓                  |
| ours              | Tutorial for devices | ✓ ✓ ✓        | ✓ ✓         | ✓                  |

Figure 10. Query type distribution of the AssistSR dataset.

Figure 11. Distribution for $\frac{\text{Area(bbox)}}{\text{Area(img)}}$.

type 41%
v type 9%
t+v type 50%

In the following, we firstly introduce the rating score definition for auditing, followed by details for different sub-tasks in annotation.

10.4.1 Rating score definition

1, bad: The sourced data / annotation is of low quality. E.g., the annotator sourced unqualified videos for subsequent annotations; asked unnatural/uninteresting questions too often, chunked the videos mistakenly; wrong query type. 2, medium: occasional error. E.g., less than 20% questions are unnatural/uninteresting, but can be improved. The main purpose of this intermediate level is to reflect the need of improvement for the annotator while acknowledging the annotator’s correct understanding of the guideline. 3, good: source high-quality videos. Good questions, segmentations, etc.

10.4.2 Video collection

1, bad: videos without voice (thus cannot obtain transcripts by ASR); videos without salient action (only scene text; only introduce the features by words); videos cannot support any interesting questions; more than 20% of videos do not meet the requirements. 2, medium: 10%-20% videos do not meet the requirements in the guideline. 3, good: satisfies the requirements in our guideline: videos with voice (someone is explaining how to use the device/finish a task; transcripts can be automatically generated); videos with salient action (someone is showing how to use the item); videos can support asking affordance-centric questions.

10.4.3 Question collection

We firstly define unnatural questions and uninteresting questions with examples. Unnatural questions refer to the ones a user would not naturally ask in daily life. For example, for a knob on a microwave, a user might not ask “What would happen if I turn this clockwise?”. Instead, he would ask “How to set the heating time to 90 seconds” for using purpose. Uninteresting questions refer to the ones which...
are easy, e.g., “How to turn it on?”, the answer of which is obvious. 1, bad: more than 20% questions are unnatural or uninteresting; miss v type questions if it is obvious that one video is found to support asking several questions of this type; mistakes in annotating query type: the correct type should be t type but annotated as t+v type; the correct type should be t+v type but v type but annotated as t type, the correct type should be t type but annotated as v type, the correct type is v type but annotated as t type. 2, medium: occasionally missed the segmentation adjusting the segmentation based on the questions. 2, medium: occasionally missed the segmentation adjusting the segmentation based on the questions. 3, good: Annotation satisfies the requirements in our guideline: ask natural and interesting questions, the textual and visual parts of the query make the question clear and easy to understand; understand the meaning of the query type and annotate correctly.

10.4.4 Video segmentation

1, bad: non-consistent semantic level within one segment; there exists short segments (obviously less than 30s) which should be merged with other segments; do not refine the segmentation based on the questions. 2, medium: occasionally missed the segmentation adjusting the segmentation. 3, good: follow the three principles for video segmentation.

10.4.5 Answer annotation

1, bad: miss important contents to answer that question; mistakenly include totally irrelevant segments. 3, good: the annotator correctly annotates all the relevant segments.

11. Supp: More Details of DME Method

Fig. 7 gives an overview of DME for AQVSR. Formally, we define the inputs into the model as: a candidate video segment associated with transcripts and a user question composed of an image with or without referring regions, and a textual query.

In our method, we represent each chunked video segment v_i as a list of consecutive short clips, i.e., v_i = [c_{i,1}, c_{i,2}, ..., c_{i,l}], where l is the length of video segment v_i (#clips). In AssistSR, each short clip is also associated with temporally aligned transcripts.

Model architecture DME is composed of two multimodal encoders based on the transformer network [41]. In AQVSR, both user question and a context video segment are multimodal, i.e., both of them are composed of vision input and language input, requiring a model for learning joint contextualized representations. Inspired by the recent success of the vision and language model [25, 26, 38], we follow [25] to use the self-attention mechanism within the Transformer to implicitly align elements of the input text and regions in the question and model the contextual information of the input video segment.

Input representations. To represent a video segment, we consider appearance features. For each frame, we used Resnet-50 [13] with weights from grid-feat [15], which is trained on Visual Genome [19] for object detection and attribute classification and produces effective features for image VQA tasks [4, 11]. We resize a frame to 448 × 746 and max-pool feature map after the C5 block to get the 2048D representation. We extract 2048D ResNet-50 features at 10 FPS and max-pool the features every 1.5 seconds to obtain a clip-level feature. To represent transcripts, we extract contextualized text features using a 6-layer pretrained DistilBert [36]. We used the implementation of DistilBert from [42] to extract contextualized token embeddings from its second-to-last layer [23]. After extracting the token-level embedding, we then max-pool them every 1.5 seconds to get a 768D clip-level feature vector. We use a 768D zero vector if encountering no transcripts.

To represent the image and referring region in the question, we used Resnet-50 [13] with weights from grid-feat [15]. For the image, similar to video appearance, we feed it into the Resnet-50 and max-pool the feature map after the C5 block to obtain a 2048D feature. For a referring region located by a bounding box, we apply RoI pooling [9, 10] on the feature map after Resnet-50 C5 block to obtain a 2048D region feature. To represent the text query in the question, we directly used the extracted token embeddings from the second-to-last layer of DistilBert [36].

Following [25], all the extracted visual features are projected into 768D features via a linear layer. Without ambiguity, we use the used the symbols by denoting the processed symbols as E^{Q_t} \in \mathbb{R}^{l_q \times d}, E^{Q_v} \in \mathbb{R}^{N_v \times d}, E^{V_f} \in \mathbb{R}^{l_f \times d} and E^{V_i} \in \mathbb{R}^{l_i \times d}$, where $Q_t$ represents the textual query of the question, $l_q$ is the length of this query; $Q_v$ means the visual feature of the question input, $N_v$ is the number of regions, including the whole image and regions being referred to; $V_f$ and $V_i$ means frames and transcripts of the video segment, $l$ is the number of clips in this video segment and $d$ is the hidden size ($d$ is set to be 768 in our experiments).

For multimodal encoders, we inject different types of input attribute into E by adding two additional embedding layers: (1) token type encoding that informs the type of information: using [vis] for visual features in the question and appearance feature in the video segment, while using [txt] for features of textual query in the question and transcript feature in video segment. (2) position encoding that is used to inject signals of the token ordering. For textual query in the question, position encoding is following the sequence order. For appearance and transcript
features in the video segment is following the order of clips sequence. As for visual features in the question, position encoding is used when alignments between words and bounding regions are provided as part of the input, and is set to the sum of the position embeddings corresponding to the aligned words as in [25].

These layers are trainable to enable models to learn the dynamics of input features and are modeled to have the same feature dimension \( d \). We combine all encoding layers through element-wise summation for each modality in a multimodal encoder. Specifically for \( m \in \{Q_t, Q_v, V_j, V_t\} \), the result representation is:

\[
Z^m = E^m + E_{tok}^m + E_{pos}^m.
\]

For input representation of each encoder, we concatenate the text feature and visual feature to create a single sequence embedding:

\[
Z^Q = [Z^Q_t; Z^Q_v] \quad \quad Z^V = [Z^V_t; Z^V_j].
\]

**Multimodal encoding** Given the input features \( Z^Q, Z^V \), as shown in Fig. 7, we use two encoders to compute their representation respectively. For each multimodal encoder, a learned [cls] token [5] is concatenated to the beginning of the input feature, which is used to produce the final output representation of the transformer. We denote the output of the Query Encoder as \( H^Q \in \mathbb{R}^d \) for the question \( Q \), and the output of the Video Encoder as \( H^V \in \mathbb{R}^d \) for the video segment \( V \).

**Training and Inference** During training, one relevant segment is randomly sampled to be paired with a question. For training loss, we employ [46] for segment retrieval setting, where matching question-segment pairs in a batch are treated as positives, and all other pairwise combinations in the same batch are treated as negatives. We maximize the score between positive pairs and minimize the score between negative pairs. We minimise the sum of two losses:

\[
L_1 = -\frac{1}{B} \sum_i^B \log \frac{\exp \left( H_i^V H_i^Q / \sigma \right)}{\sum_{j=1}^B \exp \left( H_i^V H_j^Q / \sigma \right)},
\]

\[
L_2 = -\frac{1}{B} \sum_i^B \log \frac{\exp \left( H_i^Q H_i^V / \sigma \right)}{\sum_{j=1}^B \exp \left( H_i^Q H_j^V / \sigma \right)},
\]

where \( H_i^V \) and \( H_i^Q \) here are the normalized embeddings of the \( i \)-th video segment and the \( j \)-th question in a batch of size \( B \) and \( \sigma \) is the temperature. The overall loss function is \( L = L_1 + L_2 \) for DME.

At inference time, the DME model requires only the dot product between the multimodal question embedding and candidate video segment embeddings. This retrieval inference is of trivial cost since questions and video segments are indexable and therefore it is scalable to large scale retrieval.

### 12. Supp: More Details of Experiment

#### 12.1. Implementation Details

We use the same set of offline extracted features for DME, TVQA [22] and XML [24]. We use ResNet-50 [13] as feature extractor for image query and use a 6-layer DistilBert [36] for text feature extraction when modifying ClipBERT [21]. For all transformer-based encoders, we set the number of hidden layers to be 4 and set hidden size to be 768. We keep other settings the same as the original implementation for TVQA\(^1\), XML\(^2\) and ClipBERT\(^3\). For DME, We use Adam [17] optimizer and a learning rate of \( 3 \times 10^{-5} \). The hyper-parameter \( \sigma \) for the loss is set to be 0.05.

#### 12.2. More Experimental Results

In Tab. 5, we show the performance (mAP/Recall@1/Recall@5/Recall@10/Recall@50) of different methods on AssistSR. We show the performance on all questions, seen scenarios (related to 73.7% questions) and unseen scenarios (related to 26.3% questions). We can observe that (1) DME clearly outperforms other methods for video-related tasks, indicating that models designed for other tasks can not generalize well to the AQVSR task. (2) For DME, ClipBERT [21] and XML [24], performance on seen scenarios is better than that on unseen scenarios, indicating that for these methods, they are able to capture something in common within the same scenario but fail to generalize well to unseen scenarios. This is because for items within the same scenario could share similar knowledge, such as structure and functionality, while for items for totally different scenarios, models may fail to

\(^1\)https://github.com/jayleicn/TVQA  
\(^2\)https://github.com/jayleicn/TVRetrieval  
\(^3\)https://github.com/jayleicn/ClipBERT

| Method               | mAP  | R@1  | R@5  | R@10 | R@50 |
|----------------------|------|------|------|------|------|
| TVQA [22]            | 10.13| 10.06| 10.32| 2.77 | 3.51 |
| ClipBERT [21]        | 14.36| 17.14| 7.28 | 5.66 | 1.75 |
| XML/VR [24]          | 12.56| 14.00| 8.89 | 3.47 | 4.83 |
| XML/VR + ML [24]     | 14.56| 16.30| 10.13| 3.71 | 4.48 |
| DME (ours)           | 20.35| 22.28| 15.43| 9.65 | 28.47|

Table 5. Val set results on various methods for AQVSR. “All” means all queries; “Seen” means queries for seen scenarios; “Unseen” means queries for unseen scenarios.
bridge the gap of the difference in appearance, language style, etc.

References

[1] How long should videos be? the ideal length for marketing. [online]. 2020. https://www.talkingtreecreative.com/blog/video-marketing-2/the-impact-of-video-length-on-engagement/. 9

[2] Video length: 4 tips that will help you boost engagement. [online]. 2020. https://meetmaestro.com/blogs/how-long-should-your-video-be/. 9

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 5, 6, 12, 13

[4] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Self-bridging convolutional neural networks for 3d object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5803–5812, 2017. 1, 3, 11

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015. 12

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019. 6, 13

[7] Ana Garcia del Molino and Michael Gugli. Phd-gifs: Personalized highlight detection for automatic gif creation. pages 600–608, 10 2018. 3

[8] James J Gibson. The theory of affordances. Hilldale, USA, 1(2):67–82, 1977. 1

[9] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE conference on computer vision, pages 5267–5275, 2017. 3

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014. 12

[11] Danna Gurari, Qi Liu, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3608–3617, 2018. 12

[12] Michael Gugli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. pages 1001–1009, 06 2016. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 5, 6, 12, 13

[14] Yunseeok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2758–2766, 2017. 2

[15] Huaiyu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10267–10276, 2020. 6, 7, 12

[16] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. pages 2016–2022, 08 2017. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2015. 13

[18] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision, pages 706–715, 2017. 3, 11

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 6, 12

[20] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvahighlights: Detecting moments and highlights in videos via natural language queries. arXiv preprint arXiv:2107.09609, 2021. 3

[21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In CVPR, 2021. 6, 7, 13

[22] Jie Lei, Tamara L Berg, and Mohit Bansal. Tvqa: Localized, compositional video question answering. In EMNLP, 2018. 2, 4, 6, 7, 9, 11, 13

[23] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. pages 8211–8225, 01 2020. 12

[24] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In ECCV, 2020. 3, 4, 6, 7, 9, 11, 13

[25] Liumin Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Byoung-Tak Zhang, Diederik P. Kingma, and Jimmy Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2015. 13

[26] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In CVPR, 2021. 6, 7, 13

[27] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6884–6893, 2017. 2

[28] Bertram F Malle, Louis J Moses, and Dare A Baldwin. Intentions and intentionality: Foundations of social cognition. MIT press, 2001. 2, 9
