AVQA: A Dataset for Audio-Visual Question Answering on Videos

Pinci Yang
ypc19@mails.tsinghua.edu.cn
Tsinghua-Berkeley Shenzhen
Institute, Tsinghua University

Xin Wang*
xin_wang@tsinghua.edu.cn
Tsinghua University

Xuguang Duan
dxg18@mails.tsinghua.edu.cn
Tsinghua University

Hong Chen
h-chen20@mails.tsinghua.edu.cn
Tsinghua University

Runze Hou
hrz21@mails.tsinghua.edu.cn
Tsinghua-Berkeley Shenzhen
Institute, Tsinghua University

Cong Jin
jincong0623@cuc.edu.cn
Communication University of China

Wenwu Zhu*
wwzhu@tsinghua.edu.cn
Tsinghua University

ABSTRACT

Audio-visual question answering aims to answer questions regarding both audio and visual modalities in a given video, and has drawn increasing research interest in recent years. However, there have been no appropriate datasets for this challenging task on videos in real-life scenarios so far. They are either designed with questions containing only visual clues without taking any audio information into account, or considering audio with restrictions to specific scenarios, such as panoramic videos and videos about music performances. In this paper, to overcome the limitations of existing datasets, we introduce **AVQA**, a new audio-visual question answering dataset on videos in real-life scenarios. We collect 57,015 videos from daily audio-visual activities and 57,335 specially-designed question-answer pairs relying on clues from both modalities, where information contained in a single modality is insufficient or ambiguous. Furthermore, we propose a Hierarchical Audio-Visual Fusing module to model multiple semantic correlations among audio, visual, and text modalities and conduct ablation studies to analyze the role of different modalities on our datasets. Experimental results show that our proposed method significantly improves the audio-visual question answering performance over various question types. Therefore, AVQA can provide an adequate testbed for the generation of models with a deeper understanding of multimodal information on audio-visual question answering in real-life scenarios. (The dataset is available at https://mn.cs.tsinghua.edu.cn/avqa)

CCS CONCEPTS

• Computing methodologies → Computer vision; **• Information systems → Question answering**.

KEYWORDS

dataset, audio-visual question answering, multimodal

ACM Reference Format:
Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. AVQA: A Dataset for Audio-Visual Question Answering on Videos. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3503161.3548291

1 INTRODUCTION

Some psychological and cognitive science studies [1, 16, 34] show that multimodal data plays a crucial role in the human brain's cognition system to form a whole coherent perception. Humans can naturally integrate what they see and hear to understand the surrounding environment. Audio-visual question answering is a task that aims to model this multimodal system for machines. Formally, given a stream of video, audio-visual question answering aims to answer natural language questions by integrating information from both audio and visual modalities, which has been an emerging research topic in recent years. For example, given a video showing that a flock of birds suddenly fly away from trees, and the question "why do these birds fly away?", it requires to combine the visual information "a bird flying away" and the audio information of a train whistle to answer the question as "the whistle sound shocks the bird". To achieve an accurate reasoning process and get the correct answer, it is essential to extract cues and contexts from both audio and visual modalities and discover their inner correlations.

Although most of the video question answering datasets [2, 4, 9, 24, 25, 37, 41, 44, 46] provide access to features of three modalities (audio, visual, and text), few of them have carefully considered the information from the audio modality in the creating process, thus most of the questions in these datasets could be answered with only the visual modality [41, 44]. Due to the severe noise in background sounds, some works [32, 49] discover that audio modality in these datasets provides little information and may be useless or even harmful to the question answering task. Few datasets [26, 45] are specifically proposed for the audio-visual question answering task. Specifically, Pano-AVQA [45] targeting panoramic videos takes audio-visual relationships into account, while those questions are

[*] Corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548291

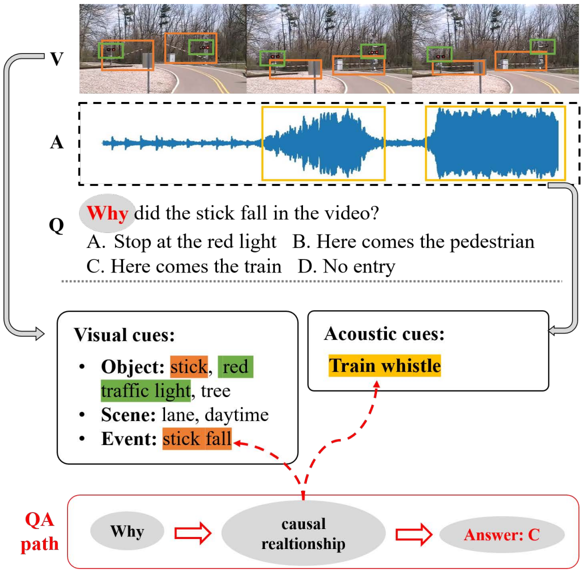

V
A
Q
Why did the stick fall in the video?
A. Stop at the red light B. Here comes the pedestrian
C. Here comes the train D. No entry
Visual cues:
• Object: stick, red traffic light, tree
• Scene: lane, daytime
• Event: stick fall
Acoustic cues:
Train whistle
QA path
Why
causal realtionship
Answer: C

Figure 1: AVQA is an audio-visual question answering dataset for the multimodal understanding of audio-visual objects and activities in real-life scenarios on videos. AVQA provides diverse sets of questions specially designed considering both audio and visual information, involving various relationships between objects or in activities. Here is an example from AVQA dataset and the top two lines show the video clip from both visual frames and audio waveform.

mainly limited to existence or location judgment using spatial attributes, neglecting the function of modeling different types of relationships such as temporal relations [6]. MUSIC-AVQA [26] further regards spatial-temporal association between modalities in the audio-visual question answering task. However, visual scenes are limited to music performance, in which the questions are only about instrument relationships, lacking the exploration of more real-life scenarios.

In short, although most researchers find it urgent to study the problem of audio-visual question answering, the lack of systematic datasets on real-life scene videos considering audio-visual correlations limits the research progress. In this paper, we propose the **AVQA** dataset, the first dataset designed for audio-visual question answering on general videos of real-life scenarios. AVQA is collected from VGG-Sound [3] dataset and selects the suitable classes for question answering tasks and annotates the original data by crowdsourcing. After the collection of the raw data, we post-process samples to improve the quality of the dataset and make it more suitable for training and testing models. Given that videos in AVQA are real-life scenarios, an intelligent system must reach answers with complete modalities. AVQA is designed as a practical yet challenging testbed for audio-visual question answering.

Besides, we propose a Hierarchical Audio-Visual Fusing (HAVF) module to integrate information from audio, visual, and text modalities for question answering. We conduct dense experiments by integrating our HAVF module into state-of-the-art video question answering models and benchmark the audio-visual question answering task on our dataset. Experimental results show significant performance increase and demonstrate the important role of the audio modality in our real-life scene video datasets.

We summarize the contributions of our paper as threefold:

- We introduce AVQA, a new video dataset annotated with specifically-designed questions about real-life audio-visual activities. Activities in the videos have distinctive acoustic characteristics, urgently requiring the capability of integrating multimodal cues and understanding relations among the three modalities.
- We propose a Hierarchical Audio-Visual Fusing (HAVF) module that can be flexibly combined with existing video question answering models.
- We conduct extensive experiments to demonstrate the effectiveness of our method, which significantly improves the accuracy of audio-visual question answering and further demonstrates the superiority of our question design.

2 RELATED WORK

2.1 Related Datasets

Based on whether audio data is accessible, we can classify existing datasets into two categories: datasets with and without access to audio. For those datasets published with audio or with links to raw videos, we can extract audio features or transcribed speech text as an extra input in addition to video frames. For example, **MSRVTT-QA** [41] dataset and **ActivityNet-QA** [44] dataset collect real-life scene data from online videos (like YouTube) and provide links to original videos. With the support of these datasets, researchers can obtain both audio and visual data according to the needs of their models. Besides, datasets [2, 4, 5, 9, 24, 25, 35, 37, 46] generated from movies, tutorial videos or social scenes contain a large number of conversations or instructions, which are usually transcribed as extra text inputs in addition to subtitles and question-answer pairs. Closest to our work are **Pano-AVQA** [45] and **MUSIC-AVQA** [26], which proposed benchmarks for audio-visual question answering on panoramic videos and music performance scenes. Some other datasets are generated from silent animated GIFs [17] and synthetic videos [36, 43] or published without links to raw videos [10, 28], thus audio information is not available and models can only utilize the visual features extracted from them.

Although the first type of dataset contains audio data, most of them are still not suitable for the audio-visual question answering task. Sound utilized in datasets [2, 4, 5, 9, 24, 25, 37, 46] is mainly human speaking and hard to model interactions with visual modality. It is usually actually introduced to models in the form of text [24, 25] rather than natural sounds, which loses the unique characteristics of sounds. In this way, the video question answering task degenerates into textual plot understanding or actor dialogue comprehension problems on these datasets. Datasets [41, 44] are towards understanding spatial and temporal relationships in real-life scenarios. Questions are mostly visual-based while sounds are accompanied by severe noise for the audio-visual mismatching. Several recently-released Video QA datasets [11, 27, 31, 38, 39] focus more on understanding richer relation types and performing more complex reasoning processes, such as multi-step reasoning [5], understanding temporal [39] and causal relationships [27, 31, 39], composition reasoning[11, 38], and commonsense reasoning [27].

Although these new problem settings involve a variety of complex reasoning, none of these problems take into account the participation of audio and its benefits during the inference process. Questions considering audio-visual relationships in [45] are generated from fixed question templates. For example, when given questions like "Who/what is making [sound]?" or "Which sound is [object] making?", it rarely requires models to go beyond recognizing the location and existence of objects and sound. MUSIC-AVQA [26] explores more complex relationships between sounds of the same or different instruments, such as spatial and temporal relative location, counting, and comparison of acoustic characteristics, but is limited in monotonous music performance scenes. Moreover, the data scale of datasets [26, 45] is small, which is inadequate for question answering models training. Therefore, both audio-visual question answering datasets limit the training and evaluation of the model reasoning ability.

Compared with the aforementioned datasets, our proposed AVQA has two distinctive characteristics: (1)questions in AVQA contain a richer set of audio-visual objects and activities in daily life; (2) AVQA has a larger scale number of videos, which benefits models to learn from and understand multiple audio-visual relationships.

2.2 Related Video Question Answering Methods

Recently, there has been significant progress in video question answering. We review the existing approaches as follows: (1) models considering audio modality. Zhuang *et al.* [49] fuse audio and visual modalities at the input stage and feed the multimodal representation to the proposed attention memory unit. Miyanishi *et al.* [32] propose a modulated multi-stream 3D ConvNets, in which three bottlenecks take motion, appearance, and audio modality as inputs, respectively. Yun *et al.* [45] propose a transformer-based model to encode multimodal features. Besides, Li *et al.* [26] introduce spatial and temporal grounding modules to reason spatio-temporal associations between audio and visual modalities under a question query. (2) models without considering audio modality. The research on traditional settings of the video question answering task has mainly proceeded along three different dimensions: encoder-decoder models [17, 42, 48], memory-based models [7, 8, 20, 21], and graph-based models [12, 15, 18, 19].

3 AVQA DATASET

The AVQA dataset centers around understanding various audio-visual relationships in real life. We introduce our dataset with an example in Figure 1, and more detailed examples can be found in supplementary materials.

3.1 Video Collection and Preprocessing

We aim to evaluate the reasoning ability of question answering models in real-life audio-visual scenarios, so the video corpus should have a considerable scale and contain rich and generic classes. Therefore, we choose the audio-visual dataset VGG-Sound [3], which consists of 200k videos for 309 audio classes.

Our dataset is designed for learning objects and activities in daily life using both audio and visual information. More specifically, we expect that videos focus on natural sounds or common human activities, and have the potential to annotate diverse questions based on them. Towards achieving this, we shuffle the initial dataset and remove some videos belonging to monotonous and multiple scene classes. After this process, we select 100k videos of 165 categories for the labeling company to annotate.

3.2 Annotation Process

We build a web application and design a four-step annotation process for creating this dataset. In order to reduce the complexity and achieve better annotation quality, we contact a professional data labeling company and conduct on-site training for 13 annotators. We introduce some tips for attention in the annotating process and give some positive and negative annotation examples to help annotators better understand our idea. Each annotator is asked to log into the web application, get the annotating task assigned by the backend, and complete it. For each question, one annotator only participates in one of these steps.

3.2.1 *Video Evaluation.*

In the first stage, every video is evaluated whether it has enough visual or audio information for proposing questions. The videos with monotonous or static scenes are skimmed. Videos that pass this step enter the next step.

3.2.2 *Annotation Collection.*

In the second stage, annotators are asked to label question-answer pairs. For each video, annotators can design one or a series of related questions. For each question, we ask annotators to write the correct answer to the question as well as the type information (video type, question type, etc.).

3.2.3 *Quality Control.*

Three other annotators check each question to ensure that it is closely related to the video and has the appropriate complexity to be answered. The question cannot enter the next stage if any of the three annotators think the question is incorrect or not complicated enough.

3.2.4 *Choice Completion.*

Questions that pass the quality control stage are completed with three options. Each annotator can write one to three options until the number of them reaches three. The option should have the same text class with the correct answer so that the question is more difficult to be answered. In particular, we think that a good option could be an answer only inferred from visual information or audio information, which is distinguishable. In order to have balanced question distribution and answer distribution, we ask annotators to consider the answers that often appear in other videos as choice candidates.

The raw questions and choices are written in Chinese. We translate them into English by Neural network translation tools from Baidu[1]. In addition, during annotators' working days, we regularly check the questions and answers and give feedback to annotators to ensure the completion quality of the annotating task.

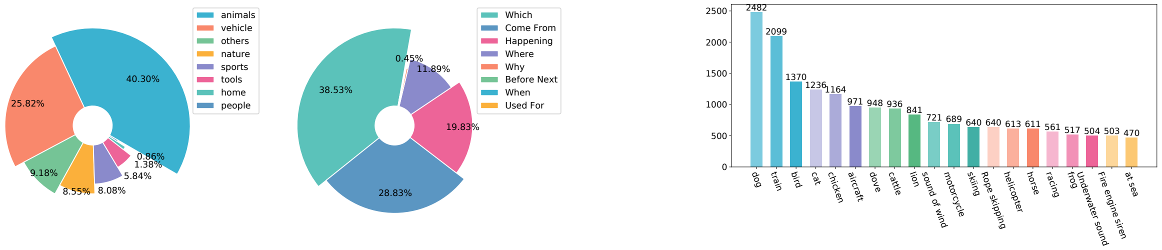
3.3 Data Balancing

Data bias exists widely in all kinds of datasets. In the video question answering task, it is manifested as the different frequency in answer candidates. Machine learning models can reach a relatively good performance only by fitting the dataset biases. To prove this, we train models only to learn from bias on MSVD-QA, MSRVTT-QA, three subsets *Action*, *Transition*, and *FrameQA* of TGIF-QA and

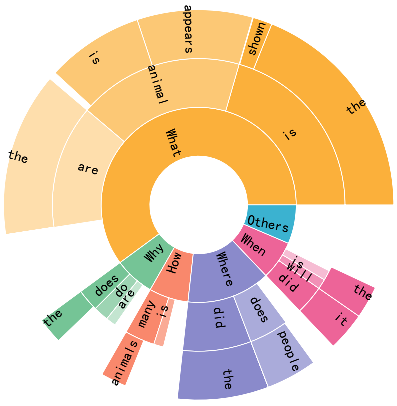
[1] https://api.fanyi.baidu.com/

Table 1: Comparison with other related video datasets. Our AVQA dataset aims to reason about multiple audio-visual relationships in real-life scenarios. (B-Background sound, S-Human speech, O-Object sound.)

Dataset	Questions for audio	Sound type	Visual scene type	Audio-visual relationship type						Video number	QA pairs number
				Existential	Location	Counting	Temporal	Causal	Purpose		
MSRVTT-QA [41]	×	B	Real-life	×	×	×	×	×	×	10K	243.7K
ActivityNet-QA [44]	×	B	Real-life	×	×	×	×	×	×	5.8K	58K
MovieQA [37]	×	S	Movie	×	×	×	×	×	×	6.8K	6.5K
TVQA [24]	×	S	Movie	×	×	×	×	×	×	21.8K	152.5K
TVQA+ [25]	×	S	Movie	×	×	×	×	×	×	4.2K	29.4K
DramaQA [4]	×	S	Movie	×	×	×	×	×	×	23.9K	18.0K
LifeQA [2]	×	S	Real-life	×	×	×	×	×	×	0.3K	2.3K
KnowIT VQA [9]	×	S	Movie	×	×	×	×	×	×	12.1K	24.3K
Social-IQ [46]	×	S	Social	×	×	×	×	×	×	1.3K	7.5K
Pano-AVQA [45]	✓	O	Panoramic	✓	✓	×	×	×	×	5.4K	51.7K
MUSIC-AVQA [26]	✓	O	Music	✓	✓	✓	✓	×	×	9.3k	45.9K
AVQA (Ours)	✓	O	Real-life	✓	✓	✓	✓	✓	✓	**57.0K**	57.3K



(a) Distribution of audio categories. (b) Distribution of question types. (c) Distribution of top-20 frequent answers.



Question Type	Top-5 frequent candidates
Which	dog, bird, cat, chicken, cattle
Come From	train, aircraft, sound of wind, motorcycle, helicopter
Happening	rope skipping, skiing, rowing, riding, machine gun fire
Where	at sea, field, eabed, highway, aquatic
Why	I'm hungry, decompression, roller coaster ride, frightened, motorcyle
Before Next	volcanic explosion, setting off fireworks, tornado, sharpen the knife, set off firecrackers
When	evening, chicken, train, lion, sound of wind
Used For	decompression, train, dog, turkey, protect your eyes

(d) Distribution of questions by their first three words. (e) Top-5 frequent candidates per qustion type.

Figure 2: Illustrations of AVQA dataset statistics. (a) Distribution of audio categories. (b-e) Distributions of collected question-answer pairs.

our proposed dataset. We then define a new metric, Unbalancing Factor (UF):

$$UF = \frac{Acc_r}{Acc_t} - 1 \tag{1}$$

where Acc_r is the accuracy of answers in reality, and Acc_t is the theoretical accuracy, *i.e.*, the probability that each answer is selected (the inverse of candidate number). UF evaluates the degree of data bias, and it decreases to 0 when there is no bias in the dataset.

Table 2: Data bias evaluation on different datasets. (OE-Open ended, MC-Multiple choice)

Dataset	Question type	Candidate number	Acc_t	Acc_r	UF
MSVD-QA [41]	OE	1.9K	0.054%	16.71%	309.4
MSRVTT-QA [41]	OE	6.2K	0.016%	9.95%	621.9
TGIF-QA_FrameQA [17]	OE	1.5K	0.065%	25.69%	395.2
MUSIC-AVQA [26]	OE	94	1.064%	18.96%	17.82
TGIF-QA_Action [17]	MC	5	20%	46.44%	2.322
TGIF-QA_Transition [17]	MC	5	20%	50.10%	2.505
AVQA(Ours initial)	MC	4	25%	72.37%	2.895
AVQA(Ours final)	MC	4	25%	41.79%	**1.672**

Table 3: Notations in Algorithm 1.

Notation	Role
A	All annotations
C_list	Unique candidate matrix, 4 columns are [$Candidates$, ANS_count, CAN_count, U_count]
$Candidates$	Text of answer candidates
ANS_count	Count number that candidates appear in correct options
CAN_count	Count number that candidates appear in confusion options
U_count	Unbalancing count = $ANS\_count - CAN\_count$
S_dict	Dictionary of candidate synonyms
B_A	Balanced annotations
th_A	Threshold for ANS_count
th_U	Threshold for CAN_count
U_list	Unbalancing sorted candidate matrix
B_items	Items need to be balanced
[]	Operations that extract some rows from the matrix
['']	Operations that get certain column from the matrix

```
Algorithm 1: CCR balancing algorithm
Input      : $A, C\_list, S\_dict$
Output     : $B\_A$
Parameters : $th_A, th_U$
1  $U\_list$ = sort( $C\_list$, key=$U\_count$)
2  $B\_items$ = $U\_list$[ $ANS\_count > th_A$ and $U\_count > th_U$]
3  for each $A_i$ in $A$ do
4      for item in $A_i$['multi-choice'] except $A_i$['answer'] do
5          if item in $B\_items$['candidates'] then
6              $B\_index$ = item index in $B\_items$['candidates']
7              $item\_B\_count$ = $B\_items[B\_index]$['$U\_count$']
8              for $S\_item$ in $S\_dict[item]$ do
9                  $S\_index$ = index of $S\_item$ in $U\_list$
10                 $item\_S\_count$ =
                     $U\_list[S\_index]$['$U\_count$']
11                 if $item\_B\_count < -item\_S\_count$ then
12                     $item \leftarrow S\_item$
13                     $B\_items[B\_index]$['$U\_count$'] -= 1
14                     $U\_list[S\_index]$['$U\_count$'] += 1
15                     break
16                 end
17             end
18         end
19     end
20 end
21 $B\_A$ = shuffle($A$)
22 return $B\_A$
```

We set all visual and question input features to zero vectors and let models only learn from data bias in answer distribution. The results are reported in Table 2. We can see that most of the datasets have great data bias. We find that our initial dataset has the largest data bias among datasets of multiple choice. Although our fully manual annotation process ensures that the questions and answer candidates are reasonable, we still introduce large annotator bias into our dataset. We then propose a data balancing algorithm called **Confusion Candidates Replacing (CCR)** multi-choice video question answering dataset balancing algorithm to decrease our dataset's data bias while keeping the advantage of reasonable question-answer pairs. The operation of CCR is presented algorithmically in Algorithm 1. Table 3 summarizes the notations used across these presentations.

The CCR balancing algorithm first sorts C_list according to the descending order of the fourth column U_count. Then we filter candidates whose ANS_count is more than threshold th_A, and U_count is more than threshold th_U (line 1-2 of Algorithm 1). We consider the number of U_count to select seriously unbalanced answer candidates and the number of ANS_count to ensure that data bias will decrease after balancing this candidate. Then, for each annotation A_i in A, confusion options of A_i in the column $Candidates$ of B_items means items that need to be balanced. If so, we then get the unbalancing count from the column U_count of B_items. We traverse the synonym list of this confusion option through S_dict and get their unbalancing count from the column U_count of U_list. Suppose there is one synonym whose unbalancing count is negative and absolute value is more than unbalancing count of the candidate itself. In that case, we can use this synonym to replace this candidate for balancing. So here we replace it and change the unbalancing count of each other.

After doing this one by one, we manually balance some items that the CCR balancing algorithm cannot solve. This process is simple because the algorithm has balanced most candidates. Finally, we find 363 items that are hard to be balanced, so we removed them from our dataset. From the results of Table 2, our dataset reaches the least data bias. Models trained on our dataset will have more ability learned from video-text-audio contextual representing to reason rather than more ability from data bias. Finally, we split the AVQA dataset by randomly selecting 70% of the samples as the training set and the rest as the validation set.

3.4 Dataset Exploration

3.4.1 Dataset Statistics. Our AVQA dataset contains 57,335 question-answer pairs and 57,015 videos for over 158 hours. We compare AVQA with existing video question answering datasets that provide audio modality access in Table. 1. Note that our AVQA provides specially designed questions about a wide variety of audio-visual relationships in real-life scenes and is of the largest scale on video numbers among these datasets. Video clips in our dataset involve eight categories of audio-visual objects and activities (see Figure 2(a)). We classify questions into eight categories based on semantics, and we visualize the distribution of question types and their first three words in Figure 2(b) and (d). Figure 2(c) and (d) show the distribution of top-20 frequent answers for the overall dataset and top-5 frequent candidates for each question type.

3.4.2 Dataset Highlights. We summarize the highlights of AVQA as follows:1) **Large-scale.** AVQA has the most videos with little noise compared to related video datasets. Besides, AVQA outperforms two other audio-visual question answering datasets in both the number of videos and the number of question-answer pairs. Meanwhile, AVQA has a broader scene scope than panoramic videos and music performance videos. A large number of videos bring a wider variety of life scenarios, which is more suitable for model training and evaluation in multimodal reasoning ability. 2) **High quality.** Our annotations are of high quality, especially compared with datasets entirely generated by NLP tools like MSVD-QA [41] and MSRVTT-QA [41]. Even though they may have more QA pairs, due to the limitation of inaccurate annotations, the current highest accuracy in both datasets is around 40%. The fully manual annotation also makes AVQA superior to semi-manual annotated datasets like TGIF-QA [17]. This is because our confusion options are more relevant to the videos, bringing models better fine-grained discrimination ability. 3) **Novelty and broader applicability.** There are a considerable number of questions in AVQA that can only be answered from both audio and visual information. Therefore, models trained on our dataset need to integrate the three modalities for reasoning, which is closer to the way humans get external information in real life. 4) **Small data Bias.** From Table 2, we can see that the data bias in our dataset becomes the least after the CCR algorithm, so that models are more likely to learn natural relations rather than data bias.

4 AUDIO-VISUAL QUESTION ANSWERING ON AVQA

4.1 Formulation

The basic video question model could be summarized as:

$$a = \arg\max_{a\in\mathcal{A}} \Pr(a|V,Q) = p_\phi\left(a|\alpha_\theta(\mathbf{h}_v,\mathbf{h}_q);\mathcal{A}\right), \qquad (2)$$

where $\mathbf{h}_v \in \mathbb{R}^{d_v}$, $\mathbf{h}_q \in \mathbb{R}^{d_q}$ is the visual feature and question feature, respectively, which could be extracted with corresponding deep neural network models [13, 14, 40]. α_θ is a feature fusion model used to fuse features from both question and video modalities, and $p_\phi(a|\cdot;\mathcal{A})$ is the answer decoder probability model to decode the final answer a from the answer set $\mathcal{A}$, with learnable weights θ and ϕ, respectively.

However, when it comes to the more challenging audio-visual question answering problem, Eq. 2 becomes more complicated that we have to model the following probability considering all three modalities:

$$a = \arg\max_{a\in\mathcal{A}} \Pr(a|V,A,Q) \qquad (3)$$

Based on current two-modalities fusing techniques, we could firstly fuse any two modalities and then fuse the results with the third one, or we could directly fuse all the three modalities directly. In our preliminary experiments, we find some modalities contribute little to the final answer. For example, the fusion of V and A theoretically does not correlate with the final answer a if the dataset is fully balanced. In the following, we introduce three important fusing methods, which could be seen as variants of Eq. 2, and we show how these variants finally contribute to Eq. 3.

Early Audio-Visual Fusion. Early Audio-Visual Fusion (EAVF) could be formularized as follows:

$$\Pr^{(e)}(a|V,A,Q) = p_\phi^{(e)}\left(a|\alpha_\theta^{(q-av)}(\alpha_\theta^{(av)}(\mathbf{h}_v,\mathbf{h}_a),\mathbf{h}_q));\mathcal{A}\right), \qquad (4)$$

where $\alpha_\theta^{(av)}$ is the audio-visual fusion model and their output is further fused with the question using $\alpha_\theta^{(q-av)}$ to generate the final answer probability. This fusion method could be explained as: regarding audio and visual features as two complementary modalities, and their fused result forms a full video feature, and the final answer could be reasoned out from the question feature and the full video feature.

Middle All Fusion. Middle all Fusion (MF) could be formularized as:

$$\Pr^{(m)}(a|V,A,Q) = p_\phi^{(m)}\left(a|\alpha_\theta^{(qav)}(\mathbf{h}_v,\mathbf{h}_a,\mathbf{h}_q);\mathcal{A}\right), \qquad (5)$$

where α_θ^{qav} is a comprehensive fusion model to fuse information from all three modalities. This fusion method treats all three modalities equally and fuses them all.

Late Audio Fusion. Late Audio Fusion (LAF) could be formularized as follows:

$$\Pr^{(l)}(a|V,A,Q) = p_\phi^{(l)}\left(a|\alpha_\theta^{(a-qv)}(\alpha_\theta^{(qv)}(\mathbf{h}_v,\mathbf{h}_q),\mathbf{h}_a));\mathcal{A}\right), \qquad (6)$$

where $\alpha_\theta^{(qv)}$ is the basic video-question fusion model as used in most existing works, while $\alpha^{(a-qv)}$ fuses the audio feature into the results of $\alpha_\theta^{(qv)}$. This fusion method treats the audio modality as a supplementary modality that provide extra information to answer the question.

Generally, both of the three fusions methods sound reasonable and have their advantages in different question types and baseline models (See Exp. 5.2 for more details). In this paper, we try to combine them all to form a more powerful and more complement model with Hierarchical Audio-Visual Fusing module (HAVF):

$$\Pr(a|V,A,Q) = \mathcal{H}\left(\Pr^{(e)},\Pr^{(m)},\Pr^{(l)}\right)(a|V,A,Q), \qquad (7)$$

where $\mathcal{H}(\cdots)$ takes the prediction of $Pr^{(e)}$,$Pr^{(m)}$,$Pr^{(l)}$ as input, and normalize the result as output probability.

4.2 Implementation

In this section, we briefly introduce the implementation details of the model, including the feature extraction model, feature fusion model, answer model and the hierarchical ensemble model.

Feature Extraction. For audio features, we choose a pre-trained audio tagging model [22] and take the output of the upper layer of the sigmoid layer as the audio embedding. For visual features, we follow previous work [23] and take a consistent approach [13, 14, 40] to extract appearance and motion features separately. For text features, we use a pre-trained GloVe [33] to transfer each word to a 300-dimensional feature vector and use an LSTM encoder to obtain text representation for questions and candidates.

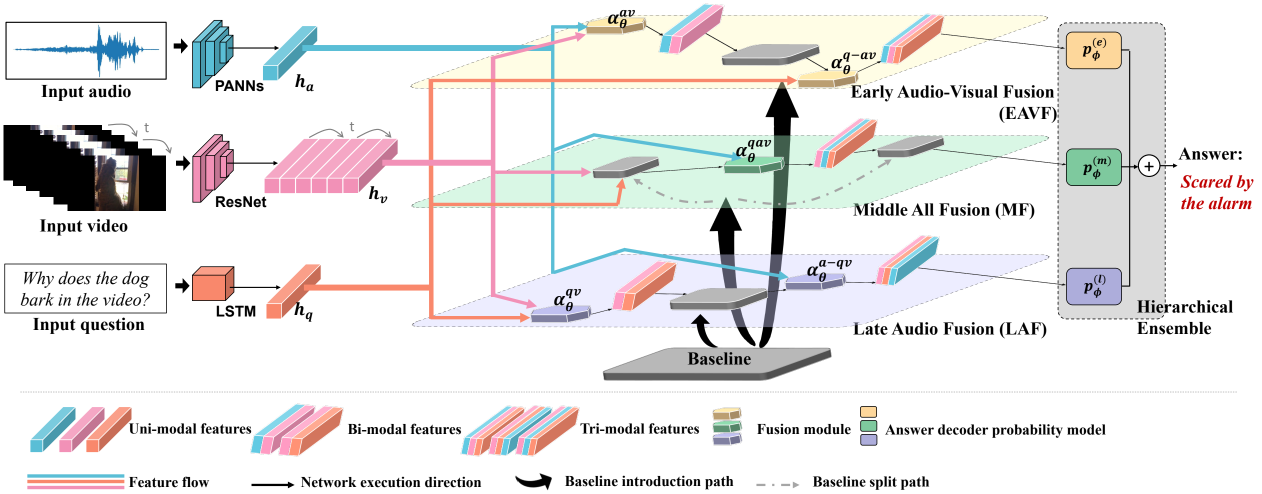


Figure 3: Our Hierarchical Audio-Visual Fusing module. We take pre-trained PANNs [22] and ResNet [14] to extract audio features and visual features and obtain the question embedding through an LSTM model for the feature extraction stage. Then, we combine three fusion modules (EAVF, MF, LAF) with the baseline model to fuse features generated from audio, visual, and text modalities. Finally, the Hierarchical Ensemble module combines the advantages of three fusing methods and uses integrated tri-modal information to predict the answer to the input question.

Table 4: Results of preliminary experiments on HCRN

Fusion Strategy	vq (basic)	$v-aq$	$q-av$ (Eq. 4)	$a-qv$ (Eq. 6)	qav (Eq. 5)
Total accuracy	82.5	82.6	86.5	87.1	**87.8**

Feature Fusion. For each variants of fusion methods, we try four strategies for different baseline models: (1) concatenation (2) element-wise addition (3) element-wise multiplication (4) modified conditional relation unit [23].

Answer Model and Hierarchical Ensemble. We take an averaging strategy to ensemble three fusing methods. After averaging the outputs of the answer model, we finally choose the candidate with the highest averaged prediction score as the answer.

Training Details. We follow the same hyperparameter settings as in the baseline papers [7, 18, 23, 29, 30, 47] and list more details in the supplementary materials.

5 EXPERIMENTS

5.1 Experimental Setup

We evaluate our dataset and our Hierarchical Audio-Video Fusing (HAVF) module with six well-known and state-of-the-art video question answering models.

We follow the feature fusion strategy as in the original model to fuse features from the audio modality (feature concatenation, feature elementary-wise operation, or attention). We also find that feature concatenation works better for the EAVF (Eq. 4) fusion, and element-wise multiplication works better for MF (Eq. 5) and LAF (Eq. 6) fusion. We basically introduce how we integrate our HAVF into these models here and put more implementation and experiment details in the supplementary materials.

HME [7]: we use an elementary-wise multiplication strategy for EAVF, MF, and LAVF fusion.
PSAC [30]: We choose concatenation for EAVF, and choose element-wise multiplication for both MF and LAF.
LADNet [29]: the same strategy as PSAC.
ACRTransformer [47]: the same strategy as PSAC.
HGA [18]: we use a concatenation strategy for EAVF, MF, and LAF.
HCRN [23]: We choose concatenation, modified conditional relation unit, and element-wise multiplication for EAVF, MF, and LAF, repectively.
For each model, we then use our proposed HAVF to obtain a final result.

5.2 Preliminary Experiments

There are theoretically four choices to fuse three modalities without considering relative order, *i.e.*, considering fusing any two of the three modalities first (audio, visual, and question), and then fusing the results with the left one or directly fusing all the three modalities. In this section, we briefly summarize our experimental results of these fusing methods, namely: $v-aq$, $q-av$, $a-qv$, qav fusions, and the baseline fusion vq.

As shown in Table 4, the $v-aq$ fusion brings almost no improvement compared with the no audio baseline, which is discarded in our further study. Except for $v-aq$, all the other variants could get 4% to 5% absolute accuracy improvement, which demonstrates the effectiveness of audio. Therefore, we choose the other three variants to form our final Hierarchical Audio-Visual Fusing module.

5.3 Experimental Results and Analysis

The performance for each model on the AVQA validation set is shown in Figure. 4 and Table. 5. Performance is reported for the baseline (without audio modality) model, baseline+EAVF, baseline+MF, baseline+LAF, and baseline+HAVF.

Table 5: Fine-grained Testing Performance(%) of baselines and our proposed Baseline-HAVF method. The best performance over each question type is highlighted in bold form. The largest performance increase caused by our HAVF module over each question type is highlighted in underline form.

Methods	Which	Come From	Happening	Where	Why	Before Next	When	Used For	Total Accuracy
HME [7]	82.2	85.9	79.3	76.6	57.0	80.0	57.1	76.5	81.8
HME+HAVF	85.6 (+3.4)	88.3 (+2.4)	83.1 (+3.8)	83.5 (+6.9)	61.6 (+4.6)	80.0 (+0.0)	**57.1** (+0.0)	**88.2** (+11.7)	85.0 (+3.2)
PSAC [30]	78.7	80.0	77.0	79.4	44.2	76.0	42.9	58.8	78.6
PSAC+HAVF	89.0 (+10.3)	91.1 (+11.1)	83.2 (+6.2)	81.7 (+2.3)	61.6 (+17.4)	**82.0** (+6.0)	52.4 (+9.5)	76.5 (+17.7)	87.4 (+8.8)
LADNet [29]	81.1	87.1	76.6	81.8	67.4	78.0	47.6	76.5	81.9
LADNet+HAVF	84.2 (+3.1)	89.0 (+1.9)	79.1 (+2.5)	81.4 (-0.4)	**68.6** (+1.2)	82.0 (+4.0)	52.4 (+4.8)	76.5 (+0.0)	84.1 (+2.2)
ACRTransformer [47]	82.5	82.8	79.4	82.5	54.7	80.0	47.6	58.8	81.7
ACRTransformer+HAVF	88.5(+6.0)	91.7(+8.9)	83.9(+4.5)	**84.9**(+2.4)	50.0(-4.7)	**82.0**(+2.0)	**57.1**(+9.5)	64.7(+5.9)	87.8(+6.1)
HGA [18]	82.1	84.3	79.5	83.1	59.3	82.0	57.1	88.2	82.2
HGA+HAVF	88.6 (+6.5)	92.2 (+7.9)	83.8 (+4.3)	82.6 (-0.5)	61.6 (+2.3)	78.0 (-4.0)	52.4 (-4.7)	82.4 (-5.8)	87.7 (+5.5)
HCRN [23]	83.7	84.1	80.2	80.9	52.3	74.0	57.1	70.6	82.5
HCRN+HAVF	**89.8** (+6.1)	**92.8** (+8.7)	**86.0** (+5.8)	84.4 (+3.5)	57.0 (+4.7)	80.0 (+6.0)	52.4 (-4.7)	82.4 (+11.8)	**89.0** (+6.5)

Table 6: Experimental results(%) of different input modality combinations for HCRN-MF.

Modalities	Accuracy
Question	48.0
Audio	83.2
Visual	80.8
Audio+Question	83.5
Visual+Question	81.6
Visual+Audio+Question	**87.8**

Overall Performance. In Figure 4, we show the overall performance of different fusing methods, from which we could find that different baseline models prefer different fusing methods: EAVF works better for HME, PSAC, while MF works better for ACRTransformer, while LAF works for LADNet. But among all the fusing methods and all models, our proposed HAVF works the best.

Performance across Different Question Categories. In Table 5, we further show the performance of all models with/without our HAVF module on all fine-grained question categories. For almost all categories, our HAVF could bring more than 5% absolute accuracy increase, which verifies the importance of audio modality on one hand and demonstrate the effectiveness of our proposed HAVF module on the other hand.

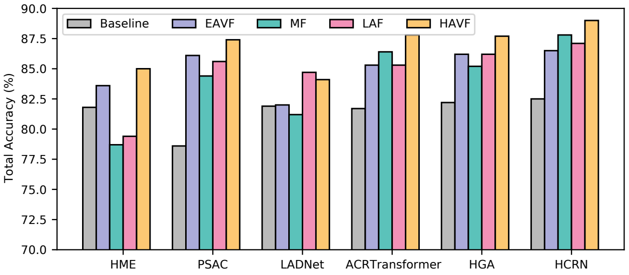


Figure 4: Overall testing accuracy(%) of variants for the proposed Hierarchical Audio-Visual Fusing module.

The Role of Multimodal Signals. In Table 6, we validate the effectiveness of different modality combinations. Audio is the most representative modality in AVQA and a little more efficient than visual features, which differs from the findings in existing video question answering datasets. The best result shown in the model with full modality inputs further demonstrates the rationality and efficiency of our question design settings of our AVQA dataset.

6 CONCLUSION AND FUTURE WORK

In this paper, we introduce a challenging dataset AVQA, which is a large-scale real-life dataset annotated with questions considering different audio-visual relationships. We present building process and propose a Hierarchical Audio-Visual Fusing module to tackle the challenges of audio-visual question answering task in real-life scenarios. We conduct extensive experiments to show the effectiveness of our model and give in-depth analysis. In addition, we identify two critical challenges that we believe are essential to be addressed in future researches. 1) **Temporal Relation Modeling.** In this work, we use a video-level audio embedding, which erases the temporal property of audio modality. Therefore more advanced model that could capture complex temporal associations between two modalities is expected to further improve performance in the temporal aspect. 2) **Explainable Framework.** If we can make the reasoning process explainable, it will help models be more reliable and flexible to be aggregated in applications in everyday life. We believe that the AVQA dataset has the potential to promote the community of audio-visual question answering and empower high multimodal understanding ability in models.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China No. 2018AAA0102001 and National Natural Science Foundation of China No. 62250008, No. 62102222.

REFERENCES

[1] David A Bulkin and Jennifer M Groh. 2006. Seeing sounds: visual and auditory interactions in the brain. *Current opinion in neurobiology* 16, 4 (2006), 415–419.

[2] Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. 2020. LifeQA: A Real-life Dataset for Video Question Answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4352–4358. https://aclanthology.org/2020.lrec-1.536

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 721–725.

[4] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. 2020. Dramaqa: Character-centered video story understanding with hierarchical qa. *arXiv preprint arXiv:2005.03356* (2020).

[5] Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. 2019. TutorialVQA: Question answering dataset for tutorial videos. *arXiv preprint arXiv:1912.01046* (2019).

[6] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems* 31 (2018).

[7] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1999–2007.

[8] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6576–6585.

[9] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. KnowIT VQA: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10826–10834.

[10] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11287–11297.

[11] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11287–11297.

[12] Mao Gu, Zhou Zhao, Weike Jin, Richang Hong, and Fei Wu. 2021. Graph-Based Multi-Interaction Network for Video Question Answering. *IEEE Transactions on Image Processing* 30 (2021), 2758–2770.

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6546–6555.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[15] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11021–11028.

[16] Robert A Jacobs and Chenliang Xu. 2019. Can multisensory training aid visual learning? A computational investigation. *Journal of vision* 19, 11 (2019), 1–1.

[17] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2758–2766.

[18] Pin Jiang and Yahong Han. 2020. Reasoning with Heterogeneous Graph Alignment for Video Question Answering. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. 11109–11116.

[19] Weike Jin, Zhou Zhao, Xiaochun Cao, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. 2021. Adaptive Spatio-Temporal Graph Enhanced Vision-Language Representation for Video QA. *IEEE Transactions on Image Processing* 30 (2021), 5477–5489.

[20] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. 2019. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8337–8346.

[21] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836* (2017).

[22] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2019. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. (2019). http://arxiv.org/abs/1912.10211

[23] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical Conditional Relation Networks for Video Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9969–9978.

[24] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*.

[25] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. TVQA+: Spatio-Temporal Grounding for Video Question Answering. In *Tech Report, arXiv*.

[26] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. *arXiv preprint arXiv:2203.14072* (2022).

[27] Jiangtong Li, Li Niu, and Liqing Zhang. 2022. From Representation to Reasoning: Towards both Evidence and Commonsense Reasoning for Video Question-Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21273–21282.

[28] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200* (2020).

[29] Xiangpeng Li, Lianli Gao, Xuanhan Wang, Wu Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song. 2019. Learnable Aggregating Net with Diversity Learning for Video Question Answering. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. 1166–1174.

[30] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 8658–8665.

[31] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2928–2937.

[32] Taiki Miyanishi and Motoaki Kawanabe. 2021. Watch, Listen, and Answer: Open-Ended VideoQA with Modulated Multi-Stream 3D ConvNets. In *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 706–710.

[33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[34] Ladan Shams and Aaron R Seitz. 2008. Benefits of multisensory learning. *Trends in cognitive sciences* 12, 11 (2008), 411–417.

[35] Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*. 493–502.

[36] SVQA-founder. 2018. Synthetic Video Question Answering. https://github.com/SVQA-founder/SVQA.

[37] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4631–4640.

[38] 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)*.

[39] 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)*. 9777–9786.

[40] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5987–5995.

[41] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*. 1645–1653.

[42] Hongyang Xue, Zhou Zhao, and Deng Cai. 2017. Unifying the video and question attentions for open-ended video question answering. *IEEE Transactions on Image Processing* 26, 12 (2017), 5656–5666.

[43] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442* (2019).

[44] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9127–9134.

[45] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. 2021. Pano-AVQA: Grounded Audio-Visual Question Answering on 360deg Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2031–2041.

[46] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8807–8817.

[47] Jipeng Zhang, Jie Shao, Rui Cao, Lianli Gao, Xing Xu, and Heng Tao Shen. 2020. Action-Centric Relation Transformer Network for Video Question Answering. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[48] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision* 124, 3 (2017), 409–421.

[49] Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. 2020. Multichannel attention refinement for video question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 1s (2020), 1–23.

A DATASET

In this section, we introduce the details of AVQA dataset construction process and show more examples from AVQA.

A.1 Video Filtering

We select 165 subcategories from the 8 audio categories of the original VGG-Sound dataset for labeling question-answer pairs. The conditions that the video clip passes the video filtering process are: (1) The video clip contains common audio-visual objects or activities in real-life scenarios. (2) The content of the video clip is informative enough to annotate question-answer pairs. (3) The background of the video clip changes dynamically over time, which is not static or monotonous.

A.2 QA Annotation

In order to ensure a high annotation quality, we design a four-stage annotation process. We train annotators to check and evaluate video content richness, question complexity, and completeness of options in different annotation stages. The flow chart of the QA annotation process is shown in Figure 5.

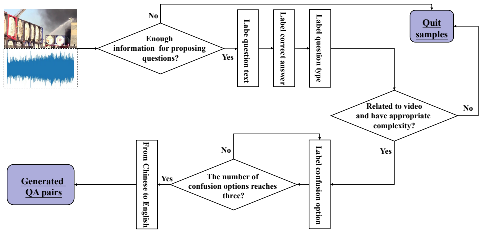


Figure 5: QA annotation process.

B MODEL IMPLEMENTATION AND TRAINING DETAILS

We describe the model implementation and training details in this section. The code is implemented based on the Pytorch library[2].

B.1 Implementation Details

To better introduce how we integrate our HAVF module to different SOTA video question answering baseline models, we show model implementation details of all fusing methods in Table. 9. We finally choose the optimal fusing strategies for each of the EAVF, MF, and LAF methods to form our HAVF module.

B.2 Hyperparameter Settings

We follow the same hyperparameter settings as the official code version of each baseline method to train and evaluate our HAVF modules, and details are listed in Table. 7.

C DETAILED EXPERIMENTAL RESULTS

We conduct extra calculations and statics for each baseline model on model complexity and efficiency. We evaluate the model accuracy and model inference time (on a TITAN XP GPU) and show the results in Table. 8 for illustration. The inference time includes: (1) a shared feature extraction time (ResNet 101 for video feature and PANNs for audio feature), which is shared by all models (about 220 ms); (2) per-branch inference, which depends on the backbone Video QA model complexity and fusion strategy.

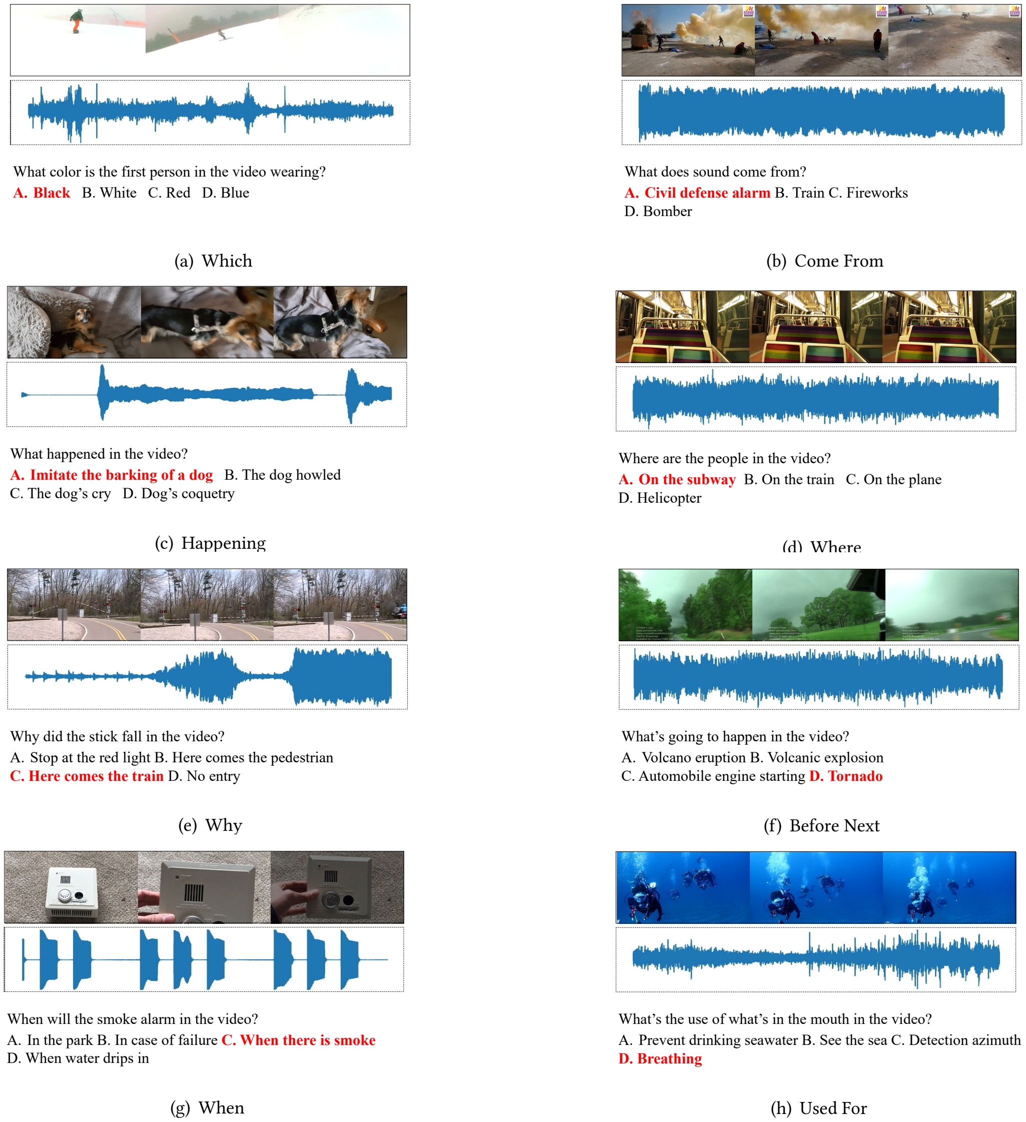
Table 7: Hyperparameter settings for each baseline model

Baseline method	Hyperparameter settings
HME	hidden_size = 512, num_layers = 2, num_epochs = 1000, batch_size = 32, num_workers = 4, learning_rate = 0.001, momentum = 0.9, weight_decay = 1e-4
PSAC	epochs = 100, num_hid = 512, max_len = 20, char_max_len = 15, num_frame = 36, batch_size = 32, seed = 1000, vid_enc_layers = 1
LADNet	epochs = 50, num_hid = 1024, max_len = 36, batch_size = 64, seed = 1000, scale = [256, 512, 1024], reasonSteps = 1, sub_nums = 8, lambda = 0.01
ACRTransformer	SENTENCE_LEN = 12, MIN_OCC = 1, NUM_HIDDEN = 3072, NUM_LAYER = 1, EMB_DROPOUT = 0.3495, FC_DROPOUT = 0.3495, L_RNN_DROPOUT = 0.3495, ANSWER_LEN = 16, LEARNING_RATE = 0.0095, MOMENTUM = 0.9, topk = 35, BATCH_SIZE = 128, NUM_GLIMPSE = 2, POOLING_SIZE = 5, C3D_SIZE:1024, RES_SIZE = 2048, VIDEO_LEN = 35, MID_DIM = 1024, SEED = 1111, EPOCHS = 30, NUM_PROPOSAL = 6, warmup = 2000
HGA	num_workers = 2, batch_size = 64, lr = 0.0001, dropout = 0.3, hidden_size = 512, max_epoch = 50, momentum = 0.9, q_max_length = 35, v_max_length = 80, rnn_layers = 1, birnn = 0, gcn_layers = 2, tf_layers = 1, max_n_videos = 100000, lr_list = [10, 20, 30, 40], cycle_beta = 0.01, two_loss = 0, weight_decay = 0
HCRN	lr = 0.0001, batch_size = 32, max_epochs = 50, vision_dim = 2048, audio_dim = 2048, word_dim = 300, module_dim = 512, k_max_frame_level = 16, k_max_clip_level = 8, spl_resolution = 1

Table 8: Comparison of Parameter Numbers and Inference Time between all baselines and baseline+HAVF models.

Method	Total Accuracy	Parameters	Inference Time
HME	81.8	165.01M	398.363ms
HME+EAVF	83.6	166.06M	403.087ms
HME+MF	78.7	167.11M	401.868ms
HME+LAF	79.4	167.11M	404.205ms
HME+HAVF	85.0	247.51M	761.215ms
PSAC	78.6	166.99M	227.993ms
PSAC+EAVF	86.1	195.94M	228.841ms
PSAC+MF	84.4	179.58M	228.098ms
PSAC+LAF	85.6	176.43M	228.227ms
PSAC+HAVF	87.4	299.17M	240.691ms
LADNet	81.9	144.25M	228.336ms
LADNet+EAVF	82.0	150.04M	229.525ms
LADNet+MF	81.2	144.25M	228.322ms
LADNet+LAF	84.7	146.35M	228.44ms
LADNet+HAVF	84.1	187.86M	241.403ms
ACRTransformer	81.7	377.62M	252.345ms
ACRTransformer+EAVF	85.3	409.08M	252.488ms
ACRTransformer+MF	86.4	383.91M	252.799ms
ACRTransformer+LAF	85.3	383.91M	252.343ms
ACRTransformer+HAVF	87.8	924.13M	313.29ms
HGA	82.2	244.88M	277.64ms
HGA+EAVF	84.4	267.43M	276.775ms
HGA+MF	85.2	260.61M	277.21ms
HGA+LAF	86.2	260.61M	276.812ms
HGA+HAVF	87.7	535.87M	388.964ms
HCRN	82.5	169.8M	398.925ms
HCRN+EAVF	86.5	173.21M	395.778ms
HCRN+MF	87.8	184.75M	391.616ms
HCRN+LAF	87.1	181.34M	393.908ms
HCRN+HAVF	89.0	286.53M	748.072ms

[2] https://pytorch.org/



What color is the first person in the video wearing?
A. Black B. White C. Red D. Blue

(a) Which

What does sound come from?
A. Civil defense alarm B. Train C. Fireworks
D. Bomber

(b) Come From

What happened in the video?
A. Imitate the barking of a dog B. The dog howled
C. The dog's cry D. Dog's coquetry

(c) Happening

Where are the people in the video?
A. On the subway B. On the train C. On the plane
D. Helicopter

(d) Where

Why did the stick fall in the video?
A. Stop at the red light B. Here comes the pedestrian
C. Here comes the train D. No entry

(e) Why

What's going to happen in the video?
A. Volcano eruption B. Volcanic explosion
C. Automobile engine starting **D. Tornado**

(f) Before Next

When will the smoke alarm in the video?
A. In the park B. In case of failure **C. When there is smoke**
D. When water drips in

(g) When

What's the use of what's in the mouth in the video?
A. Prevent drinking seawater B. See the sea C. Detection azimuth
D. Breathing

(h) Used For

Figure 6: Examples of different question types from AVQA.

Table 9: Implementation details of different fusing methods for each baseline model. (EME—EAVF&MF ensemble, MLE—MF&LAF ensemble, ELE—EAVF&LAF ensemble, add—element-wise addition, mul—element-wise multiplication, concat—concatenation.)

Baseline models	EAVF	MF	LAF	EME	MLE	ELE	HAVF
HME	add mul concat	mul concat	mul concat	EAVF (mul) + MF (mul)	MF (mul) + LAF (mul)	EAVF (mul) + LAF (mul)	EAVF (mul) + MF (mul) + LAF (mul)
PSAC	mul concat	add mul concat	add mul concat	EAVF (concat) + MF (mul)	MF (mul) + LAF (mul)	EAVF (concat) + LAF (mul)	EAVF (concat) + MF (mul) + LAF (mul)
LADNet	add mul concat	mul concat	add mul concat	EAVF (concat) + MF (mul)	MF (mul) + LAF (mul)	EAVF (concat) + LAF (mul)	EAVF (concat) + MF (mul) + LAF (mul)
ACRTransformer	add mul concat	add mul concat	add mul	EAVF (concat) + MF (mul)	MF (mul) + LAF (mul)	EAVF (concat) + LAF (mul)	EAVF (concat) + MF (mul) + LAF (mul)
HGA	add mul concat	concat	mul concat	EAVF (concat) + MF (concat)	MF (concat) + LAF (concat)	EAVF (concat) + LAF (concat)	EAVF (concat) + MF (concat) + LAF (concat)
HCRN	concat	ccrn vcrn ccrn + vcrn	mul	EAVF (concat) + MF (vcrn)	MF (vcrn) + LAF (mul)	EAVF (concat) + LAF (mul)	EAVF (concat) + MF (vcrn) + LAF (mul)