CAD - Contextual Multi-modal Alignment for Dynamic AVQA

Asmar Nadeem[1], Adrian Hilton[1], Robert Dawes[2], Graham Thomas[2], Armin Mustafa[1]
[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom.
[2]BBC Research and Development, United Kingdom.

Abstract

In the context of Audio Visual Question Answering (AVQA) tasks, the audio and visual modalities could be learnt on three levels: 1) Spatial, 2) Temporal, and 3) Semantic. Existing AVQA methods suffer from two major shortcomings; the audio-visual (AV) information passing through the network isn't aligned on Spatial and Temporal levels; and, inter-modal (audio and visual) Semantic information is often not balanced within a context; this results in poor performance. In this paper, we propose a novel end-to-end Contextual Multi-modal Alignment (CAD) network that addresses the challenges in AVQA methods by i) introducing a parameter-free stochastic Contextual block that ensures robust audio and visual alignment on the Spatial level; ii) proposing a pre-training technique for dynamic audio and visual alignment on Temporal level in a self-supervised setting, and iii) introducing a cross-attention mechanism to balance audio and visual information on Semantic level. The proposed novel CAD network improves the overall performance over the state-of-the-art methods on average by 9.4% *on the MUSIC-AVQA dataset. We also demonstrate that our proposed contributions to AVQA can be added to the existing methods to improve their performance without additional complexity requirements.*

1. Introduction

Audio-visual inputs have been used extensively in the literature to improve the performance of various tasks including video captioning [37,67,97,109], speech recognition [1,34,69,87,89], speaker recognition [15,19,27,75,81,83,88], action recognition [5,12,26,43,72,74,80,94,100,104,118], emotion recognition [9–11, 16, 31, 63], sound localization [8, 35, 36, 71, 84], saliency detection [38, 64, 96, 99, 101], event localization [21, 58, 59, 98] and finally, question-answering [3,29,33,53,82,86,115]. However, training AV or multimodal methods is a challenging task and needs to be

Correspondence to: Asmar Nadeem <asmar.nadeem@surrey.ac.uk>.

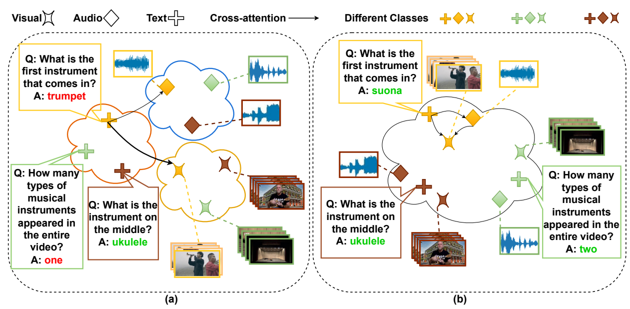

Figure 1. Visualization of audio, text and visual representations learned by (a) state-of-the-art ST-AVQA [53] and (b) CAD (ours).

addressed on Spatial, Temporal, and Semantic levels [102]. For example, in Figure 1 - yellow class, the spatial location of the instrument can be determined using audio and visual inputs but to answer 'which comes first?', we also need to simultaneously hear the sound of the instrument and see it in visual input.

At this stage, any temporal misalignment between sound and visual appearance of the instrument can hamper learning. We also need to relate both the sound and appearance of the instrument with the label of the instrument on the Semantic level. In Figure 1a, the state-of-the-art method ST-AVQA [53] predicts the wrong answer which is 'trumpet' and our method in Figure 1b predicts the right answer 'suona'. Similarly, for green class, there are two instruments xylophone and piano and the question is about 'appeared' not 'played' as only the xylophone was played' and to answer correctly, the model requires Semantic level learning which ST-AVQA [53] lacks. ST-AVQA [53] claims to assign larger weights to audio and visual segments which are more relevant to the asked question, by performing spatial grounding using audio and visual segments. This output is then temporally grounded with audio and question embedding. However, for scenarios as in the above examples, audio input is not available for both spatial and temporal groundings, which hampers the learning. As shown in Figure 1b, our contributions help align the same class modalities in a unified space with audio and text modalities used as queries as well as the chain of cross-attention, which helps in balancing AV information signals. In contrast, Figure 1a represents the state-of-the-art method ST-AVQA [53], where different modalities are misaligned for the same class.

Other multimodal methods [2,20,61,92,93,95,120] suffer from a major limitation of misaligned multi-modal information on Spatial and Temporal levels, leading to errors in the output. In AV learning, the audio and visual data are input separately to the network, and sampled at different rates. Due to the limitation of computing power and larger training time, the visual data is often more sparsely sampled as compared to audio data which requires less memory [53]. This leads to two challenges: 1) misalignment between audio and visual streams; and 2) imbalance between audio and visual semantics within a context. To achieve the best performance, ideally, the audio and visual information should be perfectly aligned and also, semantically balanced within a context. In this paper, we focus on the task of AVQA with multi-modal learning and address both the challenges in the existing methods.

Some methods [2,59,65,68] use self-attention and cross-attention to look for the most co-related parts of the audio and visual features, however, the problem is only partially solved. Recently, transformer-based methods [2, 20, 61, 92, 93, 95, 120] have become popular for their ability to use unlabeled data in a self-supervised manner by using contrastive learning which limits multi-modal class differentiation in the feature space. [2] employed the DropToken technique which drops features randomly to deal with the redundancy in audio and visual streams. Other methods [7,8,47,66] synchronise audio and visual streams as part of the self-supervised learning and [47] report a significant increase in performance on UCF101 [90] and HMDB51 [48] benchmarks. However, these methods do not resolve the problem of imbalance in AV semantics and misalignment between AV streams and do not work efficiently for more challenging datasets with dynamic scenarios [53].

In this paper, we aim to address the above challenges by introducing (1) a spatio-temporal Contextual block for visual features to mitigate the **AV spatial** misalignment; (2) **AV temporal** alignment of AV features as a pre-training task in a self-supervised manner; and (3) a chain for the cross-attention modules in our architecture for creating **AV semantic** balance to handle dynamic scenarios.

To summarise, we propose a **novel end-to-end network** for dynamic AVQA with three main contributions:

- Parameter-free contextual block to identify most context-related parts in the visual stream, reducing the spatial misalignment with the audio stream and reducing network complexity.
- AV fine temporal alignment uses self-supervised learning to pre-train the model for dynamically aligning audio and visual information over time. This helps the network understand and represent the temporal aspects of the data effectively.
- A chain of three cross-attention modules where initially audio and visual streams are cross-attended by question queries and then, the output of the audio cross-attention block queries the visual stream. This ensures that the network processes and balances the semantic (meaning-related) information from both modalities effectively.

Our contributions can be added to the existing methods to improve the performance without increasing any complexity (see Section 4.4).

2. Related Work

Previous research has primarily focused on question-answering tasks utilizing different modalities, such as audio, text, and vision, as input [3, 6, 39, 79, 103, 117]. In the context of visual question-answering (VQA) tasks [30, 62], the goal is to generate answers based solely on visual information. However, these methods do not incorporate spatio-temporal dynamics inherent in visual content. Recent advances address this limitation by enhancing the spatio-temporal reasoning ability through the integration of video-centered dynamics into the question-answering task [23, 45, 56, 105, 113]. This progress showcases the promising nature of AVQA as a developing research area [53, 115].

To facilitate AVQA research, multiple datasets have been introduced. The Pano-AVQA dataset [115] consists of 5.4k videos with questions based on dynamic audio, visual, and AV scenes. However, the AV questions in this dataset only cover existential and location categories. The MUSIC-AVQA dataset [53] gives a comprehensive set of questions for AV scenes, including comparative, existential, counting, spatial, and temporal aspects in dynamic scenarios with Audio, visual, and AV modalities.

Li et al. [53] employ spatial and temporal grounding techniques to perform the AVQA task using cross-attention. Interestingly, their ablations on the MUSIC-AVQA dataset demonstrate that using only audio as input yields better performance than using only visual in an AV scenario. This observation highlights a unique scenario where the same question (Q) is used as the query for both A and V, but V is not queried for A. As a result, the direct correlation between A and V appears to be missing, leading to the under-utilization of A-related information within the V modality.

AV learning using Transformers Transformer is a powerful architecture that has achieved exceptional performance on various language tasks including question-answering [44]. Popular models based on transformers include BERT, RoBERTa, and GPT versions, with ChatGPT being particularly well-known [18, 60, 70, 77, 78]. Transformers typically follow a two-stage learning process. In the first stage, pre-training is performed using a large-scale dataset in either a supervised or self-supervised manner. This pre-training significantly enhances the performance of large transformer models in both vision and language tasks. Masking techniques have been introduced to process visual and textual streams during pre-training, yielding promising results [18, 32, 54, 60, 106]. Transformers employ a self-attention mechanism to improve their performance. Self-attention measures the relevance of different components within a sequence, such as patches of an image or words in a sentence, and models their interactions to optimize the output [44].

On the other hand, self-supervised learning (SSL) is used to train transformer models using unlabeled data, unlike supervised learning that relies on labeled data [22]. SSL has the advantage of enabling training on large datasets that would be costly or time-consuming to label manually. This makes SSL a promising approach for developing AI systems that can tackle real-world problems. However, there are still challenges to overcome in SSL, including the development of more efficient algorithms and better methods for transferring learned representations across tasks [22].

Transformer models can be represented as graph neural networks, where self-attention processes the input as a fully-connected graph in a global fashion [110]. This enables transformers to effectively handle multiple modalities as nodes of the graph, leading to AV learning. AV learning involves tokenizing multiple modalities and representing them in a feature or embedding space. The design of this embedding space, which can have different granularity levels (e.g., fine or coarse), aims to establish class differentiation, often based on contrastive learning [42]. However, contrastive learning faces challenges when dealing with AV tasks, as its performance relies on the quality and quantity of negative examples. Insufficient negative examples can prevent contrastive learning from converging and result in poor performance due to a lack of differentiation in the embedding space [13].

To address the integration of multiple modalities in a single transformer architecture, cross-attention is used, allowing each modality to be processed with respect to the query of the other without significantly increasing computational cost. However, cross-attention tends to focus on global information, overlooking the fine-grained details within each modality, primarily due to the absence of a self-attention mechanism for individual modalities [57]. Additionally, multimodal pre-training approaches utilize diverse large-scale multimodal datasets to train transformer-based models. These models, when trained on such datasets, outperform in multiple downstream tasks and demonstrate strong generalization abilities [14, 55, 61, 91, 95]. The pre-training tasks aim to capture cross-modal interactions through either general or goal-oriented objectives. Unlike the alignment between co-occurring modalities in AV tasks, which has an inherent nature, transformer-based alignment techniques primarily leverage large amounts of data for vision-language tasks [40, 52, 65, 76, 108, 110]. Preserving the alignment between modalities while mapping them into a shared embedding space remains a challenging task. Using contrastive learning to address this proves to be ineffective [13].

3. Method

This section gives an overview of the proposed method, followed by a problem statement and detailed explanations of three key components: (1) spatio-temporal visual Contextual block, (2) AV fine temporal alignment, and (3) Cross-Attention based network.

3.1. Overview:

As shown in Figure 2, our method takes audio and visual streams extracted from the video as input. The question query is encoded using [73]. Each input stream is passed through separate pre-trained backbone encoder networks - one for audio [46] trained on AudioSet [28] and another for visual input [20] trained on ImageNet [17], to extract respective audio and visual features. The visual features are fed into the parameter-free Contextual block to choose features relevant to the query. This decreases the size of the features reducing the overall complexity of the network.

The audio and visual (refined) features are input to a novel network setting of three cross-attention modules. These modules align the inputs and integrate information across modalities, and help in dealing with dynamic scenarios. The outputs of these modules are concatenated and passed through a classifier for output prediction for both training and testing. In the pre-training stage, each of these outputs is sent to separate classifiers, each responsible for predicting the time label.

3.2. Problem Statement

For challenging datasets such as MUSIC-AVQA [53] with dynamic scenarios (types of questions including comparative, existential, counting, spatial and temporal aspects, in both audio, visual, and AV formats), existing networks suffer from two challenges.

The first one is that the audio and visual streams are often misaligned, which makes it difficult for the model to answer questions that require spatio-temporal reasoning. Another challenge is that the semantic information in the audio and visual streams is imbalanced within a context.

To address these challenges, our method incorporates a pre-training task where the audio and visual inputs are segmented into 60 cues, each lasting for one second. These cues align the audio and visual streams temporally. The visual cues are input to the spatio-temporal stochastic Contextual block in pre-training as well as training, which

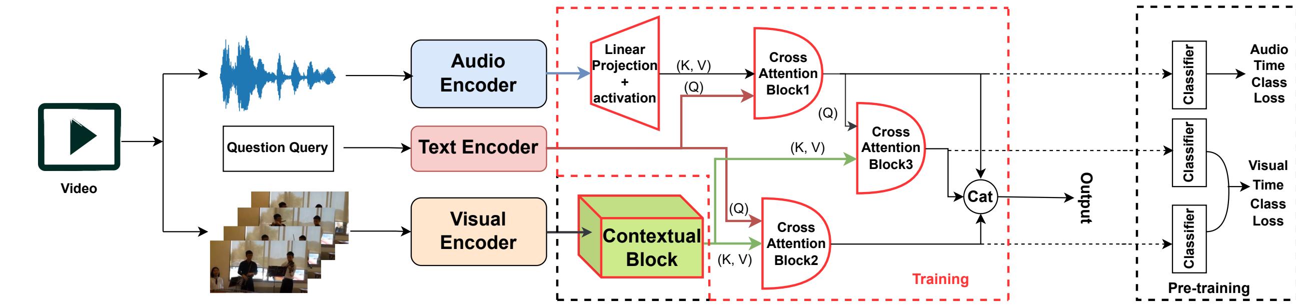

Figure 2. The overview of the proposed novel end-to-end **CAD** - **C**ontextual multi-modal **A**lignment for **D**ynamic AVQA. The novel parts of the network are highlighted in dotted lines.

aligns the audio and visual inputs spatially and also, reduces the overall complexity of the network, enhancing its efficiency. Our approach also employs three cross-attention modules as part of the pre-training process, which aligns and balances the audio and visual semantics. Following the pre-training phase, we initialize our network with the pre-trained weights and proceed to train it in an end-to-end fashion, adopting a supervised learning approach. For a detailed description of the training procedure, please refer to Algorithm **1**.

Algorithm 1 The proposed CAD method

Require: a$\in A^{d_a}$,t$\in T^{d_t}$,v$\in V^{d_v}$ ▷ Audio (a), Text (t) and Visual (v) features
  Initialize model weights from the pre-trained weights
  for batch iteration n = 1, 2, ..., N **do**
    {*Contextual block starts*}
    if training **then** ▷ Sample 80% of v features
      $f_m \leftarrow Mean(\text{v}, dim = c)$ ▷ c=channel
      $M \leftarrow Mask(f_m, th = Max(f_m) * 0.9)$
      $C_f \leftarrow sigmoid(f_m)$ ▷ contextual features
      $R \leftarrow Random(C_f, M, select\_prob = 0.9)$
      v $\leftarrow$ v$*R$
    end if
    {*Contextual block ends*}
    $a_t \leftarrow CAB1(query = t, key = a, value = a)$
    $v_t \leftarrow CAB2(query = t, key = v, value = v)$
    $v_{a_t} \leftarrow CAB3(query = a_t, key = v, value = v)$
    {*CAB = Cross Attention Block*}
    $answer \leftarrow FC(Fuse\{a_t, v_t, v_{a_t}\})$
    $Loss \leftarrow CrossEntropy(answer, ground\_truth)$
  end for

3.3. Contextual Block

In this section, we explain our stochastic visual Contextual block, which plays a significant role in both the pre-training and training phases. In AV learning, the audio and visual streams contribute complementary information for the AVQA task. However, the learning process struggles in extracting spatial information from audio and visual streams concurrently, inadvertently limiting the network's ability to effectively learn from both streams [53].

The Contextual block extracts spatio-temporal visual context and allows the network to explicitly incorporate visual information most relevant for Spatial level learning. This facilitates better learning and improves the overall performance of the network when dealing with AV tasks. This block identifies contextually relevant regions in the visual input. To achieve this, we employ a series of steps that highlight these regions. At first, we randomly sample 80% of the visual features for this process, as outlined in Algorithm **1**.

In the subsequent step, we average the visual features along the channel dimension. Then, we create a mask M with all the values greater than threshold th to be zero and otherwise, one. Similarly, we extract context C_f using sigmoid attention. Next, we introduce a stochastic selection between 90% of either M or C_f, for robustness, which is then masked out to zero. This helps to effectively reduce the complexity and training time. Further details are provided in the Appendix. The feature space $\mathcal{R}$ is defined as:

$$\mathcal{R} = \{\{A^{d_a}, T^{d_t}, V^{d_v}\} | \forall a \in A^{d_a}, t \in T^{d_t}, v \in V^{d_v}\} \tag{1}$$

Here, A^{d_a}, T^{d_t} and V^{d_v} represent the audio, text, and visual feature spaces, respectively. Similary, a, t and v denote the respective feature embeddings, with feature dimensions d_a, d_t and d_v. The visual features are represented by $\mathcal{V}^{[B*t*s]}$, where B, t and s denote the batch size, number of time frames, and spatial size, respectively. The complexity of the network is directly influenced by $[t * s]$.

The output of this block is input to two cross-attention modules sequentially which enables the network to focus on the most relevant information for Spatial level learning.

3.4. AV Fine Temporal Alignment as pre-training

The contextual block takes as input visual features, and it outputs a new set of features that are more spatially relevant to the audio stream. This makes it easier for the model to learn the interactions between the audio and visual streams. In this section, we pre-train our neural network to classify

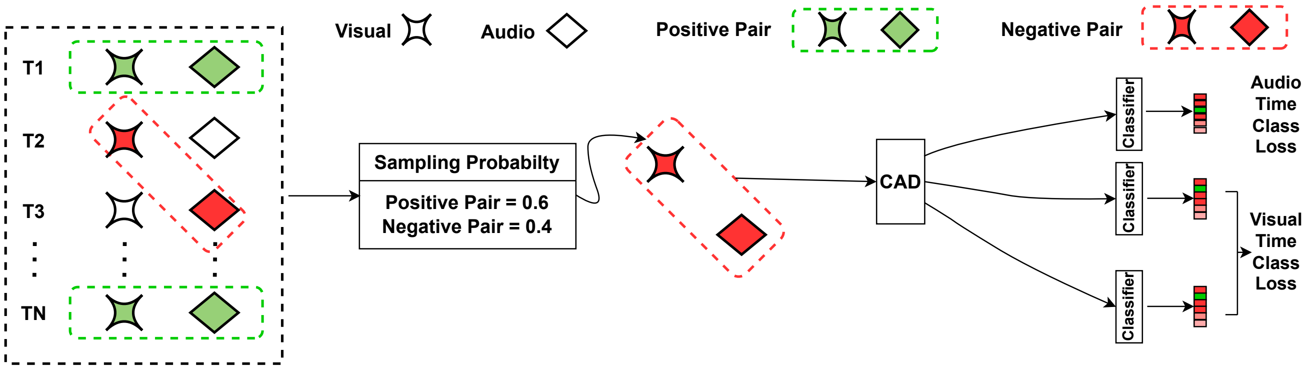

Figure 3. The overview of the proposed AV Fine Temporal Alignment where encoded audio-visual positive and negative pairs are sampled with a probability of 0.6 and 0.4 respectively before passing these through our CAD network. Time labels (N=60) for audio and visual are predicted separately via Softmax classifiers.

the time label or class of both audio and visual streams for better Temporal level learning. We propose an objective for the pre-training task to temporally align the audio and visual streams for question answering. We input audio and visual features to the network as a combination of positive and negative pairs. Positive pairs are pairs of audio and visual features that are aligned, and negative pairs are pairs of audio and visual features that are not aligned as shown in Figure 3.
This helps the network to learn to distinguish between aligned and non-aligned pairs. Audio and visual features are input stochastically to the network in a combination of positive and negative pairs with an overall share of 60% and 40% respectively which is selected after thorough experimentation provided in the Appendix. As shown in Figures 2 and 3, three classifiers are added at the end of the network to predict audio and visual time labels. Visual stream is predicted by two classifiers with both text and attended (output of audio cross-attention block) audio as queries. We use the cross-entropy loss for all three classifiers. The only difference in the training phase is removing the three classifiers and the concatenation of outputs of the cross-attention modules prior to it (see Figure 2).

3.5. Cross Attention Blocks

Our approach aims to effectively address the semantic information imbalance in the audio and visual streams, by incorporating three cross-attention blocks within the network, as seen in Figure 2. Two cross-attention blocks take the audio features and post-Contextual block visual features as input keys and values while utilizing question features as queries. These blocks enable the network to capture relevant information from both the audio and visual modalities and align it with the query, facilitating a comprehensive understanding of the question. The third cross-attention block incorporates visual features as keys and values, with queries sourced from the output of the audio cross-attention module. This configuration ensures that information from the audio stream is effectively propagated to the visual stream, allowing for enhanced integration and alignment of features. By incorporating all three cross-attention blocks, we provide a robust framework that covers all the aforementioned dynamic scenarios, resulting in improved overall performance. Algorithm **2** gives the technical details about the cross-attention block.

Algorithm 2 The proposed Cross-Attention Blocks

Require: k,v,q ▷ key (k), value(v) and query (q)
Initialize weights from the pre-trained weights
$f_a \leftarrow Multi\_Headed\_Attention$(q,k,v)
$fc1 \leftarrow FC(f_a)$ ▷ fully connected layer (512,512)
$r1 \leftarrow relu(fc1)$ ▷ Activation
$fc2 \leftarrow FC(r1)$ ▷ fully connected layer (512,512)
$r2 \leftarrow relu(fc2)$ ▷ Activation
$f_f \leftarrow f_a + r2$ ▷ Addition
$f_n \leftarrow norm(f_f)$ ▷ Layer Norm
return f_n

3.6. End-to-end CAD

In this section, we explain the end-to-end architecture employed in our method. The outputs from three cross-attention blocks are concatenated together and fed into a fully-connected layer that learns the answer embedding. This enables the network to make predictions on the most likely answer for a given input.

$$\mathcal{L}_{avqa} = -\sum_{n=1}^{N} A_n log P_n \tag{2}$$

A_n is the actual answer embedding and P_n is the output of the classifier and $\mathcal{L}_{avqa}$ is the AVQA loss. During the training phase, we employ the cross-entropy loss function to train our network. This loss function effectively measures the dissimilarity between the predicted and the ground truth labels, facilitating the optimization process. It is worth noting that the network, prior to the fusion, is initialized with weights obtained from the pre-training phase. This initialization provides a starting point that already incorporates valuable knowledge obtained during pre-training, allowing for more efficient learning during the training phase.

4. Experimental Results and Implementation

4.1. Dataset

We train and test our model on MUSIC-AVQA [53], a large-scale dataset that contains question-answer pairs for audio, visual, and AV questions about musical performance. This dataset contains 9,290 YouTube videos, 45,867 question-answer pairs and 9 types of diverse, complex and dynamic AV questions. We employ the ACAV100M [50] dataset for the pre-training task. We detail our pre-training stage dataset sampling, preprocessing, and labeling strategy in the Appendix.

4.2. Implementation

The audio input is sampled at 32kHz, which is a standard sampling rate for audio. We use PANNs [46] to extract features from audio data. The visual input is sampled at 15 frames per second in the pre-training and training. We use ViT [20], which is a transformer-based network, to extract features from images. The textual/question input is embedded using GLoVE [73]. The hidden dimension and cross-attention dimension size are both set to 512. The cross-attention modules employ 8 heads. The other training parameters are a learning rate of 0.0001, 25 epochs, a batch size of 64, and the ADAM optimizer. The training is done using one NVIDIA GeForce RTX 2080 Ti GPU. We use the same parameters for pre-training except the epochs are 10. Further details are provided in the Appendix.

4.3. Results and Comparison

Quantitative Evaluation. We compare the quantitative performance of the proposed method with the state-of-the-art approaches. For MUSIC-AVQA [53] dataset, the evaluation encompasses three multi-modal scenes, namely audio, visual, and audio-visual, and covers a total of nine different types of questions (refer to Table 1). For a fair comparison, we use the prediction accuracy of the answer as the evaluation metric where the answer vocabulary contains 42 possible answers. Notably, our method showcases superior performance across all categories and question types when compared to state-of-the-art techniques. Specifically, in the AV scene, our method achieves an improvement of 10.7%, followed by 7.7% and 5.5% improvements in the V and A scenes, respectively. These results affirm the effectiveness of our approach in tackling diverse audio-visual scenarios and addressing various types of questions. Our method showcases improvements, particularly in addressing AV temporal, localization, and comparative questions, where we observe enhancements of 20.5%, 14.7%, and 13.4% respectively. This demonstrates the effectiveness of our approach in capturing and understanding temporal dynamics, accurately localizing elements, and facilitating comparative reasoning within AV contexts. Additionally, we achieve notable improvements of 9.3% and 9.4% in V counting and A comparative questions, further highlighting the versatility and robustness of our method across different modalities. While the improvement is relatively lower in AV existential questions, our method still achieves a modest enhancement of 2%, indicating its capability to handle and reason about the existence of AV elements. It is worth mentioning that our method belongs to the AVQA task in Table 1 and all the listed methods are trained on the AVQA benchmark dataset as reported in [53]. These tasks are categorized based on the input modality or modalities. We also train and test our CAD method on MSRVTT-QA [107] and ActivityNet-QA [113] benchmark datasets and demonstrate significant improvement in comparison to the existing methods by optimizing AV learning as shown in Table 2. These results provide compelling evidence of the efficacy and applicability of our method in various question types, with substantial advancements in addressing specific challenges related to temporal, spatial, and comparative aspects in the AV domain.

Qualitative Evaluation. In Figure 4, we demonstrate the results of our method and do a comparison with the state-

Task	Method-Year	Audio			Visual			Audio-Visual						Avg
		Count	Comp	Avg	Count	Localis	Avg	Exist	Localis	Count	Comp	Temp	Avg	
AudioQA	FCNLSTM [24]-TASLP2019	70.45	66.22	68.88	63.89	46.74	55.21	82.01	46.28	59.34	62.15	47.33	60.06	60.34
	CONVLSTM [24]-TASLP2019	74.07	68.89	72.15	67.47	54.56	60.94	82.91	50.81	63.03	60.27	51.58	62.24	63.65
VisualQA	GRU [6]-ICCV2015	72.21	66.89	70.24	67.72	70.11	68.93	81.71	59.44	62.64	61.88	60.07	65.18	67.07
	BiLSTM Attn [119]-ACL2016	70.35	47.92	62.05	64.64	64.33	64.48	78.39	45.85	56.91	53.09	49.76	57.10	59.92
	HCAttn [62]-NeurIPS2016	70.25	54.91	64.57	64.05	66.37	65.22	79.10	49.51	59.97	55.25	56.43	60.19	62.30
	MCAN [114]-CVPR2019	77.50	55.24	69.25	71.56	70.93	71.24	80.40	54.48	64.91	57.22	47.57	61.58	65.49
VideoQA	PSAC [56]-AAAI2019	75.64	66.06	72.09	68.64	69.79	69.22	77.59	55.02	63.42	61.17	59.47	63.52	66.54
	HME [23]-CVPR2019	74.76	63.56	70.61	67.97	69.46	68.76	80.30	53.18	63.19	62.69	59.83	64.05	66.45
	HCRN [49]-CVPR2020	68.59	50.92	62.05	64.39	61.81	63.08	54.47	41.53	53.38	52.11	47.69	50.26	55.73
AVQA	AVSD [82]-CVPR2019	72.41	61.90	68.52	67.39	74.19	70.83	81.61	58.79	63.89	61.52	61.41	65.49	67.44
	PanoAVQA [115]-ICCV2021	74.36	64.56	70.73	69.39	75.65	72.56	81.21	59.33	64.91	64.22	63.23	66.64	68.93
	ST-AVQA [53]-CVPR2022	78.18	67.05	74.06	71.56	76.38	74.00	81.81	64.51	70.80	66.01	63.23	69.54	71.52
	CAD (Ours)	**82.91**	**73.34**	**78.13**	**78.21**	**81.19**	**79.70**	**83.42**	**73.97**	**76.37**	**74.88**	**76.16**	**76.96**	**78.26**

Table 1. Comparison against state-of-the-art methods. The best results in each category are in bold.

of-the-art ST-AVQA [53] and the ground-truth. In terms of audio-visual (AV) and visual (V) categories, our method demonstrates better results than the ST-AVQA [53](see sub-figures i, v in Figure 4).

Modality	Method	MSRVTT-QA	ActivityNet-QA
V+Q	QueST [41]	34.6	-
	ClipBERT [52]	37.4	-
	JustAsk [111]	41.5	38.9
	MV-GPT [85]	41.7	39.1
	MERLOT [116]	43.1	41.4
	SINGULARITY [51]	43.5	43.1
	VIOLET [25]	44.5	-
	FrozenBiLM [112]	47.0	43.2
	Flamingo [4]	47.4	-
	CAD (Ours)	<u>47.53</u>	<u>46.92</u>
V+A+Q	**CAD (Ours)**	**49.06**	**48.81**

Table 2. Comparison against state-of-the-art methods on MSRVTT-QA [107] and ActivityNet-QA [113] datasets. The best results in each category are in bold.

Both of these results demonstrate the ability of our method, as shown in quantitative evaluation, to perform better than the state-of-the-art in AV and V scenes and also, perform equally well in the audio (A) category.
In Figure 4**i**, the state-of-the-art method ST-AVQA [53] predicts the wrong answer which is 'trumpet' and our method predicts the right answer 'suona'. The spatial location of the 'suona' can be determined using audio and visual inputs but to answer which comes first, we also need to simultaneously hear the sound of the 'suona' and see it in visual input. At this stage, any temporal misalignment between sound and visual appearance of the 'suona' can hamper learning. Our method successfully learns the Spatial and Temporal levels and then, it also relates both the sound and appearance with the word 'suona' on the Semantic level. In Figure 4**v**, there are two instruments i.e. xylophone and piano and the question is about 'appeared' not 'played' as only the xylophone was played and to answer correctly, the model requires Semantic level learning which ST-AVQA [53] lacks and predicts 'one' which is the wrong answer. These results demonstrate the ability of our method, as shown in quantitative evaluation, to perform better than the state-of-the-art in AV and V scenes and also, performs equally well in the audio (A) category.

4.4. Ablation Results and Discussion

In this section, we discuss the effect of our contributions. This includes the Contextual block, AV fine temporal alignment and the network of three cross-attention modules.

Contextual Block. Section 3.3 emphasizes the importance of incorporating both audio and visual streams in learning from video data, as they offer complementary information. To achieve better Spatial level learning, a Contextual block is introduced, enabling the network to extract spatio-

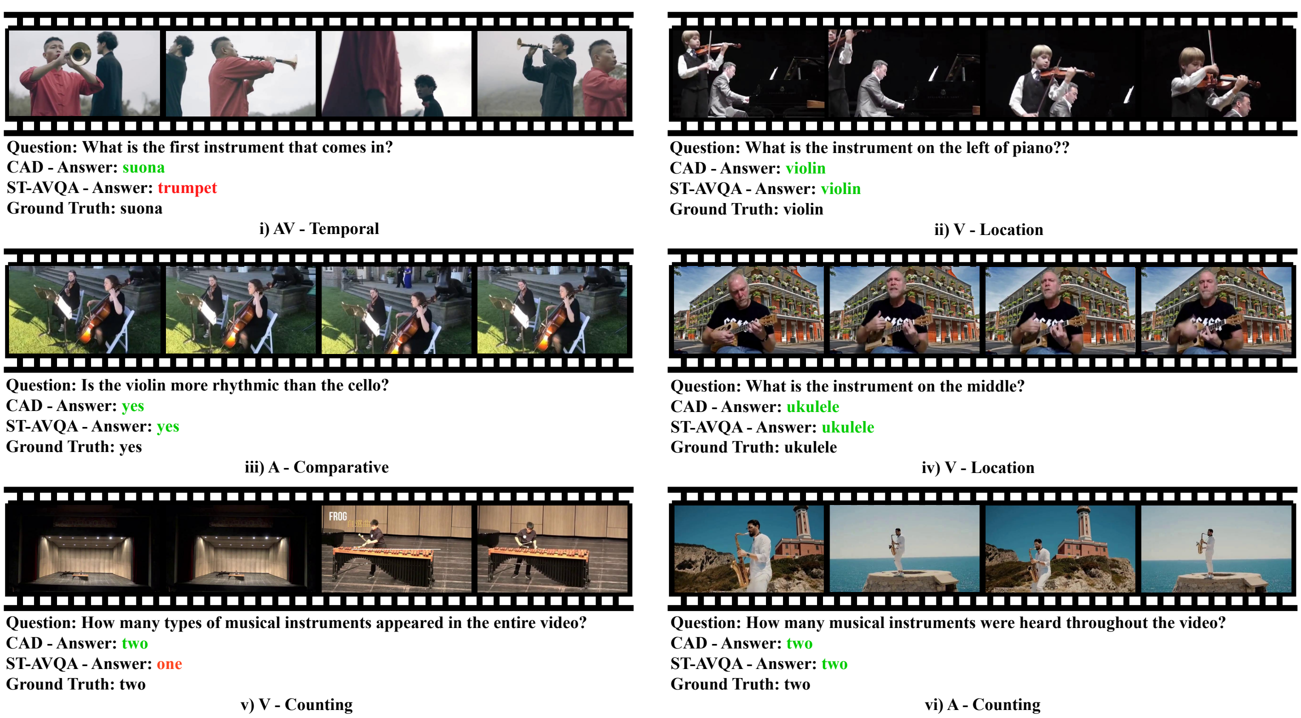

Figure 4. Qualitative results comparison against ground-truth and ST-AVQA [53].

Task	Audio			Visual			Audio-Visual						Avg
	Count	Comp	Avg	Count	Localis	Avg	Exist	Localis	Count	Comp	Temp	Avg	
w/o Pre-training	79.17	69.81	74.49	74.53	78.06	76.30	81.91	67.47	72.94	70.23	67.88	72.09	74.29
w/o Contextual block	81.13	70.05	75.59	75.91	79.62	77.77	82.24	70.41	73.14	71.37	71.26	73.68	75.68
w 3CA only	78.8	69.07	73.94	74.11	76.98	75.55	81.05	66.53	71.1	69.28	66.41	70.87	73.45
w 2CA	82.04	73.21	77.63	73.34	76.68	75.01	82.21	66.9	72.13	69.74	64.99	71.19	74.61
w 4CA	80.04	71.57	75.81	76.01	80.11	78.06	81.35	70.25	73.13	71.7	72.04	73.69	75.85
w ST-AVQA [53] + Pre-training	79.04	68.13	73.59	72.33	76.51	74.42	81.99	65.41	71.45	68.02	64.61	70.30	72.77
w ST-AVQA [53] + Contextual block	78.56	67.92	73.24	71.95	76.62	74.29	81.9	65.13	71.31	67.55	64.89	70.16	72.56
w Q	71.11	63.49	67.30	54.06	56.44	55.25	70.32	56.64	61.33	58.77	50.17	59.45	60.67
w A + Q	81.51	72.18	76.85	61.42	60.2	60.81	78.89	61.15	68.3	67.56	59.53	67.09	68.25
w V + Q	74.57	67.24	70.91	74.66	76.37	75.52	80.12	69.4	64.53	63.68	62.17	67.98	71.47
w A + V + Q - (CAD - Ours)	**82.91**	**73.34**	**78.13**	**78.21**	**81.19**	**79.70**	**83.42**	**73.97**	**76.37**	**74.88**	**76.16**	**76.96**	**78.26**

Table 3. Ablation results. The best results in each category are in bold.

temporal visual context. The output of the Contextual block is fed to the cross-attention modules, facilitating the unified representation of class features from different modalities. In Table 3, 'w/o Contextual block' clearly demonstrates the effectiveness of the Contextual block. Similarly, 'w ST-AVQA [53] + Contextual block' shows the effectiveness of the Contextual block when it is added to the existing method ST-AVQA [53]. It is important to note that while reimplementing ST-AVQA [53] with own our contributions, we employ training parameters and feature extractors used by our method for a fair comparison.

Pre-training - AV Fine Temporal Alignment (AVFA). Section 3.4 demonstrates the significance of fine temporal alignment between audio and visual streams. A pretraining task based on AV fine alignment plays an important role in improved performance. In Table 3, entry 'w/o Pre-training' shows the effectiveness of this. Similarly, 'w ST-AVQA [53] + Pre-training' shows the effectiveness of AVFA when re-implemented with ST-AVQA [53].

Three Cross-Attention Blocks (3CA). The three cross-attention blocks in our architecture play a significant role in Semantic level learning and also, in addressing questions in dynamic scenarios. Two blocks capture information from audio and visual modalities, aligning it with the query to facilitate a comprehensive understanding of the question. The third block propagates audio information to the visual stream, enhancing integration and feature alignment. By incorporating all three modules, we create a robust framework that improves overall performance by covering audio, visual, and AV dynamic scenarios effectively. In Table 3, entry 'w 3CA only' shows the effectiveness of this contribution without the other two contributions. Also, 'w 2CA' shows the limitation of the network with two cross-attention blocks using text as query and audio and visual as key, value for each, while using both AVFA and the Contextual block. Similarly, 'w 4CA' shows decreased performance when using another cross-attention module in addition to 'w 3CA' where visual semantic is employed as a query and audio as both key and value.

Effect of Input. The last 4 rows of the table show the effect of different input combinations on our proposed method. 'w Q' is when we only send questions (Q) as input. The model uses information available within the question to predict the answers. This is also the manifestation of Figure 1 where the same class but different modalities are closer to each other and even one modality as an input can help in an accurate answer. Rows 'w A + Q' and 'w V + Q' show results for audio and visual category questions which perform well when given their respective inputs i.e., audio + question (A + Q) and visual + question (V + Q) but perform adversely for the inverse case. The performance is balanced for the audio-visual question category for these two rows. The last row demonstrates performance with all the modalities given as input, demonstrating the best performance.

5. Conclusion

We introduce a novel CAD network for AVQA task. We proposed a parameter-free spatial alignment block, temporally aligned pre-training, and semantic audio-visual balance. The CAD network boosts performance on MUSIC-AVQA dataset against state-of-the-art methods, showcasing enhanced robustness and efficiency. Our work improves on the AVQA task and the contributions can find applications in other tasks such as video captioning, speech/speaker recognition, action recognition, etc. Our work addresses the challenge of AV misalignment at the Spatial, Temporal and Semantic level, which introduces inaccuracies in the understanding of the AV information, potentially resulting in biased decision-making. These contributions can potentially improve accessibility for individuals with sensory impairments.

6. Acknowledgement

This research was partly supported by the British Broadcasting Corporation Research and Development (BBC R&D), Engineering and Physical Sciences Research Council (EPSRC) Grant EP/V038087/1 “BBC Prosperity Partnership: Future Personalised Object-Based Media Experiences Delivered at Scale Anywhere”.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 1

[2] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 2

[3] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019. 1, 2

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 7

[5] Saghir Alfasly, Jian Lu, Chen Xu, and Yuru Zou. Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20208–20217, 2022. 1

[6] 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. 2, 6

[7] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 2

[8] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 1, 2

[9] Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 2021. 1

[10] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515, 2019. 1

[11] Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, 2020. 1

[12] Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1910–1921, 2022. 1

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 3

[15] Joon Son Chung, Bong-Jin Lee, and Icksang Han. Who said that?: Audio-visual speaker diarisation of real-world meetings. *arXiv preprint arXiv:1906.10042*, 2019. 1

[16] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955*, 2020. 1

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[19] Yifan Ding, Yong Xu, Shi-Xiong Zhang, Yahuan Cong, and Liqiang Wang. Self-supervised learning for audio-visual speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4367–4371. IEEE, 2020. 1

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 6

[21] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4013–4022, 2021. 1

[22] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022. 3

[23] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 2, 6

[24] Haytham M Fayek and Justin Johnson. Temporal reasoning via audio question answering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2283–2294, 2020. 6

[25] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study

of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22898–22909, 2023. 7

[26] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 1

[27] Israel D Gebru, Sileye Ba, Xiaofei Li, and Radu Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1086–1099, 2017. 1

[28] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 3

[29] Shijie Geng, Peng Gao, Moitreya Chatterjee, Chiori Hori, Jonathan Le Roux, Yongfeng Zhang, Hongsheng Li, and Anoop Cherian. Dynamic graph representation learning for video dialog via multi-modal shuffled transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1415–1423, 2021. 1

[30] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[31] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access, 2018. 1

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[33] Chiori Hori, Anoop Cherian, Tim K Marks, and Takaaki Hori. Joint student-teacher learning for audio-visual scene-aware dialog. In *INTERSPEECH*, pages 1886–1890, 2019. 1

[34] Di Hu, Xuelong Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582, 2016. 1

[35] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9844–9859, 2021. 1

[36] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022. 1

[37] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959, 2020. 1

[38] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3520–3527. IEEE, 2021. 1

[39] Yunseok 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

[40] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[41] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11101–11108, 2020. 7

[42] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in Neural Information Processing Systems*, 35:30291–30306, 2022. 3

[43] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 1

[44] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2, 3

[45] Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10106–10115, 2020. 2

[46] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 3, 6

[47] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[48] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video

database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2

[49] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 6

[50] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 6

[51] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 7

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 3, 7

[53] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 1, 2, 3, 4, 6, 7, 8

[54] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 3

[55] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[56] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019. 2, 6

[57] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198*, 2020. 3

[58] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006. IEEE, 2019. 1

[59] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 34:11449–11461, 2021. 1, 2

[60] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 3

[61] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2, 3

[62] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016. 2, 6

[63] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2554–2562, 2021. 1

[64] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020. 1

[65] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020. 2, 3

[66] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 2

[67] Asmar Nadeem, Adrian Hilton, Robert Dawes, Graham Thomas, and Armin Mustafa. Sem-pos: Grammatically and semantically correct video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2605–2615, 2023. 1

[68] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 2

[69] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning. *Applied intelligence*, 42:722–737, 2015. 1

[70] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 2

[71] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1

[72] Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. Adamml: Adaptive multi-modal learning for efficient video

recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7576–7585, 2021. 1

[73] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3, 6

[74] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Cross-domain first person audio-visual action recognition through relative norm alignment. *arXiv preprint arXiv:2106.01689*, 2021. 1

[75] Yanmin Qian, Zhengyang Chen, and Shuai Wang. Audio-visual deep neural network for robust person verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1079–1092, 2021. 1

[76] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[77] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[78] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9, 2019. 2

[79] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 2

[80] Faegheh Sardari, Armin Mustafa, Philip JB Jackson, and Adrian Hilton. Pat: Position-aware transformer for dense multi-label action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2988–2997, 2023. 1

[81] Leda Sarı, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf. A multi-view approach to audio-visual speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6194–6198. IEEE, 2021. 1

[82] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12548–12558, 2019. 1, 6

[83] Gregory Sell, Kevin Duh, David Snyder, Dave Etter, and Daniel Garcia-Romero. Audio-visual person recognition in multimedia data from the iarpa janus program. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3031–3035. IEEE, 2018. 1

[84] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 1

[85] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 7

[86] Ankit Shah, Shijie Geng, Peng Gao, Anoop Cherian, Takaaki Hori, Tim K Marks, Jonathan Le Roux, and Chiori Hori. Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7732–7736. IEEE, 2022. 1

[87] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022. 1

[88] Suwon Shon, Tae-Hyun Oh, and James Glass. Noise-tolerant audio-visual online person verification using an attention-based neural network fusion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3995–3999. IEEE, 2019. 1

[89] Qiya Song, Bin Sun, and Shutao Li. Multimodal sparse transformer network for audio-visual speech recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1

[90] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[91] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3

[92] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 2

[93] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 2

[94] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 1

[95] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2, 3

[96] Hamed R Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*, 2019. 1

[97] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. Audio-visual interpretable and controllable video captioning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*, 2019. 1

[98] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 1

[99] Antigoni Tsiami, Petros Koutras, and Petros Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4766–4776, 2020. 1

[100] Cheng Wang, Haojin Yang, and Christoph Meinel. Exploring multimodal video representation for action recognition. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1924–1931. IEEE, 2016. 1

[101] Guotao Wang, Chenglizhao Chen, Deng-Ping Fan, Aimin Hao, and Hong Qin. From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15119–15128, 2021. 1

[102] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022. 1

[103] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015. 2

[104] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 1

[105] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021. 2

[106] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3

[107] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 6, 7

[108] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pretraining for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[109] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. Learning multimodal attention lstm networks for video captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 537–545, 2017. 1

[110] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*, 2022. 3

[111] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021. 7

[112] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022. 7

[113] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 2, 6, 7

[114] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 6

[115] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021. 1, 2, 6

[116] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 7

[117] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022, 2016. 2

[118] Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees GM Snoek. Audio-adaptive activity recognition across video domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13791–13800, 2022. 1

[119] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212, 2016. 6

[120] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 2