CAD - Contextual Multi-modal Alignment for Dynamic AVQA

Asmar Nadeem¹, Adrian Hilton¹, Robert Dawes², Graham Thomas², Armin Mustafa¹
¹Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, United Kingdom.
²BBC 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 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
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: $a \in A^{d_a}, t \in T^{d_t}, v \in V^{d_v}$ $\triangleright$ 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 $\triangleright$ Sample 80% of v features

$f_m \leftarrow \text{Mean}(v, \text{dim} = c)$ $\triangleright$ c-channel

$M \leftarrow \text{Mask}(f_m, \text{th} = \text{Max}(f_m) * 0.9)$

$C_f \leftarrow \text{sigmoid}(f_m)$ $\triangleright$ contextual features

$R \leftarrow \text{Random}(C_f, M, \text{select}_{\text{prob}} = 0.9)$

$v \leftarrow v \ast R$

end if

{Contextual block ends}

$a_t \leftarrow \text{CAB1}(\text{query} = t, \text{key} = a, \text{value} = a)$

$v_t \leftarrow \text{CAB2}(\text{query} = t, \text{key} = v, \text{value} = v)$

$v_{a_t} \leftarrow \text{CAB3}(\text{query} = a_t, \text{key} = v, \text{value} = v)$

{Cross Attention Block

$answer \leftarrow \text{FC}(\text{Fuse}(a_t, v_t, v_{a_t}))$

$Loss \leftarrow \text{CrossEntropy}(answer, \text{ground truth})$

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 $R$ is defined as:

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

(1)

Here, $A^{d_a}$, $T^{d_t}$ and $V^{d_v}$ represent the audio, text, and visual feature spaces, respectively. Similarly, $a$, $t$ and $v$ denote the respective feature embeddings, with feature dimensions $d_a$, $d_t$ and $d_v$. The visual features are represented by $\mathcal{V}^{B \times t \times s}$, 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 $[t \ast s]$.

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
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 \(\triangleright\) key (k), value(v) and query (q)
Initialize weights from the pre-trained weights
\(f_n \leftarrow\) Multi_Headed_Attention(q,k,v)
\(fc_1 \leftarrow FC(f_n)\) \(\triangleright\) fully connected layer (512,512)
\(r_1 \leftarrow relu(fc_1)\) \(\triangleright\) Activation
\(fc_2 \leftarrow FC(r_1)\) \(\triangleright\) fully connected layer (512,512)
\(r_2 \leftarrow relu(fc_2)\) \(\triangleright\) Activation
\(f_f \leftarrow f_n + r_2\) \(\triangleright\) Addition
\(f_n \leftarrow norm(f_f)\) \(\triangleright\) 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
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-
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 subfigures i, v in Figure 4).

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

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 4i, 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 4v, 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-
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 re-implementing ST-AVQA [53] with our own 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 pre-training 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 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


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


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


[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. 1


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


[71] 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


