Q-TriM: Question-Guided Tri-Modal Attention for Audio–Visual Question Answering

SungHun Kim
Dept. of Computer Science and Engineering
Korea University
Seoul, South Korea
sainthun12@korea.ac.kr

SeungJun Baek*
Dept. of Computer Science and Engineering
Korea University
Seoul, South Korea
sjbaek@korea.ac.kr

Abstract. Audio-Visual Question Answering (AVQA) extends classical VQA by requiring joint reasoning over video and synchronized audio. However, many AVQA systems rely on deeply stacked layers of self- and cross attention across text, video, and audio. Such sequential stacking may incur loss of information such as subtle inter-modal cues over the layers, causing errors to accumulate across sequential attention layers during the fusion. We introduce Q-TriM which performs multi-modal fusion in a shallow and parallel manner instead of a deep and sequential manner. For Q-TriM, we propose a novel framework for attention operation incorporating video and audio conditioned on text. As a result, we obtain not only standard cross attention outputs but also Tri-Modal Attention representations in which Query, Key, and Value come from distinct modalities. These attention representations are combined in parallel at a single stage, thus avoiding the multi-modal fusion with deep stacks in order to mitigate error accumulation and depth-induced issues. Q-TriM achieves state-of-the-art performance on three AVQA benchmarks, including substantial gains on MUSIC-AVQA-R, which demonstrates its robustness and out-of-distribution generalization. Code is available at https://github.com/Sunghun95/Q-TriM

Keywords: Audio-Visual Question Answering · Attention Model· Multimodal Fusion

1 Introduction

AVQA(Audio-Visual Question Answering) is the task of answering a natural-language question given a video with synchronized audio, extending classical VQA [1] by jointly leveraging acoustic and visual streams [37]. AVQA spans many use cases: (1) scene-aware dialog/assistive agents that reason over live audio–video streams [32], (2) long-form egocentric "memory" systems that find evidence moments for open-ended questions [5], and (3) first-person action settings where EPIC-Fusion aligns sounds with egocentric visuals before answering [12].

* Corresponding author.

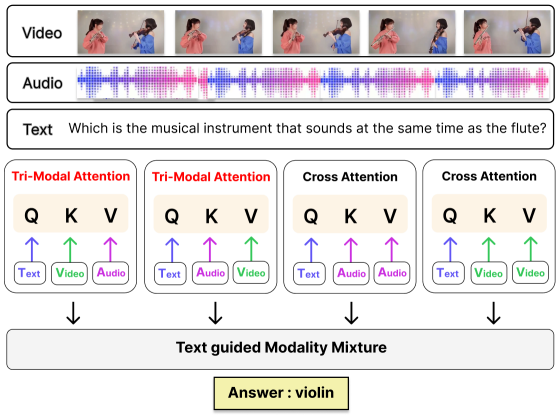

Fig. 1. Overview of Q-TriM. Given input video, audio, and text, Q-TriM combines representations from Tri-Modal Attention, together with cross attention, to perform a shallow, parallel fusion of the modalities for reasoning of the answer.

In AVQA, extracting salient evidence from both video and audio is essential. Most AVQA systems rely on *stacked* attention layers across text, video, and audio [13, 17, 22, 39] and deepen the attention stack to capture inter-modal dependencies. In general, neural layers perform the abstraction and compression of information. On the one hand, these operations are necessary to selectively extract essential information, e.g., Information Bottleneck (IB) principle [35]. On the other hand, they accompany the loss of information [11, 25, 31], which may be problematic for fusing information from many modalities. For example, a modality processed in earlier layers may lose important inter-modal cues, making them unavailable for fusion at deeper layers. A simple remedy is to design architectures that achieve similar effects with shallow attention layers. However, since AVQA must integrate three modalities, stacking attention is often unavoidable. In standard cross-attention [36], only two modalities interact at a time; a Query attends to a single Key–Value source, and thus at most two modalities interact simultaneously. For example, prior methods [17, 18] first process text and video with cross-attention, and then combine its output with audio representation via another cross-attention. Thus, coordinating text, video, and audio typically requires multiple fusion stages in a sequential manner.

In this paper, we propose **Q-TriM** (Question-guided Tri-Modal Attention and Modality Mixture) as illustrated in Figure 1. To avoid stacked attention, we go *shallow* and *parallel* rather than *deep* and *sequential*. Instead of sequential layering of blocks, we adopt a parallel fusion step to compute mixed features across all three modalities. To this end, we propose a novel framework for multi-modal attention. Our framework induces a joint distribution of video and audio representations per frame conditional on the input text. The fused representation is computed as a conditional expectation from the distribution. Unrolling these expectations (see Sec. 3.3 for details) reveals not only standard

cross-modal attention but also a new attention form in which the query, keys, and values are drawn from different modalities; we refer to this operator as **Tri-Modal attention**. Importantly, Tri-Modal attention models the three-way interactions among text, video, and audio in a *direct* manner within a single stage, thereby avoiding potential loss of inter-modal cues during the fusion. Finally, we compute a question-aware mixture of these outputs to produce the most relevant fused representation for reasoning. In addition, to ensure that only question-relevant patches participate in learning, we propose a *question-guided token filtering* module. Without introducing any extra attention, the token filtering module uses simple dot-product scoring and a straight-through estimator (STE) [3], thereby preventing further deepening of the attention stack. Consequently, our method achieves state-of-the-art performance on the standard AVQA benchmarks, MUSIC-AVQA [19] and MUSIC-AVQA-R [27]. On MUSIC-AVQA-2.0 [23], it further attains SoTA in three out of four evaluation categories.

Our main contributions are summarized as follows:

- We present **Q-TriM**, an AVQA model that performs inter-modal attention in a shallow, parallel fusion stage rather than by deep, sequential stacks. By consolidating fusion horizontally, Q-TriM achieves the effect of deeper cross-modal pipelines while curbing error propagation through deep layers.
- We propose a new modality-fusion framework that produces a question-conditioned joint distribution of video and audio frame tokens. Within this framework, we derive a Tri-Modal attention operator that allows distinct Query, Key, and Value sources across modalities, enabling direct and explicit three-way fusion without resorting to deep or sequential stacks.
- Q-TriM achieves state-of-the-art performance across three representative AVQA benchmarks. Notably, on the MUSIC-AVQA-R dataset [27], Q-TriM achieves superior performance over baselines, demonstrating its robustness.

2 Related Work

We group prior Audio–Visual Question Answering (AVQA) research into three axes: question awareness, spatial perception, and temporal perception. These axes respectively target how text guides reasoning, where to look and listen in the scene, and when salient evidence occurs across time.

2.1 Question Awareness

In AVQA, it is important to condition the model on the question so that both feature extraction and fusion remain text-guided. Co-attention and question-guided fusion from VQA (e.g., [20,26,39]) inspire AVQA systems that ground sounding regions and fuse audio–visual cues under textual guidance (e.g., [15,19,21]). Recent work further strengthens conditioning via prompt reformulation or question-aware experts: TSPM turns questions into declaratives to align with visual

semantics before temporal/spatial selection [17], and [13] introduces question-aware Gaussian experts. Dialog-style conditioning also appears in scene-aware settings [32]. Overall, these methods filter modalities through question semantics to suppress irrelevant content and to guide cross-modal reasoning [17, 19, 22].

2.2 Spatial Perception

Another line of work localizes question-relevant regions and sound sources, often by coupling detection with cross-modal grounding. Panoramic grounding requires spherical spatial reasoning in 360° videos [40]. Progressive pipelines (e.g., PSTP-Net) first select key segments and then pick relevant patches with audio-guided visual attention [18]. Complementary strategies include object-aware learning [21], causal regularization against spurious cues [15], and parameter-efficient ViT-based learners [22]. In practice, many approaches pair these spatial modules with question-aware fusion to maintain textual guidance throughout [17, 19].

2.3 Temporal Perception

Temporal perception targets frames or segments that matter for the question, ensuring that evidence is aggregated over the correct intervals. PSTP-Net explicitly performs segment selection before spatial filtering [18], while TSPM adds a prompt-guided temporal module [17]. Earlier video-/audio-QA studies explore temporal reasoning with recurrent or attention-based models (e.g., [7, 16, 20]), and dialog-style AV understanding leverages temporal context as well [32]. AVQA systems often integrate temporal selection with spatial grounding and question-aware fusion for end-to-end reasoning [19].

Overall, inter-modal attention is typically stacked in the aforementioned works (at least two layers). Notably, QA-TIGER [13], the state-of-the-art method at the time of writing, uses four cross-attention blocks which are sequentially connected.

3 Method

Q-TriM adopts a shallow and parallel architecture that combines cross-attention with Tri-Modal attention, instead of stacking attention blocks deeply. An overview of the architecture is shown in Figure 2.

Notations. A vector $\mathbf{v}$ is defined as a *row* vector by default. The dot-product between vectors $\mathbf{u}$ and $\mathbf{v}$ is thus $\mathbf{u}\mathbf{v}^\top$. A matrix $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times D}$ can be viewed as having N rows of D-dimensional vectors $\mathbf{x}_i$, $i = 1, \ldots, N$. Define the conventional dot-product attention for representations $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ of proper

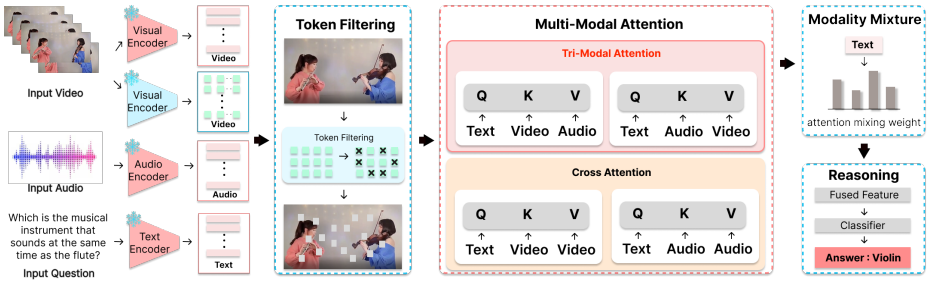

Fig. 2. Overall architecture of Q-TriM. First, question-guided token filtering keeps only the top tokens most relevant to the text and drops the rest. Next, Question-aware modality mixture computes four paths: TMA(Text,Audio,Video), TMA(Text,Video,Audio), CA(Text,Video,Video), and CA(Text,Audio,Audio). TMA stands for Tri-modal attention, and CA stands for cross-attention. Then question-conditioned weights are applied to the outputs. Finally, fused-feature reasoning combines these outputs and feeds the result to a classification layer to produce the answer.

dimensions[1]:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) := \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V} \tag{1}$$

3.1 Input Representation

We project three modalities, i.e., text, video, and audio, into a *common* embedding space. Such aligned representations help modeling direct interactions across modalities via attention. Firstly, we project text and video into CLIP embedding spaces [29]. Secondly, we align audio with text and video; there are candidate methods extending CLIP to audio, e.g., AudioCLIP [10], ImageBind [8]. We choose a recent method: ImageBind. We segment each original video into T contiguous *2-second* clips and extract frame-aligned video and audio representations per segment. The text consists of a single question for the entire video. The details of input representation is explained as follows.

Text representation. The question text $\mathbf{t}$ is encoded as a single sentence-level embedding from CLIP [29] given by

$$\mathbf{t} \in \mathbb{R}^D.$$

Video representation. We extract frame-level global video embeddings from CLIP given by

$$\mathbf{V} = \{\, \mathbf{v}_t \,\}_{t=1}^{T} \in \mathbb{R}^{T \times D}.$$

[1] The attention originally contains normalization in the softmax, e.g., $QK^\top$ is scaled by $1/\sqrt{D}$ if the dot-product is of dimension D. We omit $1/\sqrt{D}$ for notational simplicity, and assume that such normalization is performed whenever necessary for implementation.

Additionally, in order to retain spatial information, we obtain patch-level tokens using CLIP model given by

$$\mathbf{V}^{\mathrm{p}} = \{\, \mathbf{v}_t^{\mathrm{p}} \,\}_{t=1}^{T} \in \mathbb{R}^{T\times P\times D}.$$

Audio representation. We extract frame-level audio embeddings aligned with text and image CLIP using ImageBind [8] given by

$$\mathbf{A} = \{\, \mathbf{a}_t \,\}_{t=1}^{T} \in \mathbb{R}^{T\times D}.$$

3.2 Token Filtering

Overview. The entire visual representation $\mathbf{V}^{\mathrm{p}}$ from Sec. 3.1 may contain irrelevant or redundant tokens [30, 38]. This can incur unnecessary computational overhead and produce noise. We therefore introduce a lightweight token-filtering module that, without any attention, retains only text-relevant patches and preserves salient evidence. Because the token selection is discrete, we train it end-to-end with a straight-through estimator (STE) [3], which allows gradients to pass through the hard selection mask.

Scoring and STE-based Top-K filtering. Let $k \in (0, 1]$ denote the *keep ratio* defined as the fraction of patches to retain. Let K be the number of patches retained per frame. We first project the question to the patch channel and compute per-patch compatibility scores:

$$\mathbf{t}' = W_t\mathbf{t} + \mathbf{b}_t, \qquad \mathbf{s}_t \;=\; \frac{\mathbf{t}'\,(\mathbf{v}_t^{\mathrm{p}})^{\top}}{\sqrt{D'}} \;\in \mathbb{R}^{P},$$

We select the indices of the top-K scores:

$$I_t \;=\; \mathrm{TopK}(\mathbf{s}_t, K),$$

Next, we gather the corresponding patch tokens

$$X_t^{\mathrm{hard}} \;=\; \mathbf{v}_t^{\mathrm{p}}[I_t] \;\in\; \mathbb{R}^{K\times D'}.$$

To keep the forward path identical to the hard selection while allowing a gradient flow into the scoring path, we adopt a straight-through estimator (STE) [3]. Specifically, let $\tau > 0$ be a temperature and define the soft weights and their top-K slice:

$$\mathbf{w}_t \;=\; \mathrm{softmax}\Big(\tfrac{\mathbf{s}_t}{\tau}\Big) \in \mathbb{R}^{P}, \qquad \mathbf{w}_t^{\mathrm{top}} \;=\; \mathbf{w}_t[I_t] \in \mathbb{R}^{K}.$$

We form a softly reweighted selection by broadcasting $\mathbf{w}_t^{\mathrm{top}}$ across feature dimensions:

$$X_t^{\mathrm{soft}} \;=\; X_t^{\mathrm{hard}} \;\odot\; \mathbf{w}_t^{\mathrm{top}} \;\in\; \mathbb{R}^{K\times D'}.$$

where $\odot$ denotes elementwise multiplication. Finally, we use the STE composition using stop-gradient $\mathrm{sg}(\cdot)$:

$$\mathbf{v}_t^{\mathrm{p}} \;=\; \mathrm{sg}\big(X_t^{\mathrm{hard}}\big) \;+\; \Big(X_t^{\mathrm{soft}} \;-\; \mathrm{sg}\big(X_t^{\mathrm{soft}}\big)\Big).$$

Thus, the forward pass satisfies $X_t \equiv X_t^{\text{hard}}$, whereas the backward pass routes gradients through $\mathbf{w}_t^{\text{top}}$ into the scores $\mathbf{s}_t$ and ultimately into W_t. In all experiments, we set the keep ratio to $k{=}0.5$ (i.e., $K = P/2$).

Self-attention refinement. After the filtering stage, we enhance the representations by applying self-attention to each stream—the selected patch-token set $\mathbf{V^p}$, the frame-level video $\mathbf{V}$, the audio $\mathbf{A}$, and the question $\mathbf{t}$.

3.3 Attention Framework for Modality Fusion

Overview. We propose a new framework for multi-modal attention which fuses video and audio frame representations conditional on the input text. The key task for AVQA is modeling the relation between video and audio given the input text (query). For that purpose, we define Conditional Similarity Score (CSS) as follows. The CSS between i-th video frame and j-th audio frame conditional on text query $\mathbf{t}$ is denoted by $s(i, j|\mathbf{t}) \in \mathbb{R}$. We define CSS matrix $\mathbf{S} \in \mathbb{R}^{T\times T}$ that has $s(i, j|\mathbf{t})$ as the (i, j)-th element, i.e.,

$$[\mathbf{S}]_{i,j} = s(i, j|\mathbf{t}) \tag{2}$$

Thus, the row direction of $\mathbf{S}$ represents video frames, and column direction of $\mathbf{S}$ represents audio frames. We define a probability distribution from $\mathbf{S}$ by applying a 2-D softmax to $\mathbf{S}$, i.e.,

$$P_{\mathbf{S}}(i, j|\mathbf{t}) = \frac{\exp(s(i, j|\mathbf{t}))}{\sum_{i'} \sum_{j'} \exp(s(i', j'|\mathbf{t}))} \tag{3}$$

Thus, $P_{\mathbf{S}}(i, j|\mathbf{t})$ can be considered as a relative probability mass (weight) encoding the importance of i-th video and j-th audio frame conditional on question $\mathbf{t}$. We consider three types of CSS $s(i, j|\mathbf{t})$ as follows.

1) Default CSS. Let $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^T$ and $\mathbf{A} = \{\mathbf{a}_k\}_{k=1}^T$ be $T \times D$ matrices of video and audio representations respectively. The first CSS $s(i, j|\mathbf{t})$ for i-th video and j-th audio frames is defined as

$$s(i, j|\mathbf{t}) = \mathbf{v}_i\mathbf{t}^\top + \mathbf{a}_j\mathbf{t}^\top \tag{4}$$

The CSS is a combination of the similarity between $\mathbf{v}_i$ and query $\mathbf{t}$ and the similarity between $\mathbf{a}_j$ and query $\mathbf{t}$. We derive the video embedding as the weighted sum with respect to the distribution $P_{\mathbf{S}}(i, j|\mathbf{t})$. Equivalently, it can be considered as a *conditional expectation*:

$$\mathbb{E}\left[\mathbf{V}|\mathbf{t}\right] := \sum_{i,j} P_{\mathbf{S}}(i, j|\mathbf{t}) \cdot \mathbf{v}_i \tag{5}$$

We have that

$$\mathbb{E}\left[\mathbf{V}|\mathbf{t}\right] = \text{Attn}(\mathbf{t}, \mathbf{V}, \mathbf{V}) \tag{6}$$

because

$$\mathbb{E}\left[\mathbf{V}|\mathbf{t}\right] = \sum_{i,j} \mathbf{v}_i \cdot P_{\mathbf{S}}(i,j|\mathbf{t}) \tag{7}$$

$$= \sum_{i,j} \frac{\mathbf{v}_i \exp(\mathbf{v}_i \mathbf{t}^\top + \mathbf{a}_j \mathbf{t}^\top)}{\sum_{i'} \sum_{j'} \exp(\mathbf{v}_{i'} \mathbf{t}^\top + \mathbf{a}_{j'} \mathbf{t}^\top)} \tag{8}$$

$$= \frac{\sum_i \mathbf{v}_i \exp(\mathbf{v}_i \mathbf{t}^\top) \cdot \sum_j \exp(\mathbf{a}_j \mathbf{t}^\top)}{\sum_{i'} \exp(\mathbf{v}_{i'} \mathbf{t}^\top) \cdot \sum_{j'} \exp(\mathbf{a}_{j'} \mathbf{t}^\top)} \tag{9}$$

$$= \sum_i \mathbf{v}_i \cdot \frac{\exp(\mathbf{v}_i \mathbf{t}^\top)}{\sum_{i'} \exp(\mathbf{v}_{i'} \mathbf{t}^\top)} = \mathrm{Attn}(\mathbf{t}, \mathbf{V}, \mathbf{V}) \tag{10}$$

Similarly, we can show

$$\mathbb{E}\left[\mathbf{A}|\mathbf{t}\right] = \mathrm{Attn}(\mathbf{t}, \mathbf{A}, \mathbf{A}) \tag{11}$$

Thus, with CSS given by (4), we obtain the conventional *cross attention* for video and audio with input text given by (6) and (11).

2) Cross-Reference CSS. The second type of CSS is

$$\bar{s}(i,j|\mathbf{t}) = \mathbf{a}_i \mathbf{t}^\top + \mathbf{v}_j \mathbf{t}^\top \tag{12}$$

Importantly, we use i-th *audio* frame $\mathbf{a}_i$ to represent the i-th *video* frame for similarity score. Similarly, we use j-th *video* frame $\mathbf{v}_j$ to represent j-th *audio* frame. The intuition is that, audio-visual questions are likely to address the *co-occurrence* of video and audio events [2, 34]. If the i-th frame contains an important clue to the audio-visual question, *both* the video and audio information at the i-th frame will be important. In that case, the video information at i-th frame can be closely related to i-th audio $\mathbf{a}_i$ given query $\mathbf{t}$.

The CSS matrix for (12) is simply the *transpose* of CSS matrix $\mathbf{S}$ in (2), because $\bar{s}(i,j|\mathbf{t}) = s(j,i|\mathbf{t})$. Thus, we will denote its CSS matrix by $\mathbf{S}^\top$. We obtain the softmax distribution by applying 2D-softmax to $\mathbf{S}^\top$ given by $P_{\mathbf{S}^\top}$. The expected representations of video and audio under this distribution can be derived similarly to (10):

$$\mathbb{E}\left[\mathbf{V}|\mathbf{t}\right] = \mathrm{Attn}(\mathbf{t}, \mathbf{A}, \mathbf{V}) \tag{13}$$

and

$$\mathbb{E}\left[\mathbf{A}|\mathbf{t}\right] = \mathrm{Attn}(\mathbf{t}, \mathbf{V}, \mathbf{A}) \tag{14}$$

The attentions in (13) and (14) may appear somewhat unconventional; three different modalities are input to Q, K and V connections of the attention function. However, we can interpret (13) as follows. The attention weights are derived from text and audio ($\mathbf{t}$ and $\mathbf{A}$ are connected to Q and K of the attention function). The weights "select" important audio frames given the question. These weights are used to combine video frames. Thus, $\mathrm{Attn}(\mathbf{t}, \mathbf{A}, \mathbf{V})$ is a combination of video frames where the frames are considered important from the perspective of audio and text. We can expect these representations are useful for audio-visual queries. We call (13) and (14) *Tri-Modal* attention.

3) Mixed CSS. The third type of CSS we use is

$$\hat{s}(i,j|\mathbf{t}) = \max(\mathbf{v}_i\mathbf{t}^\top + \mathbf{a}_j\mathbf{t}^\top, \mathbf{a}_i\mathbf{t}^\top + \mathbf{v}_j\mathbf{t}^\top) \tag{15}$$

(15) can be considered as *mixing* the previous CSSs (4) and (12) by taking their maximum. We chose the max function for mixing scores due the following reasons. The max function is widely adopted to combine or pool feature scores, e.g., max pooling [14] (keeps the most salient activation) or maxout [9] (outputs the maximum over multiple linear responses). In particular, the max function is effective for suppressing irrelevant noise. For example, suppose one of feature score is high (large positive number) and the other is low (large negative number). Then this feature is potentially important, similar to a neuron output with a high activation value. However, if we mix the scores using sum or product, the irrelevant score can affect the result. Thus, the max operation is good at suppressing such irrelevant noise.

An issue with the max function is that it is not a smooth function. Instead, we use Log-Sum-Exp [4] as a smooth approximation of the max function:

$$\max(a,b) \approx \log(\exp(a) + \exp(b)) \tag{16}$$

If we apply (16) to CSS (15) and compute the expectations of video and audio representations, we obtain the following (a derivation is provided in the Supplementary Materials):

$$\mathbb{E}[\mathbf{V}|\mathbf{t}] \approx \tfrac{1}{2}\,\mathrm{Attn}(\mathbf{t},\mathbf{A},\mathbf{V}) + \tfrac{1}{2}\,\mathrm{Attn}(\mathbf{t},\mathbf{V},\mathbf{V}) \tag{17}$$

and

$$\mathbb{E}[\mathbf{A}|\mathbf{t}] \approx \tfrac{1}{2}\,\mathrm{Attn}(\mathbf{t},\mathbf{V},\mathbf{A}) + \tfrac{1}{2}\,\mathrm{Attn}(\mathbf{t},\mathbf{A},\mathbf{A}). \tag{18}$$

Interestingly, these approximate representations can be obtained from the *mixture* of softmax distributions derived from previous CSSs. For example, (17) is equal to

$$\sum_{i,j}\left(\frac{P_{\mathbf{S}}(i,j|\mathbf{t}) + P_{\mathbf{S}^\top}(i,j|\mathbf{t})}{2}\right)\mathbf{v}_i$$

which means that (17) is the expectation with respect to the mixture of distributions $P_{\mathbf{S}}$ and $P_{\mathbf{S}^\top}$.

Implementation: Modality Mixture. We obtained representations which are cross attentions ($\mathrm{Attn}(\mathbf{t},\mathbf{V},\mathbf{V})$ and $\mathrm{Attn}(\mathbf{t},\mathbf{A},\mathbf{A})$ in (6) and (11)), Tri-Modal Attentions ($\mathrm{Attn}(\mathbf{t},\mathbf{A},\mathbf{V})$ and $\mathrm{Attn}(\mathbf{t},\mathbf{V},\mathbf{A})$ in (13) and (14)), and their mixtures ((17) and (18)). Thus, these attentions constitute the main components of the proposed framework for modality fusion. For practical implementation, we combine these representations as follows: (i) compute the mixture weights for the four attention representations; (ii) concatenate the weighted attention representations. The concatenated representation will be used for the final reasoning of the answer. This approach is consistent with our horizontal fusion of modalities; we compute four types of attentions directly from audio, video and text representations, and merge them in parallel without deep sequential stacking.

Denote these cross- and Tri-Modal attention outputs by $r_1 = \mathrm{Attn}(\mathbf{t}, \mathbf{V}^p, \mathbf{V}^p)$, $r_2 = \mathrm{Attn}(\mathbf{t}, \mathbf{A}, \mathbf{A})$, $r_3 = \mathrm{Attn}(\mathbf{t}, \mathbf{A}, \mathbf{V}^p)$ and $r_4 = \mathrm{Attn}(\mathbf{t}, \mathbf{V}, \mathbf{A})$, respectively. Note that we used patch-level representations of video ($\mathbf{V}^p$) for the Value input to the attentions to better exploit spatial information. The attention outputs are fused using text-dependent weights. A precise weighting scheme is described as follows. For stability of those features, we modulate attention keys and values by text using FiLM [28]:

$$\begin{aligned} \mathrm{FiLM}(X; \mathbf{t}) \;&=\; X \odot \gamma(\mathbf{t}) + \beta(\mathbf{t}) \\ \gamma(\mathbf{t}) = \mathbf{1} + W_\gamma \mathbf{t}, \;\; &\beta(t) = W_\beta \mathbf{t}, \end{aligned}$$

(zero-initialized so initially identity). For r_1, r_2, r_3, and r_4, we use

$$K' = \mathrm{FiLM}(K; \mathbf{t}), \qquad V' = \mathrm{FiLM}(V; \mathbf{t}),$$

and apply attention with (K', V') in place of (K, V) inside Attention Module. From t we produce four scalars

$$(q_1, q_2, q_3, q_4) = W_g \mathbf{t} + b_g, \qquad g_i \;=\; 1 + \tfrac{1}{2} \tanh(q_i).$$

Each attention is gated by g_i:

$$h_i \;=\; g_i \, r_i, \qquad i = 1, \ldots, 4.$$

This approach allows the question $\mathbf{t}$ to determine how much each attention result r_i should matter by assigning weights g_i that control their contribution.

3.4 Reasoning

For the reasoning of the answer, we begin with concatenating the text embedding and four gated attention representations:

$$\mathbf{H} \;=\; \big[\mathbf{t} \;;\; h_1 \;;\; h_2 \;;\; h_3 \;;\; h_4\big] \in \mathbb{R}^{5 \times D}$$

Next, we apply a self-attention block to allow global interaction among all results:

$$\mathbf{F} \;=\; \mathrm{SA}(\mathbf{H}, \mathbf{H}, \mathbf{H}) \in \mathbb{R}^{5 \times D}$$

Finally, the result passes through a classification layer to produce the answer:

$$\hat{\mathbf{y}} \;=\; \mathrm{CLS}(\mathbf{F}) \;\in\; \mathbb{R}^C$$

where CLS denotes the classifier layer and C denotes the number of classes. The model is trained by minimizing the cross-entropy between the predicted label $\hat{\mathbf{y}}$ and the ground-truth label $\mathbf{y}$:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c \,. \tag{19}$$

Table 1: Dataset statistics.

Dataset	Videos	Train	Valid QA	Test QA
MUSIC-AVQA	9,288	31,927	4,568	9,129
MUSIC-AVQA-R	9,288	–	–	211,572
MUSIC-AVQA-v2.0	10,512	37,429	5,345	10,816

4 Experiments

4.1 Datasets

We evaluate on MUSIC-AVQA [19], MUSIC-AVQA-R [27], and MUSIC-AVQA-v2.0 [23], adopting the official train/val/test splits as in Table 1. The base MUSIC-AVQA [19] provides a standard benchmark over 22 instruments with audio-only, visual-only, and audio-visual questions. MUSIC-AVQA-R [27] builds on MUSIC-AVQA and upgrades the setting by concentrating on rare and out-of-distribution cases (e.g., uncommon instruments and atypical contexts), explicitly probing robustness beyond head-class priors and frequency-driven shortcuts. MUSIC-AVQA-v2.0 [23] upgrades the benchmark to address dataset biases by enriching ensemble and multi-instrument scenes, rebalancing question types (e.g., existence, location, temporal) across modalities. It provides paired bias-balanced evaluation splits, which enable controlled analysis of spurious correlations versus genuine cross-modal reasoning.

4.2 Implementation Details

We extract common-space representations by using CLIP [29] for video and text embeddings, and ImageBind [8] for audio embeddings. Video frames are sampled every 2 seconds with the hidden size given by $D = 1024$, and we obtain one CLIP embedding per frame. However, this frame-wise CLIP encoding may not fully represent spatial information. Thus, we additionally extract CLIP-based patch-level embeddings for video: for each frame, $P = 256$ patch tokens with dimensionality 1280. All attention modules use 8 heads. Dropout 0.1 is applied to every attention module and to the final classification layer. Optimization uses AdamW [24] with a step scheduler that multiplies the learning rate by 0.1 every 8 epochs. The loss employs label smoothing [33] 0.1. We apply gradient clipping of 0.1 (global norm). Training is performed on a single NVIDIA A6000 GPU.

4.3 Quantitative Results and Analysis

MUSIC-AVQA-R. On the debiased MUSIC-AVQA-R [27] benchmark, our model attains 70.89% overall as in Table 2. It performs strongly on both frequent (Head) and rare (Tail) answer categories, e.g., Audio QA Count (H/T: 85.68/76.12%) and Visual QA Local (H/T: 89.45/78.95%). For Audio-Visual QA, Count (H/T: 81.43/44.21%) and Local (H/T: 66.27/63.51%) remain competitive, suggesting robust question-conditioned temporal–spatial reasoning even under long-tail distributions.

Table 2: Experimental results (%) on the MUSIC-AVQA-R test set, with H and T representing performance on Head (frequent) and Tail (rare) answer categories, respectively. **Bold** denotes the best performance, and underline indicates the second-best performance.

Method	Audio QA				Visual QA				Audio-Visual QA										Avg
	Count		Comp		Count		Local		Exist		Count		Local		Comp		Temp		
	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T	
FCNLSTM [7]	66.23	36.48	64.78	51.24	61.75	5.31	54.86	51.06	64.76	78.52	62.69	7.23	46.66	57.30	43.13	71.67	37.02	30.78	54.12
BiLSTM [41]	73.68	46.32	21.51	**77.58**	64.30	0.00	53.92	42.01	**87.51**	21.14	62.85	2.18	35.16	43.75	27.61	74.38	17.58	31.32	48.84
HCAttn [26]	61.67	41.63	59.09	47.14	56.52	9.20	67.01	53.16	66.57	61.13	59.53	12.48	37.05	42.48	48.81	60.12	33.82	39.26	51.90
MCAN [39]	75.02	60.16	58.89	50.90	64.58	26.69	66.62	65.25	61.29	67.29	64.76	25.28	46.11	61.61	50.57	52.04	34.64	58.05	57.27
PSAC [20]	53.01	56.68	57.41	48.12	49.55	26.43	72.96	60.69	50.56	55.54	56.70	19.58	41.98	52.30	38.13	58.92	26.68	46.24	50.45
HME [6]	62.60	53.95	54.97	58.29	50.95	16.46	73.25	58.06	65.74	66.49	63.18	17.18	39.59	45.03	53.20	69.57	39.35	41.57	53.66
HCRN [16]	55.53	53.31	47.17	32.44	41.87	23.55	39.40	51.27	41.81	65.45	54.58	19.57	36.62	42.72	33.33	36.87	40.47	44.13	43.92
AVSD [32]	54.00	47.84	60.61	47.79	63.25	10.07	74.70	61.43	66.28	61.98	42.61	18.06	33.00	40.35	51.98	66.00	40.14	41.52	52.33
Pano-AVQA [40]	50.57	43.45	50.78	43.64	47.28	15.50	76.19	65.51	52.37	22.04	52.21	21.52	44.35	61.69	45.61	40.49	40.05	39.43	47.40
ST-AVQA [19]	56.40	41.48	62.22	55.79	59.86	12.94	63.31	54.00	49.78	56.41	38.11	19.41	35.05	44.09	53.30	62.44	40.42	35.85	52.80
LAVISH [22]	61.73	43.99	66.00	50.38	55.33	10.21	70.71	68.78	77.83	79.46	49.38	14.47	41.76	41.01	59.26	65.10	41.84	46.26	57.63
TSPM [17]	81.65	71.80	67.66	49.56	78.29	47.56	58.80	73.18	69.15	**82.79**	77.09	38.64	42.24	57.37	52.07	68.86	39.23	49.36	66.30
MCCD [27]	84.32	67.23	64.68	62.18	75.09	48.42	80.47	66.38	77.22	67.58	55.15	**82.23**	**70.12**	39.83	61.26	58.17	43.67	58.33	66.95
QA-Tiger [13]	82.67	75.82	**71.75**	43.11	81.30	54.59	84.76	75.59	72.84	78.56	76.70	33.55	64.65	**64.65**	37.55	**80.47**	36.85	**62.96**	67.99
Q-TriM	**85.68**	**76.12**	55.22	63.54	**86.26**	**57.91**	**89.45**	**78.95**	72.29	81.97	**81.43**	44.21	66.27	61.35	**61.35**	56.67	**43.94**	48.26	**70.89**

Table 3: Experimental results (%) on the MUSIC-AVQA test set.

Method	Audio QA			Visual QA			Audio-Visual QA						Avg
	Count	Comp	Avg	Count	Local	Avg	Exist	Count	Local	Comp	Temp	Avg	
FCNLSTM [7]	70.45	66.22	68.88	63.89	46.74	55.21	82.01	59.34	46.28	62.15	47.33	60.06	60.34
BiLSTM [41]	70.35	47.92	62.05	64.64	64.33	64.48	78.39	56.91	45.85	53.09	49.76	57.10	59.92
HCAttn [26]	70.25	54.91	64.57	64.05	66.37	65.22	79.10	59.97	49.51	55.25	56.43	60.19	62.30
MCAN [39]	77.50	55.24	69.25	71.56	70.93	71.24	80.40	64.91	54.48	57.22	47.57	61.58	65.49
PSAC [20]	75.64	66.06	72.09	68.64	69.79	69.22	77.59	63.42	55.02	61.17	59.47	63.52	66.54
HME [6]	74.76	63.56	70.61	67.97	69.46	68.76	80.30	63.19	53.18	62.69	59.83	64.05	66.45
HCRN [16]	68.59	50.92	62.05	64.39	61.81	63.08	54.47	53.38	41.53	52.11	47.69	50.26	55.73
AVSD [32]	72.41	61.90	68.52	67.39	74.19	70.83	81.61	63.89	58.79	61.52	61.41	65.49	67.44
Pano-AVQA [40]	74.36	64.56	70.73	69.39	75.65	72.56	81.21	64.91	59.33	64.22	63.23	66.64	68.93
ST-AVQA [19]	78.18	67.05	74.06	71.56	76.38	74.00	81.81	70.80	64.51	66.01	63.23	69.54	71.52
MCCD [27]	83.87	**71.04**	**79.14**	79.78	76.73	78.24	80.87	51.63	71.46	64.67	64.60	67.13	72.20
COCA [15]	79.35	67.68	75.42	75.10	75.43	75.23	**83.50**	66.63	69.72	64.12	65.57	69.96	72.33
PSTP-Net [18]	73.97	65.59	70.91	77.15	77.36	77.26	76.18	72.23	71.80	**71.79**	69.00	72.57	73.52
LAVISH [22]	82.09	65.56	75.97	78.98	81.43	80.22	81.71	75.51	66.13	63.77	67.96	71.26	74.46
APL [21]	82.40	70.71	78.09	76.52	82.74	79.69	82.99	73.29	66.68	64.76	65.95	70.96	74.53
TSPM [17]	84.07	64.65	76.91	82.29	84.90	83.61	82.19	76.21	71.85	65.76	**71.17**	73.51	76.79
QA-Tiger [13]	84.86	67.85	78.58	83.96	86.29	85.14	83.10	78.58	**72.50**	63.94	69.59	73.74	77.62
Q-TriM	**85.05**	65.49	77.84	**84.46**	**87.02**	**85.76**	82.29	**80.87**	71.85	62.58	69.83	**73.78**	**77.68**

MUSIC-AVQA. Our model achieves an overall accuracy of 77.68% on the MUSIC-AVQA [19] test set, surpassing prior work such as QA-TIGER (77.62%) as shown in Table 3. By modality, we obtain a strong Visual QA average of 85.76% (Count 84.46%, Local 87.02%) and an Audio-Visual QA average of 73.78% (Exist 82.29%, Count 80.87%, Local 71.85%, Comp 62.58%, Temp 69.83%). Audio QA averages 77.84%.

MUSIC-AVQA-v2.0. We further evaluate on the MUSIC-AVQA-v2.0 [23] benchmark. When trained on the biased dataset, Q-TriM achieves state-of-the-art performance on both the bias test (78.40%) and the balanced test (75.31%). When trained on the balanced dataset, it remains state-of-the-art performance on the bias test (76.92%), but on the balanced test it falls short by a small margin (76.29%), as shown in Table 4.

Overall, Q-TriM attains strong results across all three benchmarks. This suggests that adopting a shallow and parallel design, rather than deep and sequen-

Table 4: Experimental results (%) on the MUSIC-AVQA-v2.0 for (a) bias and (b) balanced test sets. Each evaluation is conducted with models trained on bias and balanced training sets, respectively.

Training	Method	A-QA	V-QA	AV-QA	Avg
Bias	ST-AVQA [19]	76.86	77.70	69.59	73.07
	LAVISH [22]	76.73	80.96	70.80	74.59
	QA-TIGER [13]	**79.13**	84.83	72.37	76.93
	Ours	78.93	**86.01**	**74.53**	**78.40**
Balance	ST-AVQA [19]	76.18	77.20	67.96	71.92
	LAVISH [22]	75.56	80.83	69.27	73.51
	LAST [23]	77.10	82.99	70.86	75.24
	LAST-Att [23]	77.29	83.47	71.05	75.45
	QA-TIGER	77.07	85.93	71.20	76.57
	Ours	**78.75**	**86.45**	**71.69**	**76.92**

(a) Bias test set

Training	Method	A-QA	V-QA	AV-QA	Avg
Bias	ST-AVQA [19]	73.34	76.82	64.51	69.40
	LAVISH [22]	73.14	79.70	65.01	70.39
	QA-TIGER [13]	77.57	84.84	67.43	73.91
	Ours	**77.67**	**85.05**	**69.85**	**75.31**
Balance	ST-AVQA [19]	75.50	77.67	66.32	71.02
	LAVISH [22]	76.15	81.32	68.28	73.18
	LAST [23]	78.08	83.29	69.72	74.85
	LAST-Att [23]	78.56	84.07	**70.30**	75.44
	QA-TIGER [13]	79.90	**86.95**	70.22	**76.43**
	Ours	**80.14**	86.63	70.03	76.29

(b) Balance test set

Table 5: Ablation study of the proposed framework on MUSIC-AVQA-R dataset.

Method	A-QA	V-QA	AV-QA	Avg
w/o All Modules(Base)	70.48	80.83	58.43	66.96
w/o Token Filtering	74.54	80.36	61.54	69.21
w/o Tri-Modal Attention	73.13	81.02	61.10	68.92
w/o Modality Mixture	72.08	78.21	60.96	67.85
Ours	**74.74**	**83.64**	**62.85**	**70.89**

tial, for inter-modal attention can be effective for learning cross-modal interactions compared to deeper stacked architectures as in baseline methods.

4.4 Ablation Study

We conduct ablations to assess the contribution of each component of the proposed model (Table 5).

Token filtering. We first removed the token filtering component. Token filtering not only reduces computational cost, but also suppresses patches that are irrelevant to the question, thereby easing question grounding and answer selection. Without this module, the model attends more to background regions and exhibits degraded accuracy.

Tri-Modal attention. Next, we removed the Tri-Modal attention. This module is particularly important for audio-visual question types in AVQA that require reasoning over all three modalities. Eliminating it forces the model to rely on cross attention layers only, which weakens cross-modal alignment and lowers performance.

Modality mixture. Next, we removed the modality mixture component. This module aggregates the text-conditioned attention outputs from different modalities into a coherent representation. Without this component, the model struggles to reconcile complementary cues across modalities, leading to a further drop in accuracy.

Table 6: Evaluation of token filtering with varying keep ratios on MUSIC-AVQA-R. Consistent with prior findings, a keep ratio of 0.5 achieves the highest accuracy.

Keep Ratio	A-QA	V-QA	AV-QA	Avg
0.1	75.28	82.60	61.79	70.11
0.3	75.30	83.05	62.37	70.56
0.5(ours)	74.74	**83.64**	**62.85**	**70.89**
0.7	74.28	82.71	62.56	70.39
0.9	**75.35**	83.22	62.15	70.50

4.5 Keep ratio of Token Filtering

We conducted token filtering experiments to determine how many patches to retain, controlled by the hyperparameter k (Table 6). Using the *MUSIC-AVQA-R* dataset, we evaluated five values of the keep ratio: $0.1, 0.3, 0.5, 0.7$, and 0.9. Here, a keep ratio of 0.5 means that only 50% of the total patches are kept while the remaining ones are discarded. Consistent with the paper's findings, a keep ratio of 0.5 yielded the best performance. Notably, when the keep ratio was set to 0.1. Thus, by using only 10% of the original patches, the model still achieved a reasonable 70.11% accuracy, indicating that token filtering effectively compresses patches. This suggests that, even without using attention in the filtering stage, the token-filtering module is effectively learned via simple scoring and a straight-through estimator (STE) [3] in order to retain only the question-relevant tokens. In addition, we explored a strategy that makes the keep-ratio adaptive, where the model dynamically selects how many video patches to retain conditioned on the question. Related discussion and experiments are provided in the Supplementary Materials.

5 Conclusion

We introduce Q-TriM, an AVQA model that replaces deeply stacked attention with a shallow, parallel inter-modal and question-aware fusion. The pipeline consists of three stages. First, we apply token filtering to retain only the semantically meaningful patches that are relevant to the given question. Next, Tri-Modal Attention and cross attention are applied in parallel to obtain text-relevant video and audio representations. Finally, a modality mixture head aggregates features with question-aware weights, yielding a compact representation that emphasizes information most pertinent to the query. Q-TriM attains state-of-the-art performance on three benchmark datasets. Especially, we demonstrate the robustness of our model by achieving a substantial performance margin over prior state-of-the-art methods on the MUSIC-AVQA-R dataset. As a direction for future work, we plan to explore various CSS scores in the proposed attention framework and extend them to various multi-modal tasks.

Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2026-RS-2020-II201819, ICT Creative Consilience Program, 10%, and No.RS-2026-25507543, Development of AI Co-Scientist based on Scientific Causal World Model, 80%) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT)(RS-2022-NR070834,RS-2026-25482946).

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV. pp. 2425–2433 (2015)
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV. pp. 609–617 (2017)
3. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
4. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge university press (2004)
5. Datta, S., Dharur, S., Cartillier, V., Desai, R., Khanna, M., Batra, D., Parikh, D.: Episodic memory question answering. In: CVPR. pp. 19119–19128 (2022)
6. Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., Huang, H.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: CVPR. pp. 1999–2007 (2019)
7. Fayek, H.M., Johnson, J.: Temporal reasoning via audio question answering. TASLP **28**, 2283–2294 (2020)
8. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR. pp. 15180–15190 (2023)
9. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. In: ICML. pp. 1319–1327 (2013)
10. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: ICASSP. pp. 976–980 (2022)
11. Hayou, S., Doucet, A., Rousseau, J.: On the impact of the activation function on deep neural networks training. In: ICML. pp. 2672–2680 (2019)
12. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV. pp. 5492–5501 (2019)
13. Kim, H., Jung, I., Suh, D., Zhang, Y., Lee, S., Hong, S.: Question-aware gaussian experts for audio-visual question answering. In: CVPR. pp. 13681–13690 (2025)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS. pp. 1097–1105 (2012)
15. Lao, M., Pu, N., Liu, Y., He, K., Bakker, E.M., Lew, M.S.: Coca: Collaborative causal regularization for audio-visual question answering. In: AAAI. pp. 12995–13003 (2023)
16. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: CVPR. pp. 9972–9981 (2020)

17. Li, G., Du, H., Hu, D.: Boosting audio visual question answering via key semantic-aware cues. In: ACM MM. pp. 5997–6005 (2024)
18. Li, G., Hou, W., Hu, D.: Progressive spatio-temporal perception for audio-visual question answering. In: ACM MM. pp. 7808–7816 (2023)
19. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: CVPR. pp. 19108–19118 (2022)
20. Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C.: Beyond rnns: Positional self-attention with co-attention for video question answering. In: AAAI. pp. 8658–8665 (2019)
21. Li, Z., Guo, D., Zhou, J., Zhang, J., Wang, M.: Object-aware adaptive-positivity learning for audio-visual question answering. In: AAAI. pp. 3306–3314 (2024)
22. Lin, Y., Sung, Y., Lei, J., Bansal, M., Bertasius, G.: Vision transformers are parameter-efficient audio-visual learners. In: CVPR. pp. 2299–2309 (2023)
23. Liu, X., Dong, Z., Zhang, P.: Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In: WACV. pp. 4478–4487 (2024)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
25. Lou, Y., Mingard, C.E., Hayou, S.: Feature learning and signal propagation in deep neural networks. In: ICML. pp. 14248–14282 (2022)
26. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: NeurIPS. pp. 289–297 (2016)
27. Ma, J., et al.: Look, listen, and answer: Overcoming biases for audio-visual question answering. In: NeurIPS. pp. 1–25 (2024)
28. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
30. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. In: NeurIPS. pp. 13937–13949 (2021)
31. Schoenholz, S.S., Gilmer, J., Ganguli, S., Sohl-Dickstein, J.: Deep information propagation. In: ICLR (2017)
32. Schwartz, I., Schwing, A.G., Hazan, T.: A simple baseline for audio-visual scene-aware dialog. In: CVPR. pp. 12548–12558 (2019)
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
34. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV. pp. 247–263 (2018)
35. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. In: Allerton. pp. 368–377 (1999)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
37. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: Avqa: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia (ACM MM). pp. 3480–3491 (2022)
38. Yin, H., Vahdat, A., Alvarez, J.M., Mallya, A., Kautz, J., Molchanov, P.: A-vit: Adaptive tokens for efficient vision transformer. In: CVPR. pp. 10809–10818 (2022)
39. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: CVPR. pp. 6281–6290 (2019)
40. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-avqa: Grounded audio-visual question answering on 360deg videos. In: ICCV. pp. 2031–2041 (October 2021)

41. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: ACL. pp. 207–212 (2016)