

AVQACL: A Novel Benchmark for Audio-Visual Question Answering Continual Learning

Kaixuan Wu^{1,4}, Xinde Li^{1,2,3,4,*}, Xinling Li^{1,4}, Chuanfei Hu^{1,2,4}, Guoliang Wu^{1,4}

¹School of Automation, Southeast University, Nanjing, China

²Nanjing Center for Applied Mathematics, Nanjing, China

³Southeast University Shenzhen Research Institute, Shenzhen, China

⁴Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China

{kaixuanwu, xindeli, xinlingli, chuanfeihu, guoliangwu}@seu.edu.cn

Abstract

In this paper, a novel benchmark for audio-visual question answering continual learning (AVQACL) is introduced, aiming to study fine-grained scene understanding and spatial-temporal reasoning in videos under a continual learning setting. To facilitate this multimodal continual learning task, we create two audio-visual question answering continual learning datasets, named *Split-AVQA* and *Split-MUSIC-AVQA* based on the AVQA and MUSIC-AVQA datasets, respectively. The experimental results suggest that the model exhibits limited cognitive and reasoning abilities and experiences catastrophic forgetting when processing three modalities simultaneously in a continuous data stream. To address above challenges, we propose a novel continual learning method that incorporates question-guided cross-modal information fusion (QCIF) to focus on question-relevant details for improved feature representation and task-specific knowledge distillation with spatial-temporal feature constraints (TKD-STFC) to preserve the spatial-temporal reasoning knowledge acquired from previous dynamic scenarios. Furthermore, a question semantic consistency constraint (QSCC) is employed to ensure that the model maintains a consistent understanding of question semantics across tasks throughout the continual learning process. Extensive experimental results on *Split-AVQA* and *Split-MUSIC-AVQA* datasets illustrate that our method achieves state-of-the-art audio-visual question answering continual learning performance. The code is available at <https://github.com/kx-wu/AVQACL>.

1. Introduction

Audio-Visual Question Answering (AVQA) is a complex AI task requiring systems to interpret and answer questions based on both audio and visual information from videos

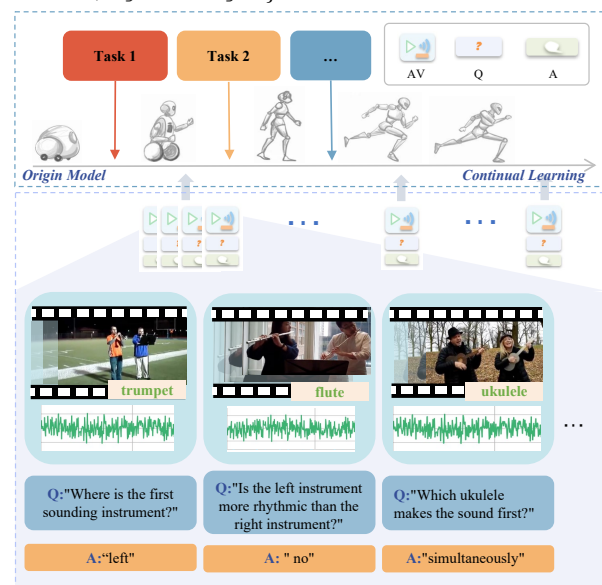


Figure 1. The illustration of an AVQA system under a continual learning setting.

[20, 39]. Unlike single-modality tasks [7, 13], AVQA integrates both auditory and visual cues, facilitating a more comprehensive understanding of scenes. This capability makes AVQA particularly crucial for applications such as intelligent assistants, autonomous driving, and embodied intelligence, where seamless multimodal integration is essential [10, 24, 38, 44].

However, most AVQA methods [21, 26, 40, 41] assume a static learning environment, training on fixed datasets without the capacity to adapt to new tasks or data over time. This limitation is problematic in dynamic real-world scenarios, where data is constantly evolving. Without adaptability, models risk becoming outdated and impractical to retrain from scratch due to high computational costs and overfitting risks. To address these issues, multi-modal continual learning (MCL) has emerged as a key approach. MCL combines

*Corresponding author

diverse data sources, such as visual, auditory, and textual information, while retaining learned knowledge across tasks [35, 43, 46]. By preventing catastrophic forgetting, MCL enables AVQA systems to adapt to new tasks while preserving prior knowledge, enhancing resilience and flexibility in dynamic environments.

To bridge the gaps between MCL and AVQA, We present AVQACL, a benchmark based on Split-AVQA and Split-MUSIC-AVQA datasets to study the challenge of learning multimodal tasks in a continual learning setting. In this setting illustrated in Fig. 1, the model initially begins with an origin model trained on Task 1 and subsequently encounters new tasks, each requiring it to learn new questions, audio-visual categories, and answer updates in a sequential manner. As the model progresses through tasks, it continuously learns new audio-visual content, adapts to evolving questions, and provides accurate answers without forgetting previously acquired knowledge. Each example displays a specific question, associated audio-visual content, and the corresponding answer, emphasizing the model’s ability to adapt to diverse and evolving question-answer pairs through continual learning. We repurpose several baselines [2, 4, 18, 25, 30, 32] and introduce network for AVQA continual learning research. It is promising that the proposed method could foster research on AVQACL with fine-grained scene understanding and spatial-temporal reasoning.

Through the development of the AVQACL benchmark and analysis of our experimental results, we have identified several unique challenges within AVQACL. (1) In audio-visual question answering tasks, different questions often require attention to specific details within multi-modal data. However, during continual learning, models can become distracted by irrelevant information, leading to a lack of focus on question-relevant details and compromising the quality of feature representations. (2) AVQA tasks frequently involve complex spatial-temporal information, making models vulnerable to catastrophic forgetting of previously acquired spatial-temporal reasoning knowledge such as scene transitions and event sequences when learning new tasks. This forgetting effect reduces the model’s spatial-temporal reasoning ability, weakening its performance in responding to questions within dynamic scenes. (3) In continual learning settings, different question types may share similar semantic patterns or structures. When learning new tasks, models may drift from previously learned question semantics, resulting in a decline in semantic consistency.

Inspired by the challenges in AVQACL, we propose a novel method comprising three key components: Question-guided Cross-modal Information Fusion (QCIF), Task-specific Knowledge Distillation with Spatial-Temporal Feature Constraints (TKD-STFC), and Question Semantic

Consistency Constraint (QSCC). The QCIF module leverages the question as a guiding query to selectively extract visual and audio features, utilizing spatial-temporal attention to enhance focus on question-relevant information within each modality. TKD-STFC addresses the challenge of preserving critical knowledge from prior tasks while learning new information in a continual learning framework; it enforces spatial-temporal consistency by distilling knowledge from previous tasks and minimizing feature discrepancies between the old and new models. Finally, QSCC maintains consistent semantic representation of the question across tasks, mitigating catastrophic forgetting and enhancing stability in question understanding as tasks evolve. Together, these components enable our model to perform robustly in cross-modal fusion, spatial-temporal reasoning, and semantic stability, advancing the effectiveness of continual learning in audio-visual question answering. We conduct experiments on two newly created datasets to assess the effectiveness of our method in AVQACL. Experimental results demonstrate that our proposed method achieves state-of-the-art performance in audio-visual question answering continual learning on both datasets.

In summary, our work makes three key contributions: (1) We introduce a novel benchmark based on Split-AVQA and Split-MUSIC-AVQA to bridge multi-modal continual learning and audio-visual question answering. This benchmark tackles challenges in fine-grained scene understanding and spatial-temporal reasoning, advancing research in audio-visual question answering for continual learning. (2) We propose a novel approach for AVQACL that integrates question-guided cross-modal information fusion to emphasize the most relevant details for better feature representation. Additionally, we incorporate task-specific knowledge distillation with spatial-temporal feature constraints to retain spatial-temporal reasoning knowledge from previous dynamic scenarios. Furthermore, we apply a question semantic consistency constraint, ensuring that the model consistently understands the semantics of the question across different tasks throughout the continual learning process. (3) Extensive experiments including six re-purposed, representative continual learning methods on the Split-AVQA and Split-MUSIC-AVQA datasets, demonstrate that our method achieves the state-of-the-art performance in audio-visual question answering for continual learning, underscoring the effectiveness of our approach.

2. Related works

2.1. Audio-Visual Question Answering

The field of Audio-Visual Question Answering (AVQA) has rapidly evolved, driven by the need for AI systems that can effectively integrate auditory and visual information to understand complex video content [1, 3, 21, 26, 31]. Re-

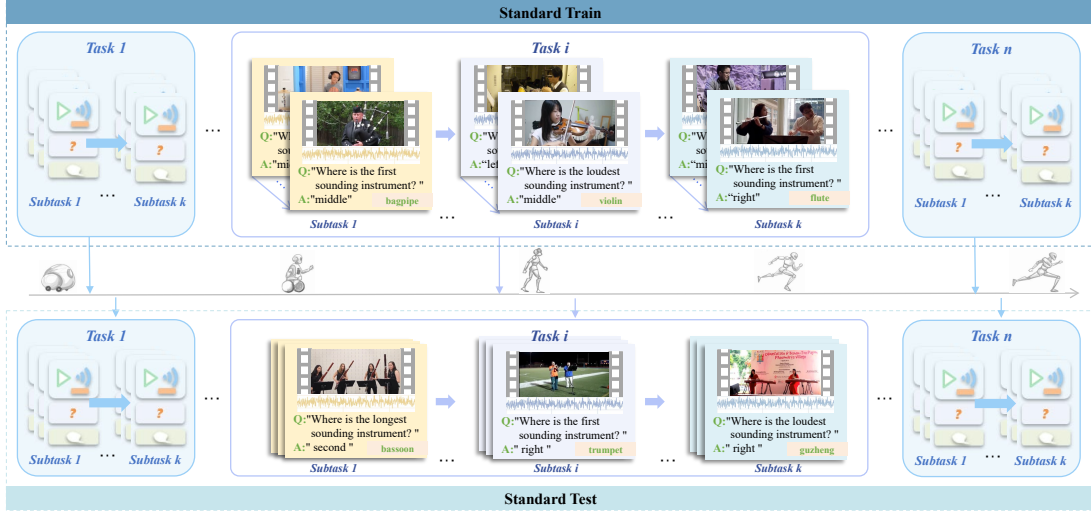


Figure 2. The proposed AVQACL setting, which is divided into Standard Train and Standard Test phases.

cently, audio-visual question answering (AVQA) has advanced through new datasets and models designed for complex cross-modal reasoning. Several methods [20, 39] then focused on integrating audio-visual cues in real-world and musical contexts, respectively. These datasets spurred the development of architectures for multi-level reasoning and advanced frameworks [21], enhancing alignment and attention-based spatial-temporal selection. Together, these developments illustrate the field’s progress toward a robust understanding of complex multimodal scenarios, setting a foundation for future AVQA applications that demand sophisticated reasoning capabilities and real-world generalizability [12, 19, 22, 40]. This work investigates audio-visual question answering in a continual learning setting, where the model incrementally learns from new tasks.

2.2. Unimodal Continual Learning

Unimodal Continual Learning (UCL) has seen various methods aimed at addressing the challenge of retaining knowledge while learning new tasks [23, 37]. Parameter Regularization method [4, 5, 18, 25], involves evaluating the importance of model parameters, assigning higher importance to certain weights to protect them during updates. This enables selective updating, allowing less important parameters to be modified while preserving the core knowledge embedded in significant parameters. Knowledge Distillation [9, 11, 33, 42] is another commonly used method, which helps to maintain previously learned information by aligning the current model’s outputs or feature representations with those of an earlier version. This alignment can take the form of minimizing distances between feature spaces or reducing divergence between probability distributions, effectively retaining learned knowledge during continual learning. In Memory Replay methods

[2, 6, 8, 27, 32], a small buffer stores samples from prior tasks, which the model replays to prevent forgetting.

2.3. Multimodal Continual Learning

Traditional continual learning primarily addresses unimodal data, such as images, with models learning incrementally across tasks. Extending this to multimodal data, particularly audio-visual, introduces added complexity due to the synchronization and cross-modal dependencies. A major challenge in multimodal continual learning (MCL) is retaining knowledge from both modalities while adapting to new tasks, as catastrophic forgetting is exacerbated by these intricate interdependencies [28, 35, 46]. To tackle this, recent AV-CIL [30] approaches include dual audio-visual similarity constraints and visual attention distillation for maintaining cross-modal consistency. The CIGN [28] uses audio-visual class tokens and distillation to preserve class-specific knowledge, while the HAD [46] framework leverages segment-wise augmentations to retain both low and high level features. And the CavRL method [45] further improves transferability with memory buffers and contrastive distillation. Video-based benchmarks like vCLIMB [35] and PIVOT [36] add temporal consistency, with PIVOT using prompt mechanisms for video tasks. The VQACL [43] framework introduces a dual-level structure to manage evolving questions and visual content, supporting generalization in dynamic multimedia settings.

3. AVQA Continual Learning Benchmark

3.1. Problem Definition

In the AVQACL learning tasks $T = \{T_1, T_2, \dots, T_n\}$, our goal is to train a model F_θ with parameters θ that can effectively extract relevant information from audio x^a , visual x^v , and question text x^q modalities within a contin-

Table 1. AVQACL dataset statistics.

Dataset	Split-AVQA	Split-MUSIC-AVQA
Scene type	Real-life	Music
Video length	10s	60s
# Videos	39,314	7,324
# QA pairs	41,008	20,774
# Question types	4	5
# Candidates	2135	40

uous video data stream to accurately predict the answer y across various types of questions. During the training process of AVQACL task T_t , where t denotes a incremental step and its corresponding training set can be denoted as $D_t = (x_{t,i}^a, x_{t,i}^v, x_{t,i}^q, y_{t,i})_{i=1}^{n_t}$, where $x_{t,i}^a$, $x_{t,i}^v$ and $x_{t,i}^q$ denote the i^{th} input sample’s audio, visual and question text modalities respectively in D_t , and $y_{t,i} \in A_t$ is the corresponding answer, where A_t denotes the label space of task T_t . In our settings, storing a small fixed size of exemplar memory set of data from previous answers is permitted according to [30], which is denoted as M_t . Therefore, all accessible training data at incremental step t can be denoted as $D_t^m = D_t \cup M_t$. Thus, the training process at incremental step t can be formulated as:

$$\Theta_t = \arg \min_{\Theta_{t-1}} (x) \mathbb{E}_{(x^a, x^v, x^q, y) \sim D_t^m} [\mathcal{L}(\mathcal{F}_{\Theta_{t-1}}(x^a, x^v, x^q), y)] \quad (1)$$

where $\mathcal{L}$ is the loss function between the model’s outputs and the corresponding answers. After the training process on D_t^m , the entire model is validated on the test set, which contains data belonging to all classes learned in previous tasks.

3.2. Datasets and Task Setting

Datasets. To benchmark this task and support our study, we selected two widely-used public datasets from the audio-visual question answering domain: AVQA [39] and MUSIC-AVQA [20], which focus on questions about visual objects, sounds, and their interactions. From these, we derived subsets named Split-AVQA and Split-MUSIC-AVQA to create the AVQACL dataset. As shown in Tab. 1, Split-AVQA consists of 39,314 ten-second videos and 41,008 question-answer pairs across four question types (“where,” “which,” “happening,” and “come from”) and 120 sound categories. To enhance the evaluation of reasoning in continual learning settings, we converted its 4-choice answering format to an open-ended one, requiring selection from 2,135 possible answers. Split-MUSIC-AVQA contains 7,324 sixty-second video clips and 20,774 question-answer pairs, representing diverse musical performances, including solos, homogeneous ensembles, and heterogeneous ensembles. It features five question types (“counting,” “location,” “comparative,” “temporal,” and “existential”) across 19 templates, with open-ended answers from a

set of 40 possibilities.

Tasks Setting. The goal of our proposed AVQACL benchmark is to enable models to integrate text, sound, and image data from the real world, allowing for cognitive processing and reasoning similar to human continual learning, while retaining previously learned knowledge as new information is acquired. Inspired by [43], the standard training and standard testing procedures for AVQACL are illustrated in the Fig. 2. In the standard training process, we construct a continuous data stream by nesting visual, audio, and question text data. First, we divide the training data into T tasks based on different question types. Then, within each language-driven task, we further divide the data into K subtasks according to the categories of objects and their sounds. The total number of categories in the dataset, C , is evenly divided into M groups. we define $T = 4$, $K=6$, $C=120$, $M=20$ for Split-AVQA and $T = 5$, $K=11$, $C=22$, $M=2$ for Split-MUSIC-AVQA.

3.3. Evaluation Metrics

AVQACL evaluates the continual learning models and learning algorithm with two standard continual learning metrics [30], Mean Accuracy (MA) and Average Forgetting (AF). MA is the average of the testing accuracy of each audio-visual question answering task, which can be denoted as:

$$MA = \frac{1}{T} \sum_{t=1}^T a_t \quad (2)$$

where a_t denotes testing accuracy of all seen classes after completing the training on task T_t .

AF is used to measure the extent of catastrophic forgetting over previously learned tasks, which can be denoted as:

$$AF = \frac{1}{T-1} \sum_{t=2}^T \left(\frac{1}{t-1} \sum_{i=1}^{t-1} \left[\max_{\gamma < t} (a_{\gamma,i} - a_{t,i}) \right] \right) \quad (3)$$

where T represents the total number of tasks, while t denotes the index of the current task, ranging from the 2nd task to the T^{th} task. $a_{\gamma,i}$ is the accuracy of the model on the i^{th} sample of task γ , and $a_{t,i}$ is the performance on the i^{th} sample during the current task t .

4. Proposed Method

4.1. Overall Architecture

The overall architecture of our proposed method is illustrated in Fig. 3, comprising a visual encoder, an audio encoder, a question encoder, a question-guided spatial-temporal attention module, a question-guided temporal attention module, and a classifier. The visual and audio encoders are frozen pre-trained models, while the question encoder is incrementally trained as new questions are added.

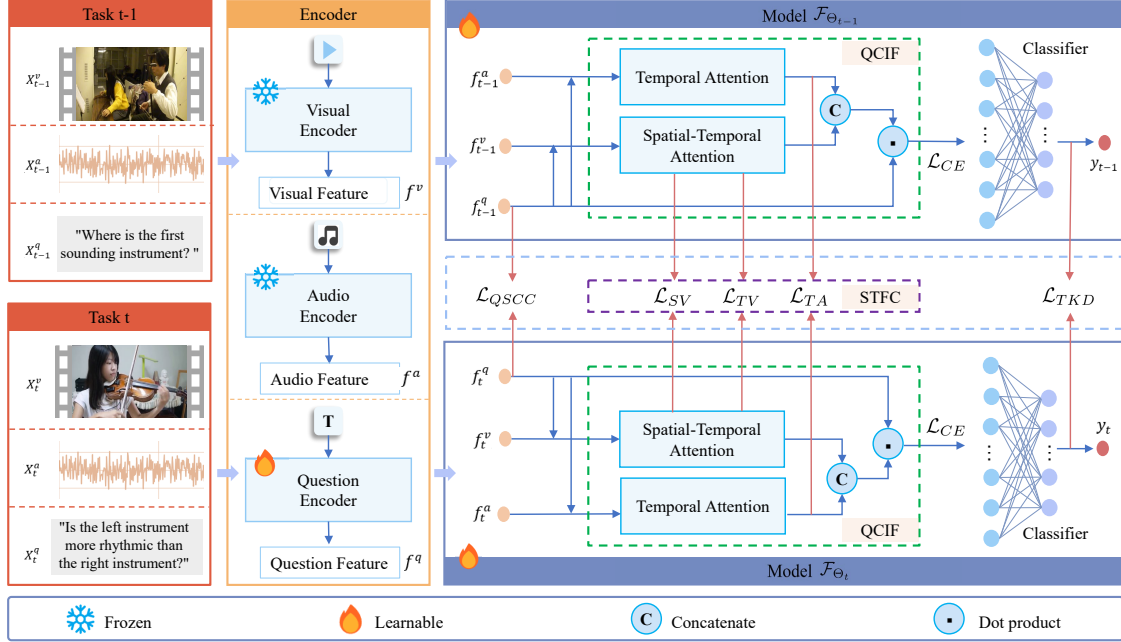


Figure 3. The overall architecture of our proposed method, which mainly consists of Multimodal Embedding Representations, Question-guided Cross-modal Information Fusion (QCIF), Task-specific Knowledge Distillation with Spatial-Temporal Feature Constraints (TKD-STFC), and Question Semantic Consistency Constraint (QSCC).

4.2. Multimodal Embedding Representations

In the continual learning framework for audio-visual question answering, input data is processed from three modalities: audio, visual, and textual. For audio representation, each audio segment A_t is encoded into a feature vector f_a^t using a pre-trained VGGish model, which processes transformed audio spectrograms. The audio features are extracted offline, with the VGGish model [14] remaining frozen during this process. For visual representation, a fixed number of frames is sampled from each video segment, and a pre-trained ResNet-18 model [15] extracts visual feature maps f_v^t for each segment. Similar to the audio features, the ResNet-18 model is used without fine-tuning to preserve the original visual representations. For textual representation, an LSTM encoder is utilized to process the input question text, generating a feature vector f_q from the final hidden state. The question encoder is trained from scratch to ensure effective capture of the question’s semantic features, optimized for downstream tasks in the continual learning setup.

4.3. Question-guided Cross-modal Information Fusion

In our proposed AVQACL, audio and video data often contain substantial amounts of background information unrelated to the question. This irrelevant information can interfere with the model’s ability to extract essential details. To address these challenges, we propose the Question-guided Cross-modal Information Fusion (QCIF) module,

which leverages the question as a guiding query to selectively extract features from the visual and audio inputs. This approach uses attention mechanisms tailored to spatial and temporal contexts, enabling the model to focus on question-relevant information within each modality. The fusion process can be broken down into three primary stages: spatial-temporal attention on visual features, temporal attention on audio features, and question-guided feature fusion.

Spatial-Temporal Attention on Visual Features. The visual feature sequence, $V \in \mathbb{R}^{T_v \times F \times d_v}$ where T_v denotes the number of frames and F denotes the number of spatial visual tokens per frame, is attended by x_q using a multi-head attention mechanism [34]. The question-guided attention helps focus on spatial regions within the visual input relevant to the question:

$$V_{att} = \text{MultiHeadAttention}(x_q, V, V) \quad (4)$$

We apply a residual connection, followed by a linear transformation and layer normalization, to integrate the attention output:

$$V_{out} = \text{LayerNorm}(V_{att} + \text{Dropout}(\text{ReLU}(W_v V_{att}))) \quad (5)$$

where W_v denotes a learnable weight matrix used to perform the linear transformation on V_{att} .

Temporal Attention on Audio Features. Similarly, we apply question-guided attention on the audio feature sequence $A \in \mathbb{R}^{T_a \times d_a}$, where T_a denotes the number of audio tokens. This attention enables temporal alignment with question-

relevant sounds:

$$A_{att} = \text{MultiHeadAttention}(x_q, A, A) \quad (6)$$

We then apply a residual connection, followed by a linear transformation and layer normalization:

$$A_{out} = \text{LayerNorm}(A_{att} + \text{Dropout}(\text{ReLU}(W_a A_{att}))) \quad (7)$$

where W_a denotes a learnable weight matrix used to perform the linear transformation on A_{att} .

Question-guided Feature Fusion. After obtaining the attended visual and audio features, we concatenate them and pass them through a non-linear transformation:

$$f = \text{Tanh}([A_{out}; V_{out}]) \quad (8)$$

This fused feature vector f , is then combined with the question feature x_q through element-wise multiplication, emphasizing the question-relevant aspects:

$$f_{combined} = \text{Tanh}(f \odot x_q) \quad (9)$$

Finally, this combined feature vector $f_{combined}$ serves as the enriched cross-modal representation for downstream reasoning tasks. This fusion mechanism enhances the model's ability to focus on question-relevant aspects across modalities, improving the alignment between question intent and multimodal content.

4.4. Task-specific Knowledge Distillation with Spatial-Temporal Feature Constraints

In our proposed task-specific knowledge distillation with spatial-temporal feature constraints (TKD-STFC), we tackle the challenge of preserving essential knowledge from prior tasks while simultaneously learning new information within a continual learning framework. The primary goal of TKD-STFC is to maintain the spatial-temporal reasoning capabilities developed during previous tasks by distilling relevant knowledge from past experiences and ensuring consistent feature representations over time and space. Task-specific Knowledge Distillation (TKD) [30] is a technique often used in continual learning involving minimizing the difference between the outputs of the old and new models. The Kullback-Leibler divergence is employed to measure similarity in our proposed method. The Spatial-Temporal Feature Constraints (STFC) are defined to measure and minimize the differences in spatial-temporal feature changes between the new model and the old model, ensuring consistency in how the model interprets spatial and temporal variations over time. STFC utilizes a set of spatial-temporal constraints, which calculates feature discrepancies over time and space between the current and prior model states. The $\mathcal{L}_{STFC}$ consists of three components: temporal audio distance loss ($\mathcal{L}_{TA}$), which measures the discrepancy between the current and old audio features to preserve temporal audio patterns across tasks; temporal visual

distance loss ($\mathcal{L}_{TV}$), which ensures temporal consistency of visual features; and spatial visual distance loss ($\mathcal{L}_{SV}$), which addresses spatial consistency by aligning spatial feature changes in visual data, thus retaining spatial structural knowledge across tasks. Let f^{new} and f^{old} denote the new and old feature tensors, respectively. These components are combined in the unified loss function $\mathcal{L}_d$, defined as:

$$\mathcal{L}_d = \frac{1}{|\mathcal{I}_d|} \sum_{i \in \mathcal{I}_d} (\Delta_d f_i^{new} - \Delta_d f_i^{old})^2 \quad (10)$$

where $\Delta_d f_i = f_{i+\delta_d} - f_i$ is the difference of features along dimension d at index i . δ_d is a unit step in dimension d . $\mathcal{I}_d$ is the set of valid indices where the differences are computed along dimension d . $|\mathcal{I}_d|$ the total number of such indices. The $\mathcal{L}_{STFC}$ combines the three components, scaled by a weighting coefficient λ_{STFC} :

$$\mathcal{L}_{STFC} = \lambda_{STFC}(\mathcal{L}_{TA} + \mathcal{L}_{TV} + \mathcal{L}_{SV}) \quad (11)$$

The overall $\mathcal{L}_{TKD-STFC}$ expressed as follows combines $\mathcal{L}_{TKD}$ and $\mathcal{L}_{STFC}$, ensuring that the model retains previously learned task-specific knowledge while acquiring new information in a manner that maintains spatial-temporal consistency across tasks. This combined loss helps the model perform continual learning without sacrificing the reasoning capabilities developed in earlier tasks.

$$\mathcal{L}_{TKD-STFC} = \mathcal{L}_{TKD} + \mathcal{L}_{STFC} \quad (12)$$

4.5. Question Semantic Consistency Constraint

The question semantic consistency constraint (QSCC) is designed to ensure that the semantic features of a given question remain consistent across different tasks or models. In a continual learning setup, this consistency helps to prevent catastrophic forgetting, which occurs when a model forgets previously learned knowledge while learning new tasks. By enforcing consistency in the semantic understanding of questions over time, the QSCC module aids in knowledge retention and stability. The $\mathcal{L}_{QSCC}$ is calculated using cosine similarity between the current question features and the features from the previous task. By minimizing the difference between the two sets of features, the model is encouraged to maintain consistency in how it understands and represents the same question across tasks. The $\mathcal{L}_{QSCC}$ is then defined as:

$$\mathcal{L}_{QSCC} = 1 - \frac{1}{N} \sum_{i=1}^N \cos_sim(q_{new}^i, q_{old}^i) \quad (13)$$

where N is the number of question instances, q_{new}^i and q_{old}^i represent the i^{th} question's feature vector at the current and previous tasks, respectively. The $\mathcal{L}_{QSCC}$ multiplied by a coefficient λ_{QSCC} , is added to the cross-entropy loss $\mathcal{L}_{CE}$

Table 2. Experimental results of various methods on the audio-visual question answering continual learning datasets, Split-AVQA and Split-MUSIC-AVQA.

Method	Split-AVQA			Split-MUSIC-AVQA		
	M	MA	AF	M	MA	AF
Vanilla	None	18.34	40.94	None	28.79	33.17
LwF	None	21.53	30.49	None	30.93	30.14
EWC	None	18.89	40.69	None	28.76	32.58
MAS	None	24.47	8.91	None	29.33	31.15
iCaRL	5000	25.28	33.76	700	28.46	33.16
SSIL	5000	30.50	2.68	700	32.16	28.05
AV-CIL	5000	27.49	2.91	700	31.22	29.65
Ours	5000	32.05	2.47	700	33.64	27.08
Upper Bound	-	55.42	-	-	65.57	-

[16], resulting in the new classification loss with the Question Semantic Consistency Constraint. The formula is as follows:

$$\mathcal{L}_{CE-QSCC} = \mathcal{L}_{CE} + \lambda_{QSCC} \mathcal{L}_{QSCC} \quad (14)$$

4.6. The Overall Objective

The overall objective of our proposed method is to effectively integrate and retain knowledge across multiple tasks while preserving task-specific features. By incorporating the QCIF, TKD-STFC, and QSCC modules, we formulate the total loss function as:

$$\mathcal{L}_{total} = \mathcal{L}_{TKD-STFC} + \mathcal{L}_{CE-QSCC} \quad (15)$$

5. Experiments

5.1. Implementation details

All experiments are conducted using PyTorch [29]. During training, the self-supervised pre-trained visual and audio encoders are kept frozen, while the remaining parts of the model are fine-tuned. For each incremental step, the number of training epochs is set to 120. The batch size is set to 128 for Split-AVQA and 64 for Split-Music-AVQA. The networks are optimized with the Adam optimizer [17], with a learning rate and weight decay of $1e-3$ and $1e-4$, respectively. For the Split-AVQA dataset, the memory size is set to 5000, with the weights of the loss terms λ_{STFC} and λ_{QSCC} set to 0.5 and 0.7, respectively. For the Split-MUSIC-AVQA dataset, these values are adjusted to 700, 0.3 and 0.9, respectively.

5.2. Comparison and Analysis with baselines

Baselines. We evaluate our proposed method against several representative and state-of-the-art continual learning approaches, including Vanilla, LwF [25], EWC [18], MAS[4], iCaRL [32], SS-IL [2], and AV-CIL [30]. Vanilla serves as the simplest incremental training strategy, in

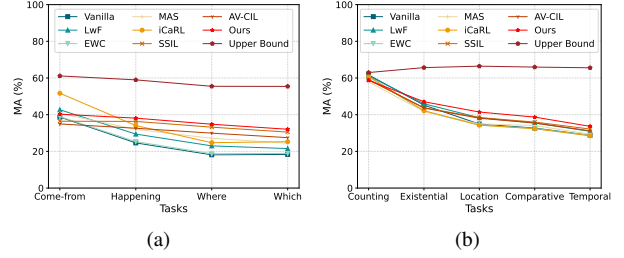


Figure 4. Fine-grained experimental results at the end of each language-driven task on (a) Split-AVQA and (b) Split-MUSIC-AVQA datasets.

which the model is initialized with parameters from the previous task and then retrained on each new task without applying any constraints or specialized techniques to mitigate catastrophic forgetting. Additionally, we include results from an Upper Bound model, which assumes access to all training data from previously seen classes. To ensure fair and consistent comparisons, all baseline methods are implemented with the same backbone architecture as our proposed method, including identical configurations for the visual encoder, audio encoder, question encoder, and classifier layer.

Comparison and Analysis. We present the main experimental results in Tab. 2, where it is evident that our proposed method outperforms all comparative methods, achieving the best performance overall. Specifically, on the Split-AVQA dataset, our method improves the state-of-the-art Mean Accuracy and Average Forgetting by 1.55 and 0.21, respectively. On the Split-MUSIC-AVQA dataset, our approach achieves enhancements of 1.48 and 0.97 in Mean Accuracy and Average Forgetting, respectively over the current best results. These findings underscore the effectiveness of our proposed method in continual learning for audio-visual question answering. Additionally, we provide more fine-grained results in Fig. 4(a) and 4(b), comparing our method and all baselines at the end of each language-driven task. Further analysis reveals that our approach attains the best testing accuracy in each incremental language-driven task on both datasets, with the exception of the first task.

5.3. Ablation study

Effect of Each Component. Our proposed approach comprises three key components: Question-guided Cross-modal Information Fusion (QCIF), Task-specific Knowledge Distillation with Spatial-Temporal Feature Constraints (TKD-STFC), and Question Semantic Consistency Constraint (QSCC). In this section, we conduct an ablation study by systematically removing one or more of these modules (QCIF, TKD-STFC or QSCC) across two datasets. As presented in Tab. 3, each module positively impacts the final performance, confirming the effectiveness and gener-

Table 3. Ablation study for our proposed method on Split-AVQA and Split-MUSIC-AVQA datasets.

QCIF	TKD-STFC	QSCC	Split-AVQA		Split-MUSIC-AVQA	
			MA	AF	MA	AF
×	×	×	18.34	40.94	28.79	33.17
✓	×	×	27.97	3.30	30.39	29.77
×	×	✓	29.08	2.54	30.73	30.19
✓	✓	×	30.77	2.92	31.19	29.85
✓	×	✓	27.85	4.36	31.87	28.75
✓	✓	✓	32.05	2.47	33.64	27.08

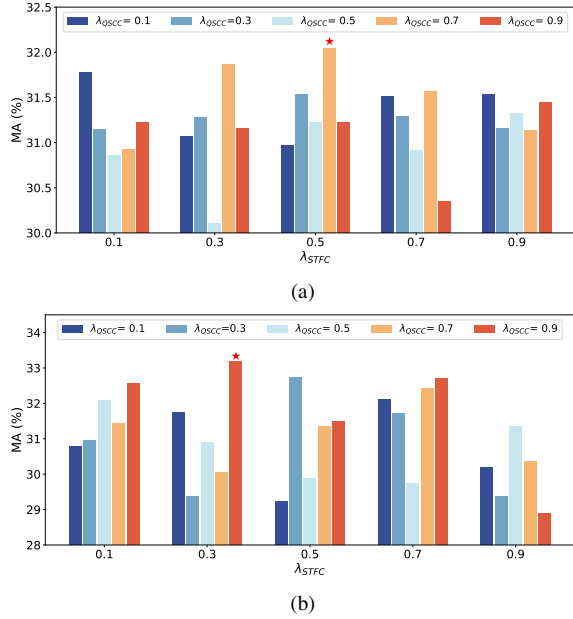


Figure 5. Parameter variation with different λ_{STFC} and λ_{QSCC} (a) Split-AVQA and (b) Split-MUSIC-AVQA datasets.

alizability of the proposed QCIF, TKD-STFC and QSCC. **Sensitive of Hyperparameter.** We investigated the hyperparameters λ_{STFC} and λ_{QSCC} in Equations (11) and (14), respectively, that influence the performance of the proposed method on the Split-AVQA and Split-MUSIC-AVQA datasets. During model training, the values of λ_{STFC} and λ_{QSCC} are selected from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The models are trained using different parameter combinations, and the corresponding test results are shown in the Fig. 5(a) and 5(b). It can be observed that the best performance on the Split-AVQA dataset is achieved when λ_{STFC} and λ_{QSCC} are set to 0.5 and 0.7, while on the Split-Music-AVQA dataset, the optimal results are obtained with different values for 0.3 and 0.9. Therefore, in our experiments, we set λ_{STFC} and λ_{QSCC} to the values that yielded the best performance on the respective datasets.

Impact of Memory Size. In this section, we investigate the impact of memory size on the performance of the proposed method. We conducted experiments with different memory sizes and present the results for the Split-AVQA and Split-MUSIC-AVQA datasets in Fig. 6(a) and 6(b), respec-

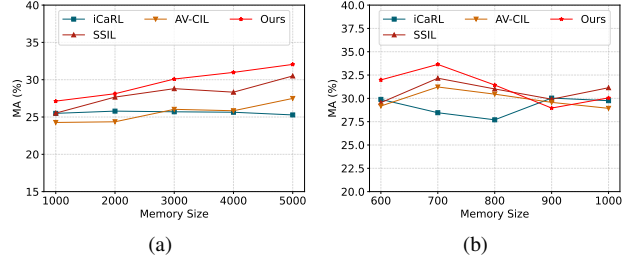


Figure 6. Experimental results of our proposed method under different memory size settings on (a) Split-AVQA and (b) Split-MUSIC-AVQA datasets.

tively. As shown in Fig. 6(a), our model exhibits a gradual performance improvement on the Split-AVQA dataset as the memory size increases from 1000 to 5000, outperforming the other replay-based methods, which highlights the effectiveness of our approach. However, on the Split-MUSIC-AVQA dataset, as the memory size increases from 600 to 1000, the performance fluctuates, as illustrated in Fig. 6(b). This suggests that as the memory size grows, our model begins to overfit, and further increases in memory size do not lead to sustained improvements in model performance.

6. Conclusion and Limitation

In this paper, we introduce a novel benchmark for audio-visual question answering continual learning to explore fine-grained scene understanding and spatial-temporal reasoning within videos under a continual learning framework. To support this multimodal continual learning task, two datasets termed Split-AVQA and Split-MUSIC-AVQA are developed based on the AVQA and MUSIC-AVQA datasets, respectively. We propose a novel continual learning approach that includes question-guided cross-modal information fusion, task-specific knowledge distillation with spatial-temporal feature constraints and question semantic consistency constraint. Extensive comparative and ablation experiments demonstrate that our method achieves state-of-the-art performance on the Split-AVQA and Split-MUSIC-AVQA datasets, establishing a robust approach to audio-visual question answering in a continual learning setting.

Recently, we have observed that large multimodal models have outperformed traditional deep learning methods in the field of audio-visual question answering. However, due to limitations in computational resources, we have not explored methods for continual learning in audio-visual question answering based on large multimodal models in this paper. Therefore, we hope that this research will inspire future studies on both traditional and large multimodal models, advancing the frontiers of what is achievable in real-world applications.

7. Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This work was co-supported by the National Natural Science Foundation of China under Grant 62233003 and 62073072, and in part by the Key Projects of Key RD Program of Jiangsu Province, China under Grant BE2020006 and Grant BE2020006-1 and in part by Shenzhen Science and Technology Program, China under Grant JCYJ20210324132202005 and JCYJ20220818101206014.

References

- [1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020. [2](#)
- [2] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 844–853, 2021. [2](#), [3](#), [7](#)
- [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. [2](#)
- [4] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. [2](#), [3](#), [7](#)
- [5] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11254–11263, 2019. [3](#)
- [6] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019. [3](#)
- [7] 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. [1](#)
- [8] Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018. [3](#)
- [9] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5138–5146, 2019. [3](#)
- [10] Yilin Dong, Xinde Li, Jean Dezert, Rigui Zhou, Changming Zhu, Lei Cao, Mohammad Omar Khyam, and Shuzhi Sam Ge. Multisource weighted domain adaptation with evidential reasoning for activity recognition. *IEEE Transactions on Industrial Informatics*, 19(4):5530–5542, 2022. [1](#)
- [11] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16*, pages 86–102. Springer, 2020. [3](#)
- [12] Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. *Advances in Neural Information Processing Systems*, 36, 2024. [3](#)
- [13] 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. [1](#)
- [14] 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. [5](#)
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [5](#)
- [16] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. [7](#)
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [7](#)
- [18] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. [2](#), [3](#), [7](#)
- [19] Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M Bakker, and Michael S Lew. Coca: Collaborative causal regularization for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12995–13003, 2023. [3](#)
- [20] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Jirong 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](#), [3](#), [4](#)
- [21] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7808–7816, 2023. [1](#), [2](#), [3](#)
- [22] Guangyao Li, Henghui Du, and Di Hu. Boosting audio visual question answering via key semantic-aware cues. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5997–6005, 2024. [3](#)

- [23] Qiwei Li, Yuxin Peng, and Jiahuan Zhou. Fcs: Feature calibration and separation for non-exemplar class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28495–28504, 2024. 3
- [24] Xianghui Li, Xinde Li, and Hong Pan. Multi-scale vehicle detection in high-resolution aerial images with context information. *IEEE Access*, 8:208643–208657, 2020. 1
- [25] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2, 3, 7
- [26] Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3306–3314, 2024. 1, 2
- [27] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12245–12254, 2020. 3
- [28] Shentong Mo, Weiguo Pian, and Yapeng Tian. Class-incremental grouping network for continual audio-visual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7788–7798, 2023. 3
- [29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7
- [30] Weiguo Pian, Shentong Mo, Yunhui Guo, and Yapeng Tian. Audio-visual class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7799–7811, 2023. 2, 3, 4, 6, 7
- [31] Zi Qian, Xin Wang, Xuguang Duan, Hong Chen, and Wenwu Zhu. Dynamic spatio-temporal modular network for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4466–4477, 2022. 2
- [32] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2, 3, 7
- [33] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On learning the geodesic path for incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1591–1600, 2021. 3
- [34] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5
- [35] Andrés Villa, Kumail Alhamoud, Victor Escorcia, Fabian Caba, Juan León Alcázar, and Bernard Ghanem. vclimb: A novel video class incremental learning benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19035–19044, 2022. 2, 3
- [36] Andrés Villa, Juan León Alcázar, Motasem Alfarra, Kumail Alhamoud, Julio Hurtado, Fabian Caba Heilbron, Alvaro Soto, and Bernard Ghanem. Pivot: Prompting for video continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24214–24223, 2023. 3
- [37] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3
- [38] LI Xinde, HUANG Xinhan, and WANG Min. Robot map building from sonar sensors and dsmt. *Inf. & Sec.: An Int. J., Bulg. Acad. of Sci*, 2006. 1
- [39] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022. 1, 3, 4
- [40] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Pacs: A dataset for physical audiovisual commonsense reasoning. In *European Conference on Computer Vision*, pages 292–309. Springer, 2022. 1, 3
- [41] 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
- [42] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafetin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1131–1140, 2020. 3
- [43] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19102–19112, 2023. 2, 3, 4
- [44] Zhentong Zhang, Xinde Li, Heqing Li, Fir Dunkin, Bing Li, and Zhijun Li. Dual-branch sparse self-learning with instance binding augmentation for adversarial detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1
- [45] Boqing Zhu, Changjian Wang, Kele Xu, Dawei Feng, Zemin Zhou, and Xiaoqian Zhu. Learning incremental audio-visual representation for continual multimodal understanding. *Knowledge-Based Systems*, 304:112513, 2024. 3
- [46] Yukun Zuo, Hantao Yao, Liansheng Zhuang, and Changsheng Xu. Hierarchical augmentation and distillation for class incremental audio-visual video recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3