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

Supplementary Material

In this supplementary material, we provide an in-depth explanation of aspects in the following that were omitted from the main paper.

- Section A: We provide additional details on the construction of the AVQACL benchmark and the comparison methods employed.
- Section B: Additional fine-grained experimental results and comprehensive ablation studies are provided to further validate the effectiveness and superiority of our proposed method.

A. More Details about the AVQACL Benchmark

In this paper, we established a novel benchmark for audio-visual question answering continual learning to study fine-grained scene understanding and spatial-temporal reasoning in videos under a continual learning setting.

Datasets. The benchmark for this study is constructed from portions of two well-established and representative datasets in the AVQA domain, namely Split-AVQA and Split-MUSIC-AVQA. For the Split-AVQA dataset, it encompasses six types of audio-visual relationship classes. Based on the question taxonomy provided in [8], we divided the Split-AVQA dataset into four language-driven tasks: come from, happening, where, and which. Detailed data regarding the training, validation, and test sets are shown in Tab. 1. Subsequently, according to the class of objects and their corresponding sounds present in the Split-AVQA dataset, each language-driven task was randomly and evenly divided into six subtasks. The detailed grouping data is presented in Tab. 2.

Similarly, for the Split-MUSIC-AVQA dataset, we followed the categorization approach from [4], organizing it into five language-driven tasks: counting, existential, location, comparative, and temporal. Detailed data for the training, validation, and test sets are shown in Tab. 3. Following this, based on the class of objects and their corresponding sounds in the Split-MUSIC-AVQA dataset, each language-driven task was randomly and evenly divided into 11 subtasks. Detailed grouping data is provided in Tab. 4.

Comparison Methods. To the best of our knowledge, there is currently no related work in the literature that investigates continual learning in the context of audio-visual question answering. Therefore, we selected five representative continual learning methods from the traditional image classification domain, including three regularization-based methods (LwF [5], EWC [3], MAS [2]) and two rehearsal-based methods (iCaRL [7], SSIL [1]). Additionally, we included a multimodal continual learning method designed for audio-visual classification, AV-CIL [6]. For a fair comparison, all methods were implemented using official codebases and repurposed on our proposed Vanilla framework to adapt to the audio-visual question continual learning setting.

LwF [5] is a continual learning method that preserves knowledge from previous tasks by incorporating knowledge distillation. In LwF, when training on a new task, the model maintains the predictions for old tasks by adding a distillation loss. This loss encourages the current model to produce similar outputs for the previous tasks as it did before, thereby reducing the risk of overwriting previously learned knowledge. Unlike methods that rely on storing data from past tasks, LwF only requires the current task's data, making it efficient in terms of memory usage.

EWC [3] is a continual learning method that employs regularization to preserve knowledge from past tasks. EWC identifies parameters essential for previous tasks by leveraging the Fisher Information Matrix, which quantifies the importance of each parameter. To prevent significant changes in these key parameters while learning new tasks, EWC adds an L_2 regularization term. This additional loss penalizes adjustments to crucial parameters, helping the model retain prior knowledge and reduce catastrophic forgetting as it acquires new information.

MAS [2] is another regularization-based approach designed to preserve knowledge from prior tasks in continual learning. MAS achieves this by discouraging significant modifications to parameters critical for previous tasks, using an additional L_2 loss as a penalty. To assess parameter importance, MAS calculates the sensitivity of the model's output function to changes in each parameter, thereby identifying which parameters should be protected during the training of new tasks. This approach helps mitigate catastrophic forgetting by retaining essential knowledge from earlier tasks.

iCaRL [7] is a continual learning approach that combines class-incremental learning with representation learning. The core idea of iCaRL is to maintain a compact, evolving model that can incorporate new classes without forgetting old ones. It achieves this by storing a small set of representative data samples for each class, which are used to update the classifier as new classes are introduced. To prevent catastrophic forgetting, iCaRL also employs a nearest-class-mean classifier that assigns labels based on the nearest class mean in the feature space, rather than recalculating the decision boundaries after each new task. Additionally, iCaRL

Table 1. The data statistics for language-driven tasks in Split-AVQA dataset include the number of videos in the training, validation, and test sets, as well as examples of question.

Task	Train	Val	Test	Question Examples
Come From	9135	1800	1861	What is the source of the sound in the video? What is the main sound source of the video?
Happening	5093	902	931	What happened in the video? What are the people in the video doing?
Where	3620	712	757	Where did the video take place? Where is the car driving in this video?
Which	11513	2315	2369	What animal appears in the video? What's driving in the video?

Table 2. The data statistics for language-driven tasks in the Split-MUSIC-AVQA dataset include the number of videos in the training, validation, and test sets, along with examples of the question.

Task	Train	Val	Test	Question Examples
Counting	3304	471	947	How many instruments are sounding in the video? How many instruments in the video did not sound from beginning to end?
Existential	3134	449	893	Is the accordion in the video always playing? Is this sound from the instrument in the video?
Location	2477	353	711	Which is the musical instrument that sounds at the same time as the flute? What is the left instrument of the first sounding instrument?
Comparative	3379	486	972	Is the flute on the right louder than the piano on the left? Is the instrument on the left more rhythmic than the instrument on the right?
Temporal	2241	320	637	What is the third instrument that comes in? Which instrument makes sounds after the piano?

Table 3. The detailed grouping data statistics for each subtask under each language-driven task in the Split-AVQA dataset, encompassing 120 distinct objects and their corresponding sound classes.

Subtask	class
1	wind rustling leaves, splashing water, sheep bleating, horse clip-clop, skateboarding, dog whimpering, pig oinking, waterfall burbling, using sewing machines, lions roaring, bull bellowing, cap gun shooting, tornado roaring, coyote howling, cat purring, car engine starting, rowboat, hammering nails, train horning, volcano explosion
2	typing on typewriter, dog bow-wow, underwater bubbling, running electric fan, canary calling, bird squawking, cat hissing, black capped chickadee calling, penguins braying, spraying water, motorboat, car passing by, train wheels squealing, vehicle horn, bee, subway, pheasant crowing, airplane, bowling impact, civil defense siren
3	driving snowmobile, otter growling, machine gun shooting, sailing, donkey, car engine knocking, pigeon, chicken crowing, hail, fox barking, frog croaking, cat growling, cattle mooing, skiing, helicopter, sea lion barking, barn swallow calling, roller coaster running, elephant trumpeting, turkey gobbling
4	train whistling, race car, pumping water, snake rattling, engine accelerating, sloshing water, magpie calling, police car, wood thrush calling, ice cream truck, wind chime, driving buses, cow lowing, mynah bird singing, warbler chirping, lions growling, reversing beeps, beat boxing, chicken clucking, ambulance siren
5	owl hooting, lighting firecrackers, dog howling, horse neighing, rope skipping, squishing water, chopping food, alligators, railroad car, cat meowing, dog growling, plastic bottle crushing, driving motorcycle, raining, scuba diving, electric shaver, printer printing, gibbon howling, sea waves, airplane flyby
6	toilet flushing, fireworks banging, lathe spinning, goose honking, wind noise, lawn mowing, whale calling, snake hissing, chipmunk chirping, fire truck siren, duck quacking, cattle, crow cawing, stream burbling, planing timber, tractor digging, vacuum cleaner cleaning floors, dog barking, people eating apple, bird chirping

Table 4. The detailed grouping data statistics for each subtask under each language-driven task in the Split-MUSIC-AVQA dataset, covering 22 different musical instruments and their corresponding sound classes.

Subtask	1	2	3	4	5	6	7	8	9	10	11
class	bagpipe ukulele	suona erhu	cello congas	guitar trumpet	bassoon xylophone	tuba violin	flute guzheng	bass accordion	drum clarinet	pipa banjo	saxophone piano

uses a loss function that encourages the model to preserve the representation of old classes while learning new ones, facilitating the continual learning process.

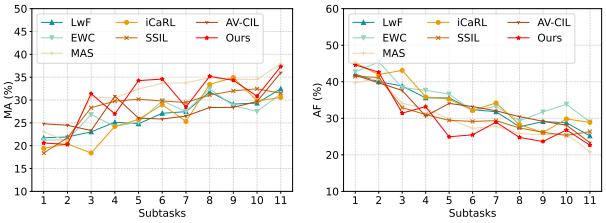
SSIL [1] is a continual learning approach that leverages self-supervised learning techniques to enable incremental task learning without forgetting previous knowledge. The method generates useful representations from unlabeled data by solving auxiliary tasks, such as contrastive learning, which helps in learning task-agnostic features. These representations are updated incrementally as new tasks are introduced, allowing the model to learn without the need for storing old data. By focusing on self-supervised objectives, SSIL reduces memory usage and mitigates catastrophic forgetting, making it an efficient solution for continual learning.

AV-CIL [6] is a multimodal continual learning approach designed to handle the challenges of learning from both audio and visual data over time. AV-CIL introduces two innovative components. First, the Dual-Audio-Visual Similarity Constraint (D-AVSC) ensures both instance-level and class-level semantic similarity between audio and visual features, facilitating robust joint learning. Second, the Visual Attention Distillation (VAD) mechanism retains audio-guided visual attention capabilities, preventing the forgetting of previously learned cross-modal correlations. The method combines audio and visual information to learn incrementally without forgetting previous tasks. This allows the model to adapt to new tasks while maintaining high performance on previous ones, making it particularly useful for applications involving dynamic audio-visual data streams.

B. More Experimental Results and Ablation Study

B.1. More Experimental Results

We provide more fine-grained experimental results on the Split-AVQA and Split-MUSIC-AVQA datasets. As shown in Tab. 5, the experimental results for the average accuracy and average forgetting rate of subtasks under the "where" language-driven task on the Split-AVQA dataset are presented, comparing our proposed method with six repurposed methods. From the table, it can be observed that our method achieves the best average accuracy in every subtask and nearly the best performance in terms of average forgetting rate. Fig. 1 presents the experimental results for the average accuracy and average forgetting rate of subtasks under the "comparative" language-driven task on the Split-MUSIC-AVQA dataset. From the figure, it is evident that the average accuracy and average forgetting rate fluctuate during the testing of subtasks across all methods. These fluctuations may be attributed to factors such as inter-subtask interference and the inherent multimodal complexity of audio-visual question answering data, which can introduce challenges in balancing feature learning and fusion. However, our proposed method achieves competitive results in the majority of subtasks.



(a) (b)

Figure 1. Fine-grained experimental results on Split-MUSIC-AVQA datasets (a) the mean accuracy for each subtask (b) the mean forgetting for each subtask.

B.2. More Ablation Study

Effect of Input Modality. In this section, we detail the continual learning experimental results of our proposed method across three distinct input modalities: audio question answering, visual question answering, and audio-visual question answering. Fig. 2 provide a comprehensive comparison of the mean accuracy and average forgetting rate achieved by the proposed method on the Split-AVQA and Split-MUSIC-AVQA datasets for these modalities. The results clearly demonstrate that, the integration of both audio and visual modalities leads to significant improvements in the inference performance of the method under the continual learning setting. Additionally, this multimodal approach effectively reduces the forgetting rate, enabling the model to retain a larger proportion of previously acquired knowledge. This highlights the advantages of leveraging multimodal information in enhancing the robustness and generalizability of continual learning models.

Effect of Frames per Video. To investigate the effect of the total number of sampled frames per input video on the testing performance of the proposed model, we conducted

Table 5. Fine-grained experimental results on Split-AVQA dataset. MA represents mean accuracy, while AF denotes average forgetting

Method	Subtask-1		Subtask-2		Subtask-3		Subtask-4		Subtask-5		Subtask-6	
	MA	AF	MA	AF	MA	AF	MA	AF	MA	AF	MA	AF
LwF	15.63	34.10	14.08	31.86	11.33	33.37	6.00	32.67	5.7	33.40	7.95	29.69
EWC	2.07	55.84	4.33	48.06	4.32	47.51	6.79	42.30	6.61	42.93	8.45	39.65
MAS	20.04	11.63	19.13	11.59	17.17	12.36	16.66	11.57	16.12	10.57	15.18	10.08
iCaRL	1.25	60.19	4.56	49.59	6.06	45.52	6.91	44.27	8.19	40.54	9.35	39.02
SSIL	30.94	4.63	27.86	5.40	27.29	5.24	25.78	5.36	25.50	4.67	24.26	4.54
AV-CIL	27.94	4.05	26.00	4.05	25.37	3.79	23.33	4.31	23.04	4.03	21.81	3.95
Ours	**31.32**	**4.39**	**30.07**	**4.12**	**28.46**	**4.87**	**27.32**	**4.35**	**26.41**	**4.23**	**25.50**	**4.00**

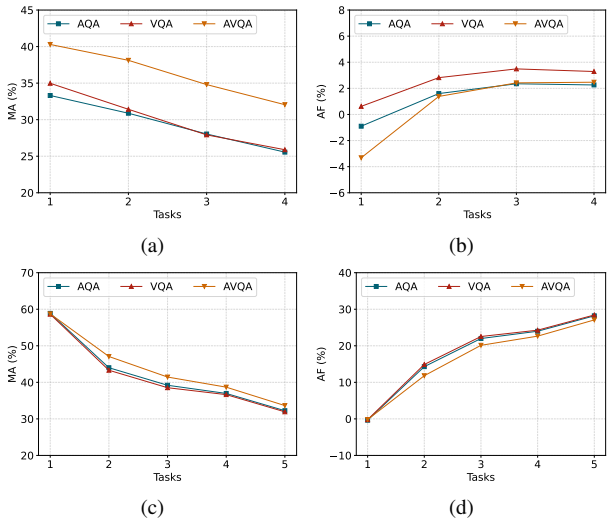


Figure 2. Effect of input modality. (a) the mean accuracy on Split-AVQA, (b) the average forgetting on Split-MUSIC-AVQA dataset, (c) the mean accuracy on Split-MUSIC-AVQA dataset and (d) the average forgetting on Split-MUSIC-AVQA dataset.

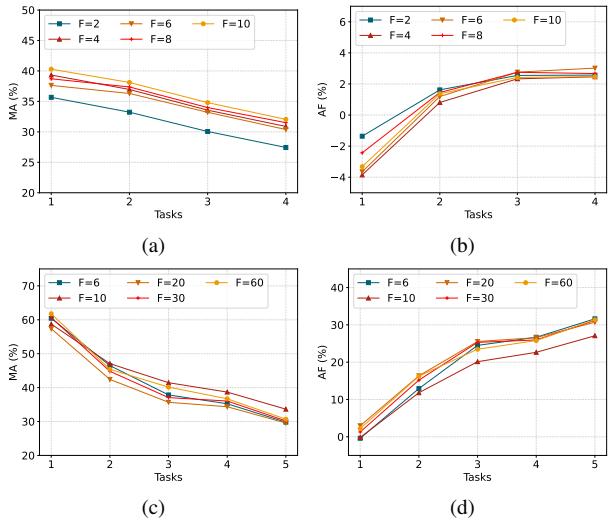


Figure 3. Effect of frames per video. (a) the mean accuracy on Split-AVQA, (b) the average forgetting on Split-MUSIC-AVQA dataset, (c) the mean accuracy on Split-MUSIC-AVQA dataset and (d) the average forgetting on Split-MUSIC-AVQA dataset.

frame count ablation experiments on both datasets. Specifically, for the Split-AVQA dataset, we set the number of input video frames per video to 2, 4, 6, 8, and 10, based on the varying lengths of the videos. Similarly, for the Split-MUSIC-AVQA dataset, the input frame counts were set to 6, 10, 20, 30, and 60. The experimental results are shown in Fig. 3, where it can be observed that the proposed method achieves the best average accuracy and the lowest average forgetting rate when the number of input frames per video is set to 10 for both datasets. Consequently, we adopted 10 input frames per video as the default setting in our experiments for both datasets.

Effect of Task Order. As shown in Fig. 4, we present the testing results of our proposed method on the Split-AVQA and Split-MUSIC-AVQA datasets when trained under different order of language-driven tasks. The figure reveals that the order of language-driven tasks significantly impacts the model's average accuracy and average forgetting rate. A plausible explanation for this observation lies in the varying levels of complexity among different language-driven tasks. For instance, tasks such as existential or comparative queries are relatively straightforward, as their answers are typically binary (e.g., "yes" or "no"). In contrast, tasks like counting or spatiotemporal reasoning are inherently more complex and diverse in their answers, requiring the model to perform higher-order reasoning to generate accurate responses.

This disparity in task complexity could lead to differences in the way the model learns and retains knowledge across task sequences, emphasizing the importance of carefully designing the order of tasks in continual learning settings to optimize overall performance. Such insights highlight the potential challenges of balancing task complexity and sequence effects in audio-visual question answering continual learning frameworks.

Effect of Class Number. In this section, we investigate the

Table 6. The experimental results of our proposed method and the comparison methods under different class numbers.

Method	Split-AVQA				Split-MUSIC-AVQA			
	6 classes × 20 subtasks		20 classes × 6 subtasks		2 classes × 11 subtasks		11 classes × 2 subtasks	
	MA	AF	MA	AF	MA	AF	MA	AF
Vanilla	18.34	40.94	10.24	42.66	28.79	33.17	33.24	34.89
LwF	21.53	30.49	10.95	34.70	30.93	30.14	34.12	32.13
EWC	18.89	40.69	10.28	44.34	28.76	32.58	33.74	34.02
MAS	24.47	8.91	10.93	37.29	29.33	31.15	33.16	33.22
iCaRL	25.28	33.76	15.03	43.00	28.46	33.16	32.93	34.61
SSIL	30.05	2.68	33.75	4.78	32.16	28.05	36.00	24.86
AV-CIL	27.49	2.91	33.59	5.34	31.22	29.65	34.20	28.47
Ours	**32.05**	**2.47**	**34.37**	**3.31**	**33.64**	**27.08**	**36.50**	**24.44**
Upper Bound	55.42	-	54.51	-	65.57	-	66.35	-

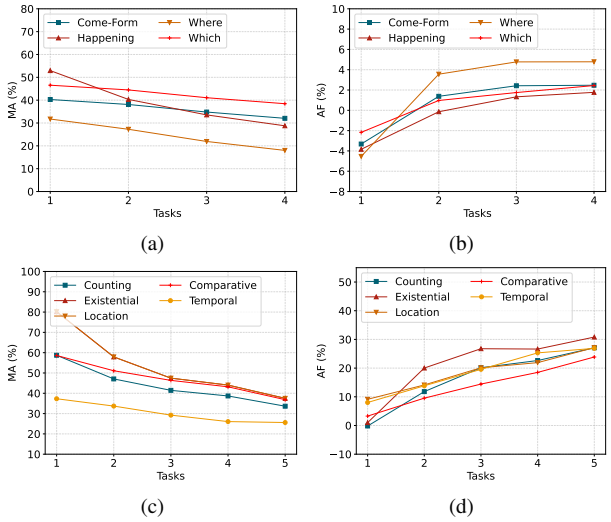


Figure 4. Effect of task order. (a) the mean accuracy on Split-AVQA, (b) the average forgetting on Split-MUSIC-AVQA dataset, (c) the mean accuracy on Split-MUSIC-AVQA dataset and (d) the average forgetting on Split-MUSIC-AVQA dataset.

effect of varying the number of object-sound class within each language-driven subtask on the performance of the proposed method and baseline approaches. Specifically, we modified the subtask learning setups in the Split-AVQA and Split-MUSIC-AVQA datasets from the original configurations of 6 classes × 20 subtasks and 2 classes × 11 subtasks to 20 classes × 6 subtasks and 11 classes × 2 subtasks, respectively.

As shown in Tab. 6 our proposed method consistently achieves the best performance across both datasets and under both subtask configurations. These results highlight the effectiveness and generalizability of the proposed approach, demonstrating its robustness in handling varying class distributions and subtask complexities within multimodal continual learning settings.

References

[1] 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, 4

[2] 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] 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

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

[5] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2

[6] 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, 4

[7] 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

[8] 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. 2