Audio-Visual Class-Incremental Learning Supplementary Material

Weiguo Pian[1†], Shentong Mo[2†], Yunhui Guo[1], Yapeng Tian[1]
[1] The University of Texas at Dallas, [2] Carnegie Mellon University
{weiguo.pian,yunhui.guo,yapeng.tian}@utdallas.edu, shentonm@andrew.cmu.edu

A. Appendix

In this appendix, we present supplementary experimental results for various class-incremental learning approaches utilizing distinct input modalities at each step, as detailed in Section A.1. Following that, we offer comprehensive visualizations of visual attention maps in Section A.2. Moreover, in Section A.3, we supply further parameter studies, and in Section A.4, we include additional results under diverse class-incremental settings. Finally, we show the experimental results compared to the existing visual attention distillation approach in Section A.5.

A.1. Effect of Input Modalities on Class-incremental Approaches

In this section, we present the experimental comparison of several class-incremental learning approaches, including Fine-tuning, LwF [4], iCaRL-NME [5], iCaRL-FC [5], SSIL [1], AFC-NME [3], and AFC-LSC [3]. We compare the performance of these approaches at each step in the context of audio-visual class-incremental learning and unimodal audio or visual modality class-incremental learning. The experimental results on the AVE-CI, K-S-CI, and VS100-CI datasets are illustrated in Figure 2, Figure 3, and Figure 4, respectively. These results indicate that training with joint audio-visual modalities consistently outperforms training using only a single audio or visual modality at each incremental step. This highlights the effectiveness and superiority of cross-modal audio-visual learning in class-incremental scenarios compared to single audio or visual modality learning. Overall, our findings suggest that integrating audio and visual modalities in class-incremental learning can enhance performance, promote efficient knowledge transfer between tasks, and prevent the model from forgetting previously acquired knowledge.

A.2. Full Visualization of Visual Attention

In this section, we present the complete visualization of audio-guided visual attention maps with and without our proposed Visual Attention Distillation (VAD) to demonstrate their vanishing and preservation at each incremental step. Figure 5 provides the full version of the figure initially shown in Section 3.4 of the main paper, illustrating the vanishing of the visual attention map. In Figure 6, we display the full version of the figure initially shown in Section 4.5, which visualizes the visual attention map after applying our proposed VAD. These visualizations reveal that implementing our VAD effectively preserves the previously learned audio-guided visual attention capabilities, preventing the model from forgetting the established audio-visual correlations in prior incremental steps. This further substantiates the effectiveness of our proposed approach

[†]Equal contribution.

Table 1: Parameter studies on AVE-CI dataset.

λ_I	λ_C	Accuracy				Mean Acc.
		step 1	step 2	step 3	step 4	
0.1	1.0	79.81	71.43	69.52	66.50	71.82
0.3	1.0	79.81	75.71	71.75	67.26	73.63
0.5	1.0	79.81	77.14	71.43	67.77	74.04
0.8	1.0	79.81	76.19	70.16	65.23	72.85
1.0	1.0	79.81	75.24	67.94	66.50	72.37
0.5	0.8	79.81	75.24	71.43	68.78	73.82
0.5	0.5	79.81	74.29	67.94	66.24	72.07
0.5	0.3	79.81	75.71	70.48	68.27	73.56
0.5	0.1	79.81	75.71	68.25	66.50	72.57

Table 2: Parameter studies on K-S-CI dataset.

λ_I	λ_C	Accuracy					Mean Acc.
		step 1	step 2	step 3	step 4	step 5	
0.1	1.0	93.01	73.64	70.19	66.47	62.00	73.06
0.3	1.0	93.01	70.13	67.76	64.40	60.32	71.12
0.5	1.0	93.01	70.26	66.64	63.56	58.53	70.40
0.8	1.0	93.01	66.10	63.78	61.55	57.05	68.30
1.0	1.0	93.01	68.05	65.08	62.01	56.03	68.84
0.1	0.8	93.01	72.99	69.50	66.73	61.64	72.77
0.1	0.5	93.01	72.47	68.63	66.21	60.88	72.24
0.1	0.3	93.01	73.38	69.41	65.44	61.44	72.54
0.1	0.1	93.01	71.95	68.80	66.34	60.27	72.07

Table 3: Parameter studies on VS100-CI dataset.

λ_I	λ_C	Accuracy										Mean Acc.
		step 1	step 2	step 3	step 4	step 5	step 6	step 7	step 8	step 9	step 10	
0.1	1.0	88.40	80.80	81.73	76.35	73.52	69.80	67.49	65.70	62.29	61.96	72.80
0.3	1.0	88.40	78.40	79.13	75.25	71.84	67.57	65.97	63.83	60.31	60.26	71.10
0.5	1.0	88.40	77.30	77.27	76.00	70.96	67.07	65.49	62.03	58.09	57.40	70.00
0.8	1.0	88.40	75.90	76.33	75.30	69.72	66.67	64.54	61.60	57.56	56.40	69.24
1.0	1.0	88.40	75.30	76.00	74.65	70.28	66.33	64.20	60.85	57.22	55.46	68.87
0.1	0.8	88.40	81.30	81.27	76.55	73.32	69.23	68.00	65.40	62.20	61.62	72.73
0.1	0.5	88.40	81.20	81.67	76.60	72.52	69.47	67.46	65.08	61.96	61.88	72.62
0.1	0.3	88.40	80.40	81.07	76.20	72.88	69.27	66.69	65.05	62.18	61.54	72.37
0.1	0.1	88.40	80.40	81.07	75.80	72.32	68.83	67.26	65.15	61.98	61.30	72.25

Table 4: Experimental results of different methods on AVE-CI and K-S-CI datasets with different class-incremental settings.

Methods	Mean Accuracy			
	AVE-CI		K-S-CI	
	7 classes × 4 steps	4 classes × 7 steps	6 classes × 5 steps	5 classes × 6 steps
Fine-tuning	42.40	37.54	41.18	37.45
LwF [4]	58.07	50.25	65.54	63.10
iCaRL-NME [5]	56.15	60.02	64.51	63.29
iCaRL-FC [5]	65.88	65.50	65.54	65.13
SS-IL [1]	61.94	65.29	69.71	68.38
AFC-NME [3]	68.46	68.61	69.13	67.74
AFC-LSC [3]	65.21	61.93	67.02	65.86
AV-CIL (Ours)	**74.04**	**71.76**	**73.06**	**73.24**
Oracle (Upper Bound)	76.85	78.23	80.43	80.50

in averting catastrophic forgetting and maintaining performance throughout multiple incremental learning steps.

A.3. Parameter Studies

In this section, we explore the impact of different settings of λ_I and λ_C on AVE-CI and K-S-CI datasets and present all of our experimental results. We also investigate the impact of λ_{VAD} and find that setting it to 0.5 performs well on both datasets, so we use this value in all our experiments. Our experimental results, which can be found in Table 1, 2, and 3 for AVE-CI, K-S-CI, and VS100-CI datasets respectively, demonstrate the effects of different settings of λ_I and λ_C on the performance of the model.

A.4. Results with Different Incremental Settings

In our main experiments, the class-incremental settings of AVE-CI and K-S-CI datasets are set to **7 classes** × **4 steps** and **6 classes** × **5 steps** respectively. In this subsection, we conduct experiments on AVE-CI and K-S-CI datasets with the class-incremental settings of **4 classes** × **7 steps** and **5 classes** × **6 steps** respectively, to investigate the performance of our proposed method compared to baselines with different class-incremental settings. The results are shown in Table 4, from which we can see that, in the incremental setting of **4 classes** × **7 steps** on AVE-CI dataset, our method outperforms the state-of-the-art method AFC-NME by **3.15**. For the incremental setting of **5 classes** × **6 steps** on K-S-CI dataset, our method outperforms state-of-the-art result by **4.86**. These results demonstrate the versatility of our proposed method, as it can achieve superior performance across different class-incremental settings.

A.5. Comparison with Existing Visual Attention Distillation Approach

In our AV-CIL, we propose the VAD, a novel audio-guided visual attention map distillation method for audio-visual class-incremental learning. Our VAD enablea the model to preserve previously learned attentive ability in future classes/tasks, effectively preventing the model from forgetting previously learned audio-visual semantic correlations. To further evaluate the effectiveness of our proposed VAD compared to existing attention distillation meth-

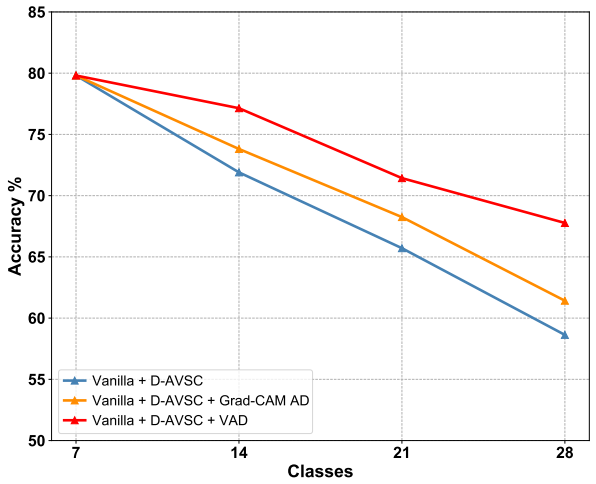

Figure 1: Testing accuracy at each incremental step of (1) Vanilla + D-AVSC, (2) Vanilla + D-AVSC + Grad-CAM AD, and (3) Vanilla + D-AVSC + VAD on AVE-CI dataset.

ods, we construct a variant of our AV-CIL, by replacing our VAD with the Grad-CAM-based visual attention distillation (Grad-CAM AD) [2]. We conduct experiments on the AVE-CI dataset with different variants of our proposed AV-CIL, in which we name Vanilla as the variant without the D-AVSC and the VAD (with only the Task-wise Knowledge Distillation and the Separated Softmax Cross-Entropy). The experimental results are presented in Table 5, where we show the Mean Accuracy of Vanilla + D-AVSC, Vanilla + D-AVSC + Grad-CAM AD, and Vanilla + D-AVSC + VAD on AVE-CI dataset. From the table, we can see that our full AV-CIL outperforms the variant with Grad-CAM AD significantly, which demonstrates the superiority and effectiveness of our proposed VAD over the Grad-CAM AD. We also show the testing accuracy at each incremental step in Figure 1, where we can see that our full AV-CIL has better performance at each incremental step compared to the variants with Grad-CAM AD, further demonstrating the effectiveness of our proposed VAD.

Table 5: Experimental results of different variants on AVE-CI dataset. Our AV-CIL performs better than the variant with Grad-CAM AD, demonstrating the superiority of our proposed VAD over the Grad-CAM AD.

Variants	Mean Acc.
Vanilla + D-AVSC	69.01
Vanilla + D-AVSC + Grad-CAM AD	70.82
Vanilla + D-AVSC + VAD (AV-CIL)	**74.04**

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 824–833, 2021. 1, 2

[2] 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

[3] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16050–16059, 2022. 1, 2

[4] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 1, 2

[5] 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. 1, 2

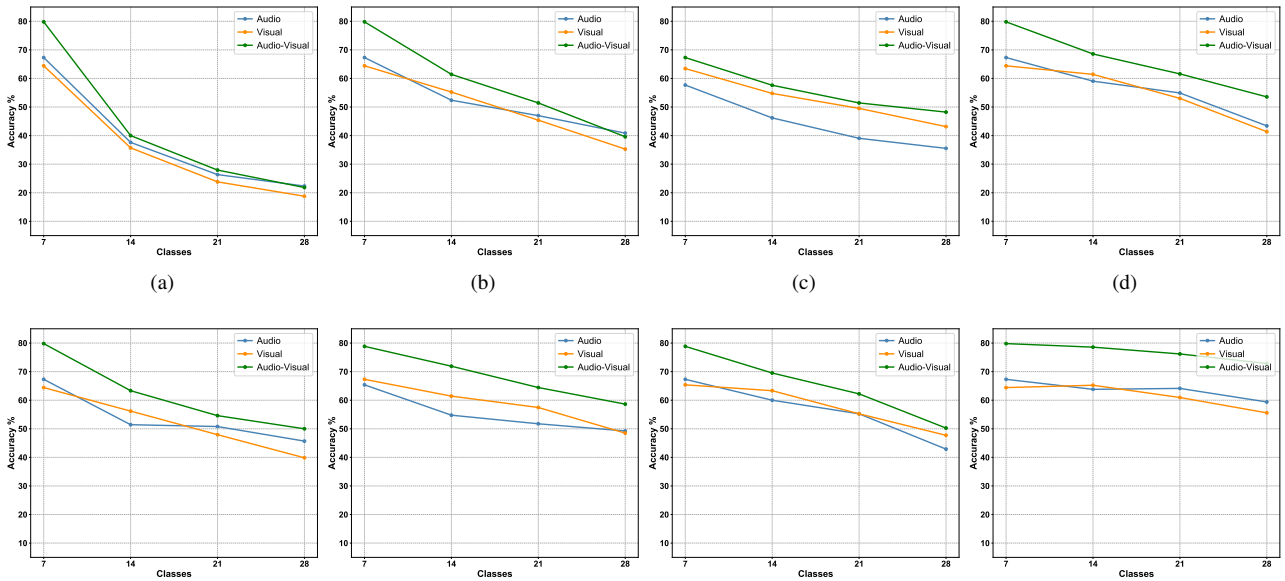

Figure 2: Testing accuracy of training with audio, visual and audio-visual modalities of (a) Fine-tuning, (b) LwF, (c) iCaRL-NME, (d) iCaRL-FC, (e) SS-IL, (f) AFC-NME, (g) AFC-LSC, and (h) upper bound on AVE-CI dataset.

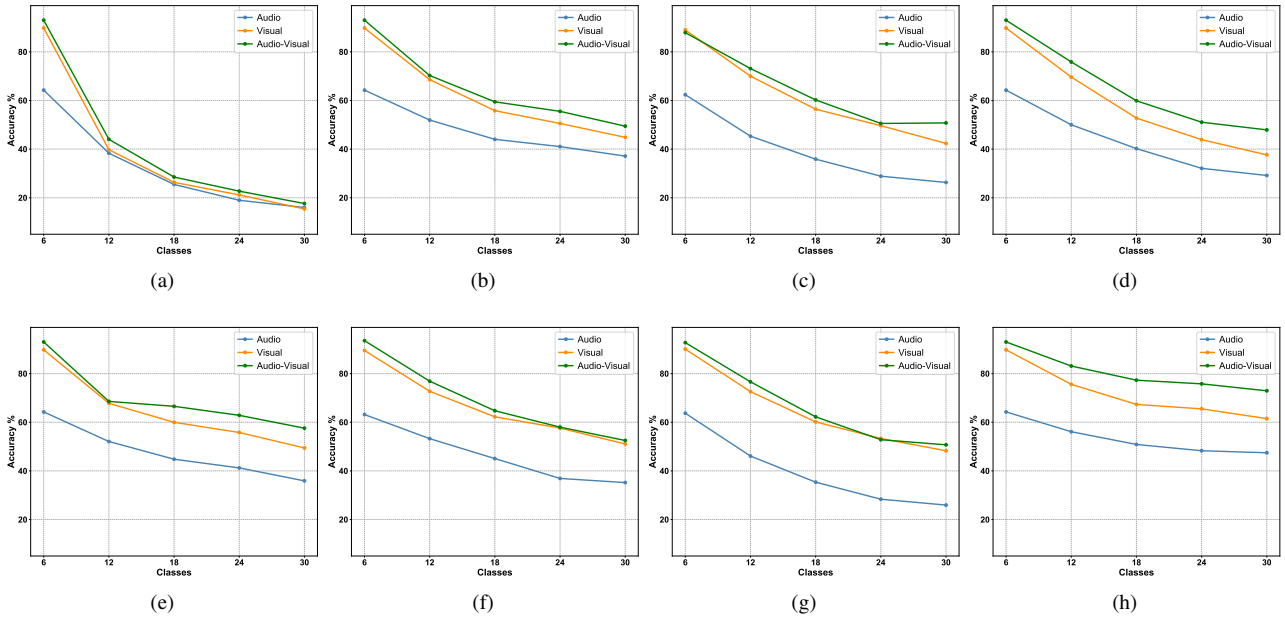

Figure 3: Testing accuracy of training with audio, visual and audio-visual modalities of (a) Fine-tuning, (b) LwF, (c) iCaRL-NME, (d) iCaRL-FC, (e) SS-IL, (f) AFC-NME, (g) AFC-LSC, and (h) upper bound on K-S-CI dataset.

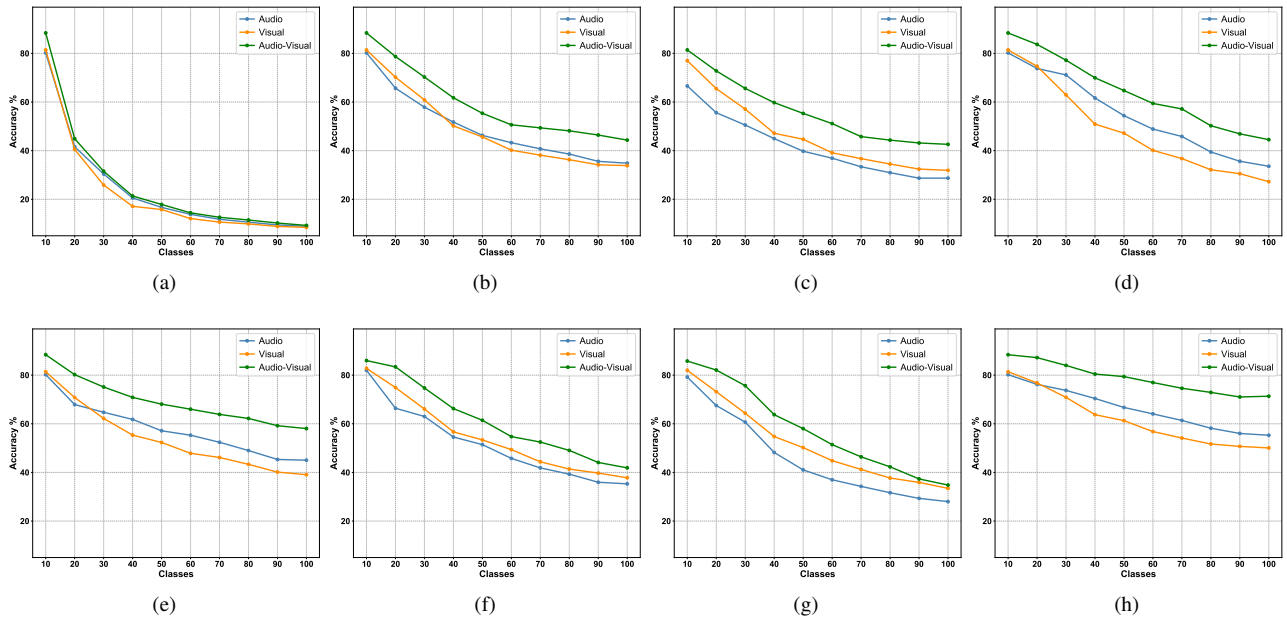

Figure 4: Testing accuracy of training with audio, visual and audio-visual modalities of (a) Fine-tuning, (b) LwF, (c) iCaRL-NME, (d) iCaRL-FC, (e) SS-IL, (f) AFC-NME, (g) AFC-LSC, and (h) upper bound on VS100-CI dataset.

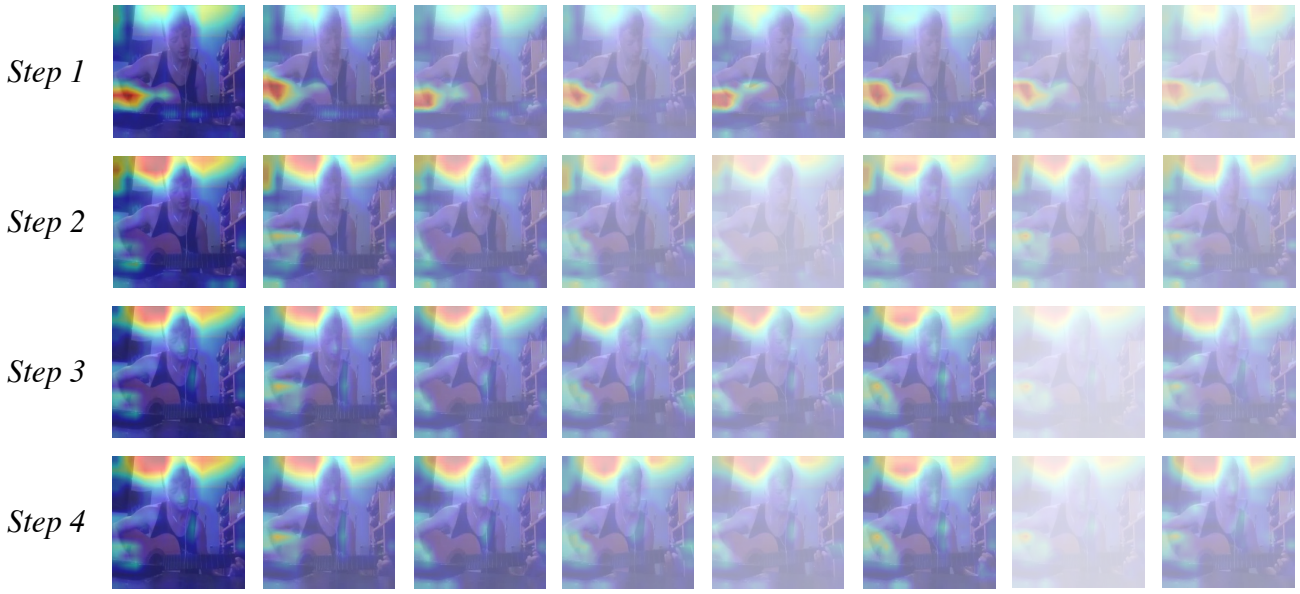

Figure 5: Full version of the visualization of the vanishing of audio-guided visual attention as the incremental step grows.

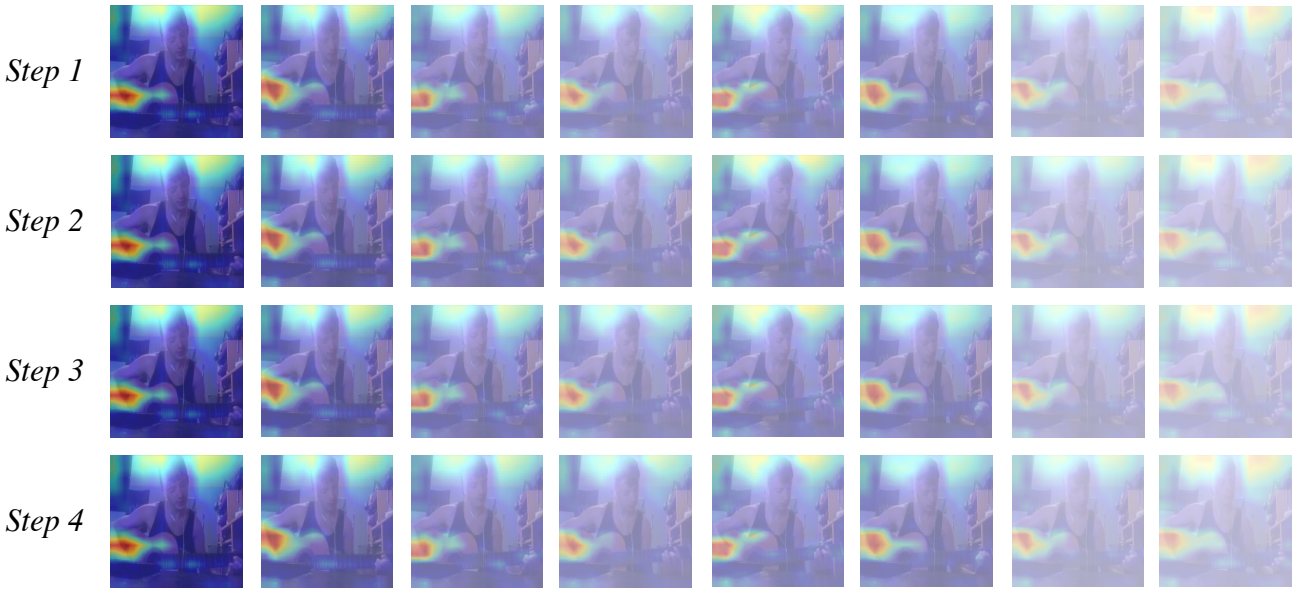

Figure 6: Full version of the visualization of visual attention maps as the incremental step grows with our proposed VAD.