Progressive Homeostatic and Plastic Prompt Tuning for Audio-Visual Multi-Task Incremental Learning

Supplementary Material

A. Tasks and Datasets

In this work, we simulate audio-visual multi-task incremental learning by treating multiple audio-visual tasks as a continuous data stream. We employ multiple different audio-visual understanding tasks, including audio-visual event localization, audio-visual video parsing, audio-visual question answering and audio-visual segmentation.

Audio-Visual Event localization (AVE) [52] is concerned with identifying events within a video that are simultaneously visible and audible across various temporal intervals. We conduct an assessment of the AVE dataset, which includes 4,143 videos across 28 event categories and one background category. These 10-second videos depict diverse scenarios like musical performances.

Audio-Visual Video Parsing (AVVP) [54] aims to parse a video into temporal event sequences and categorize them as auditory, visual, or concurrently audio-visual. We perform experiments on the *Look, Listen, and Parse* (LLP) dataset, which consists of 11,849 10-second video clips across 25 real-life categories. We utilize 10,000 clips with weak annotations for training, and 1,849 clips with detailed annotations for testing and verification.

Audio-Visual Question Answering (AVQA) [26] is to answer questions by leveraging the correlations between visual objects and auditory cues. Experiments are conducted on the *MUSIC-AVQA* dataset, which features over 45,000 Q&A pairs in 9,288 videos totaling 150+ hours.

Audio-Visual Segmentation (AVS) [87] employs the Single Sound Source (S4) subset of AVSBench, comprising 4,932 videos (5 seconds each) with single sound-emitting objects. Each video aligns five 1-second audio clips and image frames, spanning 23 categories (e.g. human voice, instruments), and provides pixel-level annotations.

B. Experimental Setup

Metrics. In this work, we generally use accuracy to represent the performance of our model on three different tasks, which is divided into five metrics: A_{mean}, A_{final}, F_{mean}, A_{single}, and A_{multi}. A_{mean} represents the average accuracy of the task over three incremental settlements. A_{final} indicates the accuracy of the initial task after the incremental process. F_{mean} denotes the average forgetting rate of the initial task after the incremental process. A_{single} expresses the performance when the task is trained individually. A_{multi} is the accuracy of the final task after training on multiple tasks. With these five indicators, our aim is to effectively demonstrate: 1) The performance of the model on a particular task. 2) The degree of forgetting of the model during the multi-task incremental process. 3) The beneficial extent of previous tasks to subsequent tasks during the multi-task incremental process.

Implements details. Our model builds upon the pre-trained CLIP and CLAP architectures as its backbone for handling three audio-visual downstream tasks. Specifically, we utilize a frozen CLIP-trained ViT [11] for visual encoding and a frozen CLAP-trained HTS-AT [7] for audio encoding, leveraging their pre-trained parameters for robust multi-modal feature extraction. The prompts and adapters are strategically injected into both ViT and HTS-AT layers to facilitate audio-visual cross-modal correspondence while preserving knowledge from previous tasks. During training, we set the batch size to 3 and train each task for 10 epochs to ensure convergence. For the optimization process, we employ the Adam optimizer with a learning rate of 3e-4 and a weight decay of 2e-4. The learning rate is scheduled with a cosine decay strategy. We conduct all experiments on a single NVIDIA 3090 GPU with 24GB memory.

C. Normalized Penalty-aware Difference

To better evaluate the improvement of different methods in multi-task learning scenarios, we propose a novel evaluation metric called normalized penalty-aware difference ($Diff$). This metric is designed to address two key challenges in performance evaluation: (1) the difficulty of achieving improvements upon higher baseline performance, and (2) the unfair advantage of methods with lower baseline performance showing larger absolute improvements.

The metric is defined by incorporating both normalized improvement and baseline performance through a quadratic penalty term:

$$Diff = \frac{A_{multi} - A_{single}}{100 - A_{single}} \times (1 + \frac{A_{single}}{100})^2 \times 100\% \quad (1)$$

where A_{multi} and A_{single} represent the performance of multi-task and single-task training respectively. The metric consists of three key components. First, the normalized improvement term $\frac{A_{multi}-A_{single}}{100-A_{single}}$ considers the relative improvement potential by normalizing the performance gain against the remaining improvement headroom. Second, the quadratic baseline penalty $(1 + \frac{A_{single}}{100})^2$ introduces a penalty that grows quadratically with the baseline performance, reflecting the increasing difficulty of achieving improvements as the baseline performance gets higher. Finally, the multiplication by 100% converts the score into a percentage form for better interpretability.

To handle boundary cases where baseline performance approaches 100%, we introduce a small positive number ϵ (e.g., 0.001) to prevent division by zero:

$$Diff = \frac{A_{multi} - A_{single}}{\max(100 - A_{single}, \epsilon)} \times (1 + \frac{A_{single}}{100})^2 \times 100\% \quad (2)$$

In Table 2 in the main text, our metric provides a more nuanced evaluation of different methods' capabilities. For example, while Fine-tune shows significant performance degradation on AVE (from 57.47% to 18.22%) but improvement on AVVP (from 52.64% to 59.00%), its overall Diff score of -58.16% indicates severe negative transfer. By comparison, EWC (-2.87%) and PC (-3.94%) demonstrate more stable performance but still fail to achieve positive knowledge transfer. Notably, despite PC achieving high performance on AVQA (69.85%), its overall transfer capability remains negative. In contrast, our method is the only approach showing positive knowledge transfer (+7.79%), with particularly strong performance on AVE (improving from 70.45% to 72.52%). These results clearly demonstrate that our progressive prompting approach effectively leverages knowledge from previously learned tasks to enhance performance on new tasks while maintaining robustness across both cross-task and cross-modal scenarios.

D. More Detailed Experiments of Comparison Methods

Here we provide more detailed experimental results to demonstrate the performance of different methods under various task orders. We evaluate six methods (Fine-tune, EWC, L2P, S-prompt, Dualprompt, and PC) across six different task ordering scenarios to track their performance through three stages of incremental learning. These comprehensive results allow us to analyze both the forgetting resistance and knowledge transfer capabilities of each method in detail.

Tab.1-Tab.6 present the complete performance metrics for each method. The progression of performance scores reveals distinct patterns in how different approaches handle catastrophic forgetting and knowledge transfer. When examining the performance trajectory, we observe that traditional methods like fine-tuning exhibit severe forgetting, with performance often deteriorating dramatically as new tasks are learned. For instance, in the AVE→AVVP→AVQA sequence, fine-tuning's performance on AVE drops from 56.77% to 19.48% after learning AVVP, and further declines to 17.79% after learning AVQA, demonstrating significant knowledge erosion of the initial task. In contrast, the PC method demonstrates greater resilience to catastrophic forgetting. In the same AVE→AVVP→AVQA sequence, PC maintains the performance on AVE at 54.50% after learning all three tasks, compared to the original 69.90%. However, PC is not immune to forgetting, as evidenced in the AVVP→AVQA→AVE sequence where the performance on AVVP drops from 45.66% to 22.90% after learning all tasks. Notably, the task order significantly impacts performance retention, with certain sequences like AVQA→AVVP→AVE allowing PC to maintain relatively stable performance on the first task (69.44% from an initial 69.72%).

Table 7- 19 provide an extensive analysis of different component configurations and ordering strategies across various task sequences. Table 11 shows the detailed performance of the model without TMA-TMDG-TMI components under different task orders. Table 9- 11 demonstrate the performance of models with TMDG-TMI components, TMA-TMI components, and TMA-TMDG components removed, respectively. Table 12- 14 focus on ablation studies of individual components. Removing TMI (Table 12) affects task-specific representation refinement, while removing TMDG (Table 13) impairs cross-modal integration capabilities. Table 14 shows that without TMA, the model struggles to establish foundational audio-visual correspondences, particularly affecting performance on earlier tasks in the sequence. Table 15- 19 examine different ordering arrangements of our components across network depths: D-M-S (Table 15), M-D-S (Table 16), D-S-M (Table 17), M-S-D (Table 18), and S-D-M (Table 19) configurations, where D represents deep layers, M represents middle layers, and S represents shallow layers. The results consistently demonstrate that our proposed progressive S-M-D ordering (shallow-middle-deep) achieves optimal performance across different metrics and task sequences. This validates our key design principle: universal representations should be learned at shallow layers, task-specific cross-modal features at middle layers, and fine-grained task-modality details at deeper layers. These comprehensive results strongly support our architectural design choices and demonstrate the effectiveness of our progressive prompting strategy for audio-visual multi-task incremental learning.

Table 1. Performance analysis of fine-tuning under different task orders. Each column within an order shows the performance of tasks as they are incrementally learned.

fine-tune

Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	56.77			58.16			51.95		
2	19.48	50.16		17.77	54.27		46.21	36.79	
3	17.79	63.01	54.18	7.69	53.02	59.71	63.01	17.79	54.08
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	53.33			54.16			54.22		
2	63.01	54.21		54.24	18.11		54.00	58.70	
3	13.58	53.16	18.21	52.50	8.98	58.28	54.47	10.05	18.23

Table 2. Performance analysis of EWC under different task orders. Each column within an order shows the performance of tasks as they are incrementally learned. The first task's performance is tracked across all stages to evaluate forgetting resistance, while the final task's performance demonstrates transfer capability.

EWC

Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	17.79			17.79			63.01		
2	9.68	63.01		1.07	54.97		5.51	17.79	
3	1.57	0.92	54.51	3.18	54.95	63.01	0.69	3.93	54.76
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	63.01			54.44			58.95		
2	1.61	52.13		54.45	17.79		59.12	62.74	
3	1.47	52.51	17.74	54.30	4.55	62.32	59.10	62.55	18.26

Table 3. Performance analysis of L2P under different task orders. Each column within an order shows the performance of tasks as they are incrementally learned.

L2P

Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	70.32			69.28			48.69		
2	34.83	48.32		69.35	60.03		35.47	65.45	
3	34.73	48.32	60.20	33.76	60.14	47.22	39.74	65.70	59.87
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	48.74			59.30			60.28		
2	49.20	60.02		59.22	68.11		60.20	49.33	
3	26.71	59.99	68.41	59.31	68.21	42.72	60.28	38.64	66.69

Table 4. Performance analysis of S-prompt under different task orders. Each column within an order shows the performance of tasks as they are incrementally learned.

S-prompt

Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	60.95			63.66			45.66		
2	51.27	42.50		64.08	58.75		34.10	48.71	
3	51.09	42.31	59.85	54.40	58.76	42.96	31.99	17.74	49.29
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	51.17			59.30			58.95		
2	50.94	58.04		59.22	59.75		59.12	52.64	
3	28.27	58.08	54.25	59.31	51.94	42.72	59.10	43.32	54.53

Table 5. Performance analysis of Dualprompt under different task orders. Each column within an order shows the performance of tasks as they are incrementally learned.

Dualprompt

Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	68.21			68.08			46.17		
2	58.26	45.85		67.26	64.69		29.14	68.46	
3	56.84	42.59	64.57	59.35	64.82	46.03	31.07	67.24	63.77
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	45.66			63.48			63.15		
2	46.31	64.06		63.54	68.28		63.19	44.75	
3	25.52	64.06	67.91	63.75	54.35	46.35	63.19	34.97	67.04

Table 6. Performance analysis of PC under different task orders. Each column within an order shows the performance of tasks as they are incrementally learned.

PC

Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	69.90			69.43			44.42		
2	55.50	45.89		68.38	69.24		36.90	66.57	
3	54.50	45.53	69.86	54.38	69.04	44.56	39.06	66.97	69.84
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	45.66			69.35			69.72		
2	45.48	69.67		69.51	71.00		69.78	44.47	
3	22.90	69.62	68.86	69.47	56.27	44.65	69.44	30.61	67.21

Table 7. Performance analysis of DCNet under different task orders. Each column within an order shows the performance of tasks as they are incrementally learned.

DCNet

Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	58.13			60.52			54.34		
2	20.62	48.83		17.79	54.25		46.73	38.06	
3	17.79	63.01	53.93	7.71	53.57	59.48	63.01	17.79	54.01
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	50.07			54.23			54.13		
2	63.01	54.37		54.47	17.71		53.14	59.57	
3	15.33	50.38	18.28	54.20	5.92	55.71	53.52	7.71	19.83

Table 8. Ablation study: Detailed Performance of model without TMA-TMDG-TMI components under different task orders. Each column within an order shows the performance of tasks as they are incrementally learned.

w/o TMA-TMDG-TMI

Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	70.67			70.87			48.14		
2	57.46	45.80		70.62	69.39		32.35	69.03	
3	58.96	45.76	69.82	57.56	69.61	46.49	34.33	69.25	70.17
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	46.90			69.28			69.85		
2	47.04	69.94		69.60	69.78		69.48	47.64	
3	35.34	69.98	68.06	69.36	57.11	45.20	69.76	30.29	70.80

Table 9. Ablation study: Detailed Performance without TMDG-TMI components under different task orders.

w/o TMDG-TMI									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	68.91			69.45			45.53		
2	60.25	46.49		68.88	70.06		29.88	69.70	
3	62.79	51.95	70.17	59.65	69.72	48.55	35.57	63.48	70.27
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	46.81			69.72			70.08		
2	49.34	70.71		69.27	71.49		69.53	48.97	
3	34.19	70.55	70.99	69.10	59.65	46.77	69.10	34.37	70.82

Table 10. Ablation study: Detailed Performance without TMA-TMI components under different task orders.

w/o TMA-TMI									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	70.03			71.19			46.95		
2	60.32	47.45		70.97	69.21		43.97	70.20	
3	61.77	47.13	70.07	59.15	69.30	47.32	39.70	69.88	69.49
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	45.76			69.69			69.57		
2	45.53	69.47		69.36	69.13		69.73	48.28	
3	32.58	69.65	69.58	69.65	59.58	46.58	69.64	30.70	69.83

Table 11. Ablation study: Detailed performance without TMA-TMDG components under different task orders.

w/o TMA-TMDG									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	71.79			70.65			50.12		
2	62.31	49.01		70.57	69.55		39.01	71.42	
3	63.06	48.55	68.76	55.30	69.69	49.29	41.53	71.12	69.20
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	48.88			69.95			69.08		
2	48.60	68.99		69.53	72.26		69.15	50.02	
3	34.65	68.90	71.37	69.69	62.36	48.23	70.81	36.99	73.21

Table 12. Ablation study: Detailed Performance without TMI component under different task orders.

w/o TMI									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	69.65			68.09			48.88		
2	58.16	47.50		68.33	69.68		45.07	71.37	
3	59.78	48.92	70.47	57.81	69.54	48.80	45.98	66.34	70.09
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	47.87			70.10			70.05		
2	50.12	70.02		69.44	71.29		69.81	50.21	
3	33.73	69.92	71.22	69.55	57.44	47.18	69.23	36.90	70.95

Table 13. Ablation study: Detailed Performance without TMDG component under different task orders.

w/o TMDG									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	69.25			69.85			49.52		
2	58.38	47.73		68.33	70.09		40.71	70.52	
3	58.33	49.24	70.94	56.19	69.93	50.53	42.59	63.61	70.68
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	47.41			70.17			70.16		
2	46.54	71.01		69.41	73.04		68.83	48.88	
3	35.66	70.39	72.96	69.49	62.24	48.97	68.93	35.20	71.47

Table 14. Ablation study: Detailed Performance without TMA component under different task orders.

w/o TMA									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	70.47			70.17			49.38		
2	63.31	49.11		70.37	70.01		31.02	69.75	
3	61.00	48.10	69.73	60.42	69.81	49.01	37.08	69.58	69.95
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	50.12			70.13			69.56		
2	50.02	69.77		70.05	71.69		69.43	49.75	
3	34.88	69.66	71.19	69.84	60.95	48.69	69.81	30.93	70.67

Table 15. Ablation study: Detailed Performance with D-M-S component ordering under different task orders.

D-M-S									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	51.12			51.82			52.69		
2	31.57	51.17		18.21	63.36		35.89	50.65	
3	14.42	59.29	63.97	26.19	56.71	50.76	43.23	18.31	63.94
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	51.72			63.52			63.43		
2	63.10	63.91		58.20	50.03		58.42	53.60	
3	46.90	60.08	51.64	56.67	31.49	50.71	59.41	35.75	51.57

Table 16. Ablation study: Detailed Performance with M-D-S component ordering under different task orders.

M-D-S									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	55.10			54.81			48.60		
2	34.55	52.32		18.07	63.94		31.99	52.84	
3	12.51	63.01	64.52	29.58	58.89	55.21	62.92	19.23	64.51
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	50.48			63.22			63.51		
2	63.39	63.97		58.55	52.81		61.02	55.12	
3	37.31	60.97	51.54	61.25	36.42	51.63	61.46	40.29	55.05

Table 17. Ablation study: Detailed Performance with D-S-M component ordering under different task orders.

D-S-M									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	56.72			57.44			51.26		
2	37.76	50.90		16.00	62.67		44.21	56.84	
3	18.98	63.01	54.44	30.67	57.64	53.69	51.10	19.35	63.31
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	50.80			54.34			54.44		
2	62.23	64.80		54.34	50.03		54.44	52.96	
3	40.16	61.06	54.18	54.34	31.49	50.94	54.44	38.18	55.72

Table 18. Ablation study: Detailed Performance with M-S-D component ordering under different task orders.

M-S-D									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	55.47			56.39			52.78		
2	39.29	52.13		17.74	62.55		40.80	56.89	
3	15.60	60.53	63.01	22.91	56.70	54.75	61.96	17.71	62.56
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	55.12			62.83			62.70		
2	61.40	62.94		62.55	46.07		61.10	53.51	
3	36.53	60.05	53.31	58.27	32.99	54.89	61.12	36.12	51.07

Table 19. Ablation study: Detailed Performance with S-D-M component ordering under different task orders.

S-D-M									
Stage	AVE→AVVP→AVQA			AVE→AVQA→AVVP			AVVP→AVE→AVQA		
1	46.77			48.31			53.01		
2	35.70	53.74		31.52	62.91		32.31	50.32	
3	28.06	53.83	62.70	28.88	60.49	54.20	51.77	30.70	62.95
Stage	AVVP→AVQA→AVE			AVQA→AVE→AVVP			AVQA→AVVP→AVE		
1	52.50			63.60			62.65		
2	56.13	61.54		59.11	45.15		60.05	52.13	
3	28.68	60.64	48.26	61.20	35.77	52.59	59.89	49.47	49.20