
Titans-as-a-Layer: Test-Time Memory for Conversational Speech Emotion Recognition

Daniel Chen¹ Qicong Hu¹ Yang Xiao² Ting Dang² Hong Jia¹

Abstract

Speech emotion recognition (SER) is commonly formulated as utterance-level classification, although conversational emotion depends on a speaker’s usual vocal range and the emotional context established by previous utterances. Speech-language models provide strong pretrained acoustic and semantic representations, and can adapt them to SER labels via finetune, but this mechanism still missing per-dialogue state. We study whether test-time neural memory can supply this missing context while leaving the large audio language models (LALMs) backbone intact. Building on Titans, we introduce a plug-and-play Memory-as-a-Layer (MAL) adapter that writes dialogue history into a small neural memory and reads it back as an audio-token-aligned residual update, avoiding changes to the host model’s token positions. Across different audio LLMs and emotion recognition datasets evaluations, our design improves SER performs across different evaluation metrics, supporting test-time memory as a residual contextual mechanism for conversational SER.

1. Introduction

Speech emotion is conveyed not only by lexical content, but also by the manner in which that content is produced. In conversational speech emotion recognition (SER), the acoustic evidence supporting an emotion label is therefore rarely contained within an isolated utterance. Prosodic and paralinguistic cues, including pitch, intensity, speaking rate and voice quality, must be interpreted relative to the speaker’s *recent vocal behaviour* and the *affective evolution of the*

dialogue. We refer to this short-term vocal reference as the speaker baseline. For instance, an increase in pitch may signal anger for one speaker, enthusiasm for another, or no salient affective change if it lies within that speaker’s recent expressive range. Conversational SER is thus inherently both local and contextual: it depends on the current waveform while also requiring information from preceding utterances in the same interaction.

Large-scale speech pretraining and audio–text alignment provide a strong foundation for SER, as contemporary speech-language models encode both linguistic content and non-lexical attributes such as prosody, energy and speaking style (Radford et al., 2023; Tang et al., 2024; Chu et al., 2024; Kong et al., 2024). In standard SER pipelines, however, these representations are commonly used at the utterance level: each speech segment is encoded and classified independently (Jia et al., 2025; Zhang et al., 2025). Consequently, evidence that is distributed across speakers and dialogue history cannot be retained intrinsically, and must instead be introduced through explicit context tokens, concatenated inputs, or an additional mechanism for maintaining dialogue state.

LoRA (Hu et al., 2022) efficiently adapts LALMs to SER without full-model fine-tuning, but the learned update is static and dialogue-independent. It improves task alignment, yet cannot maintain speaker baselines, affective trajectories, or previous-utterance information. Extending the attention context may also disturb the positional structure learned during pretraining. These limitations motivate inference-time memory with a fixed backbone. Titans (Behrouz et al., 2024) treats memory as a compact neural module updated during inference, storing prior information as model state rather than explicit context tokens. For conversational SER, this separates task adaptation from dialogue-state modelling: LoRA aligns the model to emotion recognition, while test-time memory preserves dialogue-specific evidence for subsequent utterances.

Motivated by these, we introduce a plug-and-play *Memory-as-a-Layer* (MAL) adapter based on Titans for LALMs. MAL inserts a Titans-style neural memory branch within the language-model stack and injects its output as an audio-token-aligned residual update. The adapter does not alter

¹Department of Computer Science, University of Auckland, Auckland, New Zealand ²School of Computer and Information Technology, Melbourne, Australia. Correspondence to: Ting Dang <ting.dang@unimelb.edu.au>.

Published at ICML 2026 Workshop on the Impact of Machine Learning for Audio, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

the pretrained LALMs: it neither prepends memory tokens to the LALMs input, nor shifts audio-token positions, nor requires the backbone to process the dialogue history as an extended context window. Instead, dialogue information is written into and retrieved from a separate memory state. Across multiple datasets, LALMs backbones and evaluation metrics, MAL consistently improves conversational SER performance, demonstrating that test-time memory provides a robust mechanism for incorporating dialogue-level context without compromising the pretrained model interface.

Our contributions are:

- We introduce a Titans-style memory branch that updates audio-token representations in place, preserving the host model’s token layout while adding an explicit pathway for dialogue-level state.
- We introduce a two-stage optimisation scheme in which LoRA first provides task adaptation for SER, after which the LoRA-tuned model is frozen and MAL is trained to supply dialogue-specific contextual refinement. This design encourages memory to act as an additive state mechanism rather than as a competing task adapter.
- We evaluate multiple LALMs across diverse SER datasets and metrics, demonstrating that the proposed method achieves state-of-the-art SER performance.

2. Related Work

2.1. Memory in large language models

Existing memory mechanisms for sequence models address the limited context available to standard Transformers. Transformer-XL introduces segment-level recurrence (Dai et al., 2019); Compressive Transformer stores compressed representations of past activations (Rae et al., 2020); Recurrent Memory Transformer uses learned memory tokens (Bulatov et al., 2022); and linear-attention variants improve the efficiency of long-range sequence modeling (Yang et al., 2024). Titans (Behrouz et al., 2024) extends this direction by representing long-term memory as a small neural module whose parameters are updated during inference, together with sliding-window attention and learned persistent memory tokens. Titans-style memory has also been adapted beyond text: VideoTitan (Park et al., 2025) applies the mechanism to video sequence modeling and reports strong results on long video benchmarks, including WeatherBench (Rasp et al., 2020). However, these approaches do not directly address frozen speech-language models for conversational SER, where the memory mechanism must preserve the audio-token interface while carrying information across utterances.

2.2. Static task adapters and adaptive memory.

Adapter-based fine-tuning methods such as LoRA (Hu et al., 2022) adapt pretrained models by adding trainable low-rank updates while keeping the backbone fixed. For SER, such adapters can learn a task-specific mapping from pretrained speech representations to emotion labels with far fewer trainable parameters than full fine-tuning. This makes LoRA an appropriate task-adaptation baseline for large speech-language models. However, a trained adapter is static at test time: the same learned update is applied to every dialogue, regardless of the speaker’s vocal range or the preceding emotional context. This leaves open the question of whether an adaptive memory component can add dialogue-specific state after task adaptation has been learned.

3. Method

3.1. Test-time memory

We consider conversational speech emotion recognition, where a dialogue is an ordered utterance stream $\mathcal{D} = (x_1, \dots, x_N)$ with utterance-level labels (y_1, \dots, y_N) . Each utterance x_i is encoded into T_i audio-token embeddings and inserted into the language-model sequence at the audio-token positions.

We use *test-time memory* to denote a dialogue-specific state that is updated online during inference while the backbone and trained adapters remain fixed. For utterance x_i , the model predicts using the current utterance and the memory accumulated from previous utterances:

$$\hat{y}_i = f_{\Theta}(x_i, \mathcal{S}_{i-1}), \quad \mathcal{S}_i = U_{\Phi}(\mathcal{S}_{i-1}, x_i). \quad (1)$$

Here \mathcal{S}_i is the dialogue memory state after utterance x_i , f_{Θ} is the fixed inference model, and U_{Φ} is the learned memory update rule. The memory state is reset at dialogue boundaries. At test time, no model parameters are optimized; only the dialogue-specific state \mathcal{S}_i evolves.

For MAL, the memory state is layer-specific:

$$\mathcal{S}_i = \{\mathcal{S}_i^{(\ell)}\}_{\ell=0}^{L-1}, \quad (2)$$

where $\mathcal{S}_i^{(\ell)}$ is the NeuralMemory state for language-model block ℓ after processing utterance x_i .

3.2. Memory integration ablation

We compare three ways of integrating a Titans memory branch into a frozen LALM. This ablation uses Ultra-vox v0.4 on IEMOCAP and attaches a single memory branch, so that the comparison isolates the injection mechanism. Let $h \in \mathbb{R}^{T \times D}$ be the audio-token hidden states in the language-model residual stream, let $P \in \mathbb{R}^{N_p \times D}$ be persistent memory tokens, and let \mathcal{M} be a Titans NeuralMemory module. We write $[\cdot; \cdot]$ for sequence concatenation

and $[N_p+1:]$ for removing the persistent-token prefix. The audio-aligned memory output is

$$m = \mathcal{M}([P; h])[N_p+1:], \quad m \in \mathbb{R}^{T \times D}. \quad (3)$$

Memory as Context (MAC) prepends memory to the language-model input:

$$H_{\text{MAC}} = [P; m; h]. \quad (4)$$

This increases the sequence length and shifts the positional indices of the original tokens.

Memory as Gating (MAG) keeps the sequence length fixed and gates the residual stream:

$$H_{\text{MAG}} = h \odot \sigma(W_g m), \quad (5)$$

where W_g is learned, $\sigma(\cdot)$ is the sigmoid function, and \odot denotes elementwise multiplication.

Memory-as-a-Layer (MAL) instead applies an additive memory update to the audio-token hidden states while preserving the original token positions. The zero-shot baseline obtains 44.52 WF1. MAC improves this to 49.90 WF1, MAG obtains 50.32 WF1, and MAL reaches 57.48 WF1. Therefore, we use MAL as the main memory integration strategy.

3.3. Memory-as-a-Layer (MAL) for LALM

For utterance x_i and language-model block ℓ , let $h_{i,\ell} \in \mathbb{R}^{T_i \times D}$ denote the hidden states at the audio-token positions before the block. MAL modifies these states in place: it does not add, remove, or reorder tokens in the host sequence. The memory branch uses an internal dimension d_m , so the audio states are projected into memory space and projected back afterward.

For each block ℓ , MAL computes

$$z_{i,\ell} = [P_\ell; W_{\text{in}}^{(\ell)} h_{i,\ell}], \quad (6)$$

$$(r_{i,\ell}, S_i^{(\ell)}) = \mathcal{M}_\ell(z_{i,\ell}; S_{i-1}^{(\ell)}), \quad (7)$$

$$\delta_{i,\ell} = W_{\text{out}}^{(\ell)} r_{i,\ell}[N_p+1:], \quad (8)$$

$$\tilde{h}_{i,\ell} = h_{i,\ell} + \tanh(\alpha_\ell) \delta_{i,\ell}. \quad (9)$$

Here $W_{\text{in}}^{(\ell)}: D \rightarrow d_m$ and $W_{\text{out}}^{(\ell)}: d_m \rightarrow D$ are learned projections, $P_\ell \in \mathbb{R}^{N_p \times d_m}$ are persistent memory tokens, and \mathcal{M}_ℓ is a Titans NeuralMemory module (Behrouz et al., 2024). The module reads the previous dialogue state $S_{i-1}^{(\ell)}$, processes the current memory input $z_{i,\ell}$, returns memory output $r_{i,\ell}$, and updates the state to $S_i^{(\ell)}$ for subsequent utterances.

After the memory branch, we discard the N_p persistent-token prefix and keep only the T_i audio-aligned outputs. The

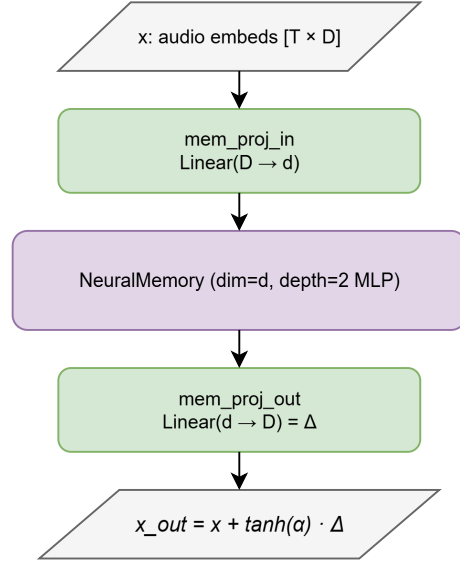


Figure 1. Memory-as-a-Layer branch architecture. Audio embeddings are projected down to the memory dimension d , passed through a Titans NeuralMemory module (depth-2 MLP), projected back to D , and added to the original embeddings through a zero-initialised residual gate $\tilde{h} = h + \tanh(\alpha) \cdot \delta$.

scalar gate $\alpha_\ell \in \mathbb{R}$ is initialized to zero, so $\tanh(\alpha_\ell) = 0$ at initialization and $\tilde{h}_{i,\ell} = h_{i,\ell}$. Thus, adding MAL initially preserves the frozen host computation.

We attach independent MAL branches to every language-model layer. In the placement ablation, dense every-layer placement with smaller branch capacity ($d_m=128$, $N_p=8$) outperformed sparser placements, such as every-8th layer at quartile positions $\{7, 15, 23, 31\}$ with $d_m=256$.

Trainable parameters. The trainable MAL parameters are

$$\{W_{\text{in}}^{(\ell)}, W_{\text{out}}^{(\ell)}, \mathcal{M}_\ell, \alpha_\ell, P_\ell\}_{\ell=0}^{L-1}. \quad (10)$$

All backbone, audio-encoder, and projector parameters remain frozen.

3.4. Sequential training schedule

We use a sequential schedule to separate static task adaptation from dialogue-state memory learning. Jointly optimizing LoRA and MAL did not improve over the LoRA-only baseline, likely because both modules modify the same residual stream while only MAL maintains state across utterances. Diagnostic ablations showed that removing LoRA after joint training had little effect, whereas removing Titans memory substantially degraded performance. We therefore train in two phases.

Phase 1: task adaptation. We train LoRA with MAL removed, so the model learns the SER decision boundary without memory updates. LoRA is trained for three epochs

Table 1. Model Performance Results Across Benchmarks. **Bold** = best and underline = second-best within each fully evaluated backbone-dataset block.

Model	IEMOCAP			MELD		Multidialog	
	WA (%)	UA (%)	WF1 (%)	WF1 (%)	Macro-F1 (%)	WF1 (%)	Macro-F1 (%)
Qwen2-Audio	58.92	59.82	58.25	39.05	20.56	28.27	17.65
+ LoRA	<u>82.40</u>	<u>82.61</u>	<u>82.42</u>	<u>54.45</u>	<u>38.32</u>	<u>56.37</u>	<u>37.16</u>
+ Titans+LoRA	83.66	84.20	83.67	56.84	40.99	57.54	37.94
Audio Flamingo 3	76.57	76.45	76.31	45.14	30.11	31.64	21.08
+ LoRA	<u>84.59</u>	<u>85.02</u>	<u>84.57</u>	<u>57.81</u>	<u>44.43</u>	<u>55.11</u>	<u>34.78</u>
+ Titans+LoRA	85.21	85.37	85.20	58.18	44.66	56.07	35.82
Ultravox-v0.4	45.29	47.09	44.53	25.33	22.70	34.24	19.78
+ LoRA	<u>63.89</u>	<u>62.84</u>	<u>63.78</u>	<u>48.58</u>	<u>31.97</u>	<u>54.85</u>	<u>30.74</u>
+ Titans+LoRA	65.49	65.43	65.61	49.63	32.64	55.04	33.32

WA = Weighted Accuracy; UA = Unweighted Accuracy; WF1 = Weighted F1.

with rank $r=32$, scaling $\alpha_{\text{LoRA}}=64$, AdamW, and learning rate 2×10^{-5} .

Phase 2: memory training. We select the best LoRA checkpoint, then freeze LoRA, the LALM backbone, audio encoder, and projector. MAL is trained for two epochs while all other components remain fixed. The projection and NeuralMemory parameters $\{W_{\text{in}}, W_{\text{out}}, \mathcal{M}\}$ use learning rate 1×10^{-4} and gradient clipping threshold 1.0; the residual gates $\{\alpha_{\ell}\}$ use learning rate 2×10^{-4} and clipping threshold 2.0.

4. Experimental setup

4.1. Datasets

We use three conversational speech emotion recognition datasets for the evaluated settings. For IEMOCAP (Busso et al., 2008), we used its standard four-class configuration (angry/happy/neutral/sad), with 5,531 utterances across 5 sessions and two consistent speakers per session, evaluated under leave-one-session-out (LOSO) cross-validation. For MELD (Porcia et al., 2019), we tested on the seven-class multi-party TV-dialogue setting with the standard train/dev/test split, totalling 13,706 utterances. For MultiDialog (Park et al., 2024), we used the gold subset of 9 emotion-accurate actors, containing 6,681 dialogues and 47,004 utterances with the same seven-class schema as MELD.

4.2. Models

We use three models to assess whether the memory effect persists across different backbone designs: Ultravox v0.4 (Fixie.ai, 2024), Qwen2-Audio (Chu et al., 2024), and Audio Flamingo 3 (Goel et al., 2025).

5. Results

The results in Table 1 show that incorporating Titans consistently improves performance over both the original backbone models and standard LoRA fine-tuning across the fully evaluated settings. On IEMOCAP, Titans+LoRA achieves the best results for all three backbones, improving over LoRA by up to 1.60% in WA, 2.59% in UA, and 1.83% in WF1. These gains are particularly evident for Ultravox-v0.4, where Titans+LoRA increases WF1 from 63.78% to 65.61%, indicating that Titans provides additional modelling benefits beyond parameter-efficient fine-tuning alone. Similar improvements are observed on MELD, where Titans+LoRA improves Audio Flamingo 3 from 57.81% to 58.18% WF1 and from 44.43% to 44.66% Macro-F1, while also improving Ultravox-v0.4 from 48.58% to 49.63% WF1 and from 31.97% to 32.64% Macro-F1. On Multidialog, Titans+LoRA further improves Ultravox-v0.4, especially in Macro-F1, increasing performance from 30.74% to 33.32%. This suggests that Titans is particularly effective in improving recognition of underrepresented classes. Overall, the consistent gains across datasets, metrics, and backbone architectures demonstrate that Titans complements LoRA by enhancing the model’s ability to capture task-relevant emotional and conversational representations.

6. Conclusion

Through this work, we study whether a NeuralMemory module (Titans) can augment frozen, LoRA-tuned LALMs with cross-utterance memory at test time. The evaluated settings show additive gains over the corresponding LoRA baselines, with the clearest evidence on IEMOCAP and additional support from the evaluated MELD runs. The key finding is that a sequential training schedule improves mem-

ory in a LoRA-tuned model for speech emotion recognition. These results support test-time memory as a lightweight complement to parameter-efficient fine-tuning for conversational emotion recognition, while broader coverage across all backbone–dataset combinations and comparisons against other memory architectures remain important next steps.

References

- Behrouz, A., Zhong, P., and Mirrokni, V. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.
- Bulatov, A., Kuratov, Y., and Burtsev, M. S. Recurrent memory transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 11079–11091, 2022.
- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4): 335–359, 2008.
- Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., and Zhou, J. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2978–2988, 2019.
- Fixie.ai. Ultravox: An open-source speech–language model. <https://github.com/fixie-ai/ultravox>, 2024. v0.4 release.
- Goel, A., Ghosh, S., Kim, J., Kong, Z., Kumar, S.-g., Lee, S.-g., Valle, R., Ping, W., and Catanzaro, B. Audio Flamingo 3: Advancing audio intelligence with fully open large audio–language models. *arXiv preprint arXiv:2507.08128*, 2025.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.
- Jia, H., Fu, S., Xia, F., Kostakos, V., and Dang, T. Beyond scale: Small language models are comparable to gpt-4 in mental health understanding. *arXiv preprint arXiv:2507.08031*, 2025.
- Kong, Z., Goel, A., Badlani, R., Ping, W., Valle, R., and Catanzaro, B. Audio Flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- Park, S. J., Kim, S., Choi, Y., Lee, H.-G., Kang, J.-H., Shin, J. W., and Kim, S.-G. Let’s go real talk: Spoken dialogue model for face-to-face conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16334–16348. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.860. URL <https://aclanthology.org/2024.acl-long.860>.
- Park, Y.-J., Seo, M., and Jeon, H.-G. VideoTitans: Scalable video prediction with integrated short- and long-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL <https://openreview.net/forum?id=86enCXORIV>.
- Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 527–536, 2019.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.
- Rae, J. W., Potapenko, A., Jayakumar, S. M., Hillier, C., and Lillicrap, T. P. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations (ICLR)*, 2020.
- Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.
- Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., and Zhang, C. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.
- Yang, S., Wang, B., Shen, Y., Panda, R., and Kim, Y. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- Zhang, T., Xue, X., Ruan, L., Fu, S., Xia, F., D’Alfonso, S., Kostakos, V., Dang, T., and Jia, H. Menta: A small language model for on-device mental health prediction. *arXiv preprint arXiv:2512.02716*, 2025.