PRESERVING ATTENTION FUNCTIONS FOR REPLAY-FREE CONTINUAL SPEECH RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual adaptation aims to learn from evolving data distributions while retaining previously acquired knowledge. In automatic speech recognition (ASR), variations in noise, recording conditions, and speaking styles make it difficult to adapt to new domains without degrading recognition on earlier ones, particularly when past utterances and transcripts are unavailable. Constraining updates on representations shared across speech domains can also restrict useful adaptation. We propose preserving attention functions (PAF), a method for replay-free continual ASR that regularizes composite attention operators during low-rank adaptation (LoRA). After each task, PAF summarizes attention inputs using their means and principal component bases. On the resulting affine supports, it jointly regularizes a score operator formed by query and key projections for context selection and a transport operator formed by value and output projections for information transformation. This objective permits coordinated changes in individual projections that leave the composite operators unchanged on the stored supports. PAF stores score references directly and reconstructs transport references from the merged model and a stored history of low-rank updates to the value and output projections. Preservation is evaluated without past examples or a separate teacher model. The learned updates are merged into a single ASR model, requiring no task identifiers or additional modules at inference. Experiments across multiple domain sequences and ASR models show lower final average word error rates than the reported replay-free baselines and reduced forgetting relative to sequential LoRA adaptation.

1 INTRODUCTION

Continual adaptation enables a model to learn from changing data distributions while retaining previously acquired knowledge (Kirkpatrick et al., 2017; Zenke et al., 2017; Saha et al., 2021). This problem is central to automatic speech recognition (ASR), where deployed systems encounter variations in noise, recording conditions, and speaking styles (Sim et al., 2018). Pretraining at scale improves generalization (Chan et al., 2021; Radford et al., 2023; Hsu et al., 2021). Unfamiliar acoustic conditions continue to motivate adaptation (Dong et al., 2025; Lin et al., 2022; Lee & Chang, 2024). Adapting to each new domain can degrade recognition on earlier ones (Chang et al., 2021; Takashima et al., 2022; Plantinga et al., 2023; Lee et al., 2022; Vander Eeckt & Van hamme, 2023). We study replay-free continual ASR, in which adaptation uses only data from the current domain without retaining previous utterances or transcripts, and inference uses a single model without task identifiers.

Parameter efficiency in continual learning has been studied extensively for large language models (Wang et al., 2023; Zhao et al., 2024; Qin & Joty, 2022; Razdaibiedina et al., 2023). Low-rank adaptation (LoRA) reduces the number of trainable parameters and allows learned updates to be merged into the pretrained model (Hu et al., 2022). Sequential LoRA updates can cause forgetting (Li et al., 2025; Qiao & Mahdavi, 2026). Existing methods control interference through the adaptation space or the combination of learned branches: O-LoRA encourages orthogonality between task subspaces (Wang et al., 2023), InfLoRA constructs subspaces to balance adaptation and retention (Liang & Li, 2024), and C-LoRA penalizes elementwise overlap between accumulated previous updates and the current LoRA update (Smith et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

GainLoRA introduces gates that reduce the contribution of a new LoRA branch to earlier inputs (Liang et al., 2025). Its authors acknowledge that accumulating constraints can hinder new learning and identify overlapping tasks as a setting requiring further investigation. This issue is relevant to ASR because different acoustic domains can share representations useful for recognition (Meng et al., 2017; 2018). When these representations also support adaptation to the current domain, suppressing new responses to them can limit useful adaptation. This motivates a preservation criterion that evaluates the effect of an update on earlier computations while allowing learning through shared representations.

Speech recognition requires contextual acoustic processing and the association of acoustic representations with output symbols. Self-attention integrates acoustic or linguistic context, and cross-attention links the two (Gulati et al., 2020; Karita et al., 2019; Chan et al., 2016; Sperber et al., 2018; Chorowski et al., 2015; Shim et al., 2022). In cross-attention, QK determines which acoustic context contributes at each output position, and VO transforms the selected context (Vaswani et al., 2017; Mittal et al., 2022). Joint regularization permits coordinated projection changes, including along shared input directions, that preserve these compositions on retained supports (Tran et al., 2025). Fixed feature coordinates give operator dimensions independent of utterance length, with attention weights computed separately for each utterance. This connects retention of earlier acoustic context selection and transformation with adaptation through shared speech representations. Applying this criterion without past utterances requires a representation of earlier attention computations that can be evaluated from retained summaries. Attention distillation methods such as SEEKR use replayed inputs to evaluate past behavior (He et al., 2024), while Attention-Spectrum preserves spectral statistics of attention maps between modalities without replay (Zhao et al., 2026).

We propose preserving attention functions (PAF), which records attention input means and principal component bases after each domain and expresses QK and VO operators on the resulting affine supports. Adaptation penalizes changes in these fixed coordinates using stored score references and transport references reconstructed from the merged model and low-rank update histories. PAF trains Q/K/V/O LoRA parameters and merges the updates for inference without task identifiers or additional modules. Experiments across domain sequences and ASR models show improved final recognition and retention. Component comparisons support joint preservation, while a controlled intervention identifies a learned projection component that supports new domain prediction and leaves the composite operators on stored supports unchanged.

2 Background

2.1 Replay-Free Continual ASR with LoRA

We study speech domains $\{\mathcal{D}_t\}_{t=1}^{T}$ containing paired speech and transcripts, where T is the number of domains. At stage t, training uses only $\mathcal{D}_t$, while evaluation covers all domains, including unseen ones. Previous utterances, transcripts, individual features, and logits are unavailable; retained information comprises input statistics, operator summaries, and parameter update histories. Inference uses one merged model without domain identifiers. For each attention projection $X \in \{Q, K, V, O\}$, LoRA adds a low-rank update to the previously merged weight as follows:

$$W_{X,\theta} = W_X^{[t-1]} + \gamma A_{X,t} B_{X,t}, \tag{1}$$

where r is the LoRA rank, α_{LoRA} is the scaling hyperparameter, and $\gamma = \alpha_{\mathrm{LoRA}}/r$. The factors have dimensions $A_{X,t} \in \mathbb{R}^{d_{\mathrm{out},X} \times r}$ and $B_{X,t} \in \mathbb{R}^{r \times d_{\mathrm{in},X}}$, where $d_{\mathrm{in},X}$ and $d_{\mathrm{out},X}$ are the input and output dimensions of projection X, respectively. Only the current LoRA factors, denoted collectively by θ, are trained; merged weights, biases, and all remaining parameters are frozen. Preservation is evaluated using these deterministic effective weights.

2.2 Attention as a Composite Function

Using column vectors, let $x \in \mathbb{R}^{d_q}$ be a query input and $z_j \in \mathbb{R}^{d_{kv}}$ a key/value input. For head h, we define the query, key, and value projections as follows:

$$q_\theta(x) = W^h_{Q,\theta} x + b^h_Q, \tag{2}$$

$$k_\theta(z_j) = W^h_{K,\theta} z_j + b^h_K, \tag{3}$$

$$v_\theta(z_j) = W^h_{V,\theta} z_j + b^h_V, \tag{4}$$

where h selects the corresponding projection rows, each output has dimension d_h, and absent biases are represented by zero. For a nonempty set J of allowed key positions, we define the attention weights as follows:

$$\alpha_{\theta,j} = \left[\mathrm{softmax}\left(\left(\frac{q_\theta(x)^\top k_\theta(z_k)}{\sqrt{d_h}} + M_k\right)_{k\in J}\right)\right]_j, \tag{5}$$

where M_k is the finite additive attention mask value at allowed key position k; masked positions are excluded from J. The contribution of each head is expressed as follows:

$$f_{\theta,h}(x, \{z_j\}) = O_{\theta,h} \sum_{j\in J} \alpha_{\theta,j} v_\theta(z_j), \tag{6}$$

where d_o is the attention output dimension and $O_{\theta,h} \in \mathbb{R}^{d_o \times d_h}$ contains the columns of the output projection corresponding to head h. The full output before residual addition is $\sum_h f_{\theta,h} + b_O$, with the output bias added once. QK routing determines attention scores and context selection; VO transport maps aggregated values to the output space. Its linear part is $O_{\theta,h} W^h_{V,\theta}$. PAF regularizes both compositions.

3 Methodology

PAF compares attention functions in a fixed coordinate system for each previous domain. The method consists of collecting input summaries, constructing the QK and VO operators in those coordinates, and regularizing their changes during subsequent adaptation. Reference storage combines directly retained QK operators with VO operators reconstructed from the merged weights and update history. Figure 1 summarizes the procedure.

3.1 Input Subspaces from Previous Domains

Speech utterances vary in length and temporal structure. PAF therefore defines its preservation criterion in the feature space of the attention inputs. After training on domain s, we collect input statistics for each attention block using the selected checkpoint and the available data $\mathcal{D}_s$. We estimate query and key/value means, μ^q_s and μ^{kv}_s, and obtain orthonormal bases, U^q_s and U^{kv}_s, from the leading m components of centered PCA. The bases are shared across heads within a block. We suppress the block index ℓ for clarity. The query and key/value inputs on the retained affine subspaces are represented as follows:

$$x = \mu^q_s + U^q_s u, \tag{7}$$

$$z_j = \mu^{kv}_s + U^{kv}_s w_j, \tag{8}$$

where $U^q_s \in \mathbb{R}^{d_q \times m}$ and $U^{kv}_s \in \mathbb{R}^{d_{kv} \times m}$ have orthonormal columns, and $u, w_j \in \mathbb{R}^m$ are subspace coordinates. We use the augmented coordinates $\bar{u} = [u^\top, 1]^\top$ and $\bar{w}_j = [w_j^\top, 1]^\top$ to incorporate input means and projection biases. The stored radii $a_s, b_s \geq 1$ specify the regions satisfying $\|\bar{u}\|_2 \leq a_s$ and $\|\bar{w}_j\|_2 \leq b_s$. We call these bounded regions the stored supports. Their bases, means, and radii remain fixed during subsequent training. Self-attention can share a support when Q, K, and V receive the same inputs. Cross-attention uses separate query and key/value supports. In either case, the retained statistics provide a fixed coordinate system for comparing attention computations across stages. This representation is independent of utterance length and temporal alignment between utterances. Individual utterances and features are discarded after the summaries have been constructed.

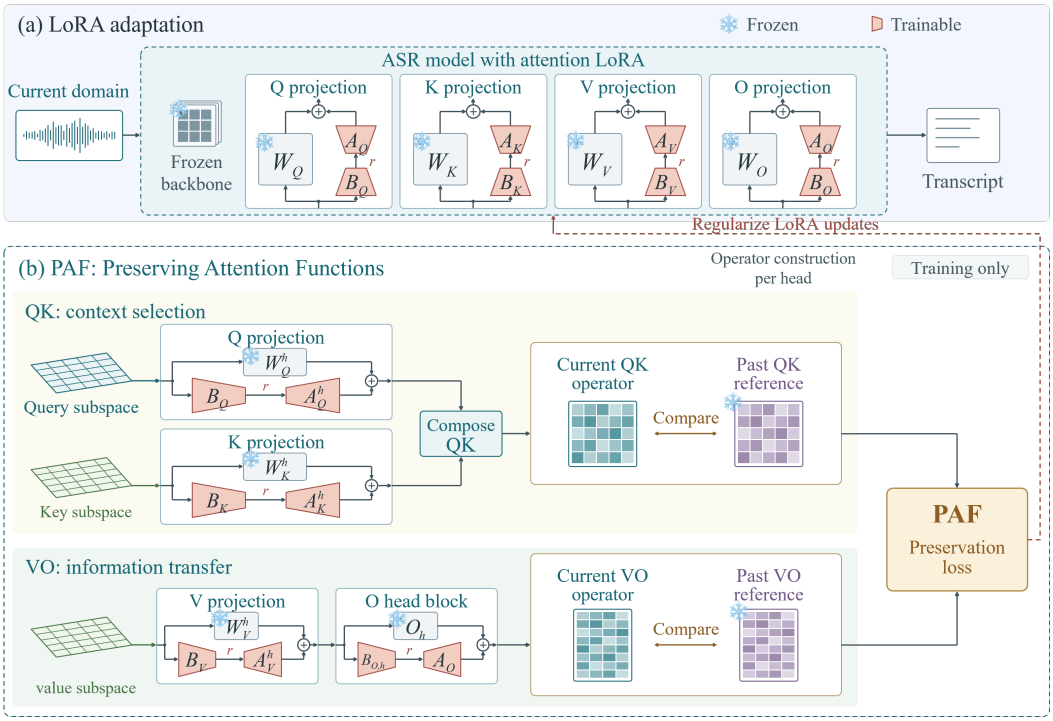

Figure 1: Overview of Preserving Attention Functions (PAF). (a) For the current domain, LoRA factors adapt the Q, K, V, and O attention projections while the pretrained backbone remains frozen; the learned updates are subsequently merged into the ASR model. (b) On the retained query and key/value subspaces, PAF constructs QK routing and VO transport operators for each head from the current effective projections and compares them with the corresponding frozen references from previous domains. The resulting preservation loss regularizes the LoRA updates during training, while inference uses only the final merged model.

3.2 QK and VO Operators on the Retained Subspaces

For each stored support, we express the current projections in its fixed coordinate system as follows:

$$\widetilde{Q}_{\theta,s,h} = \left[W^h_{Q,\theta}U^q_s,\ W^h_{Q,\theta}\mu^q_s + b^h_Q\right], \tag{9}$$

$$\widetilde{K}_{\theta,s,h} = W^h_{K,\theta}U^{kv}_s, \tag{10}$$

$$\widetilde{V}_{\theta,s,h} = \left[W^h_{V,\theta}U^{kv}_s,\ W^h_{V,\theta}\mu^{kv}_s + b^h_V\right], \tag{11}$$

where $\widetilde{Q}_{\theta,s,h}$ and $\widetilde{V}_{\theta,s,h}$ have shape $d_h \times (m+1)$, and $\widetilde{K}_{\theta,s,h}$ has shape $d_h \times m$. We define the QK operator for relative attention scores as follows:

$$P_{\theta,s,h} = \frac{\widetilde{Q}^\top_{\theta,s,h}\widetilde{K}_{\theta,s,h}}{\sqrt{d_h}}, \tag{12}$$

where $P_{\theta,s,h} \in \mathbb{R}^{(m+1)\times m}$. The bilinear form $\bar{u}^\top P_{\theta,s,h} w_j$ determines the relative score of key j for a given query. The query mean and bias are included because they affect differences between key scores. The key mean and bias add the same constant to every score for a fixed query and therefore cancel under softmax. Consequently, P captures the relative scores used for context selection. We compose the value and output projections to define the VO transport operator as follows:

$$T_{\theta,s,h} = O_{\theta,h}\widetilde{V}_{\theta,s,h}, \tag{13}$$

where $T_{\theta,s,h} \in \mathbb{R}^{d_o\times(m+1)}$. This operator includes the value projection of the input mean and the value bias and maps aggregated coordinates to the attention output space. Because the attention

weights sum to one, the final augmented coordinate remains one after aggregation. Using these operators, we express the attention weights on the stored support as follows:

$$\alpha_{\theta,s,h,j} = \left[\mathrm{softmax}\left(\left(\bar{u}^\top P_{\theta,s,h} w_k + M_k\right)_{k\in J}\right)\right]_j, \tag{14}$$

where J and M_k denote the allowed key positions and additive mask values, respectively. The resulting head output is expressed as follows:

$$f_{\theta,s,h} = T_{\theta,s,h} \sum_{j\in J} \alpha_{\theta,s,h,j} \bar{w}_j. \tag{15}$$

The dimensions of P and T are independent of sequence length. For fixed inputs and masks on a retained support, they determine the head output. Projection changes preserving both operators leave the preservation penalty unchanged.

3.3 Regularizing QK and VO Operator Changes

Let $P_{s,h}$ and $T_{s,h}$ denote the reference operators obtained from the model selected after training on domain s. We compute the current operators on the same stored support and define the reference transport coefficient as follows:

$$c_{s,h} = \|T_{s,h}\|_2^2, \tag{16}$$

where $\|\cdot\|_2$ denotes the matrix spectral norm. We define the preservation loss for each head as follows:

$$\ell_{s,h}(\theta) = 2b_s^2 \|T_{\theta,s,h} - T_{s,h}\|_F^2 + 2c_{s,h} a_s^2 b_s^4 \|P_{\theta,s,h} - P_{s,h}\|_F^2, \tag{17}$$

where $\|\cdot\|_F$ is the Frobenius norm. The two terms control VO transport and QK routing changes through the following output bound. The radii and reference coefficients remain fixed during subsequent adaptation.

Proposition 1 (Attention output on a stored support) For a fixed support and head, let $\Delta P = P_{\theta,s,h} - P_{s,h}$ and $\Delta T = T_{\theta,s,h} - T_{s,h}$. Assume that the current and reference heads receive identical inputs on this support, satisfying $\|\bar{u}\|_2 \le a_s$ and $\|\bar{w}_j\|_2 \le b_s$ for every allowed key, and use the same nonempty set of allowed keys and the same additive mask. We bound the difference between their outputs as follows:

$$\|f_{\theta,s,h} - f_{s,h}\|_2 \le b_s \|\Delta T\|_F + \|T_{s,h}\|_2 a_s b_s^2 \|\Delta P\|_F, \tag{18}$$

$$\|f_{\theta,s,h} - f_{s,h}\|_2^2 \le \ell_{s,h}(\theta), \tag{19}$$

where $f_{s,h}$ is the reference head output on these inputs. This bound justifies the joint objective and its coefficients for fixed inputs on the retained support. The proof is provided in Appendix C.1. We average the loss over previous domains and attention heads as follows:

$$\mathcal{L}_{\mathrm{PAF},t}(\theta) = \frac{1}{(t-1)N_H} \sum_{s=1}^{t-1} \sum_{\ell\in\mathcal{B}} \sum_{h=1}^{H_\ell} \ell_{s,\ell,h}(\theta), \tag{20}$$

where $t > 1$, $\mathcal{B}$ is the set of regularized attention blocks, H_ℓ is the number of heads in block ℓ, and $N_H = \sum_{\ell\in\mathcal{B}} H_\ell$. We set $\mathcal{L}_{\mathrm{PAF},1} = 0$ for the first domain. We define the training objective as follows:

$$\mathcal{L}_t(\theta) = \mathcal{L}_{\mathrm{ASR}}(\mathcal{D}_t; \theta) + \lambda \mathcal{L}_{\mathrm{PAF},t}(\theta), \tag{21}$$

where $\lambda \ge 0$ controls preservation. PAF uses effective weights and retained summaries without past utterances or teacher predictions. With adaptation confined to attention projections, Appendix C.2 extends the head bound to complete model logits and transcript loss, accounting for support residuals and propagation through the network. Exact operator agreement and coverage of reference attention inputs within the bounded stored supports, for every evaluated utterance and decoding prefix, preserve transcriptions and WER under the same deterministic decoder.

3.4 Storage and Reconstruction of Reference Operators

PAF retains the input means, bases, and radii, together with each QK reference and transport norm coefficient. To avoid storing a full VO matrix for every head and domain, it records the low-rank V/O updates. Earlier V and O weights are recovered by subtracting later recorded updates from the frozen merged model, excluding the current LoRA update, and are then composed to obtain the transport references. These targets remain fixed within each stage. Appendix C.3 gives the reconstruction equations.

3.5 Sequential Optimization and Inference

At each stage, we freeze the previously merged model and initialize new LoRA factors for Q, K, V, and O. We minimize the objective in Equation (21) using data from the current domain and the retained preservation information. After training, we collect input statistics from the selected checkpoint, construct its reference summary, and record the V/O updates. We then merge the learned updates into the model as follows:

$$W_X^{[t]} = W_X^{[t-1]} + \Delta W_{X,t}, \tag{22}$$

where $X \in \{Q, K, V, O\}$ and $\Delta W_{X,t} = \gamma A_{X,t} B_{X,t}$. The input summaries, reference operators, recorded updates, and merged weights are all obtained from the same checkpoint. Only V/O updates from the second domain onward are needed for subsequent reconstruction. The reference summaries and update history are used only during adaptation. Inference uses the resulting merged model without additional gates, domain identifiers, or separate branches for individual domains. The preservation criterion is independent of the transcription loss and applies to self-attention and cross-attention blocks of the form defined in Section 2.2.

4 Experiments

4.1 Datasets

We evaluated six speech domains using the task sequences in Table 1. TED-LIUM2 (TED2) provides presentation speech (Rousseau et al., 2014), CHiME3 (CH3) provides speech recorded in noisy environments (Barker et al., 2015), and LibriSpeech-other (LS) represents challenging audiobook speech (Panayotov et al., 2015). Common Voice 15.0 English (CV) adds crowdsourced recordings with diverse speakers and recording conditions (Ardila et al., 2020). GigaSpeech-POD and GigaSpeech-YOU provide podcast and YouTube speech, respectively (Chen et al., 2021). We split the available adaptation data into training and development subsets in a 9:1 ratio and used the development subset for validation. All methods and seeds used the same disjoint training, development, and test subsets.

Table 1: Domain orders.

Order	Task sequence
1	CHiME3 → LibriSpeech-other → TED2
2	TED2 → LibriSpeech-other → CHiME3
3	LibriSpeech-other → CHiME3 → TED2
4	TED2 → GigaSpeech-YOU → LibriSpeech-other → CV → GigaSpeech-POD → CHiME3
5	TED2 → CHiME3 → GigaSpeech-YOU

4.2 Evaluation Metrics

We used identical text normalization across methods within each protocol. We define word error rate (WER) at the corpus level as follows:

$$\mathrm{WER} = 100 \frac{S + D + I}{N}, \tag{23}$$

where S, D, and I count substitutions, deletions, and insertions, and N is the number of reference words. Let $E_{i,t}$ denote WER on domain i after task t in a T-task sequence, with domains indexed

by arrival order. We evaluated all domains after every task, including those awaiting adaptation. We define the final average WER and the average WER across the complete evaluation matrix as follows:

$$\text{Last} = \frac{1}{T}\sum_{i=1}^{T} E_{i,T}, \qquad \text{Avg} = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{i=1}^{T} E_{i,t}. \tag{24}$$

For $T > 1$, we measure deterioration from each domain's best WER attained from its arrival through the penultimate stage as follows:

$$\text{Forgetting}_{\text{WER}} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left[E_{i,T} - \min_{i\le u<T} E_{i,u}\right]_+, \tag{25}$$

where $[\cdot]_+$ clips negative values to zero. Lower values of WER, Last, Avg, and forgetting (F) indicate better performance. WER, Last, and Avg are percentages; forgetting is measured in percentage points. These definitions apply to all methods and models.

4.3 Experimental Setup

We used Whisper-small (Radford et al., 2023), S2T-Large trained on LibriSpeech (Wang et al., 2020), and OWSM v3.1 Small (Peng et al., 2024) in ESPnet (Watanabe et al., 2018), retaining their frontends and tokenizers. Experiments ran on a workstation with four NVIDIA RTX 3090 GPUs. Results with standard deviations report means $\pm$ standard deviations across three training seeds with matched initialization, training schedules, and decoding settings. Single values report one training seed. Checkpoint diagnostics, the PAF versus Individual weight comparison, the support dimension sweep, and efficiency profiles follow this single seed protocol. Bootstrap confidence intervals use utterance resampling.

PAF trains fresh Q/K/V/O LoRA branches in all attention blocks with the backbone parameters frozen ($r = 4$, $\alpha_{\text{LoRA}} = 8$, $m = 16$). Training used six epochs per task, one utterance per minibatch, and AdamW with a peak learning rate of 5×10^{-4}, 1,500 linear warmup updates, and inverse square root decay. PAF, its component variants, and Individual use $\lambda = 1.0$ unless varied explicitly. O-LoRA uses $\lambda = 0.5$, including within GainLoRA, and Attention-Spectrum uses $\beta = 0.5$ in Table 2. After each task, supports were constructed from its full training set in evaluation mode with teacher forcing. We retained input means and the leading m centered PCA components, with projection biases included in the affine maps. All positions, including Whisper padding and decoder prefixes, received equal weight. Radii were the maximum norms of PCA coordinates augmented with a constant 1. Individual used the same LoRA configuration and Q/K/V supports, plus a separate support from O inputs, and penalized squared Frobenius deviations from stored individual affine references without radius weighting. We compared PAF against six baselines: sequential LoRA adaptation (IncLoRA), EWC (Kirkpatrick et al., 2017), O-LoRA (Wang et al., 2023), C-LoRA (Smith et al., 2024), GainLoRA (Liang et al., 2025), and Attention-Spectrum adapted to ASR (Zhao et al., 2026). The EWC baseline operates in LoRA mode and applies separate EWC penalties to the factors A and B in Equation (1), with the backbone parameters frozen.

5 Results

5.1 Continual ASR Performance

Table 2 shows consistent improvements in final and average recognition performance with PAF. PAF achieved the lowest mean Last and Avg in all four orders, reducing final average WER relative to the best baseline in each order by 0.83, 0.64, 1.89, and 0.85 percentage points, respectively. PAF also achieved the lowest mean forgetting in Orders 1 and 3. Relative to IncLoRA, forgetting decreased from 12.17 to 5.03 in Order 1, 2.15 to 0.90 in Order 2, 10.65 to 3.67 in Order 3, and 11.02 to 2.79 in Order 4. These results connect the final recognition gains with stronger retention than sequential LoRA adaptation across all four orders. Figure 2 presents recognition performance after each adaptation stage. Each curve averages WER over all evaluation domains, including those awaiting adaptation, and Avg in Table 2 summarizes these values across stages. PAF's lowest mean

Avg in every order indicates improved recognition over the complete adaptation sequence. In the six domain sequence, PAF achieved a Last of 17.48% and an Avg of 18.46%, compared with the best baseline values of 18.33% for C-LoRA and 19.01% for GainLoRA, respectively. The separate forgetting metric quantifies deterioration on previously learned domains. Together, these measures show that PAF combines improved recognition across adaptation stages with stronger retention than IncLoRA, including in the longer domain sequence.

Table 2: Continual ASR results on Whisper-small for three domains (Orders 1, 2, and 3) and six domains (Order 4). Bold marks the lowest mean for each metric within an order.

Order	Method	Last (↓)	Avg (↓)	F (↓)
1	IncLoRA	22.13 ± 0.48	19.52 ± 0.31	12.17 ± 0.36
	EWC	17.96 ± 0.43	18.18 ± 0.21	6.39 ± 0.32
	O-LoRA	19.99 ± 0.23	19.28 ± 0.17	9.07 ± 0.26
	C-LoRA	23.29 ± 0.45	19.18 ± 0.26	7.85 ± 0.41
	GainLoRA	18.78 ± 0.25	18.90 ± 0.21	7.45 ± 0.29
	Attention-Spectrum	22.48 ± 0.35	20.46 ± 0.22	12.78 ± 0.40
	PAF (ours)	$\mathbf{17.13 \pm 0.31}$	$\mathbf{17.58 \pm 0.24}$	$\mathbf{5.03 \pm 0.28}$
2	IncLoRA	15.23 ± 0.31	16.24 ± 0.25	2.15 ± 0.22
	EWC	15.21 ± 0.28	16.05 ± 0.23	1.85 ± 0.21
	O-LoRA	15.43 ± 0.22	16.23 ± 0.15	2.20 ± 0.19
	C-LoRA	15.47 ± 0.18	15.67 ± 0.14	$\mathbf{0.55 \pm 0.12}$
	GainLoRA	16.20 ± 0.24	16.08 ± 0.09	2.25 ± 0.15
	Attention-Spectrum	15.27 ± 0.23	16.18 ± 0.17	2.10 ± 0.22
	PAF (ours)	$\mathbf{14.57 \pm 0.25}$	$\mathbf{15.48 \pm 0.18}$	0.90 ± 0.20
3	IncLoRA	21.67 ± 0.42	19.44 ± 0.22	10.65 ± 0.21
	EWC	19.18 ± 0.41	18.60 ± 0.26	8.20 ± 0.30
	O-LoRA	18.37 ± 0.30	17.89 ± 0.20	6.35 ± 0.35
	C-LoRA	21.13 ± 0.33	19.46 ± 0.25	4.03 ± 0.23
	GainLoRA	19.63 ± 0.33	18.53 ± 0.33	7.80 ± 0.45
	Attention-Spectrum	21.22 ± 0.35	19.57 ± 0.23	9.37 ± 0.34
	PAF (ours)	$\mathbf{16.48 \pm 0.45}$	$\mathbf{17.25 \pm 0.33}$	$\mathbf{3.67 \pm 0.44}$
4	IncLoRA	25.80 ± 0.34	21.72 ± 0.26	11.02 ± 0.30
	EWC	24.97 ± 0.32	21.58 ± 0.30	11.21 ± 0.38
	O-LoRA	18.79 ± 0.26	19.49 ± 0.18	4.71 ± 0.11
	C-LoRA	18.33 ± 0.22	20.03 ± 0.15	$\mathbf{1.53 \pm 0.10}$
	GainLoRA	20.40 ± 0.28	19.01 ± 0.20	6.74 ± 0.23
	Attention-Spectrum	24.81 ± 0.41	21.77 ± 0.35	9.38 ± 0.40
	PAF (ours)	$\mathbf{17.48 \pm 0.21}$	$\mathbf{18.46 \pm 0.24}$	2.79 ± 0.25

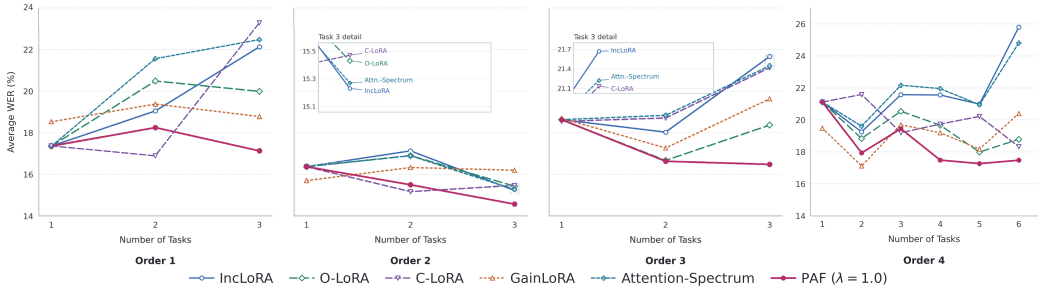

Figure 2: Average WER across all evaluation domains after each task: three domains in Orders 1–3 and six domains in Order 4.

5.2 Results on Additional ASR Models

Table 3 shows that PAF achieved the lowest mean Last and forgetting on both additional ASR models, extending the gains observed with Whisper-small. On OWSM in Order 2, PAF also achieved the lowest mean Avg, improving recognition at the final checkpoint and on average across adaptation stages. Its Last and Avg were both 13.04%, compared with 14.15% and 13.84% for O-LoRA,

respectively. On S2T-Large in Order 5, PAF reduced Last from GainLoRA's 33.16% to 31.69% and forgetting from 3.73 to 1.12 percentage points. These improvements connect better final recognition with stronger retention of earlier domains. PAF recorded an Avg of 35.75%, and GainLoRA recorded the lowest mean Avg of 35.47%. The results support PAF's gains in final recognition and retention across the three evaluated ASR models.

Table 3: Continual ASR results on additional models and their respective domain orders.

	OWSM (Order 2)			**S2T-Large (Order 5)**		
Method	Last (↓)	Avg (↓)	F (↓)	Last (↓)	Avg (↓)	F (↓)
IncLoRA	14.47 ± 0.23	14.06 ± 0.18	2.46 ± 0.22	35.89 ± 0.33	37.55 ± 0.24	7.56 ± 0.29
O-LoRA	14.15 ± 0.18	13.84 ± 0.18	2.25 ± 0.15	33.93 ± 0.25	36.93 ± 0.11	4.79 ± 0.07
GainLoRA	14.46 ± 0.20	14.11 ± 0.08	2.32 ± 0.08	33.16 ± 0.24	$\mathbf{35.47 \pm 0.10}$	3.73 ± 0.09
PAF (ours)	$\mathbf{13.04 \pm 0.12}$	$\mathbf{13.04 \pm 0.09}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{31.69 \pm 0.18}$	35.75 ± 0.16	$\mathbf{1.12 \pm 0.14}$

Table 4: Component comparisons on Whisper-small in Orders 2 and 3.

	Order 2			**Order 3**		
Variant	Last (↓)	Avg (↓)	F (↓)	Last (↓)	Avg (↓)	F (↓)
IncLoRA	15.23 ± 0.31	16.24 ± 0.25	2.15 ± 0.22	21.67 ± 0.42	19.44 ± 0.22	10.65 ± 0.21
P only	15.93 ± 0.24	16.37 ± 0.18	2.10 ± 0.20	20.62 ± 0.27	18.83 ± 0.22	9.50 ± 0.24
T only	14.77 ± 0.21	15.87 ± 0.17	1.55 ± 0.17	18.50 ± 0.23	18.33 ± 0.20	6.32 ± 0.21
Individual	15.00 ± 0.22	16.06 ± 0.19	1.75 ± 0.18	18.50 ± 0.24	18.03 ± 0.18	5.99 ± 0.20
PAF (ours)	$\mathbf{14.57 \pm 0.25}$	$\mathbf{15.48 \pm 0.18}$	$\mathbf{0.90 \pm 0.20}$	$\mathbf{16.48 \pm 0.45}$	$\mathbf{17.25 \pm 0.33}$	$\mathbf{3.67 \pm 0.44}$

5.3 Component Analysis and Efficiency

Table 4 shows that jointly preserving the score and transport operators yielded the lowest mean Last, Avg, and forgetting in both orders. Preserving T alone improved all three metrics relative to preserving P alone, and joint preservation further improved all three metrics, supporting the complementary roles of the two operators. In Order 3, joint preservation reduced Last from 18.50% to 16.48% and forgetting from 6.32 to 3.67 percentage points compared with T only. PAF also improved all three mean metrics over Individual, which preserves each projection separately, in both orders. These comparisons support the composite preservation objective under the evaluated settings; Appendix A.4 examines sensitivity to the preservation weight. In a diagnostic SGD step, removing the Q/K/V update component along stored LS and CH3 input directions reduced the TED2 NLL improvement by 45.13% after matching the total update norm (Appendix A.3). A separate intervention identified a learned O component that supported TED2 prediction and preserved the stored composite operators (Appendix A.2). These results support useful adaptation through shared inputs and coordinated projection changes. In Order 3, PAF achieved an overall decoding real time factor (RTF) of 0.389 (Appendix B.4). Merged LoRA updates retain the original inference computation within a single ASR model.

6 Conclusion

We introduced PAF for replay-free continual ASR, jointly preserving QK context selection and VO information transport on compact summaries of earlier attention inputs. Across the evaluated domain orders and ASR models, PAF improved final recognition and retention relative to sequential LoRA adaptation while using a single merged model for inference. Component comparisons supported joint preservation. A controlled intervention identified a learned component that supported new domain prediction while leaving the composite operators on stored supports unchanged. These results support composite attention preservation as a practical objective for continual speech recognition.

AI Assistance Disclosure

ChatGPT (OpenAI) was used only for grammar correction and clarity refinement. The authors reviewed and are fully responsible for all content.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, 2020. URL https://aclanthology.org/2020.lrec-1.520/.

Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 504–511, 2015. URL https://www.merl.com/publications/TR2015-136.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4960–4964, 2016. URL https://research.google/pubs/listen-attend-and-spell-a-neural-network-for-large-vocabulary-conversational-speech-recognition/.

William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc V. Le, and Mohammad Norouzi. SpeechStew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*, 2021. URL https://arxiv.org/abs/2104.02133.

Heng-Jui Chang, Hung-yi Lee, and Lin-shan Lee. Towards lifelong learning of end-to-end ASR. In *Proceedings of Interspeech*, 2021. URL https://arxiv.org/abs/2104.01616.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proceedings of Interspeech*, pp. 3670–3674, 2021. doi: 10.21437/Interspeech.2021-1965. URL https://www.isca-archive.org/interspeech_2021/chen21o_interspeech.html.

Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, volume 28, pp. 577–585, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/1068c6e4c8051cfd4e9ea8072e3189e2-Abstract.html.

Jiaheng Dong, Hong Jia, Soumyajit Chatterjee, Abhirup Ghosh, James Bailey, and Ting Dang. E-BATS: Efficient backpropagation-free test-time adaptation for speech foundation models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/html/2506.07078v3.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015. URL https://www.isca-archive.org/interspeech_2020/gulati20_interspeech.html.

Jinghan He, Haiyun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, and Jinqiao Wang. SEEKR: Selective attention-guided knowledge retention for continual learning of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3254–3266, 2024. doi: 10.18653/v1/2024.emnlp-main.190. URL https://aclanthology.org/2024.emnlp-main.190/.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Proceedings of Interspeech*, pp. 721–725, 2021. doi: 10.21437/Interspeech.2021-236. URL https://www.isca-archive.org/interspeech_2021/hsu21_interspeech.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. A comparative study on transformer vs. RNN in speech applications. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2019. doi: 10.1109/ASRU46091.2019.9003750. URL https://arxiv.org/abs/1909.06317.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://arxiv.org/abs/1612.00796.

Jae-Hong Lee and Joon-Hyuk Chang. Continual momentum filtering on parameter space for online test-time adaptation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BllUWdpIOA.

Jae-Hong Lee, Chae-Won Lee, Jin-Seong Choi, Joon-Hyuk Chang, Woo Kyeong Seong, and Jeonghan Lee. CTRL: Continual representation learning to transfer information of pre-trained for WAV2VEC 2.0. In *Proceedings of Interspeech*, pp. 3398–3402, 2022. doi: 10.21437/Interspeech.2022-10063. URL https://www.isca-archive.org/interspeech_2022/lee22i_interspeech.html.

Xinlong Li, Weijieying Ren, Wei Qin, Lei Wang, Tianxiang Zhao, and Richang Hong. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889361. URL https://doi.org/10.1109/ICASSP49660.2025.10889361.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liang_InfLoRA_Interference-Free_Low-Rank_Adaptation_for_Continual_Learning_CVPR_2024_paper.html.

Yan-Shuo Liang, Jia-Rui Chen, and Wu-Jun Li. Gated integration of low-rank adaptation for continual learning of large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/html/2505.15424v2.

Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. Listen, adapt, better WER: Source-free single-utterance test-time adaptation for automatic speech recognition. In *Proceedings of Interspeech*, pp. 2198–2202, 2022. doi: 10.21437/Interspeech.2022-600. URL https://www.isca-archive.org/interspeech_2022/lin22b_interspeech.html.

Zhong Meng, Zhuo Chen, Vadim Mazalov, Jinyu Li, and Yifan Gong. Unsupervised adaptation with domain separation networks for robust speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2017. URL https://arxiv.org/abs/1711.08010.

Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang Juang. Adversarial teacher-student learning for unsupervised domain adaptation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018. URL https://www.microsoft.com/en-us/research/publication/adversarial-teacher-student-learning-unsupervised-domain-adaptation/.

Sarthak Mittal, Sharath Chandra Raparthy, Irina Rish, Yoshua Bengio, and Guillaume Lajoie. Compositional attention: Disentangling search and retrieval. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=IwJPj2MBcIa.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://www.danielpovey.com/files/2015_icassp_librispeech.pdf.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee-weon Jung, and Shinji Watanabe. OWSM v3.1: Better and faster open whisper-style speech models based on E-Branchformer. In *Proceedings of Interspeech*, pp. 352–356, 2024. doi: 10.21437/Interspeech.2024-1194. URL https://www.isca-archive.org/interspeech_2024/peng24b_interspeech.html.

Peter Plantinga, Jaekwon Yoo, and Chandra Dhir. Continual learning for end-to-end ASR by averaging domain experts. *arXiv preprint arXiv:2305.09681*, 2023. URL https://arxiv.org/abs/2305.09681.

Fuli Qiao and Mehrdad Mahdavi. Merge before forget: A single LoRA continual learning via continual merging. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=i1Rj7yU6eF.

Chengwei Qin and Shafiq Joty. LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of T5. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.07298.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2301.12314.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 3935–3939, 2014. URL https://aclanthology.org/L14-1079/.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3AOj0RCNC2.

Kyuhong Shim, Jungwook Choi, and Wonyong Sung. Understanding the role of self attention for efficient speech recognition. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AvcfxqRy4Y.

Khe Chai Sim, Arun Narayanan, Ananya Misra, Anshuman Tripathi, Golan Pundak, Tara N. Sainath, Parisa Haghani, Bo Li, and Michiel Bacchiani. Domain adaptation using factorized hidden layer for robust automatic speech recognition. In *Proceedings of Interspeech*, pp. 892–896, 2018. doi: 10.21437/Interspeech.2018-2246. URL https://www.isca-archive.org/interspeech_2018/sim18_interspeech.html.

James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with C-LoRA. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.06027.

Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, and Alex Waibel. Self-attentional acoustic models. In *Proceedings of Interspeech*, pp. 3723–3727, 2018. doi: 10.21437/Interspeech.2018-1910. URL https://www.isca-archive.org/interspeech_2018/sperber18_interspeech.html.

Yuki Takashima, Shota Horiguchi, Shinji Watanabe, Leibny Paola Garcia Perera, and Yohei Kawaguchi. Updating only encoders prevents catastrophic forgetting of end-to-end ASR models. In *Proceedings of Interspeech*, pp. 2218–2222, 2022. doi: 10.21437/Interspeech.2022-11282. URL https://www.isca-archive.org/interspeech_2022/takashima22_interspeech.html.

Viet-Hoang Tran, Thieu Vo, An Nguyen, Tho-Huu Tran, Minh-Khoi Nguyen-Nhat, Thanh Tran, Duy-Tung Pham, and Tan Nguyen. Equivariant neural functional networks for transformers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/29962c2c9daf1fbd92530a7c958dfc2b-Abstract-Conference.html.

Steven Vander Eeckt and Hugo Van hamme. Rehearsal-free online continual learning for automatic speech recognition. In *Proceedings of Interspeech*, pp. 944–948, 2023. doi: 10.21437/Interspeech.2023-788. URL https://www.isca-archive.org/interspeech_2023/vandereeckt23_interspeech.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. Fairseq S2T: Fast speech-to-text modeling with Fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 33–39, 2020. doi: 10.18653/v1/2020.aacl-demo.6. URL https://aclanthology.org/2020.aacl-demo.6/.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023. doi: 10.18653/v1/2023.findings-emnlp.715. URL https://aclanthology.org/2023.findings-emnlp.715/.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pp. 2207–2211, 2018. doi: 10.21437/Interspeech.2018-1456. URL https://www.isca-archive.org/interspeech_2018/watanabe18_interspeech.html.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017. URL https://proceedings.mlr.press/v70/zenke17a.html.

Chuangxin Zhao, Canran Xiao, Siyuan Ma, Mengyao Lyu, Yanbiao Ma, Jun Xia, Guiguang Ding, and Yang Liu. Attention-spectrum regularization for replay-free continual multimodal LLMs. *arXiv preprint arXiv:2606.23063*, 2026. doi: 10.48550/arXiv.2606.23063. URL https://arxiv.org/abs/2606.23063.

Weixiang Zhao, Shilong Wang, Yulin Hu, Yanyan Zhao, Bing Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. SAPT: A shared attention framework for parameter-efficient continual learning of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11641–11661, 2024. doi: 10.18653/v1/2024.acl-long.625. URL https://aclanthology.org/2024.acl-long.625/.

A Attention Preservation and Adaptation

Unless specified otherwise, Whisper-small diagnostics use one training seed of Order 3 (LibriSpeech-other → CHiME3 → TED2), with $r = 4$, $m = 16$, $\lambda = 1$, and six epochs per task. Inputs and gradients serve diagnostic analysis; LS validation denotes the LibriSpeech-other validation split.

A.1 Joint Score and Transport Preservation

Figure 3 shows that joint preservation controls both attention operators and gives the lowest LS WER among the three variants. Starting from the same model after LS adaptation, we compared P only, T only, and PAF on the stored LS PCA16 support using the relative operator change defined as follows:

$$\delta_{C,\ell,h}(U) = 100 \frac{\|C^{(U)}_{\text{final},\ell,h} - C^{(U)}_{\text{LS},\ell,h}\|_F}{\|C^{(U)}_{\text{LS},\ell,h}\|_F}, \qquad C \in \{P, T\}, \tag{26}$$

where ℓ and h index blocks and heads. Evaluation uses fixed input means, basis U, and nonzero reference norms.

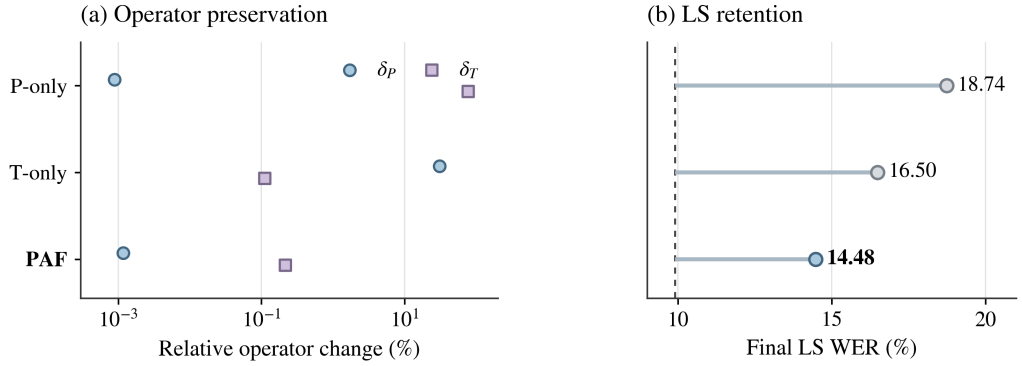

Figure 3: Complementarity of score and transport preservation. (a) Relative operator changes on the stored LS PCA16 support, on a logarithmic scale. (b) Final LS WER. The dashed line marks the common WER of 9.9154% after LS adaptation.

With P only, transport changed by 76.9823%; with T only, routing changed by 30.6158%. PAF reduced these changes to 0.2168% and 0.001182%, respectively, combining stable context selection and information transport with stronger LS retention.

A.2 Useful Projection Changes under Composite Preservation

Table 5: Projection changes, composite operator drift, and LS recognition in Order 3. Q/K/V changes use reduced coefficients on the fixed LS support; O changes use the output weights for each head. Changes are mean relative Frobenius differences across 432 heads, expressed in percent and normalized by their respective references.

	Projection change				Operator drift		
Method	Q	K	V	O	QK	VO	LS WER
Individual	0.309069	1.613118	2.124470	34.295602	1.587720	36.8698	16.44
PAF (ours)	0.007481	0.014978	0.210778	22.226498	**0.001182**	**0.21677**	**14.48**

Table 5 shows that PAF combines a mean O weight change of 22.23% with a VO drift of 0.21677% on the retained LS support. We compared changes from the common LS checkpoint to the final checkpoint using FP64 arithmetic across 36 blocks and 432 heads. Individual preserves affine Q/K/V/O maps; PAF preserves their QK and VO compositions. We tested whether an O component preserving the stored transport operators contributes to TED2 prediction. For each head, the

O change acquired from CH3 to TED2 and the concatenated reduced value maps are defined as follows:

$$\begin{aligned} \Delta O_h &= O_h^{\text{final}} - O_h^{\text{CH3}}, \\ Z_h &= \left[\widetilde{V}_{\text{final,LS},h},\ \widetilde{V}_{\text{final,CH3},h}\right]. \end{aligned} \tag{27}$$

Here, the reduced value maps use final V weights, stored means and PCA16 bases, and affine biases. With Π_{Z_h} the orthogonal projector onto $\mathrm{col}(Z_h)$, the removable component and intervened projection are defined as follows:

$$N_h = \Delta O_h(I - \Pi_{Z_h}), \qquad O_h^{(\alpha)} = O_h^{\text{final}} - \alpha N_h, \tag{28}$$

where α is the removal strength and I is the identity on the head output projection's input space. We evaluated $\alpha \in \{0, 0.5, 1\}$ across all 432 heads with all other parameters fixed. Projectors used FP64 singular value decomposition with a relative threshold of 10^{-12}. An exact projector satisfies $N_h Z_h = 0$, and the resulting transport invariance is expressed as follows:

$$\begin{aligned} T_{s,h}^{(\alpha)} &= \left(O_h^{\text{final}} - \alpha N_h\right)\widetilde{V}_{\text{final},s,h} \\ &= T_{s,h}^{(0)}, \qquad s \in \{\text{LS}, \text{CH3}\}. \end{aligned} \tag{29}$$

The unchanged QK and VO operators make the PAF loss for LS and CH3 constant along this intervention in exact arithmetic. To evaluate Individual's response on the PAF checkpoint, we express its affine O map as follows:

$$\mathcal{A}_{O,s}(\theta) = \left[O_\theta U_s^O,\ O_\theta \mu_s^O + b_O\right], \tag{30}$$

where U_s^O and μ_s^O are the stored O input basis and mean, and b_O is the fixed output bias. Its squared Frobenius deviation from the LS reference defines the block penalty, whose value and second derivative are expressed as follows:

$$e_\ell(\alpha) = \|E_\ell - \alpha G_\ell\|_F^2, \qquad e_\ell''(\alpha) = 2\|G_\ell\|_F^2, \tag{31}$$

where E_ℓ is the initial LS affine deviation, N_ℓ concatenates the head interventions, and $G_\ell = N_\ell[U_{\ell,\text{LS}}^O, \mu_{\ell,\text{LS}}^O]$. The reference, basis, mean, and bias are fixed. For $G_\ell \neq 0$, positive rescaling retains the positive curvature of Individual's penalty and the constant PAF loss. NLL used 64 fixed TED2 validation utterances with teacher forcing. Random controls matched the learned component's norm within the same null space; confidence intervals used 5,000 utterance bootstrap resamples.

Table 6: Effect of removing the learned O component in Order 3. The LS O penalty is the Individual criterion evaluated on the PAF checkpoint. Block counts compare each removal condition with the original checkpoint.

Removal (%)	LS O penalty	Blocks decreased	TED2 NLL
0	0.02615	—	0.329
50	0.02223	**36/36**	0.389
100	0.02066	**36/36**	0.571

Table 6 shows that removing this component at $\alpha = 0.5$ and $\alpha = 1$ increased TED2 NLL and reduced Individual's LS O penalty in all 36 blocks. Every fixed positive weighted sum of these block penalties therefore decreased. Full removal changed the Individual affine O map by 4.93% and lowered its penalty by 20.99%, with an LS VO change of approximately $2 \times 10^{-6}\%$. The NLL increase at full removal had a 95% confidence interval of $[0.169, 0.316]$. Full removal produced an NLL 0.239 higher than the mean norm matched random control (95% confidence interval: $[0.167, 0.311]$). These results identify a learned component that supports TED2 prediction while preserving the composite operators on the stored supports.

Compensating Q/K changes in Order 2. Table 7 shows compensation between learned Q and K changes in PAF. We analyzed Whisper-small checkpoints from Order 2 (TED2 $\rightarrow$ LS $\rightarrow$ CH3), using seed 2022 and the fixed $m = 16$ input supports saved with the Task 2 checkpoint. The analysis compared the checkpoints after LS and CH3 adaptation across 36 attention blocks and 432 heads. For each head, $\widetilde{Q}^{[i]}$ and $\widetilde{K}^{[i]}$ are the reduced coefficients from Equations (9) and (10), evaluated at

checkpoint $i \in \{2, 3\}$ in the same stored coordinates. We decompose their contributions to the score operator change symmetrically as follows:

$$C_Q = \frac{(\widetilde{Q}^{[3]} - \widetilde{Q}^{[2]})^\top (\widetilde{K}^{[3]} + \widetilde{K}^{[2]})}{2\sqrt{d_h}},$$
$$C_K = \frac{(\widetilde{Q}^{[3]} + \widetilde{Q}^{[2]})^\top (\widetilde{K}^{[3]} - \widetilde{K}^{[2]})}{2\sqrt{d_h}}. \tag{32}$$

These terms satisfy $\Delta P = P^{[3]} - P^{[2]} = C_Q + C_K$. Their interaction determines the combined change energy as follows:

$$\|\Delta P\|_F^2 = \|C_Q\|_F^2 + \|C_K\|_F^2 + 2\langle C_Q, C_K\rangle_F, \tag{33}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. A negative cross term represents compensation between the two contributions. We summarize their aggregate effect using the energy ratio as follows:

$$R(\omega) = \frac{\sum_j \omega_j \|C_{Q,j} + C_{K,j}\|_F^2}{\sum_j \omega_j \big(\|C_{Q,j}\|_F^2 + \|C_{K,j}\|_F^2\big)}, \tag{34}$$

where j indexes heads across the evaluated blocks. The unweighted ratio uses $\omega_j = 1$, and the ratio with PAF weighting uses the fixed score coefficients in Equation (17).

Table 7: Q/K compensation between the Task 2 and Task 3 checkpoints in Order 2. Energy ratios compare the combined score change energy with the sum of the separate Q and K contribution energies.

Statistic	Value
Heads with $\langle C_Q, C_K\rangle_F < 0$	$345/432$ (79.86%)
Unweighted energy ratio	0.5933
Energy ratio with PAF weighting	0.7959

Negative cross terms occurred in 79.86% of the heads. With PAF weighting, the combined Q/K change energy was 20.41% smaller than the sum of the separate contribution energies. These results identify compensating Q/K changes in trained checkpoints and provide empirical support for preserving their composition while allowing changes in the constituent projections.

A.3 Adaptation in Shared Input Directions

Figure 4 shows that speech domains share input structure and that past attention supports capture current task gradient directions. We measured overlap in temporally averaged frozen conv2 representations, the inputs to GainLoRA's gates. From 1,015 validation utterances, we fitted centered PCA16 bases to 156 utterances per domain and evaluated the remaining utterances. For orthonormal bases U_i and U_j, overlap is defined as follows:

$$\mathrm{Overlap}(U_i, U_j) = \frac{1}{m}\|U_i^\top U_j\|_F^2, \tag{35}$$

where $m = 16$; zero and one denote orthogonal and identical subspaces, respectively.

The overlaps were 0.6641 for LS and CH3, 0.6979 for LS and TED2, and 0.6419 for CH3 and TED2. Each conv2 basis explained approximately 72% to 81% of the variation in evaluation utterances from another domain. The LS attention support captured 63.2% of CH3 Q/K/V gradient energy; LS+CH3 supports captured 62.7% of TED2 gradient energy. To test the contribution of shared directions, we computed one SGD step from the ASR loss alone at the PAF checkpoint after CH3. The step used 32 TED2 utterances, with NLL measured on 64 separate utterances from an initial value of 1.325216. Reconstructed updates provided the corresponding operator changes. For each Q/K/V projection, we construct a joint basis from the retained bases and means as follows:

$$R = \mathrm{orth}[U_{\mathrm{LS}}, \mu_{\mathrm{LS}}, U_{\mathrm{CH3}}, \mu_{\mathrm{CH3}}]. \tag{36}$$

Query and key/value spaces were constructed separately, yielding rank 34 for all 108 Q/K/V projections. With $\Pi = RR^\top$, each weight update decomposes as follows:

$$\Delta W_{\mathrm{past}} = \Delta W \Pi, \qquad \Delta W_{\mathrm{remain}} = \Delta W (I - \Pi), \tag{37}$$

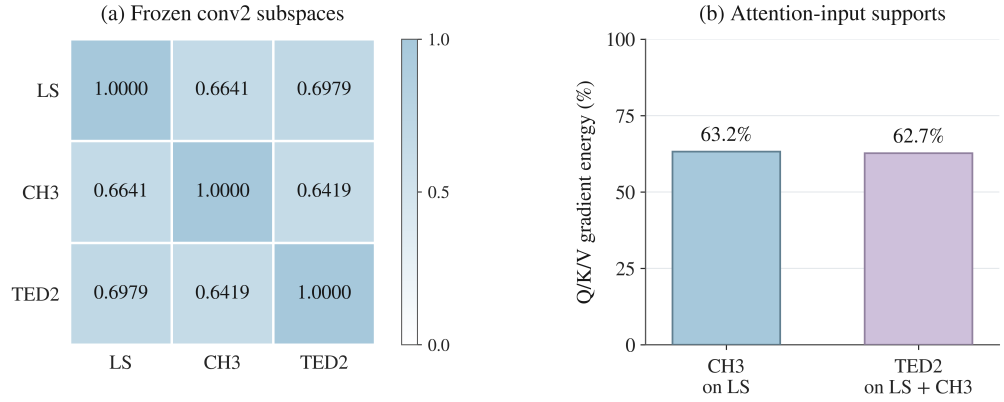

Figure 4: Shared input structure across speech domains. (a) Overlap between PCA16 subspaces of frozen conv2 representations. (b) Q/K/V gradient energy for the current task captured by past attention input supports.

where $\Delta W_{\text{remain}} U_s = 0$ and $\Delta W_{\text{remain}} \mu_s = 0$ preserve past affine Q/K/V maps in exact arithmetic with fixed biases. Table 8 compares the original update, removal of the past component, removal followed by matching the original total update norm, and the past Q/K/V component alone. Both removal conditions retained the O update. Operator changes use stored LS and CH3 PCA16 supports and the common checkpoint after CH3 as their reference.

Table 8: New domain improvement and changes in stored attention operators under the same diagnostic updates. NLL decrease was measured on TED2. The removed space spans past PCA bases and input means. Operator changes are relative Frobenius differences in percent from the common checkpoint after CH3, aggregated over 432 heads for each stored support. Here ≈ 0 denotes a value below $5 \times 10^{-9}\%$. The diagnostic SGD step used the ASR loss alone.

	TED2	LS support		CH3 support	
Update	NLL decrease (↑)	ΔP	ΔT	ΔP	ΔT
Original	**0.009832**	0.005346	0.025375	0.005334	0.025054
Remove past Q/K/V	0.005307	≈ 0	0.007588	≈ 0	0.007733
Remove and match norm	0.005395	≈ 0	0.007713	≈ 0	0.007861
Past Q/K/V only	0.004516	0.005346	0.024179	0.005334	0.023844

The original step improved TED2 NLL with all four operator changes below 0.026%. Removing the past Q/K/V component and matching the total update norm reduced the NLL gain by 45.13% (paired utterance bootstrap 95% confidence interval: $[42.95\%, 48.26\%]$). This component contained 3.2% of the update energy and improved NLL when applied alone, identifying useful adaptation through shared input directions. The retained O update contributes to VO changes in the removal conditions.

Effect of GainLoRA gate constraints. Table 9 shows that releasing the gate constraint improved final CH3 recognition in a GainLoRA control. This control used seed 2022 in Order 2 (TED2 → LS → CH3). Both conditions adapted to CH3 for two epochs from the same checkpoint after LS, with the remaining training settings fixed. The constrained condition projected both gate gradients and the actual AdamW updates onto the complement of the retained gate input subspace. The released condition removed these projections. Evaluation used the existing checkpoints and their saved recognition outputs for the full test sets. We extracted the same fixed 200 utterance IDs per domain from both conditions and rescored the corresponding hypotheses. The reported values are corpus WERs on these fixed subsets, and domain averages are arithmetic means of the corresponding WERs.

Releasing the constraint reduced CH3 WER by 3.2895 percentage points and the mean WER across all three domains by 1.0613 points; the mean WER on the previous two domains increased by 0.0528

Table 9: GainLoRA gate constraint control in Order 2, evaluated by rescoring a fixed subset of 200 test utterances per domain. Changes are released minus constrained. Bold marks the lower WER in each row.

Evaluation	Constrained	Released	Change (pp)
CH3 (new domain)	26.8840	**23.5945**	−3.2895
Previous two domains (mean)	**10.5384**	10.5912	+0.0528
All three domains (mean)	15.9869	**14.9256**	−1.0613

points. The fraction of gate displacement energy within the retained input subspace increased from approximately zero to 64.97%, confirming that the released condition used directions protected by the constraint. Together with the shared direction diagnostics above, this controlled comparison provides direct evidence that restricting gate updates in past input directions can limit adaptation to a new speech domain.

A.4 Retention and Adaptation across Preservation Weights

Table 10 shows that every evaluated PAF setting achieved lower Last, Avg, and forgetting than every Individual setting. We compared $\lambda \in \{0.1, 0.5, 1.0\}$ in one seed of Order 3 with the same LoRA configuration and Q/K/V supports. Individual used its separate O input support (Section 4.3); each objective retained its own scale and weights. PAF used compact reference storage at $\lambda = 1.0$ and dense storage at $\lambda = 0.5$ and 0.1.

Table 10: Comparison of composite and individual preservation across weights in Order 3. Bold marks the lowest observed value in each metric.

Method	λ	Last (↓)	Avg (↓)	F (↓)
Individual	1.0	18.50	18.03	5.99
Individual	0.5	19.31	18.29	7.24
Individual	0.1	18.38	17.97	6.14
PAF (ours)	1.0	**16.63**	**17.38**	**3.81**
PAF (ours)	0.5	17.40	17.82	4.84
PAF (ours)	0.1	16.94	17.55	4.39

Table 11 evaluates average recognition on previous domains and recognition on TED2 at the same final checkpoint, with the metrics defined as follows:

$$\mathrm{OldFinal} = \frac{1}{T-1}\sum_{i=1}^{T-1} E_{i,T}, \qquad \mathrm{NewFinal} = E_{T,T}, \tag{38}$$

where $E_{i,t}$ follows Section 4.2. At both $\lambda = 0.5$ and 0.1, PAF improved OldFinal and NewFinal over Individual, combining stronger LS and CH3 retention with improved TED2 recognition.

Table 11: Final recognition on previous domains and the final new domain in Order 3. OldFinal averages LS and CHiME3; NewFinal reports TED2. Bold marks the lower value at each weight.

	OldFinal (↓)		NewFinal (↓)	
λ	Individual	PAF	Individual	PAF
0.5	24.04	**21.44**	9.83	**9.34**
0.1	22.79	**20.84**	9.56	**9.16**

B Retained Supports, Parameter Sensitivity, and Cost

B.1 PCA versus Random Supports

Figure 5 shows that PCA16 captured more input variance and recognition gradient energy than random supports of equal dimension. We evaluated 108 Q/K/V projections after LS adaptation using NLL gradients from 32 utterances and their reference transcripts, with measurements averaged over three random bases within this fixed training run.

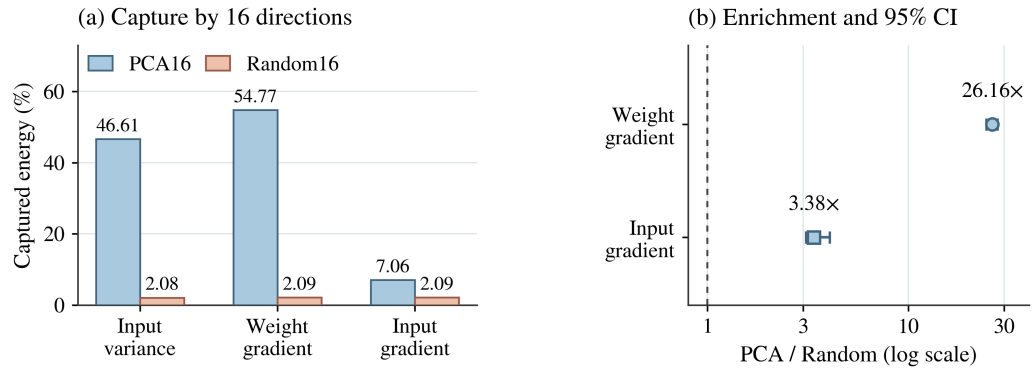

Figure 5: Recognition relevance of PCA16 supports. (a) Captured input variance, weight gradient energy, and input gradient energy for PCA16 and Random16. (b) PCA to random gradient energy ratios with utterance bootstrap 95% confidence intervals.

PCA16 captured 46.61% of input variance, 54.77% of weight gradient energy, and 7.06% of input gradient energy, compared with approximately 2.09% for Random16 in each measurement. The PCA to random ratios for weight and input gradient energy were 26.16 and 3.38, with 95% confidence intervals of $[24.48, 27.80]$ and $[3.09, 4.07]$. These gradient diagnostics connect supports constructed from input statistics with parameter update directions and recognition sensitivity.

B.2 Support Dimension and Input Coverage

Figure 6 shows that increasing m from 4 to 64 reduced QK and VO changes on a common LS PCA64 evaluation support. The single seed Order 3 sweep used $m \in \{4, 8, 16, 32, 64\}$ with fixed $r = 4$, $\lambda = 1$, and training schedule; each dimension used its corresponding radii and transport coefficients. At the default $m = 16$, QK and VO changes on PCA64 were 1.58% and 18.42%.

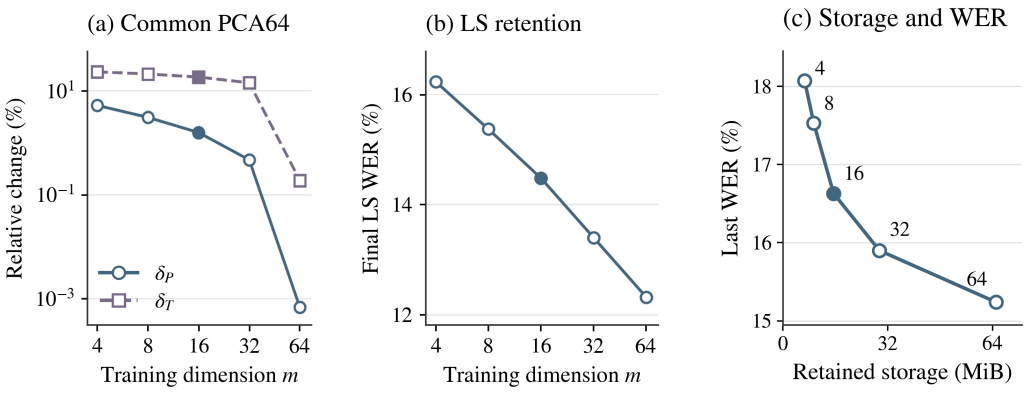

Figure 6: Effect of support dimension. (a) Relative operator changes on the common LS PCA64 support. (b) Final LS WER. (c) Last and retained storage after three tasks. Labels indicate the training dimension m; filled markers identify the default $m = 16$.

To assess input coverage, we collected attention inputs from all 7,678 test utterances at the final checkpoint using teacher forcing with reference transcripts. The training means and bases remained fixed, and aggregates included padding and prefix positions. For nonzero centered inputs, mean relative residual and explained energy are defined as follows:

$$\bar{r} = \text{mean}\left[\frac{\|z - UU^\top z\|_2}{\|z\|_2}\right], \qquad \eta = 100\frac{\sum \|U^\top z\|_2^2}{\sum \|z\|_2^2}, \tag{39}$$

where $z = x - \mu$ denotes a centered input.

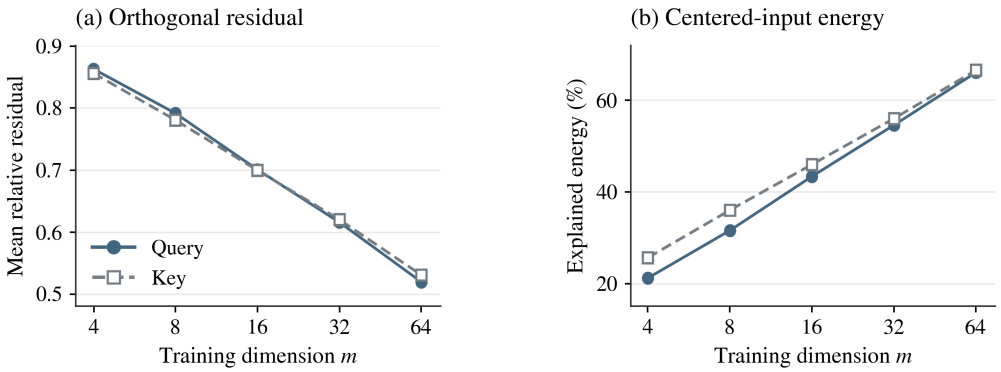

Figure 7: Input variation in the evaluation data captured by the retained supports. (a) Mean relative residual and (b) explained centered input energy at the final checkpoint. Query and key inputs are reported separately, with support dimension on a logarithmic scale with base 2.

Figure 7 shows broader input coverage at larger support dimensions. At $m = 16$, query and key residuals were 0.7010 and 0.6998. Across $m = 16, 32, 64$, explained query energy increased from 43.38% to 54.55% and 65.99%, and key energy increased from 45.98% to 55.98% and 66.48%. Final LS WER decreased from 14.48% to 13.40% and 12.32%, accompanying broader coverage and smaller operator changes. At $m = 16$, 99.97% of query and 99.99% of key augmented PCA coordinates lay within the stored radii. Radius inclusion describes projected coordinate magnitudes; explained energy quantifies retained centered input variation. These statistics cover all evaluation domains.

B.3 Preservation Weight

Table 12 shows that $\lambda = 1.0$ achieved the lowest mean Last, Avg, and forgetting in all three orders, supporting a common default across these sequences. We evaluated $\lambda \in \{0.1, 0.5, 1.0\}$ on Whisper-small with three training seeds, fixed $r = 4$, and $m = 16$. Appendix A.4 gives the separate single seed comparison with Individual.

Table 12: Sensitivity to the preservation weight on Whisper-small in Orders 1, 2, and 3.

Order	λ	Last (↓)	Avg (↓)	F (↓)
1	0.1	18.34 ± 0.18	17.87 ± 0.16	6.48 ± 0.17
	0.5	17.80 ± 0.17	17.74 ± 0.15	5.95 ± 0.14
	1.0	$\mathbf{17.13 \pm 0.31}$	$\mathbf{17.58 \pm 0.24}$	$\mathbf{5.03 \pm 0.28}$
2	0.1	14.60 ± 0.17	15.74 ± 0.20	1.20 ± 0.15
	0.5	14.80 ± 0.20	15.76 ± 0.17	1.10 ± 0.14
	1.0	$\mathbf{14.57 \pm 0.25}$	$\mathbf{15.48 \pm 0.18}$	$\mathbf{0.90 \pm 0.20}$
3	0.1	16.94 ± 0.18	17.55 ± 0.16	4.39 ± 0.18
	0.5	17.40 ± 0.16	17.81 ± 0.15	4.85 ± 0.16
	1.0	$\mathbf{16.48 \pm 0.45}$	$\mathbf{17.25 \pm 0.33}$	$\mathbf{3.67 \pm 0.44}$

B.4 Storage and Computational Cost

Table 13 shows the recognition gains obtained with larger supports in the single seed sweep. The default $m = 16$ improved Last, Avg, and forgetting over smaller supports using 15.48 MiB of storage; larger dimensions further improved Last and retention. Storage includes input summaries, score references, transport coefficients, and low-rank update histories.

Table 13: Recognition performance and retained storage across support dimensions. Storage is measured in MiB.

m	Last (↓)	Avg (↓)	F (↓)	Storage
4	18.07	18.34	5.22	6.64
8	17.53	17.94	5.05	9.43
16	16.63	17.38	3.81	15.48
32	15.90	17.48	2.46	29.48
64	**15.24**	**16.83**	**1.91**	65.08

Support bases grow linearly with m, and score references grow quadratically. Low-rank V/O histories provide compact reconstruction of transport references for past domains. Single runs on a shared server measured training times of 13.64 to 13.99 hours across support dimensions.

Training GPU memory and decoding speed. Table 14 and Figure 8 show that PAF matched O-LoRA's overall decoding RTF with similar peak reserved GPU memory. This single seed Order 3 profile used $\lambda = 1.0$ for PAF and evaluated each final checkpoint on all domains. Dashes denote unavailable logs.

Table 14: Training GPU memory and decoding speed on Whisper-small in Order 3. T1, T2, and T3 correspond to LS, CH3, and TED2 adaptation. Reserved memory includes allocator caches; Max is the largest stage peak. Decoding uses the final model, and overall RTF is total decoding time divided by total audio duration. Bold marks the lowest recorded maximum memory and overall RTF.

	Peak reserved memory (GiB)				**RTF of the final model**			
Method	T1	T2	T3	Max	LS	CH3	TED2	Overall
IncLoRA	—	—	—	—	—	—	—	—
EWC	2.754	9.209	2.518	9.209	0.428	0.412	0.402	0.415
O-LoRA	2.652	9.102	2.414	**9.102**	0.391	0.382	0.397	**0.389**
C-LoRA	2.754	9.963	3.512	9.963	0.433	0.443	0.427	0.435
GainLoRA	2.732	9.287	2.734	9.287	0.916	0.882	0.918	0.903
Attention-Spectrum	2.754	9.221	2.531	9.221	0.390	0.391	0.396	0.392
PAF (ours)	2.754	9.234	2.570	9.234	0.392	0.379	0.398	**0.389**

PAF's maximum reserved memory was 9.234 GiB, 0.132 GiB (1.45%) above O-LoRA and within 0.6% of EWC, GainLoRA, and Attention-Spectrum. Training optimizes the current Q/K/V/O LoRA factors and evaluates preservation from effective weights and retained supports. Each QK operator has $(m + 1)m$ entries and each VO operator has $d_o(m + 1)$ entries, independently of acoustic and decoder sequence lengths. Reserved memory covers current domain activations, temporary tensors, preservation computations, and allocator caches; retained summaries and histories occupy 15.48 MiB separately.

PAF and O-LoRA both achieved an overall RTF of 0.389 at the reported precision. GainLoRA's RTF of 0.903 corresponds to 2.32 times the decoding time, giving PAF a 56.9% reduction on the same audio. PAF merges learned updates into existing weights for inference with a single ASR model; supports, operator references, and histories are used only during adaptation. This design retains the original inference computation and is consistent with the measured decoding cost.

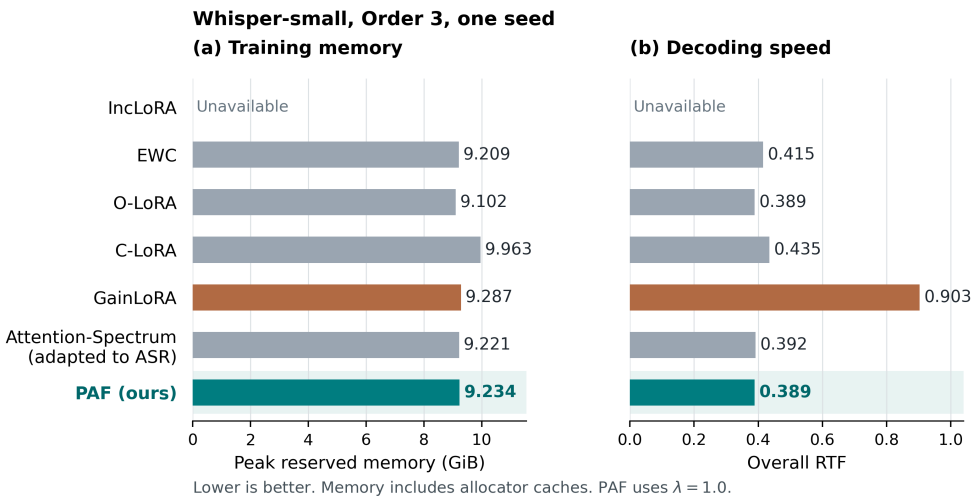

Figure 8: Efficiency on Whisper-small in Order 3. (a) Maximum training GPU peak reserved memory across the three stages. (b) Overall RTF of the final model. Lower values are better. Missing IncLoRA measurements are marked as unavailable.

C Output Bound and Reference Reconstruction

C.1 Proof of Proposition 1

To relate operator preservation to attention output stability on the stored support, write σ for softmax. For $p = \sigma(z)$, we apply the Cauchy–Schwarz inequality as follows:

$$\begin{aligned} \|D\sigma(z)v\|_1 &= \sum_j p_j |v_j - p^\top v| \\ &\le \left(\sum_j p_j v_j^2 - (p^\top v)^2 \right)^{1/2} \\ &\le \|v\|_\infty. \end{aligned} \tag{40}$$

Integrating along the segment between two logit vectors yields $\|\sigma(z) - \sigma(z')\|_1 \le \|z - z'\|_\infty$. Identical inputs within the stored radii and a shared nonempty allowed key set and additive mask give $\|\alpha_\theta - \alpha_s\|_1 \le a_s b_s \|\Delta P\|_F$. Let $g_\theta = \sum_j \alpha_{\theta,j} \bar{w}_j$ and define g_s similarly. Then $\|g_\theta\|_2 \le b_s$ and $\|g_\theta - g_s\|_2 \le a_s b_s^2 \|\Delta P\|_F$. We therefore bound the head output difference as follows:

$$\begin{aligned} \|f_{\theta,s,h} - f_{s,h}\|_2 &= \|\Delta T g_\theta + T_{s,h}(g_\theta - g_s)\|_2 \\ &\le b_s \|\Delta T\|_F + \|T_{s,h}\|_2 a_s b_s^2 \|\Delta P\|_F. \end{aligned} \tag{41}$$

Squaring with $(x+y)^2 \le 2x^2 + 2y^2$ and using $c_{s,h} = \|T_{s,h}\|_2^2$ gives the squared bound with $\ell_{s,h}(\theta)$ defined in Equation (17).

C.2 Stability of the Complete ASR Model

We extend Proposition 1 to the complete recognizer. The current and reference models differ only in the regularized attention projections and operate in evaluation mode. For the same acoustic input a, decoder prefix p, and attention masks, let $g_\theta(a, p)$ and $g_s(a, p)$ denote their next token logits. Let $I_{s,\ell}$ contain the reference query and key/value inputs of attention block ℓ, and let $\widehat{I}_{s,\ell}$ be their projections onto the bounded stored supports. Define $\rho_{s,\ell}(a, p) = \|I_{s,\ell} - \widehat{I}_{s,\ell}\|$, using the maximum Euclidean norm over positions and input streams. This distance accounts for both components outside the affine subspaces and coordinates exceeding the stored radii.

Let $L_{\theta,\ell}$ and $L_{s,\ell}$ bound the input sensitivities of the current and reference attention maps between these inputs and their support projections, using the maximum Euclidean norm over output positions.

Let κ_ℓ bound the sensitivity of the final logits, in infinity norm, to perturbations of the attention output of block ℓ in this norm along the hybrid replacements used below. These constants may depend on the model pair, acoustic input, and prefix.

Proposition 2 (Complete model stability) Under finite sensitivity bounds, the logit deviation is bounded as follows:

$$\|g_\theta(a,p) - g_s(a,p)\|_\infty \le B_s(a,p), \tag{42}$$

where we define the accumulated deviation bound as follows:

$$B_s(a,p) = \sum_{\ell\in\mathcal{B}} \kappa_\ell \Bigg[\sum_{h=1}^{H_\ell} \sqrt{\ell_{s,\ell,h}(\theta)} + (L_{\theta,\ell} + L_{s,\ell})\rho_{s,\ell}(a,p) \Bigg]. \tag{43}$$

For a transcript y containing n tokens, including the end token, let $\mathrm{NLL}_\theta(a,y)$ denote the mean token negative log-likelihood evaluated with the transcript prefixes. Its increase is bounded as follows:

$$\mathrm{NLL}_\theta(a,y) - \mathrm{NLL}_s(a,y) \le \frac{2}{n}\sum_{i=1}^{n} B_s(a, y_{<i}). \tag{44}$$

Proof. At the projected inputs, Proposition 1 bounds each head deviation by $\sqrt{\ell_{s,\ell,h}(\theta)}$. Summing over heads and applying the triangle inequality through $\widehat{I}_{s,\ell}$ bounds the attention output deviation at $I_{s,\ell}$ by $\sum_h \sqrt{\ell_{s,\ell,h}(\theta)} + (L_{\theta,\ell} + L_{s,\ell})\rho_{s,\ell}$. Replace current attention maps with reference maps in topological order. Each replacement receives reference upstream inputs and retains current computations downstream; κ_ℓ bounds its effect on the final logits by the corresponding attention output deviation. The computational state retains encoder memory and residual streams, accounting for cross-attention and residual connections. The downstream sensitivity includes subsequent attention, feedforward, normalization, and output computations. Summing the changes over these replacements yields Equation (42). For token cross-entropy, the logit gradient is $\mathrm{softmax}(g) - e_y$, where e_y is the unit vector for the target token, and its ℓ_1 norm is at most 2. Thus, token negative log-likelihood is 2-Lipschitz in the logit infinity norm, giving Equation (44).

Exact recognition preservation. Suppose every head preservation loss is zero and all reference attention inputs lie within their bounded stored supports for every prefix evaluated by the reference decoder. With the same deterministic decoding procedure and tie breaking, the attention outputs and subsequent computations agree in topological order. Equality of the logits preserves the evaluated prefixes, decoded transcription, and its WER relative to the reference checkpoint. If these conditions hold for every evaluated utterance of each domain s at every subsequent stage $t > s$, then $E_{s,t} = E_{s,s}$ throughout the sequence, and the forgetting metric in Equation (25) is zero.

C.3 Reference Reconstruction

PAF represents past attention operators using compact summaries and low-rank V/O update histories. For each domain and block, the retained summary is defined as follows:

$$\mathcal{C}_s = \left\{\mu_s^q, U_s^q, \mu_s^{kv}, U_s^{kv}, a_s, b_s, \{P_{s,h}, c_{s,h}\}_{h=1}^{H}\right\}, \tag{45}$$

where H is the number of heads. Retained factors and scales specify the merged updates $\Delta W_{X,j} = \gamma_{X,j} A_{X,j} B_{X,j}$. At stage t, the weights for domain s are reconstructed as follows:

$$\widehat{W}_X^{[s]} = W_X^{[t-1]} - \sum_{j=s+1}^{t-1} \Delta W_{X,j}, \tag{46}$$

where $X \in \{V, O\}$ and the frozen anchor $W_X^{[t-1]}$ contains the updates merged through stage $t-1$. The sum is zero for $s = t-1$. We compose the reconstructed V and O weights to obtain the transport reference as follows:

$$\widehat{T}_{s,h} = \widehat{O}_h^{[s]} \left[\widehat{W}_V^{[s],h} U_s^{kv},\ \widehat{W}_V^{[s],h} \mu_s^{kv} + b_V^h\right], \tag{47}$$

where $\widehat{O}_h^{[s]}$ is the reconstructed output block and b_V^h is the fixed value bias. The target $\widehat{T}_{s,h}$ in Equation (17) remains fixed within each stage, and gradients pass through the current operators to the current LoRA factors.

D Retained Information and Preservation Mechanisms

Table 15 distinguishes the preservation targets in the Whisper-small implementations. PAF compares composite attention operators in fixed input coordinates; the other preservation methods compare LoRA factors, effective updates, gate gradients, or statistical descriptors. Following Equation (1), $\Delta W_t = \gamma A_t B_t$, where A_t and B_t are the output and input factors, $\gamma = \alpha_{\mathrm{LoRA}}/r$, and $s < t$ indexes an earlier task.

Table 15: Retained information, current comparison targets, and preservation mechanisms in the Whisper-small implementations. Attention-Spectrum denotes the adaptation to ASR summarized in the table.

Method	**Retained information**	**Current comparison target**	**Preservation mechanism**
IncLoRA	Model with previous updates merged	None	Sequential adaptation with fresh LoRA factors; no explicit preservation penalty.
EWC	Reference LoRA factors and parameter importance estimates	Current LoRA factors A and B	Applies importance weighted quadratic penalties separately to deviations of A and B from their references.
O-LoRA	Previous output factors A_s	Current output factor A_t	Penalizes $\|A_s^\top A_t\|_F^2$ to encourage orthogonality between output subspaces.
C-LoRA	Previous task factors A_s and B_s	Current effective update ΔW_t	Penalizes elementwise overlap between accumulated previous updates and the current update.
GainLoRA	Previous LoRA branches and gates; GPM bases of gate input representations	Current output factors and gate gradients	Applies O-LoRA to the new branch and projects gate gradients away from retained input subspaces.
Attention-Spectrum (adapted to ASR)	One transcribe prototype: 54D mean and diagonal variance; 8D mean angular profile	Current 54D spectral descriptor and 8D angular profile	Confidence weighted combination of dimension averaged squared Mahalanobis distance and symmetric angular KL divergence.
PAF (ours)	Input means, bases, and radii; QK references; transport norm coefficients and V/O update histories	Current QK and VO operators on each stored support	Penalizes weighted operator deviations from fixed references in the same input coordinates.

E Limitations

The complete model bound depends on support coverage and network sensitivities, and exact recognition preservation requires zero preservation error and complete coverage across decoding prefixes. The experiments cover English ASR sequences of up to six domains. Retained storage and preservation computation grow with the number of domains. Future work will extend evaluation to multilingual speech and longer sequences and allocate support dimensions under a fixed memory budget.