Navigating the Reality Gap: Privacy-Preserving On-Device Continual Adaptation of ASR for Clinical Telephony

Anonymous ACL submission

Abstract

Automatic Speech Recognition (ASR) holds immense potential to assist in clinical documentation and patient report generation, particularly in resource-constrained regions. However, deployment is currently hindered by a technical deadlock: a severe **"Reality Gap"** between laboratory performance and noisy, real-world clinical audio, coupled with strict privacy and resource constraints. Such adaptation is essential for clinical telephony systems, where patient speech is highly variable and transcription errors can directly impact downstream clinical workflows. We quantify this gap, showing that a robust multilingual model (IndicWav2Vec) degrades up to a **40.94% WER** on rural clinical telephony speech from India, rendering it unusable. We demonstrate consistent improvements on these helpline interactions without transmitting raw patient data off-device via an on-device continual adaptation framework using Low-Rank Adaptation (LoRA). We conduct an investigative study of stabilization strategies, characterizing the trade-offs between data-driven and parameter-driven approaches. Our results demonstrate that multi-domain Experience Replay (ER) yields the primary performance gains, achieving a **17.1% relative improvement** in target WER and reducing catastrophic forgetting by **55%** compared to naive adaptation. Furthermore, we investigate a stabilized importance estimation strategy (Absolute Fisher) to ensure robust convergence against the high-variance gradients common in clinical telephony speech. Finally, we verify via a domain-specific spot check that acoustic adaptation is a fundamental prerequisite for usability in healthcare settings which cannot be bypassed by language models alone.

1 Introduction

The recent surge in Self-Supervised Learning (SSL) has propelled Automatic Speech Recognition (ASR) to near-human performance on standardized benchmarks. Foundational models like Meta's Wav2Vec 2.0 (Baevski et al., 2020) and OpenAI's Whisper (Radford et al., 2023) promise a future where automated transcription can digitize patient reports, allowing clinics to improve throughput with reduced operational costs. However, for specialized, high-impact clinical domains such as rural health helplines and remote clinical services, this promise remains unfulfilled. The "reality gap," representing the disparity between clean training corpora and the chaotic, noisy, and privacy-constrained environments of real-world clinical telephony, renders these state-of-the-art models practically unusable.

Our baseline analysis reveals that even robust multilingual models like IndicWav2Vec (Javed et al., 2022) degrade to a prohibitively high **40.94% Word Error Rate (WER)** when exposed to real-world clinical telephony audio from rural India (Bhanushali et al., 2022). This failure is compounded by a technical deadlock: patient data privacy laws preclude the use of cloud-based adaptation services (Leroy et al., 2019), while rural infrastructure lacks the high-end compute required for traditional model retraining. This creates a scenario where models cannot improve because data cannot leave the local environment, and localized compute is too constrained for standard fine-tuning.

To break this deadlock, we propose an on-device adaptation framework for localized, stream-based learning. We focus on data residency: ensuring that raw patient audio never leaves the local device. By leveraging Low-Rank Adaptation (LoRA) (Hu et al., 2022), we enable the model to fine-tune on incoming clinical data streams using a fraction of the trainable parameters. However, sequential adaptation in long-running clinical deployments introduces the risk of *catastrophic forgetting*, where the model loses its foundational linguistic capabilities (McCloskey and Cohen, 1989). While recent benchmarks on real-world Indian speech, such as NIRANTAR (Javed et al., 2025), indicate that

standard regularization methods (EWC, MAS) often fail to prevent this forgetting in diverse domains, we hypothesize that this instability is exacerbated by the acoustic mismatch inherent in low-bandwidth clinical telephony. To counteract this, we integrate a multi-domain experience replay mechanism (Chaudhry et al., 2019b), interleaving small buffers of general-domain data with the incoming clinical stream to anchor the learning trajectory.

Furthermore, we investigate parameter-regularization strategies to stabilize this pipeline. While standard quadratic formulations (EWC) can be sensitive to the high-variance gradients of noisy telephony speech, we observe that a balanced importance estimation (Absolute Fisher) (Benzing, 2021) provides a more resilient regularization signal for this localized application. Our results demonstrate that these strategies yield a 17.1% relative reduction in WER on the target clinical domain, effectively stabilizing the model's performance while respecting rigid data residency constraints. The primary research contributions of our work are:

- **On-Device Clinical Adaptation:** A localized continual adaptation framework for clinical telephony ASR that respects healthcare data residency constraints while maintaining computational efficiency.

- **Empirical Validation on Clinical Speech:** Parameter-efficient adaptation (LoRA) combined with multi-domain memory (Replay) yields a 17.1% relative improvement in WER on real-world rural clinical speech without centralized retraining or raw data exfiltration.

- **Investigation of Optimization Robustness:** The interplay between data-driven replay and parameter-driven regularization, validating that linearized Fisher importance promotes more reliable convergence in noisy, multi-speaker clinical deployments.

- **Quantification of the Clinical Reality Gap:** Establish a rigorous benchmark for 8kHz clinical telephony, providing insights into the technical prerequisites to close the gap between foundational ASR models and frontline healthcare needs.

2 Related Work

Designing a privacy-preserving, adaptive ASR pipeline for low-resource medical settings requires synthesizing advancements in self-supervised acoustic modeling, parameter-efficient adaptation, and continual learning. We address critical gaps regarding adaptation efficiency, data scarcity, and optimization stability.

2.1 Self-Supervised Acoustic Foundations and the Domain Gap

Low-resource speech recognition increasingly relies on Self-Supervised Learning (SSL) to leverage unlabeled audio. We build upon Wav2Vec 2.0 (Baevski et al., 2020), which achieves high data efficiency through contrastive learning; Baevski et al. (2020) showed that fine-tuning on just ten minutes of data yields competitive performance on benchmarks like Librispeech (Panayotov et al., 2015). To support linguistic diversity, we utilize IndicWav2Vec (Javed et al., 2022), scaling the architecture to cover 40 Indian languages.

The Gap: Despite these capabilities, a "usability gap" exists between laboratory benchmarks and real-world deployment. Clinical ASR errors are particularly risky because they affect medically salient information such as dosages, negations, and symptom reporting (Chiu et al., 2018a,b). This is exacerbated in rural healthcare by noisy telephonic audio and diverse regional dialects. Recent large-scale efforts such as NIRANTAR (Javed et al., 2025) have begun characterizing these challenges across 22 Indian languages. We utilize the Gram Vaani ASR dataset (Bhanushali et al., 2022) to represent these challenges; Bhanushali et al. (2022) highlight that standard models struggle to generalize to such 8kHz telephony audio without targeted adaptation.

While signal-level enhancement strategies, such as bandwidth expansion and speech super-resolution (Lin et al., 2023; Li et al., 2019), offer an alternative by reconstructing high-frequency components, they introduce significant computational overhead. Deploying a separate BWE model (often GAN or diffusion-based) alongside the ASR system contradicts the strict low-latency requirements of edge devices. Furthermore, privacy constraints preclude cloud-based services (Leroy et al., 2019), creating a deadlock where models cannot improve because data cannot leave the local environment.

2.2 Parameter-Efficient Fine-Tuning (PEFT) and the Efficiency Gap

Bridging the domain gap typically requires fine-tuning on target data. However, for Large Audio Models (LAMs), full fine-tuning is computationally prohibitive and prone to overfitting. This presents an *efficiency gap*: high-end GPUs are unavailable on the resource-constrained edge devices of rural hospitals.

To address this, we adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022), which enables local fine-tuning on modest clinical hardware. This approach is essential for meeting healthcare **Data Residency** requirements, as it allows for localized, high-performance adaptation without needing the centralized data storage often required by federated or cloud-based approaches (Rieke et al., 2020). Recent work by Song et al. (2024) on LoRA-Whisper validates this approach for multilingual ASR, demonstrating that PEFT can effectively adapt large Transformers without catastrophic interference. By constraining optimization to a low-dimensional subspace, LoRA resolves deployment challenges in privacy-sensitive clinics.

While PEFT solves efficiency, self-improving systems face the *stability gap* of Continual Learning (CL), specifically *Catastrophic Forgetting*, where new training erodes previously learned capabilities (McCloskey and Cohen, 1989). In healthcare, ensuring model stability over time is a critical safety requirement for clinical evaluation cycles (De Lange et al., 2022). In ASR, naive updates on recent telephony transcripts lead to overfitting on specific speakers or acoustic conditions. Benchmarks like NIRANTAR (Javed et al., 2025) further underscore this, demonstrating that no single CL strategy currently yields consistent performance across the diverse shifts encountered in real-world Indian speech.

To mitigate this, we employ a hybrid continual learning strategy that synergizes data-level and parameter-level consolidation. First, to address the distribution shift between segments, we implement **Experience Replay (ER)** (Chaudhry et al., 2019a). Building on recent regularized CL studies for Indic ASR (T and Nirmala, 2025) and episodic memory frameworks (Yang et al., 2022), we utilize a *multi-domain experience replay* mechanism. While Yang et al. (2022) focused on gradient projection, we find that a direct data mixing strategy effectively mitigates catastrophic forgetting when combined with PEFT. Our buffer retains a mixture of "hard" examples (high-loss) from the target domain (Gram Vaani) and gender-balanced samples from a clean auxiliary domain (Kathbath).

Complementing this, we constrain parameter drift using **Online LoRA-EWC** (Xiang et al., 2023), utilizing the **Absolute Fisher** importance estimation (Benzing, 2021). We estimate importance using the *absolute value* of accumulated gradients ($|g|$) rather than the standard squared Fisher Information. While this linear scaling draws inspiration from the numerical stability of Memory Aware Synapses (MAS) (Aljundi et al., 2018), Absolute Fisher explicitly ties the importance to the likelihood of the adaptation task rather than unsupervised output sensitivity. By combining this regularized signal with multi-domain replay, our approach prevents catastrophic forgetting through a dual mechanism: replay maintains the data distribution, while the regularization penalty preserves critical adapter parameters.

3 Methodology

We propose an on-device, adaptive ASR framework designed to bridge the "reality gap" and ensure deployment-ready reliability within the acoustic constraints of rural clinical telephony. Our approach focuses on Continual Learning (CL), enabling a pre-trained model to adapt sequentially to incoming clinical data streams ($\mathcal{D}_{stream}$) while maintaining strict data residency and preventing catastrophic forgetting.

3.1 Base Acoustic Backbone

We utilize **IndicWav2Vec** as our acoustic backbone. This model is built on the Wav2Vec 2.0 architecture and pre-trained on a massive corpus of diverse Indian languages. The network consists of a convolutional feature encoder $f(x)$ mapping raw audio to latent representations, followed by a Transformer context network optimized via Connectionist Temporal Classification (CTC) loss:

$$\mathcal{L}_{CTC} = -\log P(y|x) \quad (1)$$

To enable efficient adaptation on edge devices, we freeze the base model and inject trainable Low-Rank Adaptation (LoRA) matrices into the query and value projection layers. This parameter-efficient approach serves as a stability safeguard, preventing large-scale corruption of pre-trained

weights while allowing for specialized clinical adaptation.

3.2 Data-Driven Stability: Multi-Domain Experience Replay

To mitigate catastrophic forgetting, we employ Experience Replay (ER), explicitly grounding the model's optimization trajectory with historical data. We implement a **Multi-Domain Replay** strategy that maintains a dual-source buffer $\mathcal{B}$:

- **General Domain Anchor ($\mathcal{B}_{gen}$):** A gender-balanced subset of high-resource, standard Hindi samples (sourced from Kathbath) to preserve foundational phonetic robustness, which is essential for handling diverse patient demographics in a remote clinical setting.
- **Target Domain History ($\mathcal{B}_{spec}$):** A sliding window of "hard" examples (high loss) and random samples from previous clinical segments, ensuring retention of recent domain-specific adaptations.

During training, the incoming clinical stream $\mathcal{D}_{stream}$ is concatenated with samples from $\mathcal{B}$. The optimization objective for Experience Replay is:

$$\mathcal{L}_{ER} = \gamma \mathbb{E}_{x \sim \mathcal{D}_{stream}}[\mathcal{L}_{CTC}(x)] + (1-\gamma)\mathbb{E}_{x \sim \mathcal{B}}[\mathcal{L}_{CTC}(x)] \quad (2)$$

where γ represents the mixing ratio between new data and replayed data.

3.3 Parameter-Driven Stability: Elastic Weight Consolidation (EWC)

As a standalone strategy (V4.5) and a component of our hybrid approach, we implement Elastic Weight Consolidation (EWC) to prevent drift in critical parameters. Unlike ER, which requires data storage, EWC regularizes the model by penalizing changes to weights that are important for previous tasks.

We compute the importance of each LoRA parameter θ_i using the diagonal of the Fisher Information Matrix (F). To ensure robust optimization against high-variance gradients, we approximate importance using accumulated **absolute gradients** rather than squared gradients. This "Absolute Fisher" approach (Benzing, 2021) functions as a reliability optimization. We distinguish this from Memory Aware Synapses (MAS) (Aljundi et al., 2018); while MAS assesses importance based on the sensitivity of the output function, Absolute Fisher derives importance from a linearized measure of the likelihood's Fisher Information, preventing transient acoustic artifacts from drowning out meaningful weights:

$$F_i = \frac{1}{N}\sum_{j=1}^{N} |\nabla_{\theta_i} \log P(y_j|x_j)| \quad (3)$$

This linear accumulation prevents outliers from dominating the importance metric. The EWC regularization loss is then applied as a quadratic penalty:

$$\mathcal{L}_{EWC}(\theta) = \frac{\lambda}{2}\sum_i F_i(\theta_i - \theta_i^*)^2 \quad (4)$$

where θ_i^* represents the frozen optimal parameters from the previous segment and λ controls the regularization strength.

3.4 Hybrid Optimization Framework (V5.1)

To leverage the synergistic effects of data-driven grounding and parameter-driven constraints, we propose a Hybrid ER + EWC optimization framework (Strategy V5.1). This approach combines the Multi-Domain replay buffer with the EWC regularization penalty.

The total optimization objective minimizes the transcription error over the mixed batch while simultaneously constraining parameter drift:

$$\mathcal{L}_{Total} = \mathcal{L}_{ER}(\mathcal{D}_{stream}, \mathcal{B}) + \mathcal{L}_{EWC}(\theta) \quad (5)$$

By using $\mathcal{L}_{ER}$ to provide the necessary gradients for acoustic adaptation and $\mathcal{L}_{EWC}$ to define a "safe" optimization region, this hybrid mechanism aims to maximize target domain accuracy while minimizing catastrophic forgetting.

4 Experiments and Results

We conducted a comprehensive evaluation to validate the effectiveness of our adaptive pipeline. The experimental design focuses on two key aspects: the ability to adapt to a specific clinical domain (Gram Vaani) and the ability to retain general linguistic knowledge (Kathbath) to prevent catastrophic forgetting.

4.1 Experimental Setup

4.1.1 Metrics

We evaluate performance using two standard metrics: Word Error Rate (WER) and Character Error Rate (CER). WER measures transcription accuracy at the word level, while CER provides a finer-grained analysis of phonetic accuracy, particularly useful for agglutinative languages and dialectal variations.

$$\text{WER} = \frac{S_w + D_w + I_w}{N_w}$$

$$\text{CER} = \frac{S_c + D_c + I_c}{N_c}$$

4.1.2 Datasets

- **Gram Vaani (Clinical Helpline Proxy)** (Bhanushali et al., 2022): This dataset consists of rural telephonic speech (originally 8kHz, upsampled to 16kHz) and serves as a proxy for audio encountered in frontline healthcare helplines. We evaluate on the official 3-hour Evaluation set (GV_Eval_3h). Note that prior work (Patel and Scharenborg, 2022) reported 30.3% WER on the 5-hour Development set using full Conformer models. Our result (33.94%) reflects performance on the unseen Evaluation partition under strictly constrained on-device conditions (LoRA vs. full tuning). To strictly simulate a continual clinical learning scenario, we partition the 103 hours of training data into sequential segments, processing them one by one to mimic a live data stream while ensuring patient data residency.

- **Kathbath (General Domain)**: A high-quality, read speech dataset representing standard Hindi (Javed et al., 2023). We utilize a subset of the training set (approx. 25,800 samples) to populate the experience replay buffer, ensuring the model retains knowledge of standard Hindi. The complete validation set (3,151 samples) is used exclusively to measure catastrophic forgetting after adaptation.

4.2 Continual Adaptation Paradigms

To characterize the optimal pathway for localized adaptation, we investigate four distinct continual learning paradigms. For all experiments, we utilize a base `IndicWav2Vec` model, which provides a pre-training baseline of **40.94% WER** on the target rural clinical domain and **11.57%** on the general Kathbath domain.

4.2.1 Paradigm 1: Naive Continual Fine-tuning (Baseline)

This represents the simplest form of adaptation, where the model is fine-tuned sequentially on incoming clinical data streams without explicit regularization. We utilize a LoRA rank of 16 and $\alpha = 32$.

- **Outcome (V1.1)**: While the model achieves a significant reduction in target WER (down to **34.00%**), it suffers from severe catastrophic forgetting, with general domain error increasing by an absolute **5.93%**.

4.2.2 Paradigm 2: Experience Replay (ER) Strategies

We investigate two ER configurations to stabilize the adaptation trajectory via data rehearsal. We increase the LoRA capacity to rank 24 ($\alpha = 48$) to accommodate the dual-task nature of rehearsal.

- **Single-Domain ER (V2.1)**: Replays 400 samples from the target-domain history per segment using a "60% hard, 40% random" selection strategy.

- **Multi-Domain ER (V3.1)**: Replays a balanced buffer of 300 target samples and 300 general-domain (Kathbath) samples. This strategy achieved our lowest target WER of **33.94%** and reduced forgetting by **55%** relative to the naive baseline.

4.2.3 Paradigm 3: Elastic Weight Consolidation (EWC)

We examine whether parameter-regularization can implicitly protect the model's foundational knowledge without replaying data. We utilize the **Absolute Fisher importance** (F_i) (Benzing, 2021) to penalize changes to weights deemed critical for the source domain.

- **Outcome (V4.5)**: Initial experiments (V4.1–V4.3) with high λ values caused the model to "freeze," preventing any adaptation to the noisy clinical domain. By relaxing the constraint ($\lambda = 10$) and utilizing linearized importance, V4.5 achieved a competitive **33.94% WER** on the target domain, though it remained less efficient at mitigating forgetting than the ER strategies.

Paradigm	Strategy	Final Target WER	Improvement (%)	Final General WER	Forgetting
Baseline	Pre-trained IndicWav2Vec	40.94%	-	11.57%	-
Naive	V1.1 Naive Fine-tuning	34.00%	+17.0%	17.50%	+5.93%
Experience Replay	V2.1 Single-Domain ER	33.98%	+17.0%	14.90%	+3.33%
Experience Replay	**V3.1 Multi-Domain ER**	**33.94%**	**+17.1%**	**14.23%**	**+2.66%**
EWC	V4.5 EWC ($\lambda = 1e1$)	33.94%	+17.1%	15.15%	+3.58%
Hybrid	V5.1 ER + EWC	34.51%	+15.7%	14.14%	+2.57%

Table 1: Comprehensive performance comparison of the investigated continual learning paradigms. Improvement is relative to the IndicWav2Vec baseline on rural clinical data. Forgetting is the absolute increase in WER on the general Kathbath domain.

4.2.4 Paradigm 4: Hybrid ER + EWC Optimization

Finally, we evaluate the hybrid framework (V5.1) described in Section 3.4, which combines the data-driven guidance of Multi-Domain ER with the parameter-driven Absolute Fisher constraints.

- **Outcome (V5.1)**: This paradigm yielded our most stable model, achieving the lowest absolute forgetting rate with a final general WER of **14.14%** (+2.57% increase). However, this stability came at a slight cost to plasticity, with a final target WER of **34.51%**.

5 Analysis and Discussion

5.1 Comparative Performance Summary

Table 1 summarizes the final metrics for the five primary investigative strategies. All proposed paradigms successfully bridge the "Reality Gap," reducing error on the clinical domain by at least 15% relative.

5.2 Stability-Plasticity Pareto Efficiency

The efficiency of our adaptation strategies is best characterized through a Pareto analysis of retention versus adaptation (Figure 3). We observe that with the current amount of continual learning, multiple successful strategies converge to a range around 34-35% WER. Below this threshold, further reductions in target WER begin to incur disproportionately high stability costs. For instance, while Multi-Domain ER (V3.1) successfully navigates this trade-off, attempts to achieve deeper adaptation in the Hybrid model (V5.1) encountered a stability bottleneck. This behavioral ceiling suggests that for a given model architecture and domain mismatch, there exists an optimal frontier of acoustic adaptation. Crucially, the V3.1 trajectory has not plateaued but has entered a logarithmic long-tail learning phase, suggesting that further gains are possible but require significantly more data or linguistic context to unlock.

5.3 Mechanism Analysis: Data vs. Parameter Anchors

Our investigation reveals a fundamental difference between data-driven and parameter-driven regularization in lightweight models.

- **Data Anchors (ER)**: Replaying Kathbath samples provides an explicit gradient signal toward a generalized linguistic center. This was the most efficient stabilizer, as shown by the V3.1 Pareto trajectory.
- **Parameter Anchors (EWC)**: While EWC protects weights, it is "acoustically blind," penalizing drift without reconciling new acoustic features. This is evidenced by V2.1 (ER) providing superior stabilization (+3.33% forgetting) compared to V4.5 (EWC, +3.58%) despite similar target performance.
- **The LoRA-Regularization Bottleneck**: We observe a conflict in the Hybrid model (V5.1), where combining LoRA's low-rank space with EWC's parameter-level penalties overly restricts optimization paths, leading to higher target WER.

5.4 Convergence Dynamics

While the Naive baseline (V1.1) exhibits erratic loss curves, the ER and Hybrid paradigms demonstrate controlled, monotonic trends (Figure 1). This stability is a critical requirement for clinical deployment, ensuring the system does not regress after extended use. However, we observe a "soft floor" around 34% WER, which we characterize as the **Acoustic Bottleneck of Clinical Telephony**. The mismatch between 8kHz telephony and the 16kHz model expectation creates a ceiling for acoustic-only adaptation, suggesting that further gains may

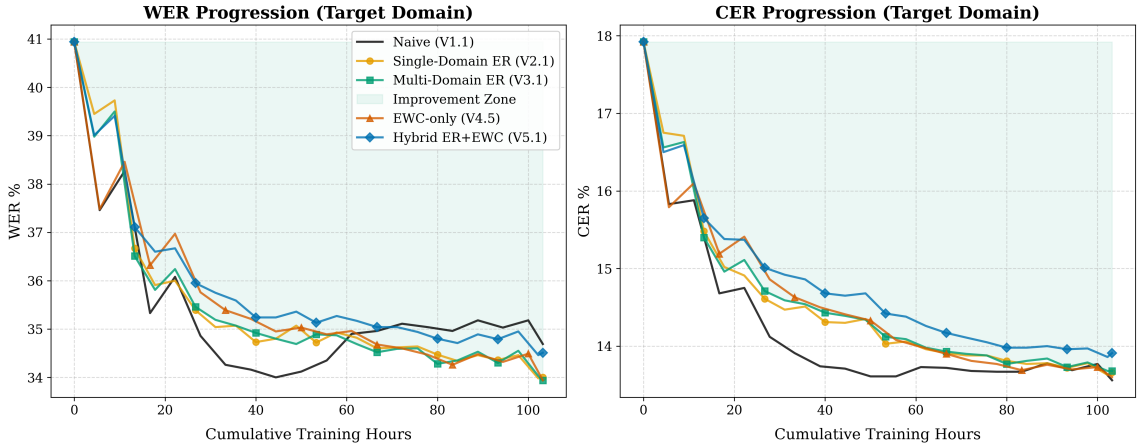

Figure 1: Progression of recognition performance on the target rural clinical data over 100 cumulative training hours. All investigated paradigms successfully bridge the "Reality Gap," with Multi-Domain ER (V3.1) achieving the deepest adaptation.

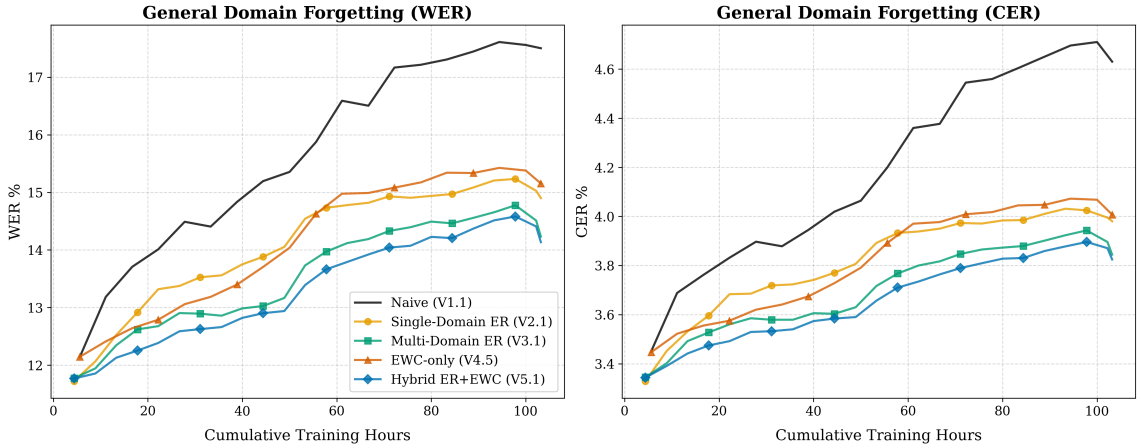

Figure 2: Catastrophic forgetting analysis on the general Kathbath domain. The plot illustrates how knowledge retention is preserved over time. Standard Naive adaptation (V1.1) and EWC-only (V4.5) show significant drift, whereas Multi-Domain ER and Hybrid models successfully flatten the error trajectory.

require specialized front-end processing or linguistic context.

5.5 The Necessity of Acoustic Adaptation

To verify that acoustic adaptation is a *prerequisite* for clinical deployment, we paired the baseline and adapted models with a domain-specific 4-gram LM (Table 2). The unadapted baseline+LM achieves only **34.96% WER**, while V3.1+LM reaches **30.26% WER**. This confirms that **Acoustic Data Residency** is the primary hurdle; solving for local acoustic mismatch is fundamental before linguistic correction can be effectively applied.

Model	WER (%)	CER (%)
Baseline + LM	34.96	17.59
V3.1 Adapted + LM	**30.26**	**15.08**

Table 2: LM Spot Check: Acoustic adaptation remains essential even with LM assistance.

5.6 Clinical Implications and Deployment Feasibility

The **17.1% relative improvement** in WER demonstrated by our Multi-Domain ER strategy (V3.1) represents a significant bridge between "raw" and "usable" clinical ASR. In a clinical dictation setting, this reduction minimizes the manual correction burden on healthcare workers, whose time is a critical resource in rural India. By enabling

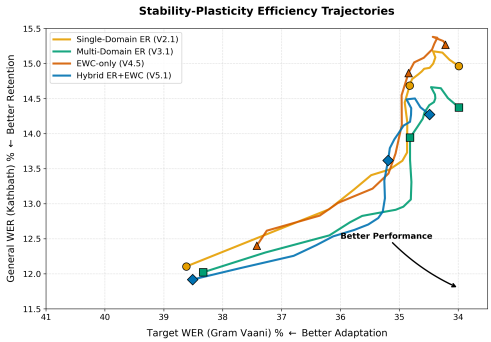

Figure 3: Stability-Plasticity Pareto Analysis: The trajectories illustrate the "cost" of adaptation. Notably, Multi-Domain ER (V3.1) achieves the best overall balance, outperforming other paradigms by maintaining the lowest error rates on both stability and plasticity metrics, whereas Hybrid and EWC approaches incur higher stability costs for similar plasticity gains.

the system to stabilize on local dialects and noisy line conditions without transmitting patient data to central servers, we provide a technical blueprint for scalable, privacy-compliant medical transcription. We note that prior work on the Gram Vaani dataset (Patel and Scharenborg, 2022) achieved a lower WER (30.3%) using heavy-duty architectures (Conformer/TDNN) and external Language Models. However, our framework (33.94%) operates under strict on-device constraints without external LMs during training, reducing computational footprint while ensuring adherence to rigid data residency requirements. Furthermore, the ability to run these adaptations on mobile GPUs ensures that hospitals with limited digital infrastructure can still leverage state-of-the-art ASR foundational models.

5.7 Ablation Studies

5.7.1 Sensitivity to EWC Regularization (λ)

A critical component of our investigation was identifying the optimal regularization strength for EWC. We explored a spectrum of λ values to characterize the trade-offs between model freezing and forgetting. Initial configurations with high regularization ($\lambda \geq 10^3$) resulted in extreme gradient dominance by the EWC constraint, effectively "freezing" the LoRA parameters and preventing any meaningful adaptation to the noisy clinical domain. Conversely, we found that a lower value of $\lambda = 10$ (V4.5) provided the optimal balance, allowing the model to bridge the reality gap while providing sufficient parameter protection.

5.7.2 LR Warmup and Convergence Speed

We evaluated the impact of learning rate warmup schedules on the stability of continual adaptation. By comparing a conservative 100-step warmup against an aggressive 10-step schedule, we observed that the latter significantly accelerated convergence in the early phases of each training segment. Crucially, this aggressive schedule did not result in an increase in catastrophic forgetting ($< 0.05\%$ difference across all runs). This suggests that in the context of LoRA-based ASR adaptation, the model can safely utilize high initial learning rates to rapidly escape local minima from previous segments without overwriting foundational linguistic features.

6 Conclusion

In this work, we investigated the "Reality Gap" that hinders the deployment of state-of-the-art ASR in rural clinical settings. By quantifying this disparity at 40.94% WER, we demonstrated the necessity for localized, privacy-preserving adaptation. Through a rigorous comparative study of four continual learning paradigms, we established that a Multi-Domain Experience Replay strategy provides the most efficient balance of plasticity and stability. Our results show a 17.1% relative improvement in target accuracy and a 55% reduction in catastrophic forgetting compared to naive baselines. Furthermore, our characterization of the trade-off between low-rank optimization and acoustic mismatch provides a technical roadmap for understanding the interplay between stability and plasticity. These findings establish a viable blueprint for building self-improving ASR systems that remain robust and reliable in high-impact, real-world environments.

7 Privacy, Ethics, and Clinical Safety Considerations

The deployment of ASR in frontline healthcare settings introduces unique ethical and safety considerations which our framework explicitly addresses.

Data Governance and Residency: Our primary focus on on-device adaptation is a direct response to healthcare data residency requirements. By ensuring that raw patient audio never leaves the local environment, we mitigate the risk of centralized data breaches and maintain compliance with emerging data protection laws in various jurisdictions (Leroy et al., 2019).

Clinical Safety and Error Risks: While our framework significantly reduces WER, automated transcription in medical contexts is never risk-free. ASR errors in medically salient keywords (e.g., negations or dosage numbers) can lead to serious diagnostic errors. We advocate for a **Clinician-in-the-Loop** model, where our adaptive system serves as a draft-generation tool that reduces manual workload while maintaining human oversight for final clinical validation.

Algorithmic Bias and Equity: By prioritizing adaptation on rural dialects and low-bandwidth telephony, this work contributes to healthcare equity. Standard ASR models often fail on marginalized populations due to accent mismatch; our framework allows localized clinics to "correct" these biases in real-time by adapting specifically to their local community's speech patterns.

8 Limitations and Future Directions

While our framework demonstrates significant potential for clinical deployment, several limitations remain:

1. **Metrics as a Clinical Proxy:** Our evaluation relies on WER and CER, which are standard in ASR but do not explicitly weight the clinical significance of errors. An error in a patient's name may be less critical than an error in a medication dosage; developing medical-concept-aware metrics remains an essential future direction.

2. **Reliance on Clinician Supervision:** Our continual learning assumption assumes a stream of corrected transcripts (e.g., from medical professionals). In under-staffed rural clinics, this supervision may be sparse or delayed, necessitating future work into uncertainty-aware or semi-supervised adaptation.

3. **Acoustic-Only Training:** While we validated the necessity of acoustic adaptation, we did not integrate Language Models (LMs) into the training loop (e.g., via shallow fusion during replay). Capturing the synergy between acoustic and linguistic adaptation is critical for handling specialized medical terminology.

4. **Language and Dialect Scope:** While we focus on rural Hindi dialects, the efficacy of our replay strategy on tonal or Dravidian languages remains to be validated.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, page 144–161, Berlin, Heidelberg. Springer-Verlag.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Frederik Benzing. 2021. Unifying regularisation methods for continual learning.

Anish Bhanushali, Grant Bridgman, Deekshitha G, Prasanta Ghosh, Pratik Kumar, Saurabh Kumar, Adithya Raj Kolladath, Nithya Ravi, Aaditeshwar Seth, Ashish Seth, Abhayjeet Singh, Vrunda Sukhadia, Umesh S, Sathvik Udupa, and Lodagala V. S. V. Durga Prasad. 2022. Gram vaani asr challenge on spontaneous telephone speech recordings in regional variations of hindi. In *Interspeech 2022*, pages 3548–3552.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019a. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019b. Continual learning with tiny episodic memories. *CoRR*, abs/1902.10486.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, and Xuedong Zhang. 2018a. Speech recognition for medical conversations. In *Interspeech 2018*, pages 2972–2976.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, and Xuedong Zhang. 2018b. Speech recognition for medical conversations. In *Interspeech 2018*, pages 2972–2976.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2022. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu

Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Tahir Javed, Kaushal Bhogale, and Mitesh M. Khapra. 2025. NIRANTAR: Continual Learning with New Languages and Domains on Real-world Speech Data. In *Interspeech 2025*, pages 918–922.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh Khapra. 2023. Indicsuperb: A speech processing universal performance benchmark for indian languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:12942–12950.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards building asr systems for the next billion users. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10813–10821.

David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. 2019. Federated learning for keyword spotting. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6341–6345.

Xinyu Li, Venkata Chebiyyam, and Katrin Kirchhoff. 2019. Speech audio super-resolution for speech recognition. In *Interspeech 2019*, pages 3416–3420.

Yin-Tse Lin, Bo-Hao Su, Chi-Han Lin, Shih-Chan Kuo, Jyh-Shing Roger Jang, and Chi-Chun Lee. 2023. Noise-robust bandwidth expansion for 8k speech recordings. In *Interspeech 2023*, pages 5107–5111.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Tanvina Patel and Odette Scharenborg. 2022. Using cross-model learnings for the gram vaani asr challenge 2022. In *Interspeech 2022*, pages 4880–4884.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus H. Maier-Hein, Sébastien Ourselin, Micah J. Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. 2020. The future of digital health with federated learning. *npj Digit. Medicine*, 3.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR. In *Interspeech 2024*, pages 3934–3938.

Gokul Adethya T and S. Jaya Nirmala. 2025. A study on regularization-based continual learning methods for indic asr. *Preprint*, arXiv:2508.06280.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. 2023. Language models meet world models: Embodied experiences enhance language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Muqiao Yang, Ian Lane, and Shinji Watanabe. 2022. Online continual learning of end-to-end speech recognition models. In *Interspeech 2022*, pages 2668–2672.

A Experimental Setup and Reproducibility

To ensure the reproducibility of our findings, we provide the full configuration details for all experimental paradigms. All models were trained for **3 epochs per data segment** to ensure local convergence.

A.1 Hyperparameter Configurations

Table 4 details the parameters used, utilizing a splittable format to distinguish between universal settings and those tailored to specific continual learning strategies.

A.2 Algorithm and Buffer Management

The technical contribution of our framework lies in the efficient integration of LoRA, linearized EWC, and prioritized experience replay. Algorithm 1 below details the localized adaptation loop, and Algorithm 2 describes our multi-domain buffer management strategy.

Algorithm 1: LoRA-based Hybrid Adaptation Loop
Require: Base model θ_{base}, Replay buffer $\mathcal{B}$, Regularization λ
1: Initialize LoRA adapters $\theta_0 \subset \theta_{base}$ and importance $F \leftarrow \mathbf{0}$
2: for segment $k = 1, \ldots, K$ **do**
3: Receive clinical stream $\mathcal{D}_k$
4: $\mathcal{D}_{train} \leftarrow \text{Sample}(\mathcal{D}_k) \cup \text{Sample}(\mathcal{B})$
5: **Adaptation Step:**
$\theta_k \leftarrow \text{argmin}_\theta \mathcal{L}_{CTC}(\mathcal{D}_{train})$
$+\frac{\lambda}{2}\sum F_i(\theta_k - \theta^*_{k-1})^2$
6: **Importance Estimation:**
$F_{new} \leftarrow \frac{1}{N}\sum_{x \in \mathcal{D}_{train}} |\nabla_\theta \mathcal{L}_{CTC}(x)|$
7: **Asynchronous Consolidation:**
$F \leftarrow \frac{F \times (k-1) + F_{new}}{k}$
8: $\theta^*_k \leftarrow \text{detach}(\theta_k)$ *Checkpointing*
9: $\mathcal{B} \leftarrow \text{UpdateBuffer}(\mathcal{D}_k, \mathcal{L}_{inst})$
10: end for

Algorithm 2: Buffer Management
Require: Segment data $\mathcal{D}_k$, Buffer $\mathcal{B}$, Threshold τ
1: Compute $\mathcal{L}_{inst}(x)$ for all instances $x \in \mathcal{D}_k$
2: Hard Example Mining:
Identify $\mathcal{S}_{hard} \leftarrow \{x \in \mathcal{D}_k \mid \mathcal{L}_{inst} > \tau\bar{\mathcal{L}}\}$
3: Buffer Update:
$\mathcal{B}_{gv} \leftarrow \text{Sample}(60\% \text{ from } \mathcal{S}_{hard}$
$\cup\ 40\% \text{ Random})$
$\mathcal{B}_{gen} \leftarrow \text{Sample}(300 \text{ Balanced from Kathbath})$
4: $\mathcal{B} \leftarrow \mathcal{B}_{gv} \cup \mathcal{B}_{gen}$

A.3 Computational Efficiency and Hardware Setup

A core goal of our work is to prove that high-performance, privacy-preserving adaptation is possible on standard mobile workstations. We evaluated our pipeline on the configurations detailed in Table 5. These benchmarks demonstrate the technical feasibility of localized, self-improving ASR systems in environments where high-end compute clusters are unavailable.

B Detailed Experimental Results

This appendix provides the detailed training progression for all experimental configurations. Each plot summarizes the recognition performance (WER/CER) on the target Gram Vaani domain alongside the CTC loss convergence trend.

Gradient Stability: A critical challenge encountered during the development of the EWC-based paradigms was gradient explosion. As shown later in Figure 13, our proposed Linearized EWC (L-EWC) strategy successfully stabilizes the gradient norm throughout the 24 adaptation segments compared to standard quadratic formulations.

Forgetting Analysis: The preservation of foundational general-domain knowledge (Kathbath) is evaluated across all experimental paradigms, with detailed comparisons provided in Figures 9 through 12.

C Dataset Characteristics

The Gram Vaani dataset serves as a rigorous proxy for rural clinical environments due to its telephonic acquisition (originally 8kHz upsampled to 16kHz) and focus on medical/agricultural discussions. Table 3 summarizes the key characteristics.

Table 3: Characteristics of the partitioned Gram Vaani dataset used for continual adaptation.

Metric	Value
Total Duration	103.2 Hours
Number of Segments (k)	24
Samples per Segment	$\approx$ 1,600
Avg. Duration per Sample	9.4 s
Acoustic Condition	Noisy
Primary Dialect	Rural Hindi

Parameter	Value
Optimizer	AdamW
Base Learning Rate	3×10^{-4}
Weight Decay	0.01
Warmup Steps	10 (Aggressive)
LoRA Target Modules	query, value (Attention blocks)
Batch Size (Effective)	64
Max Audio Duration	30.0 seconds
Training Epochs	3 per segment
Strategy-Specific Settings	
V1.1 Naive	$r = 16, \alpha = 32$
V2.1 Single-Domain ER	$r = 24, \alpha = 48$, Buffer: 400 Target
V3.1 Multi-Domain ER	$r = 24, \alpha = 48$, Buffer: 300 Target + 300 General
V4.5 EWC	$r = 24, \alpha = 48, \lambda = 10$
V5.1 Hybrid	$r = 24, \alpha = 48, \lambda = 100$, Buffer: 300 Target + 300 General

Table 4: Comprehensive hyperparameter settings. r and α represent the LoRA rank and scaling factor, respectively. λ denotes the EWC regularization strength.

Configuration	CPU	GPU	Training Time (per segment)
Config A (High-End)	Intel i7-13700H (14 cores)	NVIDIA RTX 4050 (35W)	25–30 minutes
Config B (Mid-Range)	Intel i5-12500H (12 cores)	NVIDIA RTX 3050 (70W)	50–60 minutes

Table 5: Hardware configurations and training time benchmarks for on-device adaptation.

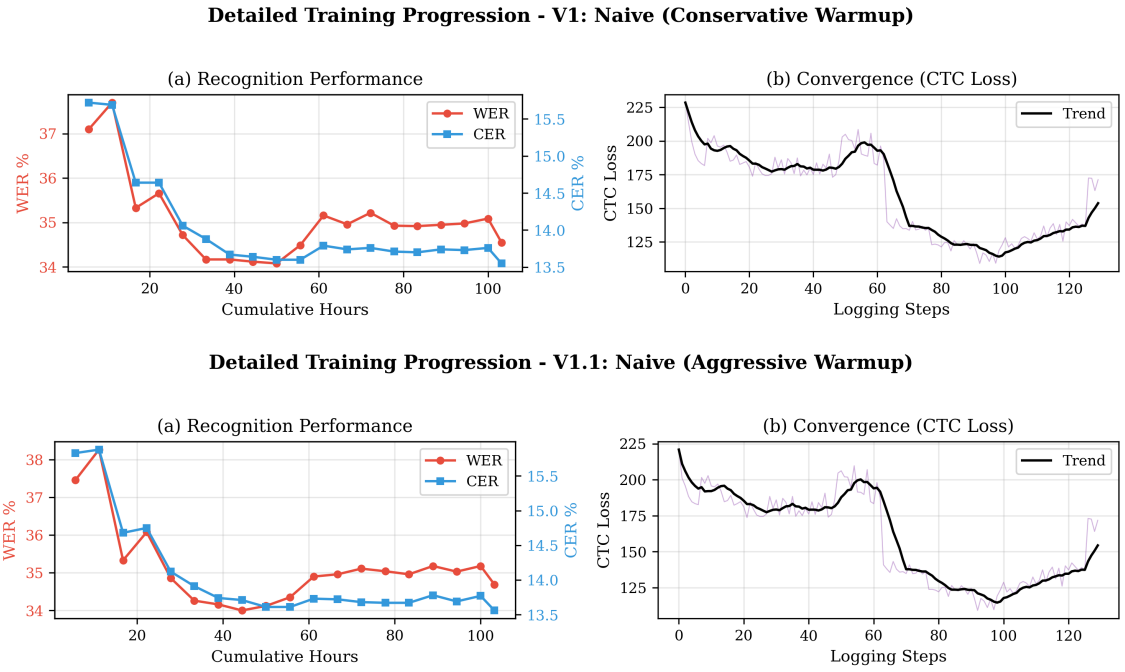

Figure 4: **Naive Baseline Dynamics (V1, V1.1):** Showing rapid initial adaptation but significant volatility, highlighting the need for regularization.

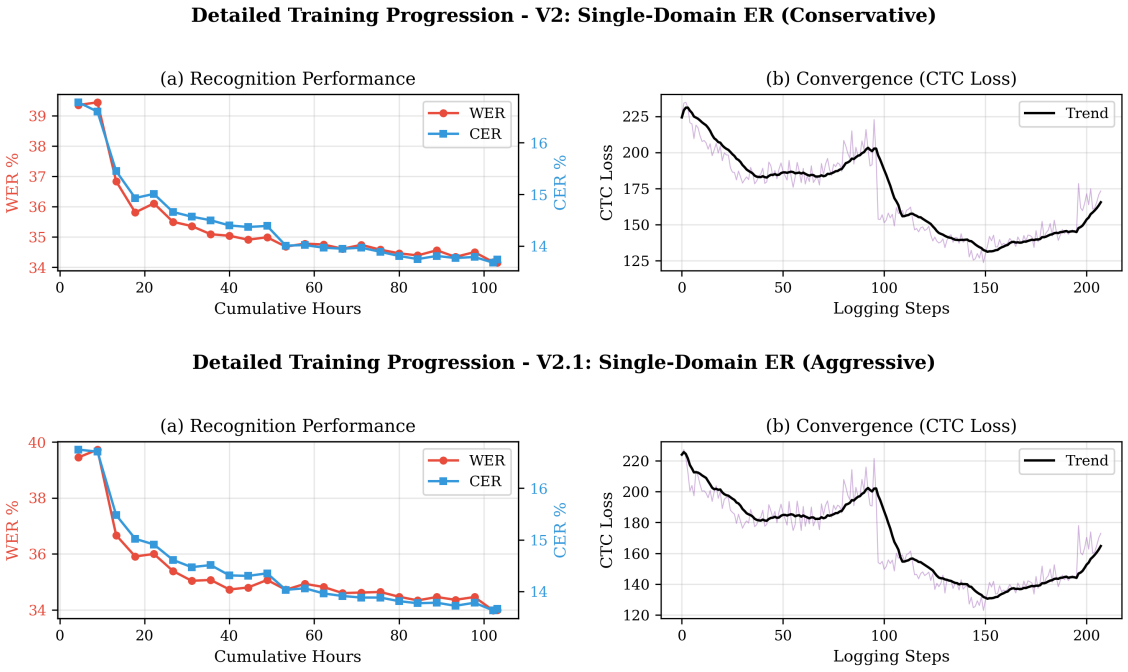

Figure 5: **Single-Domain Replay Dynamics (V2, V2.1):** Incorporating target-domain replay buffers stabilizes the learning trajectory compared to naive fine-tuning.

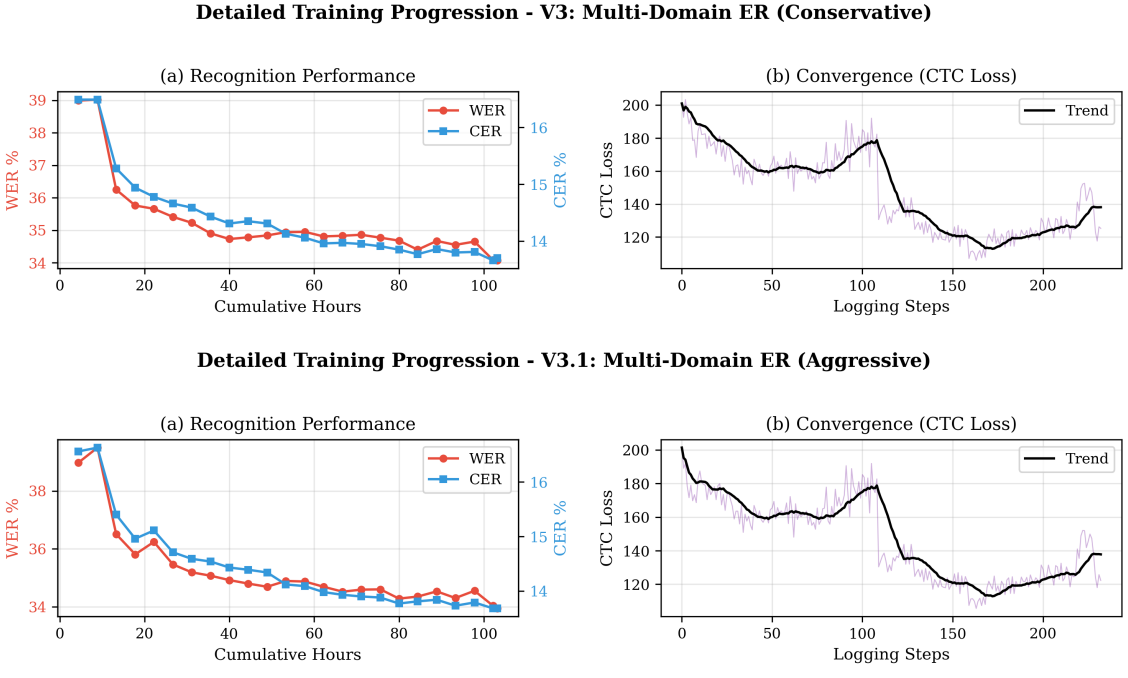

Figure 6: **Multi-Domain Replay Dynamics (V3, V3.1):** Balancing target and general domain samples in the replay buffer offers a tradeoff between plasticity and stability.

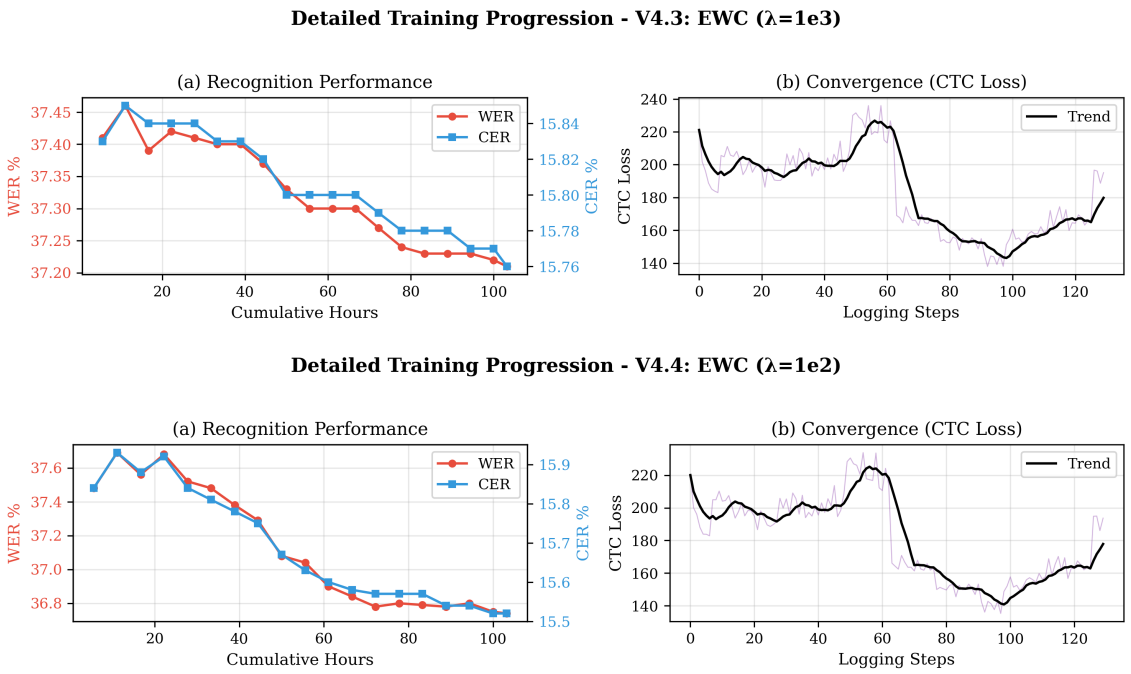

Figure 7: **EWC Regularization (V4.3, V4.4):** High regularization (λ) dampens plasticity, while moderate values allow for adaptation, though standard EWC shows instability.

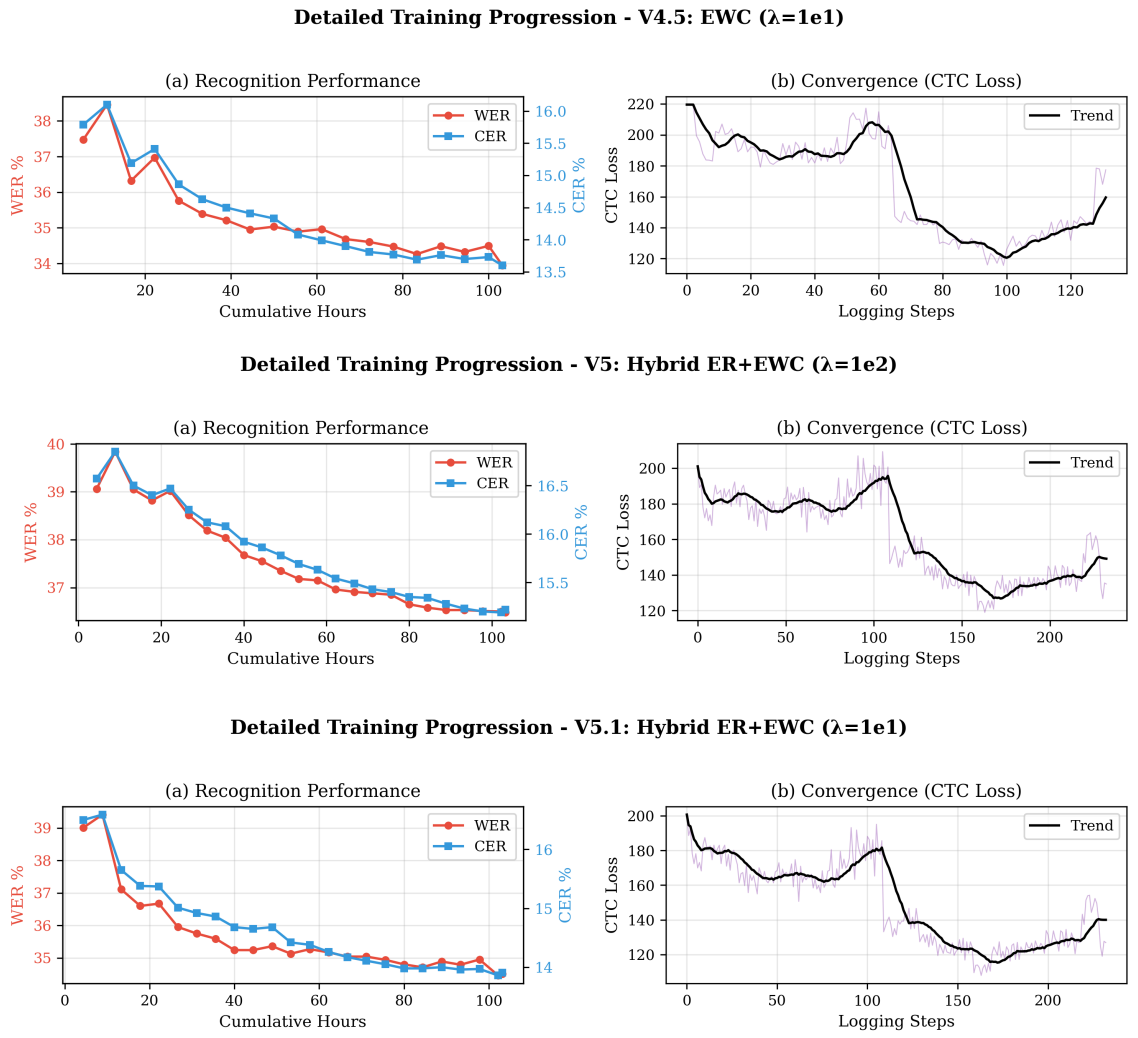

Figure 8: **Optimal Hybrid Strategies (V5, V5.1):** Combining Linearized EWC with Replay (V5.1) achieves the lowest final WER with stable convergence.

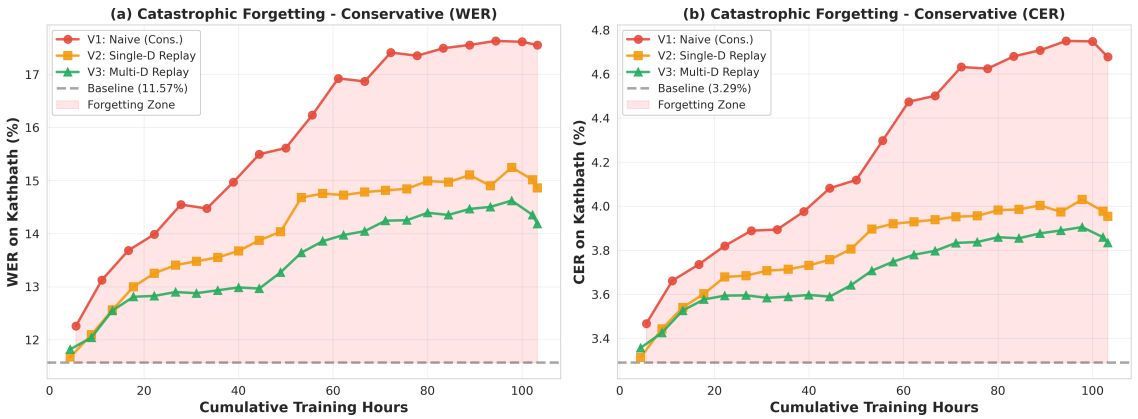

Figure 9: **Forgetting Analysis (Conservative):** Baseline strategies (V1-V3) show varying degrees of knowledge retention on the Kathbath dataset.

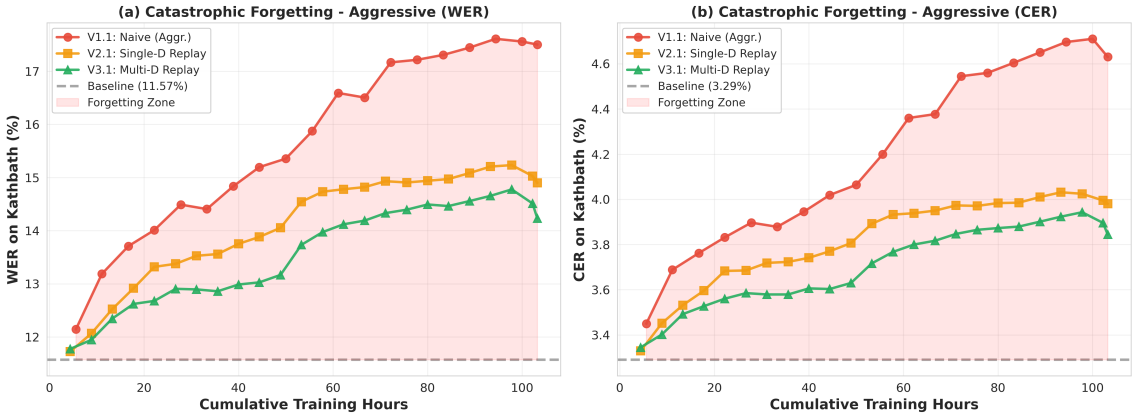

Figure 10: **Forgetting Analysis (Aggressive):** Increased learning rates in V1.1-V3.1 accelerate adaptation but risk higher catastrophic forgetting.

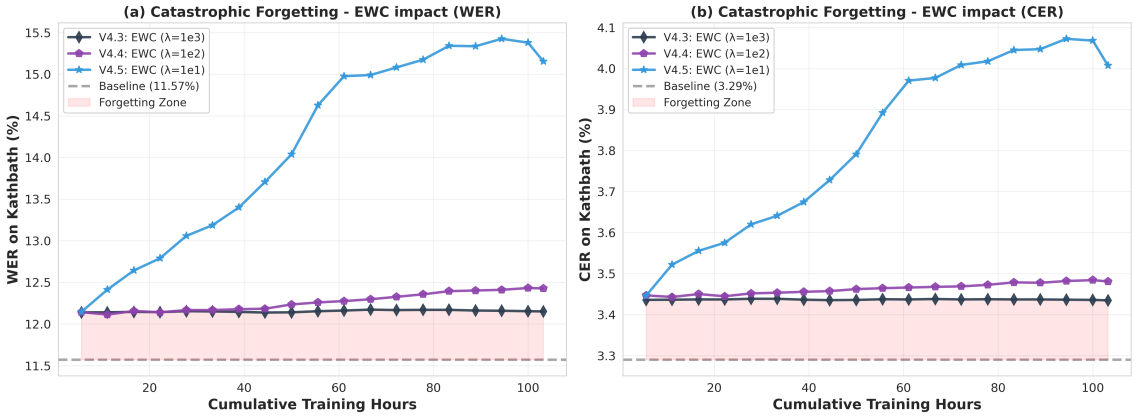

Figure 11: **EWC Forgetting Analysis:** Stronger regularization (λ) effectively reduces forgetting but may hinder adaptation speed.

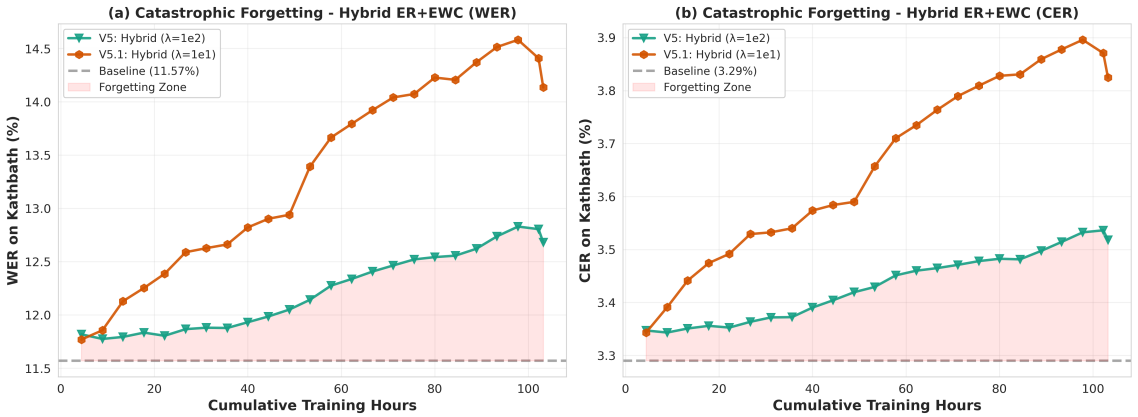

Figure 12: **Hybrid Forgetting Analysis (V5, V5.1):** The hybrid approach demonstrates the best balance, minimizing forgetting while maintaining plasticity.

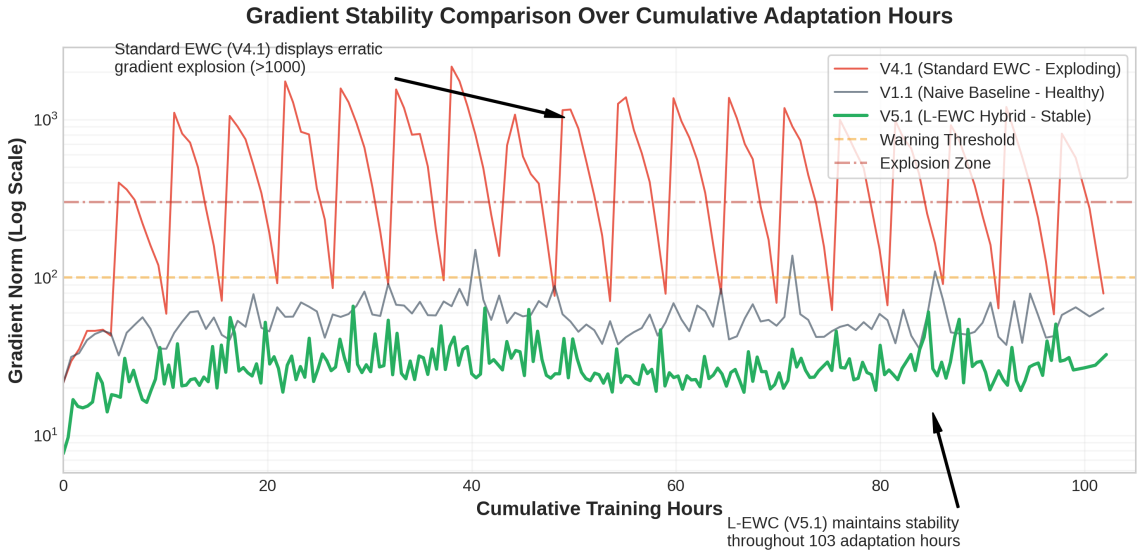

Figure 13: **Gradient Stability Analysis:** Comparing Standard EWC (V4.1), Naive Baseline (V1.1), and our proposed Hybrid L-EWC (V5.1). The log-scale plot demonstrates how the linearized importance estimation in L-EWC prevents the gradient explosion seen in standard quadratic formulations.