X-AuT: Progressive Audio-Encoder Compression for Speech LLMs with Cross-Scale Distillation

Anonymous authors
Paper under double-blind review

Abstract

Reducing audio-encoder depth lowers the inference cost of speech large language models, but removing complete blocks perturbs the embeddings consumed by the decoder and can cause deletion and premature end-of-sequence errors. We introduce X-AuT, a progressive framework that selects layer combinations through short behavioral probes and restores the pruned model through representation alignment, cross-scale distillation, scheduled student-policy supervision, and LoRA finetuning. The language-model backbone remains frozen, while attention LoRA adapters and the tied output embedding adapt during distillation. Training uses the highest-agreement tier from a transcript-consistency pipeline, followed by source reweighting during finetuning. On ten public Chinese–English benchmarks, compressing Qwen3-ASR-0.6B from 18 to 16 audio-encoder layers reduces macro-average error from 5.61% to 5.27%. The 14-layer model reaches 5.75% with 20.7% fewer audio-tower parameters. Under the matched recipe, the 1.7B teacher yields 5.55% mean error, compared with 8.45% for self-distillation, and progressive 18→14 pruning outperforms direct pruning (5.75% vs. 6.73%). These single-run results establish two practical operating points and show that the accuracy effects vary across benchmarks.

1 Introduction

Speech large language models typically combine a deep audio encoder, a bridge that maps acoustic features into the language-model embedding space, and an autoregressive text decoder Shi et al. (2026); Ma et al. (2024); Radford et al. (2022). This architecture is accurate and flexible, but the audio encoder must process every input frame and therefore remains important for first-token latency in streaming, mobile, and in-vehicle systems He et al. (2019); Han et al. (2016). Reducing encoder depth is attractive because it removes complete Transformer blocks and produces a regular, deployment-friendly model.

Starting from a strong pretrained model is also substantially cheaper than training a compact speech encoder and realigning it with a decoder from scratch. We therefore study post-training depth reduction of the 18-layer audio Transformer (AuT) in Qwen3-ASR-0.6B Shi et al. (2026). Removing layers changes the audio embeddings inserted into the decoder and can trigger premature end-of-sequence (EOS) predictions and large deletion errors. Our recipe freezes the pretrained language-model weights while allowing decoder attention LoRA adapters and, during distillation, the tied output embedding to adapt. We use "frozen decoder backbone" in this sense throughout the paper.

Prior ASR compression work has explored decoder distillation, low-rank encoder compression, weight sparsity, supernet training, and encoder-layer removal Gandhi et al. (2023); Kamahori et al. (2025); Mudi et al. (2025); Xu et al. (2024); Kolluri et al. (2026). For an already-pretrained speech LLM, two practical questions remain: which combinations of layers can be removed and recovered under a fixed training budget, and how should recovery address both hidden-state mismatch and errors induced by the student's own decoding history?

These questions are coupled. A layer that appears redundant in isolation may become important once another layer is removed, because subsequent blocks receive a shifted representation and the bridge must preserve the decoder interface learned during pretraining. Static importance scores

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.

therefore cannot fully predict whether a multi-layer candidate can be recovered within a short budget. Selection and recovery must therefore be evaluated together.

X-AuT addresses these questions through progressive pruning and recovery. Before each pruning hop, short behavioral probes compare candidate layer sets under matched initialization, data, and optimization. The selected student then undergoes representation alignment, distillation with scheduled student-policy contexts, and low-rate LoRA finetuning. A Qwen3-ASR-1.7B teacher supplies cross-scale supervision; grouped layer matching and a learned bottleneck projection accommodate the depth and width differences between teacher and student.

Figure 1 compares the Stage 2 16- and 14-layer models with the unpruned baseline across all ten evaluation sets. Each spoke reports accuracy preservation, $100(1-e_{\text{model}})/(1-e_{\text{base}})$, where e is CER or WER expressed as a fraction. This transformation provides a common visual reference across benchmarks; all quantitative comparisons use the macro-average error and the exact CER/WER values in Tables 2 and 3.

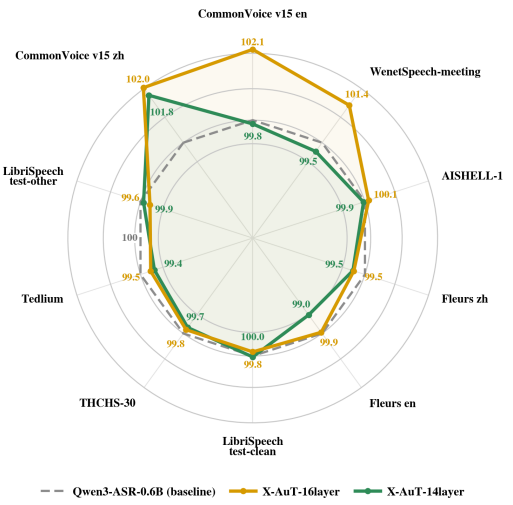


Figure 1: Per-benchmark accuracy preservation of the Stage 2 X-AuT models relative to unpruned Qwen3-ASR-0.6B. The dashed contour marks the baseline (100); higher values indicate better accuracy retention.

The two contours show different accuracy–efficiency tradeoffs. The 16-layer model remains close to or above the baseline across the suite and reaches 5.27% macro error. Further compression to 14 layers yields 5.75% macro error while reducing audio-tower parameters by 20.7% (186.4M→147.8M), with most of the additional loss concentrated on a few English benchmarks.

The benchmark variation also shows why target depth alone is not enough: layer combinations differ in recoverability, and the pruned encoder must be realigned with the decoder under both teacher-forced and student-generated contexts. Figure 2 connects the main steps, from transcript-consistency filtering and behavioral probes to progressive pruning and three-stage recovery. The reported configuration uses class 1 data in all three stages, with source reweighting during Stage 2.

Our main contributions are:

- We introduce a progressive recovery framework that combines behavioral candidate screening, cross-scale hidden-state and logit supervision, scheduled student-policy training, and LoRA finetuning while preserving the pretrained decoder backbone.
- A matched teacher-scale comparison reduces mean error from 8.45% with self-distillation to 5.55% with the 1.7B teacher, demonstrating the practical value of cross-scale supervision in the tested setting.
- Layer-pair probes expose non-additive interactions: the $\{6, 8\}$ pair combines the two strongest single removals but underperforms $\{5, 6\}$ by 0.85 pp after matched recovery.
- The 14-layer model reaches 5.75% macro error, compared with 5.61% for the baseline, while removing 20.7% of audio-tower parameters and reducing measured encoder latency by 21.4% on the in-vehicle accelerator and 11.4% on H800.

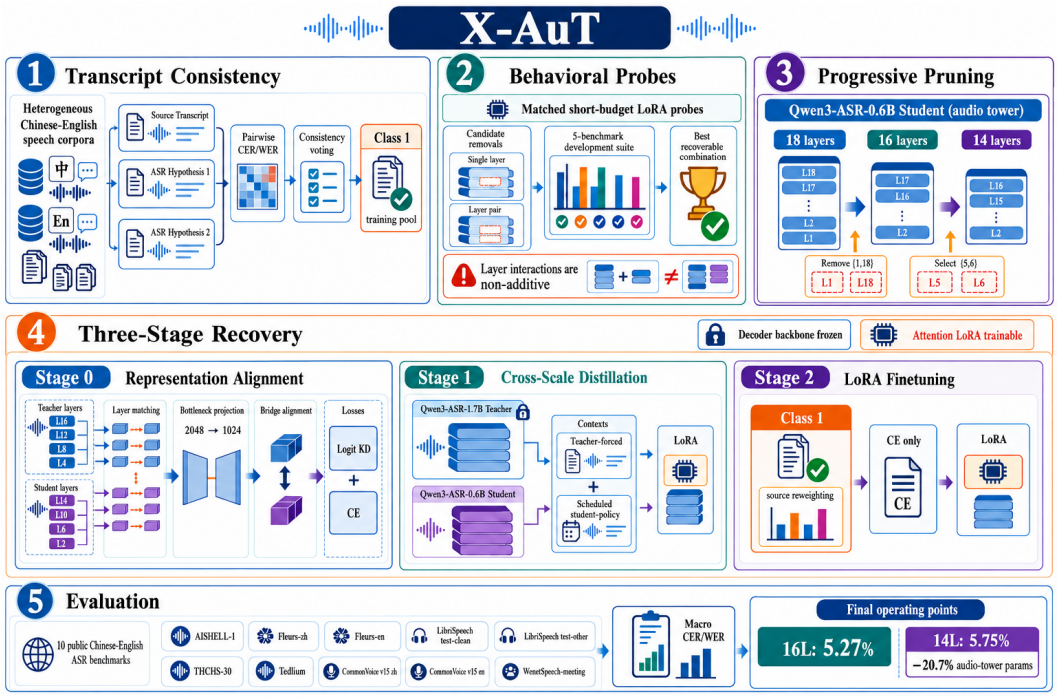


Figure 2: Overview of X-AuT. Transcript-consistency filtering constructs the class 1 training pool, matched short-budget probes select recoverable layer combinations, and two pruning hops reduce the audio tower from 18 to 14 layers. Each pruned student is recovered through representation alignment, cross-scale distillation with teacher-forced and scheduled student-policy contexts, and LoRA finetuning while the decoder backbone remains frozen. Evaluation aggregates CER/WER over ten benchmarks; Secs. 4.2 and 3.4 provide the full configuration.

2 Related Work

2.1 Speech Architectures and Encoder Compression

Modern speech systems increasingly couple an acoustic encoder with a pretrained language model. Qwen3-ASR Shi et al. (2026) inserts bridged audio representations into the Qwen3 decoder input, while SLAM-ASR Ma et al. (2024) examines lightweight connectors between speech encoders and LLMs. Qwen2-Audio Qwen Team (2024) similarly integrates acoustic representations with a general-purpose language model. Whisper Radford et al. (2022) follows an encoder–decoder architecture rather than an LLM-connector design, but remains an important reference for multilingual ASR and subsequent compression work. In each case, the encoder processes the full acoustic sequence, making its depth a direct contributor to computation and latency.

ASR compression has targeted different parts of this architecture. Distil-Whisper Gandhi et al. (2023) primarily reduces the decoder while retaining the Whisper encoder. LiteASR Kamahori et al. (2025) combines low-rank factorization with distillation for encoder matrices, whereas structured sparsity removes weights or attention heads Mudi et al. (2025). LayerDrop Fan et al. (2020) trains networks to tolerate variable depth, and Dynamic Encoder Size Xu et al. (2024) learns a supernet from which multiple encoder depths can be extracted. These approaches either introduce compression during training or reduce computation within existing blocks. Direct removal of complete blocks offers a regular architecture, but it also changes the representations consumed by downstream modules.

Kolluri et al. Kolluri et al. (2026) study Whisper encoder-layer pruning in an LLM-based SLAM-ASR model and recover the pruned network with LoRA Hu et al. (2022). X-AuT instead starts from a pretrained Qwen3-ASR audio tower, removes layers in successive hops, and uses short post-removal probes to compare candidate layer sets under a shared recovery budget. This design treats

recoverability as a property of a layer combination rather than an additive score assigned to individual layers.

2.2 Distillation and Data Selection

Knowledge distillation can transfer teacher output distributions, intermediate representations, or both. Teacher-forced logit distillation evaluates the teacher and student under gold-prefix contexts, which is efficient but differs from inference once the student conditions on its own predictions. On-policy distillation instead supplies supervision on student-generated histories Song and Zheng (2026). Ark-ASR Lin et al. (2026) develops data-efficient on-policy distillation for ASR, and ASKD-Whisper Lee et al. (2026) adjusts the strength of self-distillation during training. X-AuT combines the two regimes: teacher-forced supervision remains the default, while scheduled student-policy batches expose the teacher to the student's decoding context after an initial stabilization period. Rollout filtering returns degenerate batches to the teacher-forced objective.

Training data also shape recovery. Curriculum and data-selection methods rank, order, or filter examples according to estimated learning value or label reliability Rampp et al. (2024). Our preprocessing compares the source transcript with hypotheses from two external ASR systems and assigns a transcript-consistency tier from their pairwise agreement. The reported experiments use the highest-agreement tier throughout recovery and alter source weights during Stage 2 to emphasize target-domain data. The resulting pipeline combines confidence-based filtering with phase-specific resampling without relying on a staged mixture of progressively noisier tiers.

3 Method

X-AuT filters training examples by transcript agreement, identifies recoverable layer combinations through short behavioral probes, and restores each progressively pruned model with a three-stage training schedule (Figure 2).

3.1 Architecture and Objective

An input waveform $\mathbf{x}$ is encoded by an N-layer audio encoder E_θ and bridge B_ϕ into audio embeddings $\mathbf{e} = B_\phi(E_\theta(\mathbf{x}))$. Qwen3-ASR places these embeddings at audio-placeholder positions in the token embedding sequence; the causal language-model decoder D_ψ then predicts transcription tokens through a tied output projection H_ω. This is input-embedding conditioning rather than a separate decoder cross-attention module.

A pruning operation retains an ordered subset $\mathcal{I} \subset \{1, \ldots, N\}$ and forms $E_{\theta,\mathcal{I}}$ from those blocks. We seek a recoverable subset and parameters that minimize aggregate text error rate (TER) under a target depth:

$$\min_{\mathcal{I},\theta',\phi',\omega'} \mathrm{TER}(E_{\theta',\mathcal{I}}, B_{\phi'}, D_\psi, H_{\omega'}) \quad \text{s.t.} \quad |\mathcal{I}| = M < N. \tag{1}$$

The pretrained weights of D_ψ remain frozen. LoRA parameters attached to its q/k/v/o attention projections are trainable, and H_ω—which shares weights with the token embedding in Qwen3-ASR—is trainable during Stages 0–1 and frozen during Stage 2.

Removing encoder blocks changes the conditioning embeddings presented to the decoder. The recovery objective therefore first aligns intermediate and bridge representations, then adapts token distributions under teacher-forced and student-generated prefixes.

3.2 Transcript-Consistency Filtering

The source pool contains heterogeneous supervision. For each reference-bearing utterance, two strong ASR systems produce offline hypotheses. After language-aware normalization, we compute the three pairwise edit rates among the source transcript and the two hypotheses, using CER for Chinese and WER for English. Their maximum, $e_{\max}$, measures the largest disagreement within the transcript–hypothesis triplet. Exact agreement, Mandarin homophone agreement, and a consistency vote assign one of the nine tiers in Table 1. Lower tier numbers indicate stronger transcript agreement rather than ground-truth quality.

Table 1: Nine-tier transcript-consistency hierarchy. Threshold bands use $e_{\max}$, the maximum pairwise edit rate. The tiers measure transcript agreement rather than ground-truth quality.

Tier	Agreement band	Assignment rule
1	Full agreement	Both model hypotheses exactly match the source transcript.
2	Partial exact	At least one of the three text pairs matches exactly (i.e., at least two candidates agree).
3	Homophone	For Mandarin, at least one source–hypothesis pair has zero pinyin CER.
4	$0 < e_{\max} \leq 5\%$	The consistency vote passes and $e_{\max}$ falls in the indicated interval.
5	$5\% < e_{\max} \leq 10\%$	Same voting rule with a small transcript discrepancy.
6	$10\% < e_{\max} \leq 20\%$	Same voting rule with a moderate discrepancy.
7	$20\% < e_{\max} \leq 30\%$	Same voting rule with a relatively large discrepancy.
8	$30\% < e_{\max} \leq 50\%$	Same voting rule with a large discrepancy.
9	Failure / $e_{\max} > 50\%$	The vote fails, $e_{\max} > 50\%$, or an aligned hypothesis is empty.

The reported configuration uses class 1 for both distillation stages. Stage 2 retains the same consistency threshold but reweights sources toward cockpit and other target-domain data, separating confidence-based filtering from phase-specific source sampling.

3.3 Behavior-Driven Progressive Pruning

We prune in two hops, 18→16→14. The first hop removes original layers $\{1, 18\}$. For the second hop, every candidate starts from the same recovered 16-layer checkpoint and is trained with the same 0.3-epoch LoRA warm-up. We first evaluate each remaining layer as a single removal, then evaluate a fixed set of adjacent and non-adjacent layer pairs. Candidate selection uses the lowest aggregate TER on a fixed five-benchmark development suite (Sec. 4). This procedure is more expensive than a static score but directly measures post-removal behavior under the available recovery budget.

The pair probes are necessary because recovery after removing several layers cannot be predicted reliably from the corresponding single-layer scores. The matched comparison selects $\{5, 6\}$ for the 16→14 hop; Section A.9 reports the candidate-level results.

3.4 Three-Stage Recovery

Each hop uses the same three-stage recipe. The student is the pruned Qwen3-ASR-0.6B model; the teacher is Qwen3-ASR-1.7B with a 24-layer audio encoder. Teacher parameters are frozen and discarded at inference.

3.4.1 Stage 0: Representation Alignment

Stage 0 occupies the first 5% of the distillation epoch and combines intermediate-layer, bridge, logit, and transcript losses:

$$\mathcal{L}_{\mathrm{S0}} = \lambda_{\mathrm{layer}}\mathcal{L}_{\mathrm{layer}} + \lambda_{\mathrm{bridge}}\mathcal{L}_{\mathrm{bridge}} + \lambda_{\mathrm{logit}}\mathcal{L}_{\mathrm{logit}} + \lambda_{\mathrm{ce}}\mathcal{L}_{\mathrm{ce}}. \tag{2}$$

Both representation losses sum mean-squared error and cosine distance. Teacher layers are divided uniformly into M ordered groups, and student layer m aligns to the last teacher layer in group m. Because the teacher and student hidden widths are 2048 and 1024, respectively, a learned two-layer MLP with a 256-dimensional bottleneck projects teacher hidden and bridge features into the student space. Logit KD uses temperature-scaled KL divergence under gold prefixes.

The pruned audio encoder and bridge are fully trainable. Decoder base weights remain frozen, while rank-32 LoRA adapters on q/k/v/o projections and the tied output embedding are trained with a separate decoder learning rate.

3.4.2 Stage 1: Distillation with Scheduled Student-Policy Contexts

Stage 1 disables intermediate-layer loss and uses bridge alignment, teacher-forced logit KD, and gold-transcript CE:

$$\mathcal{L}_{\text{off}} = \lambda_{\text{bridge}}\mathcal{L}_{\text{bridge}} + \lambda_{\text{logit}}\mathcal{L}_{\text{logit}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}. \tag{3}$$

After 20% of Stage 1 has elapsed, every fifth optimizer step is scheduled for student-policy supervision. The student greedily generates a prefix; student and teacher are then evaluated on the same generated context. Their distributions are compared over the union of each model's top-k support ($k = 512$). Gold CE remains an anchor on these scheduled steps. The implementation uses no confidence reweighting (`weight_mode=none`).

Rollout safeguards prevent degenerate prefixes from entering the KD loss. Generation enforces `min_new_tokens`$= 3$, uses a duration-aware maximum capped at 256 tokens, and rejects budget-exhausted, over-long, or repetitive rollouts. If at least half of a batch is rejected, that microbatch falls back to the teacher-forced objective. Thus, approximately 20% is a scheduling target; the realized on-policy fraction can be lower after filtering.

3.4.3 Stage 2: LoRA Finetuning

Stage 2 initializes from the best Stage 1 checkpoint and optimizes gold-transcript CE for one epoch. The audio encoder, bridge, and decoder LoRA adapters remain trainable at 5×10^{-6}; the tied lm_head/embedding is frozen. The class-1 data index is reweighted toward target-domain sources. This stage contains no teacher loss:

$$\mathcal{L}_{\text{S2}} = \mathcal{L}_{\text{ce}}(\mathbf{y}^{\text{gold}}, \hat{\mathbf{y}}). \tag{4}$$

4 Experimental Setup

4.1 Models and Parameter Accounting

The student starts from Qwen3-ASR-0.6B Shi et al. (2026), whose audio tower contains 18 Transformer blocks and a convolutional/bridge frontend. Counting tensors in the released checkpoint gives 186.376M audio-tower parameters: 12.758M outside the Transformer stack and 9.645M per block. The 16- and 14-layer students therefore contain 167.085M and 147.794M audio-tower parameters, corresponding to 10.35% and 20.70% reductions. We report these exact counts rather than inferring compression from rounded labels such as 180M/140M. The cross-scale teacher is Qwen3-ASR-1.7B with a 24-layer audio tower and 2048-dimensional hidden states; the student audio hidden width is 1024.

4.2 Training Data and Quality Labels

The source pool combines public and proprietary multilingual ASR corpora, including AISHELL-1/4/5 Bu et al. (2017); Fu et al. (2021); Shi et al. (2023), CommonVoice Ardila et al. (2020), Emilia He et al. (2024), GigaSpeech Chen et al. (2021), KeSpeech Tang et al. (2021), LibriSpeech Panayotov et al. (2015), WenetSpeech Zhang et al. (2022), and cockpit-domain speech. The pool exceeds 280k hours before quality filtering. Audio is capped at 40 seconds.

Manifest construction. Each corpus is converted to a unified JSONL manifest containing an utterance identifier, audio reference, source transcript, language and split tags, duration, sampling rate, and channel count. Text normalization includes Unicode NFKC normalization, width conversion, Traditional-to-Simplified conversion for Mandarin, removal of invisible characters and numeric separators, dash canonicalization, and whitespace normalization. Original and normalized text are retained for traceability. Figure 3 summarizes the complete path from corpus ingestion through transcript agreement and quality ranking.

Reference-bearing utterances are decoded offline by Qwen3-ASR-1.7B and Qwen3.5-Omni Qwen Team (2026). Records without an inference result, a valid audio reference, or nonempty supervision are excluded. The remaining records receive the consistency labels described in Sec. 3.2. The reported distillation runs use the class-1 manifest index; the first-hop log contains approximately

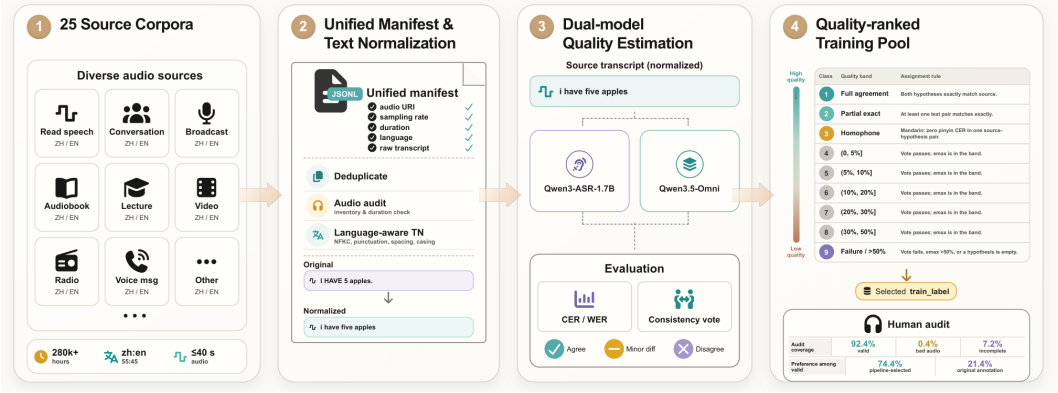


Figure 3: Four-step data pipeline: corpus unification and normalization, dual-system transcription, pairwise CER/WER and consistency voting, and quality-ranked label selection. The rightmost panel reports the WenetSpeech audit summary; audit percentages describe label-selection behavior and are not used as training weights.

299k weighted target records per epoch and the second-hop log approximately 292k. Stage 2 keeps class 1 but changes corpus weights, increasing AISHELL-4/5 and cockpit-query contributions while dropping several weakly matched web-speech sources. These counts describe the realized loader indices rather than the size of the 280k-hour source pool.

4.3 Selection and Evaluation Suites

Development selection suite. Behavior probes and checkpoint selection use five fixed development or validation subsets: AISHELL-1 (Mandarin CER), CommonVoice-en (English WER), Fleurs-en (English WER), WenetSpeech-meeting (Mandarin CER), and a proprietary cockpit-query subset (Mandarin CER). Frequent evaluation is capped at 25 utterances per benchmark, and the macro average of the five normalized error rates determines checkpoint selection. The final results are evaluated separately on the full public benchmark suite described below.

Full public evaluation suite. Final checkpoint results are reported on ten public benchmarks: AISHELL-1 (CER) Bu et al. (2017); Fleurs zh/en (CER/WER) Conneau et al. (2022); LibriSpeech test-clean/test-other (WER) Panayotov et al. (2015); THCHS-30 (CER) Wang and Zhang (2015); Tedlium (WER) Rousseau et al. (2012); CommonVoice v15 zh/en (CER/WER) Ardila et al. (2020); and WenetSpeech-meeting (CER) Zhang et al. (2022). The macro mean weights benchmarks equally, not by utterance count. The proprietary selection subset is excluded from all main-result tables.

4.4 Optimization and Reporting Protocol

The reported 0.6B runs use 32 accelerators, per-device batch size 8, gradient accumulation 2, and global batch size 512. We use AdamW with 2×10^{-5} for the audio tower and 1×10^{-4} for decoder-side LoRA plus the tied output embedding during distillation. Stage 2 uses 5×10^{-6} for all trainable parameters and freezes the tied output embedding. Weight decay is 0.01, gradient clipping is 1.0, and training uses bf16. Stage 0 and Stage 1 occupy 0.05 and 0.95 epoch; Stage 2 runs for one epoch. All configurations use seed 42.

Checkpoints are selected by the lowest observed selection-suite macro TER. Tables report the corresponding single-run checkpoint on the full suite. We did not run repeated seeds or bootstrap utterance-level confidence intervals, so boldface denotes the best observed number in a row and not statistical significance. Full hyperparameters are listed in Appendix A.1.

Efficiency is measured separately on an in-vehicle PPU and an NVIDIA H800 GPU using more than 50 utterances of varying duration. We report descriptive averages from the available benchmark output; run-to-run variability was not retained.

5 RESULTS

5.1 MAIN RESULTS

Table 2: Full-suite error rates for the 16-layer model. S1 is the Stage 1 best checkpoint and S2 is the finetuned checkpoint. Mean is the unweighted macro average. Relative mean-error change is $(\bar{e}/\bar{e}_{\text{base}} - 1) \times 100\%$; negative is better. Bold marks the best observed value, not statistical significance.

Benchmark AuT parameters	Base (18L) 186.4M	Prune-16 S1 167.1M	Prune-16 S2 167.1M
AISHELL-1 (CER)	3.33%	3.30%	**3.21%**
Fleurs-zh (CER)	**2.80%**	3.35%	3.28%
Fleurs-en (WER)	**4.17%**	4.28%	4.23%
LibriSpeech test-clean (WER)	**2.48%**	2.73%	2.65%
THCHS-30 (CER)	**3.87%**	4.10%	4.06%
Tedlium (WER)	**3.35%**	3.92%	3.79%
LibriSpeech test-other (WER)	**5.39%**	5.90%	5.79%
CommonVoice v15 zh (CER)	9.95%	8.56%	**8.12%**
CommonVoice v15 en (WER)	12.35%	10.74%	**10.50%**
WenetSpeech-meeting (CER)	8.36%	8.62%	**7.06%**
Macro mean (%)	5.61	5.55	**5.27**
Relative mean-error change (%)	–	−1.1	−6.1

The 16-layer model lowers macro-average error from 5.61% to 5.27%, an absolute change of -0.34 pp and a 6.1% relative error reduction (Table 2). It improves four benchmarks: AISHELL-1, CommonVoice zh/en, and WenetSpeech-meeting. The largest gains occur on CommonVoice zh (-1.83 pp), CommonVoice en (-1.85 pp), and WenetSpeech-meeting (-1.30 pp), while the largest degradation is 0.44 pp on Tedlium. Stage 2 improves all ten entries relative to the Stage 1 checkpoint.

Table 3: Full-suite error rates for the 14-layer model. The 147.8M audio tower has 20.7% fewer parameters than the 186.4M baseline. Formatting and reporting conventions follow Table 2.

Benchmark AuT parameters	Base (18L) 186.4M	Prune-14 S1 147.8M	Prune-14 S2 147.8M
AISHELL-1 (CER)	**3.33%**	3.52%	3.39%
Fleurs-zh (CER)	**2.80%**	3.49%	3.32%
Fleurs-en (WER)	**4.17%**	5.24%	5.10%
LibriSpeech test-clean (WER)	2.48%	3.09%	**2.45%**
THCHS-30 (CER)	**3.87%**	4.23%	4.17%
Tedlium (WER)	**3.35%**	4.07%	3.95%
LibriSpeech test-other (WER)	**5.39%**	7.00%	5.52%
CommonVoice v15 zh (CER)	9.95%	9.54%	**8.36%**
CommonVoice v15 en (WER)	**12.35%**	13.94%	12.49%
WenetSpeech-meeting (CER)	**8.36%**	10.71%	8.78%
Macro mean (%)	**5.61**	6.48	5.75
Relative mean-error change (%)	–	+15.5	+2.5

The 14-layer model contains 147.794M audio-tower parameters, 20.70% fewer than the 186.376M baseline. Its macro error is 5.75%, a 0.14-pp increase over the baseline (Table 3). CommonVoice zh improves by 1.59 pp and LibriSpeech test-clean by 0.03 pp, whereas Fleurs-en has the largest degradation at 0.93 pp. Relative to the 16-layer model, seven benchmark changes remain within 0.3 pp; CommonVoice en ($+1.99$ pp) and WenetSpeech-meeting ($+1.72$ pp) account for most of the macro-average gap. Figure 1 summarizes this benchmark-level variation, and the tables report the corresponding CER/WER values.

The two operating points expose a clear tradeoff. The 16-layer model improves the observed macro average, while the 14-layer model provides a larger structural reduction at a small average cost. Section 5.2 discusses the uncertainty associated with these single-run comparisons.

5.2 Conclusion

X-AuT compresses the Qwen3-ASR-0.6B audio tower from 18 to 14 layers through behavior-driven progressive pruning and cross-scale recovery. The 16-layer model improves macro-average error from 5.61% to 5.27%, and the 14-layer model reaches 5.75% with 20.7% fewer audio-tower parameters.

The controls identify what makes recovery work. Representation alignment and retained capacity are both necessary: Stage 0/1 alignment lets the 16-layer model surpass the baseline, but the 14-layer gap shows that alignment cannot fully replace the capacity removed by deeper pruning. The recovery signal matters, with the 1.7B teacher reaching 5.55% mean error versus 8.45% for self-distillation under the matched recipe. Layer choice is non-additive—the strongest single removals do not form the best pair—and progressive 18→16→14 pruning (5.75%) outperforms direct 18→14 pruning (6.73%) under the same nominal budget.

Because these results come from single runs with seed 42 on one model family, small differences are descriptive rather than statistically significant, and broader models with repeated trials are needed to determine how well this tradeoff generalizes.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code development, to polish the writing, and to produce some of the figures. We did not use generative AI tools to develop the research ideas or methodology, to design or interpret the experiments, or to outline the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human-subject studies. All evaluation benchmarks are publicly available, and the training data mixture consists of public and properly licensed proprietary corpora that are not redistributed. The released model weights are intended for research and evaluation only under a CC BY-NC 4.0 license and are not intended for applications that could cause harm to people. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Code, model weights, and example manifests will be released publicly upon acceptance, including a standalone inference script and a minimal LoRA finetuning recipe. Model and parameter accounting is given in Section 4, the full hyperparameter configuration in Appendix A.1, and the teacher-scale, layer-interaction, pruning-schedule, on-policy, training-dynamics, and efficiency details in Appendices A.3–A.9. The behavior-probe pipeline, the Stage 0/1 cross-scale distillation code, and the proprietary 280k-hour training data are not redistributed, but the transcript-consistency filtering and the three-stage training protocol are described in Sections 3 and 4.

REFERENCES

Rosana Ardila, Megan Branson, Kelly Lee, Michael Kohler, Rebekah Schumann, Laure Sterckx, Juan Diez Bayron, Prasanga Karunanayake, Ramón Sanabria, Andre Baas, et al. Common voice: A massively-multilingual speech corpus. In *Proc. LREC*, 2020.

Huihui Bu, Juntao Du, Xingyu Na, Yanyan Ji, and Fang Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Proc. O-COCOSDA*, 2017.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Lu, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Wang, Shuai Wu, Yong Yang, Yaqing Wang, Zhuo Yu, and Zeyu Wang. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech*, 2021.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, Noam Remez, Viktor Glavcev, Sanjay Gopala, Jian Ni, Yu-Hsiang Wu, Po-Ning Hsu, Jason Liu, Amir Sahebi, Paul-Ambroise Duquenne, Mia Chen, Vinh Chau, et al. Fleurs: Few-shot learning evaluation of universal representations of speech. In *Proc. IEEE SLT*, 2022.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations (ICLR)*, 2020.

Yihui Fu, Luyao Xu, Yukai Zhai, Yuxiang Wang, Kun Liang, Yongqiang Li, Shiliang Zhang, Yonghong Yan, and Xie Chen. Aishell-4: An open source dataset for speech recognition in multi-party conference scenario, 2021.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. In *arXiv preprint arXiv:2311.00430*, 2023.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations (ICLR)*, 2016.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuan Li, Yicheng Gu, Peiyu Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation, 2024.

Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pandey, Khe Chai Sim, Thomas Bagby, Shuo-Yin Chang, Kanishka Rao, and Alexander Gruenstein. Streaming end-to-end speech recognition for mobile devices. In *Proc. IEEE ICASSP*, pages 6381–6385, 2019.

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 (ICLR)*, 2022.

Keisuke Kamahori, Jungo Kasai, Noriyuki Kojima, and Baris Kasikci. Liteasr: Efficient automatic speech recognition with low-rank approximation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Ganesh Pavan Kartikeya Bharadwaj Kolluri, Michael Kampouridis, and Ravi Shekhar. On the role of encoder depth: Pruning whisper and lora fine-tuning in slam-asr. In *Proceedings of the SPEAKABLE Workshop, LREC*, 2026.

Junseok Lee, Nahun Kim, Sangyong Lee, and Chang-Jae Chun. Askd-whisper: Adaptive self-knowledge distillation for efficient and low-latency automatic speech recognition, 2026.

Yu Lin, Yiming Wang, Runyuan Cai, and Xiaodong Zeng. Data-efficient on-policy distillation for automatic speech recognition, 2026.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. An embarrassingly simple approach for llm with strong asr capacity, 2024.

Prasenjit K Mudi, Anshi Sachan, Dahlia Devapriya, and Sheetal Kalyani. Structured sparsity and weight-adaptive pruning for memory and compute efficient whisper models, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *Proc. IEEE ICASSP*, 2015.

Qwen Team. Qwen2-audio technical report, 2024.

Qwen Team. Qwen3.5-omni technical report, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Simon Rampp, Manuel Milling, Andreas Triantafyllopoulos, and Björn W. Schuller. Does the definition of difficulty matter? scoring functions and their role for curriculum learning, 2024.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. Ted-lium: An automatic speech recognition dedicated corpus. In *Proc. LREC*, 2012.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report, 2026.

Yuhao Shi, Luyao Xu, Shiliang Zhang, and Yonghong Yan. Aishell-5: Multi-domain mandarin speech recognition corpus with labeled data of 520 hours. In *Proc. IEEE ICASSP*, 2023.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models, 2026.

Zhiyuan Tang, Dong Wang, Xiaoxue Xu, Hongbin Zheng, Yexin Lei, Jiawen Li, Shuai Zhang, Yifan Zhu, Jingjing Meng, Haoran Li, Xuyang Xu, Yuhang Zheng, and Shihao Li. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Proc. IEEE ASRU*, 2021.

Dong Wang and Xuewei Zhang. Thchs-30: A free chinese speech corpus, 2015.

Jingjing Xu, Eugen Beck, Zijian Yang, and Ralf Schlüter. Dynamic encoder size based on data-driven layer-wise pruning for speech recognition. In *Proc. Interspeech*, pages 4563–4567, 2024.

Bin Zhang, Hui Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xie Chen, Chuang Zeng, Di Wu, and Zhendong Peng. Wenetspeech: A 10,000+ hours multi-domain mandarin corpus for asr. In *Proc. IEEE ICASSP*, 2022.

APPENDIX

A.1 FULL HYPERPARAMETER CONFIGURATION

Table 4: Main 16-layer hyperparameters. The 14-layer run uses the same recipe and initializes from the recovered 16-layer checkpoint.

Parameter	Stage 0	Stage 1	Stage 2
Fraction / epochs	0.05 epoch	0.95 epoch	1 epoch
LR (audio tower)	2×10^{-5}	2×10^{-5}	5×10^{-6}
LR (LoRA / tied head)	$10^{-4}/10^{-4}$	$10^{-4}/10^{-4}$	5×10^{-6}/frozen
Optimizer / weight decay	AdamW / 0.01	AdamW / 0.01	AdamW / 0.01
Per-device batch / accumulation	8 / 2	8 / 2	8 / 2
Global batch size	512	512	512
LoRA r/α/dropout	32 / 64 / 0.05	32 / 64 / 0.05	32 / 64 / 0.05
Temperature	1.5	1.0	–
λ_{layer}	1.0	0.0	–
λ_{bridge}	1.0	0.5	–
λ_{logit}	0.2	0.1	–
λ_{ce}	0.5	1.0	1.0
On-policy start / fraction	–	0.2 / 0.2	–
Union top-k / weight mode	–	512 / none	–
Minimum / maximum new tokens	–	3 / 256	–
Reject-batch threshold	–	0.5	–

A.2 PREMATURE-EOS SAFEGUARDS

Immediately after audio-encoder pruning, the model can emit EOS after only a few tokens, producing large deletion errors. Qwen3-ASR does not use a separate decoder cross-attention module: bridge outputs replace audio-placeholder embeddings in the causal decoder input. We hypothesize that pruning shifts these conditioning embeddings, weakening acoustic evidence for the pretrained decoder. We do not directly measure embedding mean/variance or EOS causality, so this account motivates the safeguards rather than constituting a mechanistic proof.

The implementation uses three safeguards. First, during main Stage 0/1 distillation, the tied lm_head/token embedding is trainable at the decoder-side learning rate (10^{-4}; 5×10^{-5} in the EOS ablation). Decoder Transformer weights remain frozen apart from LoRA. Second, on-policy generation enforces a fixed `min_new_tokens=` 3; its maximum is duration-aware, $\min(256, \max(128, \lceil 12d + 32 \rceil))$, where d is the longest audio duration in the batch. Third, rollout filters reject budget-exhausted, implausibly long, and repetitive sequences. If the rejected fraction reaches 0.5, the scheduled on-policy microbatch falls back to teacher-forced distillation.

Table 5: Premature-EOS safeguards on prune-16 in a 2×2 ablation. All configurations share the same Stage 0 checkpoint and Stage 1 recipe. Empty events count 100-step windows with at least one empty rollout; rejections sum budget, length, and repetition filters. TER is the best development-suite observation.

Configuration	lm_head	min_new	Empty events	TER	Rejections
A: neither	frozen	0	5	6.86	20
B: lm_head	trainable	0	3	6.83	11
C: gating	frozen	3	**0**	6.97	2113
D: both	trainable	3	**0**	**6.75**	2162

The no-mitigation condition has five windows with a nonzero empty ratio; the maximum is 0.3% and the mean is 0.018%. Training the tied output embedding reduces this to three events, while `min_new_tokens=` 3 removes empty events in both gating conditions. Gating also converts some premature terminations into longer or degenerate rollouts: configurations C and D trigger 2113 and 2162 filter rejections, compared with 20 and 11 for A and B.

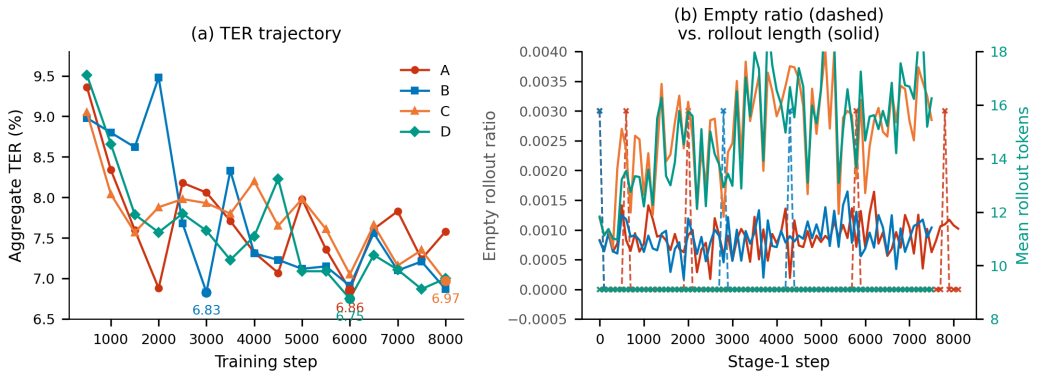


Figure 4: Premature-EOS ablation on prune-16. A disables both mitigations, B trains the tied head, C applies minimum-token gating, and D combines them. (a) Development TER. (b) Empty-rollout ratio (dashed) and mean rollout length (solid). All runs start from the same Stage 0 checkpoint.

The best observed TER is 6.75% for the combined configuration D. The tied-head-only configuration B reaches 6.83% and shows only 0.04-pp best-to-last drift; A drifts from 6.86% to 7.58%. Mean rollout length is 15.06 tokens for D and 11.05 for A. These single runs support using the tied head and gating together in the main recipe, but the 0.08-pp B–D difference is too small to interpret without uncertainty estimates.

A.3 Teacher-Scale Comparison

The self-teacher control uses the unpruned 18-layer Qwen3-ASR-0.6B model. The cross-scale condition uses Qwen3-ASR-1.7B and the required 2048→1024 bottleneck projections. Other Stage 0/1 settings are matched, and both models are evaluated at their best development checkpoint before Stage 2.

Table 6: Teacher-scale comparison for the 16-layer Stage 1 student. Schedules, data, and optimization are matched; cross-scale training additionally uses the required hidden-width projections. Values are single-run best-checkpoint results.

Benchmark	Base (18L)	Self-teacher	Cross-scale
AISHELL-1 (CER)	3.33%	4.29%	**3.30%**
Fleurs-zh (CER)	**2.80%**	4.17%	3.35%
Fleurs-en (WER)	**4.17%**	6.15%	4.28%
LibriSpeech test-clean (WER)	**2.48%**	6.00%	2.73%
THCHS-30 (CER)	**3.87%**	5.23%	4.10%
Tedlium (WER)	**3.35%**	11.28%	3.92%
LibriSpeech test-other (WER)	**5.39%**	9.89%	5.90%
CommonVoice v15 zh (CER)	9.95%	11.71%	**8.56%**
CommonVoice v15 en (WER)	12.35%	14.89%	**10.74%**
WenetSpeech-meeting (CER)	**8.36%**	10.85%	8.62%
Macro mean (%)	5.61	8.45	**5.55**
Relative mean-error change (%)	–	+50.6	−1.1

The cross-scale checkpoint is better on every public benchmark and lowers mean error from 8.45% to 5.55%. The largest gaps occur on Tedlium (11.28 vs. 3.92), LibriSpeech test-clean (6.00 vs. 2.73), and LibriSpeech test-other (9.89 vs. 5.90). Relative to the original baseline, the cross-scale checkpoint improves macro error by 1.1%, while the self-teacher checkpoint increases it by 50.6%. Relative to the original 0.6B baseline, the cross-scale model also improves CommonVoice zh/en and AISHELL-1, whereas the self-teacher is worse on all ten benchmarks. This pattern suggests that the larger teacher transfers useful acoustic behavior rather than merely restoring the pruned student to

its starting point, and is consistent with useful cross-scale transfer, especially on English and higher-error conditions. Because the comparison has one seed and includes projection modules only when dimensions differ, we do not treat it as proof that the teacher creates capabilities absent from every unpruned student.

A.4 Layer-Interaction Details

Table 7: Pair probes for 16→14 pruning after the same 0.3-epoch warm-up. TER is the macro mean over five development subsets, including proprietary SC. Δ is relative to $\{5, 6\}$. Bold marks the best observed candidate.

Candidate	Type	TER	Δ	AISHELL	CV-en	Fleurs-en	Wenet / SC
$\{5, 6\}$	Adj.	**6.93**	–	0.63	16.97	6.85	7.30 / 2.91
$\{6, 7\}$	Adj.	7.37	+0.44	0.63	16.06	6.48	8.47 / 5.23
$\{8, 9\}$	Adj.	7.75	+0.82	0.63	23.39	5.37	6.42 / 2.91
$\{6, 8\}$	Non-adj.	7.78	+0.85	0.63	20.64	6.11	8.03 / 3.49
$\{3, 6\}$	Non-adj.	8.12	+1.19	0.63	17.89	6.48	8.03 / 7.56
$\{14, 15\}$	Adj.	9.93	+3.00	0.95	23.39	11.48	8.61 / 5.23

The single-layer results used to form the interaction comparison are L8=5.97%, L6=6.11%, and L5=6.15%. Thus, the non-adjacent pair $\{6, 8\}$ contains the two best constituent removals (mean 6.04%) but reaches 7.78% as a pair. The selected adjacent pair $\{5, 6\}$ has a slightly worse constituent mean (6.13%) but reaches 6.93%. The descriptive interaction penalties are therefore 1.74 and 0.80 pp, respectively.

All three adjacent candidates in Table 7 rank above the two non-adjacent candidates, while $\{14, 15\}$ confirms that adjacency alone is not sufficient. One possible account is that removing a contiguous sub-block creates one residual-stream discontinuity whereas dispersed removal creates two. This explanation is untested; causal activation analysis and a larger factorial candidate set are needed.

A.5 Progressive versus Direct Pruning

We compare the reported progressive path with a direct 18→14 run that drops the same original layers $\{1, 18, 5, 6\}$ and uses the same nominal recovery and data budget.

Table 8: Direct 18→14 pruning versus progressive 18→16→14 pruning. Both remove $\{1, 18, 5, 6\}$ and use the same recovery/data budget. Single-run best-checkpoint results; bold marks the best observed value.

Benchmark	Base	Direct 18→14	Progressive
AISHELL-1 (CER)	**3.33%**	3.81%	3.39%
Fleurs-zh (CER)	**2.80%**	3.76%	3.32%
Fleurs-en (WER)	**4.17%**	5.94%	5.10%
LibriSpeech test-clean (WER)	2.48%	3.42%	**2.45%**
THCHS-30 (CER)	**3.87%**	4.62%	4.17%
Tedlium (WER)	**3.35%**	5.10%	3.95%
LibriSpeech test-other (WER)	**5.39%**	6.31%	5.52%
CommonVoice v15 zh (CER)	9.95%	9.79%	**8.36%**
CommonVoice v15 en (WER)	**12.35%**	15.22%	12.49%
WenetSpeech-meeting (CER)	**8.36%**	9.29%	8.78%
Macro mean (%)	**5.61**	6.73	5.75

Progressive pruning reaches 5.75% mean error, compared with 6.73% for direct pruning, and is better on all ten benchmarks. The largest direct-minus-progressive gaps are CommonVoice en (+2.73 pp), CommonVoice zh (+1.43 pp), and Tedlium (+1.15 pp). This shows that the progressive path is preferable under the tested budget. It does not establish that no alternative schedule or larger budget could improve direct pruning.

A.6 On-Policy Strategy Comparison

Figure 5 compares three Stage 1 runs on prune-16: a plain off-policy baseline, an off-policy control matched to the extra student forward and loss settings used on scheduled steps, and hybrid on-policy training. The dashboard provides the visual trajectory; numerical comparisons use the archived checkpoint logs.

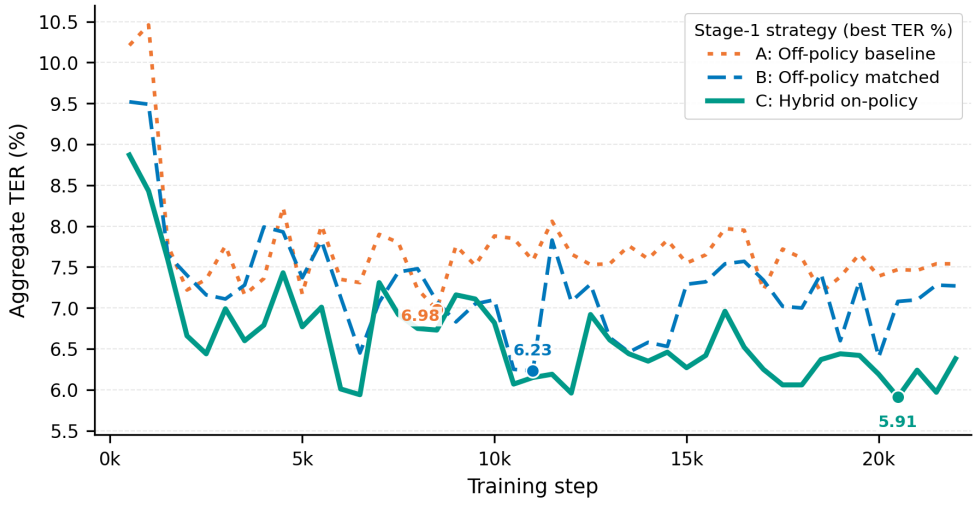


Figure 5: Stage 1 strategy dashboard for plain off-policy, matched off-policy, and hybrid on-policy training; lower TER is better. Exact best and final values in the accompanying text are taken from the archived checkpoint logs.

The best observed TER values are 6.98% for plain off-policy, 6.23% for matched off-policy, and 6.41% for hybrid on-policy. Final observations are 7.43%, 6.51%, and 6.99%, respectively. Hybrid training therefore improves over the plain baseline but does not outperform the matched off-policy control in this experiment. We retain the hybrid recipe in the main run because it directly supervises student-generated contexts and works well in the complete pipeline, while recognizing that this ablation does not establish accuracy superiority. More seeds and a sweep over rollout fraction are needed.

A.7 Inference Efficiency

Table 9: Descriptive inference measurements averaged over more than 50 utterances of varying lengths. No run-to-run variance was retained. Peak device memory includes the full model, so encoder-only pruning changes it modestly.

Model	Encoder (ms)	End-to-end (ms)	RTF	Peak memory (MB)
		In-vehicle PPU		
Baseline 18L	14	486	0.0786	1500.3
Prune-14	11	463	0.0748	1434.4
Relative change	−21.4%	−4.7%	−4.8%	−4.4%
		NVIDIA H800		
Baseline 18L	88	1672	0.256	1500.3
Prune-14	78	1629	0.245	1458.2
Relative change	−11.4%	−2.6%	−4.3%	−2.8%

The 14-layer model reduces encoder time by 21.4% on the in-vehicle PPU and 11.4% on H800. End-to-end reductions are 4.7% and 2.6%, because autoregressive decoding dominates total time. These measurements support encoder pruning as a localized efficiency improvement, especially when encoder and decoder are pipelined, but not as a large end-to-end speedup by itself.

A.8 Training Trajectories

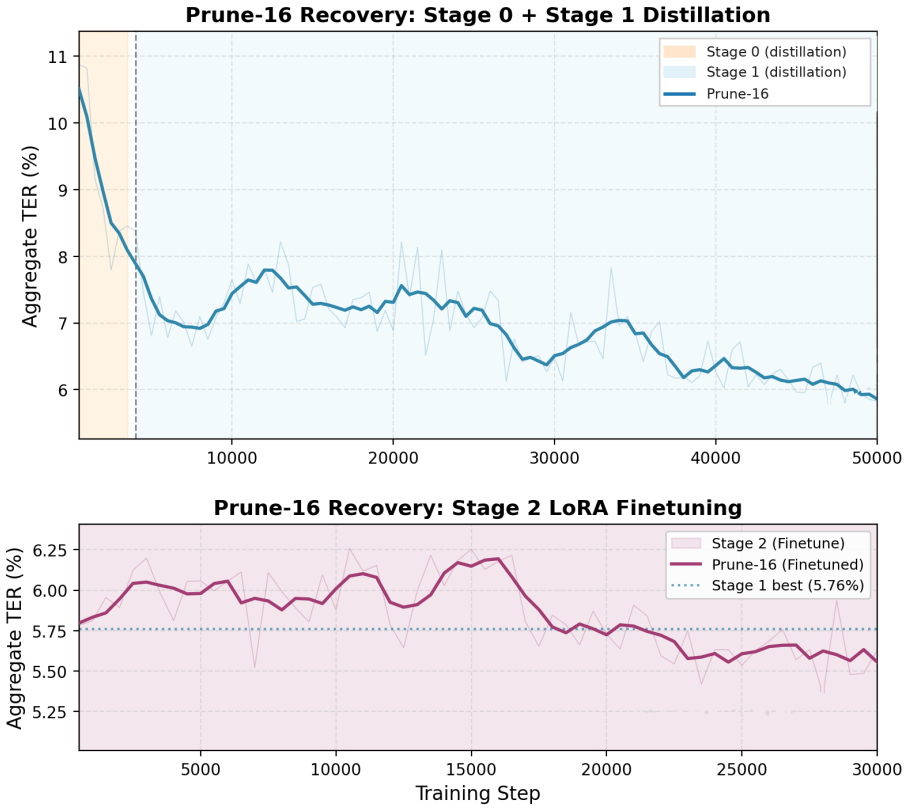


Figure 6: Selection-suite TER during 16-layer recovery. (a) Stage 0/1 distillation, with the transition near step 4000. (b) Stage 2 finetuning from the best Stage 1 checkpoint; the dashed line marks 5.76%. The displayed trends summarize checkpoint evaluations and are not uncertainty estimates.

Stage 0 TER falls from 10.88% at step 500 to 7.80% at step 2500, followed by 8.38% at the first Stage 1 evaluation near step 4000 (Figure 6). Stage 1 reaches its minimum of 5.76% at step 47,000. Starting from that checkpoint, Stage 2 reduces TER by another 0.40 pp and reaches 5.36% at step 28,000.

The trajectory shows rapid early recovery followed by slower optimization in Stage 1 and a further gain from Stage 2. Because the stage transition changes both the loss and optimizer state, the curve describes the recovery process rather than the contribution of an individual loss term.

A.9 BEHAVIOR-DRIVEN LAYER SELECTION

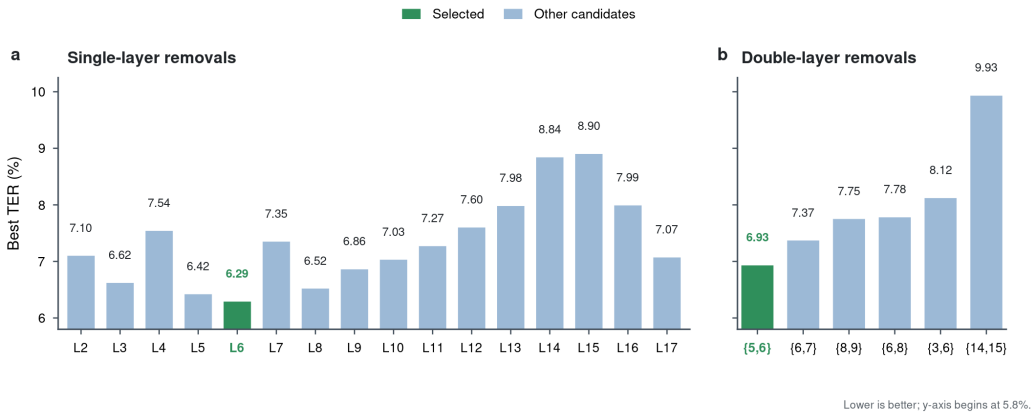


Figure 7: Layer-screening results for single- and double-layer removals after the same 0.3-epoch warm-up. Lower TER is better; green marks the candidate selected for progressive pruning.

The single-layer sweep spans 6.29–8.90% TER. L6 is the strongest single removal at 6.29%, followed by L5 at 6.42% and L8 at 6.52%. The selected $\{5, 6\}$ pair reaches 6.93%, whereas $\{6, 8\}$ reaches 7.78%. Defining the interaction penalty as pair TER minus the mean TER of its constituent removals gives 0.58 pp for $\{5, 6\}$ and 1.38 pp for $\{6, 8\}$.

The adjacent candidates $\{5, 6\}$, $\{6, 7\}$, and $\{8, 9\}$ rank ahead of the tested non-adjacent candidates $\{6, 8\}$ and $\{3, 6\}$, while the adjacent $\{14, 15\}$ pair performs worst overall. Thus, adjacency alone is not a reliable selection rule, and pair recoverability cannot be inferred from single-layer scores in isolation.