MEGA-ASR: TOWARDS *In-the-wild*^{2} SPEECH RECOGNITION VIA SCALING UP REAL-WORLD ACOUSTIC SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic speech recognition degrades nonlinearly in the wild: as acoustic corruption compounds, local confusions give way to omissions, repetitions, empty outputs, and semantically plausible but unsupported transcriptions. We present **MEGA-ASR**, a learnability-aligned framework that treats this transition as a linked problem of *coverage*, *grounding*, and *optimization*. First, **VOICES-IN-THE-WILD-2M** composes seven atomic acoustic effects into 54 controllable scenarios, providing hard but learnable supervision for compound environments. Second, **Acoustic-to-Semantic Progressive Supervised Fine-Tuning (A2S-SFT)** establishes acoustic grounding before adapting semantic recovery: it first trains the acoustic interface, then the language-side decoder, and finally aligns both end to end. Third, **Dual-Granularity WER-Gated Policy Optimization (DG-WGPO)** starts from this grounded policy and changes its preference with the error regime, favoring token-level refinement when speech remains recoverable and structural reconstruction when it does not. The three stages therefore form a capability chain rather than independent training add-ons. Compared to the baseline, Qwen3-ASR-1.7B, MEGA-ASR achieves 45.69% versus 54.01% WER on VOiCES R4-B-F and 21.10% versus 29.34% WER on NOIZEUS Station-0dB. On the mixed real/simulated compositional settings, it reduces WER by 65.8%/52.5% relative to Gemini-3-Flash and by 69.4%/69.1% relative to Whisper Large-v3.

1 INTRODUCTION

Automatic speech recognition (ASR) now reaches word error rates (WERs) near 1% on canonical clean-speech benchmarks (Shi et al., 2026; Xu et al., 2026; Gao et al., 2022; Panayotov et al., 2015). This success, however, is brittle outside the laboratory. Distance, reverberation, occlusion, background noise, recording coloration, and transmission artifacts do not occur one at a time; they co-occur and progressively erase the acoustic evidence on which transcription depends. The resulting gap remains large even for a strong modern baseline: in our evaluation, Qwen3-ASR-1.7B (Shi et al., 2026) reaches 54.01% WER on VOiCES R4-B-F and 29.34% WER on NOIZEUS Station-0dB.

Crucially, severe degradation does not merely increase the *number* of recognition errors; it changes their *form*. When parts of the signal remain accessible, the model tends to make local substitutions. Once the evidence becomes sparse or contradictory, it can omit spans, repeat phrases, collapse to an empty output, or generate fluent text that is no longer grounded in the audio. Robust ASR must therefore handle an error-regime transition: the corrective behavior that is useful for a nearly correct hypothesis is not the behavior needed to recover a structurally corrupted one.

This transition exposes three interlocking challenges in prevailing robust-ASR pipelines (Yan et al., 2025; Han et al., 2017). First, there is a *coverage gap*: training data and augmentations often isolate a small number of factors, whereas deployment failures arise from compound conditions spanning a broad difficulty range (Table 1). Second, there is a *grounding gap*: even when difficult examples are available, a model can rely on linguistic priors before it learns to extract dependable acoustic evidence, producing text that is plausible but not faithful. Third, there is an *optimization gap*: conventional WER-based training treats local correction and sentence-level recovery as variants of the same problem, even though they demand different preferences.

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.

Table 1: Acoustic-phenomenon coverage of robust-ASR datasets. Checkmarks indicate factors represented in the recordings or simulations, not independent factor control. Real recordings include controlled physical replay; VOICES-IN-THE-WILD-2M jointly controls all seven factors through simulation.

	Source type		Acoustic phenomena							
Dataset	**Real**	**Sim.**	**Noise**	**Far-field**	**Obstruction**	**Echo & Reverb.**	**Recording Coloration**	**Electronic Distortion**	**Transmission Dropout**	**Scale**
NOIZEUS (Hu and Loizou, 2007)	✗	✓	✓	✗	✗	✗	✗	✗	✗	1K
TED-LIUM (Rousseau et al., 2012)	✓	✗	✗	✓	✗	✗	✗	✗	✗	59K
CHiME-4 (Watanabe et al., 2016)	✓	✓	✓	✗	✗	✗	✗	✗	✗	15K
VOiCES (Richey et al., 2018)	✓	✗	✓	✓	✓	✓	✓	✗	✗	1M
BERSt (Tuttösí et al., 2026)	✓	✗	✗	✓	✓	✗	✗	✗	✗	4.5K
DAPS (Mysore, 2014)	✓	✗	✗	✗	✗	✗	✗	✓	✗	2K
VOICES-IN-THE-WILD-2M	✓	✓	✓	✓	✓	✓	✓	✓	✓	2M

DAPS provides aligned studio and consumer-device recordings across device/room combinations; VOiCES provides source and replay recordings with controlled room, noise, and microphone conditions. Real recording therefore does not imply the absence of paired targets or experimental control.

Our premise is that robust recognition must build three capabilities in order. *Coverage* exposes the model to compound but recoverable acoustic failures; *grounding* makes it use the remaining acoustic evidence before relying on linguistic priors; and *optimization* adapts the corrective behavior to the residual error regime. This ordering is consequential: data alone does not establish grounding, and a regime-aware reward cannot reliably discover grounding from an unprepared policy.

MEGA-ASR instantiates this capability chain. **VOICES-IN-THE-WILD-2M** supplies coverage through seven acoustic effects organized into 54 physically interpretable scenarios with calibrated shared severity. Built on this distribution, **Acoustic-to-Semantic Progressive Supervised Fine-Tuning (A2S-SFT)** first adapts the encoder–aligner, then the language side, and finally both jointly to obtain **MEGA-ASR-BASE**. **Dual-Granularity WER-Gated Policy Optimization (DG-WGPO)** then changes credit assignment with the sampled hypothesis's WER: it favors token-level refinement when speech remains recoverable and sentence-level reconstruction when local agreement is insufficient. An optional lightweight router is a deployment mechanism, rather than a fourth training stage: it retains the original backbone for clean inputs and activates the robust LoRA branch for degraded inputs. The methodological contribution is therefore the chain from compound coverage to acoustic grounding to error-regime-aware recovery, rather than any isolated use of WER, edit similarity, LCS, length control, or gating.

2 RELATED WORK

ASR Foundation Models and Robust Recognition. Modern encoder–decoder ASR systems, self-supervised speech models, and audio-language models have greatly reduced error on standard speech benchmarks (Radford et al., 2023; Xu et al., 2026; Shi et al., 2026; Gao et al., 2023; Xu et al., 2025a;b; Ding et al., 2025; Wu et al., 2025). Robust-ASR research complements this progress with multi-condition training, enhancement, adaptation, and decoding or re-ranking methods (Ko et al., 2015; 2017; Parada et al., 2022; Yan et al., 2025). These methods improve robustness, but typically target isolated perturbations. We instead address compound degradation, where errors become structural and demand coverage, grounding, and regime-aware optimization.

Robust-ASR Datasets and Acoustic Simulation. Established resources evaluate complementary aspects of adverse recognition, including noise, distant microphones, reverberation, device mismatch, and conversational or read speech (Hu and Loizou, 2007; Watanabe et al., 2016; Richey et al., 2018; Mysore, 2014; Rousseau et al., 2012; Ardila et al., 2020; Pavlichenko et al., 2021). They are essential for measuring robustness, but most provide a fixed collection of recordings or emphasize isolated acoustic factors. Likewise, standard augmentation techniques—noise mixing, RIR convolution, spectral perturbation, clipping, and codec simulation—supply valuable local perturbations (Snyder et al., 2015; Reddy et al., 2020; Ko et al., 2015; 2017; Parada et al., 2022). They rarely control effect co-occurrence or difficulty. VOICES-IN-THE-WILD-2M specifies compound scenarios and calibrated severity to provide learnability-aligned supervision.

Audio-Language Models for Grounded ASR. Audio-language models augment speech processing with contextual knowledge, audio understanding, and instruction-following capabilities (Xu et al., 2025a;b; Ding et al., 2025; Bai et al., 2024). Related work has explored LLM-based ASR correction (Lina and Aksyonov, 2024), general audio reasoning (Xie et al., 2025a), and speech-based reasoning (Xie et al., 2025b), showing that semantic capacity can be valuable when the signal is ambiguous. The remaining ASR-specific question is not whether semantics can help, but when they should be allowed to help: post-hoc linguistic correction can improve textual plausibility while weakening faithfulness to the audio. MEGA-ASR instead grounds semantic recovery within recognition: A2S-SFT stabilizes the acoustic interface before semantic adaptation, and DG-WGPO adapts correction granularity to the error regime. Together, these strands motivate a unified chain that couples compound acoustic coverage with grounded and error-aware adaptation.

3 VOICES-IN-THE-WILD-2M: COMPOUND ACOUSTIC DATA FOR LEARNABLE ROBUSTNESS

3.1 HIERARCHICAL SIMULATION WITH SHARED DIFFICULTY CONTROL

Uncontrolled natural recordings are indispensable for evaluation, but typically lack paired clean targets and make systematic factor interventions difficult. Controlled recording corpora are important exceptions: DAPS provides aligned studio and consumer-device recordings (Mysore, 2014), while VOiCES provides source and replay recordings under controlled room, noise, and microphone conditions (Richey et al., 2018). To scale compound-factor coverage and severity control, we construct **VOICES-IN-THE-WILD-2M** through signal-level simulation: two million clips spanning *8 primitive transformations*, *7 atomic effects*, and *54 compound scenarios*. An anchor–modifier rule controls effect co-occurrence, while a shared severity variable controls difficulty. Appendix A.1 details the hierarchy. Each clip retains its source transcript and a provenance record of its source audio, active transformations, scenario, and severity; Appendix A.2 documents the source audio and provenance. The resulting data exceed isolated augmentation while retaining factor-level traceability.

Waveform-domain pipeline and scope. All transformations act directly on waveforms, including noise, reverberation and echoes, filtering and bandwidth limitation, nonlinear distortion and clipping, resampling, repetition or silence replacement, and loudness adjustment. Outputs retain the original sampling rate; we neither edit STFT magnitudes nor reconstruct phase. The scenario constraints aim for physically plausible coverage rather than exact reproduction of every room, device, or time-varying process. Source-matched physical replay (Appendix A.3) measures the residual simulation-to-real gap: overall shift cosine 0.797, normalized shift error 0.353, and FD-WavLM 0.338.

From mechanisms to acoustic conditions. The simulator combines eight primitive transformations into seven recognizable acoustic effects. Primitives model individual physical or device mechanisms, whereas atomic effects express conditions such as far-field speech and transmission dropout. This separation expands coverage without replacing interpretable conditions with opaque augmentation labels. Appendix A.4 and Appendix A.5 give the complete chains and parameter ranges.

From conditions to compound worlds. We compose effects with an anchor–modifier rule. Far-field, echo&reverb, and obstruction are mutually exclusive anchors that set the propagation geometry; recording coloration, electronic distortion, noise, and transmission dropout are portable modifiers. We screen candidate combinations for plausible co-occurrence, processing order, and parameter compatibility, retaining 54 valid single- and multi-effect scenarios without mixing incompatible geometries. Appendix A.6 gives the enumeration and chain-merging procedure.

One severity variable per acoustic world. Each clip maps a latent draw to one global severity value $m \in [0, 1]$, shared by all active primitive effects. The scenario determines *which* degradations occur, whereas m determines *how difficult* the resulting acoustic world is. We compare Sqrt-Forward, Sqrt-Backward, Gaussian-Mid, and Linear mappings, and adopt Linear based on an independent development pilot on real noisy speech, separate from final testing. Appendix A.7 give the calibration results and parameterization.

Filtering for learnability. Some extreme candidates contain too little recoverable acoustic evidence for stable supervision. We therefore decode each synthesized candidate with a filtering ASR model and discard it when its WER against the clean source transcript exceeds 70%. This threshold defines a learnability criterion for grounding-first training, not a boundary on real-world difficulty.

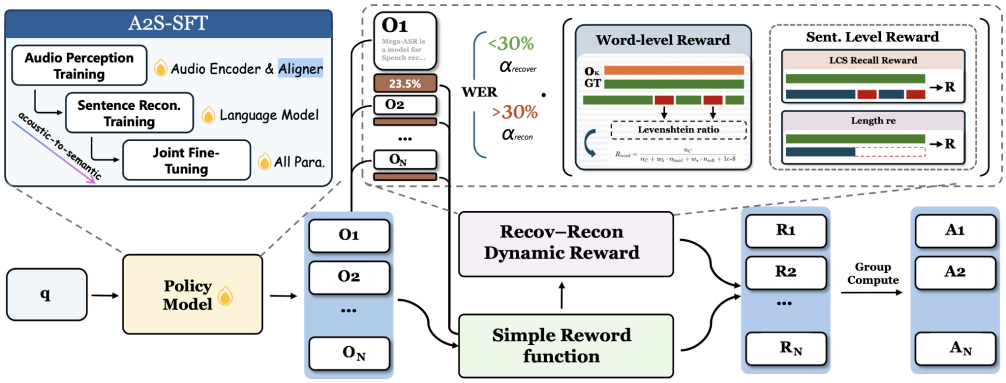


Figure 1: MEGA-ASR's capability chain. A2S-SFT establishes acoustic grounding; DG-WGPO then ranks $K = 12$ sampled hypotheses with repetition-gated, WER-regime-aware rewards.

3.2 VOICES-IN-THE-WILD-BENCH: TRANSFER BEYOND SIMULATION

Simulation is valuable only if its gains transfer beyond the simulator. We therefore release VOICES-IN-THE-WILD-BENCH, a held-out 5,000-clip English–Mandarin evaluation set designed around the same seven atomic effects. Its 3,500 synthetic clips enable controlled factor-wise analysis, while its 1,500 real-world recordings—sourced from internet materials and 16 human participants—test transfer under uncontrolled co-occurrences. The synthetic partition asks whether a model improves on a known degradation mechanism; the real partition asks whether the improvement persists when that mechanism occurs in naturally recorded audio.

Together, VOICES-IN-THE-WILD-2M supplies controllable compound conditions for learning, and VOICES-IN-THE-WILD-BENCH tests whether the resulting robustness transfers beyond simulation. Section 4 then turns this learnability-aligned distribution into an acoustically grounded policy and optimizes its remaining high-WER failures.

4 MEGA-ASR: A CAPABILITY CHAIN FOR ROBUST ASR

Section 3 provides the coverage component; MEGA-ASR builds on it to address the remaining grounding and optimization gaps. A2S-SFT first turns Qwen3-ASR (Shi et al., 2026) into the acoustically grounded MEGA-ASR-BASE. DG-WGPO then adapts this policy with WER-regime-aware preferences for local correction or structural recovery. At inference, an optional environment-aware router selects the robust adapter or original backbone to preserve clean-speech behavior. Figure 1 outlines the full capability chain.

4.1 A2S-SFT: ACOUSTIC GROUNDING BEFORE SEMANTIC RECOVERY

A2S-SFT makes semantic recovery conditional on a usable acoustic interface. Directly adapting an audio-language model to severely degraded speech can reverse this dependency: the language model may learn to produce plausible text before the encoder–aligner reliably represents the input. We therefore separate acoustic and semantic adaptation, and reunite them only after both components have been prepared. Phase I starts from Qwen3-ASR-1.7B, and Phases II and III continue sequentially from the preceding checkpoint. All phases use parameter-efficient LoRA updates; Appendix B.1 gives the stage-wise configuration.

Phase I: encoder–aligner acoustic adaptation. We update only the acoustic encoder and speech-to-LLM aligner. A 30K-utterance curriculum progressively expands the training subset from WER $<$ 30% to WER $<$ 50% and finally to WER $<$ 70%. This is the only A2S-SFT phase that uses a WER curriculum: it establishes a stable acoustic interface on comparatively recoverable examples before introducing harder compound conditions.

Phase II: LLM-side semantic adaptation. With the encoder and aligner frozen, we update the LLM-side LoRA parameters on the full targeted split. This exposes the decoder to incomplete and ambiguous acoustic evidence while preserving the interface established in Phase I.

Phase III: joint acoustic–semantic alignment. Finally, we jointly update the encoder, aligner, and LLM on the same targeted split, using a smaller learning rate on the acoustic modules than on the LLM. This end-to-end phase aligns semantic recovery with the stabilized acoustic representation and yields MEGA-ASR-BASE, the grounded policy used to initialize DG-WGPO. The staged ordering matters: acoustic-only adaptation has limited ability to recover severely corrupted utterances, whereas LLM adaptation before acoustic warm-up increases reliance on language priors (Appendix B.1).

4.2 DG-WGPO: ERROR-REGIME-AWARE POLICY OPTIMIZATION

Once the policy is acoustically grounded, its residual errors become amenable to targeted optimization. Near-reference hypotheses call for local repair, whereas severely corrupted hypotheses require sentence-level recovery and rejection of omissions, repetitions, and unsupported continuations—a distinction that WER alone cannot express. DG-WGPO initializes from LoRA-merged MEGA-ASR-BASE and trains on a split enriched with medium- and high-WER cases. It retains the audio, reference, base prediction, and base WER for targeting and diagnostics; the reference is used only by the reward. We optimize DAPO with GRPO-style group-relative advantages (Yu et al., 2025), updating the acoustic encoder, aligner, and language model through LoRA. For each prompt, $K = 12$ candidates are ranked by a rule-based WER anchor and an error-gated dynamic reward. Appendix B.2 provides the full training protocol.

Methodological contribution. DG-WGPO does not rely on any single new reward primitive; its contribution is a data–optimization co-design for compound acoustic distortions. On VOICES-IN-THE-WILD-2M, errors progress from local substitutions under recoverable degradation to omissions, hallucinations, and sentence-level damage under severe degradation. DG-WGPO therefore uses each sampled hypothesis's WER to allocate credit between local correction and global reconstruction, rather than applying one fixed reward granularity or mixture to every hypothesis. Section 5.5 compares this allocation with WER-only and ungated alternatives.

4.2.1 WER ANCHOR AND REPETITION GATE

Let H denote a sampled transcription and R its reference. The static component anchors learning to transcription accuracy, while a hard gate excludes a common degenerate mode before it can receive credit from any other reward term.

WER reward.

$$R_{\text{wer}}(H, R) = 1 - \min(\text{WER}(H, R),\, 1)\,. \tag{1}$$

Here WER is expressed as a ratio and capped at 1 (100%) only when computing this reward, ensuring $R_{\text{wer}} \in [0, 1]$. Reported evaluation WER remains uncapped.

Repetition gate. Repeated short n-grams are a frequent degenerate rollout mode. We therefore use a binary gate with a preset repetition threshold,

$$R_{\text{rep}}(H) = \begin{cases} 0, & \text{if } H \text{ contains repeated } n\text{-grams beyond threshold,} \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

The static reward is therefore

$$R_{\text{static}} = R_{\text{rep}} \cdot R_{\text{wer}}. \tag{3}$$

4.2.2 WER-GATED DUAL-GRANULARITY REWARD

The dynamic component combines two complementary views of an ASR hypothesis. R_{fine} emphasizes local, acoustically plausible correction, while R_{struc} emphasizes ordered content preservation and length consistency when local agreement is no longer sufficient.

Token-level refinement reward. We partition substitutions in the WER alignment by character-level edit similarity. Given a hypothesis token h and reference token r,

$$\text{sim}(h, r) = 1 - \frac{\text{edit}(h, r)}{\max(|h|, |r|)} \in [0, 1], \tag{4}$$

we classify a substitution as *soft* if $\mathrm{sim}(h, r) \geq 0.5$ and *hard* otherwise. Insertions and deletions are treated as hard errors because they reflect unsupported additions or missing content. The refinement reward discounts soft substitutions while retaining a stronger penalty for hard errors:

$$R_{\text{fine}} = \frac{n_C}{n_C + n_{\text{hard}} + \alpha_s \, n_{\text{soft}} + \epsilon}, \tag{5}$$

where n_C, n_{hard}, n_{soft} are the counts of correct tokens, hard errors, and soft errors respectively, $\alpha_s \in (0, 1)$ is the soft-error discount, and $\epsilon = 10^{-8}$ ensures numerical stability.

Sentence-level reconstruction reward. For heavily corrupted hypotheses, we score ordered content preservation rather than relying only on local token agreement:

$$R_{\text{struc}} = \frac{1}{2} \cdot \frac{\mathrm{LCS}(H, R)}{|R|} + \frac{1}{2} \cdot \max\left(0,\, 1 - \frac{\big||H| - |R|\big|}{|R|}\right). \tag{6}$$

Here, $\mathrm{LCS}(H, R)$ is the longest shared ordered subsequence of the normalized hypothesis and reference. Its normalized score measures ordered content coverage, while the length term penalizes truncated or overlong outputs. We compute WER and all reward terms using the same benchmark-specific normalization and evaluation units, rather than language-model subword IDs; Appendix B.2 gives the operational procedure.

WER-gated dynamic fusion. We use mirrored weights so that local refinement dominates when WER is below the gate and structural reconstruction dominates otherwise:

$$R_{\text{dynamic}} = \begin{cases} 0.75\, R_{\text{fine}} + 0.25\, R_{\text{struc}}, & \mathrm{WER}(H, R) < \tau, \\ 0.25\, R_{\text{fine}} + 0.75\, R_{\text{struc}}, & \mathrm{WER}(H, R) \geq \tau. \end{cases} \tag{7}$$

Final reward. The full reward retains the repetition gate for both components:

$$R = (1 - \alpha_{\text{dyn}})\, R_{\text{static}} + \alpha_{\text{dyn}}\, R_{\text{rep}}\, R_{\text{dynamic}}. \tag{8}$$

We use $\tau = 0.3$, $\alpha_s = 0.4$, and $\alpha_{\text{dyn}} = 0.6$. These values are selected once on an independent held-out validation split, separate from the final test sets, and then fixed for CHiME-4, VOiCES, NOIZEUS, and VOICES-IN-THE-WILD-BENCH; no test set is used for reward selection. The gate is therefore an error-regime allocation rule rather than a claim that 30% WER is a universal boundary. The dynamic component changes *what* the policy prefers without removing the direct WER anchor or rewarding repeated output.

4.3 OPTIONAL ENVIRONMENT-AWARE ROUTING

The robust LoRA branch complements, rather than replaces, the original Qwen3-ASR backbone. An *optional* environment-aware router selects the decoding branch at deployment time. Given utterance x, it predicts a degraded-speech probability $p_{\mathrm{dirty}}(x)$ and applies

$$\mathrm{route}(x) = \begin{cases} \text{MEGA-ASR LoRA branch}, & p_{\mathrm{dirty}}(x) \geq 0.5, \\ \text{Qwen3-ASR base branch}, & p_{\mathrm{dirty}}(x) < 0.5. \end{cases}$$

The router activates precomputed LoRA deltas within one Qwen3-ASR instance for degraded inputs and retains the base weights otherwise; it neither generates text nor changes the decoder. It therefore avoids loading a second ASR model and is a deployment mechanism rather than a fourth training stage. Appendix C provides implementation and diagnostic details.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP AND EVALUATION PROTOCOL

We evaluate WER on three regimes: *standard ASR* (LibriSpeech (Panayotov et al., 2015), Common Voice (Ardila et al., 2020), FLEURS (Conneau et al., 2023), AISHELL-1 (Bu et al., 2017), WenetSpeech (Zhang et al., 2022), and VoxPopuli (Pavlichenko et al., 2021)); *established robustness*

Table 2: WER (%; lower is better) on standard ASR benchmarks. LibriSpeech entries are reported as clean/other; underlining denotes the best open-source result.

Model	LibriSp.		Comm.Voice		Fleurs		AISHELL-1	WenetSp.		VoxPop.
	Dev	Test	zh	en	zh	en	test	net	meeting	en
Closed-source models										
Gemini-3-Flash	1.7|3.56	1.81|4.91	13.58	8.49	7.52	4.01	2.66	14.38	17.62	7.74
Doubao-LLM ASR	2.95|4.06	2.92|5.32	**4.60**	7.12	2.92	7.22	0.98	**4.46**	**4.90**	7.14
GPT-4o-trans.	1.52|3.29	1.75|4.23	12.61	7.22	2.62	**2.71**	3.52	15.71	31.40	7.02
Open-source models										
Canary-1B-v2	2.07|4.03	2.20|3.58	-	8.91	-	4.48	-	-	-	6.20
Parakeet-TDT-0.6B-v3	1.91|3.54	1.93|3.60	-	8.54	-	4.88	-	-	-	6.11
Voxtral-Mini-3B-2507	1.89|3.88	1.89|4.08	-	10.15	-	3.84	-	-	-	7.08
Step-Audio-2-mini	**1.21|2.50**	1.37|2.75	4.77	**7.04**	**2.48**	3.93	0.81	5.56	5.46	7.43
Kimi-Audio-7B	1.38|2.56	**1.34|2.55**	6.74	8.35	5.88	8.07	**0.76**	6.41	6.25	8.15
Whisper Large-v3	1.74|3.68	1.78|3.53	15.33	16.18	7.70	4.10	5.89	12.02	17.79	9.00
Qwen2.5-Omni-7B	2.05|4.19	2.37|4.21	5.01	8.56	4.64	4.01	1.15	6.16	9.64	6.02
Phi-4-Multimodal-Instruct	1.63|3.65	1.70|3.88	7.04	7.75	3.96	3.59	2.37	6.61	9.71	**5.92**
Qwen3-ASR-1.7B	1.62|3.07	1.62|3.40	7.42	7.57	3.93	3.19	1.52	4.99	5.80	6.25
Our model										
Mega-ASR	1.62|3.21	1.78|3.57	5.8	8.15	5.43	3.76	1.49	5.19	6.17	7.44
Mega-ASR w/ router	1.64|3.07	1.63|3.37	7.37	7.57	3.86	3.17	1.53	4.95	5.89	6.26

benchmarks (CHiME-4 (Watanabe et al., 2016), VOiCES (Richey et al., 2018), and NOIZEUS (Hu and Loizou, 2007)); and real and simulated compound conditions in VOICES-IN-THE-WILD-BENCH. The latter tests individual and mixed effects; matched-budget controls, cross-backbone transfer, and capability-chain ablations isolate the roles of coverage, A2S-SFT, and DG-WGPO, while reward analyses assess the optimization signal.

Training–test separation. We use official training splits of its clean source corpora. We enforce split- and speaker-level separation from the LibriSpeech, Common Voice, AISHELL-1, and WenetSpeech test sets. CHiME-4, VOiCES, and NOIZEUS are also absent from the generator, contributing no clean speech, noise clips, RIRs, replay recordings, or other benchmark assets to training.

Models and comparison protocol. All of our variants are initialized from Qwen3-ASR-1.7B (Shi et al., 2026). We compare with proprietary, conventional/open ASR, and audio-language systems wherever benchmark outputs are available; missing entries denote unsupported evaluation protocols. We report both the always-on robust model and the optional routed variant, which tests whether robust adaptation can coexist with the backbone's clean-domain behavior.

5.2 STANDARD AND ROBUST BENCHMARK RESULTS

General ASR (Table 2). We first ask whether robustness adaptation sacrifices ordinary recognition. The routed model preserves the Qwen3-ASR operating point on standard speech: it matches LibriSpeech test-clean (1.63 versus 1.62), improves test-other (3.37 versus 3.40), and reduces FLEURS Chinese/English WER to $3.86/3.17$ from $3.93/3.19$. We additionally include Phi-4-Multimodal-Instruct on the standard-ASR datasets for which results are available.

Established robustness benchmarks (Table 3). We next test the coverage-to-grounding chain under established adverse-condition protocols. The always-on MEGA-ASR achieves the best average WER (6.70), outperforming Qwen3-ASR (7.93), Whisper Large-v3 (10.72), Phi-4-Multimodal-Instruct (9.52), and Qwen2.5-Omni-7B (15.14). The largest gains occur in the hardest conditions: NOIZEUS 0 dB improves from 23.97 to 19.80, and VOiCES average WER from 8.94 to 7.35. Phi-4-Multimodal-Instruct obtains average WERs of 5.75, 8.64, and 14.17 on CHiME-4, VOiCES, and NOIZEUS, respectively. Appendix D.1 gives the full subset-level breakdowns.

Routing preserves the clean/robust trade-off. The always-on robust branch targets maximum performance under acoustic degradation, whereas routing returns to the base branch when robustness is unnecessary. On real test utterances, binary routing accuracy with respect to clean/degraded labels is 95.2% in the 0–5% Qwen3-ASR-WER range, 99.4% in the 5–20% range, and 100% above 20% (Appendix C.4, Table 22).

Statistical reliability. We use 10,000 paired bootstrap resamples to assess whether the aggregate gains persist across evaluation utterances; repeated greedy decoding produces identical transcripts. Appendix D.2 (Table 29) shows that MEGA-ASR reduces robust macro WER from $7.93\,[6.86, 9.06]$ for Qwen3-ASR to $6.70\,[5.73, 7.73]$, while routing preserves standard-ASR performance ($4.19\,[4.11, 4.28]$ versus $4.20\,[4.12, 4.29]$).

Table 3: Robust ASR results on established adverse-condition benchmarks (WER %; lower is better).

Model	CHiME-4			VOiCES					NOIZEUS					Avg.
	Real	Sim	Avg.	rm1	rm2	rm3	rm4	Avg.	0dB	5dB	10dB	15dB	Avg.	
Closed-source models														
Gemini-3-Flash	6.58	5.67	6.125	3.10	4.27	25.99	21.86	13.81	55.78	24.48	18.49	8.52	26.82	15.59
Doubao-LLM ASR	9.95	11.62	10.79	4.86	6.99	17.23	7.85	9.23	25.78	9.51	4.96	2.87	10.78	10.27
GPT-4o-trans.	5.36	7.57	6.47	10.97	12.56	46.68	29.38	22.65	62.40	20.56	6.15	2.64	22.94	17.35
Open-source models														
Voxtral-Mini	6.01	9.04	7.53	3.50	3.51	27.54	16.45	12.75	41.06	15.80	4.85	2.94	16.16	12.15
Kimi-Audio-7B	5.66	7.46	6.56	2.10	2.23	26.95	15.13	11.60	38.33	11.36	4.34	2.27	14.08	10.74
Whisper Large-v3	5.65	8.39	7.02	2.85	2.97	25.68	15.65	11.79	34.71	12.55	3.93	2.17	13.34	10.72
Canary-1B-v2	7.19	9.73	8.46	3.14	3.00	24.88	15.56	11.65	38.53	12.76	6.56	3.77	15.41	11.84
Parakeet-v3	6.61	8.82	7.72	3.23	3.27	19.77	13.84	10.03	38.95	14.67	5.99	3.15	15.69	11.15
Qwen2.5-Omni-7B	6.62	8.13	7.37	4.15	4.03	44.76	22.53	18.87	54.91	17.72	3.20	**0.88**	19.18	15.14
Step-Audio-2-mini	5.35	7.06	6.20	**1.81**	**1.98**	23.25	15.19	10.56	32.02	8.94	3.72	2.27	11.74	9.50
Phi-4-Multimodal-Instruct	5.17	6.33	5.75	2.27	2.28	17.54	12.46	8.64	36.42	12.14	4.75	3.36	14.17	9.52
Qwen3-ASR	4.66	6.11	5.39	2.52	2.62	19.18	11.44	8.94	23.97	8.47	3.41	1.96	9.45	7.93
Our model														
Mega-ASR	4.41	6.04	5.23	2.36	2.43	**15.13**	9.46	**7.35**	**19.80**	**6.61**	**2.79**	**0.88**	**7.52**	**6.70**
Mega-ASR w/ router	**4.38**	**5.62**	**5.00**	2.42	2.49	15.32	**9.26**	7.37	**19.80**	6.97	3.05	1.76	7.90	6.76

Table 4: Transfer to VOICES-IN-THE-WILD-BENCH by acoustic scenario and recording source (WER %; lower is better).

Model	Noise		Far.		Obst.		Echo.		Record.		Elc.Dis.		Trans.Drop.		Mixed	
	Real.	Sim.	Real.	Sim.	Real.	Sim.	Real.	Sim.	Real.	Sim.	Real.	Sim.	Real.	Sim.	Real.	Sim.
Closed-source models																
Gemini-3-Flash	7.63	10.61	5.14	1.90	3.73	2.65	8.75	14.86	8.38	19.85	3.15	7.56	5.47	7.65	7.99	9.62
Seed-ASR	8.21	8.11	3.06	3.19	3.10	2.76	16.55	18.21	18.48	23.33	3.89	5.71	7.97	7.46	6.88	9.29
GPT-4o-trans.	13.19	45.78	**1.87**	2.39	**1.57**	2.77	15.62	28.76	13.37	22.60	3.70	8.43	8.76	7.71	5.62	11.00
Open-source models																
Whisper Large-v3	16.57	18.19	3.38	6.85	3.06	6.01	25.34	39.87	18.33	31.81	3.74	8.77	7.04	8.05	8.91	14.79
Qwen2.5-Omni-7B	11.92	17.88	2.35	2.44	2.40	2.08	20.01	32.64	13.71	30.09	2.46	5.96	6.34	5.88	6.40	10.29
Kimi-Audio-7B	35.10	14.59	2.71	1.92	2.49	1.64	24.00	26.58	8.73	18.09	1.83	**2.78**	4.54	6.33	4.44	6.19
Qwen3-ASR	7.51	9.52	2.23	**1.54**	1.73	1.27	10.40	14.61	9.57	19.42	**1.54**	3.41	4.16	4.19	3.30	5.39
Our model																
Mega-ASR	6.33	8.26	2.35	1.61	1.62	**1.23**	**8.62**	12.59	7.65	14.21	1.71	3.72	**2.59**	**2.62**	2.73	4.57
Mega-ASR w/ router	**6.12**	**8.09**	2.33	1.69	1.80	1.41	8.66	**12.22**	**6.91**	**13.23**	1.60	3.35	2.72	2.88	**2.63**	**4.53**

5.3 TRANSFER TO COMPOUND ACOUSTIC CONDITIONS

The preceding benchmarks establish robustness under standard adverse-condition protocols. We now test the coverage claim more directly: whether training on controlled compound worlds transfers to the held-out real and simulated conditions in VOICES-IN-THE-WILD-BENCH.

Table 4 shows transfer beyond single-factor benchmarks: MEGA-ASR improves $12/16$ real/simulated condition cells over Qwen3-ASR. The largest gains are under recording coloration ($9.57/19.42 \rightarrow 7.65/14.21$) and mixed degradation ($3.30/5.39 \rightarrow 2.73/4.57$), with routing further improving mixed conditions to $2.63/4.53$. The external VOiCES and NOIZEUS results in Table 3 corroborate this transfer, with relative WER reductions of 17.8% and 20.4%, respectively.

Does the gain depend on mixed conditions? No. Averaged over the seven non-mixed real conditions, Qwen3-ASR, MEGA-ASR, and the routed variant obtain WERs of 5.31, 4.41, and 4.31, respectively. Thus, the robust branch retains a 16.9% relative reduction over Qwen3-ASR even after excluding the mixed condition, and routing increases the reduction to 18.8%. The four exception cells occur in the real and simulated far-field and electronic-distortion conditions: Qwen3-ASR obtains $2.23/1.54$ and $1.54/3.41$, whereas MEGA-ASR obtains $2.35/1.61$ and $1.71/3.72$, respectively. These simple isolated conditions leave limited headroom, whereas severe and combined degradations yield the larger gains targeted by the proposed simulation. The improvements concentrate where multiple effects erase local evidence and induce structural errors, with MEGA-ASR also substantially outperforming Whisper Large-v3 ($8.91/14.79$) and Gemini-3-Flash ($7.99/9.62$) on mixed degradation.

Qualitative error modes. Appendix E illustrates the same regime transition at the transcript level. Under severe degradation, Qwen3-ASR can produce empty outputs or fluent but off-audio text, whereas MEGA-ASR more often retains the semantic backbone and key entities of the reference (Figures 5 and 6). These examples complement the WER results by showing the failure modes targeted by the structural component of DG-WGPO.

Table 5: Matched-budget operator controls for Qwen3-ASR (WER %; lower is better). Each panel uses the same source utterances and adaptation budget.

(a) Spectral perturbation versus waveform dropout

Training operator	Clean	SpecAugment	Dropout	Avg.
SpecAugment	2.89	8.32	5.22	5.48
VITW Dropout	3.01	**7.66**	**3.89**	**4.85**

(b) RIR convolution versus Strong Echo

Training operator	Clean	Held-out RIR	Strong Echo	Avg.
RIR	1.25	1.46	16.41	6.37
VITW Strong Echo	**1.23**	**1.44**	**11.03**	**4.57**

5.4 Matched-Budget Operator Controls

Starting from Qwen3-ASR, each control uses the same number of source utterances and the same adaptation budget; only the operator used to create the training condition changes. We compare two direct operator pairs: SpecAugment against waveform dropout, and RIR convolution against Strong Echo. Table 5 reports the matched-budget results.

For spectral perturbation, waveform dropout lowers average WER from 5.48 to 4.85 (11.4% relative) and improves both augmented conditions, at a 0.12-point clean-speech cost. For propagation effects, Strong Echo lowers average WER from 6.37 to 4.57 (28.3% relative) and improves all three test conditions. These are operator-level controls: they show that the proposed waveform operators outperform their conventional counterparts under matched data and budget, rather than serving as a comparison of complete robust-ASR systems.

Cross-backbone transfer. The same degradation and adaptation recipe also improves two additional audio-language backbones: robust macro WER falls from 15.14 to 11.94 for Qwen2.5-Omni-7B and from 9.50 to 8.73 for Step-Audio-2-mini (Appendix D.3, Table 30). Thus, the observed gain is not specific to the Qwen3-ASR initialization.

5.5 Testing the Capability Chain

We now hold the compound training distribution fixed and test the two learned links of the method: whether A2S-SFT establishes a better grounded policy than joint SFT, and whether DG-WGPO improves that policy beyond a WER-only objective.

Appendix F.1 (Table 31) validates both links. A2S-SFT improves VOiCES/NOIZEUS WER from $8.31/8.79$ with joint SFT to $7.59/8.12$, supporting acoustic adaptation before semantic recovery (Appendix B.1). From this base, full DG-WGPO reaches $7.35/7.52$, versus $7.73/8.11$ and $7.62/7.98$ for WER-only GRPO/DAPO.

Reward efficiency and sensitivity. The rule-based reward improves VOiCES/NOIZEUS WER over the LLM judge, with a small VOiCES R4 trade-off and $3.2\times$ faster reward computation (Appendix F.2). Sensitivity is limited within the moderate ranges tested on independent development data (Appendix F.3).

6 Conclusion

Robust ASR failures are not merely more frequent word substitutions: as acoustic evidence degrades, they can become omissions, repetitions, empty outputs, and fluent but unsupported text. MEGA-ASR addresses this transition as a capability chain. VOICES-IN-THE-WILD-2M provides controllable, recoverable coverage of compound acoustic failures; A2S-SFT turns that exposure into an acoustically grounded policy; and DG-WGPO changes credit assignment from local refinement to structural recovery as the residual error regime worsens. Across standard ASR, established robustness benchmarks, and held-out compound conditions, this chain improves recognition under severe real and compositional corruption. Matched-budget controls, cross-backbone transfer, and capability ablations further isolate the contributions of the acoustic operators, staged adaptation, and regime-aware reward. The optional router retains the backbone on inputs for which robust adaptation is unnecessary. More broadly, the results suggest that robust ASR requires jointly designing training coverage, acoustic grounding, and error-aware optimization, rather than treating them as independent add-ons.

AI Use Statement

Generative AI tools were used for limited language polishing, code debugging, and formatting assistance. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

Ethics Statement

We use public datasets under their respective licenses. Recordings from human participants were collected with informed consent and released without personally identifying metadata. The resulting models may exhibit performance differences across languages, speakers, and acoustic conditions and should not be used in high-stakes settings without human review.

Reproducibility Statement

Section 5 describes the evaluation protocol. Appendices A–F document data construction, training, inference, detailed results, ablations, and hyperparameter sensitivity. Code, model weights, data metadata, and evaluation scripts are provided in the project repository.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222, 2020.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, et al. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*, 2024.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE, 2017.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*, 2022.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

Kyu Jeong Han, Seongjun Hahm, Byung-Hak Kim, Jungsuk Kim, and Ian R Lane. Deep learning-based telephony speech recognition in the wild. In *Interspeech*, pages 1323–1327, 2017.

Yi Hu and Philipos C Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech communication*, 49(7-8):588–601, 2007.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Interspeech*, volume 2015, page 3586, 2015.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5220–5224. IEEE, 2017.

Sun Lina and Konstantin A Aksyonov. Error correction for speech recognition systems using large language model reasoning capabilities. In *2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, pages 2300–2303. IEEE, 2024.

Gautham J Mysore. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8):1006–1010, 2014.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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. IEEE, 2015.

Pablo Peso Parada, Agnieszka Dobrowolska, Karthikeyan Saravanan, and Mete Ozay. pmct: Patched multi-condition training for robust speech recognition. *arXiv preprint arXiv:2207.04949*, 2022.

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. Crowdspeech and voxdiy: Benchmark datasets for crowdsourced audio transcription. *arXiv preprint arXiv:2107.01091*, 2021.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *arXiv preprint arXiv:2005.13981*, 2020.

Colleen Richey, Maria A Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, et al. Voices obscured in complex environmental settings (voices) corpus. *arXiv preprint arXiv:1804.05053*, 2018.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129, 2012.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

Paige Tuttösí, Mantaj Dhillon, Luna Sang, Shane Eastwood, Poorvi Bhatia, Quang Minh Dinh, Avni Kapoor, Yewon Jin, and Angelica Lim. Bersting at the screams: a benchmark for distanced, emotional and shouted speech recognition. *Computer Speech & Language*, 95:101815, 2026.

E Vincent S Watanabe, Michael Mandel, and Jon Barker. The 4th chime speech separation and recognition challenge. 2016.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23840–23862, 2025a.

Zhifei Xie, Ziyang Ma, Zihang Liu, Kaiyu Pang, Hongyu Li, Jialin Zhang, Yue Liao, Deheng Ye, Chunyan Miao, and Shuicheng Yan. Mini-omni-reasoner: Token-level thinking-in-speaking in large speech models. *arXiv preprint arXiv:2508.15827*, 2025b.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Kaituo Xu, Yan Jia, Kai Huang, Junjie Chen, Wenpeng Li, Kun Liu, Feng-Long Xie, Xu Tang, and Yao Hu. Fireredasr2s: A state-of-the-art industrial-grade all-in-one automatic speech recognition system. *arXiv preprint arXiv:2603.10420*, 2026.

Brian Yan, Vineel Pratap, Shinji Watanabe, and Michael Auli. Improving multilingual asr in the wild using simple n-best re-ranking. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022.

A Data Construction and Simulation Details

A.1 Three-Level Simulation Hierarchy

This appendix details the coverage component of the capability chain. Each simulated recording is specified at three levels: primitive waveform transformations implement underlying mechanisms, atomic acoustic effects define recognizable conditions, and compound scenarios define their co-occurrence.

The primitive layer contains eight controllable transformations: additive noise, echo delay, reverberation, nonlinear distortion, resampling, spectral filtering, loudness transformation, and frame-level stutter. These instantiate seven atomic conditions—*noise*, *far-field*, *obstructed*, *echo&reverb*, *recording coloration*, *electronic distortion*, and *transmission dropout*—each with a dominant mechanism and compatible auxiliary artifacts. Atomic conditions are then composed through a deterministic merge rule that preserves within-condition operator order while resolving redundant cross-condition operators. The hierarchy is

$$8 \text{ primitive transformations} \rightarrow 7 \text{ atomic acoustic effects} \rightarrow 54 \text{ compound acoustic scenarios.}$$

A.2 Source Audio and Waveform-Domain Synthesis

Rather than curate online videos—a costly and difficult-to-scale route—we construct VOICES-IN-THE-WILD-2M by applying controlled simulations to clean, transcribed speech. Clean utterances come from the official training partitions of LibriSpeech (Panayotov et al., 2015), Common Voice (Ardila et al., 2020), WenetSpeech (Zhang et al., 2022), and AISHELL-1 (Bu et al., 2017); validation and test partitions are excluded. The additive-noise pool combines MUSAN (Snyder et al., 2015), the DNS Challenge (Reddy et al., 2020), ESC-50 (Piczak, 2015), and UrbanSound8K (Salamon et al., 2014), comprising approximately $42{,}000$ clips and 129 hours of audio. Each generated clip retains the source transcript and a provenance record containing its source audio, sampled noise waveform (when applicable), active transformations, scenario, and severity.

All acoustic effects are synthesized in the *waveform domain*: filtering, convolution, nonlinear distortion, resampling, loudness transformation, and frame-level replacement operate directly on the audio signal. Thus, the simulator does not edit spectrogram magnitudes or reconstruct phase. These source pools are independent of CHiME-4, VOiCES, NOIZEUS, and the held-out VOICES-IN-THE-WILD-BENCH assets.

A.3 Source-Matched Physical-Replay Validation

The scenario constraints provide physically motivated coverage, not an exact model of every real acoustic process. We therefore test whether the simulator reproduces the direction and scale of real acoustic change in a source-matched physical-replay setting. For each source utterance, we compare its clean version, a corresponding real-world recording, and a simulated counterpart constructed from the same source. The evaluation covers Noise, Far-field, and their composition (Noise + Far-field), with 100 source-matched examples per scenario.

We extract frozen WavLM representations and compare the real and simulated shifts from the clean source. Shift cosine measures agreement in shift direction, normalized shift error measures mismatch in shift magnitude, and FD-WavLM measures feature-distribution distance. We use reference thresholds of shift cosine > 0.7, normalized shift error < 0.7, and FD-WavLM < 0.5. Table 6 reports the agreement scores.

All replayed scenarios meet the three reference thresholds. Although a simulation-to-real gap remains, the results indicate that the waveform pipeline captures the principal acoustic shifts needed for large-scale training, rather than exactly reproducing individual rooms, devices, or recording chains.

A.4 Primitive Waveform Transformations

Role in the hierarchy. The seven atomic effects denote recognizable acoustic conditions; the eight primitives below implement their reusable waveform-level mechanisms. For example, obstruction combines attenuation, filtering, and room effects, whereas transmission failures combine temporal

Table 6: Source-matched physical-replay validation against real recordings using frozen WavLM features. Higher shift cosine and lower normalized shift error / FD-WavLM indicate closer real–simulation agreement; the overall row is computed on the pooled 300 examples.

Scenario	N	Shift Cos. ↑	Norm. Shift Error ↓	FD-WavLM ↓
Noise	100	0.881	0.212	0.351
Far-field	100	0.800	0.351	0.338
Noise + Far-field	100	0.709	0.496	0.424
Overall	**300**	**0.797**	**0.353**	**0.338**

Table 7: Eight primitive waveform transformations in VOICES-IN-THE-WILD-2M, with parameter families, modeled degradations, and representative real-world counterparts.

Primitive effect	Main parameters	Simulated degradation	Typical real-world source
Additive noise	noise source, noise category, relative noise level, wet ratio	Background interference from environmental sounds, human voices, or device noise	Street, office, vehicle, crowd, household environment
Echo delay	delay time, feedback, mix ratio	Discrete delayed reflections and repeated copies of speech	Empty room, corridor, tunnel, large hall
Reverberation	room size, damping, wet level, dry level	Dense room reflections and long-tail spatial smearing	Classroom, auditorium, church, meeting room
Nonlinear distortion	drive gain, wet ratio	Overload, saturation, clipping, and harmonic artifacts	Low-quality microphone, over-amplified recorder, damaged device
Resampling	target sampling rate, wet ratio, probability gate	Bandwidth limitation and high-frequency information loss	Telephone channel, compressed audio, low-bandwidth transmission
Spectral filtering	filter type, cutoff frequency, repeat count, wet ratio	Frequency attenuation and channel coloration	Mask, door, wall, glass, narrow-band device
Loudness transformation	target LUFS	Gain mismatch, distance-induced attenuation, or abnormal recording level	Distant speaker, quiet recording, over-amplified microphone
Frame-level stutter	frame length, stutter probability, repeat probability, maximum repeats	Local dropout, repeated frames, and unstable temporal continuity	Packet loss, unstable streaming, corrupted recording

discontinuities and bandwidth loss. Table 7 summarizes their parameters, target degradations, and real-world counterparts. The shared global severity variable links their strengths, so heterogeneous mechanisms vary coherently rather than independently.

Additive noise. An external waveform is tiled or cropped to the speech duration, RMS-normalized to the target relative level, and mixed with a specified wet ratio to model environmental, interfering-speech, or device noise.

Echo delay and reverberation. Echo delay adds sparse delayed copies whose timing and feedback control perceptible repetitions. Reverberation instead produces dense room reflections, with room size, damping, and wet–dry levels controlling spatial smearing and tail decay.

Nonlinear distortion. Overdrive followed by amplitude clipping produces saturation, clipping, and harmonic artifacts; drive gain determines whether the effect is mild coloration or severe overload.

Resampling and spectral filtering. Resampling down- and then up-samples the waveform, optionally under a probability gate, to model bandwidth loss. Low- or high-pass filtering, optionally repeated, simulates muffling, thin channels, and device coloration.

Loudness transformation. The waveform is normalized to a target LUFS level to model distance attenuation and gain mismatch; clips whose loudness cannot be estimated reliably are left unchanged.

Table 8: Ordered waveform-transformation chains for the seven atomic acoustic effects.

Atomic effect	Primitive-effect chain	Dominant degradation	Representative condition
Noise	`add_noise` → `change_volume`	Low signal-to-noise ratio	Street, cafe, vehicle, crowd
Far-field	`add_reverb` → `apply_filter` → `change_volume`	Distance-induced reverberation and attenuation	Speaking to a distant microphone
Obstructed	`apply_filter` → `add_reverb` → `change_volume`	Occlusion-induced spectral loss	Speech behind a wall, door, or mask
Echo&reverb	`add_reverb` → `apply_filter` → `add_echo` → `change_volume`	Strong reflections and delayed echoes	Gymnasium, garage, large hall
Recording coloration	`add_resample` → `add_noise` → `apply_filter` → `apply_filter` → `change_volume`	Playback-recording channel degradation	Phone playback recorded by another device
Electronic distortion	`add_distortion` → `apply_filter` → `change_volume_distortion`	Clipping and nonlinear overload	Close-talking with excessive recording gain
Transmission dropout	`add_stutter_replace` → `change_volume`	Local temporal discontinuity	VoIP packet loss, unstable Bluetooth or streaming

Frame-level stutter. Short waveform segments are selectively replaced by the preceding frame or silence. Keeping total duration fixed preserves transcript alignment while modeling packet loss, unstable streaming, repetition, and local dropout.

A.5 Atomic Acoustic Effects

The primitives in Appendix A.4 are assembled into seven recognizable acoustic effects: *noise*, *far-field*, *obstructed*, *echo&reverb*, *recording coloration*, *electronic distortion*, and *transmission dropout*. Each effect is an ordered primitive chain with one dominant mechanism and only acoustically compatible auxiliary artifacts.

Table 8 lists the chains, target degradations, and representative conditions; Table 9 gives the parameter ranges. Chain order is preserved when effects are merged into compound scenarios. Parameters marked "core" are controlled by the global severity variable and determine the strength of each condition.

Noise. Additive noise at varying relative levels, followed by loudness normalization, isolates low-SNR masking without introducing confounding room effects.

Far-field. Reverberation, low-pass filtering, and attenuation capture the weaker direct path, spectral loss, and room reflections of distant-microphone speech.

Obstructed. Repeated low-pass filtering is the dominant barrier mechanism; reverberation and attenuation produce the muffled, weakened speech expected behind a wall, door, glass, or mask.

Echo&reverb. Dense reverberation and delayed echoes model highly reflective spaces, while mild high-pass filtering and loudness control keep the resulting signal realistic.

Recording coloration. Resampling, additive noise, and band-pass shaping emulate a playback–recording channel with narrowed bandwidth, noise, and channel blur.

Electronic distortion. Nonlinear distortion, low-pass filtering, and loudness control isolate clipping and harmonic overload from room or SNR effects.

Transmission dropout. Frame-level stutter creates repetitions and silences at fixed duration, modeling packet loss and unstable streaming without unnecessary spectral coloration.

Together, these effects form the scenario vocabulary of VOICES-IN-THE-WILD-2M: each emphasizes a distinct mechanism while retaining the auxiliary artifacts needed for a plausible acoustic condition.

Table 9: Severity parameterization for the seven atomic effects. Ranges are sampled, fixed settings are listed separately, and parameters marked "core" follow global severity.

Atomic effect	Primitive effect	Sampled severity parameters	Fixed parameters
Noise	add_noise	noise_db$\in [-5, 10]$ **(core)**	noise_category=filtered_wavs, wet=1.0
	change_volume	–	target_lufs=-23.0
Far-field	add_reverb	room_size$\in [0.4, 0.6]$ **(core)**, damping$\in [0.6, 0.8]$, wet_level$\in [0.4, 0.5]$	dry_level=0.5
	apply_filter	cutoff_hz$\in [3500, 4500]$ **(core)**	filter_type=lowpass, repeat=3, wet=1.0
	change_volume	target_lufs$\in [-38, -27]$ **(core)**	–
Obstructed	apply_filter	cutoff_hz$\in [1500, 2000]$ **(core)**, repeat$\in \{2, 3, 4\}$	filter_type=lowpass, wet=0.9
	add_reverb	wet_level$\in [0.5, 0.7]$	room_size=0.4, damping=0.9, dry_level=0.4
	change_volume	target_lufs$\in [-25, -15]$ **(core)**	–
Echo&reverb	add_reverb	room_size$\in [0.8, 0.95]$ **(core)**, wet_level$\in [0.6, 0.8]$	damping=0.5, dry_level=0.4
	apply_filter	cutoff_hz$\in [100, 300]$	filter_type=highpass, repeat=1, wet=1.0
	add_echo	delay_seconds$\in [0.1, 0.3]$ **(core)**, feedback$\in [0.3, 0.5]$, mix$\in [0.2, 0.3]$	–
	change_volume	target_lufs$\in [-30, -23]$ **(core)**	–
Recording coloration	add_resample	prob$\in [0, 1]$ **(core)**	target_sr=8000, wet=1.0, threshold=0.4
	add_noise	noise_db$\in [-5, 10]$ **(core)**	use_white_noise=True, wet=1.0
	apply_filter	cutoff_hz$\in [400, 600]$ **(core)**, repeat$\in \{4, 5, 6\}$	filter_type=highpass, wet=1.0
	apply_filter	cutoff_hz$\in [3500, 4500]$ **(core)**, repeat$\in \{4, 5, 6\}$	filter_type=lowpass, wet=1.0
	change_volume	–	target_lufs=-23.0
Electronic distortion	add_distortion	drive_db$\in [20, 60]$ **(core)**	wet=1.0
	apply_filter	cutoff_hz$\in [2800, 6000]$	filter_type=lowpass, repeat=1, wet=1.0
	change_volume_distortion	target_lufs$\in [-38, -27]$ **(core)**	–
Transmission dropout	add_stutter_replace	stutter_prob$\in [0.05, 0.3]$ **(core)**, max_repeats$\in \{2, 3, 4\}$	repeat_prob=0.7, frame_ms=20
	change_volume	–	target_lufs=-23.0

A.6 Physically Constrained Scenario Enumeration

To avoid indiscriminate mixing, we partition atomic effects into three mutually exclusive *scene-defining anchors—far-field*, *echo&reverb*, and *obstructed*—and four *portable modifiers—recording coloration*, *electronic distortion*, *noise*, and *transmission dropout*. Anchors define distinct propagation geometries; modifiers can accompany an anchor or one another.

The resulting design contains 54 scenarios across four effect-count groups: 7 single-effect, 18 two-effect, 13 three-effect, and 16 higher-order conditions. Table 10 gives the exact enumeration, covering both diagnostic atomic cases and realistic compound failures.

Table 10: Enumeration of the 54 physically constrained compound scenarios. Anchors set propagation geometry; modifiers add portable environmental or device artifacts.

Group	Construction rule	Number
Single-effect scenarios	Seven atomic acoustic effects	7
Two-effect scenarios	Anchor–modifier pairs: 3×4; modifier–modifier pairs: $\binom{4}{2}$	$12 + 6 = 18$
Three-effect scenarios	Anchor with one of three permitted modifier pairs: 3×3; modifier-only triples: $\binom{4}{3}$	$9 + 4 = 13$
Higher-order scenarios	Anchor with three modifiers: $3 \times \binom{4}{3}$; modifier-only four-way combination: 1; anchor with all four modifiers: 3	$12+1+3 = 16$
Total	–	**54**

Algorithm 1 Merging primitive waveform-transformation chains for a compound scenario.

Require: Atomic-effect configurations $\mathcal{C}$; selected effects $S = [s_1, \ldots, s_m]$
Require: Duplicate-allowed primitive set $\mathcal{D} = \{\texttt{add_noise}\}$
Initialize merged chain $M \leftarrow [\,]$
Initialize previously seen primitive set $V \leftarrow \emptyset$
for s_i in S **do**
    Load ordered primitive-effect chain E_i from $\mathcal{C}[s_i]$
    Initialize current-scene primitive set $U \leftarrow \emptyset$
    for each primitive effect $e \in E_i$ **do**
        if $e.\mathrm{name} \in \mathcal{D}$ **then**
            Append e to M
        else if $e.\mathrm{name} \notin V$ **then**
            Append e to M
            Add $e.\mathrm{name}$ to U
        else
            Skip e as a cross-scene duplicate
        end if
    end for
    $V \leftarrow V \cup U$
end for
return merged primitive-effect chain M

Effect-chain merge. For each scenario, we merge its ordered atomic primitive-effect chains using Algorithm 1. The procedure preserves the order within each atomic condition and removes duplicates introduced across conditions. Additive noise is the sole exception, since independent noise sources may co-occur. Repeated operators within one atomic chain, such as the high- and low-pass filters in recording coloration, remain intact.

A.7 Shared Severity Sampling and Calibration

Scenario composition determines *which* degradations are active; a shared global variable $m \in [0, 1]$ determines *how strongly* they are expressed. For each clip, we draw

$$x \sim \mathcal{U}(0, 1),$$

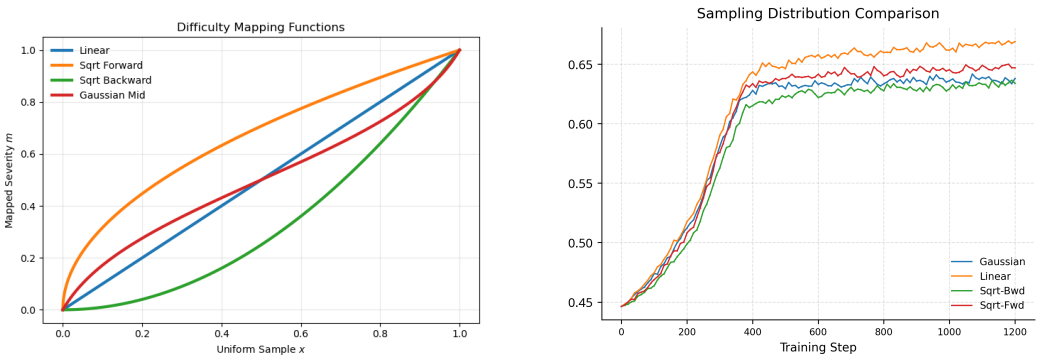
and set

$$m = f(x),$$

where $f(\cdot)$ is one of four candidate difficulty profiles:

$$m_{\text{linear}}(x) = x, \tag{9}$$

$$m_{\text{sqrt-fwd}}(x) = \sqrt{x}, \tag{10}$$



(a) Candidate severity mappings

(b) Pilot comparison on NOIZEUS 0 dB

Figure 2: Severity sampling design and pilot evaluation. (a) Candidate mappings from a uniform draw $x \in [0, 1]$ to global severity $m \in [0, 1]$. Linear is uniform; Sqrt-Forward favors harder samples, Sqrt-Backward easier samples, and Gaussian-Mid medium difficulty. (b) Pilot $1 - \text{WER}$ under the four profiles on NOIZEUS 0 dB; Linear performs best.

$$m_{\text{sqrt-bwd}}(x) = x^2, \tag{11}$$

$$m_{\text{gaussian-mid}}(x) = \text{clip}\big(\Phi^{-1}(0.05 + 0.9x; \mu = 0.5, \sigma),\ 0,\ 1\big)\,, \tag{12}$$

where $\Phi^{-1}(\cdot; \mu, \sigma)$ is the inverse Gaussian CDF with $\mu = 0.5$ and a σ chosen to emphasize the central range. Linear preserves uniform severity; Sqrt-Forward emphasizes harder samples; Sqrt-Backward emphasizes easier samples; and Gaussian-Mid concentrates samples near medium difficulty.

Each random primitive parameter has a range $[a, b]$ and a monotonicity flag. We instantiate it from the same m as

$$\theta = \begin{cases} a + (b - a)m, & \text{if larger values correspond to harder samples,} \\ b - (b - a)m, & \text{if smaller values correspond to harder samples.} \end{cases}$$

Integer-valued parameters are rounded to the nearest valid value, and categorical parameters select from an ordered candidate list using the same m. Thus, a higher severity consistently means, for example, lower low-pass cutoff, stronger reverberation or distortion, lower loudness, or more frequent stutter—rather than an arbitrary mixture of independently sampled strengths.

For this pilot, we generate 50K probe samples under the four candidate profiles and compare the resulting training data on real noisy speech. The pilot uses independent development data, separate from the final test data. Figure 2 links the candidate mappings to this comparison. Linear yields the best pilot result and is the default for VOICES-IN-THE-WILD-2M; it preserves broad, even coverage without over-emphasizing either easy or hard conditions.

B TRAINING DETAILS FOR A2S-SFT AND DG-WGPO

This appendix provides the implementation details underlying Section 4. It instantiates the two remaining components of the capability chain: A2S-SFT closes the grounding gap by establishing an acoustic interface before semantic recovery, and DG-WGPO closes the optimization gap through error-aware policy optimization.

B.1 A2S-SFT: THREE-STAGE TRAINING CONFIGURATION

This subsection details the grounding-first configuration of *Acoustic-to-Semantic Progressive Supervised Fine-Tuning* (A2S-SFT). It comprises three phases: *(i)* encoder–aligner acoustic adaptation, *(ii)* LLM-side semantic adaptation, and *(iii)* joint acoustic–semantic alignment. Phase I starts from Qwen3-ASR-1.7B; Phases II and III continue from the Phase I and Phase II checkpoints, respectively. All phases use LoRA-based parameter-efficient fine-tuning.

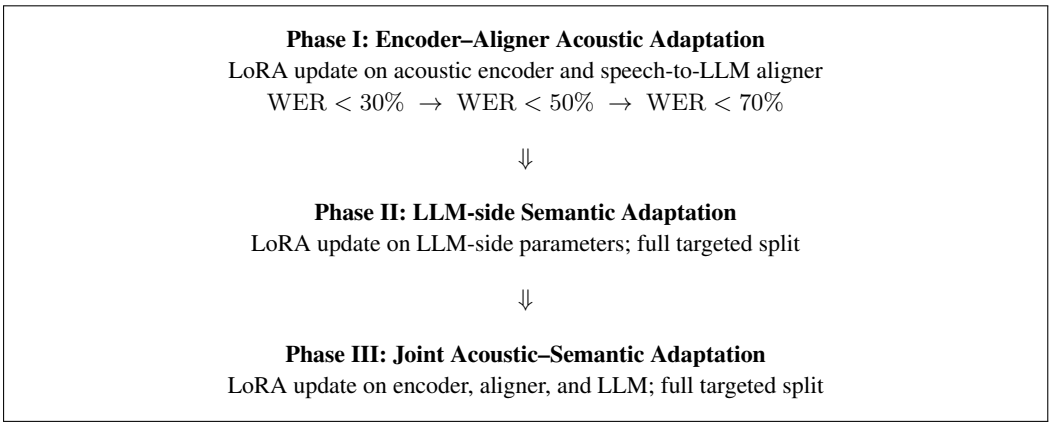


Figure 3: Three-phase A2S-SFT schedule. The WER-graded curriculum is used only for Phase I encoder–aligner adaptation; Phases II and III use the full targeted split.

Training schedule. The three phases differ in trainable scope and data schedule. In Phase I, we sample 30K utterances from the training pool and update only the acoustic encoder and speech-to-LLM aligner. This is the only phase with a WER-graded curriculum: its subset expands from $\mathrm{WER} < 30\%$ to $\mathrm{WER} < 50\%$, and then to $\mathrm{WER} < 70\%$. The schedule establishes a stable acoustic warm start before the encoder–aligner stack encounters harder compound conditions.

Phases II and III use the full targeted split produced by the same preprocessing pipeline. In Phase II, the acoustic encoder and aligner remain frozen while only the LLM-side LoRA parameters are updated for noisy transcription recovery. In Phase III, we jointly update the encoder, aligner, and LLM with LoRA to align acoustic representations and semantic decoding end to end. The validation split is disjoint from training and is used for checkpoint monitoring and failure-mode inspection. Figure 3 summarizes the three-phase schedule and the scope of the WER curriculum.

Stage-wise hyperparameters. Table 11 gives the phase-specific settings. In Phase I, LoRA adapts the upper acoustic blocks and the projection/aligner module, allowing high-level acoustic representations to change while keeping most of the pretrained backbone stable. Phase II freezes the acoustic side and updates the LLM-side LoRA parameters. Phase III jointly updates all three module groups, with smaller learning rates for the acoustic encoder and aligner to preserve the representations established in Phase I.

Table 12 then reports the implementation settings shared across all three phases.

Optimization details. All phases are trained with distributed data parallelism on two GPUs. The effective batch size is computed as

$$B_{\text{eff}} = B_{\text{device}} \times N_{\text{GPU}} \times N_{\text{accum}} = 8 \times 2 \times 8 = 128.$$

Because the model starts from a pretrained ASR–LLM checkpoint, we use low learning rates throughout. In the final joint phase, the encoder and aligner use 5.0×10^{-7}, while the LLM-side learning rate remains 1.0×10^{-6}. This asymmetric setting preserves the acoustic adaptation from Phase I while allowing the language model to adjust to the full noisy transcription distribution. Gradients are clipped to 1.0 in all phases, and checkpoints are saved every 200 optimization steps. Intermediate checkpoints store adapters only, reducing storage overhead and simplifying later merging.

Preliminary training variants. Before fixing the schedule, we examined the update orders summarized in Table 13. These development comparisons motivate staged optimization rather than defining additional final-system variants: updating all modules from the outset exposes the LLM to unreliable acoustic representations, whereas encoder–aligner-only adaptation cannot fully recover severely corrupted speech. The final schedule therefore adapts the encoder–aligner first, then the LLM, and finally both jointly.

Table 11: A2S-SFT hyperparameters by phase. The WER curriculum is used only in Phase I; Phases II and III use the full targeted split.

Setting	**Phase I**	**Phase II**	**Phase III**
Training role	Acoustic warm start	Semantic adaptation	Joint alignment
Trainable modules	Encoder + aligner	LLM	Encoder + aligner + LLM
Data schedule	WER-graded curriculum	Full targeted split	Full targeted split
WER range	$< 30\% \rightarrow < 50\% \rightarrow < 70\%$	Full targeted range	Full targeted range
Per-device batch size	8	8	8
Number of GPUs	2	2	2
Gradient accumulation	8	8	8
Effective batch size	128	128	128
Epochs	2	1	1
Encoder learning rate	1.0×10^{-6}	frozen	5.0×10^{-7}
Aligner learning rate	1.0×10^{-6}	frozen	5.0×10^{-7}
LLM learning rate	frozen	1.0×10^{-6}	1.0×10^{-6}
Warmup ratio	0.05	0.05	0.03
Weight decay	0.01	0.01	0.01
Maximum gradient norm	1.0	1.0	1.0
LoRA rank r	8	8	8
LoRA alpha	16	16	16
LoRA dropout	0.05	0.05	0.05
Checkpoint interval	200 steps	200 steps	200 steps
Saved weights	Adapter only	Adapter only	Adapter / merged adapter

Table 12: Shared implementation settings for A2S-SFT.

Item	**Setting**
Backbone model	Qwen3-ASR-1.7B
Training type	LoRA-based supervised fine-tuning
Distributed training	2-GPU training with one process per GPU
Per-device batch size	8
Gradient accumulation steps	8
Effective batch size	128
Optimizer regularization	Weight decay 0.01
Gradient clipping	Maximum gradient norm 1.0
Warmup	Linear warmup with ratio 0.05 or 0.03 depending on phase
Checkpointing	Save every 200 optimization steps
Checkpoint format	Adapter-only during intermediate phases; merged adapter for downstream use
Model selection	Validation WER together with inspection of empty, hallucinated, and off-audio outputs

Table 13: A2S-SFT schedule variants examined during development.

Variant	Observation
Direct joint SFT from the beginning	Less stable in medium- and high-WER regimes; the model could adapt to unreliable acoustic representations early in training.
Encoder-aligner only	Improved acoustic grounding, but provided limited recovery when the acoustic evidence was incomplete or severely corrupted.
LLM adaptation before acoustic warm start	Increased reliance on the language prior before the acoustic interface was sufficiently stabilized.
No WER curriculum in Phase I	Produced larger validation fluctuations during the acoustic warm-start stage.
Final three-phase schedule	Provided the most stable training behavior by separating acoustic adaptation, semantic adaptation, and final end-to-end alignment.

Table 14: Model-level adaptations that make Qwen3-ASR compatible with DG-WGPO, without implementation-specific function or class names.

Adaptation item	Purpose
Multimodal model loading	Load Qwen3-ASR as an audio-language policy model rather than a text-only decoder.
Official processor consistency	Keep the same audio preprocessing, tokenizer behavior, and prompt protocol between inference, rollout, and RL training.
Multimodal forward interface	Allow the trainer to pass text tokens, text masks, acoustic features, acoustic masks, and response labels in a unified training call.
Language-model interface alignment	Expose the inner language-model embeddings and output head to the LoRA/RLHF trainer without changing the Qwen3-ASR architecture.
Module grouping	Separate the model into language model, acoustic encoder, and aligner groups, so that update scopes can be controlled explicitly.
Rollout re-encoding	Re-encode sampled completions together with the original multimodal prompt and apply loss only on the generated response tokens.
Padding policy	Use left padding for text-side sequences and temporal padding for audio-side features, matching decoder-only generation and acoustic feature batching.
Raw-output preservation	Preserve the original ASR completion format during decoding, allowing the reward function to parse empty outputs, language prefixes, repetitions, and format irregularities consistently.

B.2 DG-WGPO: Multimodal Policy-Optimization Configuration

This subsection details Dual-Granularity WER-Gated Policy Optimization (DG-WGPO), the error-aware policy-optimization stage applied after A2S-SFT. It uses DAPO-style policy optimization with group-relative advantages. Because Qwen3-ASR is an audio-language model rather than a text-only causal language model, its acoustic encoder, speech-to-language aligner, and language model require a multimodal training interface for group-based policy optimization.

Framework and model adaptation. Qwen3-ASR consumes both an audio signal and a text-side prompt. Treating it as a pure text model would therefore break the rollout and loss construction used by GRPO/DAPO. We adapt its training interfaces while preserving its official inference behavior: audio preprocessing and prompt construction remain unchanged, the trainer receives text and acoustic inputs together, and loss is applied only to generated response tokens. Table 14 summarizes the model-level adaptations, including LoRA access and output handling for reward parsing.

Data and initialization. DG-WGPO is initialized from the A2S-SFT LoRA-merged checkpoint rather than the original Qwen3-ASR checkpoint, so the policy enters reinforcement learning with stable acoustic grounding and reasonable transcription ability. Its training and validation sets are targeted WER-aware splits: compared with general SFT data, the RL data enrich medium- and high-WER examples and reduce relatively clean utterances, preventing easy samples from dominating

Table 15: JSONL training-record schema for DG-WGPO. Absolute audio paths are omitted.

Field	Description
`messages`	System and user instructions that define the ASR task, e.g., transcribe the given audio and output plain text only.
`audios`	A list containing the audio file associated with the current prompt. Each training example uses one audio input.
`solution`	Reference transcription used for WER computation and reward evaluation.
`prediction`	Transcription generated by the initialization model. This field is used for data targeting and diagnostic comparison, not as a supervised label.
`base_wer`	WER of the initialization model on the current sample. We use it to emphasize medium- and high-WER regions in RL data construction and analysis.
`meta`	Optional metadata field for bookkeeping.

the policy update. This focuses optimization on regimes where standard supervised fine-tuning and WER-only rewards provide limited corrective signals.

Each RL example contains a multimodal prompt, an audio input, a reference transcription, the base-model prediction, and its WER, which are used for data selection and analysis. Table 15 specifies this record schema. During training, the reference transcription is used only by the reward function; the policy is optimized from sampled completions and their group-wise relative rewards.

Policy optimization setup. DG-WGPO uses GRPO-style group-relative advantage estimation with the DAPO loss. The main run uses three GPUs; the same configuration can be scaled to four or eight GPUs by increasing the number of distributed processes while retaining the per-device batch size, number of generations, and reward settings.

In the main run, the language model, acoustic encoder, and aligner are all allowed to receive LoRA updates. Although all three module groups participate in policy optimization, the total number of trainable parameters remains controlled because the update is parameter-efficient. This full-scope LoRA update is important for DG-WGPO because the reward simultaneously targets acoustic grounding and semantic reconstruction. Updating only the language model improves language-side fluency but is less effective for acoustically induced substitutions and omissions, while updating only the encoder-aligner limits sentence-level recovery in high-WER cases. Therefore, the reported DG-WGPO results use LoRA updates across the acoustic encoder, aligner, and LLM. Table 16 reports the complete optimization configuration.

Rollout and generation protocol. For each prompt, DG-WGPO samples $K = 12$ candidate transcriptions and computes group-relative rewards. Stochastic decoding supplies the intra-group diversity needed for a useful advantage signal; if completions are nearly identical, that signal collapses. Excessive exploration, however, increases hallucinations, overlong outputs, and format violations. We therefore select generation settings through a small exploratory probe rather than fixing them heuristically.

Let b_i denote the WER of the initialization model on example i, and let $H_{i,k}^{(T)}$ be the k-th sampled completion under temperature T. We use two statistics to compare temperature settings. The first measures whether the sample group contains a potentially better candidate:

$$\mathrm{CPI}_\delta(T) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\min_{1\le k\le K} \mathrm{WER}\Big(H_{i,k}^{(T)}, R_i\Big) \le b_i - \delta\right], \tag{13}$$

where CPI denotes the candidate potential indicator. The second measures whether the reward-selected candidate preserves the base transcription ability:

$$\mathrm{BAP}_\delta(T) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\mathrm{WER}\Big(H_{i,k^\star}^{(T)}, R_i\Big) \le b_i + \delta\right], \quad k^\star = \arg\max_k R\Big(H_{i,k}^{(T)}, R_i\Big), \tag{14}$$

where BAP denotes base-ability preservation. We also monitor the valid-output rate: non-empty, non-repetitive completions within the maximum length. The final temperature balances candidate potential, base-ability preservation, and output validity. Table 17 lists the resulting generation configuration, and Table 18 summarizes the temperature probe.

Table 16: DG-WGPO training hyperparameters. The effective prompt batch size corresponds to the three-GPU main run.

Hyperparameter	**Setting**
Initialization	A2S-SFT LoRA-merged Qwen3-ASR-1.7B checkpoint
Optimization method	GRPO-style advantage estimation with DAPO loss
Training type	LoRA
Trainable scope	Acoustic encoder + aligner + language model
Main number of GPUs	3
Scalable GPU settings	3 / 4 / 8 GPUs
Per-device train batch size	4
Per-device evaluation batch size	4
Gradient accumulation steps	16
Effective prompt batch size	$4 \times 3 \times 16 = 192$
Number of generations per prompt	12
Evaluation generations per prompt	4
Maximum completion length	256 tokens
Learning rate	5.0×10^{-5}
Learning-rate scheduler	Cosine decay
Warmup ratio	0.03
KL coefficient β	0.04
DAPO upper clipping parameter	0.28
Number of RL iterations	2
Dynamic sampling	Enabled
Maximum resampling times	4
Overlong filtering	Enabled
Truncation strategy	Delete overlong samples
Checkpoint interval	Every 20 steps
Logging interval	Every 5 steps

Table 17: Generation configuration for the main DG-WGPO run.

Hyperparameter	**Setting**
Number of generations	12
Evaluation generations	4
Temperature	0.50
Top-p	0.95
Top-k	50
Repetition penalty	1.08
Maximum completion length	256 tokens
Dynamic sampling	Enabled
Maximum resampling times	4
Overlong filtering	Enabled

Reward tuning and diagnostics. The main reward follows the DG-WGPO formulation in Section 4. It uses a WER-gated dynamic reward with $\tau = 0.3$, soft-error discount $\alpha_s = 0.4$, and dynamic-reward weight $\alpha_{\mathrm{dyn}} = 0.6$. Below the WER gate, the reward emphasizes token-level refinement; above it, the reward places more weight on sentence-level structural recovery. This is particularly useful for the targeted RL split, which contains many medium- and high-WER predictions for which a standard WER reward can become less discriminative.

Text normalization and sequence units. Before computing WER, edit similarity, or LCS, we apply the same benchmark-specific text normalization used by the WER evaluator to both the prediction and reference, including its prescribed treatment of case, punctuation, and whitespace. We then segment the normalized pair into the evaluation units used by that evaluator. Accordingly, a "token" in R_{fine} or R_{struc} means a normalized evaluation unit, not a language-model subword ID; $\mathrm{LCS}(H, R)$ is computed over this same ordered sequence.

Table 18: Temperature probe for DG-WGPO generation. $T = 0.50$ is selected for moderate exploration with stable base-ASR behavior.

Temperature	Observed tendency	Role in selection
0.30	Conservative decoding with low intra-group diversity	Used to check the lower-exploration regime; often yields weak advantage dispersion.
0.50	Moderate diversity with stable ASR formatting and relatively high valid-output rate	Selected as the default setting by balancing candidate potential and base-ability preservation.
0.70	Higher diversity and more possible corrections	Used to probe whether more aggressive sampling reveals better candidates, but requires stronger filtering.
0.90	Strong exploration with higher risk of hallucination, off-audio text, and over-long outputs	Used only as a stress test for reward robustness and filtering behavior.

Table 19: Reward configuration for DG-WGPO.

Reward hyperparameter	Setting
Static WER reward	$1 - \min(\mathrm{WER}, 1)$ (WER as a ratio)
Repetition gate	Enabled
Soft substitution threshold	Character-level edit similarity ≥ 0.5
Soft-error discount α_s	0.4
WER gate threshold τ	0.3
Dynamic reward weight α_{dyn}	0.6
Low-WER fusion	$0.75R_{\mathrm{fine}} + 0.25R_{\mathrm{struc}}$
High-WER fusion	$0.25R_{\mathrm{fine}} + 0.75R_{\mathrm{struc}}$
Reward scaling	Group-wise scaling
Length and overlong control	Enabled through rollout filtering and reward diagnostics

For consistency with Section 4, the repetition gate applies to both the static and dynamic components:

$$R = (1 - \alpha_{\mathrm{dyn}})R_{\mathrm{static}} + \alpha_{\mathrm{dyn}}R_{\mathrm{rep}}R_{\mathrm{dynamic}}, \qquad R_{\mathrm{static}} = R_{\mathrm{rep}}R_{\mathrm{wer}}.$$

Thus, repeated rollouts cannot receive a positive dynamic-reward contribution. Table 19 summarizes the fixed reward configuration.

Hyperparameter selection protocol. We select reward settings once on an independent held-out validation split, separate from the final test sets, and keep them fixed for CHiME-4, VOiCES, NOIZEUS, and VOICES-IN-THE-WILD-BENCH; no reward hyperparameter is selected on a test set. The validation sweeps cover $\alpha_s \in \{0.2, 0.4, 0.6\}$, $\alpha_{\mathrm{dyn}} \in \{0.4, 0.6, 0.8\}$, and $\tau \in \{0.2, 0.3, 0.4, 0.5\}$; the corresponding sensitivity results are reported in Appendix F.3. We retain equal LCS and length weights within R_{struc} and mirrored $0.75/0.25$ fine/structural fusion weights across the two regimes, so the gate changes the reward emphasis without introducing test-specific mixtures.

We tune the reward by inspecting rollout groups rather than relying only on the scalar training reward. For each diagnostic group, we compare the reference, initial prediction, sampled hypotheses, component rewards, and final ranking. We then check whether a higher reward reflects a meaningful error reduction, such as correcting acoustically plausible substitutions, recovering omitted content, reducing repeated phrases, or improving sentence-level structure. We also inspect failures in which the scalar reward prefers a shorter but incomplete hypothesis, a fluent but off-audio hypothesis, or a format-valid but semantically incorrect transcription. These diagnostics calibrate the WER gate, soft-error discount, and balance between static and dynamic rewards.

Model selection is based on validation WER together with rollout quality statistics, including empty-output rate, repetition rate, overlong-output rate, and the behavior of reward-selected candidates on medium- and high-WER samples. This avoids selecting checkpoints that over-optimize a single reward component while degrading transcription faithfulness.

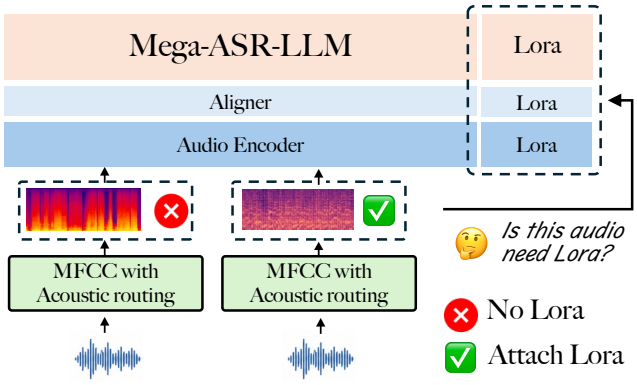


Figure 4: Clean/degraded routing at inference. The router selects the original Qwen3-ASR weights or the robust MEGA-ASR LoRA branch before decoding.

C ENVIRONMENT-AWARE ROUTING DETAILS

This appendix specifies the optional deployment component summarized in Section 4. Routing is applied only after A2S-SFT and DG-WGPO have produced the robust LoRA branch; it does not change either training stage or the ASR decoding procedure. Its sole role is to select the original Qwen3-ASR weights for clean inputs and the robust MEGA-ASR weights for degraded inputs.

C.1 CLEAN/DEGRADED ROUTING OBJECTIVE

The original Qwen3-ASR backbone retains complementary clean-domain behavior, whereas the robust LoRA branch is specialized for the compound conditions in Section 3. The router therefore makes a binary branch-selection decision, rather than attempting to assess transcription quality or generate text. Clean utterances are decoded with the base branch; degraded utterances activate the robust LoRA branch. This makes MEGA-ASR a plug-in robustness module rather than a full replacement for the original ASR system. Figure 4 illustrates this inference-time decision.

C.2 ROUTER ARCHITECTURE

The router is a lightweight audio-quality classifier operating on log-Mel acoustic features. It predicts a binary clean/degraded label with a single-layer Transformer architecture.

The model first extracts 80-dimensional log-Mel features from the waveform. A lightweight convolutional frontend maps the Mel features to a hidden dimension and performs temporal downsampling. The downsampled sequence is then augmented with sinusoidal positional encoding and passed through a single Transformer encoder layer. Finally, attention pooling aggregates frame-level features into an utterance-level embedding for a linear binary head. Table 20 summarizes the architecture.

C.3 CLEAN/DEGRADED ROUTER DATA

The router uses binary supervision: clean speech is labeled 0 and routed to the Qwen3-ASR backbone, while degraded speech is labeled 1 and routed to the MEGA-ASR LoRA branch. The clean subset is drawn from LibriSpeech, AISHELL-1, CommonVoice22, and WenetSpeech; the degraded subset is drawn from VOICES-IN-THE-WILD-2M. The full dataset contains 552,651 clean and 674,107 degraded examples, split into 1,104,084 training, 61,337 validation, and 61,337 test examples. Table 21 summarizes the composition.

On the held-out development set, the router achieves over 99.5% binary classification accuracy, indicating that the acoustic difference between clean and degraded inputs can be reliably detected by a lightweight model.

Table 20: Architecture of the lightweight environment-aware router.

Component	Configuration
Input feature	80-dimensional log-Mel spectrogram
Sample rate	16 kHz
Maximum duration	30 s
Frontend	Lightweight 1D convolutional frontend
Temporal downsampling	2×
Hidden dimension	128
Transformer layers	1
Attention heads	4
Feed-forward dimension	256
Pooling	Attention pooling
Classifier	Linear binary head
Output labels	clean / degraded

Table 21: Clean/degraded dataset for router training. Clean samples are labeled 0 and degraded samples 1.

Subset	Source	Number of samples
Clean	LibriSpeech, AISHELL-1, CommonVoice22, WenetSpeech	552,651
Degraded	VOICES-IN-THE-WILD-2M	674,107
Train split	Mixed clean/degraded	1,104,084
Validation split	Mixed clean/degraded	61,337
Test split	Mixed clean/degraded	61,337

C.4 DIFFICULTY-STRATIFIED ROUTER FAILURE ANALYSIS

The held-out classification result above evaluates the distribution used to train the router. To characterize errors in downstream use, we additionally analyze real test utterances. We stratify examples by base-model Qwen3-ASR WER and sample 500 utterances per stratum. Routing accuracy is binary accuracy with respect to the clean/degraded label: clean utterances should retain the base branch and degraded utterances should activate the robust branch. Table 22 reports the results.

As shown in Table 22, the router is least accurate in the easiest range, where clean speech and mild degradation are acoustically similar and the two ASR branches perform comparably. The few errors in the 5–20$\%$ range mostly involve accent or pronunciation variation rather than distinct environmental corruption. No routing errors are observed above 20% WER, the regime in which selecting the base branch for severely degraded speech is most consequential.

C.5 ROUTER TRAINING OBJECTIVE AND CONFIGURATION

The router is trained as a binary classifier. Given an input utterance x and a label $y \in \{0, 1\}$, where $y = 1$ denotes degraded speech, it predicts $p_\theta(y \mid x)$ and is optimized with cross entropy:

$$\mathcal{L}_{\text{router}} = -\log p_\theta(y \mid x).$$

During training, each audio file is resampled to 16 kHz, converted to mono, and truncated to at most 30 seconds. We extract log-Mel spectrogram features, pad variable-length sequences within each mini-batch, and apply lightweight random gain perturbation and weak additive noise. We use AdamW with a warmup-cosine learning-rate schedule, gradient clipping, label smoothing, and mixed-precision training. Table 23 gives the optimization and evaluation configuration.

C.6 LORA DELTA SWITCHING WITH QWEN3-ASR

We integrate the router with Qwen3-ASR-1.7B using single-model delta switching. Rather than load separate base and robust ASR models, we load one Qwen3-ASR instance and precompute the robust adapters' LoRA weight deltas. A clean routing decision retains the base weights; a degraded routing decision activates the deltas and decodes with the robust MEGA-ASR branch.

Table 22: Difficulty-stratified clean/degraded routing accuracy on real test utterances. Each stratum contains 500 sampled utterances; WER is measured by the base Qwen3-ASR model.

Qwen3-ASR WER range	Routing accuracy (%)
0–5%	95.2
5–20%	99.4
$> 20\%$	100.0

Table 23: Optimization and evaluation configuration for router training.

Item	Setting
Task	Binary clean/degraded classification
Input feature	Log-Mel spectrogram
Sample rate	16 kHz
Maximum duration	30 s
Loss	Cross entropy
Label smoothing	0.1
Optimizer	AdamW
Learning-rate schedule	Warmup + cosine decay
Warmup ratio	0.1
Gradient clipping	1.0
Mixed precision	Enabled
Validation metrics	Accuracy, precision, recall, F1, AUC
Best checkpoint criterion	Validation accuracy / AUC
Development accuracy	$> 99.5\%$

Algorithm 2 Router-guided LoRA delta switching for Qwen3-ASR inference.

Require: Input audio x, router g, threshold γ
Require: Qwen3-ASR base model with preloaded LoRA deltas
Compute dirty probability $p_{\text{dirty}} \leftarrow g(x)$
if $p_{\text{dirty}} \geq \gamma$ **then**
    Set LoRA state to active
    Decode x with the MEGA-ASR branch
else
    Set LoRA state to inactive
    Decode x with the Qwen3-ASR base branch
end if
Return transcription

Concretely, the system first runs the audio-quality predictor and obtains a dirty probability p_{dirty}. With threshold $\gamma = 0.5$, routing is defined as

$$\text{route}(x) = \begin{cases} \text{Mega-ASR LoRA branch}, & p_{\text{dirty}}(x) \geq \gamma, \\ \text{Qwen3-ASR base branch}, & p_{\text{dirty}}(x) < \gamma. \end{cases}$$

The switch adds or subtracts the precomputed LoRA delta tensors at the corresponding base weights. It therefore requires no full-model reload. Algorithm 2 gives the resulting procedure.

C.7 INFERENCE COST

The router is evaluated once before transcription, and delta switching does not reload the ASR model. For batch inference, we group utterances by routing decision before decoding the base-routed and LoRA-routed groups.

We measure inference time on CHiME-4 with the same evaluation pipeline for direct Qwen3-ASR and router-guided inference. Table 24 shows total runtimes of 374 and 371 seconds, respectively. The -0.8% difference is within normal runtime fluctuation, so this measurement shows no measurable routing overhead.

Table 24: Inference-time cost of router-guided LoRA switching on CHiME-4. Runtime is comparable to direct Qwen3-ASR inference.

System	Dataset	Total runtime	Relative difference
Qwen3-ASR-1.7B	CHiME-4	374 s	–
Qwen3-ASR + router + LoRA delta switch	CHiME-4	371 s	-0.8%

Table 25: Average WER (%; lower is better) on three robust ASR benchmarks.

Model	CHiME-4	NOIZEUS	VOiCES	Avg.
Qwen3-ASR-1.7B	5.39	9.45	8.94	7.93
Mega-ASR	5.23	**7.52**	**7.35**	**6.70**
Mega-ASR w/ router	**5.00**	7.90	7.37	6.76

Table 26: Subset-level WER (%; lower is better) on CHiME-4.

Subset	Qwen3-ASR-1.7B	Mega-ASR	Mega-ASR w/ router
dt05_bus_real	4.01	**3.55**	3.76
dt05_bus_simu	3.85	4.22	**3.71**
dt05_caf_real	4.00	**3.84**	3.90
dt05_caf_simu	5.92	5.94	**5.61**
dt05_ped_real	3.57	3.76	**3.50**
dt05_ped_simu	4.59	4.70	**4.22**
dt05_str_real	**3.79**	4.03	3.89
dt05_str_simu	5.02	5.27	**4.62**
et05_bus_real	6.59	6.21	**5.94**
et05_bus_simu	6.06	**5.47**	5.53
et05_caf_real	5.53	5.22	**4.89**
et05_caf_simu	8.36	8.36	**7.97**
et05_ped_real	4.92	**4.53**	4.71
et05_ped_simu	6.47	6.61	**6.08**
et05_str_real	4.85	**4.16**	4.42
et05_str_simu	8.64	7.74	**7.19**
Average	5.39	5.23	**5.00**

D Supplementary Experimental Analyses

This appendix provides the table-level evidence for the claims in Section 5.2 and Section 5.4. Section D.1 provides the benchmark- and condition-level results behind the established-robustness comparison; Section D.2 supplements the standard–robustness trade-off and its uncertainty; and Section D.3 supplements the cross-backbone generality test.

D.1 Detailed Results on Established Robustness Benchmarks

Table 25 separates the aggregate robust gain into three deployment regimes. Always-on Mega-ASR improves on the Qwen3-ASR backbone for every benchmark: by 0.16 WER points on CHiME-4, 1.93 points on NOIZEUS, and 1.59 points on VOiCES. It therefore gives the best three-benchmark average (6.70%), with 20.4% and 17.8% relative reductions on NOIZEUS and VOiCES, respectively. The router changes this balance. It improves most on CHiME-4 ($5.39\% \rightarrow 5.00\%$) and keeps VOiCES nearly unchanged relative to always-on adaptation (7.37% versus 7.35%), but is 0.38 points worse on NOIZEUS (7.90% versus 7.52%). Thus, always-on adaptation is preferred when maximizing aggregate robustness, whereas routing favors the mixed CHiME-4 conditions while preserving general ASR behavior.

Table 27: Subset-level WER (%; lower is better) on NOIZEUS.

Subset	Qwen3-ASR-1.7B	Mega-ASR	Mega-ASR w/ router
airport_0dB	16.12	12.80	**12.31**
airport_5dB	5.37	**3.31**	3.72
airport_10dB	2.89	**2.07**	2.89
airport_15dB	1.24	**0.41**	1.65
babble_0dB	**24.79**	25.20	27.43
babble_5dB	9.50	**5.79**	**5.79**
babble_10dB	**2.07**	**2.07**	2.48
babble_15dB	**1.24**	**1.24**	**1.24**
car_0dB	29.34	23.90	**22.47**
car_5dB	7.85	**5.79**	6.20
car_10dB	2.89	**2.07**	2.89
car_15dB	2.07	**0.83**	1.65
exhibition_0dB	16.12	**13.70**	14.20
exhibition_5dB	9.09	8.68	**5.79**
exhibition_10dB	3.31	**1.65**	2.89
exhibition_15dB	1.65	**0.83**	2.07
restaurant_0dB	23.14	18.10	**15.03**
restaurant_5dB	9.92	**7.85**	**7.85**
restaurant_10dB	2.89	2.89	**2.48**
restaurant_15dB	2.07	**0.41**	1.65
station_0dB	29.34	**21.10**	23.71
station_5dB	6.61	**5.37**	5.79
station_10dB	3.31	**2.48**	**2.48**
station_15dB	1.65	**0.83**	1.65
street_0dB	28.93	**21.90**	22.47
street_5dB	10.74	**8.26**	11.16
street_10dB	4.96	**4.13**	**4.13**
street_15dB	2.48	**1.24**	2.07
train_0dB	23.97	21.70	**20.81**
train_5dB	8.68	**7.85**	9.50
train_10dB	4.96	4.96	**4.13**
train_15dB	3.31	**1.24**	2.07
Average	9.45	**7.52**	7.90

CHiME-4. Table 26 explains why routing is strongest on CHiME-4. The routed model is best on 10 of 16 subsets: five in the development split and five in the evaluation split. Its wins span both real and simulated recordings, including the difficult evaluation simulated cafeteria ($8.36\% \rightarrow 7.97\%$) and street ($8.64\% \rightarrow 7.19\%$) conditions. Always-on MEGA-ASR is best on five subsets, particularly several real evaluation recordings, whereas the base model is best only on dt05_str_real. The remaining differences are small in already-low-WER conditions, so the routing advantage comes from consistent small improvements across the mixed condition set rather than a single large outlier.

NOIZEUS. The fine-grained results in Table 27 show that the aggregate gain is concentrated where the acoustic evidence is weakest. Across the eight 0-dB conditions, the Qwen3-ASR average is 23.97%, whereas both robust variants reduce it to 19.80%. The improvement holds in seven of these eight environments; babble is the sole exception, where the base model remains best (24.79%). The two robust branches have different strengths at 0 dB: routing is best for Airport, Car, Restaurant, and Train, while always-on MEGA-ASR is best for Exhibition, Station, and Street. At higher SNRs, the always-on model is more consistently favorable: it has the strict best score on 15 of the 24 conditions from 5 to 15 dB, while the remaining gaps are generally small because baseline WER is already low. This pattern accounts for its best NOIZEUS average (7.52%).

Table 28: Subset-level WER (%; lower is better) on VOiCES.

Subset	Qwen3-ASR-1.7B	Mega-ASR	Mega-ASR w/ router
rm1_babb_clo	2.24	**1.94**	2.16
rm1_babb_far	3.14	3.00	**2.89**
rm1_musi_clo	2.31	**2.20**	2.21
rm1_musi_far	2.77	2.72	**2.69**
rm1_none_clo	2.10	**1.95**	2.12
rm1_none_far	**2.16**	2.23	2.23
rm1_tele_clo	2.44	**2.21**	2.28
rm1_tele_far	3.00	**2.66**	2.76
rm2_babb_clo	**2.26**	2.27	**2.26**
rm2_babb_far	3.53	**3.24**	3.29
rm2_musi_clo	2.25	2.13	**2.12**
rm2_musi_far	3.14	**2.69**	2.89
rm2_none_clo	2.02	**1.96**	1.98
rm2_none_far	2.41	**2.14**	2.24
rm2_tele_clo	2.28	**2.17**	**2.17**
rm2_tele_far	3.08	**2.84**	2.97
rm3_babb_clo	7.70	6.77	**6.50**
rm3_babb_far	46.62	**36.50**	37.60
rm3_musi_clo	5.48	**4.93**	5.15
rm3_musi_far	33.83	**25.80**	26.70
rm3_none_clo	3.14	2.64	**2.62**
rm3_none_far	10.35	**8.40**	8.80
rm3_tele_clo	5.96	**4.85**	**4.85**
rm3_tele_far	40.40	31.15	**30.34**
rm4_babb_clo	2.79	**2.56**	2.71
rm4_babb_far	54.01	45.69	**43.73**
rm4_musi_clo	2.40	**1.99**	2.26
rm4_musi_far	12.43	10.54	**9.71**
rm4_none_clo	2.03	**1.91**	2.08
rm4_none_far	2.69	**2.64**	**2.64**
rm4_tele_clo	2.18	**2.02**	2.12
rm4_tele_far	12.95	**8.36**	8.80
Average	8.94	**7.35**	7.37

VOiCES. Table 28 shows that the VOiCES improvement is concentrated in the severe far-field regimes rather than in already-easy close-talking speech. Always-on MEGA-ASR improves over Qwen3-ASR on all eight far-field subsets in rooms 3 and 4, including rm3_babb_far ($46.62\% \rightarrow 36.50\%$), rm3_musi_far ($33.83\% \rightarrow 25.80\%$), and rm4_babb_far ($54.01\% \rightarrow 45.69\%$). Averaged over the eight subsets in room 3, WER falls from 19.19% to 15.13% with always-on adaptation; for room 4, it falls from 11.44% to 9.46%. The router is slightly stronger in room 4 (9.26%), notably obtaining the best scores on rm4_babb_far and rm4_musi_far, whereas always-on adaptation is stronger in room 3 and retains the best overall VOiCES average (7.35%). In rooms 1 and 2, most conditions are already near 2–3% WER, leaving limited headroom and explaining the smaller, mixed differences there.

D.2 AGGREGATE RELIABILITY AND THE STANDARD–ROBUSTNESS TRADE-OFF

Table 29 gives the aggregate comparison with paired bootstrap intervals. Step-Audio-2-mini has the best standard macro WER (3.78%), but its robust macro WER rises to 9.50%; Kimi-Audio-7B is weaker on both aggregates (4.87% standard and 10.74% robust). Qwen3-ASR provides the relevant backbone reference point: its standard and robust macro WERs are 4.20% and 7.93%, respectively. Always-on MEGA-ASR lowers the robust macro by 1.23 points to 6.70%, with the largest absolute reductions on NOIZEUS ($9.45\% \rightarrow 7.52\%$) and VOiCES ($8.94\% \rightarrow 7.35\%$), but

Table 29: Aggregate WER (%; lower is better) with 95% paired-bootstrap confidence intervals. Standard macro averages the 12 entries in Table 2; robust macro averages the three benchmark averages in Table 3.

Model	Standard macro	CHiME-4	VOiCES	NOIZEUS	Robust macro
Step-Audio-2-mini	3.78 [3.69, 3.88]	6.20 [5.78, 6.62]	10.56 [9.51, 11.67]	11.74 [8.45, 15.43]	9.50 [8.32, 10.80]
Kimi-Audio-7B	4.87 [4.76, 4.98]	6.56 [6.17, 6.97]	11.60 [10.65, 12.56]	14.08 [10.52, 17.92]	10.74 [9.50, 12.05]
Qwen3-ASR-1.7B	4.20 [4.12, 4.29]	5.39 [5.01, 5.79]	8.94 [8.13, 9.76]	9.45 [6.40, 12.71]	7.93 [6.86, 9.06]
Mega-ASR	4.47 [4.38, 4.57]	5.23 [4.88, 5.59]	7.35 [6.55, 8.17]	7.52 [4.76, 10.42]	6.70 [5.73, 7.73]
Mega-ASR w/ router	4.19 [4.11, 4.28]	5.00 [4.64, 5.37]	7.37 [6.62, 8.16]	7.90 [5.07, 10.97]	6.76 [5.78, 7.82]

Table 30: Cross-backbone transfer of the Mega-ASR degradation and adaptation recipe. Robust macro WER is averaged over CHiME-4, VOiCES, and NOIZEUS; lower is better.

Backbone	Original	+ Mega-ASR recipe	Relative reduction
Qwen2.5-Omni-7B	15.14	11.94	21.1%
Step-Audio-2-mini	9.50	8.73	8.1%

increases the standard macro to 4.47%. Routing recovers standard performance (4.19%) while retaining a 1.17-point robust-macro improvement (6.76%) and yields the best CHiME-4 score (5.00%). The NOIZEUS intervals are visibly wider than those for CHiME-4 and VOiCES, so the intervals are reported to expose this condition-dependent evaluation variability rather than to rely only on point estimates.

D.3 Cross-Backbone Transfer

Table 30 isolates the portability of the recipe by keeping each new backbone fixed and comparing it before and after adaptation on the same robust macro. Qwen2.5-Omni-7B improves from 15.14% to 11.94%, an absolute reduction of 3.20 points (21.1% relative). Step-Audio-2-mini improves from 9.50% to 8.73%, or 0.77 points (8.1% relative). The magnitude differs with the starting backbone and error level, but both comparisons move in the same direction, indicating that the recipe is not tied to the Qwen3-ASR architecture.

E Qualitative Error Analysis

This appendix provides qualitative counterparts to the compound-condition results in Section 5.3. The examples are illustrative, rather than replacements for the aggregate WER evaluation: they expose the error forms that become consequential once acoustic evidence is severely degraded. We therefore analyze the lexical evidence retained by each system, the semantic role or factual content that changes, and the residual errors in Mega-ASR, rather than treating a fluent output as evidence of a faithful transcription.

E.1 Representative Comparisons under Severe Degradation

Figure 5 compares Gemini-3.1-Pro, Qwen3-ASR, and Mega-ASR under far-field corruption and severe noise. The three cases distinguish missing output from fluent off-audio generation and from partially correct recognition with factual errors.

Far-field reconstruction. In Case 1 (peak -5.2 dB; 87% voiced frames), the reference is a long, entity-bearing narrative containing phrases such as "some honey," "Kobay," and "the tortoise." Qwen3-ASR collapses to an empty string (WER 100.0%). Gemini-3.1-Pro instead generates a coherent but unrelated family anecdote ("she met my father's sister"), yielding WER 86.1%. Thus, its surface fluency conceals a loss of acoustic grounding. Mega-ASR recovers the full reference (WER 0.0%), including the uncommon names and the narrative order.

Noise-induced content hallucination. Case 2 has the short reference "The friendly gang left the drug store." Qwen3-ASR preserves a few salient words but inserts an unsupported first sentence and changes the destination from "drug store" to "drug gang" (WER 85.7%). Gemini-3.1-Pro is

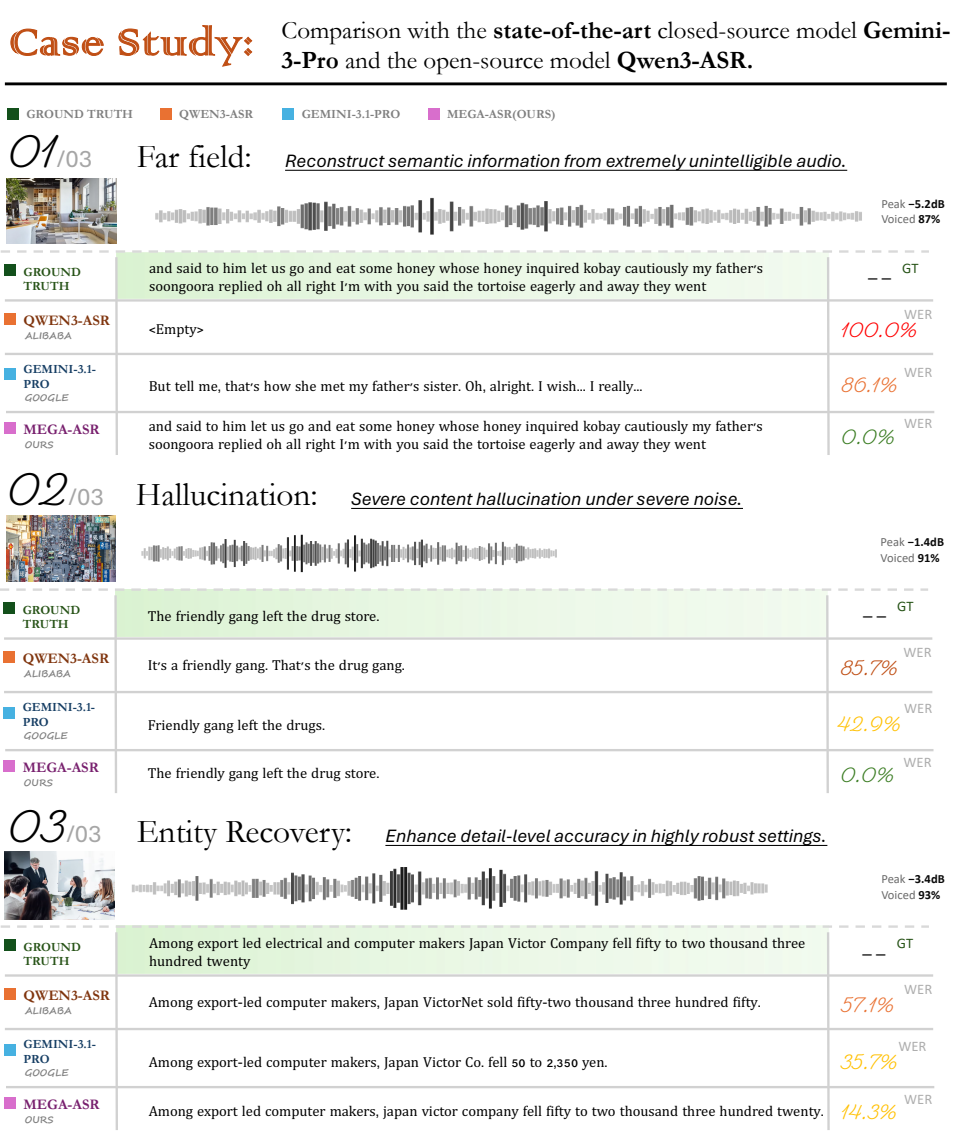


Figure 5: Comparison with GEMINI-3.1-PRO and QWEN3-ASR under severe acoustic degradation. In Case 3, "fell fifty to two thousand three hundred twenty" denotes a decline of 50 to 2,320.

closer, but its "left the drugs" still alters the event content (WER 42.9%). MEGA-ASR returns the reference exactly. This example shows that a lower WER can still mask an unsupported proposition; the relevant distinction is whether the output preserves the event and its arguments.

Entity, event, and quantity recovery. In Case 3, the reference states that Japan Victor Company *fell by fifty to two thousand three hundred twenty*: 50 is the size of the decline, and 2,320 is the resulting level. Qwen3-ASR changes the company to "Japan VictorNet," replaces the decline with "sold," and changes the quantity to fifty-two thousand three hundred fifty (WER 57.1%). Gemini-3.1-Pro preserves more of the sentence structure but changes the endpoint to 2,350 and adds an unsupported currency marker (WER 35.7%). MEGA-ASR preserves the company, the decline of 50, and the resulting level of 2,320; its remaining error is local—it omits "electrical and"—which corresponds to WER 14.3%. The progression from event replacement and entity substitution to a short modifier omission is the type of error-mode change targeted by the method.

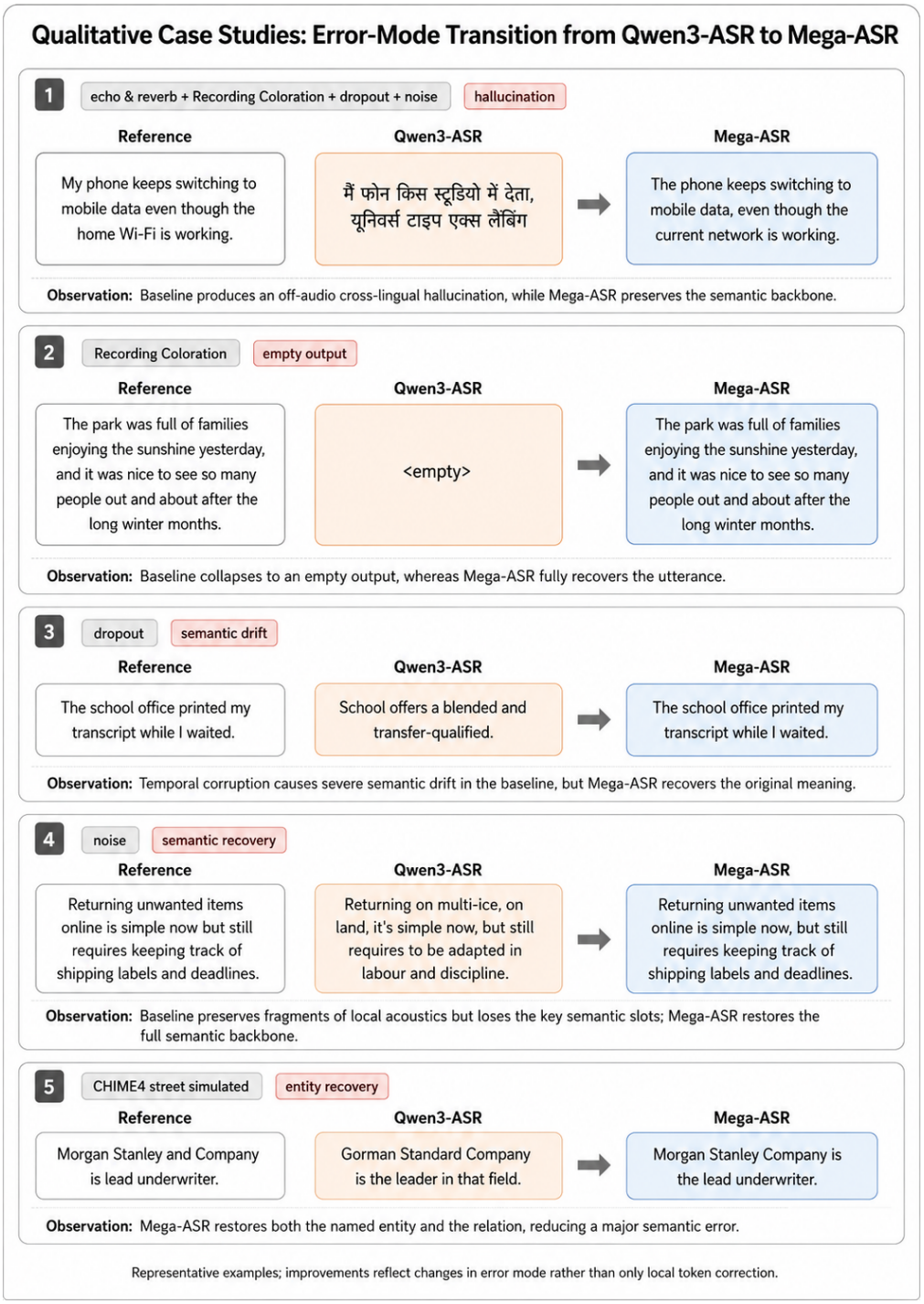


Figure 6: Qualitative comparison of Qwen3-ASR and MEGA-ASR across five adverse acoustic conditions and error modes.

E.2 ERROR MODES ACROSS ACOUSTIC CONDITIONS

Figure 6 isolates five failure modes that are not visible from an aggregate WER alone.

Compound corruption and cross-lingual hallucination. When echo/reverberation, recording coloration, dropout, and noise are composed, Qwen3-ASR emits a Hindi sentence with no lexical

connection to the English reference about a phone switching from home Wi-Fi to mobile data. In contrast, MEGA-ASR preserves the switching event and the mobile-data versus network contrast. Its wording is not identical—"home Wi-Fi" is generalized to "the current network"—but the residual error is a local loss of specificity rather than off-audio generation.

Recording coloration and empty-output collapse. For the long sentence about families enjoying sunshine after winter, Qwen3-ASR returns `<empty>`. MEGA-ASR recovers not only the main subject and predicate, but also the temporally and structurally important clauses "yesterday" and "after the long winter months." This case is useful because it tests recovery of a complete multi-clause utterance rather than a short keyword sequence.

Dropout and semantic drift. The dropout example requires preserving the agent ("the school office"), the action ("printed"), the object ("my transcript"), and the temporal relation ("while I waited"). Qwen3-ASR produces "School offers a blended and transfer-qualified," which shares the school domain but does not preserve the reference's specific agent, action, object, or temporal relation. MEGA-ASR recovers the reference, indicating that it does not merely rely on a school-related language prior.

Noise and semantic-slot corruption. Under noise, Qwen3-ASR retains isolated fragments such as "Returning" and "simple now," but replaces the key slots "unwanted items online," "shipping labels," and "deadlines" with unrelated phrases including "multi-ice" and "labour and discipline." MEGA-ASR restores the online-return task and both required tracking items. The contrast separates local acoustic fragments from preservation of the sentence-level proposition.

CHiME-4 entity–relation errors. In the CHiME-4 street-noise example, Qwen3-ASR changes "Morgan Stanley and Company" to "Gorman Standard Company" and weakens "is lead underwriter" to "is the leader in that field." MEGA-ASR restores the named entity and the underwriting relation, with only a minor formatting omission of "and." This matters because the baseline output remains fluent, yet changes both the central entity and its role.

Taken together, the cases do not establish robustness by themselves; the aggregate evaluation in Section 5.3 provides that evidence. They instead explain its practical meaning: across conditions, MEGA-ASR shifts failures from empty, unrelated, or semantically reversed outputs toward exact transcriptions or localized lexical omissions. This qualitative pattern is consistent with the lower hallucination and missing-content rates in Appendix F.2.

F CAPABILITY-CHAIN, REWARD, AND SENSITIVITY ANALYSES

This appendix provides diagnostic evidence for the capability-chain experiments in Section 5.5. It first isolates the effects of staged A2S-SFT and DG-WGPO, then tests whether the rule-based reward preserves the behavior of a more expensive LLM judge, and finally examines sensitivity to the principal reward settings.

F.1 CAPABILITY-CHAIN ABLATION

Table 31 decomposes the transition from Qwen3-ASR to the full model. Joint SFT is the appropriate control for staged A2S-SFT, while the WER-only GRPO/DAPO rows test whether policy optimization alone can replace DG-WGPO. The remaining rows remove one reward component or replace WER-gated fusion with a fixed mixture. The ablation therefore tests both learned links of the capability chain and the individual mechanisms in the policy objective.

Relative to the full method, removing repetition control, fine reward, structural reward, or WER-gated fusion degrades WER by $0.11/0.21$, $0.10/0.19$, $0.19/0.33$, and $0.06/0.16$ points, respectively; structural reward has the largest effect, consistent with severe omissions and hallucinations.

F.2 RULE-BASED REWARD AND FAITHFULNESS DIAGNOSTICS

We next ask whether a lightweight reward retains the useful preference behavior of an LLM judge, and whether the resulting model improves the failure modes targeted by DG-WGPO. Table 32

Table 31: Ablation of the capability chain and error-regime-aware reward. WER (%; lower is better) on difficult VOiCES/NOIZEUS subsets.

Variant	V	N
Qwen3-ASR	8.94	9.45
+ Joint SFT	8.31	8.79
MEGA-ASR-Base	7.59	8.12
+ GRPO (WER)	7.73	8.11
+ DAPO (WER)	7.62	7.98
DG-WGPO w/o rep.	7.46	7.73
DG-WGPO w/o fine	7.45	7.71
DG-WGPO w/o struct.	7.54	7.85
DG-WGPO w/ fixed fusion	7.41	7.68
MEGA-ASR	**7.35**	**7.52**

Table 32: Rule-based versus LLM-judge reward. WER (%; lower is better); Avg.T. is seconds per step.

Reward	VOiCES	NOIZEUS	VOiCES R4	Avg.T.
LLM-judge	7.51	7.71	**9.27**	62.23
Rule-based	**7.35**	**7.52**	9.46	**19.57**

compares the reward alternatives in both WER and per-step time. Table 33 complements WER with LLM-as-judge hallucination, missing-content, semantic-quality, and key-error measures.

Interpretation. The rule-based reward improves VOiCES and NOIZEUS WER by 0.16 and 0.19 points over the LLM-judge reward, respectively, with a 0.19-point increase on VOiCES R4, while reducing average reward time from 62.23 to 19.57 seconds per step. The independent diagnostic follows the same capability-chain progression: relative to Qwen3-ASR, MEGA-ASR-BASE already reduces hallucination and missing-content rates, and full MEGA-ASR further reduces them to 11.8 and 5.9 while improving semantic quality to 86.4. These measures link the structural reward to the qualitative failure modes illustrated in Appendix E.

F.3 SENSITIVITY TO REWARD WEIGHTS AND THE WER GATE

Tables 34 and 35 report validation sweeps on independent development data, separate from the final test sets. The selected $(\alpha_{\text{dyn}}, \alpha_s) = (0.6, 0.4)$ performs consistently across the four adverse conditions. Sensitivity is limited within the tested moderate ranges: NOIZEUS WER is 7.5–7.8% for the tested combinations with $\alpha_{\text{dyn}} \in \{0.4, 0.6\}$, whereas $\alpha_{\text{dyn}} = 0.8$ degrades performance. Varying τ over $[0.2, 0.5]$ yields WERs of 7.52–7.70%, a spread of 0.18 points, indicating limited sensitivity to the gate within this range.

Table 33: LLM-as-judge evaluation on VOiCES and NOIZEUS. Hall., Miss., Sem., and KeyE. denote hallucination, missing-content, semantic-quality, and key-error measures, respectively.

Model	**Hall.**	**Miss.**	**Sem.**	**KeyE.**
Qwen3-ASR	18.7	14.2	71.3	22.5
Mega-ASR-Base	15.4	11.6	79.8	20.1
Mega-ASR	**11.8**	**5.9**	**86.4**	**19.5**

Table 34: Sensitivity to α_{dyn} and α_s. WER (%; lower is better). V.N.R., V.F., and V.F.R. denote VOiCES near-room, far-field, and far-field–reverberant conditions.

$(\alpha_{\mathrm{dyn}}, \alpha_s)$	NOIZEUS	V.N.R.	V.F.	V.F.R.
$(0.4, 0.4)$	7.7	7.6	7.8	9.5
$(0.4, 0.6)$	7.8	7.6	7.9	9.4
$(0.6, 0.2)$	7.8	7.5	7.6	9.3
$(0.6, 0.6)$	**7.5**	7.5	**7.4**	9.3
$(0.8, 0.4)$	8.1	9.1	8.0	9.9
$(0.6, 0.4)$	**7.5**	**7.4**	**7.4**	**9.2**

Table 35: Sensitivity to the WER gate τ on NOIZEUS. WER (%; lower is better).

τ	0.2	0.3	0.4	0.5
NOIZEUS WER	7.68	**7.52**	7.66	7.70