GEOPOI: SCALABLE POST-TRAINING FOR SELECTIVE GEOGRAPHIC CONTEXTUALIZATION IN ASR

Anonymous authors
Paper under double-blind review

ABSTRACT

Recognizing point-of-interest (POI) names requires geographic knowledge, but irrelevant context can introduce transcription errors. We present GEOPOI, a framework for selective geographic contextualization with a frozen ASR backbone. Cell-specific adapters learn regional knowledge through difficulty-prioritized, geographically sharded post-training. A staged pipeline uses predicted entity spans for acoustic retrieval, teacher-anchored ranking to select relevant names, and trie-based logit biasing to guide transcription. Audio-conditioned routing selects a single cell, a GPS neighborhood, or the unmodified backbone. Training draws on over ten million utterances covering approximately four million POIs. On the evaluation set, with geographic gating enabled, POI character error rate (CER) falls from 14.07% to 5.24% (63% relative reduction). Component comparisons also show a 29% relative reduction in overall CER and increased all-target name containment from 47.8% to 83.0%, with general-speech CER essentially unchanged (4.92% versus 4.86%). We will publicly release our code, trained model weights, and evaluation dataset.

1 INTRODUCTION

The destination name is often the least replaceable part of a navigation or local-search utterance. A transcript may correctly recognize every surrounding word yet substitute a common word for an unfamiliar business, confuse two nearby homophones, or omit the qualifier distinguishing branches of the same brand. These errors are not explained by global frequency alone: a popular name can remain locally ambiguous, while a rare name may be acoustically distinctive. Location provides useful structure, but it does not establish what the speaker actually said.

This setting creates three coupled challenges. First, a recognizer needs sufficient capacity to learn geographically concentrated knowledge without updating a separate full model for every region. Second, a short name can be obscured by the rest of the utterance, and even a retrieved name may fall below the small candidate budget available to the decoder. Third, contextual evidence can be misleading: the query location may be wrong, the target may be absent, or the utterance may contain no POI at all. A system optimized only for candidate recall cannot distinguish a helpful intervention from a harmful one.

GEOPOI addresses these challenges through regional parameter memory and staged acoustic candidate learning. A span head identifies candidate entity intervals. A bounded geographic retriever combines span-conditioned queries with global queries, including a frozen fallback. A neural ranker learns corrections relative to a frozen teacher, and the selected names supply trie-based decoding evidence. Regional adaptation changes hidden representations; candidate bias changes token scores. Both branches retain the shared ASR backbone. An audio-conditioned three-state router determines the geographic scope: a single acoustically supported cell, the observed GPS neighborhood, or no contextual intervention.

Candidate relevance does not by itself establish a beneficial transcription change. We therefore evaluate the unmodified decoder, regional adaptation, candidate biasing, and their combination as separate configurations. Candidate-admitted requests use the top ten ranked names, and preservation is measured directly on non-entity text and previously correct transcripts.

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.

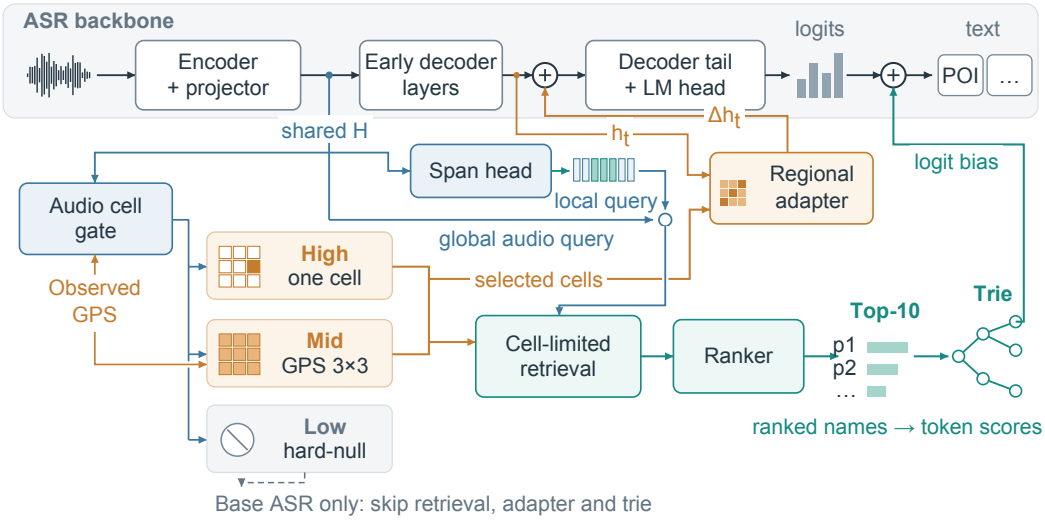


Figure 1: **Framework and deployment routing.** Shared acoustic states supply span-conditioned and global queries. Audio confidence selects one cell, the fixed GPS 3×3 neighborhood, or hard-null. The selected cells bound retrieval and address regional parameters; ranked names form a top-ten trie. Regional residuals modify hidden states, and trie rewards modify token scores. Hard-null bypasses both contextual branches and preserves the base decoding path.

Our training corpus contains over ten million utterances spanning approximately four million POIs. To learn from geographically concentrated examples, the training process selects representative audio, assigns difficult cell–name pairs to disjoint shards, and fixes negative-context assignments before partitioning. Regional learning trains the bottleneck-width factors directly and assembles the cell-specific parameters, averaging shared projections across shards.

Our contributions are: (1) a staged geographic contextualization framework combining span-aware retrieval, teacher-anchored ranking, and regional adaptation; (2) a regional training recipe with difficulty-prioritized coverage, shared negative assignments, and disjoint cell ownership; and (3) an audio-conditioned three-state deployment design with controlled evaluation of its regional and candidate components. Optional shared-basis compression and outcome-based action selection are described in Appendices B and F; the recognition comparison evaluates regional adaptation and candidate biasing.

2 Related Work

Contextual speech recognition. CLAS learns contextual phrase representations (Pundak et al., 2018); TCPGen adds prefix-tree distributions (Sun et al., 2021), including around frozen Whisper (Sun et al., 2023). Gated adapters control contextual activation (Alexandridis et al., 2023), and SeACo-Paraformer combines a frozen backbone with no-bias outputs and hotword filtering (Shi et al., 2024). Entity-category prediction selects context for language-model rescoring (Sun et al., 2024). Recent approaches learn acoustic candidate ranking (Hou et al., 2025) or combine hotword retrieval with keyword and transcription rewards for reinforcement learning (Kong et al., 2025). CTC-supervised bias-position prediction is also used as a training auxiliary removed at inference (Novitasari et al., 2026). Our span head instead supplies local queries at inference, followed by ranking and separate regional and lexical interventions.

Scalable retrieval and adaptation. BR-ASR learns speech–entity retrieval with homophonic negatives (Gong et al., 2025); we divide a bounded geographic shortlist budget between local spans and a preserved global fallback. Bottleneck adapters (Houlsby et al., 2019) and LoRA (Hu et al., 2022) establish parameter-efficient adaptation. DyLoRA trains representations usable at multiple ranks (Valipour et al., 2023); EigenLoRAx extracts principal subspaces from trained adapters or their

matrix products (Kaushik et al., 2025). Our evaluated regional recipe uses disjoint geographic ownership and full bottleneck-width residuals. The shared-subspace extension in Appendix B concerns compression of compatible trained operators and is distinct from the cell-owned assembly evaluated here.

Selection and preservation. Selective classification trades coverage for conditional risk (Geifman & El-Yaniv, 2017), while positive-congruent training reduces regressions on examples previously classified correctly (Yan et al., 2021). Our ranker retains successful teacher margins, and regional objectives penalize changes under inappropriate context. We assess preservation through complete transcripts as well as local entity scores, since candidate compatibility alone does not establish a beneficial transcription change.

3 Problem and Design

Given audio x, observed geographic context g, and a catalog $\mathcal{C}$, the system produces a transcript $\hat{y}$. Audio evidence and observed geography determine a route $\mathcal{R}(x, g)$ within a 3×3 neighborhood of Web-Mercator level-13 cells. The selected cells define the regional modules and the candidate pool $\mathcal{C}(\mathcal{R})$. Geographic inputs are held fixed across compared systems. The test set is constructed with geographic contexts; these are not evidence of naturally collected query-location distributions. Appendix A collects notation and dimensions.

Let $\mathcal{D}_a(x, g)$ denote decoding under action

$$a \in \mathcal{A} = \{\textsc{Null}, \text{regional}, \text{candidate}, \text{both}\}. \tag{1}$$

The Null action invokes the same deployed ASR backbone without regional residuals or candidate bias. The regional action activates a selected adapter; the candidate action uses selected names for trie-based logit biasing; the joint action enables both. Unsupported regional modules do not make lexical retrieval impossible: the candidate path remains a separate option.

The objective is to reduce final transcription error subject to limits on harmful intervention and contextual compute. We report whole-utterance and POI-span character error rate (CER), all-target name containment, non-POI CER, and errors introduced into previously correct content. Candidate Recall@K is a diagnostic, not a substitute for final entity recovery. Because the output is text, identical names at different physical locations do not constitute distinguishable POI identifiers.

4 Method

4.1 Audio-conditioned three-state routing

The deployment router reuses projected audio states to score GPS-local cells and a NONE alternative. High confidence selects the best audio-supported cell c^*; intermediate confidence retains the observed GPS neighborhood $\mathcal{G}(g)$; low confidence takes hard-null:

$$\mathcal{R}(x, g) = \begin{cases} \{c^*\}, & \text{high confidence,} \\ \mathcal{G}(g), & \text{medium confidence,} \\ \varnothing, & \text{low confidence.} \end{cases} \tag{2}$$

High and medium confidence use nested thresholds on application probability, the best-cell versus NONE margin, and the top-two cell margin. The middle route is a fixed nine-position window, retaining in-vocabulary cells. The selected scope determines which cells supply regional memory and POI candidates. Hard-null bypasses candidate retrieval, lexical bias, and regional residuals, returning the base ASR path. Thus audio evidence can narrow the search, GPS can retain bounded coverage under ambiguity, and weak context can be rejected without changing the backbone. Appendix C specifies the routing policy.

4.2 CELL-OWNED REGIONAL ADAPTATION

We build on Qwen3-ASR (Shi et al., 2026), freezing the pretrained ASR weights. For decoder hidden state h_t and cell c, the regional addition is

$$z_t = \mathrm{SiLU}(D\,\mathrm{LN}(h_t)), \tag{3}$$

$$h'_t = h_t + a_{\mathrm{reg}} m_t s_t \big(PC_c + \alpha R_c\big) z_t, \tag{4}$$

where $D \in \mathbb{R}^{32\times 2048}$, $P \in \mathbb{R}^{2048\times 128}$, $C_c \in \mathbb{R}^{128\times 32}$, and $R_c \in \mathbb{R}^{2048\times 32}$. The scalar s_t is a sigmoid gate computed from the causal decoder state. The mask m_t restricts valid positions and routes. The measured configuration attaches at decoder layer 24, counting from zero, and represents $R_c = U_c I_{32}$ without deployment-rank truncation.

Equation 4 gives the single-cell operator used in training, with no geographic-shift augmentation. For the medium-confidence route, the 3×3 neighborhood mixes its operators:

$$C(g) = \sum_{c\in\mathcal{G}(g)} w_c C_c, \qquad R(g) = \sum_{c\in\mathcal{G}(g)} w_c R_c. \tag{5}$$

The center, four edge neighbors, and four diagonal neighbors have initial weights 0.4, 0.1 each, and 0.05 each. Cells outside the routing vocabulary are removed and the remaining weights renormalized. Equation 4 then uses $C(g)$ and $R(g)$. The high-confidence route uses its single selected cell with weight one; hard-null bypasses the addition.

The adapter starts from zero cell cores and residuals under a common seeded initialization. Regional optimization updates D, P, C_c, the scalar gate, and the cell-owned U_c; it does not update the ASR backbone. Teacher-forced regional training restricts the adapter to annotated entity-prediction positions. Correct-, wrong-, and absent-context target views accompany random- and absent-context general views. Entity-weighted transcription loss and preservation terms compare these views; Appendix C gives the settings and supervision limitations.

4.3 COVERAGE, PARTITIONING, AND ASSEMBLY

The national catalog contains 90,760 routing cells. Each cell retains at most 32 names, prioritizing low strict-recall names from a fixed recognition probe. A cell with at least four unknown names receives at least 16 scheduled visits; one with one to three receives at least eight. Known-name anchors replace a bounded share of replay slots. This gives 66,500 trained residual cells and 875,083 scheduled training records, excluding padding.

A deterministic global schedule assigns general examples and incorrect geographic contexts from the full routing vocabulary before partitioning. Twelve shards own disjoint cell sets, preserving the global assignments. Each shard is optimized for four passes. Assembly copies each cell core and residual from its owning shard and takes the arithmetic mean of the shared projections and gate parameters. This is an empirical model-assembly procedure, not an algebraic equivalence to joint optimization: shared parameters diverge during shard training.

4.4 SPAN-AWARE ACOUSTIC EVIDENCE

A lightweight temporal head reads frozen projector states. Layer normalization, a linear projection, and a residual temporal convolution produce frame states, from which three outputs predict entity membership, start, and end. Attention pooling yields utterance-level mention and purpose features; pooled spans additionally support role, spatial-relation, and referent prediction when reliable annotations exist. Missing semantic labels are masked; auxiliary-task coverage is reported from the training annotations.

Thresholded membership predictions define connected intervals, retaining up to four proposals with two-frame padding. Start and end predictions supply confidence evidence rather than splitting adjacent intervals. The default action features use boundary confidence, availability, and local acoustic support; duration and observed truncation are reserved for a separate full-feature setting. These quantities describe acoustic intervals: they are not directly used as a hard mask on autoregressive decoder tokens. An empty span prediction does not veto retrieval, because entity detection itself can fail.

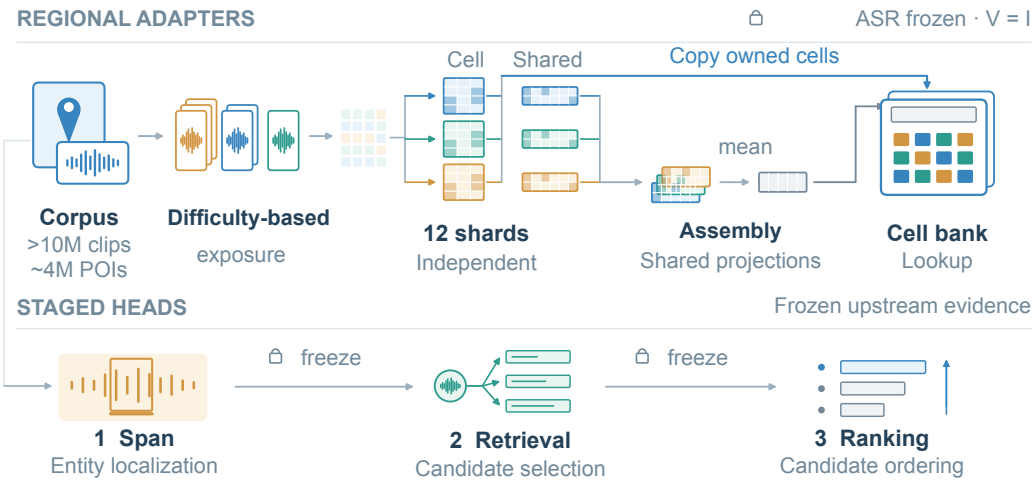


Figure 2: **Geographically sharded and staged post-training.** Difficulty-prioritized audio trains independent shards with disjoint cell ownership. Assembly copies the owned cell parameters and averages the shared projections. Contextual heads train in sequence, passing selected span and retrieval outputs to the next stage. The ASR backbone remains frozen throughout.

The span loss is $\mathcal{L}_{\text{frame}} + \mathcal{L}_{\text{mention}} + 0.2 \sum_u \mathcal{L}_{\text{CE},u}$, where frame and mention terms are masked binary cross-entropies and u indexes the four semantic tasks. Training transitions from annotated to predicted proposals for the span-conditioned semantic heads; a one-to-one soft-IoU match of at least 0.3 transfers available labels and remaps referents. Downstream retrieval and ranking use predicted spans. Alignment-derived time labels are heuristic supervision, with low-confidence labels masked.

4.5 Bounded local and global retrieval

For normalized reader states $A_\tau \in \mathbb{R}^{d_e}$, $d_e = 128$, and soft span weights $w_{s\tau}$, a local query pools valid acoustic frames τ around proposal s:

$$q_s = \text{norm}\left(W_q \sum_\tau \text{softmax}_\tau\big(v_q^\top A_\tau + \log(\epsilon + w_{s\tau})\big) A_\tau + b_q \right). \tag{6}$$

The nonzero $\epsilon = 0.02$ retains surrounding evidence. Local queries are combined with a trainable eight-query global head and a frozen eight-query fallback, both initialized from the established retriever. Catalog embeddings remain frozen during head training. A candidate's retrieval score is the maximum cosine similarity across valid queries, divided by temperature 0.1.

Retrieval searches only the admitted pool $\mathcal{C}(\mathcal{R})$ and deduplicates candidates by decoder surface. The default shortlist budget is $K = 256$, increased to 512 when the geographic pool contains at least $8{,}192$ distinct surfaces. Local queries share this request-level budget with all global queries; they do not each receive an independent full shortlist. The implementation reserves 25% for the global branches together, fills local and global quotas round-robin, and redistributes unused capacity. When no local span is available, global queries receive the whole budget. Thus increasing the number of detected entities does not multiply the shortlist size.

Training uses an entity-wise multi-positive objective. Let $\mathcal{P}_m$ contain valid alternatives for entity m, $\mathcal{N}_-$ contain non-target surfaces, and $s_\varnothing$ be a NONE score. Then

$$\ell_{\text{mp}}(s) = \frac{1}{M} \sum_{m=1}^{M} \left[\log\left(\sum_{j \in \mathcal{P}_m \cup \mathcal{N}_-} e^{s_j} + e^{s_\varnothing} \right) - \log \sum_{j \in \mathcal{P}_m} e^{s_j} \right]. \tag{7}$$

Here M counts entity groups present in the actual pool; other true entities are excluded from each entity's negative set. The retriever minimizes Equation 7 for all-query scores and adds the same objective for local-only scores whenever a valid local query exists. This auxiliary term prevents

the global branch from taking all the learning signal. If no target is present, the objective instead maximizes NONE in a softmax over candidates and NONE. Training labels identify positive surfaces in the actual pool; evaluation never inserts the gold name into the retrieved shortlist.

4.6 TEACHER-ANCHORED RANKING AND REJECTION

The ranker combines candidate-conditioned attention over audio with local-query support, global support, coarse retrieval scores, frozen reader scores, and utterance semantics. It retains a frozen teacher scorer and initializes a trainable copy from it. Candidate scores take the form

$$s_j = s_j^{\text{old}} + \eta f_\psi(u_j) + F_\phi(j) - F_{\phi_0}(j), \tag{8}$$

Here u_j concatenates utterance semantics and five candidate-evidence features, η scales the residual, and f_ψ has a zero-initialized output layer. The trainable scorer F_ϕ initially equals its frozen anchor F_{ϕ_0}, preserving initial candidate scores. The ranking model allows unbounded residual updates and learned fusion; the separately learned NONE score can change candidate acceptance at initialization.

Ranking supervision combines listwise learning, repair of teacher misses, preservation of successful teacher margins, and NONE rejection. Top-k losses target $k \in \{1, 3, 10\}$ with weights $\{0.15, 0.35, 1\}$. For M present entities, the effective target is $q_k = \max(k, M)$. Let Δ_{m,q_k} be entity m's best positive score minus its q_kth competing score. Teacher misses incur $[0.2 - \Delta_{m,q_k}]_+$; hits incur $[\Delta^{\text{old}}_{m,q_k} - \Delta_{m,q_k}]_+$, weighted by 0.5. A hit requires strictly positive teacher margin; pairs with fewer than q_k competitors are omitted. Other true entities consume slots but their competing scores are detached. Repair and preservation means are normalized separately. Positive rows also receive $0.2\ell_{\text{mp}}$; rows without a present target instead retain a 0.1 mean-squared teacher-score anchor.

The NONE head uses utterance semantics and detached pool evidence to predict an offset from the detached maximum candidate score. Binary cross-entropy supervises $\max_j s_j - s_\varnothing$ for target presence. Candidate scores are detached in this loss, and NONE in the listwise term. The optional rejection policy retains at most B distinct surfaces satisfying $s_j > s_\varnothing + \delta_{\text{none}}$. The top-ten policy in the component comparison bypasses this lexical threshold. Ranking does not replace frozen reader scores or their full-shortlist maximum used to scale decoder bias.

4.7 CANDIDATE BIASING AND MEASURED SELECTIVITY

For the component comparison, the decoder uses the ten highest-ranked surfaces without thresholding their lexical NONE scores. A prefix trie adds candidate rewards to permitted next-token logits, using the frozen reader's acoustic support to scale the reward. Ranking therefore determines which names can influence decoding, while the autoregressive prefix determines which candidate tokens receive a reward. It does not impose a vocabulary restriction on all generated text.

The regional and lexical branches are controlled independently for evaluation. The combined configuration enables regional adaptation and top-ten trie bias. The three-state geographic route controls contextual scope; acoustic ranking controls candidate identity, and token-level gates and trie rewards control the resulting intervention. Closing a gate at a later token cannot undo earlier hidden-state or text changes, so preservation requires direct measurement.

5 TRAINING RECIPE

Regional audio selection. Regional training selects representative utterances for each name from the unified training corpus, joins names to the geographic catalog, and applies the difficulty-first cell schedule above. The capacity-based sampler operates on 1,305,342 cell–utterance records. Appendix C also describes an optional allocation rule combining difficulty and popularity; the reported configuration uses difficulty-first selection.

Contextual-head stages. Span, retrieval, and ranking train sequentially with frozen acoustic features, reader, and catalog embeddings. Heads initialize from compatible task-trained checkpoints. Each stage supplies predicted evidence to the next: spans define retrieval queries, and retrieved candidates define ranking inputs. The head sampler combines POI contexts, general speech, and changed-geography stress contexts with a global batch of 256 source examples. Appendix C specifies optimization and model selection.

Table 1: **Overall training corpus and geographic coverage.** Corpus quantities are rounded; utterance counts and hours exclude repeated training presentations.

Quantity	Scale
Training utterances	$>$ 10 million
Audio duration	$>$ 16,000 hours
POI coverage	Approximately 4 million
Routing vocabulary	90,760 cells
Trained regional residuals	66,500 cells

Negative contexts and supervision. The sampler allocates 50% of presentations to POI contexts, 25% to general speech, and 25% to stress contexts. Stress contexts reuse audio under altered geography. General-speech filtering excludes transcripts matching place-name patterns to obtain negative span supervision; this heuristic is not exhaustive entity annotation. Unavailable purpose, role, relation, and referent labels are masked. Predicted spans supply downstream retrieval features.

6 Evaluation

6.1 Data scale and experimental scope

The unified training corpus combines recorded and synthesized speech, including synthesis conditioned on human voice references. It contains over ten million utterances totaling more than 16,000 hours and covers approximately four million POIs. Table 1 summarizes the overall scale. Regional and contextual-head training use task-specific sampling from this corpus (Section 5). Repeated geographic views and replayed presentations are not counted as additional recordings.

The evaluation set contains 10,967 utterances: 7,529 with annotated POI names and 3,438 without a POI, totaling 11.06 hours and 185,136 normalized reference characters. Audio-duration, catalog-coverage, and quality filters define a fixed set of utterances for the component comparison. Appendix D gives the filtering criteria, exclusion counts, and sensitivity analysis. Audio-hash checks find no duplicates within the evaluation set or overlap with the examined regional and contextual-head training inputs. Contextual heads are selected on separate validation data, and regional training uses four passes. The evaluation dataset will be publicly released with its audio, annotations, and geographic contexts.

Controls and scoring. We compare the same ASR backbone with neither enhancement, candidate bias only, regional adaptation only, and both. The regional pair carries the adapter in both arms: its bias-free arm is regional-only, not the original ASR. Audio IDs and references are matched across runs. Within each paired run the batch composition is fixed; the unadapted and regional runs use different shard counts, so cross-run differences also include batching effects. The enhanced configurations use identical candidates, candidate scores, acoustic inputs, biasing implementation, and decoding budgets. The comparison measures the individual and combined configurations of the two contextual branches; external contextual methods are not included.

The reported CER pools edit counts before division. References and hypotheses undergo NFKC normalization, case folding, and removal of Unicode separator and punctuation categories. Entity offsets are recomputed after normalizing annotated parts. "Names" is the percentage of POI-bearing utterances whose transcripts contain every normalized target name. It measures name containment, not aligned-span accuracy or physical-location identification. Appendix D defines scoring, exclusions, and uncertainty.

Geographic gating. With the geographic gate enabled, POI CER decreases from 14.07% to 5.24%, an 8.83-percentage-point decrease and a 62.8% relative reduction. Table 2 separately compares the regional and candidate components under fixed decoding configurations.

Table 2: **Evaluation-set results for the component comparison.** All values are percentages, with the same 10,967 IDs in every row. "Other" CER covers non-entity characters within the 7,529 POI-bearing utterances; "General" covers 3,438 POI-negative utterances. "Exact" uses all utterances, and "Names" uses POI-bearing utterances.

System	Whole CER	POI CER	Other CER	General CER	Exact	Names
ASR only	8.33	17.03	4.55	4.92	51.66	47.76
Candidate only	6.38	10.49	4.52	4.92	67.50	76.24
Regional only	7.99	16.10	**4.44**	4.87	53.50	50.27
Regional + candidate	**5.93**	**8.94**	4.57	**4.86**	**71.76**	**82.96**

Table 3: **Paired candidate-bias changes.** Each pair fixes its regional configuration and batch composition. Better/worse compare total character-edit counts. Correct-to-wrong counts previously exact transcripts made incorrect; the eligible baseline differs between pairs.

Candidate-bias comparison	Fewer errors	More errors	Correct → wrong
ASR → candidate	2,369	51	17
Regional → combined	2,618	86	27

6.2 Contribution of regional and candidate learning

In the component comparison, the combined configuration lowers whole-utterance CER from 8.33% to 5.93% relative to unmodified ASR, a 28.9% relative reduction; POI CER falls from 17.03% to 8.94%. All-target name containment increases by 35.20 percentage points. Candidate bias alone accounts for much of this gain. Adding regional adaptation further reduces whole CER from 6.38% to 5.93% and POI CER from 10.49% to 8.94%. In the directly paired regional run, enabling candidate bias reduces CER from 7.99% to 5.93%, while exact-transcript accuracy increases from 53.50% to 71.76%.

A 2,000-replicate paired source bootstrap gives a whole-CER change of -2.41 points $[-2.66, -2.09]$ for combined versus ASR-only, and -0.45 $[-0.68, -0.13]$ for combined versus candidate-only. These intervals describe variation within the scored test set; they do not capture shared voice/template dependence. Cross-run batching prevents attributing all cross-run differences solely to the adapter.

Preservation is not uniform. Other CER is 4.57% for combined versus 4.52% for candidate-only, while general CER is 4.86% versus 4.92%. Relative to candidate-only, combined improves 790 utterances, worsens 228, and makes 90 previously exact transcripts incorrect. Relative to ASR-only, it makes 79 exact transcripts incorrect, including 23 general utterances. These regressions accompany the aggregate recognition gains.

An auxiliary span-supervision evaluation shows localization recall at intersection-over-union (IoU) ≥ 0.5 increasing from 13.3% to 91.1% with negative-frame supervision. This is a span-localization diagnostic, separate from the recognition evaluation. On the 10,000-example contextual-head validation subset, shortlist recall is 73.22% and ranker top-ten recall is 66.96%. These denominators include deliberately incorrect or absent contexts and differ from those of the evaluation set. In the component comparison, 7,207 of 7,529 POI-bearing requests have a target in the shortlist (95.72%), and 6,615 have one in the biased top ten (87.86%). Combined POI CER is 5.24% in the latter group. For 592 requests with a shortlisted target outside the top ten, regional-only and combined POI CER are 24.67% and 28.72%; for 322 requests with no target in the shortlist, they are 35.60% and 36.25%. Candidate absence remains an identifiable failure mode.

NONE-threshold sensitivity. We also evaluate a fixed zero NONE-margin rule by selecting between the existing regional-only and combined transcripts. Whole CER changes from 5.928% to 5.909%, but POI CER increases from 8.94% to 9.05% and name containment declines from 82.96% to 82.44%. The paired whole-CER interval includes zero. This alternative does not improve POI CER or name containment; Appendix D reports the full comparison.

Training resources. Regional training uses twelve independent shards with four passes per shard. The assembled FP32 adapter occupies 19.19 GB, excluding the frozen backbone. Contextual-head training uses 16 GPUs; optimizer and batching settings appear in Appendix C. Online latency and real-time factor are not measured.

7 Limitations and Conclusion

Geographic context cannot recover a target excluded from the catalog or shortlist. Improvements in POI transcription can coexist with errors elsewhere, and the general-speech test subset does not represent all accents, recording devices, or spontaneous speech. Synthesized speech, including human-voice-conditioned synthesis, does not by itself establish robustness on human recordings. Audio-hash deduplication also does not establish speaker, prompt-family, or semantic independence. The component comparison does not isolate the effects of individual head objectives or routing thresholds; external-method comparisons and online latency measurements remain future work.

The framework combines cell-owned regional training, audio-conditioned three-state routing, and top-ten candidate biasing. The component results demonstrate complementary gains from regional adaptation and candidate biasing: combined whole CER is 5.93% versus 8.33% for unmodified ASR. With geographic gating enabled, POI CER decreases from 14.07% to 5.24%. The training recipe combines a unified large-scale speech corpus with task-specific sampling and geographically partitioned optimization.

AI Use Statement

An AI assistant was used to inspect implementation details, search and cross-check literature, draft and revise manuscript text, prepare figures, formulate the shared-subspace derivation, and provide feedback on method descriptions and evaluation design. The authors are responsible for verifying all AI-assisted technical content, citations, experimental evidence, and conclusions.

Ethics Statement

Speech and query locations can reveal identities and movement patterns. The intended application is user-directed transcription, and contextual location is not evidence of a speaker's identity. Relevant risks include unwanted insertion of place names and uneven accuracy across regions, accents, and popularity groups. A public POI catalog does not confer permission to collect or redistribute associated speech or location traces. The evaluation measures contextual errors; broader disparity analyses and data-use documentation remain necessary.

Reproducibility Statement

Our code and trained model weights will be publicly released. The code will be provided in a public repository containing training, inference, and evaluation scripts and configuration files; model weights will be provided as downloadable checkpoints. The evaluation dataset, including audio, annotations, and geographic contexts, will also be publicly released with the evaluation protocol. We will publicly document how to access the training data, but will not openly release the full training corpus. Reproducing the reported evaluation requires these resources, the specified ASR backbone and tokenizer, the POI catalog, and the evaluation audio, annotations, and geographic contexts. Reproducing training additionally requires the training corpus and alignment-derived supervision. Sections 4–5 and Appendices C–D specify model objectives, optimization, decoding, and scoring. Appendix E summarizes dependencies and reproduction settings.

REFERENCES

Anastasios Alexandridis, Kanthashree Mysore Sathyendra, Grant P. Strimel, Feng-Ju Chang, Ariya Rastrow, Nathan Susanj, and Athanasios Mouchtaris. Gated contextual adapters for selective contextual biasing in neural transducers. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095322. URL https://www.amazon.science/publications/gated-contextual-adapters-for-selective-contextual-biasing-in-neural-transducers.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Xun Gong, Anqi Lv, Wangyou Zhang, Zhiming Wang, Huijia Zhu, and Yanmin Qian. BR-ASR: Efficient and scalable bias retrieval framework for contextual biasing ASR in speech LLM. In *Interspeech 2025*, pp. 4043–4047, 2025. doi: 10.21437/Interspeech.2025-326. URL https://www.isca-archive.org/interspeech_2025/gong25_interspeech.html.

Haoxiang Hou, Xun Gong, Wangyou Zhang, Wei Wang, and Yanmin Qian. Ranking and selection of bias words for contextual bias speech recognition. In *Interspeech 2025*, pp. 5183–5187, 2025. doi: 10.21437/Interspeech.2025-646. URL https://www.isca-archive.org/interspeech_2025/hou25_interspeech.html.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Prakhar Kaushik, Ankit Vaidya, Shravan Chaudhari, and Alan Yuille. EigenLoRAx: Recycling adapters to find principal subspaces for resource-efficient adaptation and inference. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 649–659, 2025. doi: 10.1109/CVPRW67362.2025.00070. URL https://arxiv.org/abs/2502.04700.

YuXiang Kong, JunFeng Hou, Jian Tang, Bingqing Zhu, Jicheng Zhang, and Shaofei Xue. Contextual biasing for LLM-based ASR with hotword retrieval and reinforcement learning. *arXiv preprint arXiv:2512.21828*, 2025. URL https://arxiv.org/abs/2512.21828.

Sashi Novitasari, Takashi Fukuda, Gakuto Kurata, and George Saon. Contextual biasing for ASR in speech LLM with common word cues and bias word position prediction. *arXiv preprint arXiv:2604.12398*, 2026. URL https://arxiv.org/abs/2604.12398.

Golan Pundak, Tara N. Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. Deep context: End-to-end contextual speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 418–425, 2018. doi: 10.1109/SLT.2018.8639034. URL https://arxiv.org/abs/1808.02480.

Xian Shi, Yexin Yang, Zerui Li, Yanni Chen, Zhifu Gao, and Shiliang Zhang. SeACo-Paraformer: A non-autoregressive ASR system with flexible and effective hotword customization ability. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10346–10350, 2024. doi: 10.1109/ICASSP48485.2024.10446106. URL https://arxiv.org/abs/2308.03266.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*, 2026. URL https://arxiv.org/abs/2601.21337.

Chuanneng Sun, Zeeshan Ahmed, Yingyi Ma, Zhe Liu, Lucas Kabela, Yutong Pang, and Ozlem Kalinli. Contextual biasing of named-entities with large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10151–10155, 2024. doi: 10.1109/ICASSP48485.2024.10445918. URL https://arxiv.org/abs/2309.00723.

Guangzhi Sun, Chao Zhang, and Philip C. Woodland. Tree-constrained pointer generator for end-to-end contextual speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 780–787, 2021. doi: 10.1109/ASRU51503.2021.9687915. URL https://arxiv.org/abs/2109.00627.

Guangzhi Sun, Xianrui Zheng, Chao Zhang, and Philip C. Woodland. Can contextual biasing remain effective with Whisper and GPT-2? In *Interspeech 2023*, pp. 1289–1293, 2023. doi: 10.21437/Interspeech.2023-1440. URL https://www.isca-archive.org/interspeech_2023/sun23e_interspeech.html.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.239. URL https://aclanthology.org/2023.eacl-main.239/.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14299–14308, 2021. doi: 10.1109/CVPR46437.2021.01407. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yan_Positive-Congruent_Training_Towards_Regression-Free_Model_Updates_CVPR_2021_paper.html.

A NOTATION AND DIMENSIONS

Acoustic frames and autoregressive decoder positions use different indices. An acoustic span supplies retrieval evidence; it does not define a decoder-token mask. The symbols below describe a single request unless a sum over requests or cells is explicit.

Table 4: Notation used in the model, compression, and decision rule. Dimensions are symbolic unless an implementation value is explicitly stated.

Symbol	Meaning or dimension
$x, y, \hat{y};\ g$	Audio, reference text, predicted text; observed query geography.
τ, t, ℓ	Acoustic frame, decoder position, and configured decoder hook.
$s, c, \sigma;\ j, m$	Proposed acoustic span, routed cell, shard; candidate and gold-entity group. The index s differs from a scalar score.
$\mathcal{C}, \mathcal{C}(\mathcal{R})$	Catalog and its audio-routed geographic pool; a lexical surface can correspond to multiple physical POIs.
$H;\ d_{\mathrm{ac}}$	Shared projector-state sequence in Figure 1; projector width. H denotes a sequence, not a hidden dimension.
d_h, d_b, d_o	Decoder hidden width, adapter bottleneck width, and shared output rank.
h_t, z_t	Decoder state in $\mathbb{R}^{d_h}$ and bottleneck state in $\mathbb{R}^{d_b}$.
D, P, C_c	Input transform $d_b \times d_h$, output projection $d_h \times d_o$, and cell core $d_o \times d_b$.
$\mathcal{R}(x, g), \mathcal{G}(g)$	Audio-conditioned route and fixed observed-GPS neighborhood.
$w_c;\ C(g), R(g)$	Normalized geographic weights and mixed core/residual operators in Equation 5.
$R_c;\ \alpha, s_t$	Wide residual $d_h \times d_b$; residual scale and causal position gate. The gate value is recomputed at each hook.
G_σ, G	Shard and aggregate parameter Gram matrices, both $d_b \times d_b$.
$V_r, U_c;\ r_c$	Orthonormal row basis $r \times d_b$, projected cell factor $d_h \times r_c$, and a cell's retained rank.
$A_\tau, q_s;\ d_e$	Normalized reader frame and local query in $\mathbb{R}^{d_e}$; $d_e = 128$ in the implementation.
$w_{s\tau}, \epsilon$	Soft membership weight of frame τ for span s and positive smoothing constant in Equation 6.
W_q, v_q, b_q	Local-query projection $d_e \times d_e$, attention vector d_e, and projection bias d_e.
$K, B, \mathcal{S}$	Request-level retrieval budget, final surface budget, and accepted decoder surfaces ($\lvert\mathcal{S}\rvert \le B$).
$\mathcal{P}_m, \mathcal{N}_-; M$	Valid surfaces for present entity m, non-target surfaces, and number of entity groups represented in the actual pool.
$s_j, s_\varnothing$	Candidate and NONE scores within a stage. Retrieval and ranking have distinct scores and NONE heads.
$u_j;\ F_\phi, F_{\phi_0}$	Semantic and candidate-evidence vector; trainable scorer and frozen anchor in Equation 8.
$\mathcal{A}, \mathcal{D}_a$	Four intervention actions and the decoder under action a. $\mathcal{D}_a$ is distinct from matrix D.
$a_{\mathrm{reg}}, a_{\mathrm{cand}}$	Binary indicators enabling regional and candidate branches.
$p_e, p_b^a, p_h^a, \hat{g}_a$	Predicted entrance probability, action benefit/harm probabilities, and real-valued error reduction. Gain is not a probability.
τ_e, τ_b, τ_h	Entrance, benefit, and harm thresholds of the fixed policy.
$\mathcal{N}, E_a, g_a$	Scoring normalization, aligned error-location set, and observed normalized error reduction; g_a differs from geography g.
d_i, p_i, n_i	Frozen-probe difficulty, popularity percentile, and allocated clip quota for lexical entity i.

Vectors are column vectors in the equations; implementations may store batched row vectors. norm is ℓ_2 normalization with a numerical denominator floor, $\mathrm{softmax}_\tau$ excludes padded frames, and $[v]_+ = \max(v, 0)$. A detached score has its forward value preserved but receives no gradient through that occurrence. In the ranker, u_j combines utterance semantics with local support, global support, and coarse, reader, and teacher scores; the last three are divided by ten in the feature construction.

B Optional shared-basis compression

This appendix describes optional shared-basis compression of trained residual matrices. The reported recognition model retains the full bottleneck width; the projection analysis below does not constitute a compression-quality evaluation.

Numerical validation. Synthetic matrix experiments assess the numerical consistency of projection and shard aggregation. For two collections of 8×4 matrices, ranks 1, 2, and 4 reproduce the spectral-tail error identity to within 3×10^{-13} in squared-error units. The full-rank relative reconstruction error is 3.91×10^{-8} after FP32 basis storage. These checks establish numerical consistency of the operator calculation; they do not measure recognition quality after compressing the trained regional model.

Shared parameters across decoder layers. Let $\ell_1 < \cdots < \ell_L$ denote the configured decoder hooks, and let $h_{\ell t} \in \mathbb{R}^{d_h}$ be the hidden state before the regional addition at hook ℓ and decoder position t. A single parameter set is reused at every hook: $D \in \mathbb{R}^{d_b \times d_h}$, $P \in \mathbb{R}^{d_h \times d_o}$, $C_c \in \mathbb{R}^{d_o \times d_b}$, $R_c \in \mathbb{R}^{d_h \times d_b}$, and gate parameters $w_g \in \mathbb{R}^{d_h}$, $\beta_g \in \mathbb{R}$. For the selected cell c, inference applies

$$\begin{aligned} z_{\ell t} &= \mathrm{SiLU}(D\,\mathrm{LN}(h_{\ell t}))\,, \\ s_{\ell t} &= \mathrm{sigmoid}\big(w_g^\top\,\mathrm{LN}(h_{\ell t}) + \beta_g\big)\,, \\ h'_{\ell t} &= h_{\ell t} + a_{\mathrm{reg}} m_{\ell t} s_{\ell t}\,(b_\ell P C_c + \alpha R_c)\,z_{\ell t}, \qquad b_\ell = \mathbf{1}[\ell = \ell_1]. \end{aligned} \tag{B.1}$$

Here the scalar mask $m_{\ell t} \in \{0, 1\}$ restricts application to valid routes and text-prediction positions; a_{reg} is the request-level regional-action flag. Equation 4 describes active positions where $m_{\ell t} = 1$; invalid routes bypass the addition before a cell operator is accessed. The frozen base operator PC_c is applied only at the first hook, whereas the same residual R_c is applied at every hook. Shared weights do not imply identical activations or gate values across layers. Training additionally applies dropout after the bottleneck nonlinearity. For this residual-only optimization setting, the ASR backbone, D, P, C_c, and gate parameters remain frozen while cell-owned residuals are optimized. Gradients still propagate through subsequent frozen decoder layers. Section 4 instead jointly trains the adapter projections and cell parameters.

Optimal common-rank projection. Consider a fixed collection of trained residual matrices in the same input coordinates. Wide training uses a direct matrix R_c with a fixed identity right factor. Its trainable entries are initialized to zero before optimization. Compression seeks a common row-orthonormal basis $V \in \mathbb{R}^{r \times d_b}$, where $VV^\top = I_r$ and $1 \le r \le d_b$. For any fixed V, minimizing $\|R_c - U_c V\|_F^2$ gives $U_c = R_c V^\top \in \mathbb{R}^{d_h \times r}$. Thus the reconstruction is $\widetilde{R}_c = R_c \Pi$, with orthogonal projector $\Pi = V^\top V$. Writing $G = \sum_c R_c^\top R_c$ yields

$$\begin{aligned} J(V) &= \sum_c \|R_c(I - \Pi)\|_F^2 = \mathrm{tr}(G) - \mathrm{tr}(VGV^\top), \\ \min_{VV^\top = I_r} J(V) &= \sum_{i=r+1}^{d_b} \lambda_i(G), \qquad \lambda_1(G) \ge \cdots \ge \lambda_{d_b}(G) \ge 0. \end{aligned} \tag{B.2}$$

The trace identity uses $\Pi^2 = \Pi$. For orthonormal eigenvectors q_i of G, the trace is $\sum_i \lambda_i \|Vq_i\|^2$. Each weight lies in $[0, 1]$ and their sum is r, so the trace is at most $\sum_{i=1}^r \lambda_i$. Taking the leading eigenvectors as rows of V_r attains this bound. The Gram is uncentered and gives each cell matrix equal weight, without popularity weighting. Equal weighting does not normalize residual norms: larger operators can contribute more spectral energy. If the cutoff eigenvalue is repeated, the minimizing subspace need not be unique; sign and eigenspace conventions specify a reproducible representative. At full bottleneck rank, the projection reconstructs every operator exactly in real arithmetic; lower ranks discard precisely the aggregate spectral energy in the omitted directions.

Scope of the compression result. For disjoint shards covering this fixed collection, $G = \sum_\sigma G_\sigma$ follows directly from addition. Compatible shards share the frozen base, input transforms, dimensions, vocabulary interpretation, and hook semantics. This identity concerns already trained matrices; it does not equate independently scheduled optimization with joint training. Each shard

contributes a $d_b \times d_b$ statistic, allowing basis extraction without loading all regional matrices simultaneously. The original matrices are still required when projecting each shard into the selected basis. Accumulation and eigendecomposition use FP64, followed by FP32 basis storage, so floating-point order and eigenspace conventions remain relevant. With heterogeneous cell ranks r_c, deployment uses prefixes V_{r_c} of one ordered basis. Each coefficient matrix is optimal for its specified prefix, but this construction need not jointly minimize reconstruction error over all nested bases and mixed ranks. Moreover, parameter reconstruction error is not transcription loss: compression quality must be assessed by decoding, including harmful interventions.

Parameter and storage accounting. Let N count cells with a base core and n count cells with an additional residual. Put $S_r = \sum_{c=1}^{n} r_c$ and $r_{\max} = \max_c r_c$. The shared base has

$$Q_{\text{base}} = d_h d_b + d_h d_o + N d_o d_b + d_h + 1. \tag{B.3}$$

These terms count the input transform, output projection, core table, and affine scalar gate. Layer normalization has no affine parameters, and the input transform has no bias. The residual counts differ between optimization and serialization:

$$\begin{aligned} Q_{\text{wide,train}} &= n d_h d_b, \qquad Q_{\text{wide,stored}} = n(d_h d_b + d_b^2), \\ Q_{\text{factorized,stored}} &= (d_h + d_b) S_r, \qquad Q_{\text{deduplicated}} = d_h S_r + r_{\max} d_b. \end{aligned} \tag{B.4}$$

The full-width checkpoint stores a frozen identity matrix per residual cell. The factorized checkpoint format similarly stores each cell's U_c and its own copy of V_{r_c}, even though the right factors share a subspace. The final expression assumes a different serialization that stores the common basis once; it is not the factorized format's footprint. Add Q_{base} to the relevant stored count for the adapter total. There is no factor of L, because all hooks reuse these parameters. FP32 tensor payloads require four bytes per element; file headers, metadata, and configuration add overhead. Frozen ASR weights, optimizer states, and runtime activations are excluded from these adapter counts and must be measured separately.

C Training and inference details

C.1 Optimization settings

The evaluated regional configuration freezes Qwen3-ASR and trains the adapter's input/output projections, cell core table, causal scalar gate, and cell-owned residual left factors. Right factors are fixed 32×32 identities. All shards share initialization seed 20260925 and a global negative-assignment plan with data seed 20260926. Each shard uses eight data-parallel ranks, batch two per rank, and no gradient accumulation, giving 16 logical records per optimizer step. The twelve shards complete four passes each. Separate AdamW groups use learning rates 2×10^{-4} for residuals and 3×10^{-4} for the adapter base, cosine decay, 2.5% warm-up, zero weight decay, gradient norm cap 0.5, and dropout 0.05. Adapter parameters are FP32 and the frozen ASR runs in BF16. The adapter has one decoder hook, full bottleneck width 32, and shared output width 128.

Each logical regional record includes correct-, wrong-, and absent-context target views plus random- and absent-context general views. General examples have empty entity-span masks, so they do not directly supervise the regional residual or its gate under this configuration. Their role in contextual-head training is negative span and candidate supervision. General-speech recognition quality is evaluated directly.

For contextual heads, the ASR, projector, reader, and catalog embeddings stay frozen. The span head is selected first, then retrieval learns from its predicted spans, and ranking learns from the selected retriever's outputs. The global fallback and ranker teacher remain frozen. The head sampler uses seed 20260923, a global batch of 256 source examples, AdamW weight decay 0.01, and a 50/25/25 POI/general/stress allocation. Head training uses 48 distributed ranks across 16 GPUs.

A diagnostic comparison with and without negative-frame supervision shows its effect on span localization: recall at IoU ≥ 0.5 rises from 0.133 to 0.911, while the predicted positive-frame fraction falls from 98.7% to 26.7% (the annotated fraction is 22.3%). This diagnostic measures span localization; the head-selection criterion and evaluation-set recognition metrics measure different quantities.

Table 5: **Head optimization and model selection.** The two recall values include stress contexts; they are not evaluation-set recognition metrics.

Stage	Learning rate	Selection criterion	Selected value
Span	10^{-3}	Source-mean loss	0.07572
Retrieval	3×10^{-4}	Bounded shortlist recall	0.73217
Ranking	3×10^{-4}	Ranker top-10 recall	0.66959

Span proposals use membership threshold 0.5, at most four intervals, two-frame padding, and a 0.05 membership floor. The hidden width is 128 on 2,048-dimensional projector features. Semantic labels absent from the training manifests are masked. Reference-derived alignment uses a 0.15-second gap tolerance to reconnect nearby attention peaks and retains confidence filtering; these boundaries are heuristic rather than human-verified. The ranker uses residual scale $\eta = 4$, with unbounded residual updates and trainable fusion.

Component-comparison decoding settings. The four component configurations use BF16, SDPA attention, deterministic greedy generation, batch size eight, repetition penalty 1.1, and at most 128 new tokens. Candidate bias uses a top-ten prefix trie, base reward 8, relative-score scale 0.25, start-reward scale 0.5, continuation priority, and top-rank boosts 20 and 12. The regional run uses residual scale 1 and a 3×3 geographic neighborhood. In these component comparisons, the optional trie-derived token-position gate is disabled; it is distinct from the audio cell gate used for geographic routing. Both arms of the regional pair carry the adapter. The unadapted pair uses four decode shards and the regional pair sixteen. No complete online latency or real-time-factor claim is made from these offline runs.

Compute and parameter storage. Regional optimization uses twelve independent shards, eight data-parallel ranks per shard, and four passes through each shard's difficulty-based schedule. Each logical record produces five context roles. Cell-owned factors are assembled with shared parameter averaging as described in Section 4. The FP32 adapter occupies 19.19 GB (decimal), excluding the frozen ASR weights. This storage figure is for the full-width adapter; compressed-model quality and end-to-end latency require separate measurement.

C.2 Three-state geographic routing

The audio cell gate consumes projected audio states and GPS-local cell identities, before final autoregressive decoding. It does not require a transcript or a retrieved POI name as an input. Its output sets the geographic scope of the deployment path; this differs from the later lexical ranker's NONE score and the separate four-action benefit predictor.

Let l_c and l_0 be the valid trained-cell and NONE logits, respectively, and let n be the number of valid trained candidates. The application score is $p_{\text{apply}} = \text{sigmoid}(\text{logsumexp}_c\, l_c - \log n - l_0)$. The two margins are $l_{(1)} - l_0$ and $l_{(1)} - l_{(2)}$. A route is high confidence when all three high thresholds are met. Otherwise, it is medium confidence when all three relaxed thresholds are met; each relaxed threshold is no greater than its high counterpart. With one valid candidate the top-two margin is infinite; without one the policy rejects. Thresholds are calibrated for the specified gate and ASR model.

Table 6: **Three-state deployment routing.** Audio confidence controls geographic specificity and contextual admission.

Confidence	Geographic scope	Contextual behavior
High	One audio-selected cell	Retrieve and adapt within that cell
Medium	Observed GPS 3×3	Use the fixed fuzzy neighborhood
Low	Empty; hard-null	No retrieval, bias, or regional residual

The deployment interface passes the chosen cell set to candidate retrieval and the corresponding route weights to regional adaptation. The high-confidence state concentrates context on one cell;

the middle state retains local coverage when the audio does not distinguish cells reliably. Low-confidence requests execute the unmodified ASR branch. A common cell vocabulary keeps the gate, retrieval pool, and regional adapter aligned under the same geographic interface. This compatibility is an architectural property; recognition gains remain those measured under the stated evaluation controls.

C.3 Alternative difficulty–popularity allocation

An alternative allocation rule combines difficulty and popularity; the reported model uses the difficulty-based schedule. It defines difficulty from a frozen probe's strict entity recall ρ_i:

$$d_i = \begin{cases} \mathrm{clip}(1-\rho_i, 0, 1), & \rho_i < 1, \\ 0.25 \min(\mathrm{CER}_i, 1), & \rho_i = 1. \end{cases} \tag{C.1}$$

Here probe CER is a ratio, not a percentage. The popularity value is the maximum joined annual page-view count for a lexical entity; missing counts are zero. For N_{ent} scored entities, zero-based ascending ranks divided by $\max(N_{\mathrm{ent}} - 1, 1)$ yield p_i. Ties are ordered by lexical identifier, except that a constant popularity vector gives $p_i = 0$ for all entities. Thus tied counts need not otherwise have identical percentiles.

The alternative quota is $n_i = \mathrm{clip}_{[n_{\min}, n_{\max}]}(n_{\min} + \mathrm{round}((n_{\max} - n_{\min})(w_d d_i + w_p p_i)))$. It uses starting weights $(w_d, w_p) = (0.6, 0.4)$ and quotas $(n_{\min}, n_{\max}) = (1, 16)$. Rounding follows nearest-integer, ties-to-even semantics. Clips whose zero-based within-entity rank is below n_i are retained; unscored entities retain all available clips. A quota cannot create absent or invalid audio. Distinct retained clips, actual training presentations, and logical replay counts therefore require separate counts. The separate three/two/one difficulty-only replay scheme repeats training exposure and does not itself supply new acoustic diversity.

C.4 Optional action policy and missing resources

The separate outcome-based policy, which is not used for Table 2, has the following order:

1. Extract shared acoustic states, predicted spans, semantics, and geographic pool metadata. Evaluate p_e from pre-retrieval features. If $p_e < \tau_e$, dispatch NULL.
2. Otherwise, retrieve at most K distinct surfaces using the shared local/global budget. Rank them and retain at most B surfaces strictly above $s_\varnothing + \delta_{\mathrm{none}}$.
3. Route a trained cell using frozen-reader support over the full shortlist. Construct the evidence for each of the three enhancement actions, including its two branch indicators.
4. Retain actions with $p_b^a \geq \tau_b$, $p_h^a \leq \tau_h$, and $\hat{g}_a > 0$. Select the largest predicted gain, or NULL if none remain. Exact gain ties follow a fixed deterministic order: regional, candidate, then both.
5. Run the selected decoder once. NULL calls the base decoding path; other actions enable their specified branches.

An empty span set leaves the global queries available. Without GPS or explicit geographic context, the geographic scope is empty rather than expanded to a national search. The router returns no cell if the shortlist supports no trained module; this does not cancel candidate retrieval. Conversely, an empty accepted set $\mathcal{S}$ does not imply that no regional route exists, because routing uses the full shortlist.

The optional action policy evaluates all three action records and has no additional hard action-availability mask. Missing-resource behavior is part of the fixed decoder implementation, and an enabled branch can have no effective contribution. Consequently, paired labels and evaluation must use the same missing-resource behavior. Structural NULL bypass also does not establish numerical parity across different precision, kernels, batching, or padding; a parity check uses the same deployed decoding configuration.

D SCORING AND EVALUATION CONVENTIONS

Final test-set protocol. The evaluation set contains 12,000 utterances before eligibility filtering. Audio preparation excludes 794 items longer than 45 seconds and another 214 exceeding 0.75 seconds per reference character. Seventeen further items have no matching surface in the fixed catalog and are excluded by catalog eligibility. These are distinct filtering criteria; a missing catalog entry is not evidence of corrupt audio. The component-comparison predictions cover the remaining 10,975 items.

The final scoring protocol applies one further quality filter to the normalized regional-only reference r and hypothesis h:

$$|h| > \max(24, 1.8|r|) \quad \text{and} \quad \max_{v} \frac{\#\{t : h_t = v\}}{|h|} > 0.25. \tag{D.1}$$

It flags eight POI-bearing items. Their IDs are fixed once and excluded from every compared arm, leaving 10,967 test utterances. The filter detects repetitive decoding; its output is not itself an independent annotation of input-audio corruption. We therefore also report all 10,975 items under the same text normalization (Table 7). Recomputing a separate filter for each model would change the comparison set and is not the protocol used here.

Text normalization and denominators. References, hypotheses, and target names undergo Unicode NFKC normalization, case folding, then removal of characters whose Unicode category starts with Z (separator) or P (punctuation). Leading and trailing whitespace is stripped by the transcript scorer. The decoder processor first extracts the transcription from its generated output. This rule does not remove all formatting controls or symbols, convert simplified/traditional scripts, or equate spoken and written address numbers.

Each annotated reference part is normalized before assigning entity offsets. Their concatenation equals the normalized whole reference for every item. The final set contains 185,136 reference characters: 55,007 entity characters, 92,357 other characters within POI-bearing utterances, and 37,772 characters in general utterances. Whole CER pools errors before division. Exact accuracy compares normalized whole transcripts; name containment requires every normalized annotated target string to occur in the hypothesis. The latter is an all-target substring measure, not aligned-span exact accuracy or physical-POI identification. Candidate-availability buckets retain the original shortlist target indices and decoded top-ten indices; text normalization does not redefine candidate membership.

Table 7: **Sensitivity to the final eight-item exclusion.** Whole-utterance CER (%) uses identical normalization in both columns and identical IDs across systems within each column.

System	All decoded items (10,975)	Final protocol (10,967)
ASR only	8.537	8.333
Candidate only	6.585	6.378
Regional only	8.199	7.990
Regional + candidate	6.139	5.928

Prediction checks and preservation. For the component comparison, every configuration contains the same 10,975 pre-filter IDs without duplicates. Paired arms agree on audio and references. Per-example acoustic tensors, candidates, and scores agree across the unadapted and regional runs, while batch membership and some padding widths differ. Evaluated audio hashes are disjoint from the training manifests included in the overlap check.

On the final scoring set, candidate-only introduces a new aligned error location in 85 requests relative to ASR-only; combined does so in 306. Relative to candidate-only, combined has 273 such requests, including 151 with new entity errors and 126 with new non-entity errors; those categories can overlap. Within the regional pair, general-speech edit counts change from 1,838 to 1,836, despite six improving requests, six worsening requests, and four correct-to-wrong requests. Aggregate preservation and absence of individual regressions are different properties.

Paired percentile intervals use 2,000 source-ID bootstrap samples with seed 20260926. The 10,967 source clusters are singletons, so resampling cannot capture shared voice or prompt-template dependence. For combined versus candidate-only, POI CER changes by -1.56 points $[-2.12, -0.64]$, other CER by $+0.05$ $[-0.16, +0.37]$, and name containment by $+6.72$ $[+6.13, +7.30]$. These intervals are conditional on the scoring protocol and fixed exclusion list.

Fixed NONE-margin diagnostic. An offline selector uses the combined transcript only if the best ranker candidate score exceeds its NONE score; otherwise it retains the regional-only transcript. The zero threshold is fixed, without fitting to test outcomes. It admits 9,129 of 10,967 requests. All features and both transcripts are already computed, so this experiment establishes neither reduced online latency nor a trained benefit gate.

Table 8: **Offline selection with a fixed zero NONE margin.** CER and accuracy are percentages. New-error and correct-to-wrong counts are relative to regional-only.

Policy	CER	POI CER	Exact	Names	New errors	Correct → wrong
Always regional	7.990	16.102	53.50	50.27	0	0
Always combined	5.928	8.935	71.76	82.96	122	27
NONE margin > 0	5.909	9.052	71.47	82.44	113	25

Relative to always-combined, the selector changes whole CER by -0.019 points $[-0.151, +0.053]$ while worsening POI CER by $+0.116$ $[+0.076, +0.159]$. It improves eight utterances but worsens 51 and turns 34 exact transcripts into errors. Thus its slight net edit-count advantage does not demonstrate uniform improvement. The optional outcome-based action policy described below is distinct from both this diagnostic and the geographic gate.

Separate action-label normalization. The optional paired-action pipeline uses a stronger $\mathcal{N}$: supported Chinese/Arabic address-number equivalence, ASR-prefix removal, NFKC normalization, case folding, and removal of punctuation, symbols, controls, and whitespace. This normalization is not applied to the measured recognition tables and must not be mixed with their error counts. The following definitions describe that optional pipeline unless stated otherwise.

Deterministic alignment. Reference $\bar{y} = \mathcal{N}(y)$ and hypothesis $\mathcal{N}(\hat{y})$ are aligned with unit-cost Levenshtein distance. Equal-cost traceback choices prefer the diagonal move, then deletion, then insertion. A substitution or deletion at zero-based reference index i contributes the same slot $(\mathrm{ref}, i, 0)$ to the error set E. An insertion at reference gap i contributes (ins, i, k), where k distinguishes successive insertions at that gap during traceback. Thus $|E|$ equals the edit count, including multiple insertions, although slots do not retain erroneous character identities. A substitution/deletion inherits its reference character's entity annotation. An insertion inherits the preceding reference character, or the first character for a leading insertion; an empty reference has no entity positions.

Paired action labels. For the same reference, let E_0 and E_a describe NULL and action a, and let $e_{\mathrm{poi}}(E)$ count slots assigned to entity positions. The labels are

$$\begin{aligned} H_a &= \mathbf{1}[E_a \setminus E_0 \neq \varnothing], \\ B_a &= \mathbf{1}[e_{\mathrm{poi}}(E_a) < e_{\mathrm{poi}}(E_0)](1 - H_a), \\ g_a &= \frac{|E_0| - |E_a|}{\max(1, |\bar{y}|)}. \end{aligned} \tag{D.2}$$

The entrance label is the logical OR of B_a over the three forced enhancement actions. Benefit requires entity repair without newly erroneous alignment locations; a net CER reduction alone is insufficient. Gain is a signed error-rate difference, not a probability, and can exceed one because of insertions. Conversely, replacing one wrong character by another at an already erroneous location need not produce $H_a = 1$. Alignment changes can also relocate slots. These labels measure a reproducible positional notion of harm, not semantic damage or every newly hallucinated catalog name; lexical insertion checks remain separate.

Recognition denominators. For partition $q \in \{\text{all}, \text{poi}, \text{nonpoi}\}$, corpus-micro CER is

$$\text{CER}_q = 100 \frac{\sum_u e_{uq}}{\sum_u n_{uq}}, \tag{D.3}$$

Here u indexes requests, e_{uq} is the aligned error count, and n_{uq} the normalized reference-character count. Insertions contribute errors but no reference characters. Counts are pooled before division; utterance CERs are not averaged. Zero-character partitions contribute no denominator and an entirely empty partition has no defined corpus CER. For the optional normalized scoring path, exact lexical recovery checks whether the normalized canonical mention occurs as a substring of the normalized output. Alias-expanded matching, if reported, is a separate measure. Text matching cannot distinguish physical POIs sharing a name. For repeated or multiple mentions, the manifest fixes whether the denominator counts occurrences or distinct target groups; one substring match does not establish recovery of every occurrence. Unconditional candidate recall includes all reference targets, whereas pool-conditional recall includes only targets available in the geographic pool. Shortlist recall precedes gating; admitted inclusion and final mention recovery are distinct outcomes.

Request and source-cluster harm. For the optional action analysis, "new harm" is the percentage of all evaluated requests with $H_{a^*} = 1$, assigning zero to NULL. Correct-to-wrong instead conditions on a completely correct NULL transcript and counts requests made erroneous by the selected action. It is neither the same numerator nor the same denominator. Multiple GPS views of one audio source are dependent. For uncertainty, all related views remain in one source cluster; cluster harm means at least one harmful request. The conditional cluster denominator contains clusters with at least one baseline-correct request. The implemented development constraint uses the general-speech *correct-to-wrong cluster upper bound*, alongside a retrieval-rate budget, rather than the request-level new-harm percentage.

Uncertainty and scope. For m affected clusters among n eligible clusters and failure probability $\delta \in (0, 1)$, the implemented upper summary is

$$U(m, n; \delta) = \begin{cases} 1, & n = 0, \\ 1 - \delta^{1/n}, & n > 0,\ m = 0, \\ \min\left(1, \frac{m}{n} + \sqrt{\frac{\log(1/\delta)}{2n}}\right), & m > 0. \end{cases} \tag{D.4}$$

The zero-event expression is the binomial zero-event bound; the nonzero branch uses Hoeffding's inequality. Their interpretation requires a fixed policy and independent clusters representative of the assessed population. A valid future threshold-search protocol must use development data and reserve independent calibration and test sets. Repeated selection or dependent clusters prevent an automatic confidence interpretation, and these summaries do not establish unmeasured safety guarantees. Paired bootstrap resamples whole clusters with replacement, retaining their baseline/system error and character totals, and recomputes the count ratio for each draw rather than averaging per-cluster CERs. This preserves within-cluster pairing without treating repeated geographic views as independent observations.

E REPRODUCIBILITY SETTINGS

The GeoPOI code will be publicly released in a repository containing training, inference, and evaluation scripts and configuration files. Trained model weights will be publicly released as downloadable checkpoints. The evaluation dataset, including audio, annotations, and geographic contexts, will also be publicly released with the evaluation protocol. We will publicly document how to access the training data, but will not openly release the full training corpus. The repository will identify the required backbone, tokenizer, software dependencies, and model-loading procedure. Full numerical reproduction also requires the evaluation audio, annotations, geographic contexts, and POI catalog; training reproduction requires the training corpus and its supervision. Third-party backbone weights and training resources remain subject to their respective redistribution terms.

Reproducing GEOPOI requires a consistent acoustic backbone, tokenizer, projector, reader, contextual heads, geographic catalog, and regional adapter. Sections 4–5 specify their interfaces, objectives,

and training dependencies; Appendix C gives seeds, optimizer settings, precision, batch sizes, and model-selection criteria.

The geographic catalog defines cell membership and candidate surfaces. Audio annotations provide target names and alignment-derived spans. Data preparation retains source identifiers for overlap checking and records whether speech is recorded or synthesized. Collection and training permissions are distinct from redistribution rights.

Inference fixes the geographic route, shortlist budget K, surface budget B, ranking policy, trie rewards, residual scale, and missing-context behavior. The component comparison evaluates the four combinations of regional and candidate enhancement on identical audio and references. Appendix D specifies text normalization, valid test items, metric denominators, and bootstrap uncertainty. Numerical reproduction also depends on the precision, attention implementation, and batching settings in Appendix C.

Adapter storage excludes the frozen ASR backbone. A deployment latency measurement would additionally include feature extraction, cell routing, retrieval, ranking, and final decoding under stated cache and batching conditions. Such latency and the optional compression and learned action-policy extensions are outside the reported recognition comparison.

F Optional outcome-based action selection

This optional extension introduces learned entrance and benefit/harm predictors, distinct from the audio cell gate used for geographic routing. These outcome-prediction heads are not used in the reported recognition results; their objectives and validation requirements are described below.

The entrance head uses information available before acoustic candidate scoring: mention and span evidence, semantic probabilities, GPS availability and accuracy, and local pool size. Its target is whether *any* available enhancement would produce a beneficial outcome. It does not use the shortlist margin as a free pre-retrieval feature.

After retrieval, a rule-based router selects the available cell with the strongest reader-supported candidate, without accumulating votes from duplicate stores. The action head additionally reads acoustic compatibility, ranking and NONE margins, regional support, boundary evidence, and two flags identifying the proposed enhancement branches. One shared network predicts three values for each non-NULL action: benefit probability, harm probability, and normalized error reduction. Reference transcripts are used only to create offline labels.

To construct supervision, one decodes each audio under NULL and all three enhancement actions, including inputs rejected by the reference policy. Under fixed normalization $\mathcal{N}$, let E_0, E_a be aligned error-location sets. Harm is $\mathbf{1}[E_a \setminus E_0 \neq \emptyset]$; benefit requires fewer entity errors and no newly erroneous locations. The gain is $g_a = (|E_0| - |E_a|)/\max(1, |\mathcal{N}(y)|)$. These labels do not detect changed wrong characters at an already erroneous position (Appendix D). Entrance uses binary cross-entropy on any-beneficial labels. The action loss averages benefit/harm binary cross-entropies and adds $0.2\,\mathrm{SmoothL1}(\hat{g}_a, g_a)$.

For entrance score p_e and per-action predictions $(p_b^a, p_h^a, \hat{g}_a)$, the policy selects

$$a^* = \arg\max_{a \in \mathcal{A}_{\mathrm{ok}}} \hat{g}_a, \quad \mathcal{A}_{\mathrm{ok}} = \{a \neq \textsc{Null} : p_e \geq \tau_e,\ p_b^a \geq \tau_b,\ p_h^a \leq \tau_h,\ \hat{g}_a > 0\}, \tag{F.1}$$

and returns NULL if the set is empty. Thresholds must be selected on development data subject to retrieval and general-speech harm budgets, then assessed on independent calibration data. Score thresholds alone are not a formal population safety guarantee. Changing adapter weights, compression, routing, or decoder bias changes the action and requires corresponding labels and calibration to be refreshed.

Two levels of preservation. Request-level NULL dispatch structurally bypasses both enhancement branches and can be checked against the same backbone decoder. When an action is active, preservation remains an empirical property. In particular, closing a token-level position gate later in an utterance does not undo earlier token or hidden-state changes. We therefore evaluate non-entity errors and previously correct utterances directly.