ANCHORSLU: CROSS-LINGUAL SEMANTIC ANCHORING FOR LOW-RESOURCE SPEECH UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual speech foundation models have advanced rapidly, yet large resource-dependent performance gaps persist when target-language supervision is limited. Standard cross-entropy (CE) provides only *class-level supervision*: semantically different utterances within the same class receive the same target, leaving exact sentence-level cross-lingual correspondence unused. We introduce ANCHORSLU, which converts this correspondence into *representation-level supervision*. For each target utterance, we synthesize sentence-matched reference speech in eight high-resource languages, encode these reference views offline with a frozen speech backbone, and aggregate selected-layer representations into sentence-specific semantic anchors. During adaptation, the model learns from natural target speech using target-only CE and multi-layer anchor alignment; inference requires only the adapted target model. Across 20 low-resource SIB-Fleurs languages, ANCHORSLU improves Qwen2.5-Omni-7B over matched target-only CE by **4.39** points in Accuracy and **5.04** points in Macro-F1. On 60-intent Speech-MASSIVE, it improves Accuracy by **3.19** points and Macro-F1 by **3.52** points under substantially richer target supervision. The benefit further extends to a substantially larger speech backbone. Controlled comparisons show that removing exact sentence correspondence largely eliminates the benefit even when class identity is preserved, while direct TTS augmentation and output-level distillation remain weaker. Together, these results establish exact sentence-level cross-lingual correspondence as an effective source of representation-level supervision beyond class labels, suggesting that sentence-matched multilingual supervision may also benefit broader speech understanding settings where such correspondence is available.

1 INTRODUCTION

Multilingual speech foundation models now support speech recognition, translation, and instruction following across an increasingly broad set of languages (Babu et al., 2022; Radford et al., 2023; Xu et al., 2025a). Yet coverage does not imply comparable understanding: with limited target-language supervision, the same model can remain substantially weaker in low-resource languages than in high-resource ones. For multilingual speech understanding, this disparity raises a central question: **How can knowledge be transferred from high-resource to low-resource languages?**

This gap is already visible before target adaptation (Figure 1). Under the same frozen Qwen2.5-Omni-7B backbone, prompt, and seven-topic task, the eight reference languages average 83.2% Accuracy, versus 45.0% across the 20 low-resource target languages. This asymmetry suggests that useful task information is already expressed elsewhere within the same multilingual model. The challenge is to turn this cross-lingual information into effective supervision for natural target speech.

Target-only CE is too coarse for this purpose. It supervises class identity: semantically different utterances receive the same target whenever they share a label. Parallel data provide a finer relation by identifying utterances across languages that express the same sentence content. Sentence-level cross-lingual correspondence therefore provides supervision that class labels alone do not encode.

We introduce ANCHORSLU to realize this supervision at the representation level. Sentence-matched reference speech is encoded offline with a frozen speech backbone and aggregated across reference

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.

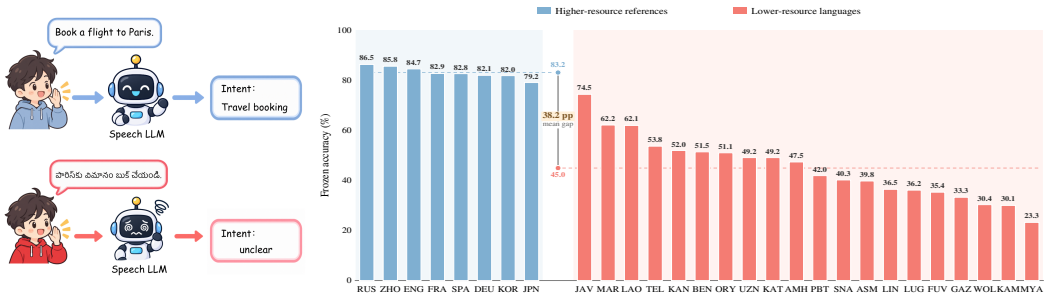

Figure 1: **Resource gap under a shared frozen model and task. Left:** Illustrative SLU predictions for higher- and lower-resource speech. **Right:** Frozen Qwen2.5-Omni-7B Accuracy on the same seven-topic task across eight reference and 20 target languages; group means are 83.2% and 45.0%.

languages at selected layers into sentence-specific semantic anchors. The adapted target model then learns from target-only CE and multi-layer anchor alignment, while inference uses only the adapted target model. The matched sentence, rather than the class itself, thus becomes the unit of cross-lingual supervision.

Across the 20-language SIB-Fleurs panel, all three ANCHORSLU runs outperform matched target-only CE in the language macro, and their mean improves both Accuracy and Macro-F1 for every target language. The gains persist under substantially richer target supervision on 60-intent Speech-MASSIVE and extend to the substantially larger Qwen3-Omni backbone, where ANCHORSLU improves matched target-only CE across every reported language group. Crucially, breaking exact sentence correspondence removes most of the benefit even when class identity is preserved, while direct TTS augmentation and output-level distillation remain weaker. These results localize the main gain to exact sentence correspondence and its representation-level use.

Our contributions are threefold:

- **Cross-lingual semantic anchoring.** We introduce ANCHORSLU, which converts exact sentence-level cross-lingual correspondence into sentence-specific semantic anchors for multi-layer representation-level supervision, with all reference-side computation performed offline.
- **Evaluation across resource regimes and backbone scales.** ANCHORSLU consistently improves over matched target-only CE across 20 low-resource SIB-Fleurs languages and seven Speech-MASSIVE target languages, spanning scarce and richer supervision, and the benefit extends to the larger Qwen3-Omni backbone.
- **Controlled analysis of the supervision signal.** Correspondence, transfer-form, and anchor-design controls show that the gain depends on exact sentence matching and representation-level supervision rather than class agreement, synthetic-data exposure, or output-level distillation alone.

2 RELATED WORK

Multilingual speech models and resource gaps. Multilingual speech models have rapidly expanded in both capability and language coverage, from self-supervised systems such as XLS-R to large-scale models such as Whisper, Qwen2.5-Omni, and Qwen3-Omni (Babu et al., 2022; Radford et al., 2023; Xu et al., 2025a;b). However, performance in low-resource languages remains strongly affected by target-language data availability and multilingual pretraining coverage (Manakul et al., 2025; Fong et al., 2025), and zero-shot SLU continues to exhibit substantial cross-language variation (Lee et al., 2026). Increasing model scale and language coverage alone therefore does not eliminate resource-dependent disparities in multilingual speech understanding.

Cross-lingual SLU and representation alignment. Cross-lingual SLU has addressed scarce target-language annotation through representation regularization, contrastive learning, multilingual or code-switched views, and representation- or prediction-level alignment (Liu et al., 2020; Liang et al., 2022; Qin et al., 2022; Zhu et al., 2023; Xing & Tsang, 2024; Zhu et al., 2024; Cheng et al., 2024). Recent methods further model inter-language information flow, slot correspondence, and multilingual

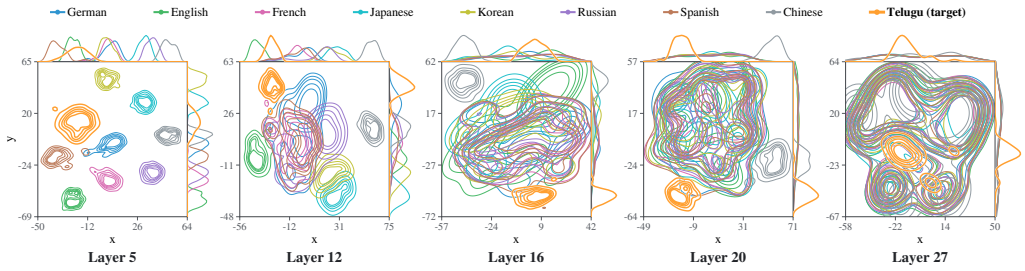

Figure 2: **Frozen target–reference representations.** Layer-wise t-SNE of Telugu target speech (orange) and eight sentence-matched reference languages. Target and reference views are separated in early and intermediate layers and increasingly interleaved in deeper layers; coordinates are fitted independently per panel.

semantic interaction (Wu et al., 2025; Xing et al., 2026), while multi-grained speech–text alignment has also been explored for end-to-end SLU (Zhu et al., 2022). ANCHORSLU takes a different view of the supervision unit: exact sentence-level cross-lingual correspondence is used to construct sentence-specific semantic anchors that directly supervise natural target-speech representations.

Synthetic speech and knowledge transfer. Synthetic speech is commonly used as additional labeled training data (Peng et al., 2021; Zhong et al., 2022; Robinson et al., 2022), with its effectiveness depending on factors such as natural-speech availability and the synthetic-to-real ratio (Hao et al., 2026). Knowledge distillation instead transfers teacher information through output distributions or intermediate representations (Hinton et al., 2015; Romero et al., 2015) and has also been applied to low-resource cross-lingual SLU (Mao & Zhang, 2023). In ANCHORSLU, synthesized reference speech serves a different role: it is used only in offline reference-side computation to construct fixed sentence-specific semantic anchors. It is neither treated as classification data for the adapted target model nor used as an output-level distillation target.

3 CROSS-LINGUAL SEMANTIC ANCHORING

At the core of ANCHORSLU is a *sentence-specific semantic anchor*: a multilingual reference point representing the shared content of a single training utterance across high-resource languages. ANCHORSLU constructs these anchors offline from sentence-matched reference speech and uses them to guide the adaptation of natural target speech. Semantic correspondence alone does not guarantee representation alignment: as shown in Figure 2, target and reference utterances expressing the same sentence remain separated in early and intermediate frozen representations, becoming increasingly interleaved only in deeper layers while retaining distinct local structure. ANCHORSLU therefore makes this existing semantic correspondence explicit in the representation space by aligning natural target-speech representations with their corresponding anchors (Figure 3).

3.1 SENTENCE-MATCHED REFERENCE SPEECH CONSTRUCTION

We consider a target language q with N_q labeled natural training utterances. Each example contains natural target speech $a_i^{(q)}$, its task label y_i, parallel text e_i, and a sentence identity s_i:

$$\mathcal{D}^{(q)} = \left\{ \left(a_i^{(q)}, y_i, e_i, s_i \right) \right\}_{i=1}^{N_q}. \tag{1}$$

The task label y_i specifies class identity, whereas s_i identifies the exact sentence shared across languages and thus defines a finer cross-lingual relation than the class label alone.

Let $\mathcal{R} = \{r_1, \ldots, r_K\}$ denote the set of high-resource reference languages. For example i and reference language $r \in \mathcal{R}$, the parallel text e_i is mapped to sentence-matched text $x_i^{(r)}$ in language r, from which the corresponding reference speech is synthesized:

$$x_i^{(r)} = g_{\text{text}}(e_i, r), \qquad \widetilde{a}_i^{(r)} = g_{\text{TTS}}\left(x_i^{(r)} \right), \qquad r \in \mathcal{R}. \tag{2}$$

Each natural target utterance is therefore paired with K cross-lingual reference views. These views differ in language and speech realization but share sentence identity s_i with the target utterance, pre-

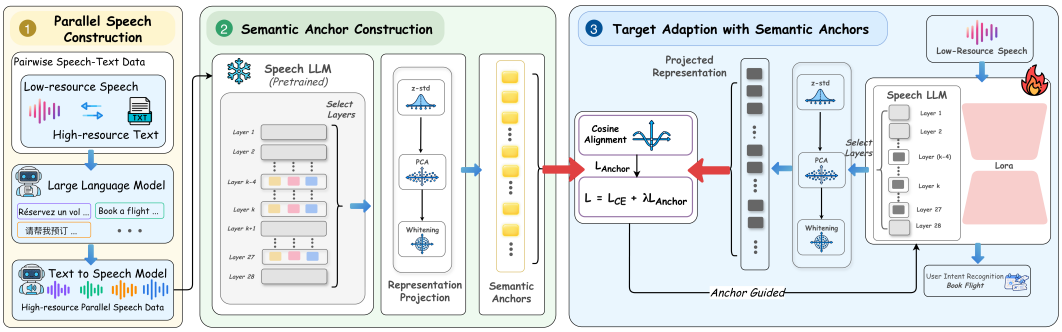

Figure 3: **Overview of ANCHORSLU.** Sentence-matched cross-lingual reference speech is encoded offline and aggregated into sentence-specific semantic anchors; the target model is then adapted on natural target speech using target-only CE and multi-layer anchor alignment. All reference-side computation is precomputed and fixed, and inference uses only the adapted target model.

serving exact sentence-level correspondence rather than only class agreement. Reference construction is entirely offline. The resulting speech is never treated as classification data for the adapted target model; its role is to define the semantic references used for anchoring.

3.2 SEMANTIC ANCHORS FROM MULTILINGUAL REFERENCE REPRESENTATIONS

We next consolidate the cross-lingual views of the same sentence into a stable representation target. Each reference utterance is encoded by a frozen speech backbone f_{θ_0}, identical to the pretrained backbone used to initialize the target model. Let $\mathcal{S}$ denote the selected layer set used for anchoring. Given task prompt p, we pool the task-prompt hidden states after the audio input and before answer generation at layer $l \in \mathcal{S}$, yielding the task-conditioned reference representation

$$\mathbf{u}_{i,r,l} = \mathrm{Pool}\left(H_{\theta_0,l}\left(\widetilde{a}_i^{(r)}, p\right)\right). \tag{3}$$

Raw hidden states can vary substantially in scale and covariance across dimensions and layers. We therefore place reference and target representations in a fixed *reference-derived comparison space* before alignment. For each selected layer l, the transformation P_l is fitted exclusively from training reference states using dimension-wise standardization, PCA, and whitening; target-language, validation, and test representations are excluded from fitting.

Each reference representation is first mapped into this comparison space, after which the sentence-matched views are aggregated:

$$\mathbf{z}_{i,r,l} = P_l\left(\mathbf{u}_{i,r,l}\right), \qquad \mathbf{c}_{i,l} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} \mathbf{z}_{i,r,l}. \tag{4}$$

The resulting $\mathbf{c}_{i,l}$ is the sentence-specific semantic anchor for example i at layer l. It aggregates only multilingual views that share sentence identity s_i, making the anchor sentence-specific rather than class-specific. Across the selected layers, $\{\mathbf{c}_{i,l}\}_{l\in\mathcal{S}}$ provides a multi-layer semantic reference for the corresponding target utterance.

3.3 ANCHOR-GUIDED TARGET MODEL ADAPTATION

Target adaptation turns the cached semantic anchors into training supervision while keeping the entire reference side fixed. We denote the adapted target model by $f_{\theta_0,\phi}$, where the pretrained parameters θ_0 remain frozen and only the adaptation parameters ϕ are trainable. The layer-specific transformations P_l and semantic anchors $\mathbf{c}_{i,l}$ are likewise fixed throughout adaptation.

For natural target speech $a_i^{(q)}$, the same task prompt, pooling rule, and fixed transformation map the target representation into the reference-derived comparison space:

$$\mathbf{u}_{i,l}^{(q)} = \mathrm{Pool}\left(H_{\theta_0,\phi,l}\left(a_i^{(q)}, p\right)\right), \qquad \mathbf{z}_{i,l}^{(q)} = P_l\left(\mathbf{u}_{i,l}^{(q)}\right). \tag{5}$$

ANCHORSLU then directly aligns these target representations with the semantic anchors for the same sentence. For training example i, the anchor loss averages cosine distance across the selected layers:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \left[1 - \cos\left(\mathbf{z}_{i,l}^{(q)}, \mathbf{c}_{i,l} \right) \right]. \tag{6}$$

Target adaptation combines two complementary supervision signals. Target-only CE constrains the task prediction and supplies class-level supervision, while $\mathcal{L}_{\text{anchor}}$ directly constrains intermediate representations using exact sentence-level cross-lingual correspondence. At optimization step t, the objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_t \mathcal{L}_{\text{anchor}}. \tag{7}$$

Only ϕ is updated. Natural target speech is the only speech passed through the adapted target model; reference speech, frozen reference states, comparison-space transformations, and semantic anchors are precomputed and fixed. ANCHORSLU therefore introduces no reference-side component at inference, which uses only the adapted target model.

4 EXPERIMENTS

Multilingual speech understanding spans settings with substantially different amounts of target-language supervision and task granularity. We evaluate ANCHORSLU in two complementary regimes: low-resource topic classification on SIB-Fleurs and richer-supervision intent recognition on Speech-MASSIVE. We additionally test whether the benefit extends to a substantially larger speech backbone through a matched Qwen3-Omni comparison. Together, these evaluations examine whether semantic anchoring remains effective across supervision conditions, task granularities, backbone scales, and linguistically diverse target languages.

4.1 EXPERIMENTAL SETUP

Datasets. SIB-Fleurs (Schmidt et al., 2025) combines natural FLEURS speech (Conneau et al., 2023) with the seven topic labels of SIB-200 (Adelani et al., 2024). We use a fixed 20-language target panel spanning nine broad language families and 11 scripts, selected to balance pre-adaptation frozen-model performance with family and script coverage. Speech-MASSIVE (FitzGerald et al., 2023; Lee et al., 2024) provides a complementary richer-supervision setting with 60 intent labels and 2,011 training recordings per target locale; its seven target locales are disjoint from the fixed eight-language reference set. For the family-level analysis in Table 2, targets in both benchmarks are organized into broad reporting groups, with the remaining families merged under *Other*. Exact family definitions, group sizes, target-panel construction, and complete language-level results are provided in Appendices B.1 and C.1.

Table 1: **Evaluation settings.**

Dataset	Targets	Task	Labels
SIB-Fleurs	20	Topic	7
Speech-MASSIVE	7	Intent	60

Models and training. Qwen2.5-Omni-7B is the primary backbone. Within each benchmark, ANCHORSLU and matched target-only CE use the same target-language data, LoRA configuration, optimization budget, checkpoint-selection rule, and evaluation protocol, isolating the contribution of semantic anchoring. Full implementation details are provided in Appendix A. Table 2 additionally reports target-only CE results from five diverse speech backbones to contextualize the two evaluation regimes. A matched Qwen3-Omni comparison in Section 4.2 evaluates ANCHORSLU at a substantially larger backbone scale.

Evaluation. Accuracy and Macro-F1 are computed separately for each target language. Overall scores are equally weighted averages across target languages, and group scores apply the same averaging within each reported broad language group. Languages from families without a dedicated column are pooled under *Other*, a reporting category rather than a linguistic family. Primary Qwen2.5-Omni-7B results are averaged over three independent runs. Group definitions, complete language-level results, and run-level variation are provided in Appendices B and C.

Table 2: **Results across resource regimes and broad language groups.** Cells report equally weighted language-macro Accuracy / Macro-F1. Bold marks the best result, and underlining marks the strongest target-only CE baseline for each metric. Group definitions and complete language-level results are provided in Appendices B.1 and C.1.

Model / Method	SIB-Fleurs				Speech-MASSIVE			
	Overall	Niger–Congo	Indo–European	Other	**Overall**	Indo–European	Turkic	Other
Target-only CE across backbones								
MiniCPM-o 2.6	56.78/52.53	50.35/44.10	65.24/61.60	56.36/53.12	55.94/49.49	68.79/61.97	60.86/55.68	41.44/34.94
LLaMA-3.1-8B-Omni	51.59/47.65	44.69/38.23	59.79/56.52	51.64/48.99	65.88/58.08	71.40/63.80	68.12/60.57	59.61/51.53
Step-Audio-2-mini	51.31/47.15	48.08/41.73	54.55/52.11	51.66/48.01	54.10/46.53	68.13/61.70	56.86/49.42	39.15/30.39
Phi-4 Multimodal	42.96/36.14	44.63/36.09	44.58/39.57	40.94/34.26	33.08/25.42	52.62/44.94	24.28/16.68	16.47/8.80
MOSS-Audio-4B-Instruct	38.19/31.72	35.45/26.15	39.77/34.69	39.14/33.77	37.16/29.43	54.65/46.72	18.93/14.21	25.75/17.21
Matched Qwen2.5-Omni-7B comparison								
Frozen	45.02/41.97	34.79/31.72	49.34/47.10	49.43/45.96	35.82/35.89	45.25/46.05	27.94/28.70	29.01/28.14
Target-only CE	55.48/51.39	48.60/43.97	60.75/57.78	57.14/52.78	67.20/60.92	75.79/70.29	66.98/59.96	58.70/51.88
ANCHORSLU	**59.87/56.43**	**52.52/47.57**	**66.03/63.20**	**61.36/58.57**	**70.39/64.44**	**78.11/73.17**	**72.42/66.53**	**62.00/55.01**

4.2 MAIN RESULTS ACROSS RESOURCE REGIMES

ANCHORSLU delivers consistent gains across resource regimes and broad language groups. Table 2 compares the matched Qwen2.5-Omni-7B results with target-only CE results from five additional speech backbones. ANCHORSLU improves both metrics over matched target-only CE and attains the strongest Accuracy and Macro-F1 in every overall and language-group column.

Low-resource topic classification. On the 20-language SIB-Fleurs panel, ANCHORSLU improves Accuracy by 4.39 points and Macro-F1 by 5.04 points over matched target-only CE. The gain extends across every reported broad language group, showing that sentence-level cross-lingual anchoring remains effective across linguistically diverse target languages. Notably, target-only CE already provides a strong improvement over the frozen backbone; ANCHORSLU therefore contributes beyond conventional target-language adaptation rather than merely compensating for weak zero-shot performance.

The result is equally consistent at finer granularity: every independent run improves the overall language macro, and the run-averaged result improves both metrics for every target language. Complete per-language and run-level results are provided in Appendix B.

Richer-supervision intent recognition. ANCHORSLU remains effective under substantially richer target-language supervision and a finer-grained label space. Each Speech-MASSIVE target locale provides 2,011 training recordings across 60 intents. Over matched target-only CE, ANCHORSLU improves Accuracy by 3.19 points and Macro-F1 by 3.52 points, with positive gains in every reported broad language group. Together with the SIB-Fleurs results, this shows that sentence-level anchoring complements class supervision across markedly different resource regimes and task granularities.

Scaling to a Larger Speech Backbone. We further evaluate ANCHORSLU with Qwen3-Omni on the fixed ten-language diagnostic panel used for the SIB-Fleurs analyses in Section 5 and detailed in Appendix D.1. As shown in Table 3, ANCHORSLU improves matched target-only CE by 2.95 points in Accuracy and 3.20 points in Macro-F1.

Table 3: **Results with Qwen3-Omni on SIB-Fleurs.** Cells report Accuracy / Macro-F1 on the fixed ten-language diagnostic panel.

Method	**Overall**	Niger–Congo	Indo–European	Other
Frozen	48.67/49.75	31.94/37.49	57.57/57.77	47.01/47.78
Target-only CE	64.88/61.03	49.74/43.27	72.75/69.48	63.46/59.76
ANCHORSLU	**67.83/64.23**	**54.45/47.95**	**75.33/71.77**	**66.31/63.17**

The improvement is positive across all three reported language groups. At the language level, Macro-F1 improves for all ten target languages, while Accuracy improves or is preserved for eight. These results show that sentence-level cross-lingual anchoring remains effective after strong target-only

adaptation on a substantially larger speech backbone. Complete language-level results are provided in Appendix B.5.

Across resource regimes and backbone scales, ANCHORSLU consistently improves matched target-only CE across linguistically diverse target languages. We next examine which properties of sentence-matched reference speech produce these gains and how semantic anchoring turns them into effective supervision for target adaptation.

5 WHAT MAKES SEMANTIC ANCHORING EFFECTIVE?

Section 4 shows that ANCHORSLU consistently improves matched target-only adaptation across resource regimes, broad language groups, and backbone scales. We now ask what produces these gains. We first examine how semantic anchoring reshapes target–reference representation geometry, then isolate the role of exact sentence correspondence, compare alternative uses of the same reference speech, and finally analyze how the comparison space and multilingual aggregation shape the semantic anchor. The SIB-Fleurs analyses use a fixed ten-language panel, while the Speech-MASSIVE correspondence controls use all seven target locales; full settings and language-level results are provided in Appendices C and D.

5.1 REPRESENTATION ALIGNMENT

ANCHORSLU reshapes low-resource target representations toward a shared multilingual semantic space. The consistent improvements across all 20 SIB-Fleurs target languages show that the benefit is not confined to a single language. Figures 2 and 4 visualize the representation change using Telugu as a representative low-resource target language. In the frozen backbone, natural target speech and its sentence-matched reference views occupy distinct regions in early and intermediate layers. Target-only CE leaves target speech locally separated from the multilingual reference region across multiple layers. In contrast, ANCHORSLU produces substantially stronger target–reference interleaving in intermediate and deeper layers while preserving coherent target-local structure.

This geometric shift follows directly from the supervision introduced by ANCHORSLU. Target-only CE constrains class predictions but does not identify which cross-lingual utterances express the same sentence. ANCHORSLU supplies this missing relation by aligning intermediate target representations with sentence-specific multilingual anchors. The resulting geometry shows how exact sentence-level correspondence is embedded directly into intermediate representations, aligning low-resource target speech with a shared multilingual semantic region that class-level supervision alone does not define.

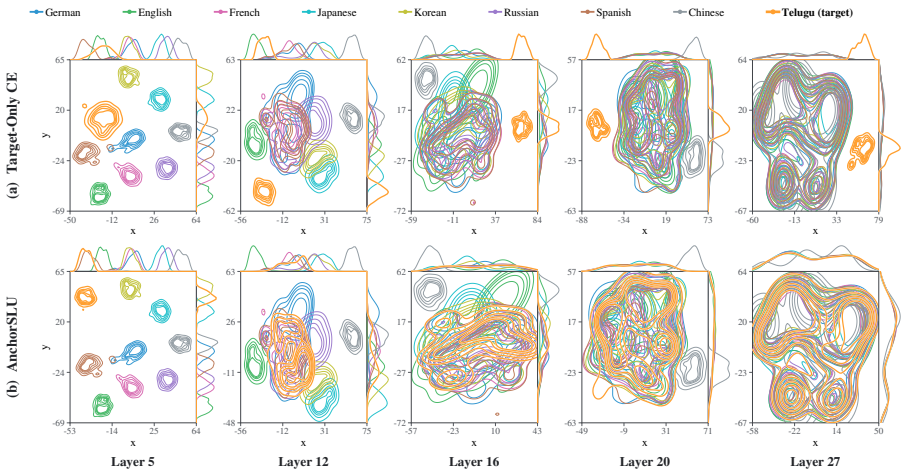

Figure 4: **Representation alignment after adaptation.** t-SNE at five selected layers for Telugu as a representative low-resource target language under target-only CE and ANCHORSLU. ANCHORSLU brings target representations substantially closer to the multilingual reference region in intermediate and deeper layers.

5.2 EXACT SENTENCE CORRESPONDENCE

Exact sentence correspondence drives the gains from ANCHORSLU. We isolate correspondence granularity while keeping reference speech, the comparison space, anchor construction, and the alignment objective fixed. Shuffled assignment removes both sentence and class correspondence; class-matched assignment retains class identity but pairs the target with a different sentence; sentence-matched assignment preserves the exact cross-lingual correspondence used by ANCHORSLU.

Table 4: **Cross-lingual correspondence granularity.** Shuffled, class-matched, and exact sentence-matched target–anchor assignments are compared while the remaining anchoring pipeline is fixed.

Target–anchor match	SIB-Fleurs		Speech-MASSIVE	
	Accuracy ↑	Macro-F1 ↑	Accuracy ↑	Macro-F1 ↑
Target-only CE	55.35	51.25	66.97	60.48
Shuffled	55.22	50.52	67.30	60.84
Class-matched	56.17	51.72	67.95	61.91
Sentence-matched (ANCHORSLU)	**60.54**	**56.53**	**70.69**	**64.77**

Exact sentence matching produces the strongest results in both evaluation settings. On SIB-Fleurs, ANCHORSLU exceeds class-matched anchoring by 4.37 points in Accuracy and 4.81 points in Macro-F1. On Speech-MASSIVE, the corresponding margins remain 2.74 points in Accuracy and 2.86 points in Macro-F1. Shuffling reduces performance to the level of target-only CE, showing that reference speech without the correct target–anchor relation cannot recover the gains from ANCHORSLU.

This contrast isolates the source of the improvement. Class identity provides only a coarse constraint shared by semantically different sentences, whereas ANCHORSLU anchors each target utterance to multilingual realizations of the same sentence content. The effective unit of cross-lingual supervision is therefore the *sentence*, rather than the class.

5.3 USING SENTENCE-MATCHED REFERENCE SPEECH

ANCHORSLU makes the most effective use of sentence-matched reference speech. Using identical reference views, we compare data-, output-, and representation-level transfer. TTS-CE treats reference speech as classification data, Reference KD transfers reference-side predictions, and Projected SupCon uses the same selected layers and reference-derived comparison space as ANCHORSLU.

Table 5: **Alternative uses of sentence-matched reference speech.** Results are reported on the fixed ten-language SIB-Fleurs diagnostic panel. Parentheses show changes relative to target-only CE.

Adaptation	Level	Accuracy ↑	Macro-F1 ↑
Target-only CE	–	55.35	51.25
TTS-CE	Data	54.65 (−0.70)	50.42 (−0.83)
Exposure-matched TTS-CE	Data	57.75 (+2.40)	53.38 (+2.13)
Reference KD	Output	57.11 (+1.76)	51.15 (−0.10)
Projected SupCon	Representation	59.64 (+4.29)	55.25 (+4.00)
ANCHORSLU	Representation	**60.54** (+5.19)	**56.53** (+5.28)

Data- and output-level transfer do not recover the gains of ANCHORSLU. Plain TTS-CE does not improve over target-only CE, and ANCHORSLU exceeds Exposure-matched TTS-CE by 2.79 points in Accuracy and 3.15 points in Macro-F1 even after natural target-speech exposure is restored. Reference KD provides no Macro-F1 improvement over target-only CE. The target-side optimization analysis further strengthens this result. Compared with Exposure-matched TTS-CE, ANCHORSLU reaches its selected checkpoint using 1,696 rather than 7,200 target-model examples, corresponding to a 76.4% reduction, while achieving higher Accuracy and Macro-F1, as detailed in Appendix D.5.

Projected SupCon provides the strictest representation-level control. It uses the same reference views, selected layers, and reference-derived comparison space as ANCHORSLU. With these factors fixed,

ANCHORSLU remains ahead by 0.90 points in Accuracy and 1.28 points in Macro-F1. It also reaches its selected checkpoint at 212 rather than 260 target-side steps, with a median stopping point of 360 rather than 440 steps, as detailed in Appendix D.5. This comparison isolates the advantage of consolidating multilingual realizations of each sentence into a single semantic anchor, enabling direct target-to-anchor alignment without introducing cross-sentence negatives.

5.4 SEMANTIC ANCHOR DESIGN

Projection and multilingual aggregation jointly strengthen semantic anchoring in ANCHORSLU. Figure 5(a) shows that the full standardization, PCA, and whitening configuration improves over raw-space cosine alignment by 4.99 points in Accuracy and 5.38 points in Macro-F1. Unlike the partial projection variants, *Full* achieves the best result on both metrics, showing that the complete configuration is both effective and stable across Accuracy and Macro-F1. The reference-derived projection normalizes scale and covariance differences, placing target and reference representations in a directly comparable space for alignment.

Multilingual aggregation provides a complementary benefit. Figure 5(b) shows that one sentence-matched reference language already improves over target-only CE. Expanding from one to eight reference languages further increases Accuracy by 1.97 points and Macro-F1 by 1.67 points, producing the strongest configuration. Aggregating multiple realizations of the same sentence reduces dependence on any single reference language and yields a more informative sentence-specific target.

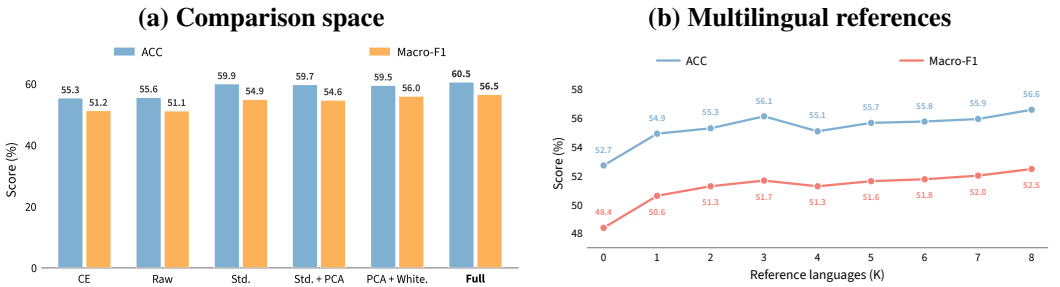

Figure 5: **Semantic-anchor design. (a)** Reference-derived comparison-space ablation; *Full* denotes standardization + PCA + whitening. **(b)** Multilingual reference-set sweep from $K = 0$ (target-only CE) to $K = 8$.

Taken together, these analyses show that the gains from ANCHORSLU arise from turning exact sentence correspondence into a sentence-specific representation target. Reference-derived projection and multilingual aggregation construct an alignable target, enabling ANCHORSLU to reshape low-resource representations more effectively than data augmentation, output distillation, or contrastive alignment.

6 CONCLUSION

This work shows that resource-dependent gaps in multilingual spoken language understanding can be narrowed by turning exact sentence-level cross-lingual correspondence into representation-level supervision. ANCHORSLU aggregates sentence-matched high-resource speech into multi-layer semantic anchors that guide natural target-speech adaptation, with all reference-side computation performed offline. Across 20 low-resource SIB-Fleurs languages, ANCHORSLU improves Accuracy by 4.39 points and Macro-F1 by 5.04 points over matched target-only CE; the gains persist on 60-intent Speech-MASSIVE and transfer to the larger Qwen3-Omni backbone. Controlled analyses confirm that exact sentence matching is essential, while projected multilingual anchors outperform data augmentation, output distillation, and contrastive alignment. Representation analyses further show that this supervision moves low-resource target representations toward a shared multilingual semantic space. Together, these results establish exact sentence correspondence as a practical source of supervision beyond class labels for multilingual speech understanding, while preserving target-only inference without additional model components.

AI USE STATEMENT

Generative AI tools were used for stylistic polishing of the manuscript and to generate the sentence-matched multilingual reference texts and synthetic speech described in Section 3.1, which are subsequently used for semantic-anchor construction in Section 3.2. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the paper.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. This work uses publicly available multilingual speech and text resources together with synthetic speech generated by TTS systems, and all data are obtained and used in accordance with their respective licenses and terms of use. Any reference-speaker speech used in the speech synthesis process is handled in accordance with the applicable privacy and usage policies. All models trained or adapted in this work are used solely for research purposes.

REPRODUCIBILITY STATEMENT

The main text specifies the training setup and evaluation protocol, and the appendix provides implementation details, complete language-level results, control experiments, and sensitivity analyses. An anonymized implementation, including data processing, prompt and TTS configurations, projection fitting, adapter settings, evaluation scripts, and machine-readable outputs, is available at https://anonymous.4open.science/r/AnchorSLU-BFDE/.

REFERENCES

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 226–245. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.14. URL https://aclanthology.org/2024.eacl-long.14/.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*, pp. 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143.

Xuxin Cheng, Zhihong Zhu, Bang Yang, Xianwei Zhuang, Hongxiang Li, and Yuexian Zou. Cyclical contrastive learning based on geodesic for zero-shot cross-lingual spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1806–1816. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.106. URL https://aclanthology.org/2024.findings-acl.106/.

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)*, pp. 798–805. IEEE, 2023. doi: 10.1109/SLT54892.2023.10023141. URL https://ieeexplore.ieee.org/document/10023141.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4277–4302. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.235. URL https://aclanthology.org/2023.acl-long.235/.

Seraphina Fong, Marco Matassoni, and Alessio Brutti. Speech LLMs in low-resource scenarios: Data volume requirements and the impact of pretraining on high-resource languages. In *Proceedings of Interspeech*, pp. 2003–2007, 2025. doi: 10.21437/Interspeech.2025-764. URL https://www.isca-archive.org/interspeech_2025/fong25_interspeech.html.

Yun Hao, Reihaneh Amooie, Wietse de Vries, Rik van Noord, and Martijn Wieling. Investigating the role of synthetic data augmentation and training strategies on improving low-resource language ASR. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 5633–5639. ELRA Language Resource Association, 2026. doi: 10.63317/4c4ad7t5i967. URL https://aclanthology.org/2026.lrec-1.442/.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. Speech-MASSIVE: A multilingual speech dataset for SLU and beyond. In *Proceedings of Interspeech*, pp. 817–821, 2024. doi: 10.21437/Interspeech.2024-957.

Beomseok Lee, Marco Gaido, Ioan Calapodescu, Laurent Besacier, and Matteo Negri. Towards zero-shot SLU: An empirical study of competing architectural paradigms. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, pp. 1–7. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.iwslt-1.1. URL https://aclanthology.org/2026.iwslt-1.1/.

Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. Label-aware multi-level contrastive learning for cross-lingual spoken language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9903–9918. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.673. URL https://aclanthology.org/2022.emnlp-main.673/.

Zihan Liu, Genta Indra Winata, Peng Xu, Zhaojiang Lin, and Pascale Fung. Cross-lingual spoken language understanding with regularized representation alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7241–7251. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.587. URL https://aclanthology.org/2020.emnlp-main.587/.

Potsawee Manakul, Guangzhi Sun, Warit Sirichotedumrong, Kasima Tharnpipitchai, and Kunat Pipatanakul. Enhancing low-resource language and instruction following capabilities of audio language models. In *Proceedings of Interspeech*, pp. 2083–2087, 2025. doi: 10.21437/Interspeech.2025-1258. URL https://www.isca-archive.org/interspeech_2025/manakul25_interspeech.html.

Tianjun Mao and Chenghong Zhang. DiffSLU: Knowledge distillation based diffusion model for cross-lingual spoken language understanding. In *Proceedings of Interspeech*, pp. 715–719, 2023. doi: 10.21437/Interspeech.2023-699. URL https://www.isca-archive.org/interspeech_2023/mao23_interspeech.html.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. Data augmentation for spoken language understanding via pretrained language models. In *Proceedings of Interspeech*, pp. 1219–1223, 2021. doi: 10.21437/Interspeech.2021-117.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 2677–2686. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.191. URL https://aclanthology.org/2022.acl-long.191/.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Nathaniel Romney Robinson, Perez Ogayo, Swetha R. Gangu, David R. Mortensen, and Shinji Watanabe. When is TTS augmentation through a pivot language useful? In *Proceedings of Interspeech*, pp. 3538–3542, 2022. doi: 10.21437/Interspeech.2022-11203.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Fabian David Schmidt, Ivan Vulić, Goran Glavaš, and David Ifeoluwa Adelani. Fleurs-SLU: A massively multilingual benchmark for spoken language understanding. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Tqj3fYqhwS.

Di Wu, Liting Jiang, Bohui Mao, Hongyan Xie, Haoxiang Su, Zhongjiang He, Ruiyu Fang, Shuangyong Song, Hao Huang, and Xuelong Li. INT: Establishing information transfer for multilingual intent detection and slot filling. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15120–15142. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.783. URL https://aclanthology.org/2025.findings-acl.783/.

Bowen Xing and Ivor W. Tsang. HC^{2}L: Hybrid and cooperative contrastive learning for cross-lingual spoken language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8094–8105, 2024. doi: 10.1109/TPAMI.2024.3402746. URL https://doi.org/10.1109/TPAMI.2024.3402746.

Bowen Xing, Libo Qin, Zhihong Zhu, Zhou Yu, and Ivor W. Tsang. DXA-Net: Dual-task cross-lingual alignment network for zero-shot cross-lingual spoken language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1052–1062, 2026. doi: 10.1109/TPAMI.2025.3597726. URL https://doi.org/10.1109/TPAMI.2025.3597726.

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. *arXiv preprint arXiv:2503.20215*, 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. URL https://arxiv.org/abs/2509.17765.

Guolong Zhong, Hongyu Song, Ruoyu Wang, Lei Sun, Diyuan Liu, Jia Pan, Xin Fang, Jun Du, Jie Zhang, and Lirong Dai. External text based data augmentation for low-resource speech recognition in the constrained condition of OpenASR21 challenge. In *Proceedings of Interspeech*, pp. 4860–4864, 2022. doi: 10.21437/Interspeech.2022-649.

Yi Zhu, Zexun Wang, Hang Liu, Peiying Wang, Mingchao Feng, Meng Chen, and Xiaodong He. Cross-modal transfer learning via multi-grained alignment for end-to-end spoken language understanding. In *Proceedings of Interspeech*, pp. 1131–1135, 2022. doi: 10.21437/Interspeech.2022-11378.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. Enhancing code-switching for cross-lingual SLU: A unified view of semantic and grammatical coherence. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7849–7856. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.486. URL https://aclanthology.org/2023.emnlp-main.486/.

Zhihong Zhu, Xuxin Cheng, Zhanpeng Chen, Xianwei Zhuang, Zhiqi Huang, and Yuexian Zou. Code-switching can be better aligners: Advancing cross-lingual SLU through representation-level and prediction-level alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 153–160. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-short.15. URL https://aclanthology.org/2024.acl-short.15/.

Appendix

Contents

A Implementation Details . . . 14
- A.1 Reference-Text Construction . . . 14
- A.2 Prompts and Representation Pooling . . . 14
- A.3 Target Adaptation Parameters . . . 14
- A.4 Reference Speech and Comparison-Space Projection . . . 14
- A.5 Optimization and Checkpoint Selection . . . 15
- A.6 Anchor Configuration and Controlled Baselines . . . 15
- A.7 Algorithms . . . 15

B Full SIB-Fleurs Results . . . 17
- B.1 Dataset and Target-Panel Construction . . . 17
- B.2 Run-Level Results . . . 20
- B.3 Cross-Backbone Target-Only CE Results . . . 20
- B.4 Per-Language Primary Results . . . 21
- B.5 Matched Qwen3-Omni-30B-A3B Comparison . . . 22

C Full Speech-MASSIVE Results . . . 22
- C.1 Dataset Details and Primary Results . . . 22
- C.2 Cross-Backbone Target-Only CE Results . . . 23
- C.3 Target-Data Sensitivity . . . 23
- C.4 Comparison-Space Sensitivity . . . 24
- C.5 Sentence-Correspondence Controls . . . 24

D Mechanism and Design Analyses . . . 25
- D.1 Diagnostic Panel . . . 25
- D.2 Full-Layer Representation Analysis . . . 25
- D.3 Sentence-Correspondence Controls . . . 27
- D.4 TTS Augmentation Controls . . . 27
- D.5 Target-Side Optimization . . . 28
- D.6 Output-Level Distillation . . . 28
- D.7 Reference-Language Sweep . . . 29
- D.8 Comparison-Space Ablation . . . 29

E Shared Multilingual Adaptation . . . 30

Unless stated otherwise, Accuracy and Macro-F1 are reported as percentages and macro-averaged equally across target languages. In paired entries, Accuracy is reported first, followed by Macro-F1. Δ denotes the difference between ANCHORSLU and the corresponding target-only CE baseline.

A Implementation Details

A.1 Reference-Text Construction

For each training example, the parallel English text e_i serves as the semantic source for constructing sentence-matched reference text in eight fixed reference languages: German, English, French, Japanese, Korean, Russian, Spanish, and Chinese. The English reference uses the original English sentence directly, while the other seven reference-language realizations are generated with Qwen2.5-7B-Instruct. The translation process preserves sentence meaning, names, numbers, and sentence boundaries, with automatic checks filtering empty, malformed, or clearly abnormal outputs.

The resulting reference-language texts are then synthesized offline with Qwen3-TTS-12Hz-1.7B-CustomVoice to obtain sentence-matched reference speech. Each target training example therefore has eight cross-lingual speech views expressing the same sentence content, which are subsequently used for sentence-specific semantic anchor construction.

A.2 Prompts and Representation Pooling

Within each benchmark, target and reference utterances use the same English system message and task prompt. The system message is fixed to "You are a helpful assistant." For SIB-Fleurs, the task prompt asks the model to classify the input speech into one of seven topics: entertainment, geography, health, politics, science/technology, sports, and travel. For Speech-MASSIVE, the task prompt asks the model to classify the input into one of 60 fixed intent categories, presented in the same order used for evaluation. Neither task uses demonstrations or additional natural-language label definitions beyond the category names.

The complete Speech-MASSIVE intent inventory is:

> alarm query, alarm remove, alarm set, audio volume down, audio volume mute, audio volume other, audio volume up, calendar query, calendar remove, calendar set, cooking recipe, datetime convert, datetime query, email addcontact, email query, email querycontact, email sendemail, general greet, general joke, general quirky, iot cleaning, iot coffee, iot hue lightchange, iot hue lightdim, iot hue lightoff, iot hue lighton, iot hue lightup, iot wemo off, iot wemo on, lists createoradd, lists query, lists remove, music dislikeness, music likeness, music query, music settings, news query, play audiobook, play game, play music, play podcasts, play radio, qa currency, qa definition, qa factoid, qa maths, qa stock, recommendation events, recommendation locations, recommendation movies, social post, social query, takeaway order, takeaway query, transport query, transport taxi, transport ticket, transport traffic, weather query.

For Qwen2.5-Omni, representations are mean-pooled at each selected layer over the task-prompt hidden states after the audio input and before answer generation. Concretely, the pooling span begins immediately after the unique audio-end token and ends before the first subsequent user-message end token, thereby retaining the ordinary task-prompt tokens while excluding the system message, boundary tokens, assistant prefix, and answer. Target and reference utterances use the same representation extraction and pooling rule.

A.3 Target Adaptation Parameters

Target adaptation updates only the LoRA parameters ϕ on top of the frozen Qwen2.5-Omni-7B backbone. LoRA adapters are inserted into the `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` linear layers of all 28 Thinker decoder layers, with rank 8, scaling factor 32, and dropout 0.05, yielding 20,185,088 trainable parameters. The audio tower and multimodal alignment modules remain frozen throughout adaptation, with no trainable bias terms or additional trainable modules.

A.4 Reference Speech and Comparison-Space Projection

All sentence-matched reference texts are synthesized offline with Qwen3-TTS-12Hz-1.7B-CustomVoice. Synthesis uses the built-in Ryan speaker without additional reference audio or instruction and produces 24-kHz mono speech. We use the model's default sampling configuration,

with top-$k = 50$, top-$p = 1.0$, temperature 0.9, and repetition penalty 1.05. Before reference-side representation extraction, synthesized utterances are checked for basic validity, including finite values, duration, RMS energy, and clipping.

Both benchmarks use the same reference-derived comparison-space construction. At each selected layer l, feature-wise means and standard deviations are estimated from training reference representations only and used for dimension-wise standardization. PCA-256 is then fitted to the standardized reference states, followed by whitening of the retained components. For numerical stability, standard deviations are lower-bounded at 10^{-6} and retained variances at 10^{-8}. The resulting layer-specific projection P_l remains fixed throughout target adaptation. All projection statistics are fitted exclusively from training reference states; target-language, validation, and test representations are excluded.

A.5 Optimization and Checkpoint Selection

For Qwen2.5-Omni-7B, AnchorSLU and matched target-only CE use the same optimization configuration and training budget. Training uses bfloat16 with a physical batch size of 1 and gradient accumulation of 8, giving an effective batch size of 8 per parameter update. The learning rate is 5×10^{-5}, with weight decay 0.01 and a cosine learning-rate schedule; each target-language adaptation runs for at most 760 optimizer updates.

For AnchorSLU, the anchor weight λ_t is linearly warmed up over the first 38 optimizer updates to a maximum of 0.4. Validation is performed every 40 updates, and checkpoints are selected by minimum assistant-token validation CE with an early-stopping patience of 4. The same checkpoint-selection criterion is used for AnchorSLU and matched target-only CE, with no test performance used for model selection. Maximum input lengths are 2,304 tokens for SIB-Fleurs and 2,048 tokens for Speech-MASSIVE.

A.6 Anchor Configuration and Controlled Baselines

The default AnchorSLU configuration uses the fixed reference set DE, EN, FR, JA, KO, RU, ES, and ZH, with sentence-specific semantic anchors constructed at the selected layer set $\mathcal{S} = \{12, 16, 20, 24, 27\}$. Each anchor aggregates only reference representations sharing the same sentence identity as the corresponding target example. Reference speech, frozen reference representations, layer-specific projections, and semantic anchors are computed before target adaptation and remain fixed thereafter.

The controlled baselines in the main text retain the same target-language data, reference speech, and evaluation protocol whenever applicable, while varying the supervision signal or anchor configuration under study. TTS-CE and Exposure-matched TTS-CE use the same sentence-matched synthetic reference speech as additional classification data; Reference KD transfers output distributions from the same reference views; sentence-correspondence controls vary only the target-to-anchor correspondence; and the comparison-space variants and reference-language sweep alter the projection construction and reference set, respectively. Full constructions are provided in Appendix D.

Projected SupCon further controls the form of representation-level supervision. It uses the same eight frozen reference views, the same five selected layers, and the same fixed reference-derived comparison space as AnchorSLU. For each target example and layer, the eight projected exact sentence-matched reference representations form the positive set, while 64 projected reference representations from other sentence identities form the negative set. Projected target and reference representations are L_2-normalized and optimized with a multi-positive InfoNCE objective at temperature 0.07. The contrastive loss uses the same maximum weight of 0.4 and 38-update linear warm-up as the anchor loss. Unlike AnchorSLU, Projected SupCon retains the matched reference views as multiple positives and introduces cross-sentence negatives rather than aggregating them into a sentence-specific semantic anchor.

A.7 Algorithms

Algorithm 1 specifies the complete computation of AnchorSLU while separating offline reference-side computation from target adaptation. The offline phase comprises the first two conceptual stages in Figure 3: sentence-matched reference speech construction and sentence-specific semantic

anchor construction. Reference representations, layer-specific projections, and semantic anchors are computed before target adaptation and remain fixed thereafter. During adaptation, natural target speech is mapped into the same reference-derived comparison space and optimized with target-only CE and multi-layer anchor alignment; only ϕ is updated.

Algorithm 1 ANCHORSLU: offline reference-side computation and target adaptation.

Require: Target-language training set $\mathcal{D}^{(q)} = \{(a_i^{(q)}, y_i, e_i, s_i)\}_{i=1}^{N_q}$; reference-language set $\mathcal{R}$; frozen backbone f_{θ_0}; reference-text constructor g_{text}; TTS model g_{TTS}; task prompt p; selected-layer set $\mathcal{S}$; anchor-weight schedule $\{\lambda_t\}$

Ensure: Adapted parameters $\phi^\star$

Stage I: Offline Reference-Side Computation

for each training example i and reference language $r \in \mathcal{R}$ **do**
  Construct sentence-matched reference text $x_i^{(r)} \leftarrow g_{\text{text}}(e_i, r)$
  Synthesize reference speech $\widetilde{a}_i^{(r)} \leftarrow g_{\text{TTS}}(x_i^{(r)})$
  for each selected layer $l \in \mathcal{S}$ **do**
    Extract $\mathbf{u}_{i,r,l} \leftarrow \text{Pool}\Big(H_{\theta_0,l}(\widetilde{a}_i^{(r)}, p)\Big)$
  end for
end for
for each selected layer $l \in \mathcal{S}$ **do**
  Fit and freeze P_l from $\{\mathbf{u}_{i,r,l}\}_{i,r}$ using standardization, PCA-256, and whitening
end for
for each training example i and selected layer $l \in \mathcal{S}$ **do**
  Project reference views $\mathbf{z}_{i,r,l} \leftarrow P_l(\mathbf{u}_{i,r,l})$ for all $r \in \mathcal{R}$
  Construct the sentence-specific semantic anchor

$$\mathbf{c}_{i,l} \leftarrow \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathbf{z}_{i,r,l}$$

end for
Cache $\{\mathbf{c}_{i,l}\}$

Stage II: Target Adaptation

Initialize ϕ while keeping θ_0 frozen
for each optimization step t **do**
  Sample minibatch $\mathcal{B} \subset \mathcal{D}^{(q)}$ and compute $\mathcal{L}_{\text{CE}}$
  for each $i \in \mathcal{B}$ and $l \in \mathcal{S}$ **do**
    Extract and project the target representation

$$\mathbf{u}_{i,l}^{(q)} \leftarrow \text{Pool}\Big(H_{\theta_0,\phi,l}(a_i^{(q)}, p)\Big), \qquad \mathbf{z}_{i,l}^{(q)} \leftarrow P_l(\mathbf{u}_{i,l}^{(q)})$$

  end for
  Compute

$$\mathcal{L}_{\text{anchor}} \leftarrow \frac{1}{|\mathcal{B}|\,|\mathcal{S}|} \sum_{i \in \mathcal{B}} \sum_{l \in \mathcal{S}} \Big[1 - \cos\Big(\mathbf{z}_{i,l}^{(q)}, \mathbf{c}_{i,l}\Big)\Big]$$

  Update only ϕ using $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_t \mathcal{L}_{\text{anchor}}$
end for
return $\phi^\star$

Algorithm 2 defines the sentence-correspondence controls used in Section 5. The sentence-matched mapping preserves the original sentence identity. The class-matched mapping assigns each target example an anchor from a different sentence in the same class, preserving class identity while removing exact sentence correspondence. The shuffled mapping removes both constraints by permuting anchor assignments across the training set. All mappings are fixed before optimization, and the class-matched condition permits neither self-matches nor cross-class fallbacks.

Algorithm 2 Sentence-correspondence controls.

Require: Sentence-matched anchors $\{\mathbf{c}_{i,l}\}$; labels $\{y_i\}$; sentence identities $\{s_i\}$; fixed random seed ξ
Ensure: Fixed anchor mappings π_{sent}, π_{class}, and π_{shuf}
Define the sentence-matched mapping $\pi_{\text{sent}}(i) \leftarrow i$ for every training example i
for each class y **do**
    Define $\mathcal{I}_y \leftarrow \{i : y_i = y\}$
    Shuffle $\mathcal{I}_y$ using ξ and search its cyclic rotations for a valid within-class derangement π_y satisfying $s_{\pi_y(i)} \neq s_i$ for every $i \in \mathcal{I}_y$
    Set $\pi_{\text{class}}(i) \leftarrow \pi_y(i)$ for every $i \in \mathcal{I}_y$
end for
Sample a fixed global derangement π_{shuf} over the training examples using ξ
for each $v \in \{\text{sent}, \text{class}, \text{shuf}\}$ **do**
    Define the corresponding anchor assignment $\mathbf{c}_{i,l}^{(v)} \leftarrow \mathbf{c}_{\pi_v(i),l}$ for every i and l
end for
Freeze all mappings before target adaptation
return $\pi_{\text{sent}}, \pi_{\text{class}}, \pi_{\text{shuf}}$

Algorithm 3 formalizes the checkpoint-selection and evaluation protocol used throughout the primary Qwen2.5-Omni-7B experiments. For each target language, model selection depends only on assistant-token validation CE; test data are never used for checkpoint selection. Accuracy and Macro-F1 are computed independently for each target language and then macro-averaged with equal weight across languages.

Algorithm 3 Validation-only checkpoint selection and evaluation.

Require: Candidate checkpoints for each target language; validation and test sets; task prompt p
Ensure: Language-macro Accuracy and Macro-F1
for each target language q **do**
    Select the checkpoint using validation CE only
    $\kappa_q^\star \leftarrow \arg\min_\kappa \mathcal{L}_{\text{CE}}^{\text{val}}(\kappa)$
    for each $(a_i^{(q)}, y_i)$ in the test set **do**
        Obtain the predicted class with the fixed task evaluator
        $\widehat{y}_i \leftarrow \text{Predict}\left(a_i^{(q)}, p; \kappa_q^\star\right)$
    end for
    Compute language-level ACC_q and MacroF1_q
end for
Equally macro-average Accuracy and Macro-F1 across target languages
return language-macro Accuracy and Macro-F1

B Full SIB-Fleurs Results

This appendix provides the complete SIB-Fleurs evidence supporting the main results, including the frozen pre-adaptation landscape, target-panel construction, run-level variation, cross-backbone context, and full language-level comparisons.

B.1 Dataset and Target-Panel Construction

SIB-Fleurs contains 102 language configurations. We evaluate all 102 with the frozen Qwen2.5-Omni-7B backbone on the same seven-topic task to characterize the pre-adaptation performance landscape before any target-language adaptation. The complete frozen results are reported below. The eight fixed reference languages are retained in this inventory for context but are excluded from target-panel construction.

For target selection, the remaining languages are stratified by frozen Accuracy into three pre-adaptation difficulty ranges: below 40%, 40%–60%, and at least 60%. Frozen Accuracy is used only to characterize starting capability before adaptation, not to estimate the expected gain from

ANCHORSLU. Within these strata, the fixed 20-language panel is constructed to broaden both language-family and script coverage. The resulting panel contains 3, 9, and 8 languages from the highest, middle, and lowest difficulty strata, respectively. Neither target-only CE performance, ANCHORSLU performance, nor their difference enters the selection procedure.

The resulting panel spans nine broad language families and 11 scripts, with 442–755 training recordings per language. It contains wol_Latn, lao_Laoo, amh_Ethi, mya_Mymr, tel_Telu, asm_Beng, ben_Beng, ory_Orya, gaz_Latn, uzn_Latn, jav_Latn, kan_Knda, mar_Deva, lin_Latn, pbt_Arab, lug_Latn, fuv_Latn, kam_Latn, sna_Latn, and kat_Geor. This fixed panel is used unchanged for all primary SIB-Fleurs experiments.

Table 6: Selection metadata for the fixed 20-language SIB-Fleurs panel.

Language	Script	Broad family	Train	Frozen Acc.	Diagnostic
wol Latn	Latin	Niger–Congo	643	30.37	Yes
lao Laoo	Lao	Kra–Dai	591	62.07	Yes
amh Ethi	Ethiopic	Afro-Asiatic	752	47.45	Yes
mya Mymr	Myanmar	Sino-Tibetan	746	23.28	Yes
tel Telu	Telugu	Dravidian	658	53.78	Yes
asm Beng	Bengali	Indo-European	730	39.84	Yes
ben Beng	Bengali	Indo-European	742	51.50	Yes
ory Orya	Odia	Indo-European	442	51.12	—
gaz Latn	Latin	Afro-Asiatic	574	33.33	—
uzn Latn	Latin	Turkic	734	49.19	Yes
jav Latn	Latin	Austronesian	740	74.53	—
kan Knda	Kannada	Dravidian	660	52.04	—
mar Deva	Devanagari	Indo-European	749	62.24	—
lin Latn	Latin	Niger–Congo	755	36.51	—
pbt Arab	Arabic	Indo-European	701	41.98	Yes
lug Latn	Latin	Niger–Congo	691	36.19	—
fuv Latn	Latin	Niger–Congo	752	35.37	—
kam Latn	Latin	Niger–Congo	752	30.07	—
sna Latn	Latin	Niger–Congo	689	40.26	—
kat Geor	Georgian	Kartvelian	557	49.19	Yes

Full frozen pre-adaptation inventory. The following inventory reports the frozen Qwen2.5-Omni-7B performance for all 102 SIB-Fleurs languages, ordered by Accuracy. The eight fixed reference languages are marked with †; they are shown for completeness but are not eligible for inclusion in the target panel.

#	Language	Eval. n	Frozen Acc.	Frozen Macro-F1
1	ita_Latn	175	89.14	87.15
2	zho_Hant	172	88.95	86.32
3	nld_Latn	123	88.62	87.06
4	glg_Latn	174	86.78	84.53
5	rus_Cyrl†	392	86.48	84.58
6	por_Latn	177	85.88	83.90
7	zho_Hans†	485	85.77	83.33
8	arb_Arab	133	85.71	84.06
9	eng_Latn†	320	84.69	82.83
10	ast_Latn	177	83.62	80.91
11	ind_Latn	167	83.43	81.56
12	vie_Latn	176	82.95	80.62
13	fra_Latn†	327	82.87	82.18
14	spa_Latn†	465	82.80	79.92
15	cat_Latn	177	82.49	79.67
16	deu_Latn†	442	82.13	80.51
17	kor_Hang†	200	82.00	79.69
18	tha_Thai	176	81.82	78.16
19	hin_Deva	132	81.25	81.34
20	kea_Latn	175	81.14	77.80
21	bel_Cyrl	177	80.23	77.42
22	zsm_Latn	171	80.12	76.93
23	tur_Latn	164	79.27	75.06
24	jpn_Jpan†	337	79.23	77.01

Continued on next page

#	Language	Eval. n	Frozen Acc.	Frozen Macro-F1
25	oci_Latn	177	76.84	74.14
26	ukr_Cyrl	164	76.83	75.05
27	urd_Arab	120	76.54	75.72
28	ron_Latn	177	76.27	71.62
29	hrv_Latn	176	75.00	73.33
30	mkd_Cyrl	177	74.58	70.72
31	jav_Latn	171	74.53	72.32
32	srp_Cyrl	164	74.39	72.27
33	ltz_Latn	176	73.86	70.11
34	tgl_Latn	176	72.16	70.32
35	slv_Latn	174	71.26	66.43
36	ceb_Latn	149	70.47	68.24
37	mlt_Latn	176	70.45	65.95
38	snd_Arab	177	68.93	68.99
39	slk_Latn	169	68.64	64.62
40	ces_Latn	172	68.60	63.73
41	pol_Latn	165	68.48	65.22
42	bul_Cyrl	176	68.18	64.30
43	bos_Latn	177	67.23	63.86
44	guj_Gujr	177	67.12	65.75
45	azj_Latn	174	66.09	60.45
46	afr_Latn	95	64.21	63.10
47	swe_Latn	168	63.69	58.85
48	mar_Deva	177	62.24	60.00
49	lao_Laoo	132	62.07	57.12
50	pan_Guru	143	61.77	60.34
51	lvs_Latn	174	58.62	55.12
52	lit_Latn	177	57.63	52.49
53	kir_Cyrl	177	57.17	54.45
54	dan_Latn	177	57.06	54.61
55	npi_Deva	175	56.59	54.39
56	nob_Latn	127	55.91	51.25
57	kaz_Cyrl	176	55.48	53.14
58	tgk_Cyrl	163	55.21	52.17
59	tel_Telu	153	53.78	53.53
60	est_Latn	176	52.84	49.41
61	kan_Knda	174	52.04	51.12
62	ben_Beng	176	51.50	49.41
63	fin_Latn	175	51.43	47.44
64	ory_Orya	168	51.12	49.55
65	heb_Hebr	175	49.71	42.41
66	mal_Mlym	174	49.48	50.33
67	uzn_Latn	175	49.19	47.10
68	kat_Geor	177	49.19	44.40
69	ell_Grek	168	48.21	41.92
70	cym_Latn	177	48.02	46.15
71	amh_Ethi	149	47.45	46.31
72	pes_Arab	165	46.67	44.24
73	hun_Latn	177	45.20	34.63
74	swh_Latn	154	42.68	43.52
75	pbt_Arab	144	41.98	40.84
76	sna_Latn	176	40.26	38.11
77	asm_Beng	176	39.84	35.70
78	nso_Latn	169	39.65	34.26
79	ckb_Arab	176	39.20	36.02
80	nya_Latn	169	38.96	32.84
81	gle_Latn	176	38.64	37.49
82	hye_Armn	177	37.85	35.50
83	khk_Cyrl	177	37.29	29.67
84	xho_Latn	177	36.81	32.09
85	lin_Latn	139	36.51	32.20
86	lug_Latn	173	36.19	33.38
87	som_Latn	177	36.16	29.08
88	hau_Latn	166	36.01	31.49
89	fuv_Latn	166	35.37	33.00
90	ibo_Latn	177	34.92	29.91
91	khm_Khmr	168	33.51	26.47
92	gaz_Latn	17	33.33	22.08
93	zul_Latn	175	31.71	30.15
94	tam_Taml	169	31.44	29.93
95	yor_Latn	172	30.88	26.25
96	wol_Latn	123	30.37	26.74
97	kam_Latn	177	30.07	26.91
98	luo_Latn	133	27.82	25.04
99	umb_Latn	108	27.78	22.87
100	mri_Latn	176	25.20	23.08

Continued on next page

#	Language	Eval. n	Frozen Acc.	Frozen Macro-F1
101	mya_Mymr	175	23.28	19.67
102	isl_Latn	23	21.74	21.77

† Fixed high-resource reference language; included in the frozen inventory but excluded from target-panel construction.

B.2 Run-Level Results

Across all three independent runs, AnchorSLU improves over matched target-only CE on both language-macro Accuracy and Macro-F1. Averaged over the three runs, the improvement is 4.39 Accuracy points and 5.04 Macro-F1 points, exactly matching the primary SIB-Fleurs comparison in Table 2. The positive gain in every run further shows that the aggregate improvement is not driven by a single optimization instance.

Table 8: Run-level results on the fixed 20-language SIB-Fleurs panel. Accuracy and Macro-F1 are language-macro averages. Mean, standard deviation, and Δ are computed from unrounded values.

	Target-only CE		AnchorSLU		Δ	
Run	**Acc.**	**Macro-F1**	**Acc.**	**Macro-F1**	**Acc.**	**Macro-F1**
1	54.80	51.08	59.98	56.82	+5.19	+5.74
2	55.73	51.57	60.32	56.72	+4.58	+5.15
3	55.92	51.50	59.32	55.74	+3.40	+4.24
Mean $\pm$ SD	55.48 ± 0.60	51.39 ± 0.27	$\mathbf{59.87 \pm 0.51}$	$\mathbf{56.43 \pm 0.60}$	$\mathbf{+4.39 \pm 0.91}$	$\mathbf{+5.04 \pm 0.76}$

B.3 Cross-Backbone Target-Only CE Results

The additional speech backbones provide target-only CE context on the same fixed 20-language SIB-Fleurs panel used in the primary experiments. Their language-macro results match the corresponding SIB-Fleurs entries in Table 2. The Qwen2.5-Omni-7B results are the same three-run per-language means used in the primary comparison, whereas each additional backbone contributes one complete evaluation. Because these models differ in pretraining, architecture, speech front end, tokenizer, and adaptation implementation, they provide model-level context rather than matched estimates of the AnchorSLU effect.

Table 9: Per-language target-only CE results on the fixed 20-language SIB-Fleurs panel. Each entry reports Accuracy followed by Macro-F1. Qwen2.5-Omni-7B uses the three-run per-language mean; the remaining backbones use one complete evaluation.

Language	Qwen2.5-Omni-7B	MiniCPM-o 2.6	LLaMA-3.1-8B-Omni	Step-Audio-2-mini	Phi-4 Multimodal	MOSS-Audio-4B-Instruct
kam_Latn	42.27/39.38	45.69/39.51	43.12/35.66	44.29/39.45	40.09/30.88	31.24/18.45
fuv_Latn	49.39/43.38	50.30/44.79	43.60/39.33	49.09/43.11	45.43/37.03	34.15/22.52
wol_Latn	44.68/37.49	51.31/44.34	41.88/33.71	49.74/38.45	41.36/34.13	34.55/22.78
amh_Ethi	57.25/54.30	56.86/53.39	46.67/41.88	52.55/47.94	49.02/46.42	47.84/44.16
mya_Mymr	36.00/26.25	32.59/29.82	27.49/22.52	33.48/27.21	34.15/16.12	24.83/9.25
lao_Laoo	66.50/64.03	55.17/51.93	56.65/54.72	42.36/37.66	30.05/20.72	52.71/48.74
pbt_Arab	53.18/47.47	54.96/49.45	56.49/51.65	57.25/52.28	40.46/38.56	38.93/32.71
ory_Orya	59.64/57.40	63.45/60.95	63.45/61.05	48.88/48.12	42.60/38.46	37.89/36.47
kan_Knda	68.58/65.09	70.98/67.63	68.11/63.63	60.43/57.06	45.32/39.66	42.69/35.79
tel_Telu	64.71/64.23	70.59/69.98	68.49/66.60	60.08/59.81	47.48/47.20	51.26/48.24
mar_Deva	72.51/70.49	76.69/73.99	62.81/61.08	63.78/62.75	47.21/40.11	43.35/37.32
asm_Beng	52.31/49.09	61.77/57.28	53.52/49.08	47.48/43.88	42.86/35.38	36.82/31.18
lin_Latn	58.23/54.94	55.19/52.90	52.70/46.54	57.26/55.45	49.38/47.10	43.15/36.10
lug_Latn	47.00/42.17	49.06/39.07	42.09/37.88	46.11/38.41	42.09/27.63	32.17/26.17
sna_Latn	50.04/46.45	50.54/43.99	44.75/36.29	41.97/35.53	49.46/39.80	37.47/30.90
ben_Beng	66.09/64.43	69.31/66.35	62.66/59.75	55.36/53.50	49.79/45.32	41.85/35.79
gaz_Latn	29.63/21.59	50.00/45.31	22.22/28.57	44.44/44.22	33.33/25.19	16.67/13.17
jav_Latn	81.12/80.02	68.02/64.32	67.75/64.25	60.98/57.22	43.09/36.51	39.02/34.02
kat_Geor	53.05/46.44	50.41/47.01	50.61/45.68	52.85/49.57	46.95/43.04	38.41/33.36
uzn_Latn	57.43/53.04	52.66/48.69	56.81/53.04	57.74/51.42	39.03/33.47	38.80/37.20
Language macro	**55.48/51.39**	**56.78/52.53**	**51.59/47.65**	**51.31/47.15**	**42.96/36.14**	**38.19/31.72**

B.4 Per-Language Primary Results

The aggregate improvement extends across the full target panel. Under the three-run mean, AnchorSLU improves over matched target-only CE in both Accuracy and Macro-F1 for all 20 target languages, consistent with the main-text result. The magnitude of the gain nevertheless varies substantially across languages, ranging from modest improvements to gains above ten points. The language-macro average therefore summarizes a broadly consistent direction of improvement without implying a uniform effect size.

Table 10: Per-language SIB-Fleurs results on the fixed 20-language target panel. Values are mean ± sample SD over three independent runs.

Language	Target-only CE Acc.	ANCHORSLU Acc.	Δ Acc.	Target-only CE Macro-F1	ANCHORSLU Macro-F1	Δ Macro-F1
kam_Latn	42.27±1.50	45.22±1.42	+2.95±0.13	39.38±1.66	41.89±1.21	+2.51±0.79
fuv_Latn	49.39±2.00	52.44±1.06	+3.05±3.00	43.38±2.23	45.69±0.99	+2.31±3.12
wol_Latn	44.68±1.84	53.06±1.51	+8.38±2.28	37.49±1.40	46.67±1.06	+9.18±2.41
amh_Ethi	57.25±0.40	58.56±1.77	+1.31±1.59	54.30±0.72	55.12±2.03	+0.82±1.53
mya_Mymr	36.00±3.87	36.36±3.85	+0.37±2.65	26.25±5.62	31.33±0.59	+5.07±5.20
lao_Laoo	66.50±1.31	71.43±0.49	+4.93±1.71	64.03±2.26	70.24±0.57	+6.21±2.54
pbt_Arab	53.18±0.96	56.49±2.67	+3.31±1.72	47.47±1.00	50.32±1.67	+2.85±2.30
ory_Orya	59.64±0.23	65.69±2.33	+6.05±2.34	57.40±0.52	63.43±2.29	+6.02±2.56
kan_Knda	68.58±1.73	72.42±1.90	+3.84±3.48	65.09±0.99	70.60±1.79	+5.51±2.77
tel_Telu	64.71±1.45	75.35±1.47	+10.64±0.87	64.23±1.54	74.40±1.29	+10.17±0.44
mar_Deva	72.51±0.11	76.30±1.68	+3.79±1.73	70.49±0.17	75.15±2.07	+4.66±2.13
asm_Beng	52.31±0.41	59.76±0.20	+7.44±0.61	49.09±1.58	56.91±0.37	+7.82±1.35
lin_Latn	58.23±1.33	61.55±1.46	+3.32±0.83	54.94±0.71	58.49±0.75	+3.55±0.24
lug_Latn	47.00±1.57	50.22±1.57	+3.22±0.54	42.17±0.86	43.88±3.33	+1.71±3.27
sna_Latn	50.04±0.62	52.60±0.81	+2.57±0.21	46.45±1.18	48.80±0.79	+2.35±0.68
ben_Beng	66.09±1.41	71.89±2.32	+5.79±2.18	64.43±1.69	70.17±1.55	+5.74±1.50
gaz_Latn	29.63±3.20	35.19±6.41	+5.55±9.62	21.59±3.03	31.88±10.50	+10.30±13.54
jav_Latn	81.12±2.11	85.46±0.83	+4.34±1.41	80.02±1.34	84.14±0.68	+4.12±0.66
kat_Geor	53.05±1.06	55.56±0.71	+2.51±1.69	46.44±0.61	50.73±0.03	+4.29±0.63
uzn_Latn	57.43±0.93	61.89±2.41	+4.47±2.37	53.04±0.63	58.67±2.58	+5.63±2.29
Language macro	55.48±0.60	**59.87±0.51**	**+4.39±0.91**	51.39±0.27	**56.43±0.60**	**+5.04±0.76**

B.5 MATCHED QWEN3-OMNI-30B-A3B COMPARISON

The larger-backbone comparison provides complementary evidence that the benefit of semantic anchoring is not confined to the primary Qwen2.5-Omni-7B setting. On the fixed ten-language diagnostic panel, ANCHORSLU improves the language-macro result by 2.95 Accuracy points and 3.20 Macro-F1 points over matched target-only CE. The effect remains language-dependent in Accuracy, while Macro-F1 improves for all ten languages. Because this experiment uses one additional backbone from the same model family and a smaller target subset, we treat it as supporting evidence rather than a general claim of cross-backbone or scale generalization.

Table 11: **Results with Qwen3-Omni on SIB-Fleurs.** Cells report Accuracy / Macro-F1 on the fixed ten-language diagnostic panel.

Language	Target-only CE Acc.	ANCHORSLU Acc.	Δ Acc.	Target-only CE Macro-F1	ANCHORSLU Macro-F1	Δ Macro-F1
amh_Ethi	61.18	60.00	-1.18	56.93	57.13	+0.20
asm_Beng	71.03	78.07	+7.04	69.01	75.09	+6.08
ben_Beng	83.48	84.55	+1.07	81.69	82.45	+0.76
kat_Geor	58.74	61.18	+2.44	53.77	56.86	+3.09
lao_Laoo	78.33	79.80	+1.47	76.40	77.78	+1.38
mya_Mymr	40.35	40.35	+0.00	32.93	34.81	+1.88
pbt_Arab	63.74	63.36	-0.38	57.75	57.76	+0.01
tel_Telu	75.21	81.93	+6.72	73.75	80.18	+6.43
uzn_Latn	66.97	74.60	+7.63	64.77	72.26	+7.49
wol_Latn	49.74	54.45	+4.71	43.27	47.95	+4.68
Language macro	64.88	**67.83**	**+2.95**	61.03	**64.23**	**+3.20**

C FULL SPEECH-MASSIVE RESULTS

This appendix provides the complete Speech-MASSIVE evidence supporting the main results, including the primary language-level comparison, target-only CE backbone context, target-data sensitivity, anchor-design sensitivity, and sentence-correspondence controls.

C.1 DATASET DETAILS AND PRIMARY RESULTS

Speech-MASSIVE provides a complementary setting with substantially richer target-language supervision than SIB-Fleurs. The evaluation covers seven target locales—Arabic (`ar-SA`), Hungarian (`hu-HU`), Dutch (`nl-NL`), Polish (`pl-PL`), Portuguese (`pt-PT`), Turkish (`tr-TR`), and Vietnamese (`vi-VN`)—with 2,011 training recordings per locale and a 60-intent label space. None of these target locales overlaps with the fixed eight-language reference set.

The language-level results support the aggregate improvement reported in the main text. Averaged across the seven target locales, ANCHORSLU improves matched target-only CE by 3.19 Accuracy points and 3.52 Macro-F1 points. Accuracy improves for six of the seven locales, while Macro-F1 improves for all seven, indicating a broadly consistent benefit despite language-level variation in effect size.

Table 12: Per-locale Speech-MASSIVE results on the fixed seven-locale target panel. Values are mean ± sample SD over three independent runs. Differences are computed from unrounded means before formatting.

Locale	Target-only CE Acc.	ANCHORSLU Acc.	Δ Acc.	Target-only CE Macro-F1	ANCHORSLU Macro-F1	Δ Macro-F1
ar-SA	63.33±0.57	64.77±3.81	+1.45	55.98±1.11	57.42±4.59	+1.44
hu-HU	42.26±1.16	49.67±0.39	+7.42	33.23±1.52	40.14±0.37	+6.90
nl-NL	81.66±0.42	81.19±0.81	−0.47	76.84±0.78	76.87±0.64	+0.03
pl-PL	66.43±2.82	73.39±0.22	+6.96	58.38±4.27	66.73±0.60	+8.35
pt-PT	79.28±1.05	79.75±0.71	+0.47	75.64±1.73	75.92±0.54	+0.29
tr-TR	66.98±1.00	72.42±0.51	+5.44	59.96±0.96	66.53±0.47	+6.57
vi-VN	70.50±1.65	71.56±2.64	+1.06	66.43±2.00	67.46±1.77	+1.03
Language macro	67.20	**70.39**	**+3.19**	60.92	**64.44**	**+3.52**

C.2 CROSS-BACKBONE TARGET-ONLY CE RESULTS

The target-only CE comparison across speech backbones provides the same model-level context as in Table 2, now resolved by target locale. All models are evaluated on the same fixed seven-locale Speech-MASSIVE panel, but they differ in pretraining, architecture, speech front end, tokenizer, and adaptation implementation. These results therefore contextualize the primary Qwen2.5-Omni-7B setting rather than providing matched evidence that ANCHORSLU generalizes across these backbones. The language-macro results exactly reproduce the corresponding Speech-MASSIVE entries in Table 2.

Table 13: Per-locale target-only CE results on the fixed seven-locale Speech-MASSIVE panel. Each entry reports Accuracy followed by Macro-F1. Qwen2.5-Omni-7B uses the same three-run per-locale means as the primary comparison; the remaining backbones use one complete evaluation.

(a) Qwen2.5-Omni-7B, MiniCPM-o 2.6, LLaMA-3.1-8B-Omni, and Step-Audio-2-mini

Locale	Qwen2.5-Omni-7B	MiniCPM-o 2.6	LLaMA-3.1-8B-Omni	Step-Audio-2-mini
ar-SA	63.33/55.98	43.95/37.68	54.88/47.22	67.85/59.03
hu-HU	42.26/33.23	33.76/27.02	61.16/52.22	28.75/18.92
nl-NL	81.66/76.84	71.18/64.72	73.37/65.65	72.76/66.12
pl-PL	66.43/58.38	67.92/60.05	73.54/64.40	62.71/54.90
pt-PT	79.28/75.64	67.28/61.13	67.28/61.36	68.93/64.08
tr-TR	66.98/59.96	60.86/55.68	68.12/60.57	56.86/49.42
vi-VN	70.50/66.43	46.60/40.12	62.78/55.14	20.85/13.23
Language macro	**67.20/60.92**	**55.94/49.49**	**65.88/58.08**	**54.10/46.53**

(b) Phi-4 Multimodal and MOSS-Audio-4B-Instruct

Locale	Phi-4 Multimodal	MOSS-Audio-4B-Instruct
ar-SA	18.43/10.30	19.27/9.92
hu-HU	15.47/8.13	20.41/11.31
nl-NL	48.22/38.57	55.58/46.96
pl-PL	28.51/18.67	45.26/36.05
pt-PT	81.14/77.59	63.11/57.16
tr-TR	24.28/16.68	18.93/14.21
vi-VN	15.50/7.98	37.56/30.41
Language macro	**33.08/25.42**	**37.16/29.43**

C.3 TARGET-DATA SENSITIVITY

The benefit of ANCHORSLU persists across a wide range of target-language supervision budgets. Both language-macro Accuracy and Macro-F1 improve over matched target-only CE at every tested fraction from 10% to 100%. The gain is not monotonic with the amount of target data: improvements remain positive in the lowest-data regimes, while larger gains appear at several intermediate and

higher fractions. This analysis therefore supports robustness across supervision budgets rather than a claim that semantic anchoring becomes increasingly effective as target labels become scarcer.

All fractions use one fixed seed shared by target-only CE and ANCHORSLU. The subsets are nested, intent-stratified prefixes constructed by sorting semantic identities with a stable hash, so smaller fractions are contained within larger ones and the same target identities are used for both methods. The representation preprocessing is refitted independently at each fraction. The 100% condition reuses the corresponding single primary run; it is therefore distinct from the three-run mean reported in Table 12 and Table 2.

Table 14: Speech-MASSIVE gains over matched target-only CE across target-data fractions on the fixed seven-locale panel. All conditions use the same fixed seed; Δ denotes ANCHORSLU minus target-only CE.

Target data	Δ Accuracy	Δ Macro-F1
10%	+2.02	+1.49
20%	+1.19	+0.81
40%	+1.72	+1.81
50%	+3.37	+3.34
60%	+4.34	+4.02
80%	+4.27	+4.66
100%	+3.72	+4.29

C.4 COMPARISON-SPACE SENSITIVITY

The benefit of sentence-level alignment depends strongly on the space in which target and reference representations are compared. Direct cosine alignment in the raw representation space reaches 66.95 language-macro Accuracy. Dimension-wise standardization followed by PCA-256 raises the result to 69.11, and adding whitening further improves it to 70.69. This pattern is consistent with the comparison-space analysis in the main text: sentence-level correspondence provides the semantic supervision, while the reference-derived projection makes that supervision more effective for representation alignment.

Table 15: Speech-MASSIVE comparison-space sensitivity on the fixed seven-locale panel. Results are language-macro Accuracy from one evaluation.

Comparison space	Accuracy
Raw-space cosine	66.95
Standardization + PCA-256	69.11
Standardization + PCA-256 + whitening	**70.69**

C.5 SENTENCE-CORRESPONDENCE CONTROLS

Speech-MASSIVE exhibits the same correspondence-granularity pattern as SIB-Fleurs. When exact sentence correspondence is preserved, sentence-matched anchoring reaches 70.69 language-macro Accuracy and 64.77 Macro-F1, improving target-only CE by 3.72 and 4.29 points, respectively. Preserving only class identity recovers much less of the benefit, with gains of 0.98 Accuracy points and 1.43 Macro-F1 points. Randomly shuffled anchors remain close to target-only CE. The richer-supervision setting therefore supports the same conclusion as the main SIB-Fleurs analysis: the effective cross-lingual signal is concentrated at the sentence level rather than being explained by class agreement alone.

Table 16: Speech-MASSIVE sentence-correspondence controls on the fixed seven-locale panel. Locale rows report Accuracy; the final rows report language-macro Accuracy and Macro-F1 from one evaluation.

Metric / locale	Target-only CE	Shuffled anchor	Class-matched anchor	Sentence-matched anchor (ANCHORSLU)
ar-SA	63.05	62.68	63.82	66.34
hu-HU	40.92	40.18	39.27	49.23
nl-NL	82.08	82.68	81.98	80.60
pl-PL	63.21	66.91	69.81	73.23
pt-PT	80.26	79.79	80.90	80.50
tr-TR	67.08	66.61	67.38	72.33
vi-VN	72.19	72.23	72.50	72.63
Language-macro Accuracy	66.97	67.30	67.95	**70.69**
Language-macro Macro-F1	60.48	60.84	61.91	**64.77**

D MECHANISM AND DESIGN ANALYSES

This appendix provides the detailed evidence underlying the mechanism and design analyses in Section 5, including sentence-correspondence controls, reference-speech alternatives, representation-level objectives, reference-set composition, and comparison-space construction.

D.1 DIAGNOSTIC PANEL

Computationally intensive SIB-Fleurs analyses use a fixed ten-language subset of the primary 20-language panel to make repeated representation extraction and controlled retraining tractable. The subset is fixed independently of adaptation outcomes using only pre-adaptation frozen-model difficulty, broad family coverage, and script coverage; neither target-only CE, ANCHORSLU, nor their difference enters selection.

The resulting panel contains `wol_Latn`, `lao_Laoo`, `amh_Ethi`, `mya_Mymr`, `tel_Telu`, `asm_Beng`, `ben_Beng`, `uzn_Latn`, `kat_Geor`, and `pbt_Arab`. It retains all three pre-adaptation difficulty strata, spans eight broad language families and eight scripts, and covers frozen Accuracy from 23.28% to 62.07%. The same fixed panel is used throughout the SIB-Fleurs mechanism analyses and in the matched Qwen3-Omni-30B-A3B comparison. All reported ten-language aggregates weight these languages equally.

Table 17: Pre-adaptation metadata for the fixed ten-language SIB-Fleurs diagnostic panel. Difficulty strata are defined by frozen Qwen2.5-Omni-7B Accuracy before target-language adaptation.

Language	Script	Broad family	Difficulty	Frozen Acc.
wol_Latn	Latin	Niger–Congo	< 40	30.37
lao_Laoo	Lao	Kra–Dai	≥ 60	62.07
amh_Ethi	Ethiopic	Afro-Asiatic	40–60	47.45
mya_Mymr	Myanmar	Sino-Tibetan	< 40	23.28
tel_Telu	Telugu	Dravidian	40–60	53.78
asm_Beng	Bengali	Indo-European	< 40	39.84
ben_Beng	Bengali	Indo-European	40–60	51.50
uzn_Latn	Latin	Turkic	40–60	49.19
kat_Geor	Georgian	Kartvelian	40–60	49.19
pbt_Arab	Arabic	Indo-European	40–60	41.98

D.2 FULL-LAYER REPRESENTATION ANALYSIS

The full-layer view extends the selected-layer representation analyses in Figures 2 and 4. Across all 28 layers, the frozen backbone and target-only CE retain visible local separation between Telugu target representations and the eight sentence-matched reference languages at multiple depths. After adaptation with ANCHORSLU, stronger target–reference interleaving emerges across intermediate and deeper layers while target-local structure remains visible.

This analysis is qualitative: it is intended to show that the representation pattern illustrated in the selected-layer main-text figures is not confined to those displayed layers. Because each t-SNE panel is fitted independently, absolute coordinates and distances should not be compared across layers or adaptation conditions.

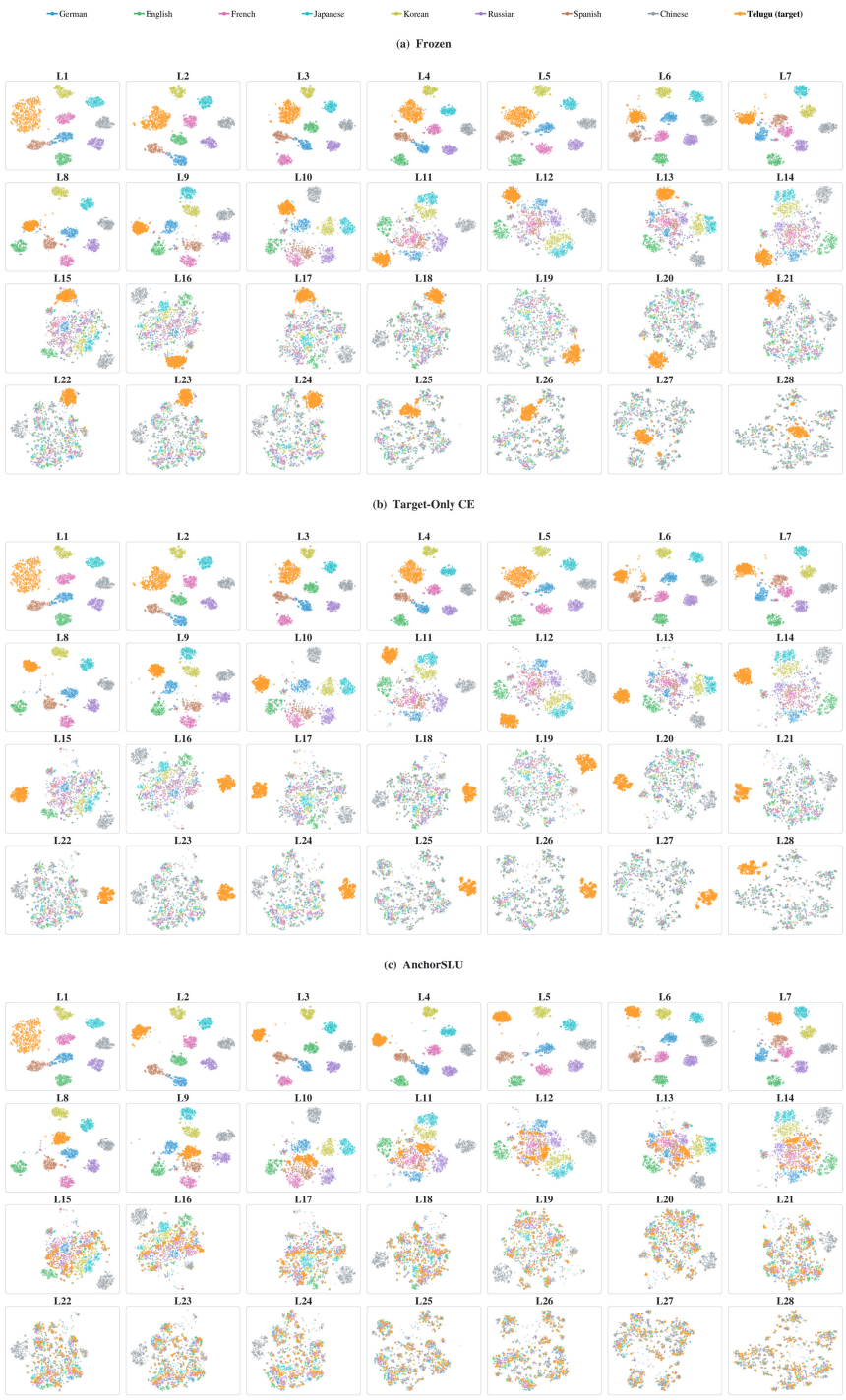

Figure 6: **Full-layer target–reference representation comparison.** Layer-wise t-SNE projections from L1–L28 for Telugu target speech and eight sentence-matched reference languages under the frozen backbone, target-only CE, and ANCHORSLU. Orange denotes the target language. Each panel uses independently fitted t-SNE coordinates; the visualization provides qualitative representation evidence only.

D.3 SENTENCE-CORRESPONDENCE CONTROLS

The language-level results reinforce the aggregate correspondence control in Table 4. Sentence-matched anchoring yields a substantially larger ten-language improvement than either coarser alternative: relative to target-only CE, ANCHORSLU improves language-macro Accuracy by 5.19 points and Macro-F1 by 5.28 points, whereas preserving class identity while replacing the matched sentence yields gains of only 0.82 and 0.47 points. Globally shuffled anchors remain near target-only CE. The language-level breakdown therefore supports the same conclusion as the aggregate analysis: class agreement alone does not account for the benefit of exact sentence-level cross-lingual correspondence.

Table 18: **SIB-Fleurs sentence-correspondence controls on the fixed ten-language diagnostic panel.** Each language entry reports Accuracy followed by Macro-F1; the final row is the equally weighted macro over all ten languages.

Language	Target-only CE	Shuffled anchor	Class-matched anchor	Sentence-matched anchor (ANCHORSLU)
wol Latn	42.93/36.03	45.55/37.36	45.03/36.77	53.93/47.50
lao Laoo	67.49/66.39	67.00/64.92	66.01/64.97	71.43/70.27
amh Ethi	57.25/54.46	56.47/54.25	56.08/53.07	56.86/54.04
mya Mymr	39.25/29.03	39.91/28.52	40.58/29.51	40.80/31.96
tel Telu	65.55/65.62	65.97/65.08	65.13/64.29	76.89/75.77
asm Beng	51.91/48.26	49.70/44.57	57.34/52.75	59.96/57.04
ben Beng	66.31/64.65	64.38/62.45	66.74/65.03	74.46/71.74
uzn Latn	57.27/52.72	58.89/54.47	58.43/54.87	59.12/55.72
kat Geor	52.44/46.90	52.44/46.63	51.83/46.16	55.49/50.73
pbt Arab	53.05/48.44	51.91/46.94	54.58/49.76	56.49/50.54
Language macro	**55.35/51.25**	**55.22/50.52**	**56.17/51.72**	**60.54/56.53**

D.4 TTS AUGMENTATION CONTROLS

Directly using the sentence-matched synthetic reference speech as additional classification data does not recover the benefit of semantic anchoring. *TTS-CE* mixes natural target speech with the same eight synthetic reference views used by ANCHORSLU while retaining the 760-step training ceiling of target-only CE and ANCHORSLU. Because the resulting 1:8 natural-to-synthetic mixture reduces the number of natural target examples seen during training, *Exposure-matched TTS-CE* expands the training schedule and validation grid by a factor of nine while preserving the same mixture. Validation and test sets contain only natural target speech.

This additional target exposure recovers part of the gap: Exposure-matched TTS-CE improves target-only CE by 2.40 Accuracy points and 2.13 Macro-F1 points. It nevertheless remains 2.80 Accuracy points and 3.15 Macro-F1 points below ANCHORSLU. The comparison therefore separates semantic anchoring from a direct synthetic-data augmentation explanation: additional sentence-matched reference speech helps when sufficient target exposure is restored, but does not reproduce the gain obtained by using the same cross-lingual correspondence as representation-level supervision.

Table 19: **TTS augmentation controls on the fixed ten-language SIB-Fleurs diagnostic panel.** Each language entry reports Accuracy followed by Macro-F1; the final row is the equally weighted macro over all ten languages.

	Target-only	Exposure-matched		
Language	CE	TTS-CE	TTS-CE	ANCHORSLU
wol Latn	42.93/36.03	39.27/35.03	50.79/43.79	53.93/47.50
lao Laoo	67.49/66.39	71.43/68.86	71.92/69.73	71.43/70.27
amh Ethi	57.25/54.46	52.55/49.67	58.82/54.71	56.86/54.04
mya Mymr	39.25/29.03	39.69/28.10	36.59/30.03	40.80/31.96
tel Telu	65.55/65.62	65.97/63.83	73.53/70.48	76.89/75.77
asm Beng	51.91/48.26	51.91/50.03	47.89/46.95	59.96/57.04
ben Beng	66.31/64.65	60.73/58.33	67.60/65.20	74.46/71.74
uzn Latn	57.27/52.72	56.35/51.70	57.51/52.11	59.12/55.72
kat Geor	52.44/46.90	56.30/50.00	54.07/48.50	55.49/50.73
pbt Arab	53.05/48.44	52.29/48.61	58.78/52.31	56.49/50.54
Language macro	**55.35/51.25**	**54.65/50.42**	**57.75/53.38**	**60.54/56.53**

Across the fixed ten-language panel, the paired bootstrap interval for the Accuracy difference between ANCHORSLU and target-only CE is $[+2.89, +7.69]$ points. Exposure-matched TTS-CE improves

Accuracy over target-only CE by 2.40 points, with interval $[-0.006, +4.85]$, whereas its difference from ANCHORSLU is -2.80 points with interval $[-5.48, -0.50]$. These intervals characterize variation across the fixed language panel rather than optimizer-seed uncertainty.

D.5 TARGET-SIDE OPTIMIZATION

The stronger Exposure-matched TTS-CE control requires substantially more target-model optimization to restore the natural target-speech exposure lost under synthetic-speech mixing. Its selected checkpoint occurs at 900 steps on average, compared with 212 steps for ANCHORSLU, while its language-macro performance remains lower. Thus, the advantage of semantic anchoring is not explained by a larger target-side optimization budget.

Table 20: Target-side optimization on the fixed ten-language SIB-Fleurs diagnostic panel. Selected steps are language means and stopping steps are language medians. Target-model examples equal selected steps multiplied by the effective batch size of eight.

Adaptation	Selected step	Stop step	Target-model examples	Acc. / Macro-F1
Target-only CE	116	280	928	55.35/51.25
TTS-CE	332	480	2,656	54.65/50.42
Exposure-matched TTS-CE	900	2,340	7,200	57.75/53.38
Reference KD	128	280	1,024	57.11/51.15
Projected SupCon	260	440	2,080	59.64/55.25
ANCHORSLU	212	360	1,696	**60.54/56.53**

Relative to Exposure-matched TTS-CE, ANCHORSLU reaches the selected checkpoint with 76.4% fewer target-model training examples while obtaining higher language-macro Accuracy and Macro-F1. This quantity is intentionally limited to target-side adaptation: it does not include offline reference-speech synthesis, frozen reference encoding, projection fitting, or anchor construction, and should therefore not be interpreted as an end-to-end compute comparison.

D.6 OUTPUT-LEVEL DISTILLATION

Reference KD tests whether the information in sentence-matched reference speech can be transferred sufficiently through output distributions alone. It uses the same eight sentence-matched reference views as ANCHORSLU. For each target sentence, the frozen backbone produces an output distribution over the seven task labels for every reference view; these distributions are averaged across reference languages at temperature $T = 2$. Target adaptation then combines target-only CE with unit-weight KL divergence to this fixed reference distribution.

The resulting reference distribution is itself informative, predicting the gold class with 86.38% Accuracy on the training cache. Output-level transfer, however, recovers only a limited fraction of the downstream improvement. Reference KD raises language-macro Accuracy from 55.35 to 57.11 but does not improve Macro-F1 over target-only CE, whereas ANCHORSLU reaches 60.54 Accuracy and 56.53 Macro-F1. Thus, access to informative sentence-matched reference predictions alone does not account for the gain from directly using the same correspondence as intermediate representation-level supervision.

Table 21: **Reference KD on the fixed ten-language SIB-Fleurs diagnostic panel.** Each language entry reports Accuracy followed by Macro-F1; the final row is the equally weighted macro over all ten languages.

Language	Target-only CE	Reference KD	ANCHORSLU
wol Latn	42.93/36.03	46.07/36.80	53.93/47.50
lao Laoo	67.49/66.39	67.00/63.38	71.43/70.27
amh Ethi	57.25/54.46	57.25/53.17	56.86/54.04
mya Mymr	39.25/29.03	41.46/28.42	40.80/31.96
tel Telu	65.55/65.62	68.49/65.94	76.89/75.77
asm Beng	51.91/48.26	56.34/49.98	59.96/57.04
ben Beng	66.31/64.65	64.81/62.82	74.46/71.74
uzn Latn	57.27/52.72	58.43/53.64	59.12/55.72
kat Geor	52.44/46.90	55.49/49.22	55.49/50.73
pbt Arab	53.05/48.44	55.73/48.18	56.49/50.54
Language macro	**55.35/51.25**	**57.11/51.15**	**60.54/56.53**

D.7 REFERENCE-LANGUAGE SWEEP

A single sentence-matched cross-lingual view already provides a substantial fraction of the benefit of semantic anchoring. Starting from target-only CE at $K = 0$, adding the first reference language raises language-macro Accuracy from 55.35 to 58.57 and Macro-F1 from 51.25 to 54.86. Subsequent nested reference sets remain within a comparatively narrow range before the full eight-reference configuration reaches the highest result, at 60.54 Accuracy and 56.53 Macro-F1.

The sweep follows the fixed nested order EN, ZH, RU, KO, JA, ES, DE, and FR. Consequently, increasing K changes both the number and identity of the included reference languages. The intermediate fluctuations should therefore not be interpreted as a monotonic language-count effect. Rather, the analysis shows that one exact sentence-matched cross-lingual view already supplies a strong supervision signal, while the full multilingual reference set provides the strongest endpoint.

Table 22: Reference-language sweep on the fixed ten-language SIB-Fleurs diagnostic panel. $K = 0$ denotes target-only CE; for $K \geq 1$, reference languages are added cumulatively in the fixed order EN, ZH, RU, KO, JA, ES, DE, FR.

K	0	1	2	3	4	5	6	7	8
Accuracy	55.35	58.57	59.11	59.57	59.10	59.65	59.73	59.64	**60.54**
Macro-F1	51.25	54.86	55.69	55.82	55.38	55.98	55.88	55.87	**56.53**

D.8 COMPARISON-SPACE ABLATION

Exact sentence correspondence alone is insufficient when alignment is performed directly in the raw representation space. Raw-space cosine alignment remains near target-only CE, reaching 55.55 language-macro Accuracy and 51.15 Macro-F1. Transforming the representations into a reference-derived comparison space produces substantially larger gains.

All three projection-based variants improve over raw-space alignment, while the complete dimension-wise standardization, PCA-256, and whitening pipeline achieves the strongest joint result, reaching 60.54 Accuracy and 56.53 Macro-F1. The ablation therefore complements the correspondence controls: sentence matching determines the semantic content of the supervision, while the comparison-space construction determines how effectively that signal can be used for representation alignment.

Table 23: Comparison-space ablation on the fixed ten-language SIB-Fleurs diagnostic panel. Results are equally weighted language-macro Accuracy and Macro-F1.

Comparison space	**Accuracy**	**Macro-F1**
Target-only CE	55.35	51.25
Raw-space cosine	55.55	51.15
PCA-256 + whitening	59.48	55.95
Standardization only	59.94	54.91
Standardization + PCA-256	59.73	54.65
Standardization + PCA-256 + whitening	**60.54**	**56.53**

E SHARED MULTILINGUAL ADAPTATION

The primary experiments adapt one target-specific LoRA adapter per language. As a complementary setting, we concatenate the training recordings from all 20 SIB-Fleurs target languages and train one shared multilingual adapter for each method. Shared target-only CE and shared ANCHORSLU use identical training and validation manifests, and checkpoints are selected by 20-language macro-average validation CE.

Semantic anchoring remains beneficial under this joint adaptation regime. Shared ANCHORSLU improves the language macro from 59.23 to 60.65 Accuracy and from 54.66 to 56.83 Macro-F1, corresponding to gains of 1.43 and 2.18 points, respectively. Accuracy improves for 14 languages and ties for two, while Macro-F1 improves for 16 of the 20 languages. The effect is therefore positive at the aggregate level despite language-level variation.

This experiment uses different optimization settings from the primary target-specific protocol and a single optimizer seed. We consequently treat it as complementary evidence that semantic anchoring is compatible with shared multilingual adaptation, rather than as an additional estimate of the primary method effect.

Table 24: Shared-adapter results on the fixed 20-language SIB-Fleurs panel. Each entry reports Accuracy followed by Macro-F1. Δ denotes shared ANCHORSLU minus shared target-only CE.

Language	Shared target-only CE	Shared ANCHORSLU	Δ
amh_Ethi	60.39/56.53	62.35/60.10	+1.96/+3.57
asm_Beng	58.55/55.67	59.36/54.71	+0.80/-0.96
ben_Beng	68.24/67.04	71.24/69.89	+3.00/+2.85
fuv_Latn	52.74/45.36	52.74/48.21	+0.00/+2.85
gaz_Latn	38.89/38.50	38.89/37.82	+0.00/-0.68
jav_Latn	81.84/81.46	82.38/81.76	+0.54/+0.30
kam_Latn	44.99/37.34	49.42/44.00	+4.43/+6.67
kan_Knda	69.78/65.53	73.62/69.76	+3.84/+4.23
kat_Geor	58.33/52.17	58.94/53.32	+0.61/+1.16
lao_Laoo	64.04/61.11	68.47/65.95	+4.43/+4.84
lin_Latn	58.51/55.53	60.58/58.66	+2.07/+3.13
lug_Latn	53.62/46.92	52.28/46.49	-1.34/-0.43
mar_Deva	73.41/71.71	75.53/74.31	+2.12/+2.61
mya_Mymr	41.69/32.02	40.80/33.29	-0.89/+1.27
ory_Orya	67.26/64.75	65.25/63.67	-2.02/-1.08
pbt_Arab	55.34/49.63	58.40/51.61	+3.05/+1.98
sna_Latn	55.46/48.76	54.82/49.41	-0.64/+0.65
tel_Telu	68.49/68.63	72.69/72.32	+4.20/+3.69
uzn_Latn	60.05/53.18	60.28/55.18	+0.23/+2.00
wol_Latn	52.88/41.26	54.97/46.18	+2.09/+4.91
Language macro	59.23/54.66	**60.65/56.83**	**+1.43/+2.18**