SEC: Selective Embedding Correction for Cross-Lingual Transfer in Low-Resource Languages

Anonymous ACL submission

Abstract

Enhancing NLP for low-resource languages (LRLs) is vital for supporting linguistic diversity, yet a fundamental challenge lies in the cross-lingual semantic gap. To address this, a common approach is to initialize LRL word embeddings with those from high-resource languages, thereby enriching their semantic space. However, subsequent embedding adaptation is typically performed through uniform updates across the vocabulary, which fails to allocate larger corrections to semantically difficult or infrequent tokens (e.g., culturally specific terms in Amharic or Somali) that need them most. To address this, we propose **S**elective **E**mbedding **C**orrection (SEC), a parameter-efficient framework for dynamic semantic correction. SEC models embedding adjustments as sparse low-rank residuals regularized by a group-lasso objective. This enables data-driven selectivity that refines poorly aligned tokens while preserving well-aligned representations. Experiments on named entity recognition, part-of-speech tagging, and natural language inference across multiple low-resource languages show that SEC achieves strong improvements over baselines with only about 1% additional trainable parameters[1].

1 Introduction

Advancing natural language processing for low-resource languages (LRLs) is crucial for supporting linguistic diversity. A fundamental challenge is the scarcity of training data in LRLs, which hinders the learning of high-quality, semantically rich word representations (Joshi et al., 2020; van Esch et al., 2022). To compensate for this scarcity, a common approach is to leverage representations from high-resource languages, but this leads to a semantic gap, as those representations often fail to capture the nuanced semantics and structural patterns specific to LRLs. Additionally, issues such as limited vocabulary coverage and fragmented subword representations also weaken the semantic foundation of the embedding layer (Rust et al., 2021; Petrov et al., 2023).

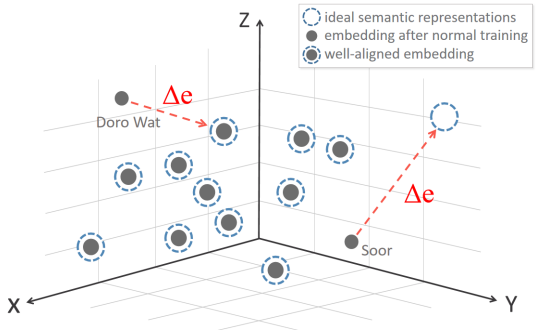

Figure 1: Conceptual illustration of selective embedding correction. Dashed blue circles denote ideal semantic representations, gray points represent token embeddings after normal training. Two rare tokens from low-resource languages are shown: "*Doro Wat*" (Amharic, a traditional spicy chicken dish) and "*Soor*" (Somali, a traditional cornmeal porridge). Both appear only a few times in the adaptation corpus. Red arrows denote learned corrections (Δe) that selectively shift such long-tail tokens toward ideal semantic representations.

To improve cross-lingual transfer for LRLs, recent approaches have increasingly focused on enriching target-language embeddings using semantic information transferred from high-resource languages. Existing methods achieve this through different forms of cross-lingual embedding alignment, including multilingual similarity-based initialization (Minixhofer et al., 2022; Dobler and de Melo, 2023; Pham et al., 2024), learned embedding mapping (Liu et al., 2024; Özeren et al., 2025), and bilingual lexical projection (Remy et al., 2024; Sakajo et al., 2025). Despite their methodological differences, these approaches generally adapt embeddings through shared update mechanisms

[1] Code is available at https://anonymous.4open.science/r/sec-code

applied uniformly across the vocabulary during language transfer.

However, not all tokens require the same degree of semantic adjustment during cross-lingual adaptation. Prior work has observed that cross-lingual transfer performance varies substantially across languages and tasks (Lauscher et al., 2020; Ebrahimi et al., 2022). Recent studies further suggest that rare or under-trained tokens may remain poorly represented even after standard training procedures, leading to unstable or degraded downstream behavior (Land and Bartolo, 2024; Chizhov et al., 2024). As illustrated in Figure 1, the Amharic token "*Doro Wat*" and the Somali token "*Soor*" may appear only a few times during training, making them difficult to align through uniform updates alone. Although relatively rare, such poorly aligned tokens can still disproportionately affect downstream tasks such as named entity recognition and natural language inference. These observations suggest that embedding adaptation should selectively focus on hard-to-align tokens rather than treating all tokens equally.

To address this limitation, we propose **S**elective **E**mbedding **C**orrection (**SEC**), which formulates embedding adaptation as a selective correction process rather than uniform updates. Specifically, we decompose the correction into a low-rank residual term added to the base embedding and enforce token-level sparsity via a group-lasso regularizer. As shown in Figure 1, this enables the model to learn correction vectors Δe that tend to shift misaligned tokens toward the ideal semantic representations while largely leaving well-aligned tokens unchanged. This selectivity arises from a trade-off induced by the sparsity objective: only tokens whose base embeddings produce insufficient task gradients are more likely to overcome the penalty and receive non-zero corrections

Our contributions are summarized as follows:

- We propose SEC, a framework that formulates embedding adaptation as a token-level selective correction problem, addressing the observation that alignment difficulty varies across tokens.
- We introduce a group-lasso regularized low-rank parameterization that achieves sparsity-driven token selection with only about 1% additional parameters.
- Extensive evaluations across NER, POS, and NLI tasks demonstrate that SEC outperforms strong baselines on multiple low-resource languages.

2 Related Works

To enable cross-language vocabulary transfer, existing approaches typically first initialize subword embeddings and then refine them through adaptive pre-training on the target language.

2.1 Vocabulary Expansion and Initialization

Embedding initialization is critical for cross-lingual transfer (Yamaguchi et al., 2024). One line of work uses static alignment, projecting target subword embeddings into a shared semantic space via multilingual static word vectors. Methods such as WECHSEL (Minixhofer et al., 2022), FOCUS (Dobler and de Melo, 2023), UniBridge (Pham et al., 2024), and LGSE (Teklehaymanot et al., 2026) exploit morphological and subword-level signals without bilingual supervision. However, these approaches remain inherently static and cannot dynamically correct tokens that remain misaligned after the initialization.

Another class of approaches predicts embeddings dynamically using hypernetworks trained to map external word representations into the model embedding space, such as OFA and HYPEROFA (Liu et al., 2024; Özeren et al., 2025), as well as ZeTT (Minixhofer et al., 2024) in zero-shot settings. While effective, these methods typically require sufficient training data to learn reliable mapping functions, which limits their applicability under extreme low-resource conditions.

A third category employs bilingual dictionaries or parallel resources for semantic alignment, as in Trans-Tokenizer (Remy et al., 2024) and Dictionaries to the Rescue (Sakajo et al., 2025). Although such methods often achieve strong alignment performance, their reliance on high-quality lexical resources restricts their use in many low-resource languages.

In this work, we adopt UniBridge (Pham et al., 2024) as the initialization backbone due to its strong balance between semantic and syntactic alignment without relying on external resources, providing a robust foundation for subsequent embedding-level correction.

2.2 Parameter-Efficient Cross-Lingual Transfer

Following initialization, the challenge is adapting pretrained multilingual models without overfitting

to extremely limited LRL data. Parameter-Efficient Fine-Tuning (PEFT) avoids full model updates by learning only a small number of task- or language-specific parameters (Houlsby et al., 2019). Early approaches like MAD-X (Pfeiffer et al., 2020) inserted adapter modules to capture language-specific knowledge, while LoRA (Hu et al., 2021) introduced low-rank matrix decomposition for efficient transfer.

Recent work has increasingly explored structured and selective parameter-efficient adaptation strategies. TIPA (Accou and Poelman, 2026) aggregates parameters based on linguistic typology for zero-shot transfer, but focuses on weight-space sharing rather than correcting embedding-level misalignment. SLoRA (Ding et al., 2023) and SLTrain (Han et al., 2024) employ sparse low-rank updates, and sparse subnetwork enhancement (Gurgurov et al., 2025) identifies critical subnets for underrepresented languages. While these methods improve internal weight efficiency, they do not selectively correct hard-to-align tokens in the embedding space. Circuit-targeted supervised fine-tuning (CT-SFT) (Nur'aini et al., 2026) demonstrates strong cross-lingual transfer by selectively updating task-relevant attention heads. However, its selection is at the circuit level, whereas our SEC targets the embedding token level.

In contrast, our work focuses on selective embedding correction, aiming to adapt semantically difficult or language-specific tokens while preserving already well-aligned representations.

3 Methodology

3.1 Formulation and Overview

The goal of SEC is to adapt a multilingual pretrained model to a target low-resource language (LRL) while preserving its cross-lingual alignment. Given a target vocabulary V, we aim to learn an embedding matrix $\mathbf{E} \in \mathbb{R}^{|V| \times d}$ (d: embedding dimension). We formulate the embedding $\mathbf{e}(w)$ for token $w \in V$ as an additive decomposition:

$$\mathbf{e}(w) = \mathbf{e}_{base}(w) + \Delta\mathbf{e}(w) \tag{1}$$

where $\mathbf{e}_{base}(w)$ is a fixed base embedding obtained from an initialization method (e.g., UniBridge (Pham et al., 2024)), and $\Delta\mathbf{e}(w)$ is a learnable correction.

The core design principle of our method is to ensure that $\Delta\mathbf{e}(w)$ focuses on adjusting word embeddings that remain misaligned with their ideal semantic positions after conventional multilingual adaptation. To this end, SEC combines three core mechanisms: (i) a low-rank parameterization that constrains embedding updates to a compact subspace, (ii) structured sparsity that selectively activates corrections at the word level, and (iii) a multi-objective optimization strategy that balances task performance, semantic stability, and selective correction. These components are designed to refine token representations relative to a fixed base embedding. An overview of the SEC framework is shown in Figure 2, and its components are introduced in the following sections.

3.2 Low-Rank Parameterization

We parameterize the semantic correction matrix $\Delta\mathbf{E} \in \mathbb{R}^{|V| \times d}$ using a low-rank structure. This design choice is motivated by the fact that directly learning $\Delta\mathbf{E}$ doubles embedding layer parameters, risking overfitting and memory overhead in low-resource scenarios, especially given that the embedding layer alone can constitute a substantial portion of model parameters (Tavares et al., 2025). LoRA hypothesizes that adaptation is intrinsically low-rank (Hu et al., 2021), a property recently validated across tasks (Zhang et al., 2023). By the same principle, we posit that the semantic shifts during cross-lingual embedding adaptation can also be captured within a low-dimensional subspace.

We decompose the correction matrix as:

$$\Delta\mathbf{E} = \mathbf{W}_A \mathbf{W}_B \tag{2}$$

where $\mathbf{W}_A \in \mathbb{R}^{|V| \times r}$ contains word-specific coefficients that capture language-specific semantic adjustments ($r \ll d$, e.g., $r = 8$, $d = 768$), and $\mathbf{W}_B \in \mathbb{R}^{r \times d}$ is a shared projection matrix that maps the low-rank coefficients back to the original embedding dimension. This method reduces the parameter complexity from $\mathcal{O}(|V| \cdot d)$ to $\mathcal{O}(|V| \cdot r + r \cdot d)$. Concretely, the compression ratio is:

$$\frac{|V|r + rd}{|V|d} = \frac{r}{d} + \frac{r}{|V|}$$

for typical values (e.g., $|V| \approx 10^5$, $d = 768$, $r = 8$), this ratio is approximately $0.0104 + 0.00008 \approx 1.05\%$, meaning SEC adds only about 1% additional parameters relative to the full embedding matrix. Moreover, the low-rank constraint implicitly regularizes the correction term, mitigating overfitting on limited LRL data.

Regarding the initialization, we adopt an asymmetric initialization strategy: $\mathbf{W}_A = \mathbf{0}$, $\mathbf{W}_B \sim$

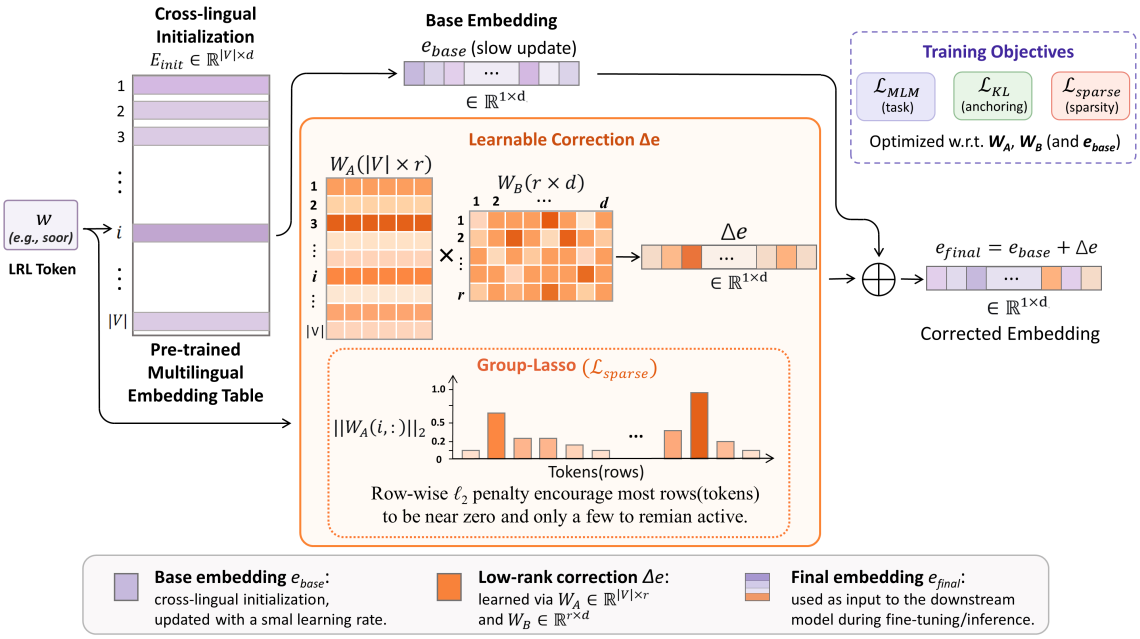

Figure 2: Overview of the SEC framework. Given a base embedding e_{base}, SEC learns a sparse low-rank residual correction $\Delta e = W_A W_B$ that selectively adjusts misaligned tokens while preserving stable cross-lingual representations. Group-Lasso regularization encourages token-level sparsity in W_A, enabling SEC to concentrate correction capacity on a small subset of hard-to-align tokens during adaptation.

$\mathcal{N}(0, \sigma^2)$. This ensures that $\Delta\mathbf{e}(w) = \mathbf{0}$ at the beginning of training, preserving the cross-lingual alignment encoded in the base embedding $\mathbf{e}_{base}$, while allowing LRL-specific offsets to emerge gradually during language-specific adaptation.

3.3 Adaptive Sparse Regularization

However, low-rank parameterization alone does not distinguish which words require correction. Thus, we need to selectively activate rather than adjust all words. To activate more meaningful language-specific tokens, we impose a group-lasso penalty (Yuan and Lin, 2006) on the rows of $\mathbf{W}_A$:

$$\mathcal{L}_{sparse} = \sum_{w \in V} \|\mathbf{W}_A(w)\|_2 \tag{3}$$

where $\mathbf{W}_A(w) \in \mathbb{R}^r$ denotes the row corresponding to token w. Since each row governs whether the associated token participates in the correction subspace, the row-wise ℓ_2 penalty encourages structured sparsity at the token level. Consequently, most tokens remain almost unchanged, while only a sparse subset receives substantial updates.

The selective behavior can be understood via the proximal gradient update:

$$\mathbf{W}_A^{(t+1)}(w) = \text{prox}_{\eta\lambda\|\cdot\|_2}\Big(\mathbf{W}_A^{(t)}(w) - \eta\nabla_{\mathbf{W}_A(w)}\mathcal{L}\Big) \tag{4}$$

where $\text{prox}_{\eta\lambda\|\cdot\|_2}(\mathbf{v}) = \max\left(0, 1 - \frac{\eta\lambda}{\|\mathbf{v}\|_2}\right)\mathbf{v}$ is the group-wise soft-thresholding operator. This operator pushes small-norm rows toward zero, so tokens with weak gradients receive negligible updates, while those with strong gradients achieve larger corrections. This trade-off concentrates correction capacity on misaligned tokens while leaving well-aligned ones mostly unchanged.

3.4 Optimization Objective and Strategy

We optimize SEC with a multi-objective loss that balances task adaptation, semantic stability, and sparsity:

$$\mathcal{L}_{total} = \mathcal{L}_{MLM} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{sparse}\mathcal{L}_{sparse} \tag{5}$$

where $\mathcal{L}_{MLM}$ is the masked language modeling loss (standard continued pre-training), $\mathcal{L}_{KL}$ penalizes KL divergence from the initial embedding distribution to prevent representational drift, and $\lambda_{KL}, \lambda_{sparse}$ are tuned on validation data.

We further employ differential learning rates: the base embeddings $\mathbf{e}_{base}$ are updated with a learning rate scaled by $\gamma \in (0, 1)$ relative to the default rate, while the correction term $\Delta\mathbf{e}$ uses the full learning rate. This slows changes to the well-aligned base while allowing flexible adaptation for the corrections.

Algorithm 1 summarizes the training procedure.

4 Experimental Setup

4.1 Base Models, Datasets and Evaluation Protocol

We evaluate SEC on mBERT (Pires et al., 2019) and XLM-R (Conneau et al., 2020) under the same experimental setting as UniBridge (Pham et al., 2024). The source languages are English, Chinese, Russian, Arabic, and Japanese. For target languages, we evaluate 14 low-resource languages from WikiANN (Rahimi et al., 2019) for named entity recognition (NER), 9 low-resource languages from Universal Dependencies (UD) (Nivre et al., 2016) for part-of-speech tagging (POS), and 10 languages from AmericasNLI (Ebrahimi et al., 2022) for natural language inference (NLI). Following prior work, each WikiANN target language contains only 100 labeled training samples.

For language adaptation, we use monolingual corpora collected from Wikipedia[2] for WikiANN and UD languages. Following UniBridge, we use up to 10K monolingual samples for target languages depending on corpus availability, and 50K samples for each source language. For AmericasNLI languages, we instead use the monolingual side of the translation data from (Mager et al., 2021). We train a SentencePiece tokenizer for each target language using its corresponding monolingual corpus.

For downstream tasks, models for NER are trained on the WikiANN training splits of all source languages and evaluated on the target-language test sets. Models for POS tagging are trained on the UD source-language training splits. For NLI, models are trained on the XNLI (Conneau et al., 2018) training splits of English, Chinese, Russian, and Arabic, and evaluated on the AmericasNLI target-language test sets. We use multi-source transfer with combined source-language training.

4.2 Evaluation Metrics

For NER on WikiANN, we report F1 score to measure the overlap between predicted and gold entity spans. For part-of-speech (POS) tagging on UD and natural language inference (NLI) on AmericasNLI/XNLI, we report token-level and instance-level accuracy, respectively. We use the same evaluation splits and metrics as prior work (Pham et al., 2024).

[2] https://huggingface.co/datasets/graelo/wikipedia

4.3 Baselines

We compare SEC against the following baselines, all built upon the same mBERT backbone. (Results on XLM-R are deferred to Appendix C.1 due to space constraints.)

- **mBERT** (Pires et al., 2019) The original multilingual BERT without multilingual adaptation, serving as a non-adapted baseline.
- **MAD-X (mBERT)** (Pfeiffer et al., 2020) An adapter-based parameter-efficient framework that enables cross-lingual transfer on top of a frozen mBERT backbone.
- **UniBridge (mBERT)** (Pham et al., 2024) A static cross-lingual embedding alignment and adaptation pipeline built upon mBERT.

4.4 Implementation Details

We broadly follow the training setup of UniBridge to maintain comparability. Learning rates and scheduling strategies are initialized from UniBridge's recommended ranges, with light tuning on each task's development set. For SEC, we choose the rank $r = 8$ across all experiments, while the sparsity weight λ_{sparse} and the DLR decay factor γ are selected using development performance. A complete hyperparameter specification and search ranges are provided in Appendix A.

5 Results and Analysis

5.1 Main Task Results

We conduct experiments on NER, POS tagging, and NLI. The results are shown in Table 1, Table 2, and Table 3, respectively.

Overall, SEC outperforms UniBridge on average across all three tasks. The largest improvement is on NER (Table 1), where SEC improves the average F1 by +9.57% (from 37.35% to 46.92%). This highlights the benefit of selective token-level correction for entity recognition. On the NLI task, SEC achieves a consistent improvement of +1.88% on average, which is expected given that this task primarily relies on sentence-level semantics rather than fine-grained lexical adjustments. SEC further improves POS tagging performance by +3.45% on average.

For NER, the greatest gains are observed in Amharic (amh) and North Frisian (frr), where entities have unique language-specific characteristics. The sparse regularizer effectively corrects

Language	amh	ang	cdo	crh	eml	frr	km	kn	lij	ps	sa	sd	si	so	AVG
mBERT	12.87	52.24	19.76	**47.81**	**39.71**	51.3	18.46	42.86	45.0	25.86	30.71	13.61	2.79	46.15	32.08
MAD-X(mBERT)	13.91	51.48	16.22	46.22	39.2	45.76	19.2	31.3	37.35	29.25	22.96	20.31	12.34	37.66	30.23
UniBridge(mBERT)	15.46	53.28	30.42	45.67	36.15	54.72	19.49	44.07	**45.49**	39.33	20.55	42.36	13.68	62.28	37.35
SEC(mBERT)	**31.08**	**61.12**	**36.99**	43.82	39.44	**59.36**	**40.17**	**55.93**	43.51	**47.91**	**32.64**	**51.88**	**48.22**	**64.84**	**46.92**

Table 1: F1 scores on 14 WikiANN target languages (mBERT backbone). Best results are shown in **bold** and second best results are underlined.

Language	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	AVG
mBERT	33.33	33.33	33.33	33.33	33.33	33.33	33.2	33.28	33.33	33.33	33.31
MAD-X(mBERT)	33.06	33.33	34.4	33.46	34.0	33.33	33.42	33.73	32.93	33.2	33.49
UniBridge(mBERT)	35.73	33.33	33.33	37.46	34.34	36.31	33.42	36.66	34.40	**34.4**	34.94
SEC(mBERT)	**37.47**	**33.73**	**36.13**	**39.60**	**35.20**	**39.57**	**35.70**	**40.93**	**35.73**	34.13	**36.82**

Table 2: Accuracy on 10 AmericasNLI target languages (mBERT backbone). Best results are shown in **bold** and second best results are underlined.

Language	amh	lij	olo	sa	sd	si	ta	tl	tt	AVG
mBERT	8.59	60.66	61.49	9.35	20.39	11.47	72.93	66.3	83.2	43.82
MAD-X(mBERT)	13.31	50.47	59.61	10.88	24.93	25.68	66.61	55.56	74.17	42.36
UniBridge(mBERT)	29.24	65.53	70.65	**12.86**	66.78	52.61	75.23	**70.65**	**84.16**	58.64
SEC(mBERT)	**34.53**	**66.17**	**75.20**	11.13	**77.88**	**71.82**	**82.74**	60.95	78.42	**62.09**

Table 3: Accuracy on 9 UD target languages (mBERT backbone). Best results are shown in **bold** and second best results are underlined.

low-frequency or morphologically complex subwords while preserving well-aligned ones.

While SEC consistently improves performance across all three tasks, we observe notable variability in POS tagging results across individual languages. SEC yields substantial gains for several languages, including Amharic (amh), Sindhi (sd), and Sinhala (si), but shows modest decreases for Sanskrit (sa), Tagalog (tl), and Tatar (tt), e.g., sa drops from 12.86% to 11.13%, tt from 84.16% to 78.42%. One reason is that Sanskrit (highly inflected) and Tatar (agglutinative) produce fragmented subword representations under standard tokenizers; with only a few thousand labeled POS samples, sparse gradient estimates can become unstable, leading to occasional degradation. Another reason is that the interaction between task gradient and sparsity penalty is sensitive to hyperparameters such as λ_{sparse} and the differential learning rate. While the chosen values work well on average, languages with extreme morphological complexity may benefit from per-language tuning or adaptive regularization. Future work could explore gated mechanisms or learnable per-token thresholds to better control correction magnitude. Despite these fluctuations, NER and NLI results remain robust, and the average POS improvement confirms SEC's overall effectiveness.

We also evaluate on XLM-R (full results in Appendix C.1). On XLM-R, SEC improves NER moderately (+1.44%) and POS marginally (+0.31%), but slightly hurts NLI (-1.47%). We attribute this mixed outcome to XLM-R's stronger multilingual pretraining and finer-grained tokenization, which already reduce embedding misalignment. Thus, selective correction yields smaller gains and may over-regularize on sentence-level tasks (NLI). Instead, these findings further clarify the scenarios where SEC is most effective: SEC brings clear gains when base embeddings are weaker (e.g., mBERT) while having little effect on stronger backbones.

5.2 Ablation Study Results

Table 4 compares SEC with its ablated variants (detailed results in Appendix C.3). The low-rank structural constraint is the most critical component: removing it (- low-rank) causes substantial performance drops, especially on NER and POS (over 5% on average), whereas NLI remains more robust due to its sentence-level nature. In contrast, removing selective activation (- $\mathcal{L}_{sparse}$) leads to a moderate but consistent degradation, suggesting that sparsity mainly focuses correction capacity rather than driving performance alone. Semantic anchoring

Models	Task		
	NER	NLI	POS
Full SEC	**46.92**	**36.82**	**62.09**
- $\mathcal{L}_{sparse}$	44.36	35.70	61.68
- $\mathcal{L}_{KL}$	45.52	35.31	57.98
- low-rank	41.36	34.55	56.61

Table 4: The performance of SEC on NER, NLI and POS tasks for different ablation experiments.

(- $\mathcal{L}_{KL}$) primarily provides stabilization, as its removal leads to a minor 1–2% degradation. Overall, SEC's effectiveness stems from its selective sparsity and low-rank design.

5.3 Token-Level Diagnostic Analysis

This section examines whether SEC fulfills its design goal of allocating embedding correction capacity selectively, i.e., concentrating the correction term $\Delta\mathbf{e}$ on tokens that are harder to align under uniform language adaptation, rather than distributing updates evenly across the vocabulary.

To identify tokens that are more likely to require explicit correction, we introduce a representation-level diagnostic that captures the mismatch between a token's static embedding and its contextual usage. Specifically, for each token w, let $\mathbf{e}(w)$ denote its static embedding obtained from the embedding layer after language-specific adaptation. Given an input context, $\mathbf{e}(w)$ is propagated through the Transformer layers to produce a contextualized representation $\mathbf{h}_{ctx}(w)$ at the final layer. We measure their discrepancy as:

$$\mathrm{CtxGap}(w) = \mathbb{E}_{ctx}\left[\|\mathbf{e}(w) - \mathbf{h}_{ctx}(w)\|_2\right]$$

although $\mathbf{e}(w)$ and $\mathbf{h}_{ctx}(w)$ reside at different processing stages, their discrepancy provides a practical proxy for contextual adaptation difficulty. Intuitively, if a token needs to be modified more extensively in different contexts, it means that the token is more difficult to align under normal circumstances.

Based on this score, we partition tokens into three groups (*Low*, *Mid*, and *High*) according to their alignment difficulty, and analyze how SEC allocates correction capacity across these groups. Specifically, we measure the activation rate of the correction term $\Delta\mathbf{e}$. As shown in Figure 3, tokens in the *Mid* and *High* groups consistently exhibit higher activation rates and larger correction magnitudes across languages, while tokens in the *Low* group are rarely activated. More detailed experimental results can be found in the Appendix C.4. These results indicate that SEC concentrates embedding corrections on tokens with higher alignment difficulty, rather than uniformly modifying the embedding space. We further complement this analysis with a sentence-level cross-lingual alignment study (Appendix C.2), which provides a direct evaluation of embedding quality.

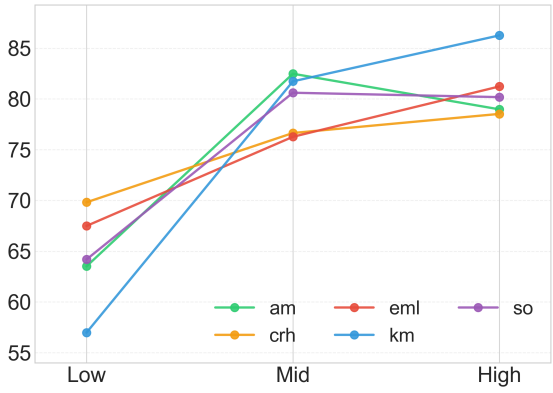

Figure 3: Activation rate of Δe across tokens grouped by CtxGap (Low/Mid/High) for five representative languages (amh, crh, eml, km, so). Tokens with higher alignment difficulty (Mid/High) receive corrections more frequently than well-aligned ones (Low).

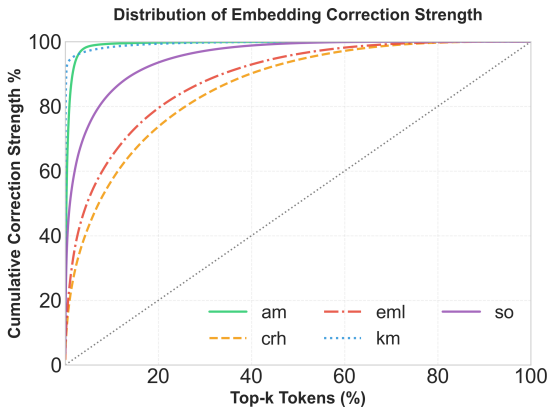

Figure 4: Cumulative distribution of embedding correction strength. Across representative languages, the top 20% of tokens account for more than 70% of the total correction strength.

5.4 Concentration of Embedding Corrections

To verify whether SEC indeed performs selective embedding correction, we analyze how the learned offset vectors Δe are distributed across the entire vocabulary.

For each token w, we compute the ℓ_2 norm of its correction vector $\Delta e(w)$, and use its squared norm $\|\Delta e(w)\|_2^2$ as a measure of correction strength. To-

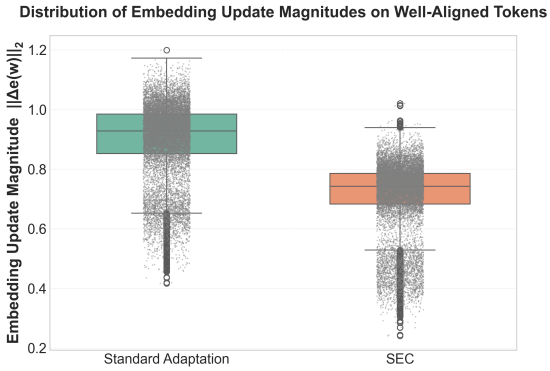

Figure 5: Distribution of embedding updates $\|\Delta \mathbf{e}(w)\|_2$ for well-aligned tokens. Compared to standard adaptation, SEC updates are consistently smaller and more concentrated.

	amh	crh	eml	km	so
Std. Adapt. $\|\Delta \mathbf{E}\|_2$	0.88	0.68	0.62	0.96	0.95
SEC (Sel.) $\|\Delta \mathbf{E}\|_2$	0.71	0.42	0.49	0.76	0.77
Reduction (%)	19.1	38.0	20.6	21.0	19.0

Table 5: Average ℓ_2 norm of embedding update magnitudes ($\|\Delta \mathbf{E}\|_2$) under well-aligned tokens, comparing standard language-specific adaptation (Std. Adapt.) and SEC across five languages.

kens are then ranked by their correction magnitude. For several representative languages, we visualize the cumulative distribution of correction strength, where the x-axis corresponds to the top $k\%$ of tokens and the y-axis represents the proportion of total correction strength contributed by these tokens. As shown in the Figure 4, a clear concentration pattern emerges: the top 20% of tokens account for more than 70% of the total correction strength across languages.

These results indicate that SEC concentrates most of its correction on a subset of key tokens, while making only minor adjustments to the majority of tokens.

5.5 Controlled Update Magnitudes

As shown in Sections 5.3 and 5.4, SEC preferentially allocates its correction power to poorly aligned tokens, resulting in highly concentrated updates. However, selectivity alone does not guarantee that well-aligned tokens are unaffected. Therefore, a supplementary question arises: does SEC avoid unnecessary updates to tokens that are already well aligned? To answer this question, we compare the overall magnitude of embedding updates induced by standard multilingual adaptation and SEC on well-aligned tokens. For each token w, we measure its embedding change relative to the ℓ_2 norm of the shared initialization. We report the mean and standard deviation of update magnitudes per language.

Table 5 summarizes the results across five languages. Across all languages, SEC yields smaller average embedding updates than standard fine-tuning, with reduction ratios ranging from 19.1% to 38.0%. More detailed experimental results can be found in the Appendix C.6. Beyond the aggregated statistics, Figure 5 visualizes the distribution of update magnitudes for well-aligned tokens. Specifically, SEC exhibits a lower median magnitude and a more compact distribution, indicating that unnecessarily large updates are effectively suppressed. This reduction aligns with the selective correction behavior observed in Sections 5.3 and 5.4. Together, these analyses characterize SEC's update strategy: while difficult tokens receive focused corrections, well-aligned tokens are largely preserved. Unlike standard multilingual adaptation, which uniformly propagates updates across the vocabulary, SEC explicitly limits redundant and unnecessary perturbations, achieving a more targeted and disciplined embedding adaptation.

6 Conclusion

In this work, we introduced SEC, a selective embedding correction framework for cross-lingual transfer in low-resource languages. Through sparse low-rank adaptation at the embedding level, SEC selectively refines semantically hard-to-align tokens while preserving already aligned representations.

Experimental results across multiple low-resource benchmarks show that uniform embedding adaptation is often suboptimal, as alignment difficulty varies substantially across tokens. By concentrating adaptation capacity on poorly aligned tokens, SEC achieves consistent improvements across diverse tasks and languages with minimal parameter overhead.

More broadly, our findings suggest that token-level selectivity is an important yet underexplored factor in cross-lingual embedding adaptation. We hope this perspective encourages future research on more adaptive and alignment-aware transfer mechanisms for low-resource multilingual modeling.

Limitations

Although SEC consistently improves cross-lingual transfer in low-resource settings, several limitations remain. First, the effectiveness of selective embedding correction varies across pretrained backbones. In particular, we observe smaller gains on models such as XLM-R, whose stronger multilingual pretraining and larger subword vocabulary may already alleviate part of the embedding misalignment problem. This suggests that the benefit of token-level correction is partially dependent on the underlying tokenizer and embedding space structure.

Second, SEC introduces an additional interaction between the task objective and the sparsity regularization term. In practice, the balance between these two signals can be sensitive to hyperparameters such as λ_{sparse} and the differential learning rate, especially for morphologically rich or highly fragmented languages. While a shared configuration works well overall, certain languages may benefit from more adaptive sparsity control or language-specific tuning strategies.

Future work may explore adaptive gating mechanisms or learnable token-wise thresholds to further improve the robustness and scalability of selective embedding correction.

References

Stef Accou and Wessel Poelman. 2026. TIPA: Typologically informed parameter aggregation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2253–2267, Rabat, Morocco. Association for Computational Linguistics.

Pavel Chizhov, Catherine Arnett, Elizaveta Korotkova, and Ivan P. Yamshchikov. 2024. BPE gets picky: Efficient vocabulary refinement during tokenizer training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16587–16604, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2023. Sparse low-rank adaptation of pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4133–4145, Singapore. Association for Computational Linguistics.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Daniil Gurgurov, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2025. Sparse subnetwork enhancement for underrepresented languages in large language models. *Preprint*, arXiv:2510.13580.

Andi Han, Jiaxiang Li, Wei Huang, Mingyi Hong, Akiko Takeda, Pratik Kumar Jawanpuria, and Bamdev Mishra. 2024. Sltrain: a sparse plus low rank approach for parameter and memory efficient pretraining. *Advances in Neural Information Processing Systems*, 37:118267–118295.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 6282–6293, Online. Association for Computational Linguistics.

Sander Land and Max Bartolo. 2024. Fishing for magikarp: Automatically detecting under-trained tokens in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11631–11646, Miami, Florida, USA. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. 2024. Zero-shot tokenizer transfer. *Advances in Neural Information Processing Systems*, 37:46791–46818.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Khumaisa Nur'aini, Ayu Purwarianti, Alham Fikri Aji, and Derry Wijaya. 2026. Mechanisms are transferable: Data-efficient low-resource adaptation via circuit-targeted supervised fine-tuning. *Preprint*, arXiv:2601.08146.

Enes Özeren, Yihong Liu, and Hinrich Schuetze. 2025. HYPEROFA: Expanding LLM vocabulary to new languages via hypernetwork-based embedding initialization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 79–96, Vienna, Austria. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Thirty-seventh Conference on Neural Information Processing Systems*, NIPS '23, pages 36963–36990, Red Hook, NY, USA. Curran Associates Inc.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3184, Bangkok, Thailand. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *Preprint*, arXiv:2408.04303.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3118–3135, Online. Association for Computational Linguistics.

Haruki Sakajo, Yusuke Ide, Justin Vasselli, Yusuke Sakai, Yingtao Tian, Hidetaka Kamigaito, and Taro Watanabe. 2025. Dictionaries to the rescue: Cross-lingual vocabulary transfer for low-resource languages using bilingual dictionaries. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25963–25976, Vienna, Austria. Association for Computational Linguistics.

Diogo Tavares, David Semedo, Alexander Rudnicky, and Joao Magalhaes. 2025. Language models can be efficiently steered via minimal embedding layer transformations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22949–22967, Suzhou, China. Association for Computational Linguistics.

Hailay Teklehaymanot, Dren Fazlija, and Wolfgang Nejdl. 2026. Lgse: Lexically grounded subword embedding initialization for low-resource language adaptation. *Preprint*, arXiv:2603.22629.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046, Marseille, France. European Language Resources Association.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024. An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6760–6785, Miami, Florida, USA. Association for Computational Linguistics.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zhong Zhang, Bang Liu, and Junming Shao. 2023. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1713, Toronto, Canada. Association for Computational Linguistics.

A Hyperparameters and Training Details

All experiments are conducted on NVIDIA RTX A6000 GPUs. Training time varies across languages and tasks; a single language typically takes between 0.5 to 2 hours. We present the hyperparameters for training and inference for SEC and all the baselines' configurations in Table 6 and Table 7.

Hyperparameter	Value
Initial vocab size v_i	7000
Maximum vocab size v_m	60000
Increased step of vocab size δ_v	1000
Threshold for stopping the algorithm ϵ_v	5.0

Table 6: The hyperparameter for vocabulary size searching process.

Hyperparameter	Value
Static embedding model	FastText
Static embedding dimension	300
Number of epoch of training	3

Table 7: The hyperparameter for embedding initialization stage.

B Additional Discussion

B.1 Why focus on low-resource languages?

Although SEC is technically applicable to any language, we specifically position it in the low-resource setting because the challenges it addresses are particularly pronounced there.

In high-resource languages, large amounts of task-specific data allow standard fine-tuning to gradually adjust token embeddings through sufficient gradient signals. However, in low-resource languages (LRLs), two issues arise:

- The pretrained embeddings often exhibit systematic semantic misalignment due to limited pretraining exposure.
- The available task data is too limited to safely perform unconstrained embedding updates without overfitting or distorting well-initialized representations.

SEC is designed precisely for this regime. By decomposing embeddings into a frozen base component and a learnable low-rank residual, and enforcing token-level sparsity via group-lasso regularization, SEC enables controlled and selective corrections. Only tokens exhibiting substantial contextual discrepancy receive updates, while stable embeddings remain untouched.

This correction mechanism is particularly beneficial when training data is scarce, as it prevents unnecessary global shifts in the embedding space.

B.2 Choice of low-rank dimension

We do not assume that $r = 8$ is universally optimal across languages with different morphological or script characteristics. In fact, the choice of the rank

Algorithm 1 Training procedure of SEC

Require: Base embedding matrix $\mathbf{E}_{base}$, monolingual corpus $\mathcal{D}$ for target LRL, rank r, sparsity weight λ_{sparse}, KL weight λ_{KL}, scaling factor γ

Initialize $\mathbf{W}_A = \mathbf{0}$, $\mathbf{W}_B \sim \mathcal{N}(0, \sigma^2)$
Set the learning rate of $\mathbf{E}_{base}$ to $\gamma \cdot \eta$ (differential LR)
for each training step **do**
  Sample a batch from $\mathcal{D}$
  Compute $\mathbf{e}(w) = \mathbf{e}_{base}(w) + (\mathbf{W}_A\mathbf{W}_B)(w)$ for all w in batch
  Forward pass through the model to compute $\mathcal{L}_{MLM}$
  Compute $\mathcal{L}_{KL}$ and $\mathcal{L}_{sparse}$ (Eq. 3)
  $\mathcal{L}_{total} = \mathcal{L}_{MLM} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{sparse}\mathcal{L}_{sparse}$
  Backward pass and update $\mathbf{W}_A$, $\mathbf{W}_B$, and $\mathbf{E}_{base}$
end for
return Corrected embedding matrix $\mathbf{E} = \mathbf{E}_{base} + \mathbf{W}_A\mathbf{W}_B$

Hyperparameter	Value
Number of epochs trained	10
Batch size	32
Learning rate	1e-5

Table 8: The configuration of the pre-trained LMs' fine-tuning on source downstream task and zero-shot transfer to target language.

Hyperparameter	Value
Invertible adapter reduced factor	2
Language adapter reduced factor	2
KL divergence weight β	1.0
Masked probability	0.15
Number of epochs trained	50
Batch size	32
Base embedding learning rate	{5e-4, 3e-4}
Correction learning rate	{5e-3, 2.5e-3, 1e-3, 5e-4}
Sparsity weight λ_{sparse}	{7.5e-5, 5e-5}
Low-rank dimension r	8

Table 9: Hyperparameters for language adaptation. Adapter dimensions are scaled by a reduction factor, with learning rates tuned individually per language to account for varying representation in the pre-trained LM.

Hyperparameter	Value
Task adapter reduced factor	16
Number of epochs trained	20
Batch size	32
Learning rate	{2e-4, 1e-3}

Table 10: Hyperparameters for task adaptation. Since languages are unevenly represented in pre-trained LMs, we employ language-specific learning rates for optimal adaptation.

is empirical rather than theoretically tied to specific linguistic properties.

In early experiments, we initially introduced a full-dimensional residual vector (i.e., same dimension as the original token embedding). While this improved performance, it was unclear whether the gains came from selective correction or simply from increased parameter capacity. To better control model capacity and isolate the effect of structured correction, we adopted a low-rank parameterization.

r	$r=4$	$r=6$	$r=8$	$r=10$	$r=12$	$r=14$	$r=16$
result	46.26	43.22	**46.92**	43.86	44.48	42.09	45.09

Table 11: Results of F1 score on the NER task with mBERT across different values of r. The best performance is achieved at $r = 8$.

We conducted a hyperparameter sweep over $r \in \{4, 6, 8, 10, 12, 14, 16\}$. Across languages and tasks, $r = 8$ provided the best average performance, offering a good balance between expressiveness and regularization. Smaller ranks limited correction capacity, while larger ranks introduced mild overfitting in low-resource settings. For illustrative purposes, Table 11 reports the results on the NER task using mBERT as the backbone; analogous sweeps on XLM-R and across other tasks (POS, NLI) consistently favored $r = 8$ as well, which we therefore adopt uniformly across all experiments.

Thus, $r = 8$ should be understood as an empirically chosen trade-off value rather than a linguistically motivated constant.

B.3 Extremely low-resource settings

To evaluate the robustness of SEC with smaller adaptation corpora, we subsample target-language monolingual data to 1K and 3K sentences for se-

lected languages. Table 12 reports the downstream performance under these reduced data settings. The results show that SEC maintains improvements over the baseline even with severely limited data, although gains are naturally smaller compared to using the full corpus. This demonstrates the method's effectiveness in low-resource scenarios.

C More Experiments and Ablation Study

C.1 Performance of SEC on XLM-R

To further evaluate the effectiveness of our proposed SEC method, we conduct additional experiments by applying it to the XLM-R backbone. The detailed results of SEC on XLM-R are presented in Tables 13, Table 14 and Table 15.

C.2 Cross-lingual Sentence-level Alignment

To directly assess the quality of learned embeddings, we evaluate cross-lingual semantic alignment using parallel sentence pairs. In low-resource settings, intrinsic evaluation of embedding quality is challenging due to the lack of reliable lexical-level benchmarks. Therefore, we adopt sentence-level cross-lingual similarity as a practical proxy: if embeddings are better aligned, semantically equivalent sentences across languages should be mapped to closer representations in the shared embedding space.

We conduct this analysis on the same 14 low-resource languages used in the NER experiments. For each language, we construct a set of English–target language sentence pairs (approximately 10–20 pairs) based on the evaluation data.

Given a sentence s, we obtain its representation $\mathbf{h}_s \in \mathbb{R}^d$ by mean-pooling the final-layer hidden states of the Transformer encoder:

$$\mathbf{h}_s = \frac{1}{L} \sum_{i=1}^{L} \mathbf{H}_i \tag{6}$$

where $\mathbf{H} \in \mathbb{R}^{L \times d}$ denotes the sequence of contextualized token representations. For each sentence pair (s_{en}, s_{target}), we compute the cosine similarity between $\mathbf{h}_{en}$ and $\mathbf{h}_{target}$.

As shown in Table 16, SEC improves cross-lingual similarity in the majority of languages, with an overall increase in average similarity. This indicates that SEC leads to better alignment between semantically equivalent sentences across languages, suggesting improved embedding quality. Notably, the gains are more pronounced in several typologically diverse languages (e.g., ang, frr, km, si), while only minor decreases are observed in a small number of cases. Overall, these results consistently demonstrate that SEC enhances cross-lingual representation quality.

C.3 Detailed Ablation Study of SEC on mBERT

We further perform an ablation study to analyze the individual contributions of SEC's components when applied to mBERT. The results are displayed in Table 17, Table 18 and Table 19.

C.4 Token-Level Alignment Difficulty Statistics

To investigate whether SEC selectively adjusts poorly aligned tokens, we analyzed the activation statistics of $\Delta\mathbf{e}$ in three misaligned groups. Detailed results for all 14 languages are shown in Table 20.

C.5 Top-Token Correction Strength Statistics

This appendix reports the contribution of the top 20% tokens to the total embedding correction strength for all languages, corresponding to the analysis in §5.4. The results are summarized in Table 21.

C.6 Controlled Update Magnitude Statistics

This appendix presents the quantitative results for the update magnitude analysis in §5.5. Table 22 reports the mean update magnitudes $\|\Delta\mathbf{E}\|_2$ on well-aligned tokens across all languages, while Table 23 provides the corresponding standard deviations. Together, these results show that SEC suppresses unnecessary updates on well-aligned tokens while preserving focused corrections for difficult tokens.

Language	amh	ang	cdo	crh	eml	frr	km	kn	lij	ps	sa	sd	si	so	AVG
1K Adaptation Samples															
UniBridge	0.0	**29.34**	5.23	15.60	23.58	40.0	0.0	27.23	21.09	**3.05**	**14.02**	**4.06**	0.0	44.78	16.32
SEC	**1.40**	27.07	**12.45**	**23.67**	**24.19**	**44.22**	**11.24**	**32.08**	**25.31**	0.0	13.77	3.10	**2.55**	**48.46**	**19.25**
Δ	+1.40	-2.27	+7.22	+8.07	+0.61	+4.11	+11.24	+4.85	+4.22	-3.05	-0.25	-1.05	+2.55	+3.68	+2.93
3K Adaptation Samples															
UniBridge	1.16	44.72	18.23	32.5	31.62	**49.77**	20.66	45.30	27.00	26.79	**24.31**	33.85	13.89	50.23	30.00
SEC	**8.65**	**57.92**	**24.60**	**40.75**	**32.03**	47.44	**31.05**	**48.76**	**37.69**	**30.69**	23.40	**43.17**	**24.67**	**53.68**	**36.04**
Δ	+7.49	+13.20	+6.37	+8.25	+0.41	-2.33	+10.39	+3.46	+10.69	+3.90	-0.91	+9.32	+10.78	+3.45	+6.03

Table 12: F1 scores on 14 WikiANN target languages with different amounts of monolingual adaptation data using mBERT as the backbone. Best results are shown in **bold**.

Language	amh	ang	cdo	crh	eml	frr	km	kn	lij	ps	sa	sd	si	so	AVG
XLM-R	43.31	52.71	22.04	44.62	40	44.17	40.69	45.34	40.45	46.0	**41.28**	43.13	50.03	50.23	43.14
MAD-X(XLM-R)	39.3	46.59	17.32	36.63	33.86	39.51	**50.0**	45.24	38.13	42.66	19.93	39.06	39.55	49.6	38.38
UniBridge(XLM-R)	**49.6**	43.24	**42.91**	46.03	40.15	50.67	42.67	**48.72**	**45.16**	46.09	29.74	**51.32**	52.86	54.22	45.95
SEC(XLM-R)	47.46	**54.09**	40.35	**48.86**	**40.68**	**61.68**	48.34	47.22	37.30	**46.61**	37.15	41.51	52.09	**60.09**	**47.39**

Table 13: F1 scores on 14 WikiANN target languages (XLM-R backbone). Best results are shown in **bold** and second best results are underlined.

Language	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	AVG
XLM-R	36.26	38.53	36.4	37.33	**37.33**	39.43	36.89	37.6	35.86	34.66	37.03
MAD-X(XLM-R)	39.46	36.8	38.93	39.73	35.86	40.78	33.42	37.46	**39.06**	36.53	37.80
UniBridge(XLM-R)	**52.13**	36.8	**40.26**	**59.59**	35.86	46.88	**42.38**	**59.86**	35.6	36.4	**44.58**
SEC(XLM-R)	40.93	**38.93**	36.67	56.93	36.13	**50.81**	36.63	54.43	38.4	**41.20**	43.11

Table 14: Accuracy on 10 AmericasNLI target languages (XLM-R backbone). Best results are shown in **bold** and second best results are underlined.

Language	amh	lij	olo	sa	sd	si	ta	tl	tt	AVG
XLM-R	46.02	39.15	60.69	32.9	70.01	**76.25**	**85.53**	67.45	59.89	59.77
MAD-X(XLM-R)	**47.72**	58.28	69.48	36.1	71.2	73.86	83.85	69.01	65.83	63.93
UniBridge(XLM-R)	40.88	73.75	**81.45**	**38.94**	**71.73**	63.52	83.5	**72.62**	**81.3**	67.52
SEC(XLM-R)	44.71	**77.24**	79.57	33.91	70.84	69.66	85.10	70.94	78.46	**67.83**

Table 15: Accuracy on 9 UD target languages (XLM-R backbone). Best results are shown in **bold** and second best results are underlined.

Language Pair	Baseline	SEC	Δ
en–amh	0.3798	0.3876	+0.0078
en–ang	0.3544	0.3877	+0.0333
en–cdo	0.2664	0.2664	+0.0000
en–crh	0.4491	0.4265	-0.0226
en–eml	0.3294	0.3418	+0.0124
en–frr	0.3352	0.3624	+0.0272
en–km	0.4415	0.4760	+0.0345
en–kn	0.5177	0.5276	+0.0099
en–lij	0.5049	0.5110	+0.0061
en–ps	0.4891	0.4918	+0.0027
en–sa	0.4348	0.4409	+0.0061
en–sd	0.4772	0.4864	+0.0092
en–si	0.4511	0.4779	+0.0268
en–so	0.5026	0.4859	-0.0167
Average	**0.4238**	**0.4336**	**+0.0098**

Table 16: Sentence-level cross-lingual similarity as a proxy for embedding quality.

Language	amh	ang	cdo	crh	eml	frr	km	kn	lij	ps	sa	sd	si	so	AVG
SEC(mBERT)	**31.08**	61.12	36.99	43.82	39.44	**59.36**	**40.17**	**55.93**	**43.51**	**47.91**	32.64	**51.88**	48.22	64.84	**46.92**
- $\mathcal{L}_{sparse}$	30.15	52.31	28.08	**48.06**	37.63	58.93	33.06	50.64	41.91	38.06	33.10	**51.72**	51.01	66.07	44.36
- $\mathcal{L}_{KL}$	23.39	**63.49**	**40.53**	45.59	33.56	54.31	39.31	52.77	36.5	41.94	34.03	47.65	**53.11**	**71.11**	45.52
- low-rank	19.66	51.32	35.44	36.8	**40.31**	49.11	28.95	54.46	41.88	40.34	**34.04**	48.59	37.82	60.27	41.36

Table 17: The results of the F1 score for SEC and its ablated variants with mBERT as a backbone are reported on WikiANN across 14 low-resource languages. Best results are shown in **bold** and second best results are underlined.

Language	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	AVG
SEC(mBERT)	**37.47**	33.73	**36.13**	39.60	**35.20**	39.57	**35.70**	**40.93**	**35.73**	34.13	**36.82**
- $\mathcal{L}_{sparse}$	33.33	**33.87**	**36.13**	38.0	33.73	**40.24**	35.03	35.87	**35.73**	**35.07**	35.70
- $\mathcal{L}_{KL}$	33.33	32.67	34.0	**42.4**	34.27	37.40	35.03	34.27	35.47	34.27	35.31
- low-rank	33.73	33.33	34.13	39.06	33.33	38.35	33.29	33.73	34.27	32.27	34.55

Table 18: The results of accuracy for SEC and its ablated variants with mBERT as a backbone are reported on AmericasNLI across 10 low-resource languages. Best results are shown in **bold** and second best results are underlined.

Language	amh	lij	olo	sa	sd	si	ta	tl	tt	AVG
SEC(mBERT)	**34.53**	66.17	**75.20**	**11.13**	77.88	71.82	**82.74**	60.95	78.42	**62.09**
- $\mathcal{L}_{sparse}$	33.74	66.03	74.73	11.11	77.18	71.48	81.81	59.97	**79.03**	61.68
- $\mathcal{L}_{KL}$	32.99	**69.10**	63.84	10.60	77.37	**72.72**	62.50	**63.52**	69.17	57.98
- low-rank	34.23	64.44	60.35	10.54	75.55	66.82	72.91	58.49	66.14	56.61

Table 19: The results of accuracy for SEC and its ablated variants with mBERT as a backbone are reported on the UD dataset across 9 low-resource languages. Best results are shown in **bold** and second best results are underlined.

Language	amh	ang	cdo	crh	eml	frr	km	kn	lij	ps	sa	sd	si	so
Low	60.81	67.27	59.98	69.14	67.42	63.81	59.38	66.29	63.66	64.26	64.26	63.21	61.49	64.86
Mid	83.76	78.14	81.21	77.09	78.06	79.79	80.46	80.31	79.86	81.06	81.36	80.91	82.26	78.67
High	80.40	79.57	83.78	78.75	79.50	81.38	85.14	78.38	81.46	79.65	79.35	80.85	81.23	81.46

Table 20: Activation rates of $\Delta\mathbf{e}$ for 14 low-resource languages, with tokens grouped into low, medium, and high alignment-difficulty buckets as measured by contextual discrepancy (CtxGap).

Language	amh	ang	cdo	crh	eml	frr	km	kn	lij	ps	sa	sd	si	so
Top-20% Strength (%)	99.72	82.22	86.03	73.70	79.44	99.03	99.35	97.40	97.09	99.34	93.15	99.08	96.80	93.56

Table 21: Contribution of the top 20% tokens to the total embedding correction strength across all languages.

	amh	ang	cdo	crh	eml	frr	km	kn	lij	ps	sa	sd	si	so
Std. Adapt.	0.880	0.468	0.582	0.681	0.627	0.939	0.967	0.857	0.820	0.764	0.866	0.682	0.841	0.950
SEC (Sel.)	0.712	0.623	0.582	0.423	0.498	0.679	0.764	0.720	0.631	0.716	0.704	0.671	0.720	0.770
Reduction (%)	19.1	-33.2	-0.04	38.0	20.6	27.7	21.0	15.9	23.0	6.2	18.7	1.6	14.3	19.0

Table 22: Mean ℓ_2 norm of embedding update magnitudes ($\|\Delta\mathbf{E}\|_2$) on well-aligned tokens under standard language-specific adaptation and SEC. Reduction is computed based on mean values; negative values indicate increased update magnitude under SEC.

	amh	ang	cdo	crh	eml	frr	km	kn	lij	ps	sa	sd	si	so
Std. Adapt. (Std)	0.057	0.062	0.056	0.057	0.055	0.064	0.070	0.053	0.058	0.056	0.058	0.049	0.058	0.065
SEC (Sel.) (Std)	0.047	0.047	0.056	0.048	0.047	0.047	0.056	0.045	0.047	0.053	0.049	0.044	0.048	0.056

Table 23: Standard deviation of embedding update magnitudes ($\|\Delta\mathbf{E}\|_2$) on well-aligned tokens under standard language-specific adaptation and SEC across languages.