

Enhancing Cross-lingual Transfer via Phonemic Transcription Integration

Hoang H. Nguyen¹, Chenwei Zhang², Tao Zhang¹, Eugene Rohrbaugh³, Philip S. Yu¹

¹ Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

² Amazon, Seattle, WA, USA

³ Harrisburg University of Science and Technology, Harrisburg, PA, USA

{hnguy7, tzhang90, psyu}@uic.edu, cwzhang@amazon.com, gene.rohrbaugh@gmail.com

Abstract

Previous cross-lingual transfer methods are restricted to orthographic representation learning via textual scripts. This limitation hampers cross-lingual transfer and is biased towards languages sharing similar well-known scripts. To alleviate the gap between languages from different writing scripts, we propose **PhoneXL**, a framework incorporating phonemic transcriptions as an additional linguistic modality beyond the traditional orthographic transcriptions for cross-lingual transfer. Particularly, we propose unsupervised alignment objectives to capture (1) local one-to-one alignment between the two different modalities, (2) alignment via multi-modality contexts to leverage information from additional modalities, and (3) alignment via multilingual contexts where additional bilingual dictionaries are incorporated. We also release the first phonemic-orthographic alignment dataset on two token-level tasks (Named Entity Recognition and Part-of-Speech Tagging) among the understudied but interconnected Chinese-Japanese-Korean-Vietnamese (CJKV) languages. Our pilot study reveals phonemic transcription provides essential information beyond the orthography to enhance cross-lingual transfer and bridge the gap among CJKV languages, leading to consistent improvements on cross-lingual token-level tasks over orthographic-based multilingual PLMs.¹

1 Introduction

Despite recent advances in cross-lingual pre-trained language models (PLM) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), PLMs remain heavily biased towards high-resourced languages due to the skewed amount of available pre-training data under parameter capacity constraints. This heavily affects the downstream task performance of less-represented languages during pre-training. In addition, as most

¹Our code and datasets are publicly available at https://github.com/nhhoang96/phonemic_xlingual

Table 1: Orthographic and Phonemic representations² of name entities across CJKV languages. **Blue** and **Red** denote pre-segmented phrases that share similar meanings.

	Orthographic	Phonemic
EN	electronic industry	ilektrɔnik ɪndəstri
ZH (src)	电子 行业	tʃɛn tsu xɑŋ jɛ
VI (tgt)	Công nghiệp Điện tử	kɔŋ nɪəp ðiən tu
JA (src)	電子 産業	dɛnʃi sɑɛnju
KO (tgt)	전자 산업	ʧɛɑnjə sɑɛniawp
EN	Vietnam News Agency	viɛtnam nuz ɛjʒənsi
ZH (src)	越南 通讯社	ʈɛ nan tʰoŋ cʏn sʏ
VI (tgt)	Thông tấn xã Việt Nam	tʰoŋ tʏn sɑ viət nam
JA (src)	ベトナム 通信社	bitɔnamu tsɑufɪmjə
KO (tgt)	베트남 통신사	betunəm tɔŋmsɔ

high-resourced languages share the similar scripts (i.e. mostly Latin scripts), PLMs tend to perform well on languages sharing similar scripts with those languages (Pires et al., 2019; Muller et al., 2021; Fujinuma et al., 2022). As most challenging low-resourced languages do not share similar scripts with common high-resourced languages, this limitation leaves them significantly behind.

To alleviate the challenges on low-resource or zero-resource target languages, recent works attempt to transfer knowledge from high-resourced languages (typically English (EN)) to low-resourced target languages via augmentations from bilingual dictionaries and parallel corpora. However, these approaches are restricted to English source language and results in less significant performance gain on languages that are more distant from English (Yang et al., 2022; Fujinuma et al., 2022). Languages can be considered distant due to the differences in orthography, phonology, morphology and grammatical structures. The fact that performance drop occurs on distant languages

²For the sake of simplicity and clearer comparison of phonemic representation similarity between source and target languages, we omit the tonal IPA characters. Tonal IPA characters are preserved as a part of the phonemic inputs for tokenization and training purposes.

when transferring from English indicates that additional works are needed to exploit connectivity between closely related languages, especially under extreme parameter constraints.

Besides purely relying on the orthographic representation in the format of written scripts, additional modality of languages such as articulatory signals can provide essential information beyond the written scripts to enhance textual representations (Bharadwaj et al., 2016; Chaudhary et al., 2018). Phonemic transcriptions which capture linguistic articulatory signals are beneficial to understanding non-Latin-based languages when integrated with PLMs (Sun et al., 2021). They can also facilitate for knowledge transfer between languages sharing lexical similarities in phonology but possessing different writing scripts. As demonstrated in Table 1, despite differences in orthographic representations, the terms “电子” (ZH) and “Điện tử” (VI) possess significant phonemic similarities when encoded into International Phonetic Alphabet (IPA). Similarly, although “ベトナム” (JA) and “베트남” (KO) are different, their phonemic representations (“bɛtənamu” and “bɛtunæm” respectively) are almost identical in terms of articulatory features.

Motivated by the inherent lexical similarities in terms of phonology among CJKV languages, we propose a novel cross-lingual transfer framework to integrate and synthesize two specific linguistic modalities (1) textual orthographic input scripts, (2) phonemic transcription, represented in International Phonetic Alphabet (IPA) format. Our unified cross-lingual transfer framework aims to effectively (1) align both orthographic and phonemic transcriptions via multi-modality learning, (2) capture additional alignment between the two modalities via contextual information, (3) enhance cross-lingual transfer of the two modalities with additional bilingual dictionary. Our work specifically targets Chinese-Vietnamese-Japanese-Korean languages which are not well-studied in cross-lingual transfer and possess lexical similarities with one another in terms of phonology. Our contributions can be summarized as follows:

- We provide the first pre-processed orthographic-phonemic transcription alignment dataset for token-level tasks (i.e. Part-of-Speech Tagging (POS) and Named Entity Recognition (NER)) among CJKV languages (Chinese-Japanese-Korean-Vietnamese).

- We propose a multi-modality learning paradigm with unsupervised alignment objectives to fuse the knowledge obtained from both modalities/ transcriptions to enhance cross-lingual transfer.
- Our proposed framework yields consistent improvements over the orthographic-based multilingual PLMs (mBERT and XLM-R) on both POS and NER tasks.

2 Related Work

Cross-lingual Transfer Recent works in Cross-lingual transfer focus on generating multilingual contextualized representation for different languages based on the Pre-trained Language Models (PLM) via bilingual dictionaries (Qin et al., 2021) and/or machine translation approaches (Fang et al., 2021; Yang et al., 2022). Qin et al. (2021) proposes a comprehensive code-switching technique via random selection of languages, sentences, and tokens to enhance multilingual representations, leading to improved performance on target languages on different downstream tasks. On the other hand, other approaches leverage parallel corpora generated by Machine Translation to (1) distill knowledge from source languages to target languages (Fang et al., 2021) or augment source language data with target language knowledge during training (Yang et al., 2022; Zheng et al., 2021). However, current cross-lingual efforts concentrate on single source language (EN) to multiple target languages. Under parameter capacity constraints, cross-lingual transfer has been shown to be biased towards high-resourced languages which share similar scripts and possess larger corpora of unlabeled data during pre-training (Fujinuma et al., 2022). Unlike previous works, we specifically target enhancing performance of low-resourced languages by exploiting inherent linguistic similarities between closely-connected languages (Nguyen and Rohrbaugh, 2019; Zampieri et al., 2020).

Multi-modality Learning Multi-modality learning (Radford et al., 2021; Li et al., 2021, 2022) was initially proposed for the task of Visual-Question Answering (Goyal et al., 2017). The objective is to find alignment between the given images and textual input (i.e. caption). The two aligned modalities are trained to maximize the agreement with ground truth textual-image alignment. Despite its simple objectives, the CLIP (Radford et al., 2021)

pre-training mechanism is considered the state-of-the-art in multi-modality representation learning. Motivated by multi-modality learning, we integrate multi-modality learning approaches in unifying two modalities of transcriptions (orthographic and phonemic) for better representation enrichment.

3 Problem Formulation

In this work, we study the problem of Cross-lingual Transfer in a bilingual setting where there exists annotated data collection of high-resource language S, namely $D_S^{train} = \{(X_i^{(S)}, Y_i^{(S)})\}_{i=1}^{N_s}$, and unlabeled data collection of low-resource target language T, denoted as $D_T^{test} = \{(X_j^{(T)})\}_{j=1}^{N_t}$. N_s, N_t denote the sample size of source language training data and target language inference dataset respectively.

Formally, given an i -th input source language utterance with the length of M orthographic tokens $x_i^{(S)} = [x_{i,1}^{(S)}, x_{i,2}^{(S)}, \dots, x_{i,M}^{(S)}]$ and the corresponding phonemic transcriptions $z_i^{(S)} = [z_{i,1}^{(S)}, z_{i,2}^{(S)}, \dots, z_{i,M}^{(S)}]$ and token-level labels $y_i^{(S)} = [y_{i,1}^{(S)}, y_{i,2}^{(S)}, \dots, y_{i,M}^{(S)}]$, the overall training objective is summarized as:

$$\theta_S = \underset{\theta}{\operatorname{argmin}} \frac{1}{N_s} \sum_{i=1}^{N_s} l(F(x_i^{(S)}, z_i^{(S)}; \theta), y_i^{(S)}) \quad (1)$$

where $F(\cdot)$ denotes the transformation function that takes an input of both $x_i^{(S)}, z_i^{(S)}$ to output probability prediction of label $y_i^{(S)}$ for individual tokens. θ denotes the parameters of the transformation framework and $l(\cdot)$ is the token-level cross-entropy loss.

The overall trained framework is then evaluated in a zero-shot setting on target language T as follows:

$$p(y^{(T)} | x^{(T)}, z^{(T)}) = \underset{k}{\operatorname{argmax}} F(x^{(T)}, z^{(T)}; \theta_S) \quad (2)$$

where k denotes the label space of token-level tasks. In our cross-lingual transfer settings, no labeled target data or parallel data between source and target language is leveraged during training.

4 Proposed Framework

In this section, we introduce our proposed Cross-lingual Transfer Framework, namely **PhoneXL**, with 3 major learning objectives: (1) Orthographic-Phonemic Alignment (\mathcal{L}_{align}), (2) Contextual

Cross-modality Alignment (\mathcal{L}_{MLM}), (3) Contextual Cross-lingual Alignment (\mathcal{L}_{XMLM}). The overall learning objective is summarized as follows:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha \mathcal{L}_{align} + \beta \mathcal{L}_{MLM} + \gamma \mathcal{L}_{XMLM} \quad (3)$$

where \mathcal{L}_{task} denotes the corresponding downstream token-level task and λ, β, γ correspond to weights of the respective losses for balanced loss integration. For \mathcal{L}_{task} is computed based on the training objective in Equation 1 as we leverage the generic CRF layer on top of the sequence output from PLM to generate the probability distribution of each token over the token-level class labels.

4.1 Orthographic-Phonemic Alignment

Traditional PLMs such as BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) encode the pre-tokenized input texts into 3 separate trainable embeddings: (1) token embedding (\vec{w}_t), (2) positional embedding (\vec{w}_p), (3) segment embedding (\vec{w}_s) where $\vec{w}_t, \vec{w}_p, \vec{w}_s \in \mathbb{R}^D$ and D denotes the hidden dimensions of corresponding PLM. In our work, we name the token embedding as orthographic embedding (OE) to distinguish from (1) phonemic embedding, (2) unified embedding from both phonemic and orthographic inputs. The overall representation of individual tokens is computed as a summation of three types of embedding $\vec{w} = \vec{w}_t + \vec{w}_p + \vec{w}_s$.

With the goal of enhancing textual representations via both orthographic and phonemic transcriptions, we introduce the Phonemic Embedding (PE) to directly capture phonemic transcriptions. Phonemic embedding, namely $\vec{w}_{PE} \in \mathbb{R}^D$, encodes the representations of phonemic transcription inputs. Phonemic Embedding is integrated with orthographic embedding, positional and segment embedding to form the token representations ³.

Motivated by previous works (Conneau and Lample, 2019; Chaudhary et al., 2020), we introduce additional Language Embedding ($\vec{w}_l \in \mathbb{R}^D$) to encode the input language types. These signals are beneficial to training objectives in cross-lingual settings with code-switched inputs introduced in Section 4.3.

The final word representation for the PLM Encoder is $\vec{w} = \vec{w}_t + \vec{w}_p + \vec{w}_s + \vec{w}_{PE} + \vec{w}_l$. We

³To ensure the alignment between length-variable phonemic and orthographic input resulted from tokenization, we meanpool the embedding of sub-tokens of individual inputs to construct representation for token-level tasks.

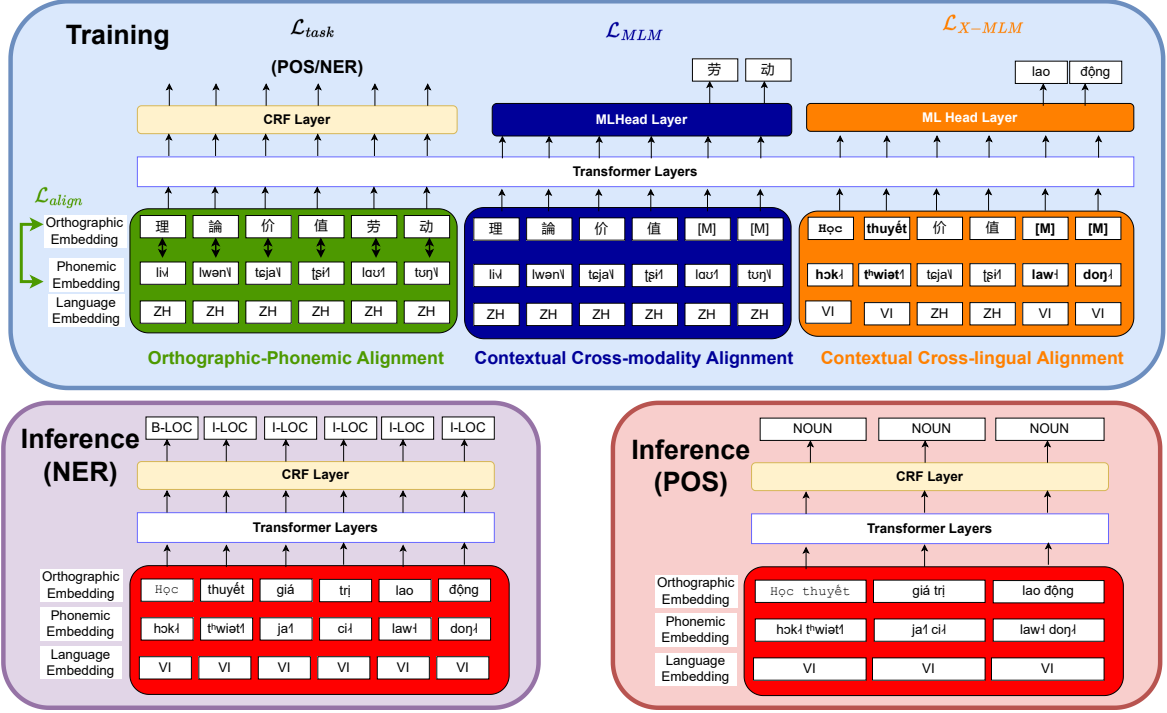


Figure 1: Illustration of the Proposed **PhoneXL** Model Overview. The model is trained with labeled source language data (ZH) and tested on target language data (VI). Orthographic embedding is identical to Token Embedding in PLMs and receives orthographic inputs. Phonemic Embedding maps phonemic inputs in IPA format to the embedding space. **Green, Blue, Orange** denote \mathcal{L}_{align} , \mathcal{L}_{MLM} , \mathcal{L}_{X-MLM} training objectives respectively. **Bold** represents target language words that are existent in bilingual dictionary to generate code-switching inputs. Parameters of Transformer Layers are shared across different inputs and unsupervised learning objectives.

denote $\vec{v} = Q(\vec{w})$ as the word representation produced by PLM where $Q(\cdot)$ denotes the PLM Encoder function.

To encourage the alignment between the orthographic textual input and its corresponding phonemic representation, we leverage cross-modality alignment and propose the computation of the phonemic-orthographic alignment loss:

$$\mathcal{L}_{OtoP} = CrossEntropy(sim_{OtoP}, labels) \quad (4)$$

The similarity matrices between phonemic and orthographic inputs (sim_{OtoP}) is computed as:

$$sim_{OtoP} = \sum_m^M \frac{\vec{w}_{m,PE}}{\|\vec{w}_{m,PE}\|} * \frac{\vec{w}_{m,t}}{\|\vec{w}_{m,t}\|} * \tau \quad (5)$$

where τ denotes the learnable soft temperature parameter and $\|\cdot\|$ is L2-normalization. $\vec{w}_{m,PE}$, $\vec{w}_{m,t}$ denote OE and PE of the m -th token in a sentence of length M .

Similarly to text-image alignment in cross-modality learning (Radford et al., 2021), the alignment is computed as a bi-directional interaction between orthographic and phonemic transcriptions. Therefore, the overall alignment loss is as follows:

$$\mathcal{L}_{align} = (\mathcal{L}_{OtoP} + \mathcal{L}_{PtoO})/2 \quad (6)$$

4.2 Contextual Cross-modality Alignment

The introduced alignment from 4.1 is restricted to 1-1 alignment between IPAs and tokens. However, contexts can significantly affect alignment between IPA and tokens. For instance, the orthography of 行 usually corresponds to [c̄iŋ]. However, the orthography of 行业 corresponds to [xáŋ jè]

To overcome the challenges, we propose introducing additional Masked Language Modeling to further align the two modalities. In other words, we randomly mask $\mu\%$ of input orthographic tokens and train the models to predict the masked tokens via (1) contextual/ non-masked orthographic tokens, (2) all of the phonemic transcriptions (including those of masked tokens). This objective encourages the alignment between phonemic and orthographic inputs via contextual information from both modalities of languages. Specifically, given a masked orthographic input and its corresponding phonemic representation, the model aims at predicting the masked tokens correctly. The loss is summarized as follows:

$$\mathcal{L}_{MLM} = - \sum_{j \in C} \log(P(y_j | \vec{v}_j; \theta)) \quad (7)$$

Table 2: Details of processed PANX and UDPOS datasets. We report statistics of source language training set and target language testing set for ZH-VI language pair.

	PANX		UDPOS	
	Source	Target	Source	Target
# Labels	7	7	18	18
# Samples	20000	10000	13320	1710
Avg Token Length	25.88	21.47	20.88	10.97
Avg Tokenized Orthographic Length	25.88	21.47	32.15	25.92
Avg Tokenized Phonemic Length	47.61	45.03	59.71	67.94

where y_j, \vec{v}_j denote the j -th location ground truth MLM label and input representation produced by PLM as introduced in 4.1 respectively. C denotes the number of masked tokens in the input training sentence.

4.3 Cross-lingual Contextual Alignment

Up to this point, the model does not leverage any specific cross-lingual signals for knowledge transfer between source and target languages. Therefore, we further introduce the Cross-lingual Contextual Alignment objective via bilingual dictionary. Similarly to Contextual Multi-modal Alignment as introduced in Section 4.2, we leverage MLM objectives to encourage the recognition of source language orthographic tokens given the phonemic inputs and multilingual orthographic contexts. The major difference between XMLM and MLM is that the input to XMLM is the code-switched input utterances which contain a mix of tokens from both source and target languages. Specifically, following (Qin et al., 2021), we conduct random code-switching of tokens of the source input utterances with ratio of $r\%$ where r is a hyperparameter. The code-switched inputs follow the same procedure of MLM as introduced in Section 4.2. The XMLM training objective is summarized as follows:

$$\mathcal{L}_{XMLM} = - \sum_{j \in C'} \log(P(y_j | \vec{v}_j; \theta)) \quad (8)$$

where \vec{v}_j is the PLM representation of j -th token for the code-switched input sentences of source language and C' is the number of masked tokens based on percentage of code-switched source language inputs. Depending on the selected tokens for code-switching and its corresponding target language tokens, the absolute values of C and C' are not necessarily the same since the number of tokens in source samples and code-switched samples might not be identical.

5 Experiments

5.1 Datasets & Preprocessing

We evaluate our proposed framework on token-level tasks, including Named Entity Recognition (NER) and Part-of-Speech Tagging (POS) among four languages: Chinese (ZH), Vietnamese (VI), Japanese (JA) and Korean (KO). Based on linguistic structural similarities (SVO vs SOV structural orders) and lexical similarities in terms of phonemic representations, we divide the four languages into 2 groups: (JA,KO) and (ZH,VI) where the former in each pair is considered a high-resourced language and the latter is a low-resourced counterpart. During training, only high-resourced languages are leveraged and we conduct zero-shot evaluation on the target low-resourced languages.

To evaluate our proposed framework on token-level tasks, we first construct a new dataset by preprocessing the alignment between orthographic and phonemic transcriptions. Specifically, we leverage NER and POS datasets (namely PANX and UDPOS) from XTREME benchmark datasets (Hu et al., 2020). Given input utterances from the datasets, we generate the corresponding phonemic transcriptions on token-level. As phonemic transcriptions can either **Romanized Transcriptions** (i.e. Pinyin for ZH, Romaji for JA, Romaja for KO) or **IPA Transcriptions**, we generate both types of phonemic transcriptions and conduct empirical study on both in Section 6.

Generating Romanized Transcriptions As VI is written in Latin, we preserve the original characters as the corresponding Romanized transcriptions. For ZH, JA, KO, we directly obtain the corresponding Romanized transcriptions via [dragonmapper](#), [pykakasi](#) and [korean_romanizer](#) respectively.

Generating IPA Transcriptions As PanPhon (Mortensen et al., 2016) does not support some of our targeted languages (JA, KO), we leverage external open-source tools to generate IPA transcriptions for individual languages. Specifically, we use [dragonmapper](#), [viphoneme](#) to generate IPA transcriptions for ZH, VI respectively. As there are no direct open-source IPA transcription tools available for JA,KO, we generate IPA transcriptions via a two-stage process. First, we generate Romanized transcriptions for JA, KO via [pykakasi](#) and [korean_romanizer](#) respectively. Then, as the aforementioned Romanized transcriptions are in Latin-based format, we treat them as Latin char-

Table 3: NER and POS Experimental Results on PANX and UDPOS **test** datasets respectively.

Model	PANX				UDPOS			
	ZH->VI		JA->KO		ZH->VI		JA->KO	
	Source (ZH)	Target (VI)	Source (JA)	Target (KO)	Source (ZH)	Target (VI)	Source (JA)	Target (KO)
mBERT	78.10 ± 0.25	49.94 ± 1.44	69.85 ± 0.17	26.64 ± 0.17	89.93 ± 0.02	48.62 ± 0.66	86.24 ± 0.13	43.63 ± 1.28
PhoneXL (full)	80.42 ± 0.07	52.28 ± 0.98	72.90 ± 0.37	29.25 ± 0.59	90.53 ± 0.04	50.71 ± 0.40	90.00 ± 0.15	46.75 ± 0.09
PhoneXL (w \mathcal{L}_{align})	79.71 ± 0.21	51.09 ± 0.42	72.01 ± 0.11	28.23 ± 0.32	90.42 ± 0.03	50.29 ± 0.13	89.56 ± 0.07	45.96 ± 0.47
PhoneXL (w \mathcal{L}_{MLM})	79.70 ± 0.17	50.23 ± 1.63	72.62 ± 0.02	27.90 ± 0.11	90.44 ± 0.03	50.49 ± 0.67	89.53 ± 0.16	45.94 ± 0.45
PhoneXL (w \mathcal{L}_{XMLM})	79.69 ± 0.15	50.83 ± 0.63	72.57 ± 0.57	28.85 ± 0.71	90.40 ± 0.05	50.20 ± 1.63	89.63 ± 0.16	45.25 ± 0.73
XLM-R	75.31 ± 0.46	35.68 ± 0.66	66.31 ± 0.06	14.80 ± 0.97	91.28 ± 0.04	50.40 ± 0.51	89.94 ± 0.17	46.16 ± 0.24
PhoneXL (full)	77.00 ± 0.24	38.88 ± 0.15	69.02 ± 0.24	16.39 ± 0.13	91.43 ± 0.24	52.73 ± 0.86	90.06 ± 0.04	48.82 ± 0.43
PhoneXL (w \mathcal{L}_{align})	76.41 ± 0.09	37.04 ± 0.68	68.76 ± 0.25	15.34 ± 0.13	91.39 ± 0.02	52.46 ± 0.17	90.01 ± 0.12	47.96 ± 0.62
PhoneXL (w \mathcal{L}_{MLM})	76.70 ± 0.07	37.29 ± 0.34	67.62 ± 0.13	15.16 ± 0.58	91.14 ± 0.02	51.88 ± 1.53	90.02 ± 0.05	47.83 ± 0.39
PhoneXL (w \mathcal{L}_{XMLM})	76.52 ± 0.15	37.15 ± 0.30	68.68 ± 1.39	15.89 ± 0.79	91.04 ± 0.05	51.15 ± 1.40	89.90 ± 0.37	47.85 ± 0.56

Table 4: NER and POS Baseline Results on PANX and UDPOS **test** datasets respectively. **Dict** denotes the assumptions of available bilingual dictionary and **MT** refers to the assumptions of available Machine Translations between source and target languages. Cross-lingual Transfer methods leverage either Dict or MT or both.

Model	Assumption		PANX				UDPOS			
	Dict	MT	ZH->VI		JA->KO		ZH->VI		JA->KO	
			Source (ZH)	Target (VI)	Source (JA)	Target (KO)	Source (ZH)	Target (VI)	Source (JA)	Target (KO)
mBERT			78.10 ± 0.25	49.94 ± 1.44	69.85 ± 0.17	26.64 ± 0.17	89.93 ± 0.02	48.62 ± 0.66	86.24 ± 0.13	43.63 ± 1.28
CoSDA-ML	✓		78.48 ± 0.34	47.82 ± 1.43	70.42 ± 0.50	25.76 ± 1.75	89.76 ± 0.19	49.84 ± 0.49	87.63 ± 0.14	41.19 ± 1.16
X-MIXUP		✓	78.87 ± 0.17	52.98 ± 0.05	68.10 ± 0.69	26.41 ± 1.06	89.41 ± 0.10	50.05 ± 0.95	87.58 ± 0.17	48.47 ± 0.37
PhoneXL (full)	✓		80.42 ± 0.07	52.28 ± 0.98	72.90 ± 0.37	29.25 ± 0.59	90.53 ± 0.04	50.71 ± 0.40	90.00 ± 0.15	46.75 ± 0.09
XLM-R			75.31 ± 0.46	35.68 ± 0.66	66.31 ± 0.06	14.80 ± 0.97	91.28 ± 0.04	50.40 ± 0.51	89.94 ± 0.17	46.16 ± 0.24
FILTER		✓	72.55 ± 0.11	40.17 ± 1.35	62.92 ± 0.26	18.60 ± 1.02	90.57 ± 0.05	55.85 ± 0.27	90.81 ± 0.19	43.25 ± 1.52
xTune	✓	✓	77.48 ± 0.08	40.94 ± 0.87	68.02 ± 0.26	21.95 ± 1.02	91.75 ± 0.10	51.91 ± 0.74	89.75 ± 0.31	51.03 ± 1.26
X-MIXUP		✓	75.89 ± 0.46	38.22 ± 0.72	65.33 ± 0.69	16.43 ± 2.98	90.67 ± 0.06	50.30 ± 1.23	88.48 ± 0.23	50.63 ± 0.95
PhoneXL (full)	✓		77.00 ± 0.24	38.88 ± 0.15	69.02 ± 0.24	16.39 ± 0.13	91.43 ± 0.24	52.73 ± 0.86	90.06 ± 0.04	48.82 ± 0.43

acters and input them to PanPhon to generate IPA transcriptions for JA and KO. Details of our constructed datasets are provided in Table 2.

5.2 Implementation Details

For evaluation of both NER and POS tasks, we report the F-1 score for different individual language pairs. We report performance on both development and test set of both source and target languages.

As IPA involves unique characters that are outside the typical orthographic vocabulary (i.e. /ŋ/, /ç/, /ʃ/, /ʒ/), we extend PLM vocabulary to account for these special characters. Therefore, both OEs and PEs are resized to account for the newly extended vocabulary. The impact of extended vocabulary will be further discussed in Section 6.

For each language pair (ZH-VI vs JA-KO) and token-level tasks (NER vs POS), we tune hyperparameters of our framework based on the development set of the source language. Specifically, we conduct grid search for λ , β , γ over the space [0.1, 0.01, 0.001, 0.0001]. Mask ratio (μ) and cs_ratio (r) are tuned over the space [0.1, 0.4] inclusive with a step of 0.05. Hyperparameter details for each task and language pair are reported in Table 5.

We train our model with the batch size of 32 and 16 for mBERT and XLM-R respectively. Both

Table 5: Hyperparameters for PANX and UDPOS datasets (NER and POS tasks respectively) on experimental language pairs ZH->VI and JA->KO

	PANX		UDPOS	
	ZH->VI	JA->KO	ZH->VI	JA->KO
λ	0.01	0.1	0.01	0.1
β	0.01	0.001	0.001	0.01
γ	0.01	0.001	0.01	0.01
μ	0.20	0.25	0.10	0.05
r	0.40	0.30	0.40	0.30

multilingual base versions (L=12, H=12, D=768 where L,H,D denote the number of hidden layers, the number of attention heads per layer and hidden dimension respectively) are used as backbone PLM architectures for our experiments. Both training and inference of our framework are conducted on NVIDIA TITAN RTX and NVIDIA RTX 3090 GPUs. We report our experimental results as the average performance of 3 runs from different random seed initializations with standard deviations. Due to space constraints, we report our empirical studies on test sets in Table 3 and 4. Additional results on development (dev) sets for both datasets are summarized in the Appendix A.

In our study, we leverage publicly available MUSE bilingual dictionary (Conneau et al., 2017). As bilingual dictionary is only available between

EN and target languages, we construct bilingual dictionaries for our language pairs (ZH-VI and JA-KO) by leveraging EN as a bridge for semantic alignment between source and target languages.

5.3 Baseline

We compare our method with previously proposed cross-lingual transfer frameworks for token-level tasks. We conduct experiments with both mBERT and XLM-R backbone PLM architecture. We compare our work with both approaches leveraging Machine Translation (MT) and/or Bilingual Dictionary (Dict), including:

- CoSDA-ML (Qin et al., 2021): Multi-level Code-switching augmentation for various cross-lingual NLP tasks, including POS and NER tasks.
- FILTER (Fang et al., 2021): Cross-lingual Transfer via Intermediate Architecture Disentanglement with Knowledge Distillation objectives from source to target languages.
- xTune (Zheng et al., 2021): Two-stage augmentation mechanisms with four exhaustive augmentation methods for cross-lingual transfer.
- XMIXUP (Yang et al., 2022): Cross-lingual transfer via Manifold Mixup Augmentation and Machine Translation Alignment between source and target languages.

As *FILTER*, *xTune* and *XMIXUP* require training parallel corpora, we generate translation of training data from source languages (ZH, JA) to target languages (VI, KO) via third-party MT package⁴. Ground-truth label sequence of MT data is the same as the original source language data.

As *CoSDA-ML* and *xTune* both require bilingual dictionary for cross-lingual transfer, we leverage the reconstructed MUSE bilingual dictionary for our setting as introduced in 5.2.

6 Result & Discussion

Our experimental results for NER and POS tasks are summarized in Table 3 and 4. Based on the empirical study, our proposed *PhoneXL* framework consistently outperforms by the backbone PLM architecture of mBERT and XLM-R in both evaluated token-level tasks for low-resourced target

languages VI and KO. For instance, with ZH-VI pair, we observe the target language’s F1 evaluation metric improvement of 2.34 points and 2.01 points for NER and POS tasks respectively as compared to the fine-tuned backbone mBERT architecture. This improvement implies the phonemic information provides essential information beyond orthographic representation to further bridge the gap between the source and target languages.

In NER task, the larger and state-of-the-art multilingual PLM XLM-R yields worse performance than mBERT in both of our language pair performance on source and target languages. On the other hand, for POS task, XLM-R based architecture only results in marginal performance gains when compared with mBERT. Interestingly, our mBERT-based framework achieves competitive performance with XLM-R backbone architecture on POS task despite leveraging smaller vocabulary size and less pre-training data. We hypothesize this might be due to the fact that XLM-R has been trained with more languages, leading to biases towards certain types of languages during pre-training that might not share common properties with CJKV languages.

Despite the performance gain over XLM-R based architecture observed in Table 3, our *PhoneXL* framework does not consistently outperform the previous baselines such as *FILTER*, *xTune* and *XMIXUP* in Table 4. However, these baselines require extra parallel corpora obtained from machine translation which might not always be readily available for all languages, especially for low-resourced ones. On the other hand, our proposed method achieves state-of-the-art performance among methods leveraging only bilingual dictionary. In addition, despite its impressive performance, *xTune* requires two-stage training procedures, four different exhaustive augmentation methods as well as the knowledge of both machine translation and bilingual dictionary. Therefore, it is deemed more time-consuming and resource-intensive than our approach.

Romanized vs IPA Phonemic Transcriptions

As observed in Table 6, leveraging Romanized transcriptions from individual languages instead of IPA degrades the downstream POS task performance on both low-resourced target languages (averaged 0.93 and 1.57 points of performance drop on VI and KO respectively from *PhoneXL-IPA (full)* to *PhoneXL-Romanized (full)*). We hypothesize it might be due

⁴<https://pyapi.org/project/googletrans/>

Table 6: Ablation study of the impact of vocabulary extension and IPA embedding on target language F1-score in POS task (VI,KO respectively) with mBERT backbone architecture.

	ZH->VI	JA->KO
mBERT (w/o PE, w/o extension)	48.62 ± 0.66	43.63 ± 1.28
mBERT (w PE, w/o extension)	48.95 ± 0.52	43.85 ± 0.14
mBERT (w PE, w extension)	49.14 ± 0.98	44.42 ± 0.36
PhoneXL-IPA (w/o Lang Embedding)	49.74 ± 0.28	45.84 ± 0.07
PhoneXL-Romanized (full)	49.78 ± 1.99	45.18 ± 2.03
PhoneXL-IPA (full)	50.71 ± 0.40	46.75 ± 0.09

to the lack of phonemic consistency among Romanized transcriptions of different languages. In addition, as certain low-resourced languages might not have their own Romanized transcriptions, IPA phonemic transcriptions provide a more generalized and consistent pathway to generate phonemic transcriptions across different language families.

Impact of Phonemic Embeddings Based on Table 6, we also observe that the introduction of IPA embedding, even without any unsupervised objectives, also provide additional improvements as compared to the backbone orthographic-based mBERT architecture. However, further training with our introduced objectives provide stronger boost in target language performance improvements on the downstream task.

Impact of Vocabulary Extension As observed in Table 6, vocabulary extension is important to learn effective Phonemic Embedding. Vocabulary extension allows the model to differentiate unique IPA characters when encoding tokens, leading to 0.19 and 1.42 points of F1 score improvements on mBERT for VI and KO respectively. It is intuitive since additional special phonemic characters possess different meanings than typical orthographic characters. However, we still observe a significant gap between *mBERT with PE* and *PhoneXL* framework. It is due to the lack of alignment between embedding of phonemic and orthographic inputs.

Impact of Unsupervised Objectives As observed in Table 3, each introduced alignment objective between phonemic and orthographic inputs provides additional performance gain over the original backbone mBERT and XLM-R PLMs on both language groups. Additionally, from Table 6, the existence of performance gap between simple introduction of PE (*mBERT (w PE, w extensions)*) and *PhoneXL* (i.e. 1.57 points on VI and 2.33 points on KO in POS task) implies that the unsupervised alignment objectives are crucial to bringing about

the benefits of PE.

Impact of Language Embedding Language Embedding is crucial to our framework performance, leading to consistent performance gain on both target languages in POS task. In fact, Language Embedding is especially important to \mathcal{L}_{XMLM} as inputs are code-switched sentences which are made up of tokens from different languages. Without language indication from Language Embedding, the model is unable to predict the correct masked tokens in the correct language.

7 Conclusion & Future Work

In our work, we propose **PhoneXL**, a novel mechanism to integrate phonemic transcription with orthographic transcription to further enhance representation capability of language models in cross-lingual transfer settings. By encouraging the alignment between the two linguistic modalities via direct one-to-one alignment, indirect contextual alignment and additional code-switching via bilingual dictionaries, our proposed **PhoneXL** yields consistent performance improvements over the backbone orthographic-based PLM architecture in downstream cross-lingual token-level task among the CJKV languages. We also release the first aligned phonemic-orthographic datasets for CJKV languages for two popular token-level tasks (NER and POS). In future work, we plan to train our proposed unsupervised objectives with larger CJKV corpora as pre-training mechanisms to evaluate effectiveness of the representations in multi-granularity downstream tasks (i.e. sentence-level classification tasks to Question-Answering tasks). Further extensions towards few-shot learning settings (Nguyen et al., 2020; Xia et al., 2020) where a small number of target language examples can be leveraged to exploit orthographic-phonemic similarity between source and target languages is a promising direction for our future work.

Limitations

Our approach is heavily dependent on the quality of the pre-processed orthographic-phonemic transcription data as it provides the ground-truth for unsupervised alignment objectives. Generating phonemic transcriptions and aligning them correctly with orthographic representations can be costly. Despite our significant efforts, the alignment is still far from perfect optimality.

Secondly, our approach might not be effective in improving performance on randomly chosen language pairs. As our framework aims to exploit phonemic similarities of languages with different orthographic representations, the methods are only effective in cross-lingual transfer between lexically similar languages in terms of phonology such as CJKV languages. Languages that do not fall into this category might observe little to no performance gains with our proposed framework.

References

- Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. [Phonologically aware neural model for named entity recognition in low resource transfer settings](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.
- Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries.
- Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R Mortensen, and Jaime G Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. *arXiv preprint arXiv:1808.09500*.
- 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 and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. Filter: An enhanced fusion method for cross-lingual language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12776–12784.
- Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. [Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.
- David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL.
- Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *NAACL-HLT 2021-2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Hoang Nguyen and Gene Rohrbaugh. 2019. Cross-lingual genre classification using linguistic groupings. *Journal of Computing Sciences in Colleges*, 34(3):91–96.
- Hoang Nguyen, Chenwei Zhang, Congying Xia, and S Yu Philip. 2020. Dynamic semantic matching and aggregation network for few-shot intent detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1209–1218.

- 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.
- Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3853–3860.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. **Chinese-BERT: Chinese pretraining enhanced by glyph and Pinyin information**. 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 2065–2075, Online. Association for Computational Linguistics.
- Congying Xia, Chenwei Zhang, Hoang Nguyen, Jiawei Zhang, and Philip Yu. 2020. Cg-bert: Conditional text generation with bert for generalized few-shot intent detection. *arXiv preprint arXiv:2004.01881*.
- Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. *arXiv preprint arXiv:2205.04182*.
- Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.
- Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. 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 3403–3417.

A Additional Experiments

We provide additional experiments on dev set of PANX and UDPOS datasets of XTREME benchmark datasets in Table 7 and 8. Our observations are mostly consistent between dev and test sets on both evaluated token-level tasks.

Table 7: NER and POS Experimental Results on PANX and UDPOS **dev** datasets respectively.

Model	PANX				UDPOS			
	ZH->VI		JA->KO		ZH->VI		JA->KO	
	Source (ZH)	Target (VI)	Source (JA)	Target (KO)	Source (ZH)	Target (VI)	Source (JA)	Target (KO)
mBERT	78.54 ± 0.09	49.30 ± 0.25	69.05 ± 0.12	27.27 ± 0.19	92.72 ± 0.09	46.75 ± 0.66	92.42 ± 0.06	42.15 ± 0.74
PhoneXL (full)	79.89 ± 0.16	51.81 ± 0.62	72.20 ± 0.06	30.00 ± 0.77	93.75 ± 0.03	49.04 ± 0.29	96.41 ± 0.11	44.24 ± 0.19
PhoneXL (w \mathcal{L}_{align})	79.32 ± 0.24	50.78 ± 0.45	71.46 ± 0.09	28.80 ± 0.40	93.52 ± 0.05	48.40 ± 0.28	96.19 ± 0.06	43.65 ± 0.71
PhoneXL (w \mathcal{L}_{MLM})	79.40 ± 0.35	49.84 ± 1.72	72.12 ± 0.37	27.97 ± 0.22	93.41 ± 0.13	48.70 ± 0.73	96.29 ± 0.03	42.95 ± 0.58
PhoneXL (w \mathcal{L}_{XMLM})	79.34 ± 0.04	50.55 ± 0.73	72.07 ± 0.49	29.28 ± 0.80	93.43 ± 0.06	48.59 ± 0.46	96.12 ± 0.08	43.18 ± 0.70
XLM-R	75.33 ± 0.71	35.77 ± 0.75	66.14 ± 0.016	14.56 ± 0.80	94.92 ± 0.08	48.96 ± 0.46	96.44 ± 0.08	44.15 ± 0.17
PhoneXL (full)	76.54 ± 0.16	38.90 ± 0.31	68.85 ± 0.16	17.26 ± 0.23	94.84 ± 0.41	51.45 ± 1.55	97.18 ± 0.02	45.97 ± 0.48
PhoneXL (w \mathcal{L}_{align})	76.27 ± 0.24	36.42 ± 0.23	68.17 ± 0.14	16.08 ± 0.01	95.00 ± 0.05	51.47 ± 0.30	96.67 ± 0.08	45.00 ± 0.44
PhoneXL (w \mathcal{L}_{MLM})	76.16 ± 0.10	37.20 ± 0.11	68.12 ± 0.15	15.90 ± 0.35	94.26 ± 0.17	50.84 ± 1.56	96.70 ± 0.04	45.27 ± 0.41
PhoneXL (w \mathcal{L}_{XMLM})	76.09 ± 0.06	36.81 ± 0.28	68.26 ± 1.27	16.11 ± 0.23	94.22 ± 0.14	50.48 ± 1.06	96.67 ± 0.05	44.88 ± 0.62

Table 8: NER and POS Baseline Results on PANX and UDPOS **dev** datasets respectively. **Dict** denotes the assumptions of available bilingual dictionary and **MT** refers to the assumptions of available Machine Translations between source and target languages. Cross-lingual Transfer methods leverage either Dict or MT or both.

Model	Assumption		PANX				UDPOS			
	Dict	MT	ZH->VI		JA->KO		ZH->VI		JA->KO	
			Source (ZH)	Target (VI)	Source (JA)	Target (KO)	Source (ZH)	Target (VI)	Source (JA)	Target (KO)
mBERT			78.54 ± 0.09	49.30 ± 0.25	69.05 ± 0.12	27.27 ± 0.19	92.72 ± 0.09	46.75 ± 0.66	92.42 ± 0.06	42.15 ± 0.74
CoSDA-ML	✓		78.06 ± 0.19	47.22 ± 1.39	70.46 ± 0.43	26.10 ± 1.57	92.97 ± 0.17	48.46 ± 0.59	94.66 ± 0.18	40.85 ± 0.91
X-MIXUP		✓	78.68 ± 0.21	53.62 ± 0.45	67.70 ± 0.58	26.70 ± 0.87	94.26 ± 0.23	48.61 ± 0.62	96.02 ± 0.07	48.70 ± 0.51
PhoneXL (full)	✓		79.89 ± 0.16	51.81 ± 0.62	72.20 ± 0.06	30.00 ± 0.77	93.75 ± 0.03	49.04 ± 0.29	96.41 ± 0.11	44.24 ± 0.19
XLM-R			75.33 ± 0.71	35.77 ± 0.75	66.14 ± 0.016	14.56 ± 0.80	94.92 ± 0.08	48.96 ± 0.46	96.44 ± 0.08	44.15 ± 0.17
FILTER		✓	72.13 ± 0.16	39.76 ± 1.31	62.96 ± 0.38	19.46 ± 0.95	93.14 ± 0.19	53.89 ± 0.28	96.99 ± 0.07	39.39 ± 1.56
xTune	✓	✓	77.38 ± 0.10	40.58 ± 0.67	68.26 ± 0.38	23.05 ± 0.95	95.34 ± 0.17	50.29 ± 0.69	97.08 ± 0.08	49.54 ± 0.79
X-MIXUP		✓	73.20 ± 0.22	38.17 ± 0.78	64.46 ± 0.47	17.00 ± 2.81	94.80 ± 0.14	50.25 ± 1.05	96.42 ± 0.05	48.05 ± 0.96
PhoneXL (full)	✓		76.54 ± 0.16	38.90 ± 0.31	68.85 ± 0.16	17.26 ± 0.23	94.84 ± 0.41	51.45 ± 1.55	97.18 ± 0.02	45.97 ± 0.48