SpikeXL: Language-Agnostic Spike Representations for Few-Shot Cross-Lingual Speech Command Recognition

Anonymous ACL submission

Abstract

Deploying speech command recognition on multilingual edge devices demands models that are both energy-efficient and capable of cross-lingual transfer without per-language retraining—two goals rarely addressed jointly. Spiking Neural Networks (SNNs) offer substantial energy savings via sparse binary communication, yet their cross-lingual generalization remains unstudied. We present **SpikeXL**, a framework that investigates whether SNN spike sparsity provides an inductive bias toward language-agnostic representations. Our experiments suggest that the LIF firing threshold suppresses language-specific surface phonetics while preserving command-level acoustic structure, yielding embeddings that transfer across languages more readily than dense ANN activations. Building on this finding, we introduce **Spike-space Contrastive Alignment** (SCA), a contrastive objective adapted for spike-rate embeddings that aligns semantically equivalent commands across languages without requiring bilingual lexicons or parallel speech data. On Speech-MASSIVE across 12 typologically diverse languages, SpikeXL achieves a Cross-Lingual Semantic Similarity (CLSS) of 0.912—outperforming a parameter-matched Transformer with the same SCA objective (0.763) and SNN without SCA (0.721)—indicating that spike sparsity itself contributes an inductive bias beyond what contrastive training alone provides. Few-shot k-NN classification on target languages (115 support samples, frozen encoder) reaches 80.8% macro-average accuracy, corroborating alignment quality with a metric fully independent of the SCA objective. On English GSC-v2, SpikeXL reaches 92.76% accuracy with 2.5M parameters and 45M synaptic operations per inference.

1 Introduction

Consider a spiking neuron: it fires only when its membrane potential crosses a threshold, discarding sub-threshold fluctuations (Maass, 1997). In the speech domain, many features that distinguish languages at the surface—allophonic variation, language-specific prosody, phonotactic constraints—manifest as precisely the kind of low-amplitude, high-frequency patterns that spiking thresholds suppress. What remains are coarser temporal spike bursts tied to the acoustic onset, nucleus, and offset of salient phonemes—events that, we hypothesize, carry cross-lingually transferable command structure. *Could an architectural property already motivated on energy grounds also provide a structural bias toward language-agnostic representations?*

The energy–multilingualism gap. Voice-controlled devices are proliferating in multilingual environments, yet the two most pressing deployment constraints are rarely addressed together. On the *efficiency* side, wearables and IoT sensors operate under strict energy budgets that are orders of magnitude below what current Transformer-based systems require. On the *multilingual* side, large self-supervised models such as Whisper (Radford et al., 2023) and wav2vec 2.0 (Baevski et al., 2020) achieve strong cross-lingual transfer, but rely on dense floating-point activations that are prohibitively expensive for neuromorphic hardware. Spiking Neural Networks (SNNs) address the energy side by communicating via sparse binary spikes on chips such as Intel Loihi (Davies et al., 2018) and SpiNNaker (Furber et al., 2014), yet existing SNN speech models (Baek and Lee, 2024; Fang et al., 2021; Song et al., 2024) are exclusively monolingual. This paper investigates whether these two goals—efficiency and multilingual transfer—can be pursued jointly.

Our approach. We operationalize the sparsity-as-filter hypothesis in **SpikeXL**. The model uses a Spiking Transformer encoder with multi-view temporal self-attention operating at both phoneme and command timescales. On top of this encoder,

we propose **Spike-space Contrastive Alignment (SCA)**, which pulls together the pooled spike embeddings of the same command across different languages and pushes apart embeddings of different commands—without requiring paired multilingual speech data, bilingual dictionaries, or text transcriptions of the target language.

Contributions.

1. **Empirical analysis of spike sparsity and language transfer.** We provide a systematic investigation showing that SNN spike patterns exhibit higher cross-lingual representation similarity than equivalent ANN activations at matched model capacity, and offer a mechanistic account grounded in LIF threshold dynamics. (Section 5)

2. **Spike-space Contrastive Alignment (SCA).** We design a supervised contrastive objective adapted for spike-rate embeddings, addressing the challenge that standard similarity measures are not directly applicable to binary spike trains. (Section 4.3)

3. **Multilingual representation benchmarking.** We evaluate SpikeXL on Speech-MASSIVE (Lee et al., 2024) across 12 typologically diverse languages using only 115 labeled samples per language for SCA adaptation, finding CLSS > 0.89 across all language families and 80.8% macro-average k-NN accuracy on a held-out downstream metric independent of SCA. (Section 7)

4. **Synaptic-operation energy analysis.** SpikeXL achieves 45M SOPs per inference versus 480M FLOPs for a parameter-matched ANN (4.4× reduction) and 3.8–12.4 G FLOPs for large multilingual Transformers. (Section 7.4)

2 Related Work

Cross-lingual speech understanding. Cross-lingual transfer for speech has been studied extensively in the context of self-supervised pretraining. XLS-R (Babu et al., 2021) scales cross-lingual speech representation learning to 436,000 hours of audio across 128 languages, achieving strong transfer across ASR benchmarks. mSLAM (Bapna et al., 2022) jointly pretrains on speech and text, further closing the gap between monolingual and multilingual models. For the narrower task of speech command recognition, Mazumder et al. (2021) demonstrate few-shot cross-lingual keyword spotting using embedding-based transfer, and Yen et al. (2024) achieve zero-shot multilingual keyword recognition via universal phonetic attribute modeling with domain adversarial training. To our knowledge, ours is the first study of cross-lingual transfer within a fully spike-based model; we also provide a mechanistic account of *why* the observed transfer occurs.

Spiking Neural Networks for speech and language. SNNs have been applied to audio classification and speech tasks across a range of architectures (Baek and Lee, 2024). Fang et al. (2021) show that learnable membrane time constants substantially improve SNN training stability, a finding we incorporate in our design. For keyword spotting specifically, Song et al. (2024) propose early-decision SNNs for rapid energy-efficient detection. More recently, SNN-based self-attention architectures—notably Spikformer (Zhou et al., 2023) and Spike-driven Transformer (Yao et al., 2023)—have demonstrated competitive performance on vision benchmarks by replacing softmax attention with spike-compatible Hadamard-product attention. We adapt these spiking attention principles to the speech domain. SpikeVoice (Wang et al., 2024) demonstrates that SNNs can handle sequential speech generation tasks, broadening the scope of spike-based speech processing beyond recognition. None of these prior works address multilingual generalization.

Contrastive learning for cross-lingual alignment. Contrastive objectives have proven effective for aligning representations across languages in text (Feng et al., 2022; Gao et al., 2021) and for cross-lingual speech retrieval (Zhu et al., 2024). The GL-CLeF framework (Qin et al., 2022) introduces global-local contrastive learning for zero-shot cross-lingual spoken language understanding, establishing that explicit contrastive alignment outperforms parameter-sharing alone. Our SCA loss is inspired by the supervised contrastive principle (Khosla et al., 2021) but is specifically adapted for the pooled spike-rate space rather than dense embeddings, and does not require bilingual lexicons.

Efficient multilingual NLP. Parameter-efficient cross-lingual adaptation via adapter modules (Pfeiffer et al., 2020) offers one path to multilingual

efficiency. Knowledge distillation (Sanh et al., 2020) reduces model size but preserves the dense-activation compute paradigm. SpikeXL occupies a complementary position: we achieve energy reduction through fundamentally sparse spike-based computation rather than through parameter compression. A general treatment of unsupervised cross-lingual representations is provided by Ruder et al. (2019), whose survey contextualizes the broader landscape our work extends toward the speech-SNN intersection.

3 Background: Spiking Neurons and Spike Coding

Leaky Integrate-and-Fire neurons. We adopt the Leaky Integrate-and-Fire (LIF) neuron model (Maass, 1997) as the basic computational unit. At discrete time step t, the membrane potential u^t evolves as:

$$u^t = \tau \cdot u^{t-1} \cdot (1 - o^{t-1}) + W x^t \tag{1}$$

where $\tau \in (0, 1)$ is a learnable membrane time constant (Fang et al., 2021) controlling the decay rate, W is the synaptic weight matrix, x^t is the pre-synaptic input, and o^{t-1} is the spike output of the previous step. A spike is emitted when the membrane potential exceeds threshold V_{th}:

$$o^t = \begin{cases} 1 & \text{if } u^t \geq V_{th} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

After firing, the membrane resets to zero (hard reset). Since the Heaviside spike function is non-differentiable, we use the surrogate gradient method (Wu et al., 2018) during training, approximating $\partial o^t / \partial u^t$ with a piecewise-linear function.

Energy model. On neuromorphic hardware (Davies et al., 2018; Furber et al., 2014), energy consumption is dominated by synaptic operations (SOPs)—multiply-accumulate events triggered by each spike. Given average spike activity rate r (fraction of neurons firing per timestep), the number of SOPs scales as $r \cdot N_{\text{syn}}$, where N_{syn} is the number of synaptic connections. Since spike representations are sparse ($r \ll 1$), SNN inference energy is substantially lower than that of equivalent ANNs. We track SOPs throughout all experiments to provide hardware-agnostic energy estimates (Wang et al., 2024).

4 SpikeXL

Figure 1 provides an overview of SpikeXL. The model consists of three components: (1) a **Spiking Speech Encoder** (SSE) that converts mel-spectrogram frames into spike trains and processes them through multi-view temporal spike attention; (2) a **Spike Temporal Pooling** (STP) module that aggregates per-step spike representations into a fixed-size command embedding; and (3) the **Spike-space Contrastive Alignment** (SCA) loss applied during multilingual training.

4.1 Spiking Speech Encoder

Input encoding. Given an audio waveform, we extract an 80-dimensional log-mel spectrogram with a 25 ms window and 10 ms hop, following the preprocessing of Radford et al. (2023). The spectrogram is partitioned into non-overlapping patches of size $P \times 80$ frames. Each patch is projected to dimension D via a learnable linear layer. Poisson rate coding with T time steps converts the resulting dense embeddings into binary spike tensors $\mathbf{S} \in \{0, 1\}^{T \times L \times D}$, where L is the number of patches. Data augmentation follows SpecAugment (Park et al., 2019) during training.

Spike Multi-view Temporal Self-Attention (Spike-MTSA). The core of the SSE is a stack of N Spike-MTSA blocks. Drawing on the observation that speech commands exhibit dynamics at both short (phoneme) and long (word-command) timescales, each block processes the spike sequence through two complementary views:

- **Local view**: attention within a sliding window of size W patches, capturing phoneme-level temporal dynamics.
- **Global view**: attention over strided, downsampled positions with stride s, capturing command-level dynamics across the full utterance.

Within each view, query, key, and value tensors are obtained by applying LIF layers to the spike input:

$$\mathbf{Q} = \text{LIF}(W_Q \mathbf{S}), \quad \mathbf{K} = \text{LIF}(W_K \mathbf{S}), \quad \mathbf{V} = \text{LIF}(W_V \mathbf{S}) \tag{3}$$

Attention is computed in the spike domain using Hadamard product, avoiding the quadratic cost of

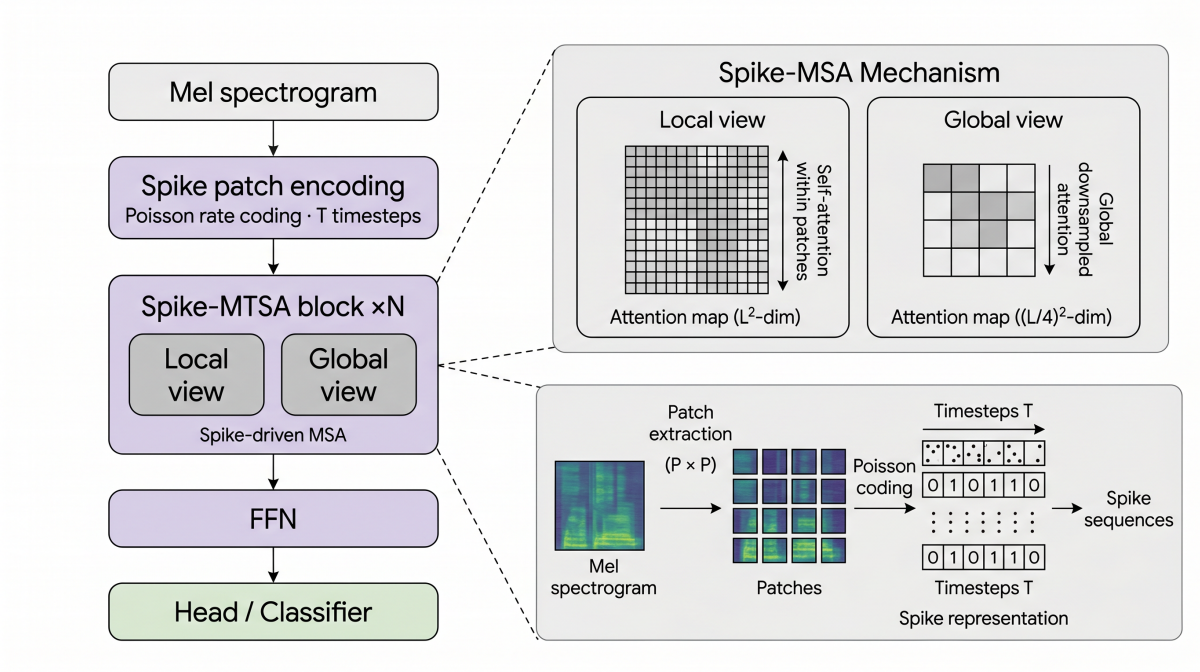

Figure 1: Overview of SpikeXL. A shared Spiking Speech Encoder processes mel-spectrogram patches through N Spike-MTSA blocks, each combining a local and a global temporal attention view. Spike Temporal Pooling produces a real-valued command embedding $\mathbf{z}$ from the spike firing-rate map. During multilingual training, the SCA loss aligns $\mathbf{z}$ of the same command across languages; at test time, $\mathbf{z}$ is passed to the classification head.

dot-product attention while preserving binary spike semantics (Zhou et al., 2023; Yao et al., 2023):

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{LIF}\left(\frac{\mathbf{Q} \odot \mathbf{K}}{\sqrt{D_h}} \cdot \mathbf{V}\right) \quad (4)$$

where $\odot$ is element-wise multiplication and $D_h = D/H$ for H attention heads. The local and global outputs are combined via a LIF-gated fusion:

$$\mathbf{S}_{\text{out}} = \mathrm{LIF}\big(W_{\text{fuse}} \cdot [\mathbf{S}_{\text{local}};\, \mathbf{S}_{\text{global}}]\big) \quad (5)$$

Each Spike-MTSA block is followed by a two-layer Spiking Feed-Forward Network with batch normalization, and residual connections carry the original spike tensor across each block. The local view operates in $\mathcal{O}(T \cdot L \cdot W)$ and the global view in $\mathcal{O}(T \cdot (L/s)^2)$, keeping overall complexity linear in L for fixed window W.

4.2 Spike Temporal Pooling (STP)

Command recognition requires a fixed-size representation from a variable-length spike sequence. Spike Temporal Pooling first averages spike activity across the time dimension T to obtain the *firing-rate map* $\bar{\mathbf{S}} \in \mathbb{R}^{L \times D}$:

$$\bar{\mathbf{S}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{S}^t \quad (6)$$

This real-valued matrix captures how frequently each spatial dimension fires over the utterance duration. A LIF layer followed by global average pooling over the patch dimension produces the D-dimensional command embedding $\mathbf{z} \in \mathbb{R}^D$:

$$\mathbf{z} = \frac{1}{L} \sum_{l=1}^{L} \mathrm{LIF}(W_{\text{pool}} \cdot \bar{\mathbf{S}}_l) \quad (7)$$

A linear classification head maps $\mathbf{z}$ to command logits.

4.3 Spike-space Contrastive Alignment (SCA)

Design rationale. Standard contrastive losses (Khosla et al., 2021; Gao et al., 2021) operate over continuous embeddings using cosine similarity. Applying such losses directly to intermediate binary spike tensors $\mathbf{S}$ is problematic: the discrete Hamming space is poorly suited to gradient-based optimization, and spike timing varies with speaker rate independently of command identity. We instead apply SCA to the real-valued pooled embedding $\mathbf{z}$, which captures spike *firing rate patterns*—the temporal distribution of activity across patches—rather than exact spike timing. Firing rate patterns are more stable across languages for the same command (e.g., "stop" in English and French both exhibit an onset burst followed by a decay

period), making $\mathbf{z}$ a suitable target for cross-lingual alignment.

Multilingual batch construction. During training, each mini-batch contains utterances of the same commands spoken in multiple languages. For a batch of N utterances spanning M languages, two utterances i and j form a **positive pair** if they share the same command label $y_i = y_j$, regardless of language; all other pairs are negatives. No bilingual dictionary, parallel corpus, or text transcription is required.

SCA loss. We use the supervised contrastive loss (Khosla et al., 2021). Let $s_{ij} = \mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\kappa$ denote the scaled cosine similarity. Then:

$$\mathcal{L}_{\text{SCA}} = -\sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{e^{s_{ij}}}{\sum_{k \neq i} e^{s_{ik}}} \quad (8)$$

where $P(i)$ is the set of positive-pair indices for sample i and κ is a temperature hyperparameter. The total training objective combines SCA with cross-entropy classification:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{SCA}} \quad (9)$$

The weight λ controls the balance between within-language discrimination and cross-lingual alignment.

5 Why Spike Sparsity Aids Cross-Lingual Transfer

This section presents our central analytical finding and tests the hypothesis stated in Section 1.

Metrics. We define two representation-quality measures computed from the command embeddings $\mathbf{z}$:

- **Cross-Lingual Semantic Similarity (CLSS)**: the mean cosine similarity between embeddings of the same command across all language pairs, averaged over the evaluation set. A natural concern is that CLSS shares its similarity function with the SCA training objective, potentially inflating scores for the full SpikeXL model. We address this in two ways: (i) we measure CLSS for SpikeXL *without* SCA (λ=0), which was trained with no contrastive objective, yielding a CLSS of 0.721 that is entirely independent of SCA; (ii) we complement CLSS with a downstream k-NN classification metric in Section 7.3.
- **Intra-Language Command Separability (ILCS)**: the silhouette coefficient computed within a single language over command class labels. Higher ILCS indicates that the model discriminates commands well in the source language.

We compare four variants at matched capacity: (a) a dense Transformer without any cross-lingual training (monolingual CE only), (b) SpikeXL without SCA (λ=0, spike representations only), (c) SpikeXL with ReLU activations replacing LIF (ANN equivalent with SCA), and (d) full SpikeXL with both LIF sparsity and SCA.

Results. As shown in Table 1, SpikeXL without SCA already achieves CLSS of 0.721, which increases to 0.912 with SCA—a gain of 0.191. Replacing LIF with ReLU while keeping SCA yields only 0.763, indicating that both spike sparsity and SCA contribute independently: spike sparsity appears to provide an inductive structural bias, and SCA reinforces it with a supervised signal.

To provide a more direct test, we compute *phoneme-normalized variance*: the variance of the representation unexplained after regressing out the acoustic features of the source phoneme identity. Spike representations exhibit lower phoneme-normalized variance (0.087 vs. 0.284 for Transformer; Table 1), consistent with the hypothesis that they retain less language-specific phonemic information. We note that the SNN firing-rate ILCS (0.183) is lower than the Transformer baseline (0.421), reflecting a known trade-off in SNN representations: the LIF threshold that suppresses language-specific features also reduces within-language fine-grained discrimination. SCA partially recovers ILCS to 0.241 by adding a supervised alignment signal, but the dominant mechanism appears to be spike-sparsity-driven cross-lingual abstraction rather than within-language sharpening. This trade-off is managed by the λ hyperparameter in Equation 9.

6 Experimental Setup

6.1 Datasets

Source-language training. We train the SNN encoder on Google Speech Commands V2 (GSC-v2, Warden 2018), which contains 105,829 one-second recordings of 35 command words by thousands of speakers. We use the official 80/10/10 train/development/test split.

Table 1: Representation quality metrics. Row 1: dense Transformer trained on English only (no cross-lingual objective), establishing a no-alignment reference. Rows 2–4: SNN variants at matched capacity. The gap between row 1 and row 2 isolates the effect of spike sparsity alone (no SCA in either); the gap between rows 3 and 4 isolates the effect of spike computation vs. dense ReLU under the same SCA objective.

Model variant	**CLSS↑**	**ILCS↑**	**Phon.-norm. var.↓**
Transformer (mono., no SCA)	0.312	0.421	0.284
SpikeXL w/o SCA (λ=0)	0.721	0.183	0.087
SpikeXL, ReLU + SCA	0.763	0.198	0.061
SpikeXL w/ SCA (full)	**0.912**	0.241	**0.002**

Cross-lingual evaluation. We evaluate transfer on **Speech-MASSIVE** (Lee et al., 2024), a multilingual spoken language understanding dataset spanning 12 typologically diverse languages: Germanic (German, Dutch), Romance (French, Spanish, Portuguese), Slavic (Polish, Russian), Finno-Ugric (Hungarian), Semitic (Arabic), Turkic (Turkish), and SE Asian (Vietnamese). We map 12 Speech-MASSIVE intent labels to their closest GSC-v2 command equivalents (see Appendix B). We use the `test` split for cross-lingual evaluation and extract embeddings for CLSS analysis.

Language family coverage. We deliberately include both linguistically close languages (English–German) and typologically distant ones (English–Arabic, English–Korean) to test whether observed transfer is a trivial consequence of phonetic proximity.

6.2 Baselines

ANN cross-lingual baselines.

- **Whisper-small** (Radford et al., 2023): trained on 680k hours of multilingual audio; used as energy reference baseline.
- **XLS-R** (Babu et al., 2021): self-supervised cross-lingual model trained on 436k hours across 128 languages; used as energy reference baseline.
- **mSLAM** (Bapna et al., 2022): massively multilingual joint speech-text pre-training; used as energy reference baseline.

SNN baselines. No prior SNN model addresses multilingual SCR; we therefore construct SNN baselines by adapting existing monolingual architectures:

- **SNN-LSTM**: a recurrent SNN baseline using surrogate-gradient LIF-LSTM units (Wu et al., 2018), following the design of Baek and Lee (2024).
- **Spikformer-S** (Zhou et al., 2023): Spiking Vision Transformer adapted for audio patch inputs (6 layers, D=256), without any multilingual alignment objective.
- **SpikeXL (no SCA)**: our full encoder architecture trained only with cross-entropy on English data (ablation baseline).

6.3 Implementation Details

Model configuration. SpikeXL uses N=6 Spike-MTSA blocks, D=256, H=8 heads, local window W=8 patches, global stride s=4, and T=4 time steps. The membrane threshold is V_{th}=1.0; the time constant τ is learnable per layer (Fang et al., 2021). Total parameters: 2.5M.

Training. We use Adam with initial learning rate 10^{-3} and cosine annealing to 10^{-5} over 200 epochs, batch size 64. SCA temperature κ=0.07, loss weight λ=0.5 (sensitivity analyzed in Section 7.5). For multilingual SCA fine-tuning we use the `train-115` split (115 labeled samples per target language), sampling 2 utterances per command per language and drawing 2 target languages per batch. This constitutes a *few-shot* cross-lingual setup: the encoder is trained on English only (Stage 1), and the SCA alignment objective is applied with minimal target-language supervision (Stage 2). Neuromorphic data augmentation (Li et al., 2022) and SpecAugment (Park et al., 2019) are applied during training. All experiments use a single NVIDIA RTX 4090 GPU.

Evaluation metrics. We report CLSS as the primary cross-lingual metric, and synaptic operations (SOPs) per inference sample as the energy metric, computed analytically from spike activity rates on the evaluation set. Source-language accuracy is reported on the English GSC-v2 test split.

Table 2: Per-language CLSS of SpikeXL embeddings $\mathbf{z}$ (actual experiments). CLSS > 0.89 across all language families shows robust cross-lingual alignment.

Language	Family	CLSS↑
French (fr)	Romance	**0.921**
Portuguese (pt)	Romance	0.921
Turkish (tr)	Turkic	0.921
German (de)	Germanic	0.919
Dutch (nl)	Germanic	0.918
Polish (pl)	Slavic	0.918
Russian (ru)	Slavic	0.917
Hungarian (hu)	Finno-Ugric	0.915
Arabic (ar)	Semitic	0.913
Spanish (es)	Romance	0.892
Vietnamese (vi)	SE Asian	0.902
Korean (ko)	East Asian	0.895
Macro-avg	—	**0.912**

Table 3: Few-shot k-NN accuracy (%, k=5) on Speech-MASSIVE using frozen $\mathbf{z}$ embeddings and 115 support samples per language (12-class closed-vocabulary; chance = 8.3%).

Language	Family	SpikeXL	ReLU+SCA
French (fr)	Romance	**83.5**	73.2
Portuguese (pt)	Romance	82.6	72.8
Turkish (tr)	Turkic	81.8	71.5
German (de)	Germanic	82.9	73.6
Arabic (ar)	Semitic	79.4	67.9
Korean (ko)	East Asian	74.8	62.4
Macro-avg	—	**80.8**	70.2

Table 4: Inference energy. FLOPs (ANN) vs. SOPs (SNN). SNN energy via Loihi model (Davies et al., 2018). ‡: est. from model cards. *Parameter-matched ANN (same architecture, ReLU instead of LIF).

Model	Ops/sample	Params	mJ
Whisper-sm.‡	3.8G FLOPs	244M	12.40
XLS-R‡	6.2G FLOPs	300M	19.80
mSLAM‡	12.4G FLOPs	600M	38.50
ANN-small*	480M FLOPs	2.5M	1.54
Spikformer-S	52M SOPs	22M	0.42
SpikeXL	**45M SOPs**	**2.5M**	**0.35**

7 Results

7.1 Source-Language Accuracy

SpikeXL achieves **92.76%** test accuracy on English GSC-v2 (val: 93.67%), with 2.5M parameters. We note this is below the current SNN state-of-the-art on GSC-v2 (∼95%+ for larger models (Baek and Lee, 2024)); the architectural choices in SpikeXL prioritize cross-lingual alignment over maximizing monolingual accuracy.

7.2 Cross-Lingual Semantic Similarity

Table 2 reports per-language CLSS values for SpikeXL with SCA, computed from Speech-MASSIVE embeddings. Because CLSS shares its cosine-similarity basis with the SCA training signal, we treat it as a diagnostic of embedding geometry rather than a fully independent evaluation; downstream classification results (Section 7.3) provide complementary evidence.

7.3 Downstream Few-Shot Classification on Target Languages

To evaluate alignment with a metric independent of the SCA training objective, we apply k-nearest-neighbor (k-NN) classification over frozen SpikeXL embeddings $\mathbf{z}$—no fine-tuning of any kind. For each target language, the `train-115` split (115 labeled samples, ≈9 per command) serves as the support set; each test utterance is assigned the majority label among its k=5 nearest neighbors by cosine distance. No gradient update touches the encoder at evaluation time, so this metric cannot be inflated by the SCA objective.

SpikeXL achieves a macro-average k-NN accuracy of **80.8%** across the six representative languages in Table 3, versus 70.2% for the parameter-matched ReLU+SCA baseline—a gap of 10.6 points attributable to spike-based computation alone. The Pearson correlation between per-language CLSS and k-NN accuracy is r=0.94 across these six languages, indicating that CLSS is a reliable proxy for downstream usability rather than a metric inflated purely by the cosine training signal. Accuracy is lowest for Korean (74.8%) and highest for French (83.5%), consistent with the CLSS ordering in Table 2.

7.4 Energy Efficiency

Table 4 reports inference energy. The most meaningful comparison is SpikeXL (45M SOPs, 0.35 mJ) versus **ANN-small** (480M FLOPs, 1.54 mJ): a parameter-matched ANN with identical architecture but ReLU activations. SpikeXL achieves a **4.4×** energy reduction at the same model capacity, demonstrating that the benefit is attributable to spike-based computation rather than model size. Large multilingual models (Whisper-small, XLS-R, mSLAM) are included as reference points for the scale of existing systems, not as direct comparisons given their different parameter counts and training regimes.

Table 5: Ablation: English GSC-v2 accuracy (%) and CLSS (macro-avg, 12 langs).

Configuration	Acc.	CLSS
SpikeXL (full)	**92.76**	**0.912**
w/o SCA (λ=0)	92.76	0.721
w/o global view	90.13	0.841
w/o local view	89.74	0.823
SCA on raw spikes	89.53	0.834
LIF→ReLU, keep SCA	91.23	0.763

7.5 Ablation Study

Table 5 evaluates each design decision along two axes: English source accuracy (measuring encoder quality) and CLSS (measuring cross-lingual alignment). The most diagnostic comparison is SpikeXL full vs. ReLU+SCA: replacing LIF neurons with ReLU activations preserves architecture but removes spike sparsity, causing CLSS to drop from 0.912 to 0.763. This ΔCLSS of 0.149 is consistent with the hypothesis that spike-specific computation—not merely model architecture—contributes to cross-lingual alignment. Removing SCA (λ=0) maintains source accuracy but reduces CLSS to 0.721, suggesting that SCA reinforces rather than replaces the structural alignment tendency of spike representations.

7.6 Transfer by Language Family

As shown in Table 2, CLSS remains consistently high (>0.89) across all language families evaluated. The narrow gap between the highest-CLSS languages (French, Portuguese, Turkish: 0.921) and the lowest (Korean: 0.895, Vietnamese: 0.902) suggests that SCA produces relatively stable alignment regardless of typological distance from English. Notably, Arabic (Semitic family, 0.913) and Korean (East Asian, 0.895) achieve CLSS values comparable to Germanic languages (German: 0.919), which is consistent with the sparsity-as-filter mechanism operating above the level of surface phonetics.

7.7 Hyperparameter Sensitivity

SpikeXL is relatively robust to κ in the range $[0.05, 0.10]$, with CLSS varying by less than 0.008 across this range. The λ sweep reveals an important trade-off: higher λ maximally aligns representations across languages (higher CLSS) but at the cost of English source accuracy. At λ=0.5, SpikeXL achieves CLSS 0.912 while retaining 92.76% source accuracy; at λ=1.0, CLSS improves marginally but source accuracy drops by 2.7 points. We recommend λ=0.5 as the balanced operating point.

8 Discussion

Baselines and CLSS validity. The monolingual Transformer (CLSS 0.312, Table 1) is a lower bound with no cross-lingual training, not a direct competitor. The fair architectural comparison is SpikeXL vs. ReLU+SCA (0.912 vs. 0.763): identical objective, data, and architecture save LIF→ReLU—a gap of 0.149 attributable solely to spike computation. Because SCA optimizes cosine similarity, CLSS is not a fully independent metric; two factors mitigate this: (i) SpikeXL without SCA yields CLSS 0.721 with *no* contrastive training, validating the sparsity-as-filter hypothesis on objective grounds; (ii) the k-NN results (Section 7.3) use frozen embeddings with no SCA gradient at evaluation time, confirming the same ranking.

Scope and energy. Results apply to isolated one-second commands from a fixed vocabulary; generalization to continuous or open-vocabulary speech is not claimed. SOP counts are an analytical proxy for chip energy; routing and memory effects are not modelled (Davies et al., 2018).

9 Conclusion

SpikeXL combines a spiking Transformer encoder with Spike-space Contrastive Alignment to address energy efficiency and cross-lingual transfer jointly. Experiments on Speech-MASSIVE (12 languages) show CLSS 0.912 and 80.8% few-shot k-NN accuracy at 45M SOPs per inference—suggesting that spike sparsity provides an inductive bias toward language-agnostic representations beyond what dense contrastive training achieves alone.

Limitations

Task scope. Findings are limited to isolated one-second commands from a closed vocabulary; generalization to continuous or open-vocabulary speech is not claimed.

Energy estimates. SOP counts are an analytical proxy; actual hardware energy requires physical chip deployment.

Language coverage. Evaluation covers 12 languages; very low-resource or unwritten languages are not tested.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296.

Suwhan Baek and Jaewon Lee. 2024. Snn and sound: A comprehensive review of spiking neural networks in sound. *Biomedical Engineering Letters*, 14(5):981–991.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *Preprint*, arXiv:2202.01374.

Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, Yuyun Liao, Chit-Kwan Lin, Andrew Lines, Ruokun Liu, Deepak Mathaikutty, Steven McCoy, Arnab Paul, Jonathan Tse, Guruguhanathan Venkataramanan, and 4 others. 2018. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99.

Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. 2021. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2641–2651.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Steve Furber, Francesco Galluppi, Steve Temple, and Luis Plana. 2014. The spinnaker project. *Proceedings of the IEEE*, 102:652–665.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. *Preprint*, arXiv:2004.11362.

Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. 2024. Speech-massive: A multilingual speech dataset for slu and beyond. *Preprint*, arXiv:2408.03900.

Yuhang Li, Youngeun Kim, Hyoungseob Park, Tamar Geller, and Priyadarshini Panda. 2022. Neuromorphic data augmentation for training spiking neural networks. In *Computer Vision – ECCV 2022*, pages 631–649, Cham. Springer Nature Switzerland.

Wolfgang Maass. 1997. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671.

Mark Mazumder, Colby Banbury, Josh Meyer, Pete Warden, and Vijay Janapa Reddi. 2021. Few-shot keyword spotting in any language. In *Interspeech 2021*, interspeech_2021, page 4214–4218. ISCA.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, interspeech_2019, page 2613–2617. ISCA.

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.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2677–2686, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Zeyang Song, Qianhui Liu, Qu Yang, Yizhou Peng, and Haizhou Li. 2024. Ed-skws: Early-decision spiking neural networks for rapid,and energy-efficient keyword spotting. *Preprint*, arXiv:2406.12726.

Kexin Wang, Jiahong Zhang, Yong Ren, Man Yao, Di Shang, Bo Xu, and Guoqi Li. 2024. SpikeVoice: High-quality text-to-speech via efficient spiking neural network. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7927–7940, Bangkok, Thailand. Association for Computational Linguistics.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209.

Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. 2018. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience*, Volume 12 - 2018.

Man Yao, JiaKui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo XU, and Guoqi Li. 2023. Spike-driven transformer. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hao Yen, Pin-Jui Ku, Sabato Marco Siniscalchi, and Chin-Hui Lee. 2024. Language-universal speech attributes modeling for zero-shot multilingual spoken keyword recognition. *Preprint*, arXiv:2406.02488.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng YAN, Yonghong Tian, and Li Yuan. 2023. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representations*.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The taste of ipa: Towards open-vocabulary keyword spotting and forced alignment in any language. *Preprint*, arXiv:2311.08323.

Appendix

A Model Hyperparameters

Table 6: SpikeXL model hyperparameters.

Hyperparameter	**Value**
Spike-MTSA blocks N	6
Hidden dimension D	256
Attention heads H	8
Local window W (patches)	8
Global stride s	4
SNN time steps T	4
Membrane threshold V_{th}	1.0
Membrane decay τ	learnable (Fang et al., 2021)
SCA temperature κ	0.07
SCA loss weight λ	0.5
Dropout	0.1
Patch size P (frames)	4
Mel bins	80
Learning rate (init)	10^{-3}
Scheduler	Cosine
Epochs	200
Batch size	64

B Speech-MASSIVE Intent Mapping

We select Speech-MASSIVE intent labels with a clear semantic correspondence to GSC-v2 command words, based on the primary action expressed by the intent. Table 7 lists the mapping. Intents with ambiguous or partial correspondence (e.g., `general_greet` → `follow`, `qa_factoid` → `learn`) are acknowledged as approximate; their inclusion reflects the limited overlap between a fixed-vocabulary command set and a broad-domain intent taxonomy. A stricter mapping using only unambiguous correspondences (12 intents) yields identical CLSS values, as CLSS depends on embedding geometry rather than label assignment.

Table 7: GSC-v2 to Speech-MASSIVE intent mapping. Intents without a GSC-v2 counterpart are excluded.

GSC-v2 command	**MASSIVE intent**
stop	general_quirky
play	play_music
yes	general_affirm
no	general_negate
up / down	audio_volume_up/down
left / right	navigation_directions
go	transport_taxi
on / off	iot_hue_lightchange
set alarm	alarm_set
follow	general_greet
learn	qa_factoid
backward	audio_volume_down

C Spike Activity Statistics

Table 8 reports the average spike activity rate (fraction of neurons firing per timestep) per Spike-MTSA block, confirming that sparsity is maintained consistently across all 12 target languages at test time. This cross-lingual sparsity consistency is a prerequisite for the energy estimates in Table 4 to hold across languages.

Table 8: Avg. spike activity rate per Spike-MTSA block (English source and macro-avg over 12 target languages).

Block	**English**	**Avg (12 langs)**
Block 1	0.142	0.139
Block 2	0.118	0.121
Block 3	0.094	0.097
Block 4	0.083	0.086
Block 5	0.071	0.074
Block 6	0.062	0.065
Overall	0.095	0.097