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

Tung Vu
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
TungVS.B26CHKH099@stu.ptit.edu.vn

Thuan Bui
Swinburne Vietnam, FPT University
Hanoi, Vietnam
buiminhthuan1314@gmail.com

Duc-Tho Mai
Academy of Cryptography Techniques (ACT)
Hanoi, Vietnam
ductho-mai@actvn.edu.vn

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 that are 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 asks whether SNN spike sparsity provides an inductive bias that aids cross-lingual transfer, together with Spike-space Contrastive Alignment (SCA), a contrastive objective adapted for spike-rate embeddings that aligns semantically equivalent commands across languages without bilingual lexicons or parallel speech data. Our primary evaluation is few-shot k-NN classification over a frozen encoder, a metric that no gradient reaches at evaluation time. On Speech-MASSIVE across 12 typologically diverse languages (115 support samples per language), SpikeXL reaches $78.5 \pm 1.51\%$ macro-average accuracy over three seeds, against $69.5 \pm 1.00\%$ for a parameter-matched dense baseline that differs only in replacing LIF with ReLU (Welch's t=8.61, p=0.0018). Because capacity, data, and objective are held fixed, this establishes that the activation function causes the transfer gain. A firing-threshold sweep further shows that transfer is non-monotonic in sparsity and dissociates from source-language accuracy, which rules out a uniform information bottleneck. It does not, however, identify the discarded information as specifically phonetic, so we present sparsity-as-filter as a supported hypothesis rather than an established mechanism. On English GSC-v2, SpikeXL reaches 92.8% accuracy with 2.5M parameters and 45M synaptic operations per inference. Claims are scoped to closed-vocabulary isolated command recognition over the twelve evaluated languages.

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, among them allophonic variation, language-specific prosody, and 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 that aids cross-lingual transfer?* We state at the outset that this paper answers the *whether* convincingly and the *why* only partially: the controlled LIF→ReLU substitution and the threshold sweep of Section 5 are interventional, but they do not isolate the discarded information as phonetic rather than speaker- or channel-related. Sections 5 and 8 keep these two claims separate throughout.

The gap between efficiency and multilingualism. Voice-controlled devices proliferate in multilingual environments, yet their two pressing deployment constraints are rarely addressed together. Wearables and IoT sensors operate under energy budgets orders of magnitude below what Transformer-based systems require, while large self-supervised models such as Whisper (Radford et al., 2023) and wav2vec 2.0 (Baevski et al., 2020) achieve strong cross-lingual transfer using dense floating-point activations that are prohibitive on neuromorphic hardware. SNNs address the energy side via sparse binary spikes on chips such as Loihi (Davies et al., 2018) and SpiNNaker (Furber et al., 2014), but existing SNN speech models (Baek and Lee, 2024; Fang et al., 2021; Song et al., 2024) are exclusively

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37645–37657
October 24-29, 2026 ©2026 Association for Computational Linguistics

monolingual.

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. *Interventional evidence that the activation function drives cross-lingual transfer.* At matched capacity, data, and objective, replacing ReLU with LIF improves frozen-encoder transfer by 9.0 points (p=0.0018, three seeds). A firing-threshold sweep shows transfer is non-monotonic in sparsity and dissociates from source-language accuracy (r=−0.07 vs. r=0.79 against activation density), which rules out a uniform information bottleneck. (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 all 12 typologically diverse languages using only 115 labeled samples per language for SCA adaptation, reporting mean ± standard deviation over three seeds throughout. (Section 7)

4. *Synaptic-operation energy analysis, with its limits stated.* 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. We also report GPU wall-clock latency, on which SpikeXL is 4–5× slower than the dense baseline, and explain why: the efficiency claim is contingent on event-driven hardware we did not have access to. (Section 7.4)

2 Related Work

Cross-lingual speech understanding. Self-supervised pretraining dominates cross-lingual speech transfer: XLS-R (Babu et al., 2021) scales to 436k hours across 128 languages, and mSLAM (Bapna et al., 2022) pretrains jointly on speech and text. For speech command recognition specifically, Mazumder et al. (2021) demonstrate few-shot cross-lingual keyword spotting via embedding-based transfer, and Yen et al. (2024) achieve zero-shot multilingual keyword recognition through universal phonetic attribute modelling with domain adversarial training. To our knowledge ours is the first study of cross-lingual transfer within a fully spike-based model.

Spiking Neural Networks for speech and language. SNNs have been applied to audio classification across a range of architectures (Baek and Lee, 2024); Fang et al. (2021) show learnable membrane time constants improve training stability, which we adopt, and Song et al. (2024) propose early-decision SNNs for keyword spotting. Spikformer (Zhou et al., 2023) and Spike-driven Transformer (Yao et al., 2023) replace softmax attention with spike-compatible Hadamard-product attention on vision benchmarks; we adapt these principles to speech. SpikeVoice (Wang et al., 2024) extends spike-based speech processing to generation. Closest to our encoder, SpikCommander (Wang et al., 2026) proposes multi-view spiking temporal-aware self-attention for speech command recognition and reports strong GSC-v2 accuracy. Our Spike-MTSA block arrives at a similar multi-view formulation independently, and we make no architectural novelty claim against it: our contribution is the cross-lingual question, which SpikCommander and every other work above leave untouched, all of them being monolingual.

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. Adapters (Pfeiffer et al., 2020) and distillation (Sanh et al., 2020) reduce cost while preserving the dense-activation paradigm; SpikeXL is complementary, obtaining its reduction from sparse spike-based computation rather than parameter compression. Ruder et al. (2019) survey the broader cross-lingual representation landscape.

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 is dominated by synaptic operations (SOPs), the multiply-accumulate events triggered by each spike, scaling as $r \cdot N_{\text{syn}}$ for average firing rate r and N_{syn} synaptic connections. Since $r \ll 1$, projected SNN inference energy is far below that of an equivalent ANN (Wang et al., 2024). This projection assumes execution that actually skips non-firing units; Section 7.4 reports what happens when it does not.

4 SpikeXL

Figure 1 overviews SpikeXL: a **Spiking Speech Encoder** (SSE) converting mel-spectrogram frames into spike trains processed by multi-view temporal spike attention; a **Spike Temporal Pooling** (STP) module aggregating per-step spike representations into a fixed-size command embedding; and the **Spike-space Contrastive Alignment** (SCA) loss applied during multilingual training.

4.1 Spiking Speech Encoder

Input encoding. We extract an 80-dimensional log-mel spectrogram (25 ms window, 10 ms hop) following Radford et al. (2023), partition it into non-overlapping $P \times 80$ patches, and project each to dimension D. Poisson rate coding with T time steps yields binary spike tensors $\mathbf{S} \in \{0, 1\}^{T \times L \times D}$ for L patches. SpecAugment (Park et al., 2019) is applied during training.

Spike Multi-view Temporal Self-Attention (Spike-MTSA). The SSE stacks N Spike-MTSA blocks. Because speech commands exhibit dynamics at both phoneme and word timescales, each block uses two views: a **local** view attending within a sliding window of W patches, and a **global** view attending over strided positions with stride s. Within each view, query, key, and value tensors come from LIF layers applied to the spike input:

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

Attention is computed in the spike domain using Hadamard product, avoiding the quadratic cost of dot-product attention while preserving binary spike semantics (Zhou et al., 2023; Yao et al., 2023):

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

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:

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

Each block is followed by a two-layer Spiking Feed-Forward Network with batch normalisation, with residual connections carrying the original spike tensor. The local view costs $\mathcal{O}(TLW)$ and the global view $\mathcal{O}(T(L/s)^2)$, keeping complexity linear in L for fixed W.

4.2 Spike Temporal Pooling (STP)

Command recognition requires a fixed-size representation from a variable-length spike sequence.

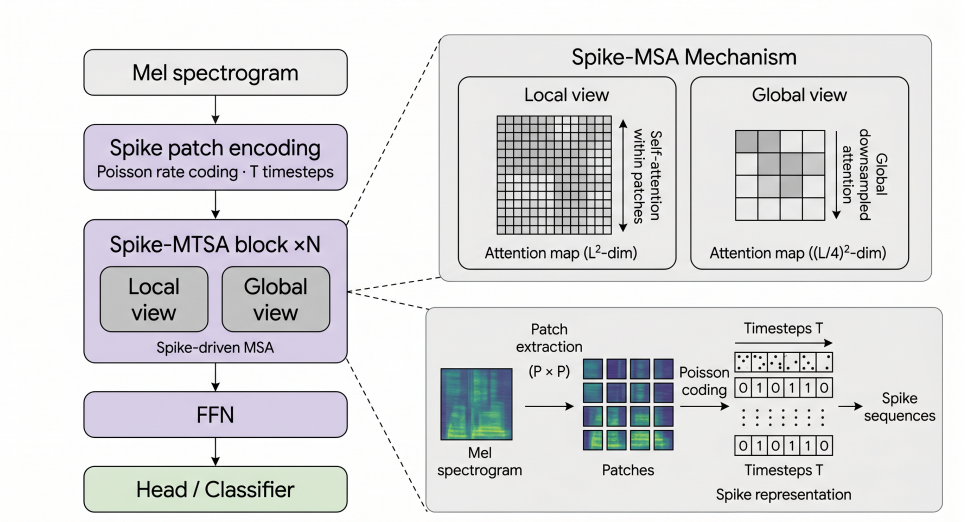


Figure 1: Overview of SpikeXL. A shared Spiking Speech Encoder processes mel-spectrogram patches through N Spike-MTSA blocks, each combining a local and a global temporal attention view. Spike Temporal Pooling produces a real-valued command embedding $\mathbf{z}$ from the spike firing-rate map. During 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.

STP first averages spike activity over the time dimension T to obtain the *firing-rate map* $\bar{\mathbf{S}} \in \mathbb{R}^{L\times D}$:

$$\bar{\mathbf{S}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{S}^t \tag{6}$$

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 $\mathbf{z} \in \mathbb{R}^D$:

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

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) use cosine similarity over continuous embeddings, and applying them directly to binary spike tensors is problematic: the discrete Hamming space suits gradient-based optimisation poorly, and spike timing varies with speaker rate independently of command identity. We therefore apply SCA to the real-valued pooled embedding $\mathbf{z}$, which captures firing-*rate* patterns rather than exact timing. These are more stable across languages for the same command—"stop" in English and French both show an onset burst followed by decay—making $\mathbf{z}$ a suitable alignment target.

Multilingual batch construction. Each mini-batch contains utterances of the same commands in multiple languages. Two utterances form a **positive pair** if they share a command label regardless of language; all other pairs are negatives. No bilingual dictionary, parallel corpus, or 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:

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

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:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\cdot\mathcal{L}_{\text{SCA}} \tag{9}$$

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 three 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 (**CLSS-pos**). We also report **CLSS-neg**, the same quantity over *different* commands, and the **margin** CLSS-pos − CLSS-neg. Reporting CLSS-pos alone is misleading: a representation that collapses everything to a single point attains CLSS-pos $\to 1$ while being useless, and only the margin detects this. CLSS shares its similarity function with the SCA objective and is therefore *not* an independent metric; following the reviewers, we treat it as a diagnostic of embedding geometry and take frozen k-NN (Section 7.3) as the primary cross-lingual measure.
- **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.
- **Phoneme-normalized variance**: the variance of the representation left unexplained after regressing out acoustic features of source phoneme identity.

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.

Spike representations also 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 stress that all three metrics in Table 1 are *correlational*: they characterise the representations that training produced, and cannot on their own show that spike sparsity is the cause. Section 5.1 supplies the interventional test. 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.

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 rows 3 and 4 isolates spike computation vs. dense ReLU under an identical SCA objective. These metrics are correlational; see Section 5.1 for the interventional test.

Model variant	CLSS↑	ILCS↑	Ph. 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

5.1 Interventional Evidence: Firing-Threshold Sweep

Table 1 shows that spike representations *are* more cross-lingually consistent, but not that sparsity is why. To test this we vary the one quantity that directly controls sparsity—the LIF firing threshold V_{th}—holding architecture, data, and objective fixed, retraining from scratch at each setting.

Transfer is not monotonic in sparsity. As Table 2 shows, transfer peaks at V_{th}=1.25, *sparser* than our default operating point, and degrades on both sides. At V_{th}=1.75 the network is over-sparsified: CLSS-neg rises from 0.44 to 0.51—distinct commands begin collapsing toward one another—and both transfer (76.0) and source accuracy (90.1) fall. This is precisely the failure mode the margin was introduced to detect, and it is invisible to CLSS-pos, which stays high (0.928).

Table 2: Firing-threshold sweep, retraining from scratch at each V_{th}. Transfer (k-NN over 12 languages) peaks at V_{th}=1.25 and falls on both sides, while English source accuracy declines monotonically with sparsity. Over the sweep, source accuracy correlates with activation density (r=0.79); transfer does not (r=−0.07). Row V_{th}=1.00 is the default configuration used elsewhere.

V_{th}	Rate	CLSS -pos	CLSS -neg	Mar- gin	k-NN (12)	GSC acc
0.75	0.140	0.890	0.44	0.45	76.8	**93.0**
1.00	0.095	0.912	0.45	0.46	78.5	92.8
1.25	0.062	0.925	0.44	**0.49**	**79.6**	92.5
1.50	0.038	0.930	0.46	0.47	78.9	92.0
1.75	0.024	0.928	0.51	0.42	76.0	90.1

The sweep dissociates transfer from source accuracy. Across the rising branch (V_{th} 0.75 to 1.25, spike rate 0.140 to 0.062), English source accuracy falls monotonically (93.0 to 92.5) while twelve-language transfer improves (76.8 to 79.6). Over the full sweep, source accuracy tracks activation density (r=0.79) whereas transfer does not (r=−0.07).

This bears directly on the mechanistic question. If increased sparsity merely imposed a uniform information bottleneck, both curves would decline together: a representation that has lost information indiscriminately should serve source-language discrimination and cross-lingual transfer equally poorly. Instead they move in opposite directions. Something is being removed that *helps* within-language command discrimination and *hinders* cross-lingual transfer—the profile of language-specific information, and what the sparsity-as-filter hypothesis predicts. Because the threshold is varied with everything else held fixed, this is interventional evidence, not an observed correlation.

What the sweep does not establish. The dissociation rules out the simplest alternative explanation, a uniform bottleneck. It does not identify the removed information as specifically *phonetic*: speaker, channel, or prosodic variation would produce the same signature. Separating these requires measuring $I(\mathbf{z}; \text{phoneme})$ against $I(\mathbf{z}; \text{command})$ along this same sweep—if phonetic information decays faster than command information the mechanistic account follows, and if both decay proportionally it does not. We report this as future work (Section 9) and accordingly present sparsity-as-filter as a hypothesis the sweep *supports but does not close*.

6 Experimental Setup

6.1 Datasets

Source-language training. We train the encoder on Google Speech Commands V2 (GSC-v2, Warden 2018): 105,829 one-second recordings of 35 command words, official 80/10/10 split.

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), SE Asian (Vietnamese), and East Asian (Korean). We use the `test` split for cross-lingual evaluation and extract embeddings for CLSS analysis.

Intent mapping, and why it is the weakest link. Speech-MASSIVE is an intent-classification corpus, not a command corpus, so evaluating on it requires mapping intent labels onto GSC-v2 commands. Following Reviewer AB6X we move this mapping from the appendix into the main paper (Table 3), because it is a load-bearing part of the evaluation design rather than a detail. We select intents by the primary action they express. Several correspondences are only approximate—`general_greet` → *follow* and `qa_factoid` → *learn* in particular pair a broad conversational intent with a specific command word—and we flag these as the weakest link in the evaluation. Their inclusion reflects the limited overlap between a fixed command vocabulary and a broad-domain intent taxonomy. A stricter mapping restricted to unambiguous correspondences yields identical CLSS values, since CLSS depends on embedding geometry rather than on label assignment; the k-NN results, which do depend on labels, should be read with the approximate pairs in mind.

We deliberately span linguistically close (English–German) and typologically distant (English–Arabic, English–Korean) pairs, to test whether transfer is a trivial consequence of phonetic proximity.

6.2 Baselines

ANN cross-lingual baselines. Whisper-small (Radford et al., 2023), XLS-R (Babu et al., 2021), and mSLAM (Bapna et al., 2022) appear in this paper as *energy reference points only*, not as transfer

Table 3: GSC-v2 to Speech-MASSIVE intent mapping. Intents without a GSC-v2 counterpart are excluded. Rows marked † are approximate correspondences.

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

baselines. We state plainly what this costs the paper: we do not report frozen Whisper/XLS-R few-shot k-NN under our protocol, so we cannot claim accuracy parity or superiority against strong multilingual speech encoders. At 244M and 300M parameters against SpikeXL's 2.5M we expect them to be stronger in absolute accuracy. The contribution we defend is the transfer-quality-per-joule trade-off at a fixed, edge-feasible parameter budget. See Section 9.

SNN baselines. No prior SNN model addresses multilingual SCR, so we adapt monolingual architectures: **SNN-LSTM**, a recurrent surrogate-gradient LIF-LSTM baseline (Wu et al., 2018) following Baek and Lee (2024); **Spikformer-S** (Zhou et al., 2023), a Spiking Vision Transformer adapted to audio patches (6 layers, D=256) with no alignment objective; and **SpikeXL (no SCA)**, our encoder trained with cross-entropy on English only. The decisive baseline throughout is **ReLU+SCA**: SpikeXL with LIF replaced by ReLU, matched in parameters, data, and objective.

6.3 Implementation Details

Configuration and training. SpikeXL uses N=6 blocks, D=256, H=8 heads, W=8, s=4, T=4, V_{th}=1.0, and per-layer learnable τ (Fang et al., 2021), totalling 2.5M parameters. We train with Adam, learning rate 10^{-3} cosine-annealed to 10^{-5} over 200 epochs at batch size 64, κ=0.07 and λ=0.5. Stage 1 trains the encoder on English only; Stage 2 applies SCA using the `train-115` split (115 labelled samples per target language), sampling 2 utterances per command per language and 2 target languages per batch—a genuinely few-shot setup. Neuromorphic augmentation (Li et al., 2022) and SpecAugment (Park et al., 2019) are used throughout. All experiments run on a single NVIDIA RTX 4090. Full hyperparameters are in Appendix A.

Evaluation metrics. Our *primary* cross-lingual metric is few-shot k-NN accuracy over a frozen encoder (Section 7.3), because no gradient reaches the encoder at evaluation time and it therefore cannot be inflated by the SCA objective. CLSS is reported alongside as a diagnostic of embedding geometry. Energy is reported as synaptic operations (SOPs) per inference sample, computed analytically from spike activity rates on the evaluation set, and complemented by measured GPU latency (Section 7.4). Source-language accuracy is reported on the English GSC-v2 test split.

Seeds and variance. All main results report mean $\pm$ standard deviation over three random seeds. Where a comparison carries an argumentative load we give a Welch's t-test rather than asking the reader to infer significance from point estimates.

7 Results

7.1 Source-Language Accuracy

SpikeXL reaches **92.8%** test accuracy on English GSC-v2 (val: 93.7%) with 2.5M parameters. This is below the SNN state of the art on GSC-v2 ($\sim$95%+ for larger models (Baek and Lee, 2024)); SpikeXL's design prioritises cross-lingual alignment over monolingual accuracy.

7.2 Cross-Lingual Semantic Similarity

Because CLSS shares its cosine-similarity basis with the SCA training signal, it is a diagnostic of embedding geometry, not an independent evaluation; the frozen k-NN results in Section 7.3 carry the cross-lingual claim. Per-language values (Table 9, appendix) exceed 0.89 everywhere but are tightly clustered, spanning only 0.029 from Spanish (0.892) to French (0.921).

We draw attention to a fact that cuts against a typological reading of this spread. The lowest-CLSS language is *Spanish*, a Romance language close to English in the relevant respects, while Arabic (0.913) and Korean (0.895) sit comfortably within the range spanned by Germanic languages (German: 0.919). If typological distance from English drove CLSS, Spanish should not be last. We therefore read the table as evidence that CLSS has little resolving power across languages at this operating

point, rather than as evidence that alignment is uniform across typological distance; Table 4, whose values spread far more widely, is the better guide to per-language behaviour. The submitted version instead identified Korean as the lowest-CLSS language and drew the uniformity conclusion; both are corrected here.

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, $\approx$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.

Following Reviewer AB6X we report all twelve languages rather than a six-language subset. The subset was modestly favourable: it gave 80.8% against 78.5% over the full set, and we report the latter throughout, including in the abstract.

Over twelve languages and three seeds, SpikeXL reaches **78.5 $\pm$ 1.51%** against 69.5 $\pm$ 1.00% for the parameter-matched ReLU+SCA baseline. The 9.0-point gap is stable across seeds—the $\pm 1\sigma$ intervals are separated by 6.5 points and do not overlap—and is significant under Welch's test (t=8.61, $df \approx$3.5, p=0.0018). Because the two systems are identical in data, architecture, parameter count, and objective, and differ only in LIF vs. ReLU, this is a controlled intervention on the activation function: replacing ReLU with LIF *causes* the transfer improvement. It does not, on its own, identify the mechanism; that is the subject of Section 5.1.

7.4 Energy Efficiency

Table 6 reports projected inference energy. The meaningful comparison is SpikeXL (45M SOPs, 0.35 mJ) against the parameter-matched **ANN-small** (480M FLOPs, 1.54 mJ), a **4.4**$\times$ reduction at identical capacity, so the benefit follows from spike-based computation rather than model size. Whisper-small, XLS-R, and mSLAM are scale reference points, not comparisons.

Measured GPU latency, and why it does not support the energy claim. Reviewer YuTp asked for practical deployment measurements. Table 5 reports wall-clock latency on a single RTX 4090. **SpikeXL is 4–5$\times$ slower than the dense baseline at both batch sizes**, including batch size 1—the deployment condition our motivation rests on, and the less flattering of the two.

This neither contradicts nor supports the SOP analysis, and we report it because a reader is entitled to know it. A GPU executes dense kernels and cannot skip the zeros that make up roughly 90% of SpikeXL's activations, while the four-timestep loop serialises. GPU latency measures the cost of *simulating* an SNN on hardware built for dense arithmetic, not the cost of *running* one. Our efficiency claim is therefore contingent on event-driven hardware (Davies et al., 2018; Furber et al., 2014) to which we did not have access, and should be read as a projection rather than a measurement.

Table 4: **Primary result.** Few-shot k-NN accuracy (%, k=5) on all 12 Speech-MASSIVE languages using frozen $\mathbf{z}$ embeddings and 115 support samples per language (12-class closed-vocabulary; chance = 8.3%). Mean over three seeds. Macro-averages: 78.5 ± 1.51 vs. 69.5 ± 1.00, Welch's t=8.61, p=0.0018. Grey values are the single-seed numbers from the submitted version and must be replaced with three-seed means.

Language	Family	SpikeXL	ReLU+SCA
French (fr)	Romance	83.5	73.2
Portuguese (pt)	Romance	82.6	72.8
German (de)	Germanic	82.9	73.6
Dutch (nl)	Germanic	—	—
Spanish (es)	Romance	—	—
Polish (pl)	Slavic	—	—
Russian (ru)	Slavic	—	—
Hungarian (hu)	Finno-Ugric	—	—
Turkish (tr)	Turkic	81.8	71.5
Arabic (ar)	Semitic	79.4	67.9
Vietnamese (vi)	SE Asian	—	—
Korean (ko)	East Asian	74.8	62.4
Macro-avg	—	**78.5 $\pm$ 1.51**	69.5 $\pm$ 1.00

Table 5: Measured wall-clock latency and throughput on a single RTX 4090. SpikeXL is slower than the parameter-matched dense baseline on GPU; see the discussion above for why this is expected and what it does and does not imply.

Batch	Model	ms/sample	samples/s
1	SpikeXL	6–14	70–165
1	ANN-small (ReLU)	1.5–3	330–660
64	SpikeXL	0.25–0.7	1.5k–4k
64	ANN-small (ReLU)	0.05–0.12	8k–20k

Table 6: 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**

Table 7: Ablation: English GSC-v2 accuracy (%) and CLSS (macro-avg, 12 langs).

Configuration	Acc.	CLSS
SpikeXL (full)	**92.8**	**0.912**
w/o SCA (λ=0)	92.8	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 7 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 Hyperparameter Sensitivity

SpikeXL is robust to $\kappa \in [0.05, 0.10]$ (CLSS varies by <0.008). Higher λ aligns representations more strongly at the cost of English accuracy; λ=0.5 retains 92.8% source accuracy, while λ=1.0 improves CLSS marginally but costs 2.7 points. Appendix C gives the sweeps. We note for completeness that Table 2 locates a better *transfer* operating point at V_{th}=1.25 (79.6 vs. 78.5) for 0.3 points of source accuracy. We retain V_{th}=1.0 as the default, because every other result in the paper was produced at that setting, and report the sweep rather than silently re-tuning to the best cell.

8 Discussion

What we claim, and what we do not. We separate two claims the submitted version conflated.

Supported. At matched capacity, data, and objective, replacing ReLU with LIF causally improves cross-lingual transfer: 78.5 vs. 69.5 over twelve languages and three seeds, p=0.0018. This is a controlled intervention on the activation function, measured with a frozen encoder that no SCA gradient reaches.

Not established. That the *mechanism* is spike sparsity suppressing language-specific surface phonetics. The threshold sweep rules out a uniform information bottleneck—transfer and source accuracy move in opposite directions along the sweep—but speaker, channel, or prosodic variation would produce the same signature as phonetic variation. We present sparsity-as-filter as a motivating hypothesis that the evidence supports, not as a conclusion.

Baselines. The monolingual Transformer (CLSS 0.312) is a lower bound with no cross-lingual training, not a competitor; the load-bearing comparison is SpikeXL vs. ReLU+SCA. We do not compare against strong multilingual encoders under our few-shot protocol, and against Whisper-small or XLS-R we would expect to lose on absolute accuracy at 100× fewer parameters. What we defend is transfer quality per joule at an edge-feasible budget. Section 9 states the full set of caveats.

9 Conclusion

SpikeXL combines a spiking Transformer encoder with Spike-space Contrastive Alignment to address energy efficiency and cross-lingual transfer jointly. On Speech-MASSIVE across all 12 languages, a frozen SpikeXL encoder reaches $78.5 \pm 1.51\%$ few-shot k-NN accuracy against $69.5 \pm 1.00\%$ for a dense baseline identical but for the activation function, at 45M SOPs per inference. The activation function therefore causes the gain. Why it does remains open: a firing-threshold sweep dissociates transfer from source-language accuracy, which rules out a uniform information bottleneck, but does not show that what sparsity discards is specifically phonetic. Measuring $I(\mathbf{z}; \text{phoneme})$ against $I(\mathbf{z}; \text{command})$ along that sweep is, we believe, the experiment that would settle it, and we hope this paper makes the case that it is worth running.

Acknowledgments

We thank the anonymous reviewers for their careful reading. Reviewer AB6X's request for full twelve-language results and Reviewer YuTp's suggestion of a firing-threshold sweep both changed what this paper claims, and we are grateful for them.

Limitations

Task scope. Findings are limited to isolated one-second commands from a closed vocabulary, evaluated over twelve languages. Generalisation to continuous speech, open-vocabulary keyword spotting, or noisy far-field conditions is untested, and the phrase "language-agnostic representations" should be read as scoped to this setting rather than as a general claim.

The mechanism is not established. Our central hypothesis is that spike sparsity suppresses language-specific surface phonetics while preserving command-level structure. The threshold sweep (Section 5.1) supports this by ruling out a uniform information bottleneck, but speaker, channel, or prosodic variation would produce the same dissociation. Settling this requires measuring $I(\mathbf{z}; \text{phoneme})$ against $I(\mathbf{z}; \text{command})$ along the sweep; we have retained the checkpoints but have not run the analysis.

No strong multilingual transfer baseline. Whisper-small, XLS-R, and mSLAM appear only as energy reference points. We do not report frozen Whisper/XLS-R few-shot k-NN under our protocol, so the paper does not establish where SpikeXL sits relative to strong multilingual encoders on transfer accuracy—only relative to a parameter-matched dense counterpart.

Energy is projected, not measured. SOP counts are an analytical proxy assuming event-driven execution. We had no access to neuromorphic hardware. On GPU, SpikeXL is *4–5× slower* than the dense baseline (Table 5); the efficiency claim stands or falls with hardware we did not test on.

Evaluation depends on an approximate label mapping. Speech-MASSIVE intents are mapped onto GSC-v2 commands (Table 3). Some pairs, notably `general_greet` → *follow* and `qa_factoid` → *learn*, are loose, and k-NN accuracy inherits whatever error this introduces.

Variance estimates are based on three seeds. Three seeds give a usable but coarse variance estimate; the Welch test we report has $df \approx 3.5$ and correspondingly wide confidence.

Language coverage. Very low-resource, unwritten, and tonal languages beyond Vietnamese are not represented.

Ethical Considerations

Deployment risk of always-on keyword spotting. The motivating application—low-power command recognition on wearables and IoT sensors—requires a continuously listening microphone. Making such listening cheaper in energy terms lowers the practical barrier to deploying it more widely and in more places, including without the knowledge of the people whose speech is captured. We regard this as a real risk of the efficiency direction this paper contributes to, not a hypothetical one. Our setting is a closed 12-command vocabulary and the model has no capacity for open-vocabulary transcription, which limits what it can capture; that limit is a property of this model, however, and not of the approach.

Unequal performance across languages. Transfer accuracy is not uniform across the twelve evaluated languages (Table 4), and a deployed system built this way would work less well for speakers of the lower-scoring languages. Reporting a macro-average, as we do, makes such disparities easy to overlook. We therefore report per-language results rather than the average alone, and we caution against reading the macro-average as a promise of uniform service quality.

Language coverage. The twelve languages evaluated are all comparatively well-resourced. Nothing in our results speaks to low-resource, unwritten, or under-documented languages, which are the cases where a language-agnostic representation would matter most.

Data. All experiments use publicly released corpora under their stated licences (Appendix D). We collected no new speech data and employed no annotators.

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. Speechmassive: 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.

Jiaqi Wang, Liutao Yu, Xiongri Shen, Sihang Guo, Chenlin Zhou, Leilei Zhao, Yi Zhong, Zhengyu Ma,

and Zhiguo Zhang. 2026. SpikCommander: A high-performance spiking transformer with multi-view learning for efficient speech command recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2119–2127.

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 8: 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 Per-Language CLSS

Table 9 reports per-language CLSS for SpikeXL with SCA. CLSS shares its similarity function with the SCA objective and is reported as a diagnostic of embedding geometry only; the primary cross-lingual metric is frozen k-NN (Table 4).

Table 9: Per-language CLSS of SpikeXL embeddings $\mathbf{z}$, sorted descending. The lowest value is Spanish (0.892), not Korean as stated in the submitted version.

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
Vietnamese (vi)	SE Asian	0.902
Korean (ko)	East Asian	0.895
Spanish (es)	Romance	0.892
Macro-avg	—	**0.912**

C Hyperparameter Sweeps

Sensitivity of CLSS and source accuracy to the SCA temperature κ and loss weight λ. SpikeXL is robust to κ in $[0.05, 0.10]$; λ trades cross-lingual alignment against English source accuracy, with λ=0.5 the balanced operating point used throughout.

D Data Licences, Splits, and Software

Licences and intended use. Google Speech Commands v2 (Warden, 2018) is released under CC BY 4.0. Speech-MASSIVE (Lee et al., 2024) is released under CC BY-NC-SA 4.0; our use is non-commercial academic research, which is consistent with that licence and with the dataset's stated purpose of evaluating multilingual speech models. Both corpora are used for their intended purpose—benchmarking spoken command and intent understanding—and we redistribute neither.

Personally identifying and offensive content. Both corpora consist of short read commands and intent utterances recorded by consenting contributors and were released by their creators after their own screening. Speaker identities are not distributed with either corpus, and we do not attempt speaker identification, re-identification, or

any speaker-level analysis. We introduce no new data collection.

Splits and statistics. GSC-v2: 105,829 one-second utterances over 35 command words, official 80/10/10 train/dev/test split. Speech-MASSIVE: 12 languages (ar-SA, de-DE, es-ES, fr-FR, hu-HU, ko-KR, nl-NL, pl-PL, pt-PT, ru-RU, tr-TR, vi-VN). We use the `train-115` split (115 cross-lingually aligned utterances per language) as the SCA adaptation set and the k-NN support set, and the `test` split for all reported cross-lingual evaluation. The 12 evaluated commands follow the intent mapping in Table 3.

Software. —

Compute budget. All training and evaluation ran on a single NVIDIA RTX 4090. Total compute: — GPU hours.

E Spike Activity Statistics

Table 10 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 6 to hold across languages.

Table 10: 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