

Low-Complexity Sign Language Recognition: A Cross-Modal Distillation Approach for State-Space Models

Victor Costa
vrcosta@inf.ufpel.edu.br
PPGC, ViTech, AEMIA - UFPel
Pelotas/RS, Brazil

Lucas Superti
lsdsilva@inf.ufpel.edu.br
PPGC, ViTech, AEMIA - UFPel
Pelotas/RS, Brazil

Brenda Salenave Santana
bssalenave@inf.ufpel.edu.br
PPGC, AEMIA - UFPel
Pelotas/RS, Brazil

Tatiana Bolivar Lebedeff
tblebedeff@gmail.com
PPGL, AEMIA - UFPel
Pelotas/RS, Brazil

Guilherme Corrêa
gcorrea@inf.ufpel.edu.br
PPGC, ViTech, AEMIA - UFPel
Pelotas/RS, Brazil

ABSTRACT

Sign Language Recognition (SLR) is a fundamental challenge for digital accessibility and interactive multimedia systems. While state-of-the-art approaches that rely on 3D Convolutional Neural Networks (e.g., I3D) achieve high accuracy by extracting dense spatiotemporal features directly from RGB videos, their high computational cost makes them prohibitive for real-time edge computing applications. Conversely, lightweight models based on spatial landmarks significantly reduce computational overhead but suffer from a severe drop in accuracy due to the loss of implicit visual semantic context, such as subtle hand depths and environmental textures. To address the dichotomy between computational efficiency and semantic richness, this paper proposes a Cross-Modal Knowledge Distillation pipeline for Brazilian Sign Language (LIBRAS) recognition. We utilize an I3D network operating on raw video as the Teacher model to transfer "soft probabilities" to a Mamba 1D Student model operating exclusively on spatial coordinates extracted via MediaPipe. Experimental results demonstrate that our cross-modal distillation approach successfully bridges the semantic gap. While a vanilla Mamba 1D model (trained without distillation) achieves only 71% signer-independent accuracy, our proposed distilled approach elevates this performance to 78%. Remarkably, this 7% absolute improvement introduces zero computational overhead during inference. The proposed distilled Mamba 1D model achieves real-time latency (0.21 ms), requires only 9.81 MB of VRAM, and consumes a mere 0.004 Joules per inference, representing a 99.7% reduction in energy consumption compared to the video-based baseline. These results demonstrate the viability of deploying highly accurate, semantically aware sign language translation systems natively on resource-constrained mobile and IoT devices.

KEYWORDS

LIBRAS, sign language, sign language recognition, neural networks, knowledge distillation, Mamba.

1 INTRODUCTION

Modern technology has the potential to bridge long-standing social gaps, yet individuals with disabilities are still often excluded from many of its benefits. Among these groups, the deaf and hard-of-hearing community continues to face significant barriers in communication, education, and access to services. In Brazil, accessibility is a legal right guaranteed by legislation [13, 14], which recognizes LIBRAS (*LIngua BRasileira de Sinais* – Brazilian Sign Language) as a legal means of communication and expression. Approximately 5% of the Brazilian population lives with some degree of hearing impairment [12], highlighting the urgency of inclusive solutions. Consequently, scientific research must actively address these inequalities by developing interactive multimedia systems that support natural and ubiquitous communication.

Sign Language Recognition (SLR) is a fundamental challenge in this domain. Historically, the Computer Vision community has approached this problem by employing robust Deep Learning models, such as 3D Convolutional Neural Networks (e.g., I3D [3]), to process raw RGB videos. For instance, in our previous work [4], we demonstrated that extracting dense spatial-temporal features directly from video frames achieves state-of-the-art performance, reaching 92.5% accuracy under a strict signer-independent evaluation protocol. However, these video-based models present a challenging barrier for real-world interactive multimedia applications: their high computational cost. Processing three-dimensional video tensors demands high-end hardware, making native execution (Edge Computing) on mobile devices or public service terminals impractical due to high latency, large memory footprint and excessive energy consumption.

To circumvent these computational bottlenecks, a common alternative in the literature is the use of spatial articulation extractors, or landmarks (such as MediaPipe[11]), which reduce the dimensionality of the problem from millions of pixels to small 1D vectors. While highly efficient, this approach introduces a severe semantic gap. By discarding the original

video stream in favor of a mathematical skeleton, the model loses a vast array of implicit visual contexts, such as subtle hand depths, facial micro-expressions, and environmental textures. In the baseline experiments presented in this paper, replacing the raw-video I3D network with a highly efficient State-Space Model (Mamba 1D) [7] operating on spatial coordinates caused the signer-independent accuracy to drop to 71%, clearly illustrating the cost of losing multimedia semantic richness.

To resolve the dichotomy between computational efficiency and semantic comprehension, this paper proposes a Cross-Modal Knowledge Distillation framework for LIBRAS recognition. The core premise is to utilize the robust I3D network as a "Teacher" model, capable of modeling the full video semantics, to train and calibrate the weights of an extremely lightweight Mamba 1D "Student" model. Unlike traditional distillation, the transfer of *soft probabilities* occurs across distinct modalities: the teacher processes RGB matrices, while the student operates exclusively on temporal sequences of 1D landmarks.

The main contributions of this paper are:

- The design and implementation of an efficient Mamba 1D architecture for LIBRAS recognition that operates in real-time, requiring only 9.81 MB of VRAM and consuming a mere 0.004 Joules per inference.
- A novel Cross-Modal Knowledge Distillation pipeline that successfully transfers rich visual semantics from a heavy video-based model to a lightweight landmark-based model.
- Empirical evidence demonstrating that our distilled approach elevates the signer-independent accuracy of the lightweight baseline from 71% to 78% (a 7% absolute improvement) without introducing any computational overhead during inference, representing a 99.7% reduction in energy consumption compared to the previous video-based state-of-the-art.

The remainder of this paper is organized as follows. Section 2 reviews related work in sign language recognition. Section 3 introduces the background on the knowledge distillation paradigm. Section 4 details the proposed cross-modal distillation framework, including data pre-processing, the Inflated 3D ConvNet, and the Mamba 1D architecture. Section 5 reports the experimental results, comparing accuracy, computational cost, and energy consumption. Finally, Section 6 concludes the paper and discusses future directions.

2 RELATED WORK

Recent literature on Sign Language Recognition (SLR) has predominantly focused on maximizing classification accuracy, often at the expense of computational efficiency. A common approach involves leveraging heavy, video-based architectures to extract spatial-temporal features directly from RGB frames. For instance, Silva et al. [5] utilized pre-computed I3D [3] features to process synthetic datasets generated from the V-LIBRASIL corpus [18]. While their method successfully

supported sentence-level translation, it did not address end-to-end fine-tuning for isolated signs or signer generalization. Similarly, Fanucchi et al. [6] explored Vision Transformers (ViTs) by fine-tuning a Video Masked Autoencoder (VideoMAE) for Augmented Reality applications. Although they achieved 84% accuracy with a subset of signers, their study highlights the inherent risk of overfitting and the severe computational demands associated with processing raw video using Transformer-based architectures.

The MINDS-Libras dataset [16], which serves as the foundation for our experiments, was initially validated by its creators, Rezende et al. [17]. They employed a 3D Convolutional Neural Network (3D CNN) on RGB videos, reporting an accuracy of 93.3%. However, this baseline was established using a random "per-sign" data split, which allows frames from the same signer to appear in both training and testing sets, potentially masking the model's true generalization capacity.

To address the generalization gap, our recent study [4] established a strict signer-independent benchmark for the MINDS-Libras dataset. By fine-tuning an Inflated 3D ConvNet (I3D) architecture directly on raw RGB videos, we achieved a state-of-the-art accuracy of 92.5% on unseen signers. Despite this high accuracy, the deployment of such a heavy 3D convolution pipeline proved to be highly impractical for edge computing, requiring substantial GPU memory and presenting prohibitive energy consumption rates for real-time mobile applications.

Seeking alternatives to heavy video processing, some researchers have pivoted to landmark-based approaches. Alves et al. [1] proposed extracting body, hand, and facial coordinates using OpenPose [2], subsequently encoding these spatio-temporal sequences into a single 2D image to feed a pre-trained ResNet-18 [8]. While conceptually innovative, the landmark extraction phase alone imposed a severe bottleneck, taking approximately 36 seconds per video sequence, precluding real-time usage.

A critical analysis of the related work reveals a persistent dichotomy in the field: models are either highly accurate but computationally prohibitive (e.g., I3D[3], VideoMAE[20]), or they rely on lightweight landmark representations that either fail to capture deep visual semantics or employ extraction pipelines too slow for real-time inference. Our current work fills this gap by introducing a Cross-Modal Knowledge Distillation [19] strategy, bridging the semantic richness of video-based models with the ultra-low latency of lightweight landmark sequence models like Mamba 1D [7].

3 KNOWLEDGE DISTILLATION PARADIGM

Knowledge Distillation (KD), originally formalized by Hinton et al. [9], is a machine learning paradigm designed to transfer knowledge from a large, highly parameterized model (the "Teacher") to a smaller, more efficient model (the "Student"). The core premise is that a trained Teacher produces not only a final discrete class prediction but also a continuous distribution of probabilities across all classes. These relative

probabilities, often referred to as “soft probabilities”, encode structural similarities between classes (e.g., the implicit visual similarity between two distinct signs) that are completely absent in standard one-hot encoded hard labels.

To extract this hidden knowledge, a temperature scalar T is introduced into the standard *softmax* function. For a given vector of raw logits z produced by a network, the softened probability q_i for class i is computed as:

$$q_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} \quad (1)$$

When $T = 1$, this equation reduces to the standard softmax function, producing a sharp probability distribution. As the temperature T increases, the probability distribution becomes softer, revealing the Teacher’s secondary and tertiary predictions, thus exposing its internal reasoning process and topological priors.

During the distillation process, the Student model is trained using a composite loss function $\mathcal{L}$ that balances two distinct objectives. The first objective is the standard Cross-Entropy loss ($\mathcal{L}_{CE}$) between the Student’s standard predictions and the ground-truth hard labels y . The second objective is the Kullback-Leibler Divergence ($\mathcal{L}_{KL}$) between the Student’s softened probabilities p_T^S and the Teacher’s softened probabilities p_T^T . The total distillation loss is defined as:

$$\mathcal{L} = (1 - \alpha) \mathcal{L}_{CE}(y, p_{T=1}^S) + \alpha T^2 \mathcal{L}_{KL}(p_T^S, p_T^T) \quad (2)$$

where α is a hyperparameter ($0 \leq \alpha \leq 1$) that controls the relative influence of the Teacher’s knowledge against the true labels. The T^2 multiplier is essential to ensure that the magnitude of the gradients produced by the soft targets matches those produced by the hard targets during backpropagation. By optimizing this composite function, the lightweight Student network learns to mimic the deep semantic understanding of the Teacher.

4 PROPOSED LOW-COMPLEXITY SLR

This section presents our Cross-Modal Knowledge Distillation pipeline for Sign Language Recognition. As illustrated in Figure 1, the proposed architecture is composed of two parallel modalities that converge during the optimization phase: a heavy Video Modality processed offline by the Teacher, and a lightweight Landmark Modality processed by the Student during training.

The pipeline begins with the **MINDS-Libras Dataset** [16], which provides the raw RGB videos and the corresponding ground truth hard labels. From this data source, the flow splits into two distinct processing paths:

- **Video Modality (Offline Extraction):** The upper block of Figure 1 represents the Teacher’s domain. The raw video pixels are fed into a robust Inflated 3D ConvNet (I3D). Because the Teacher is already fully trained and

its weights are frozen, this step can be performed offline. The I3D network outputs the raw Teacher logits (z_T). These logits are then passed through the temperature-scaled softmax function ($T = 4.0$) to generate the “Soft Targets”, which encapsulate the rich spatial-temporal soft probabilities of the original video.

- **Landmark Modality (Training Phase):** The middle block outlines the Student’s domain. To circumvent the computational cost of processing video tensors, the raw frames are processed by MediaPipe Holistic [11] to extract 1D spatial landmarks (x, y, z coordinates). This lightweight temporal sequence is fed into the trainable Mamba 1D Student model. The model produces the Student logits (z_S), which are then processed in two ways: through a standard LogSoftmax ($T = 1.0$) for hard label comparison, and through a temperature-scaled LogSoftmax ($T = 4.0$) to generate “Soft Predictions” for distillation.
- **Knowledge Distillation Optimization:** The rightmost block details the composite loss calculation. The KL Divergence Loss compares the Teacher’s Soft Targets with the Student’s Soft Predictions, while the Cross-Entropy Loss evaluates the Student’s standard predictions against the ground truth hard labels. These two metrics are aggregated into a Total Loss (L) using a balancing weight ($\alpha = 0.5$). Finally, the backpropagation updates *only* the weights of the Mamba 1D Student, as indicated by the dashed feedback loop.

The following subsections detail the specific architectural components and the dataset that instantiate this pipeline. Subsection 4.1 describes the I3D Teacher network. Subsection 4.2 introduces the Mamba 1D architecture and the landmark extraction process. Subsection 4.3 presents the dataset, and Subsection 4.4 outlines the training procedure, formalized step-by-step in Algorithm 1, along with the optimization hyperparameters.

4.1 The Teacher: Inflated 3D ConvNet (I3D)

The Inflated 3D ConvNet (I3D) architecture, introduced by Carreira and Zisserman [3], is a highly robust video classification solution. The key idea behind I3D is to extend well-established 2D convolutional neural networks (CNNs) into the spatio-temporal domain by inflating their 2D filters (e.g., 3×3) into 3D kernels (e.g., $3 \times 3 \times 3$). This enables the network to simultaneously capture spatial features (such as texture, shape, and background context) within individual frames and temporal dynamics across consecutive frames.

A major advantage of I3D is its ability to reuse pre-trained weights from large-scale image datasets, providing a strong prior for spatial structure, while temporal learning is fine-tuned on video-specific datasets like Kinetics-400 [10]. In this work, we employ an I3D model with a ResNet-50 backbone. As demonstrated in a recent study [4], this configuration effectively captures the deep visual semantics of LIBRAS signs directly from raw pixels, achieving high recognition accuracy. Consequently, it serves well as a “Teacher” network

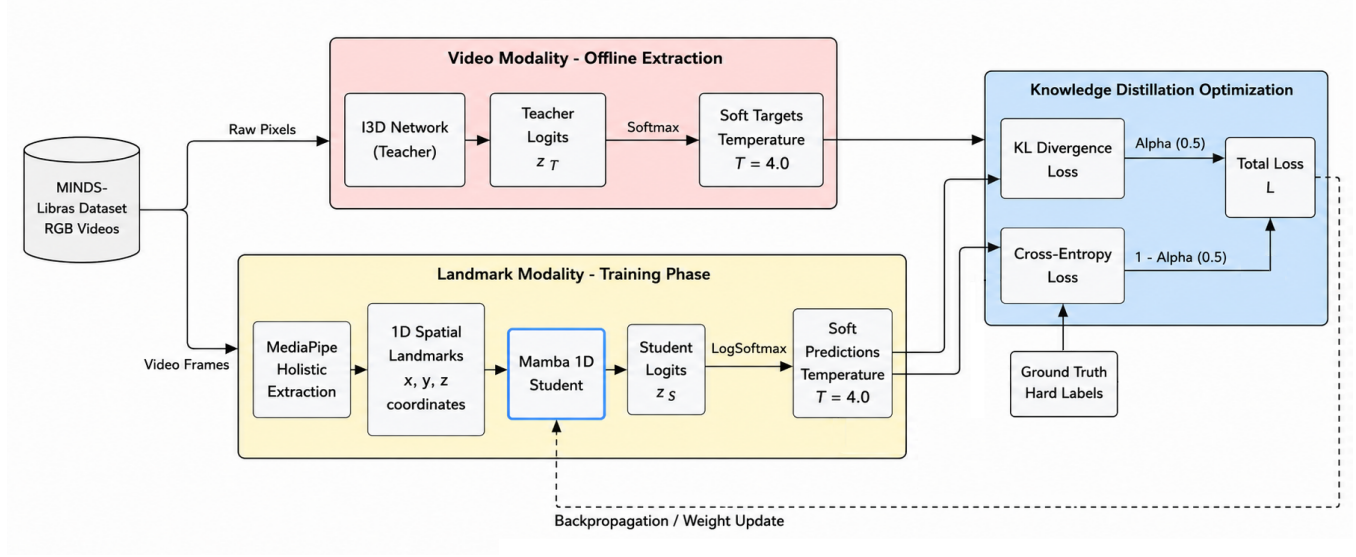


Figure 1: Overview of the proposed Cross-Modal Knowledge Distillation pipeline. The Teacher model (I3D) extracts dense semantics from raw RGB video to generate soft targets, while the lightweight Student model (Mamba 1D) processes efficiently extracted spatial landmarks. The distillation is governed by a composite loss function that balances hard labels and the Teacher’s soft probabilities.

in our distillation pipeline, encapsulating the complex motion patterns and visual richness of the signs.

Algorithm 1 Cross-Modal Knowledge Distillation for Mamba 1D Student

Require: Training dataset $\mathcal{D}_{train}$ containing tuples X, y, z^T , where X is the temporal sequence of MediaPipe landmarks, y is the ground-truth label, and z^T is the pre-computed I3D Teacher logits vector.

Require: Hyperparameters: Temperature T , distillation weight α , learning rate η , total epochs E .

Ensure: Optimized Mamba 1D Student weights θ^S .

```

1: Initialize Student model parameters  $\theta^S$ 
2: for epoch = 1 to  $E$  do
3:   for each mini-batch  $X, y, z^T \in \mathcal{D}_{train}$  do
4:     Forward pass:  $z^S \leftarrow \text{Mamba1DX}; \theta^S$  {Compute student raw logits}
5:     Compute hard loss:  $\mathcal{L}_{CE} \leftarrow \text{CrossEntropyLoss}(y, z^S)$ 

6:     Soften teacher targets:  $p^T \leftarrow \text{Softmax}(z^T / T)$ 
7:     Soften student outputs:  $p^S \leftarrow \text{LogSoftmax}(z^S / T)$ 
8:     Compute soft loss:  $\mathcal{L}_{KL} \leftarrow \text{KL\_Divergence}(p^S, p^T) \times T^2$ 
9:     Combine losses:  $\mathcal{L} \leftarrow 1 - \alpha \mathcal{L}_{CE} + \alpha \mathcal{L}_{KL}$ 
10:    Backward pass: Compute gradients  $\nabla_{\theta^S} \mathcal{L}$ 
11:    Update parameters:  $\theta^S \leftarrow \theta^S - \eta \nabla_{\theta^S} \mathcal{L}$  {Using AdamW optimizer}
12:   end for
13: end for
14: return  $\theta^S$ 

```

4.2 The Student: Mamba 1D and Landmark Extraction

While I3D is highly accurate, its reliance on heavy 3D video tensors makes it unsuitable for real-time Edge Computing. To achieve ultra-low latency and minimal energy consumption, we design a "Student" architecture based on State-Space Models (SSMs), specifically the Mamba architecture [7]. Unlike traditional Transformers that suffer from quadratic computational complexity regarding sequence length, Mamba operates with linear time complexity. It selectively processes continuous sequential data through a parameterized hidden state, making it fast and memory-efficient for long temporal sequences.

To adapt the video data for the 1D Mamba architecture, we discard the raw RGB pixels and rely exclusively on spatial coordinates. We utilize the MediaPipe Holistic framework [11] to extract structural landmarks from each video frame. Specifically, the holistic topology captures a total of 543 landmarks per frame, distributed across the facial mesh, body pose, and both hands, as detailed in Table 1. Since each landmark is represented by three-dimensional coordinates (x, y, z) , the extraction process yields a dense, flattened 1D vector of 1,629 spatial features per frame.

By converting heavy video frames into these lightweight 1D temporal vectors, the Mamba model can process the signs with a fraction of the computational cost, relying on the Teacher’s distilled knowledge to compensate for the lost visual semantics.

Table 1: Distribution of MediaPipe Holistic landmarks extracted per frame

Anatomical Region	Landmarks	Features (x, y, z)
Face Mesh	468	1,404
Pose (Body)	33	99
Right Hand	21	63
Left Hand	21	63
Total per Frame	543	1,629

4.3 MINDS-Libras Dataset

The MINDS-Libras dataset, presented in Figure 2 forms the foundation of our study. It was developed and publicly released by Rezende et al. [16] to address the scarcity of well-documented and publicly available datasets for LIBRAS research. The dataset originally comprises 20 distinct LIBRAS signs, each recorded 5 times by 12 different signers, aiming for a total of 1,200 samples. It includes multiple data modalities: RGB videos, depth videos, spatial coordinates of 25 body joints, and 1,347 facial points. Due to data loss during recording – particularly for some signs from signers #3, #4, and #9 – the final usable set contains 1,155 RGB videos.

The signs were carefully selected to ensure diversity in phonological parameters such as hand configuration, articulation point, movement, palm orientation, and non-manual expressions. Furthermore, the signers represent a mix of perceived gender, ages, and prior knowledge of LIBRAS, enhancing the dataset’s real-world representativeness.

Data were captured using a Canon EOS Rebel T5i RGB camera and a Kinect v2 RGB-D sensor, with videos provided in MP4 format. In our pipeline, the Teacher I3D model utilizes the raw RGB video modality, while the Student Mamba model relies on landmarks independently extracted from these same videos. The full list of signs included in the dataset, along with their English translations, is detailed in Table 2. Note that this work does not focus on Brazilian Portuguese to English translation; the English terms are provided solely as approximate glosses to assist non-Portuguese-speaking readers.

4.4 Training and Knowledge Distillation

The training and knowledge distillation processes were performed using an NVIDIA GeForce RTX 4090 GPU. The operating system was Ubuntu 24.04 LTS, running on a machine with an Intel® Core™ i7 processor and 48 GB of RAM. Due to the ultra-lightweight nature of the 1D landmark sequences compared to raw video tensors, we were able to significantly increase the *batch size* to 64 without encountering memory constraints, optimizing the hardware utilization.

The Student model (Mamba 1D) was trained from scratch using the AdamW optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 1×10^{-4} . To ensure stability and smooth convergence over the training cycle, a Cosine

Table 2: Sign classes in the MINDS-Libras dataset and their English translations

Label	Sign Meaning (Portuguese)	Translation (English)
W1	Acontecer	To Happen
W2	Aluno	Student
W3	Amarelo	Yellow
W4	America	America
W5	Aproveitar	To Enjoy
W6	Bala	Candy
W7	Banco	Bank
W8	Banheiro	Bathroom
W9	Barulho	Noise
W10	Cinco	Five
W11	Conhecer	To Know
W12	Espelho	Mirror
W13	Esquina	Corner
W14	Filho	Son
W15	Maçã	Apple
W16	Medo	Fear
W17	Ruim	Bad
W18	Sapo	Frog
W19	Vacina	Vaccine
W20	Vontade	Will

Annealing Learning Rate Scheduler was applied, gradually decaying the learning rate to a minimum of 1×10^{-6} over a total of 150 epochs.

The optimization was guided by the composite distillation loss described in Section 3. The hyperparameters for the Knowledge Distillation were empirically determined through a grid search over the validation set. The optimal balance between the hard labels and the Teacher’s soft probabilities was achieved with a distillation weight of $\alpha = 0.5$, indicating an equal contribution from both sources. The temperature scalar was set to $T = 4.0$, which proved to be the ideal value to soften the I3D Teacher’s logits, exposing the underlying semantic relationships between LIBRAS signs without diluting the probability distribution into uniform noise.

5 EXPERIMENTAL RESULTS

This section presents the evaluation of the proposed cross-modal knowledge distillation pipeline. To ensure a rigorous and realistic assessment of the model’s generalization capabilities to unseen individuals, we adopted a strict signer-independent train-test split, following the protocol established in our previous work [4]. Specifically, signers #11 and #12 were isolated exclusively for the test set, comprising 200 video samples, while the remaining 10 signers were used for training and validation. The experiments were executed on an NVIDIA GeForce RTX 4090 GPU, matching the hardware baseline of our aforementioned I3D study for a fair comparison.



Figure 2: Signers in the MINDS-Libras dataset. The number of stars indicates each signer’s prior knowledge of LIBRAS: one star for beginner, two stars for intermediate, and three stars for advanced.

Table 3: Overall performance and computational cost comparison during inference

Model	Acc (%)	VRAM (MB)	Lat. (ms)	Energy (J)
Teacher (I3D)	92.50	~450.00	824.00	1.330
Vanilla Mamba	71.00	9.81	0.21	0.004
Distilled Mamba	78.00	9.81	0.21	0.004

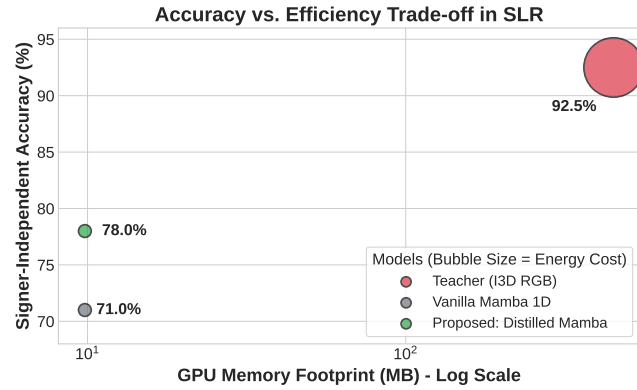


Figure 3: Accuracy vs. Efficiency trade-off. The x-axis (log scale) represents the GPU memory footprint. The bubble size corresponds to the energy consumed per inference. Our distilled model improves accuracy by 7% absolute points with zero added inference cost.

5.1 Accuracy vs. Efficiency Trade-off

The primary goal of this work is to bridge the semantic gap between heavy video-based models and lightweight landmark-based models. Table 3 summarizes the macroscopic results of our experiments, comparing the I3D Teacher, the vanilla Mamba 1D baseline, and our proposed Distilled Mamba 1D.

As anticipated, transitioning from raw RGB video (I3D) to 1D spatial landmarks (Vanilla Mamba) causes a severe drop in the signer-independent accuracy, plummeting from 92.5% to 71.0%. However, the computational gains are massive: VRAM usage drops from 450 MB to less than 10 MB, and energy consumption per inference drops from 1.33 Joules to a negligible 0.004 Joules.

By applying our cross-modal distillation framework, the Distilled Mamba 1D achieved a global accuracy of 78.0%. As illustrated in Figure 3, this represents a substantial absolute

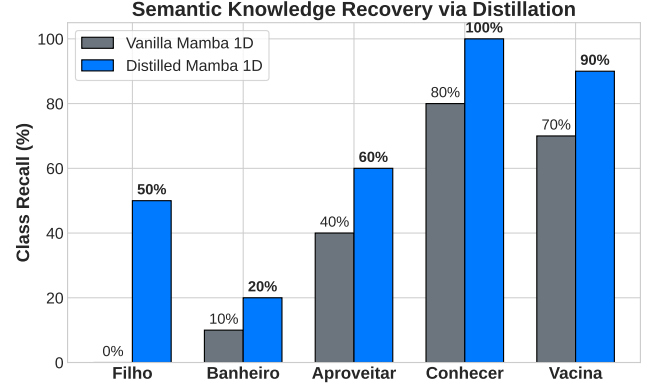


Figure 4: Class-specific recall improvement demonstrating the semantic recovery effect. Distillation allows the Student to capture subtle relational patterns previously accessible only to the video-based Teacher.

gain of 7% over the vanilla baseline. Most importantly, because the Teacher is discarded after training, this semantic recovery introduces *zero overhead* during inference. The distilled model operates with the exact same ultra-low latency (0.21 ms) and energy footprint (0.004 Joules) as the vanilla baseline, offering a highly superior accuracy-to-cost ratio for Edge Computing.

5.2 Semantic Recovery and Error Analysis

To understand how the transfer of “soft probabilities” affected the model’s internal representations, we analyzed the class-specific recall rates. Figure 4 demonstrates the semantic recovery phenomenon on selected challenging classes.

Without distillation, the vanilla model entirely failed to recognize the sign W14 (*Filho/Son*), yielding a 0% recall rate. Guided by the Teacher’s soft probabilities, the distilled model was able to extract hidden structural similarities, resurrecting the class to a 50% recall rate. Similar significant boosts were observed across other context-heavy signs such as W11 (*Conhecer/To Know*) and W5 (*Aproveitar/To Enjoy*).

Despite the 78% overall accuracy, the confusion matrix revealed specific bottlenecks where the lack of raw pixel data



Figure 5: Visual comparison of Signer #1 performing the signs (a) W2 (*Aluno/Student*) and (b) W19 (*Vacina/Vaccine*). The reliance on 1D spatial coordinates exacerbates the visual ambiguity caused by identical articulation points.

still severely limits the model. The most critical misclassification occurred with sign W2 (*Aluno/Student*), which the distilled model incorrectly predicted as W19 (*Vacina/Vaccine*) in 90% of the test instances.

This persistent error can be analyzed through the phonological parameters of sign languages [15]. As shown in Figure 5, both *Aluno* and *Vacina* share the exact same articulation point (contact with the upper arm/shoulder area) and involve a rapid, inward-directed movement. The primary contrast between these two signs lies in the fine hand configuration (a flatter hand for student versus a specific finger-pointing grip mimicking a syringe for vaccine) and the micro-texture of the contact.

While the heavy I3D network easily captures the depth shadows and the precise pixel-level occlusion of the hand against the arm, the MediaPipe [11] landmark extractor reduces this entire interaction to a sparse set of x, y, z dots hovering near the shoulder. Consequently, the Mamba 1D network receives highly ambiguous spatio-temporal sequences for both classes. Even with the Teacher’s guidance during training, if the input signals (landmarks) for W2 and W19 are mathematically almost identical, the Student is forced to guess, often defaulting to the class with the strongest topological attractor in its latent space. This highlights the inherent limitations of skeleton-only approaches and sets a clear path for future research in hybrid early-fusion architectures.

5.3 Comparison with Related Work

Table 4 presents a comparison of our proposed approach against related works evaluating the MINDS-Libras [16] dataset. While previous studies [1, 4, 17] report higher accuracies (above 92%), they achieve these results at a prohibitive computational cost, making them largely incompatible with real-time Edge Computing constraints.

For instance, while Alves et al. [1] reach 93.0% accuracy, they report a massive pre-processing bottleneck of approximately 36,000 ms per sign just for landmark extraction. Similarly, a recent study [4] proposed an end-to-end I3D architecture that eliminates the skeleton extraction phase, achieving

Table 4: Comparison with works using the MINDS-Libras dataset

Work	Accuracy (%)	Inference Latency
Fanucchi et al. [6]	84.00	Not Available
Rezende et al. [17]	93.30	Not Available
Alves et al. [1]	93.00	$\sim 36,000.00$ ms*
I3D Baseline [4]	92.50	~ 824.00 ms
Ours (Distilled Mamba 1D)	78.00	~ 0.21 ms

*Reported pre-processing time (landmark extraction).

an impressive 92.5% under a strict signer-independent protocol. However, that video-based approach still requires about 824 ms per inference on a high-end GPU, which remains computationally heavy for continuous, real-world multimedia applications.

In contrast, our proposed Distilled Mamba 1D prioritizes ultra-low latency and edge deployability. Although the absolute accuracy of 78.0% is lower than the heavy video-based models, our solution requires a mere 0.21 ms per inference. This represents a processing speed orders of magnitude faster than the I3D baseline [4], completely circumventing the heavy tensor operations required by 3D convolutions. The performance gap – with our execution being thousands of times faster than competing methods – demonstrates the practicality and necessity of cross-modal distillation for deploying semantically aware LIBRAS recognition systems on resource-constrained devices.

6 CONCLUDING REMARKS

This study addressed the critical bottleneck of deploying Brazilian Sign Language (LIBRAS) recognition systems on resource-constrained devices, bridging the gap between high-accuracy but computationally prohibitive video models and highly efficient but semantically limited skeleton-based approaches. By introducing a Cross-Modal Knowledge Distillation framework, we successfully transferred the rich, spatio-temporal “soft probabilities” of a heavy I3D network into an ultra-lightweight Mamba 1D model operating solely on spatial landmarks. Under a strict, completely unseen test set protocol, our proposed model achieved an accuracy of 78%. Crucially, it executes inferences in only 0.21 ms while consuming a mere 0.004 Joules in our experimental environment. This represents a nearly 99.7% reduction in energy cost compared to the video-based baseline.

The primary contribution of this work lies in demonstrating that the semantic degradation inherent to 1D structural representations can be significantly mitigated without incurring any computational penalty during inference. Our distillation pipeline facilitated a 7% absolute accuracy recovery over the vanilla baseline, partially rescuing context-heavy signs by guiding the Student model’s latent space with the Teacher’s topological priors. Furthermore, by maintaining a rigorous signer-independent evaluation protocol, we ensure that our ultra-efficient pipeline, which operates orders of magnitude faster than current state-of-the-art video or transformer-based

solutions, is genuinely prepared for real-world generalization in Edge Computing environments.

As future work, we intend to tackle the remaining spatial ambiguities identified in our error analysis, which stem from the complete removal of pixel-level micro-textures and hand occlusions. To this end, we plan to explore hybrid early-fusion architectures that inject lightweight, localized visual patches (e.g., dynamic hand cropping) into the Mamba 1D state-space representation, balancing semantic richness and linear computational complexity. Finally, extending this cross-modal distillation framework to Continuous Sign Language Recognition (CSLR) and evaluating its scalability across larger, unconstrained datasets remain key priorities for advancing ubiquitous and accessible technologies.

7 ETHICAL CONSIDERATIONS

This research utilized the publicly available MINDS-Libras dataset [16]. The authors did not conduct direct studies involving human subjects for this article. Therefore, this work complies with the Brazilian National Health Council (CNS) Resolution No. 510/2016.

Regarding the Brazilian General Data Protection Law (LGPD – Law No. 13.709/2018), the public data was handled exclusively for academic research purposes. The original dataset creators obtained the necessary ethical approvals and consent from all participants prior to publication. Data processing and storage were conducted securely to preserve the integrity of the information. Additionally, the images presented in this manuscript do not expose individuals who are minors or belong to vulnerable groups.

ACKNOWLEDGMENTS

The authors of this work would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) Finance Code 001, CNPq, and FAPERGS for funding this research.

Generative artificial intelligence (AI) tools, specifically Gemini 3.1 pro, were used exclusively to assist with English language editing, grammar correction, and textual revision of this manuscript. The authors reviewed and verified all modifications and assume full responsibility for the content, interpretations, and conclusions presented in the paper.

REFERENCES

- [1] Carlos Eduardo G. R. Alves, Francisco De A. Boldt, and Thiago M. Paixão. 2024. Enhancing Brazilian Sign Language Recognition Through Skeleton Image Representation. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 1–6. <https://doi.org/10.1109/SIBGRAPI62404.2024.10716301>
- [2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 43, 01 (Jan. 2021), 172–186. <https://doi.org/10.1109/TPAMI.2019.2929257>
- [3] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733. <https://doi.org/10.1109/CVPR.2017.502>
- [4] Victor Costa, Leandro Tavares, Ruhan Conceição, Luciano Agostini, Brenda Santana, Tatiana Lebedeff, and Guilherme Corrêa. 2025. A Lightweight I3D-Based Approach for Real-Time Brazilian Sign Language Recognition. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web (Rio de Janeiro/RJ)*. SBC, Porto Alegre, RS, Brasil, 501–505. <https://doi.org/10.5753/webmedia.2025.16072>
- [5] David Vinicius da Silva, Valter Estevam, and David Menotti. 2024. Less is More: Concatenating Videos for Sign Language Translation from a Small Set of Signs. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 1–6. <https://doi.org/10.1109/SIBGRAPI62404.2024.10716311>
- [6] Rodrigo Zempulski Fanucchi, Arlindo Rodrigues Galvão Junior, Gabriel da Mata Marques, Lucas Brandão Rodrigues, Anderson da Silva Soares, and Telma Woerle Lima Soares. 2024. Fine-Tuning a Video Masked Autoencoder to Develop an Augmented Reality Application for Brazilian Sign Language Interpretation. In *Proceedings of the 26th Symposium on Virtual and Augmented Reality*. 275–278.
- [7] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
- [10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV] <https://arxiv.org/abs/1705.06950>
- [11] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, Vol. 2019. Long Beach, CA, 2.
- [12] Ministério da Educação do Brasil. 2023. Dia Nacional do Surdo destaca importância da inclusão e da acessibilidade. <https://portal.mec.gov.br/component/tags/tag/33784>. Estimates that about 5% of the Brazilian population has some form of hearing impairment..
- [13] Presidência da República do Brasil. 2002. Lei nº 10.436, de 24 de abril de 2002. http://www.planalto.gov.br/ccivil_03/leis/2002/10436.htm. Recognizes the Brazilian Sign Language (Libras) as a legal means of communication and expression..
- [14] Presidência da República do Brasil. 2005. Decreto nº 5.626, de 22 de dezembro de 2005. http://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2005/Decreto/D5626.htm. Regulates Lei nº 10.436/2002 and mandates the provision of accessibility for deaf people in education, media, and public services..
- [15] Ronice Müller de Quadros and Lodenir Becker Karnopp. 2007. Língua de sinais brasileira: estudos linguísticos. (2007).
- [16] Tamires Martins Rezende. 2021. *Reconhecimento automático de sinais da Libras: desenvolvimento da base de dados MINDS-Libras e modelos de redes convolucionais*. Ph.D. Dissertation. Universidade Federal de Minas Gerais.
- [17] Tamires Martins Rezende, Sílvia Grasiella Moreira Almeida, and Frederico Gadelha Guimarães. 2021. Development and validation of a Brazilian sign language database for human gesture recognition. *Neural Computing and Applications* 33, 16 (2021), 10449–10467.
- [18] Ailton José Rodrigues. 2021. *V-LIBRASIL: uma base de dados com sinais na língua brasileira de sinais (Libras)*. Master's thesis. Universidade Federal de Pernambuco, Recife.
- [19] Fida Mohammad Thoker and Juergen Gall. 2019. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 6–10.
- [20] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.