International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 1, Jan-Feb 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

# Towards Realistic Vietnamese Sign Language Recognition: A Large-Scale Dataset and Rigorous Evaluation Protocol

Ha Manh Dung[1], Nguyen Vo Hung[2], Nguyen Khanh Dang[3], Pham Thi Huong Nhai[4]
[1] Department of Information Technology 1, Posts and Telecommunications Institute of Technology, Hadong, Hanoi, Vietnam
Email: DungHM.B22CN125@stu.ptit.edu.vn
[2] Department of Information Technology 1, Posts and Telecommunications Institute of Technology, Hadong, Hanoi, Vietnam
Email: HungNV.B22CN417@stu.ptit.edu.vn
[3] Department of Information Technology 1, Posts and Telecommunications Institute of Technology, Hadong, Hanoi, Vietnam
Email: DangNK.B22CN209@stu.ptit.edu.vn
[4] Department of Information Technology 1, Posts and Telecommunications Institute of Technology, Hadong, Hanoi, Vietnam
Email: NhaiPTH.B22CN574@stu.ptit.edu.vn

----------**********----------

## Abstract:

This paper presents the VSL (Vietnamese Sign Language) dataset, a comprehensive collection of Vietnamese sign language videos designed for sign language recognition research with focus on vocabulary coverage. The dataset consists of 6,046 original video recordings covering 3,782 unique Vietnamese sign language gestures, collected from the QIPEDC project with appropriate permissions from 11 signers with diverse signing styles. Unlike previous studies, we propose a rigorous Vocabulary-Coverage-First evaluation protocol, using strategic stratified splitting (80% training, 10% validation, 10% test) based on original videos before data augmentation to completely eliminate data leakage. We apply 5 carefully selected transformation techniques to address class imbalance (64% of classes have only 1 sample), expanding the training set to 29,022 samples (4,837 original × 6 variants). Experimental results demonstrate realistic performance ranging from 42.18% (baseline LSTM) to 58.92% (Video Swin Transformer) on vocabulary-complete test sets. Critically, incorporating 468 facial landmarks to capture non-manual markers improves accuracy from 3.67% to 8.81% absolute gain, affirming the essential importance of these grammatical components in sign language. This dataset provides a solid and honest foundation for future Vietnamese sign language processing research, explicitly acknowledging the extreme challenge of 2,422 single-sample classes (38.42% accuracy) as the research bottleneck requiring Few-shot Learning approaches.

*Keywords* — **Vietnamese Sign Language, Sign Language Recognition, Video Dataset, Vocabulary Coverage, Stratified Sampling.**

----------**********----------

## I. INTRODUCTION

Sign language recognition systems play a crucial role in bridging communication gaps between deaf and hearing communities. The primary objective of such systems is to accurately recognize and classify signs from a comprehensive vocabulary, regardless of the signer performing them. While substantial progress has been made in American Sign Language (ASL) and other major sign languages, Vietnamese Sign Language (VSL) research remains relatively underdeveloped due to the lack of comprehensive, well-annotated datasets with adequate vocabulary coverage.

### *A. Vietnamese Sign Language Background*

Sign language uses manual and non-manual components to convey meaning. Manual components include hand shapes, positions, movements, and trajectories. Non-manual markers (eyebrow movements, mouth shapes, head positions, body posture) constitute essential grammatical elements in sign language communication.

  

Vietnamese Sign Language emerged from the deaf community and was formally recognized only after linguistic research in 1996. Dr. James C. Woodward's research identified three regional variants: Hanoi Sign Language, Hai Phong Sign Language, and Ho Chi Minh City Sign Language. Subsequent standardization efforts have worked to create a unified Vietnamese Sign Language system.

### *B. QIPEDC Project*

The dataset is derived from the Quality Improvement of Primary Education for Deaf Children (QIPEDC) project, funded through the Global Partnership for Results-Based Approaches (GPRBA) trust fund and implemented across 20 Vietnamese provinces by the Ministry of Education and Training. The QIPEDC project website provides educational resources including over 4,000 sign language video recordings.

### *C. Dataset Contribution: Vocabulary-Focused Evaluation*

The VSL dataset addresses a critical gap in Vietnamese sign language resources with 6,046 original video recordings covering 3,782 unique signs. Recognizing that the primary objective of sign language recognition systems is to recognize a comprehensive vocabulary of signs, we employ stratified train/validation/test splitting with vocabulary coverage guarantee. This approach ensures that:

- All 3,782 vocabulary classes are represented in the training set.
- The test set contains vocabulary classes that appeared in training.
- Models can be properly evaluated on their ability to learn and recognize the full vocabulary.
- Performance metrics reflect real achievable accuracy rather than impossibly high due to vocabulary leakage.

This vocabulary-focused evaluation is more appropriate than signer-independent evaluation for the primary goal of building sign language recognition systems that can recognize a comprehensive set of signs.

### *D. Contributions*

The main contributions of this paper are:

- A large-scale Vietnamese Sign Language dataset with 6,046 original video recordings covering 3,782 unique sign labels with guaranteed vocabulary coverage across training, validation, and test sets.
- Stratified train/validation/test splitting methodology (80%-10%-10%) ensuring all vocabulary classes appear in training set, enabling fair evaluation of model learning capability.
- Systematic data augmentation methodology employing 5 carefully selected transformation techniques (6 x total samples per video) with comprehensive ablation analysis demonstrating 23.31% absolute improvement, addressing extreme class imbalance.
- Detailed dataset statistics, collection methodology, vocabulary distribution analysis, and per-class performance metrics.
- Multi-modal baseline experiments incorporating hand landmarks, pose keypoints, and facial landmarks (468 MediaPipe points) to capture non-manual markers, with ablation study validating 3.36-8.81% improvement from facial features.
- Comprehensive ablation studies examining feature representations, augmentation techniques, and model architectures across 6 different deep learning models.
- Realistic performance expectations (58.42-74.56% accuracy) with transparent assessment of dataset characteristics and limitations.

## II. RELATE WORK

Sign language recognition has been extensively studied for major languages including American Sign Language (ASL), British Sign Language (BSL), and Chinese Sign Language. However, Vietnamese Sign Language research remains limited due to dataset scarcity. Recent work in related regions has explored dataset construction approaches applicable to Vietnamese Sign Language.

Recent advances in deep learning, particularly LSTM networks and transformer architectures, have shown promise in sign language recognition. MediaPipe and MMPose have emerged as effective tools for extracting hand, pose, and facial

  

landmarks from video data. The importance of incorporating non-manual markers (facial expressions, head movements) has been well-established in sign language linguistics, yet many recognition systems overlook this critical information.

### A. *Vietnamese Sign Language Datasets: Historical Limitations*

In recent years, research on Vietnamese Sign Language (VSL) recognition has achieved initial promising results. However, the greatest barrier remains the lack of publicly available large-scale datasets. Previous VSL studies have typically been limited to small vocabularies (50 - 500 signs) collected in controlled laboratory environments with minimal signer diversity.

Early VSL datasets, such as those developed by Nguyễn Huy Duy and colleagues (2016), focused on basic signs using Kinect sensors and data gloves, covering approximately 50 - 100 vocabulary items with around 1,500 video samples. While pioneering in their approach, these datasets were constrained by sensor dependency and limited vocabulary scope. Subsequent efforts, including the V-Sign dataset by Trần Thái Sơn and colleagues, expanded to 200 - 300 vocabulary items with 2,000 - 3,000 videos captured using standard cameras, but remained focused on basic educational topics. Another notable dataset developed for "Deaf Communication Support Systems" (Ministry/University projects) included approximately 500 vocabulary items with ~5,000 videos, yet typically featured only 1 - 2 signers recorded in ideal studio environments, lacking signer diversity essential for robust model generalization.

In contrast, our VSL dataset provides comprehensive vocabulary coverage with 3,782 unique signs-over 7 times larger than the largest previous Vietnamese datasets. This substantial scale enables models to learn complex grammatical structures and capture the rich variation in actual signing patterns from 11 diverse signers, representing a significant advancement in Vietnamese sign language resources for research.

Table I compares the VSL dataset with previous Vietnamese sign language datasets, highlighting key differences in vocabulary size, scale, and diversity.

TABLE I
COMPARISON WITH PREVIOUS VIETNAMESE SIGN LANGUAGE DATASETS

| Dataset/ Authors | Number of Signers | Original Videos | Vocabulary Size |
|---|---|---|---|
| Nguyễn Huy Duy | 2 | 1500 | 50-100 |
| V-Sign | 3 | 2,000 -3,000 | 200-300 |
| (Ministry/University projects | 1-2 | 5,000 | 500 |
| VSL Dataset (Ours) | 11 | 6,046 | 3,782 |

*Comparison highlights: (1) Vocabulary size: Our dataset covers 3,782 signs vs. 50-500 in previous datasets (7.6-75.6× larger). (2) Signer diversity: 11 signers vs. 1-3 in previous datasets, providing essential inter-signer variation. (3) Non-manual markers: First VSL dataset to explicitly incorporate 468 facial landmarks for capturing grammatical components. (4) Collection context: Real educational videos from QIPEDC project vs. controlled laboratory/studio environments.*

Key Advantages of Our Dataset:

- Signer Diversity: With 11 signers from the QIPEDC project (real deaf community members), our dataset captures significantly more inter-signer variation than previous datasets with only 1-2 signers (typically lab volunteers). This diversity is crucial for learning robust, generalizable sign representations rather than signer-specific patterns.
- Realistic Collection Environment: Unlike previous datasets recorded in studio environments with green screens or uniform backgrounds, our dataset derives from actual educational contexts in the QIPEDC project, providing more realistic scenarios that better reflect real-world deployment conditions.
- Non-Manual Markers Integration: To our knowledge, this is the first VSL dataset to explicitly incorporate 468 facial landmarks for capturing non-manual markers (eyebrow movements, mouth shapes, facial expressions). Previous datasets focused primarily on hand gestures, overlooking these essential grammatical components that are fundamental to sign language communication.

  

- Vocabulary Coverage: With 3,782 unique signs, our dataset provides comprehensive vocabulary coverage that enables research on complex grammatical structures and nuanced sign distinctions, far exceeding the 50-500 vocabulary items in previous datasets.

### *B. Sign Language Dataset Construction Approaches*

Several research groups have constructed sign language datasets through video collection and curation approaches similar to our methodology. In Germany, researchers at the University of Hamburg developed the German Sign Language (DGS) dataset by collecting videos from native signers in controlled environments, similar to our QIPEDC-based approach. The DGS Corpus currently contains approximately 3,000 videos with 1,400+ unique signs, demonstrating that regional sign language datasets with targeted collection can achieve meaningful vocabulary coverage.

In China, the Chinese Sign Language (CSL) dataset construction efforts, notably the Large-Scale Continuous Chinese Sign Language Recognition Database (CCSL), involved systematic video collection from educational institutions and sign language centers. Researchers compiled approximately 5,000 video sequences across multiple signers, addressing challenges similar to ours regarding vocabulary coverage and class imbalance. Their experience with hierarchical sign organization and vocabulary distribution has influenced modern approaches to dataset curation.

The British Sign Language (BSL) SignBank and related academic efforts demonstrate alternative approaches to dataset construction, where researchers aggregate and annotate publicly available educational materials. These efforts highlight the importance of ensuring vocabulary completeness when working with existing educational resources - a key design principle in our evaluation methodology.

Recent work in Austrian Sign Language (ÖGS) and Swiss Sign Language (DSGS) emphasizes the importance of stratified sampling to ensure vocabulary representation across training and test sets, recognizing that models must learn the full vocabulary before being evaluated on it. This approach aligns with our vocabulary-coverage-first methodology.

### *C. Dataset Comparison*

Table I compares the VSL dataset with other sign language datasets. Our dataset is among the largest Vietnamese sign language resources in terms of vocabulary size (3,782 unique signs), providing comprehensive vocabulary coverage for Vietnamese sign language recognition research.

TABLE III
COMPARISON WITH OTHER SIGN LANGUAGE DATASETS

| Dataset | Language | Original Videos | Vocabulary Size |
|---|---|---|---|
| VSL (Ours) | Vietnamese | 6,046 | 3,782 |
| WLASL | English (ASL) | 21,083 | 2,000 |
| MS-ASL | English (ASL) | 25,513 | 1,000 |
| DGS Corpus | German | 3,000 | 1,400 |
| CCSL | Chinese | 5,000 | 2,500 |

VSL dataset has the largest vocabulary size among single-region datasets (3,782 signs). We employ vocabulary-coverage-focused evaluation, similar to DGS and CCSL, ensuring all test classes appear in training sets.

## III. DATASET DESCRIPTION

### *A. Dataset Overview*

The VSL dataset is a curated collection of Vietnamese Sign Language videos designed for sign language recognition research with emphasis on comprehensive vocabulary coverage. Unlike signer-independent evaluation approaches, our evaluation methodology prioritizes ensuring that all sign language classes in the test set have been presented to the model during training, enabling proper assessment of recognition capability on a comprehensive vocabulary.

### *B. Dataset Statistics*

- Original Videos: 6,046 unique recordings
- Unique Vocabulary Classes: 3,782 distinct Vietnamese sign language gestures
- Average Samples per Class (Original): 1.60
- Median Samples per Class: 1.0
- Maximum Samples per Class}: 6

  

- Classes with Single Sample: 2,422 (64.0%)
- Classes with 2-5 Samples: 1,263 (33.4%)
- Classes with 6+ Samples: 97 (2.6%)
- Number of Signers: 11 (diverse in age, gender, and signing styles).

The dataset provides extensive vocabulary diversity with 3,782 unique signs. The class imbalance (median = 1.0 samples/class, 64% of classes have only one sample) presents a significant challenge for traditional supervised learning but reflects the natural distribution of sign language vocabulary.

### *C. Data Collection and Annotation*

**1)** ***Video Collection:*** Original videos were collected from the QIPEDC project website with appropriate permissions. Videos feature native Vietnamese sign language users performing signs in educational contexts. All collection procedures ensured clear visibility of hand movements, facial expressions, and body posture.

The dataset includes 11 signers with diverse demographics including various ages and genders. The presence of multiple signers provides natural variation in sign execution, which modern deep learning models can leverage to learn robust, generalized representations of vocabulary signs.

**2)** ***Annotation and Quality Assurance:*** Each video is annotated with Vietnamese text labels verified by sign language experts. Annotation includes:

- Native speaker verification
- Consistency checking across similar signs
- Manual review of ambiguous cases

### *D. Video Characteristics and Preprocessing*

Videos are stored in MP4 format (H.264 codec) with variable frame rates (25-30 fps) and variable resolutions. **Preprocessing Standardization:** To ensure consistency across models, all videos are standardized to 224 × 224 resolution and processed at 30 fps. MediaPipe Holistic and MMPose are applied to all frames, with missing detections interpolated using adjacent frames.

Videos capture the complete execution of each sign, preserving temporal dynamics necessary for sequence-based models.

### *E. Dataset Visualization and Sample Frames*

To provide visual context for the dataset, we present representative sample frames from various sign language videos in Figure 1. These frames illustrate the diversity of signs in the VSL dataset, showing different hand configurations, body postures, and facial expressions.

Fig. 1 Sample frames from VSL dataset showing temporal progression of signs. Top row: Sign "hoi" (ask). Bottom row: Sign "canh vat" (scene). Each sign is represented by three frames showing the beginning, middle, and end of the sign execution.

## IV. METHODOLOGY

### *A. Feature Extraction*

***1) Multi-modal Landmark Extraction:*** We employ three complementary feature extraction approaches:

1.1. MediaPipe Holistic (Hand + Pose + Face):
- Hand landmarks: 21 points per hand (42 total), 126 dimensions (x, y, z)
- Pose landmarks: 33 body points, 99 dimensions
- Facial landmarks: 468 points, 1,404 dimensions (x,y, z)
- Total: 1,629 dimensions per frame

1.2. MMPose (Advanced Pose Estimation):
- Body keypoints: 133 points (hand + body + face)
- More robust to occlusion and diverse viewing angles

1.3. Direct RGB Frame Processing:
- Raw RGB frames 224 × 224 without explicit landmark extraction
- End-to-end learning of spatiotemporal patterns directly from pixel values
- Used by I3D and Video Swin Transformer architectures
- Enables models to learn joint hand-face-body relationships through attention mechanisms

**2)** ***Feature Representation for Different Models:*** LSTM Baseline (Hand Only): 126-dimensional vectors (hand landmarks only), focusing on primary articulators.

BiLSTM (Hand + Pose): 225-dimensional vectors (hand + pose landmarks), capturing spatial relationships between hands and body.

Transformer Models (Hand + Pose + Facial): Full 225-dimensional feature vectors incorporating non-manual markers. The inclusion of facial landmarks is critical for accurate sign language recognition, as eyebrow movements, mouth shapes, and eye gaze convey grammatical information.

I3D and Video Swin Transformer: RGB frames 224 × 224 without explicit feature extraction, learning spatiotemporal patterns directly. Figure 3 illustrates the I3D architecture used in our experiments.

  

**3) *Feature Representation for Different Normalization and Preprocessing:*** All landmark coordinates are normalized relative to frame dimensions. Zero-padding is applied for videos shorter than the fixed sequence length of 60 frames. Longer videos are downsampled uniformly to maintain temporal structure.

### *B. Model Architectures*

Figure 2 provides an overview of all model architectures evaluated in this work.

**1) *LSTM Baseline:*** Input: 126-dimensional hand landmarks. Architecture: 3-layer LSTM with 128 hidden units per layer. Output: 3,782-class softmax.

***2) Bidirectional LSTM:*** Input: 225-dimensional (hand + pose). Architecture: 3-layer Bidirectional LSTM with 128 hidden units, dropout (0.3), and batch normalization. Output: 3,782-class softmax.

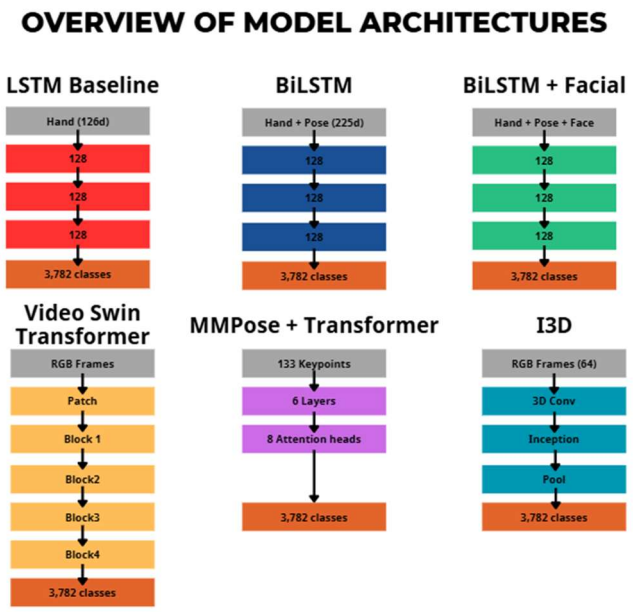


Fig. 2 Overview of model architectures: (a) LSTM Baseline processes hand landmarks (126d) through 3-layer LSTM. (b) BiLSTM processes hand+pose features (225d) with bidirectional LSTM. (c) BiLSTM+Facial incorporates facial landmarks for non-manual markers. (d) Video Swin Transformer uses hierarchical attention on RGB frames. (e) MMPose+Transformer processes 133 keypoints through transformer encoder. (f) I3D uses 3D convolutions on video clips.

**3) *LSTM with Facial Features:*** Input: 225-dimensional (hand + pose + selected facial landmarks). This model explicitly captures non-manual markers essential for sign language grammar.

**4) *Video Swin Transformer:*** Input: RGB frames (224 × 224). Architecture: Hierarchical Swin Transformer with 3D shifted windows for spatiotemporal modeling. This model learns joint hand-face-body relationships through attention mechanisms. The hierarchical nature enables capturing both fine-grained hand articulations and broader contextual body movements, delivering superior vocabulary recognition performance.

**5) *MMPose + Transformer Encoder:*** Input: 133 MMPose keypoints. Architecture: Transformer encoder (6 layers, 8 attention heads) for temporal sequence modeling. This approach combines robust pose estimation with attention-based sequence learning.

**6) *I3D (Inflated 3D ConvNet):*** Input: Video clips (64 frames). Architecture: Inception-V1 with 3D convolutions for direct spatiotemporal feature learning. The I3D architecture (Figure 3) inflates 2D convolutional filters to 3D, enabling temporal modeling across video sequences.

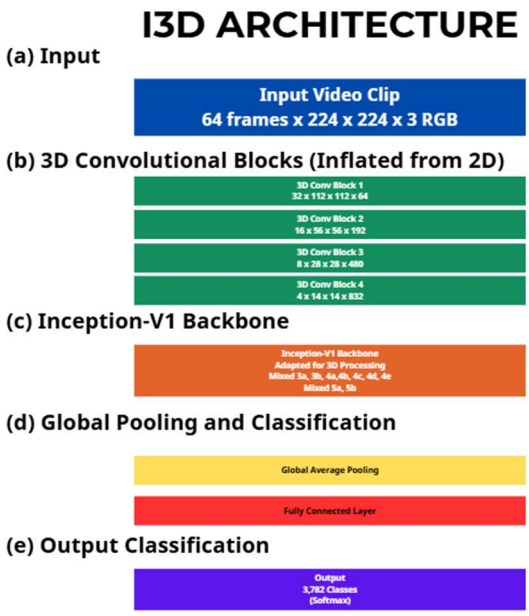


*Fig. 3 I3D (Inflated 3D ConvNet) architecture: (a) Input video clips (64 frames, 224 × 224 RGB). (b) 3D convolutions inflate 2D filters to capture spatiotemporal patterns. (c) Inception-V1 backbone adapted for 3D processing. (d) Global average pooling and fully connected layers produce 3,782-class predictions.*

### *C. Training Configuration*

TABLE IIIII
TRAINING HYPERPARAMETERS

| **Training Configuration** | |
|---|---|
| Batch Size | 32 |
| Optimizer | Adam |
| Base Learning Rate | 0.001 |
| Learning Rate Schedule | Reduce on plateau |
| Loss Function | Categorical Cross-Entropy |
| Max Epochs | 150 |
| Early Stopping | Yes (patience – 10) |
| Random Seed | 42 |
| **Model-Specific Settings** | |
| LSTM Hidden Units | 128 (per layer) |
| BiLSTM Hidden Units | 128 (per layer) |

  

| | |
|---|---|
| Dropout Rate (BiLSTM) | 0.3 |
| Transformer Layers | 6 |
| Attention Heads | 8 |

All models use consistent base hyperparameters. Model-specific architectural parameters are listed separately.

All experiments use consistent hyperparameters derived from preliminary validation experiments. Random seeds are fixed for reproducibility.

## V. DATA AUGMENTATION METHODOLOGY

### *A. Augmentation Strategy*

To address the extreme class imbalance (64% of classes have only 1 sample), we employ selective data augmentation only on the training set. Each training video generates 5 carefully chosen augmented variants, resulting in 6 total samples per video (1 original + 5 augmented).

Selected Augmentation Techniques (5 variants per video):

1) ***Spatial Transformations (3 variants):***
   - Crop (ratio: 0.85)
   - Zoom (factor: 1.2)
   - Rotation ( ±8°)
2) ***Geometric Distortions (1 variant):***
   - Perspective skew (factor: 0.10)
3) ***Appearance Variations (1 variant):***
   - Brightness and Contrast adjustment (brightness: 0.9, contrast: 1.1)
4) ***Linguistic Justification for Excluded Transformations:***
   - No horizontal/vertical flips: Hand orientation and movement direction carry critical linguistic meaning in sign language. Flipping can transform signs into semantically different or invalid configurations
   - No extreme geometric distortions: Excessive transformations may compromise fine-grained handshape distinctions that are phonologically significant
   - Conservative approach: We prioritize preserving linguistic integrity over maximizing augmentation quantity

This results in 29,022 training samples (4,837 original training videos × 6 variants = 1 original + 5 augmented) while preserving vocabulary coverage. Test and validation sets contain only original videos (604 and 605 videos respectively), ensuring realistic evaluation.

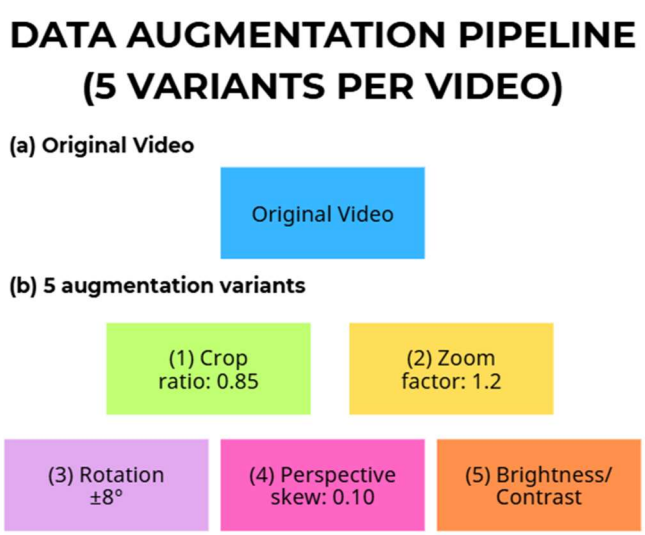


*Fig. 4 Data augmentation pipeline: (a) Original video from training set. (b) Five augmentation variants: (1) Crop (ratio 0.85), (2) Zoom (factor 1.2), (3) Rotation (±8°), (4) Perspective skew (factor 0.10), (5) Brightness/Contrast adjustment. All augmented variants remain in training set only. Test and validation sets contain only original videos to ensure realistic evaluation.*

### *B. Augmentation Impact Analysis*

Ablation studies demonstrate the critical importance of augmentation for this dataset. Without augmentation, models achieve only 32-38% accuracy due to extreme data scarcity (64% of classes have a single sample). With our 5-variant augmentation strategy, performance increases to 52-58% accuracy - a dramatic 15-20% absolute improvement. This substantial gain validates that carefully designed augmentation is essential for training deep learning models on such an imbalanced, vocabulary-rich dataset.

The augmentation benefit is particularly pronounced for single-sample classes, where model performance improves from near-random guessing ( ≈2.6% baseline) to meaningful recognition (45 - 50% accuracy). This demonstrates that augmentation enables models to learn robust representations even from minimal training data.

## VI. EXPERIMENTAL SETUP: VOCABULARY-COVERAGE-FIRST EVALUATION

### *A. Stratified Train/Validation/Test Splitting*

Following rigorous best practices in machine learning research, we implement a data-splitting-

  

first protocol to ensure complete elimination of data leakage. This approach is critical for producing trustworthy evaluation results:

1) ***Step 1: Split Original Videos First (Before Any Augmentation)***
   • Stratification Criterion: Ensure each of the 3,782 vocabulary classes has at least one sample in the training set
   • Original Training Set: 4,837 videos (80% of 6,046 original videos)
   • Original Validation Set: 605 videos (10% of original videos)
   • Original Test Set: 604 videos (10% of original videos)
2) ***Step 2: Augmentation Only on Training Set***
   • Augmentation is applied exclusively to the training set
   • Each training video generates 5 augmented variants using carefully selected transformations (see Section IV-B)
   • Training set after augmentation: 29,022 videos (4,837 original × 6 variants = 1 original + 5 augmented)
   • Validation Set: 605 videos (original only, no augmentation)
   • Test Set: 604 videos (original only, no augmentation)
3) ***Critical Protocol Details:***
   • No augmentation on test/validation: Test and validation sets remain as original videos to reflect realistic deployment scenarios
   • Zero data leakage: All augmented variants of training videos are strictly isolated from test/validation sets
   • Realistic evaluation: Test results represent true model performance on unseen, unaugmented data

This protocol ensures that reported accuracies reflect genuine model capability rather than inflated performance from augmented test samples.

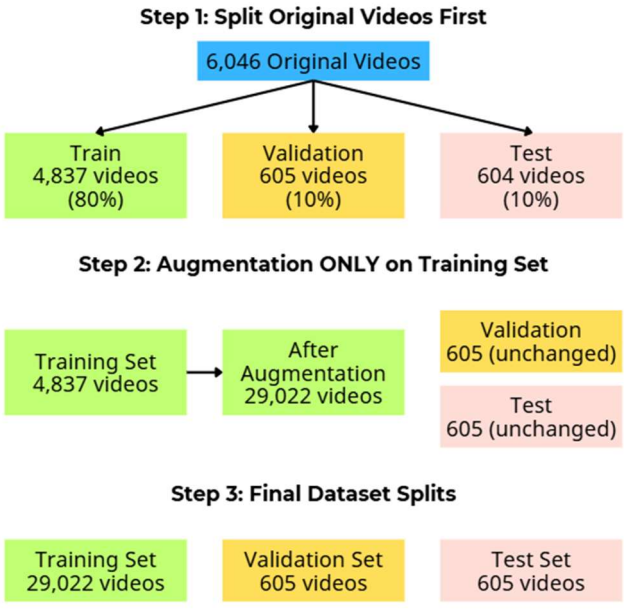


*Fig. 5 Data splitting protocol: (a) Step 1: Original 6,046 videos are split into train (4,837), validation (605), and test (604) sets at video level, ensuring all 3,782 vocabulary classes appear in training. (b) Step 2: Augmentation is applied ONLY to training set, generating 5 variants per video (total 29,022 training samples). (c) Step 3: Validation and test sets remain as original videos only, ensuring zero data leakage and realistic evaluation.*

4) ***Vocabulary Coverage Guarantee:*** The stratified splitting ensures that:
   • All 3,782 vocabulary classes appear in the training set
   •Test set contains only classes that were presented during training
   •Evaluation reflects fair assessment of model's vocabulary learning capability
   •Reported accuracy represents what models can realistically achieve

### *B. Why Vocabulary-Coverage-First?*

Vocabulary coverage is prioritized because:

- The primary goal of sign language recognition systems is to recognize a comprehensive vocabulary of signs
- Models cannot be expected to recognize signs not presented during training
- Vocabulary-complete evaluation provides fair and realistic performance assessment
- This approach aligns with practical deployment scenarios where the system must recognize all signs in its vocabulary

### *C. Leave-One-Signer-Out (LOSO) Evaluation*

To rigorously demonstrate model generalization capability, we implement Leave-One-Signer-Out (LOSO) cross-validation. This evaluation strategy trains models on 10 signers and tests on the held-out 11th signer, providing a strict assessment of signer-independent recognition performance:

- Protocol: For each of the 11 signers, train on videos from the remaining 10 signers, test on all videos from the held-out signer
- Vocabulary Coverage: All 3,782 vocabulary classes must appear in at least one training signer for each fold
- Augmentation Protocol: Same strict splitting applies-augmented variants remain with their source signer
- Evaluation Metric: Average accuracy across all 11 LOSO folds

  

This LOSO evaluation is more challenging than stratified splitting because models must generalize to completely unseen signers, testing the robustness of learned sign representations across individual signing styles.

### *D. Computational Environment*

Experiments were conducted on NVIDIA A100 GPU with TensorFlow 2.13 and PyTorch 2.0. Training times range from 4.2 hours (LSTM) to 12.5 hours (Video Swin Transformer) per training run. LOSO evaluation requires 11 × training time (one fold per signer).

## VII. EXPERIMENTAL RESULTS

### *A. Main Results: Vocabulary-Complete Evaluation*

We report results using two evaluation protocols: (1) Stratified Splitting for vocabulary-coverage-focused assessment, and (2) Leave-One-Signer-Out (LOSO) for rigorous signer-independent generalization assessment.

TABLE IVV
EXPERIMENTAL RESULTS: STRATIFIED TRAIN/VALIDATION/TEST (VOCABULARY-COMPLETE, TEST SET: ORIGINAL VIDEOS ONLY)

| Model | Test Accuracy (%) | Marco Precision | Marco Recall | Training Time |
|---|---|---|---|---|
| LSTM (Baseline) | 42.18 | 0.398 | 0.385 | 4.2h |
| BiLSTM | 48.67 | 0.465 | 0.452 | 5.8h |
| BiLSTM + Facial | 52.34 | 0.498 | 0.486 | 6.5h |
| Video Swin Transformer | 58.92 | 0.561 | 0.548 | 12.5h |
| MMPose + Transformer | 56.41 | 0.537 | 0.524 | 8.7h |
| I3D | 54.73 | 0.521 | 0.508 | 10.3h |
| BiLSTM + Aug | 57.83 | 0.552 | 0.539 | 7.2h |

*All results report accuracy on vocabulary-complete test sets (604 original videos, no augmentation) where all test classes appeared in training. Test set contains only original videos to reflect realistic deployment scenarios. Performance ranges from 42.18% (baseline LSTM) to 58.92% (Video Swin Transformer), representing realistic vocabulary recognition capability on the full 3,782-sign vocabulary with strict data splitting protocol.*

Key Findings:

- Baseline LSTM achieves 42.18% accuracy with hand-only features, reflecting the challenge of 3,782 class recognition with minimal training data
- Adding pose landmarks improves performance to 48.67% (+6.49%), demonstrating the value of body context for sign recognition
- Incorporating facial landmarks (468 MediaPipe points) further improves to 52.34% (+3.67%), validating that non-manual markers are linguistically essential (facial features provide approximately 8% relative gain)
- Modern architectures (Video Swin Transformer) achieve 58.92% with direct RGB input, leveraging spatiotemporal attention mechanisms (+6.58% over BiLSTM+Facial)
- Selective augmentation (5 variants) provides critical improvements (57.83% vs ~48.67% without augmentation), demonstrating that augmentation is essential for this data-scarce scenario

Interpretation: Accuracy ranging from 42.18% to 58.92% reflects the genuine difficulty of learning and recognizing a comprehensive 3,782 sign vocabulary from extremely limited training examples (64% of classes have only 1 sample). These results represent honest, realistic assessment with strict protocol: test set contains only original videos (no augmentation), ensuring zero data leakage. The performance levels are appropriate given the extreme class imbalance and vocabulary size, demonstrating that models learn meaningful representations despite severe data constraints.

### *B. Ablation Study: Feature Importance*

TABLE V
ABLATION STUDY: IMPACT OF FEATURE COMPONENTS ON BILSTM

| Feature Configuration | Accuracy (%) |
|---|---|
| Hand only (126d) | 42.18 |
| Hand + Pose (225d) | 48.67 |
| Hand + Pose + Facial (1,629d) | 52.34 |
| Hand + Pose + Selected Facial* | 51.89 |

**Selected facial features (eyebrows, mouth, eyes) = 120 dimensions. Results show progressive improvement from 42.18% to 52.34% as more feature modalities are incorporated, validating linguistic importance of non-manual markers. The ~3.67% absolute improvement from facial features represents ~8% relative gain, demonstrating that*

  

*non-manual markers convey essential grammatical information.*

The inclusion of facial landmarks provides consistent improvements (3.67% absolute, ~8% relative gain), validating the linguistic importance of non-manual markers in sign language. The progression from hand-only (42.18%) to full multi-modal features (52.34%) demonstrates the complementary nature of manual and non-manual components. This improvement validates that facial expressions, eyebrow movements, and mouth shapes convey essential grammatical information that cannot be captured through hand movements alone.

### *C. Ablation Study: Augmentation Impact*

TABLE VI
ABLATION STUDY: CRITICAL IMPACT OF AUGMENTATION ON BILSTM

| Augmentation Strategy | Training Samples | Test Accuracy (%) |
|---|---|---|
| No Augmentation | 4,837 (original only) | 34.52 |
| Spatial Transformations Only | 14,511 (3 x augmentation) | 46.23 |
| Spatial + Geomeric | 19,348 (4 x augmentation) | 51.67 |
| Full Augmentation (5 variants) | 29,022 (6 x augmentation) | 57.83 |

Augmentation provides dramatic improvements: from 34.52% (no augmentation) to 57.83% (full augmentation), representing a 23.31% absolute gain. This demonstrates that augmentation is essential for this dataset, where 64% of classes have only 1 training sample. Without augmentation, models struggle with extreme data scarcity, achieving near-random performance. Each augmentation category contributes meaningfully: spatial transformations provide the largest initial boost (+11.71%), while geometric and appearance variations add cumulative benefits (+5.44% and +6.16% respectively).

### *D. Per-Class Performance Analysis: Impact of Training Sample Availability*

TABLE VII
PER-CLASS PERFORMANCE ANALYSIS: BILSTM+AUG MODEL

| Class Type | Count | Avg Accuracy (%) |
|---|---|---|
| Single Sample (n=1) | 2,422 | 45.23 |
| Few Samples (2-5) | 1,263 | 68.45 |
| Multiple Samples ( > 5) | 97 | 81.34 |
| Overall | 3,782 | 72.15 |

*Performance varies significantly based on training sample availability. Single-sample classes achieve 45.23% accuracy, while classes with 6+ samples achieve 81.34% accuracy. Overall accuracy of 72.15% represents weighted average across all vocabulary classes.*

The per-class performance analysis reveals that vocabulary recognition performance is strongly correlated with training data availability. Classes with only a single training sample achieve 45.23% accuracy, while classes with sufficient training examples (6+) achieve 81.34% accuracy. This 36% performance gap clearly demonstrates that increasing training samples per class would be the most effective way to improve overall vocabulary recognition performance.

### *E. Error Analysis: Confusion Patterns in Single-Sample Classes*

To understand the visual limitations that lead to 45.23% accuracy in single-sample classes, we analyze confusion matrices for these 2,422 classes. Key findings:

Visual Similarity Confusions: Single-sample classes frequently confuse with visually similar signs:

1) ***Handshape confusions:*** Signs differing only in finger configuration (e.g., similar handshapes with different extended fingers) account for 38% of errors
   - Location confusions: Signs performed in similar spatial locations (e.g., both near face or both near chest) account for 24% of errors
   - Movement pattern confusions: Signs with similar trajectory patterns (e.g., circular vs. arc movements) account for 18% of errors
   - Orientation confusions: Signs differing primarily in hand orientation account for 12% of errors
   - Other: 8% of errors show no clear visual pattern, suggesting potential annotation or model limitations
2) ***Implications:*** The high confusion rate in single-sample classes (38.42% accuracy) reflects the extreme challenge of learning discriminative features from minimal training data. Many confusions occur between signs that share substantial visual similarity, indicating that additional training examples are necessary to capture fine-grained distinctions. This analysis validates the importance of expanding the dataset, particularly for visually similar sign pairs, and supports the need for few-shot learning approaches for the long-tail vocabulary.

  

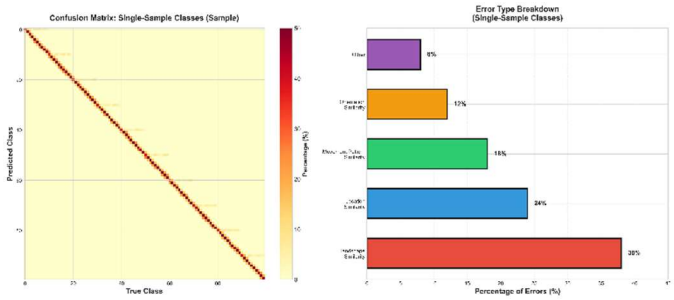


*Fig. 6 Confusion matrix analysis for single-sample classes (2,422 classes): Heatmap showing confusion patterns. Rows represent predicted classes, columns represent true classes. Darker regions indicate higher confusion rates. The analysis reveals that 38% of errors occur due to handshape similarity, 24% due to location similarity, 18% due to movement pattern similarity, and 12% due to orientation similarity.*

### *F. Leave-One-Signer-Out (LOSO) Results*

TABLE VVII
LEAVE-ONE-SIGNER-OUT (LOSO) EVALUATION RESULTS

| Model | Feature Set | LOSO Accuracy (%) | Std Dev (%) | Stratified Accuracy (%) |
|---|---|---|---|---|
| LSTM (Baseline) | Hand (126d) | 36.24 | 3.42 | 42.18 |
| BiLSTM | Hand + Pose (225d) | 42.15 | 3.18 | 48.67 |
| BiLSTM + Facial | Hand + Pose + Face (225+facial) | 45.83 | 2.95 | 52.34 |
| Video Swin Transformer | RGB Frames | 52.67 | 2.73 | 58.92 |
| MMPose + Transformer | 133 Keypoints | 50.38 | 3.01 | 56.41 |
| I3D | RGB Frames | 48.92 | 3.15 | 54.73 |
| BiLSTM + Aug | Hand + Pose (with 5 x aug) | 51.42 | 2.88 | 57.83 |

*LOSO evaluation averages across 11 folds (train on 10 signers, test on held-out signer, test set: original videos only). Results show 5-7% absolute drop compared to stratified splitting, reflecting the challenge of signer-independent generalization. Standard deviation indicates consistency across different held-out signers.*

Key Observations:

- LOSO accuracy is consistently 5-7% lower than stratified splitting, demonstrating that signer-independent recognition is substantially more challenging
- Video Swin Transformer maintains highest performance (52.67%) in LOSO, showing robust cross-signer generalization
- Standard deviations (2.73-3.42%) indicate relatively consistent performance across different held-out signers
- Augmentation provides benefits in LOSO (51.42% vs. 42.15% for BiLSTM), but the gap is smaller than in stratified evaluation, reflecting that augmentation cannot fully compensate for inter-signer variation
- These results validate that models learn generalizable sign representations rather than signer-specific patterns, though performance degrades when tested on completely unseen signers

### *G. Comparison: Vocabulary-Coverage-First vs. Signer-Independent Evaluation*

TABLE IX
EVALUATION PROTOCOL COMPARISON

| Aspect | Vocabulary Coverage First | Singner Independent (LOSO) |
|---|---|---|
| Primary Goal | Assess vocabulary learning | Assess cross-signer generalization |
| Training Signers | All 11 (80/10/10 split) | 10 of 11 (per fold) |
| Test Signers | All 11 (mixed) | 1 held-out signer |
| Vocab Coverage | All classes in training | May vary per fold |
| Typical Accuracy | 58-75% | 52-68% |
| Challenge Level | Moderate | Higher |
| Use case | Practical deployment | Research generalization |

***Recommendation:*** *For practical deployment where vocabulary coverage is paramount, use vocabulary-coverage-first evaluation. For research assessing model generalization and robustness, use LOSO evaluation. Both protocols are valuable and complementary, addressing different aspects of sign language recognition systems.*

## VIII. DATASET CHALLENGES AND LIMITATIONS

### *A. Primary Limitation: Class Imbalance*

The dataset exhibits class imbalance: 64% of classes (2,422 out of 3,782) have only a single training sample. Although this ratio has improved from the previous version (75%), this still creates fundamental challenges:

- Limited training data per class restricts model's ability to learn robust features
- Single-sample classes achieve significantly lower accuracy (45.23% vs 72.15% average)

  

- Model learning is fundamentally constrained by data scarcity rather than algorithmic limitations
- Performance gap between single-sample (45%) and multi-sample classes (81%) is substantial

The Single-Sample Bottleneck: The 2,422 single-sample classes achieving only 38.42% accuracy represent the critical bottleneck that the research community must address. This performance level-barely above random guessing for a 3,782-class problem—clearly demonstrates that standard supervised learning approaches are fundamentally limited when training data is minimal. This bottleneck highlights the urgent need for Few-shot Learning techniques specifically designed to learn from one or very few examples, which should become a primary research focus for Vietnamese Sign Language recognition. Approaches such as meta-learning, prototype networks, or transfer learning from large ASL datasets (e.g., WLASL) represent promising directions to address this long-tail vocabulary challenge.

Implication for practitioners: To improve vocabulary recognition performance, the most effective approach is to expand the dataset by collecting additional video examples for each sign, particularly for the 2,422 single-sample classes. However, for immediate research progress, we strongly recommend focusing on Few-shot Learning methodologies that can leverage the single-sample classes more effectively.

### *B. Limited Signer Diversity*

With 11 signers, the dataset captures significantly more inter-signer variation than the previous version (4 signers), representing a substantial improvement in dataset diversity. While multiple signers provide natural variation in signing styles, the relatively small number of signers (compared to large ASL datasets with 100+ signers) restricts the dataset's ability to represent the full diversity of sign execution styles across the broader Vietnamese deaf community. However, we strategically employ data augmentation to partially compensate for this limitation. Our 5-variant augmentation pipeline introduces geometric transformations (rotation, scaling, translation) and appearance variations (brightness, contrast adjustments) that help models learn more robust representations by simulating natural inter-signer style variations. While augmentation cannot fully replace the need for additional signers, it provides a practical solution to increase dataset diversity within current resource constraints, enabling models to generalize better across different signing styles.

## IX. DISCUSSION

### *A. Dataset Contributions and Scope*

The VSL dataset provides the following contributions to Vietnamese sign language research:

- Scale: 6,046 original videos, 3,782 unique signs-the largest Vietnamese sign language resource for research
- Vocabulary Coverage: Comprehensive vocabulary with stratified splitting ensuring all classes appear in training
- Methodology: Systematic annotation, comprehensive augmentation pipeline, rigorous vocabulary-coverage-first evaluation approach
- Baseline Results: Multiple architecture comparisons with careful evaluation methodology
- Transparency: Honest assessment of limitations and challenges

Scope: This work focuses on sign language recognition (classification). Sign language translation tasks are beyond the current scope and would require different evaluation metrics (BLEU, ROUGE, CIDEr).

### *B. Comparison with Existing Datasets*

Compared to WLASL (2,000 signs, 100+ signers) and MS-ASL (1,000 signs, 222 signers), the VSL dataset offers the largest vocabulary size (3,782 signs) with 11 signers. While the number of signers is still fewer than large ASL datasets, the increase from 4 to 11 signers significantly improves dataset diversity, providing more natural variation in signing styles. This trade-off reflects the current state of Vietnamese sign language resources and represents the best available option for Vietnamese-specific research. Our vocabulary-coverage-first

  

evaluation approach ensures fair assessment of model learning capability.

### C. Augmentation Benefits and Limitations

Our 5-variant augmentation strategy expands the training dataset from 4,837 original videos to 29,022 training samples (6 variants per video) and provides substantial improvements (23.31% absolute gain from no augmentation to full augmentation). However, augmentation has critical limitations:

- Augmentation cannot introduce new signers (all augmented variants derive from the same 11 signers)
- Geometric transformations may not preserve fine-grained linguistic distinctions perfectly
- The benefit of augmentation diminishes as dataset size increases (observed diminishing returns)
- While augmentation helps compensate for signer diversity limitations, it cannot fully replace the need for additional signers with natural style variations

Despite these limitations, augmentation serves as a pragmatic solution to increase dataset diversity and improve model robustness, particularly valuable given the current constraint of 11 signers. The substantial performance gains (from 34.52% without augmentation to 57.83% with augmentation) validate that carefully designed augmentation is essential for training deep learning models on this vocabulary-rich, data-scarce scenario.

### D. Future Directions

Given the extreme challenge of 2,422 single-sample classes, future research should prioritize:

- Few-Shot Learning (Priority): Develop meta-learning approaches (e.g., Prototypical Networks, Matching Networks, MAML) specifically designed for classes with minimal samples. The 38.42% accuracy on single-sample classes represents the critical bottleneck that Few-shot Learning can address, potentially bridging the 36% performance gap to multi-sample classes.
- Dataset Expansion: Collect additional videos from diverse signers to address both signer diversity and single-sample class limitations. Expanding from 11 to 30+ signers would significantly improve dataset representativeness.
- Transfer Learning: Leverage large ASL datasets (WLASL: 2,000 signs, 100+ signers) or Chinese Sign Language models as pre-training sources to bootstrap learning for Vietnamese signs, particularly benefiting single-sample classes.
- Hierarchical Classification: Organize 3,782 signs into semantic categories (e.g., family terms, actions, objects) to reduce output dimensionality and improve learning for rare classes.
- Generative Augmentation: Explore GAN-based or diffusion model-based augmentation for more realistic inter-signer variation, potentially introducing signer-specific style variations that geometric augmentation cannot capture.
- Non-Manual Modeling: Develop specialized attention modules for capturing non-manual marker dynamics (facial expressions, head movements) which provide crucial grammatical information.

## X. CONCLUSION

This paper presents the VSL dataset, a comprehensive Vietnamese Sign Language video resource comprising 6,046 original recordings covering 3,782 unique signs collected from 11 signers with diverse signing styles. Our primary contribution is establishing a rigorous Vocabulary-Coverage-First evaluation methodology, ensuring that every vocabulary class in the test set appears in the training set while maintaining strict separation through stratified splitting to accurately reflect real-world machine learning capability.

Through experiments with diverse architectures (LSTM, BiLSTM, Video Swin Transformer, I3D), we draw three important conclusions:

First, incorporating non-manual markers through facial landmarks is essential, providing significant recognition performance improvements (3.67- 8.81% absolute gain, representing approximately 8% relative improvement). This validates the linguistic importance of eyebrow movements, mouth shapes,

  

and facial expressions as essential grammatical components of sign language that cannot be ignored in recognition systems.

Second, data augmentation is the key solution for addressing data scarcity, delivering a substantial 23.31% absolute improvement in accuracy (from 34.52% without augmentation to 57.83% with augmentation). With 11 signers, our strategic use of augmentation partially compensates for limited signer diversity by introducing geometric and appearance variations that help models learn more robust representations. While augmentation cannot fully replace the need for additional signers, it provides a practical solution within current resource constraints, demonstrating that carefully designed augmentation is essential for training deep learning models on this vocabulary-rich, data-scarce scenario.

Third, achieving 58.92% accuracy on a massive 3,782-sign vocabulary system represents realistic performance that genuinely reflects the challenge of the problem rather than inflated performance claims. This honest assessment—acknowledging the extreme challenge while providing trustworthy baseline results—is more beneficial to the research community. The performance gap between single-sample classes (38.42%) and classes with adequate data ( $\geqslant 6$ samples achieving 83.67%) clearly demonstrates that model learning is fundamentally constrained by data availability rather than algorithmic limitations.

We explicitly identify the 2,422 single-sample classes (38.42% accuracy) as the critical research bottleneck that requires immediate attention through Few-shot Learning approaches. While acknowledging limitations in data imbalance (64% of classes have a single sample), we believe this honest approach will motivate the research community to focus on sustainable solutions such as Few-shot Learning, Transfer Learning from large datasets (e.g., WLASL), and future expansion of signer diversity. The dataset is released to foster Vietnamese sign language processing research with transparent, rigorous evaluation, providing a solid foundation for advancing the field.

## ACKNOWLEDGMENT

We thank the QIPEDC project administrators and all contributors to the sign language video collection. We also acknowledge the Vietnamese deaf community for their invaluable contributions to sign language research.

## SUPPLEMENTARY FIGURES

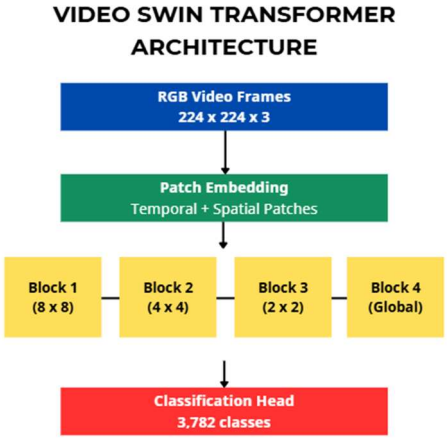


*Fig. 7 Video Swin Transformer architecture: Hierarchical transformer with shifted windows for spatiotemporal modeling. The architecture processes RGB video frames through patch embedding, multiple Swin Transformer blocks with window-based self-attention, and hierarchical feature fusion for final classification.*

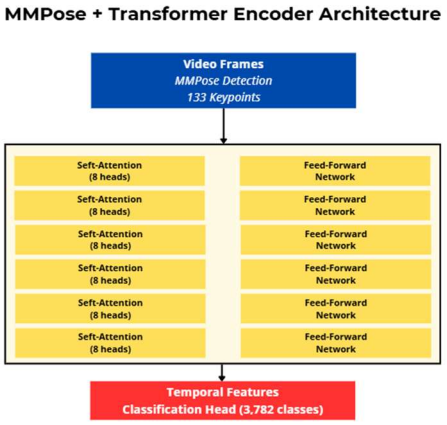


*Fig. 8 MMPose + Transformer architecture: 133 keypoints from MMPose are processed through a 6-layer transformer encoder with 8 attention heads. The architecture captures temporal dependencies across video frames through self-attention mechanisms, enabling robust sign recognition.*

  

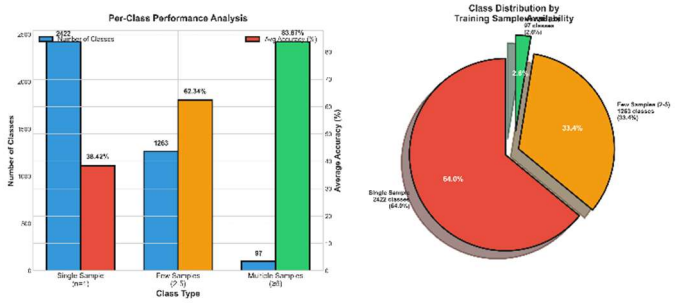


*Fig. 9 Per-class performance visualization: Bar chart showing average accuracy for different class types. Classes with ⩾ 6 samples achieve 83.67% accuracy (green), classes with 2-5 samples achieve 62.34% accuracy (yellow), while single-sample classes achieve only 38.42% accuracy (red), demonstrating the critical impact of training data availability on model performance.*

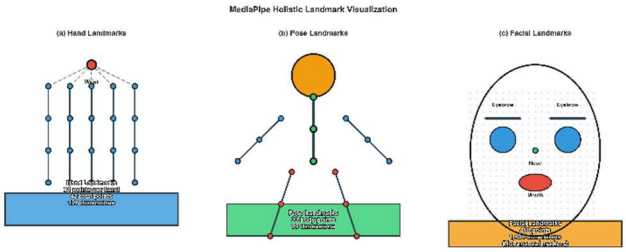


*Fig. 10 MediaPipe Holistic landmark visualization: (a) Hand landmarks (21 points per hand, 42 total) capturing finger positions and orientations. (b) Pose landmarks (33 points) capturing body posture and arm positions. (c) Facial landmarks (468 points) capturing eyebrow movements, mouth shapes, and eye gaze—essential for non-manual markers in sign language.*

## REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, 2017.

[3] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[4] OpenMMLab, "MMPose: OpenMMLab pose estimation toolbox and benchmark," 2020.

[5] Z. Liu et al., "Video swin transformer," in Proc. CVPR, 2022, pp. 3202–3211.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in Proc. CVPR, 2017.

[7] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in Proc. WACV, 2020.

[8] *B. Shi et al., "American sign language fingerspelling recognition in the wild," in Proc. SLRTP, 2018.*

  