# ONESIGN: UNIFYING SIGN LANGUAGE UNDERSTANDING TASKS WITH ONE MODEL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Sign Language Understanding (SLU) encompasses a diverse set of tasks, including Isolated Sign Language Recognition (ISLR), Continuous Sign Language Recognition (CSLR), and Sign Language Translation (SLT). Although these tasks share basic semantic and linguistic foundations, they are typically addressed with task-specific architectures and training pipelines, which hinders knowledge sharing and requires costly pretraining and finetuning for each task. In this paper, we focus on two aspects of SLU tasks: (1) training and inference pipelines are highly fragmented: most methods rely on pretraining on large-scale SL datasets followed by task- or dataset-specific finetuning, which leads to multiple specialized models rather than a single checkpoint. (2) current LLM-based methods may overlook the inherent modality discrepancy between sign and text tokens, simply concatenating them and processing both modalities with the same decoder layers. In this paper, we present OneSign, a unified framework that addresses multiple SLU tasks within a single model and a single checkpoint. OneSign reformulates ISLR, CSLR, and SLT under a single training paradigm, allowing all tasks to be trained and inferred using the same model checkpoint without dataset or task-specific pretraining or retraining. To accommodate the heterogeneous characteristics of sign and text representations, we introduce a **Modality-Adaptive Mixture-of-Experts**(MA-MoE) architecture, consisting of a shared expert and modality-specific experts for sign and text tokens. A modality router dynamically activates the corresponding experts, and their outputs are aggregated to form the final token representations. By enabling modality-dependent expert specialization while preserving a shared expert path, MA-MoE can effectively model the modality differences between continuous sign representations and discrete text tokens. Extensive experiments on multiple benchmarks demonstrate that OneSign achieves competitive or state-of-the-art performance on several benchmarks, highlighting its effectiveness as a unified SLU model.

## 1 INTRODUCTION

Sign Language (SL) Understanding (SLU) encompasses a family of closely related tasks, including Isolated SL Recognition (ISLR) (Albanie et al., 2021), Continuous SL Recognition (CSLR) (Min et al., 2021), and SL Translation (SLT) (Shi et al., 2019). ISLR focuses on recognizing a single sign instance from a short, trimmed sign clip, typically formulated as a classification problem. CSLR aims to recognize unsegmented sign videos and output a sequence of gloss labels, addressing the challenge of temporal alignment between sign sequence and gloss sequence. In contrast, SLT (Sincan et al., 2023) goes beyond recognition by translating sign language (SL) sequences directly into spoken language sentences. Despite strong semantic connections of these tasks, existing SLU methods are predominantly developed in a task-specific manner, where each task is addressed using an independent model architecture and training pipeline. For ISLR, models usually adopt a visual backbone to capture spatiotemporal features, followed by a classification head for sign prediction. For CSLR, a typical pipeline adopts a visual backbone for feature extraction (Gan et al., 2023a; 2026d), together with temporal modeling modules, and leverages CTC-based objectives to handle the latent alignment between the sign videos and the gloss labels, yielding the target gloss sequence. A similar task-specific design paradigm also dominates current SLT models. Most SLT approaches first employ a sign language backbone to extract visual representations (Zhou et al., 2021b; Gan et al.,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

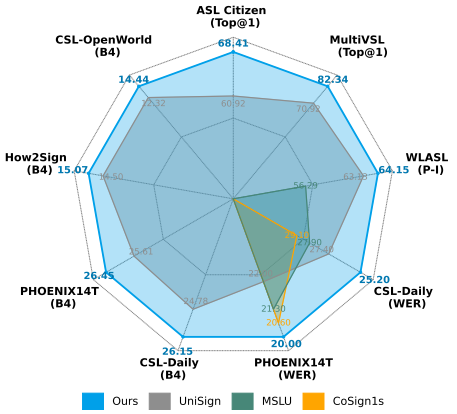


Figure 1: Overall performance comparison across multiple tasks.

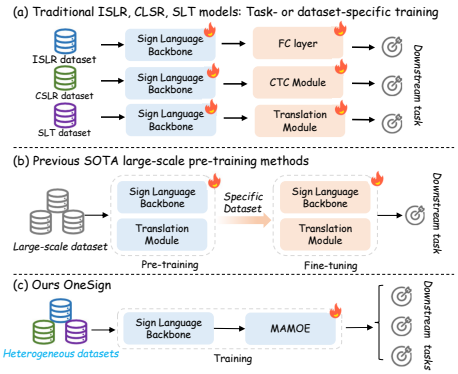


Figure 2: Comparison between previous SLU pipelines and the proposed OneSign framework.

2023a), followed by temporal encoding modules. The resulting sign features are then concatenated with textual embeddings and fed into a translation model and processed with the same transformer layer (*e.g.*, GPT-2 (Wong et al., 2024), Gemma (Guo et al., 2025) and Llama (Gan et al., 2026a)), to generate the corresponding spoken language sentence.

Recently, several unified SL models (*e.g.*, MSLU (Zhou et al., 2025), UniSign (Li et al., 2025) and ShuBERT (Gueuwou et al., 2025b)) have shown promising progress toward addressing multiple SLU tasks within a pretraining framework. These approaches typically rely on large-scale SL datasets for pretraining, followed by task- or dataset-specific finetuning to obtain dataset-dependent model checkpoints, achieving notable performance improvements. Despite their success, existing unified SLU models still exhibit two key limitations. (1) they depend on large annotated SL datasets for pretraining, which are expensive to construct. However, models still need to be finetuned separately for each downstream dataset, limiting training efficiency and model reusability. (2) many recent methods adopt decoder-only architectures, where large language models (LLMs) (*e.g.*, GPT-2 or Gemma) are used as translation modules. In these frameworks, sign tokens and text tokens are typically concatenated into a single sequence and processed by the same Transformer decoder layers, implicitly treating sign and text representations as belonging to the same modality. However, due to the inherent distributional differences between continuous SL representations and discrete textual tokens, this uniform modeling strategy overlooks the modality gap, which can hinder effective cross-modal reasoning and representation learning.

In this paper, we focus on core SLU tasks, including ISLR, CSLR, and SLT, and propose OneSign, a unified framework that aims to solve heterogeneous SLU tasks within a single model and a single checkpoint. Unlike existing frameworks that follow a rigid pretraining–finetuning paradigm, where models are first pretrained on large-scale annotated datasets or carefully designed pretext tasks, and then separately finetuned on each downstream task. OneSign reformulates SLU training under a unified paradigm (See Figure 2). This enables heterogeneous SLU datasets, including ISLR, CSLR, and SLT datasets, to be jointly trained within the same framework, without task-specific pretraining or retraining.

Furthermore, we observe that directly modeling sign tokens and text tokens within a standard LLM introduces significant challenges due to the inherent distributional gap between continuous sign representations and discrete textual tokens. To address this issue, we propose a Modality-Adaptive Mixture-of-Experts (MA-MoE) architecture, which explicitly accounts for modality heterogeneity while preserving shared semantic modeling. MA-MoE consists of an always-activated shared expert that captures modality-invariant sign–text semantics, together with modality-specific experts specialized for sign and text representations, respectively. Through unified training and modality-adaptive modeling, OneSign effectively models the difference between heterogeneous sign and text modalities, enabling efficient multi-task learning and inference across diverse SLU tasks using a single model checkpoint (Figure. 1). The main contributions are as follows:

- We construct **CSL-OpenWorld**, a large-scale Chinese SL (CSL) dataset consisting of over 100k video samples with manually curated annotations collected from diverse web sources

across multiple domains (*e.g.*, news, law, and culture), substantially expanding existing CSL datasets. The dataset will be publicly released.

- We introduce a **Modality-Adaptive Mixture-of-Experts (MA-MoE)** architecture for LLM-based SLU modeling, which combines an always-activated shared expert with modality-specific experts to explicitly model the distributional difference between continuous sign representations and discrete text tokens.
- We propose **OneSign**, the first unified framework that reformulates ISLR, CSLR, and SLT under one learning paradigm, enabling unified training and inference across heterogeneous SLU tasks with a single model checkpoint, without task-specific finetuning. Extensive experiments on multiple benchmarks show the effectiveness of OneSign across multiple SLU tasks.

## 2 METHOD

### 2.1 DATASETS

The lack of large-scale, high-diversity corpora has long been a major bottleneck in SL research. While substantial progress has been made for American SL (ASL) (Tanzer, 2025; Uthus et al., 2023)s and British SL (BSL) (Albanie et al., 2021) , Chinese SL (CSL) remains under-resourced, limiting the generality and scalability of existing methods. Moreover, existing CSL benchmarks are largely "domain-locked," consisting of either low-vocabulary laboratory recordings (*e.g.*, CSL-Daily) or highly formalized and linguistically constrained television broadcasts (*e.g.*, CSL-News). As a result, these CSL datasets fail to capture the morphological richness and environmental variability of natural signing.

To address these limitations, we introduce **CSL-OpenWorld**, a large-scale web-sourced dataset for open-domain SLT. As summarized in Table 1, CSL-OpenWorld expands the CSL vocabulary to over 12K tokens by leveraging diverse online platforms, representing a six-fold increase over CSL-Daily and more than that of CSL-News. Furthermore, as illustrated in Figures 3 and 4, CSL-OpenWorld covers a wide range of topics, camera viewpoints, and non-studio environments, providing a rigorous testbed for evaluating model robustness and real-world generalization.

Table 1: Comparison with existing SLT datasets. **Vocab**: vocabulary size; **Hours**: total video duration in hours; **Topic**: main domain of videos (*e.g.*, daily life, TV news).

| Name | Language | Vocab. | Hours | Source | Topic |
|---|---|---|---|---|---|
| VRT-RAW (Camgöz et al., 2021) | VGT | - | 100 | TV | Broadcast |
| SWISSTXT (Camgöz et al., 2021) | DGS | - | 88 | TV | News |
| PHOENIX-2014T (Camgoz et al., 2018) | DGS | 3K | 11 | TV | Weather |
| BOBSL (Albanie et al., 2021) | BSL | 77K | 1,447 | TV | Broadcast |
| How2Sign (Duarte et al., 2021) | ASL | 16K | 79 | Lab | Daily |
| OpenASL (Shi et al., 2022) | ASL | 33K | 288 | Web | Multi-Topic |
| YouTube-ASL (Uthus et al., 2023) | ASL | 60K | 984 | Web | Multi-Topic |
| SP-10 (Yin et al., 2022) | Various | 17K | 14 | Web | Multilingual |
| AfriSign (Gueuwou et al., 2023) | Various | 20K | 152 | Web | Multilingual |
| CSL-Daily (Zhou et al., 2021a) | CSL | 2K | 23 | Lab | Daily |
| CSL-News (Li et al., 2025) | CSL | 5K | 1,985 | TV | News |
| CSL-OpenWorld (Ours) | CSL | 12K | 216 | Web | Multi-Topic |

CSL-OpenWorld contains 100,458 video samples in total, with 86,387 used for training, making it substantially larger than CSL-Daily and Phoenix14t. The dataset features a high-density token vocabulary, with 12K unique tokens, enabling a challenging benchmark for evaluating the generalization ability of SLT models. The dataset is constructed through a rigorous pipeline involving web crawling, manual annotation, and quality filtering to ensure accurate video–text pairs. By incorporating diverse, open-domain web data, CSL-OpenWorld serves as a critical component in evaluating OneSign, enabling assessment of model robustness across domains and linguistic variations, and providing a useful resource for advancing future SL research.

### 2.2 PRELIMINARIES

**ISLR task formulation.** ISLR focuses on a short sign clip containing a single sign, formulated as a classification problem. Given an input sign sequence $\mathbf{X}$, the objective is to find the optimal sign label $g^* \in \mathcal{G}$ that maximizes the expected probability under the model parameters $\theta$:

$$g^* = \arg\max_{g \in \mathcal{G}} \mathbb{E}_{g' \sim p_\theta(\cdot|\mathbf{X})}[\mathbb{I}(g = g')] \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, effectively seeking the Maximum A Posteriori (MAP) estimate.

**CSLR task formulation.** CSLR aims to recognize continuous sign sequences by predicting a gloss sequence $\mathbf{G} = \{g_1, \ldots, g_L\}$. Since the mapping involves a latent alignment between signs and glosses, the objective is formulated as the expectation over all valid align paths $\pi$ that map to $\mathbf{G}$ through a many-to-one function $\mathcal{B}$:

$$\mathbf{G}^* = \arg\max_{\mathbf{G}\in\mathcal{G}^L} \mathbb{E}_{\pi\sim p_\theta(\cdot|\mathbf{X})}[\mathbb{I}(\mathcal{B}(\pi) = \mathbf{G})] \quad (2)$$

**SLT task formulation.** SLT aims to translate sign sequences into a text sequence $\mathbf{Y}^* = \{y_1, \ldots, y_U\}$. This task is commonly formulated as a conditional language generation problem, where the objective is to find $\mathbf{Y}^*$ that maximizes the expected log-likelihood (or minimizes the Bayes risk) over the target space $\mathcal{Y}^U$, expressed as:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}\in\mathcal{Y}^U} \mathbb{E}_{\mathbf{Y}'\sim p_\theta(\cdot|\mathbf{X})}[\mathbb{I}(\mathbf{Y} = \mathbf{Y}')] \quad (3)$$

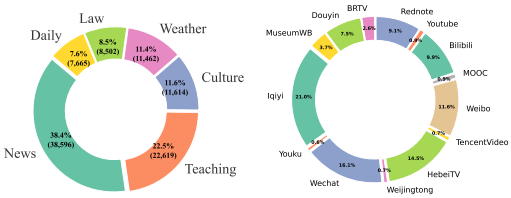


Figure 3: Distribution of topics and data sources.

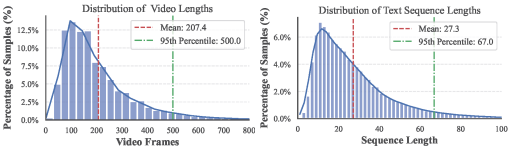


Figure 4: Distribution of video durations and text lengths.

In practice, this is often decomposed using the chain rule of probability as $p_\theta(\mathbf{Y} \mid \mathbf{X}) = \prod_{u=1}^{U} p_\theta(y_u \mid \mathbf{Y}_{<u}, \mathbf{X})$, where the model learns to predict the next token conditioned on $\mathbf{X}$ and the previously generated words $\mathbf{Y}_{<u}$.

**Unified Perspective.** Despite differences, all tasks can be unified under a single probabilistic framework. In OneSign, each sample is represented as $\{\mathbf{X}, \mathbf{G}, \mathbf{Y}\}$. The model is trained by maximizing the joint conditional likelihood:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{X},\mathbf{G},\mathbf{Y})\sim\mathcal{D}} \left[\log(p_\theta(\mathbf{Y}, \mathbf{G}) \mid \mathbf{X})\right]. \quad (4)$$

This unified formulation allows OneSign to be trained under a single maximum-likelihood objective from heterogeneous datasets. Moreover, it naturally subsumes multiple SLU tasks as special cases:

- **ISLR:** $\mathbf{G} = \mathbf{Y} = g$, where both denote a single sign label.
- **CSLR:** $\mathbf{G} = \mathbf{Y} = \{g_1, \ldots, g_L\}$, corresponding to a gloss sequence.
- **GBSLT:** $\mathbf{G}$ denotes the gloss sequence, while $\mathbf{Y}$ represents the spoken language sentence.
- **GFSLT:** $\mathbf{G}$ corresponds to a pseudo-gloss sequence, and $\mathbf{Y}$ denotes the spoken language sentence.

Accordingly, the proposed unified architecture autoregressively generates $\mathbf{Y}$ as the common prediction target across tasks, with $\mathbf{G}$ providing intermediate supervision during training.

### 2.3 Model Architecture

As illustrated in Figure 5, our model consists of three main components: (i) a visual encoder that converts an input sign sequence into a sequence of visual features, (ii) a text encoder that maps text tokens into continuous text representations, and (iii) a Modality-Adaptive Mixture-of-Experts (MA-MoE) decoder that performs unified sign-text modeling. To enable efficient training on large-scale datasets under limited GPU resources, we adopt pose-based representations instead of raw RGB videos for all tasks.

Specifically, for a sign sequence $\mathbf{X} = \{x_1, \ldots, x_T\}$, each frame $x_t$ is represented using human pose information and decomposed into four semantic parts: $x_t = \{x_t^b, x_t^f, x_t^{lh}, x_t^{rh}\}$, corresponding to *torso*, *face*, *left hand*, and *right hand*, respectively. We first project each part into a shared high-dimensional feature space using independent linear transformations: $\mathbf{h}_t^p = W_p x_t^p, \quad p \in \{b, f, lh, rh\}$, where $W_p$ denotes a learnable linear projection. This yields four part-level feature embeddings per frame.

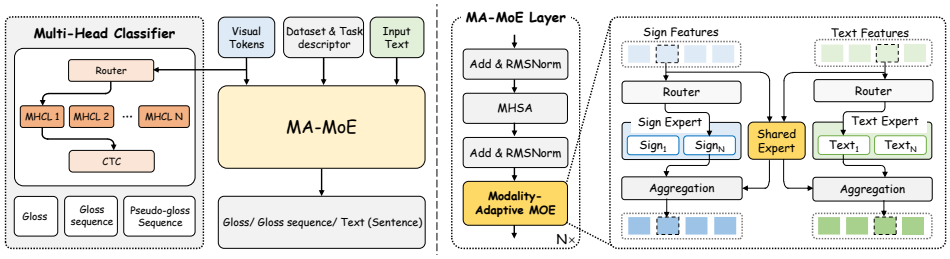


Figure 5: Left: the OneSign Architecture. Right: The MA-MoE Layer with Modality Adaptive MoE structure.

**Dynamic Graph Fusion for Body Part.** To effectively model the spatial correlations and structural dependencies between these body parts, inspired by previous pose modeling (Chen et al., 2021), we propose a simple yet effective pose fusing method: the Dynamic Part Graph Fusion (DPGF). Unlike static graph models, DPGF treats each body part as a node in a dynamic graph and learns the underlying topology on-the-fly. Specifically, sign features are concatenated to form a node matrix $\mathbf{H} \in \mathbb{R}^{4\times D}$, where each node represents a body part. We employ a self-attention mechanism to compute the dynamic adjacency matrix $\mathbf{A} \in \mathbb{R}^{4\times 4}$, which defines the edge strengths between parts for each frame:

$$\mathbf{A} = \text{Softmax}\left(\frac{Q(\mathbf{H})K(\mathbf{H})^\top}{\sqrt{D}}\right) \tag{5}$$

where $Q$ and $K$ are linear transformations for Query and Key. This allows the model to adaptively focus on different part correlations (*e.g.*, hand-face coordination) depending on the sign being performed. The node features are updated through a weighted aggregation of their neighbors followed by a GCN-style update with a residual connection:

$$\mathbf{H}_{fused} = W * (\text{MLP}(\mathbf{A}V(\mathbf{H})) + \mathbf{H}) \tag{6}$$

where $V$ is the linear projection. The updated nodes are then flattened and projected with $W$ to the final feature space $\mathbf{H}_{fused} \in \mathbb{R}^{B\times T\times C}$, providing a comprehensive representation of the sign gesture for each frame.

**Multi-Head Classifier Layers.** The visual backbone is trained with a CTC objective using $\mathbf{G}$ as supervision. When mixing multiple heterogeneous SL datasets, directly using a single shared classifier would require merging their gloss vocabularies, resulting in an excessively large output space and severe label sparsity. Inspired by previous multi-task learning (Long et al., 2017), we therefore adopt Multi-Head Classifier Layers (MHCL), in which each dataset is assigned a dedicated classification head with its own gloss vocabulary, while all heads share the same visual backbone. This strategy allows the backbone to learn unified visual representations across datasets, while each classifier head specializes in its dataset-specific label space. By decoupling feature learning from dataset-specific vocabularies, this design improves optimization stability and enables scalable multi-dataset pretraining. **MHCL is only used during training to stabilize optimization. It is removed at inference, where all outputs are generated by the unified MA-MOE without dataset-specific heads.**

### 2.4 Modality-Adaptive LLM MoE

After obtaining sign features, many approaches concatenate sign and text tokens into a single sequence and process them using standard LLM layers. Such a design implicitly treats sign and text tokens as the same modality, overlooking their inherent modality-specific characteristics and relying entirely on shared parameters across all decoder layers.

However, this design relies on an implicit assumption: *sign tokens and textual tokens are well aligned in the latent space*, such that a unified set of transformation layers can model both modalities equally well without explicit distinction. In practice, sign representations derived from pose sequences exhibit markedly different statistical properties, temporal dynamics, and levels of semantic abstraction compared to discrete text tokens. As a result, forcing both modalities to share identical transformation paths introduces a non-negligible *modality gap*, which can hinder effective cross-modal modeling and degrade overall performance.

To address this issue, we propose a **Modality-Adaptive Mixture-of-Experts (MA-MoE)** architecture that explicitly models modality-specific distributions while preserving shared semantic representations. As illustrated in Fig. 5, MA-MoE augments a standard Transformer block by introducing modality-adaptive expert routing within the feed-forward network (FFN). Specifically, the expert pool is decomposed into three disjoint groups: a shared expert, $N$ sign-specific experts, and $M$ text-specific experts. The shared expert is always activated to capture modality-invariant semantics, while modality-specific experts focus on modeling characteristics unique to sign or text tokens.

Let the input to the MA-MoE layer be a sequence of token representations: $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_N\}, \quad \mathbf{h}_i \in \mathbb{R}^D$, where each token is associated with a modality label $m_i \in \{\text{sign}, \text{text}\}$. The MA-MoE module consists of a shared expert $\mathcal{E}_{\text{shared}} = \{E_s\}$, a set of sign-specific experts $\mathcal{E}_{\text{sign}}$, and a set of text-specific experts $\mathcal{E}_{\text{text}}$. For each token $\mathbf{h}_i$, the shared expert output is first computed as

$$\mathbf{h}_i^s = E_s(\mathbf{h}_i). \tag{7}$$

Modality-specific routing is then performed within each expert group. A lightweight router produces routing logits:

$$\mathbf{r}_i = \begin{cases} \text{Router}_{\text{sign}}(\mathbf{h}_i), & \text{if } m_i = \text{sign}, \\ \text{Router}_{\text{text}}(\mathbf{h}_i), & \text{if } m_i = \text{text}, \end{cases} \tag{8}$$

which are normalized over the modality-specific expert set:

$$\alpha_{i,e} = \frac{\exp(r_{i,e})}{\sum_{e' \in \mathcal{E}_{m_i}} \exp(r_{i,e'})}, \quad e \in \mathcal{E}_{m_i}. \tag{9}$$

We denote the set of selected top-$k$ expert indices as $\tau$. The final output representation is obtained by aggregating the shared and modality-specific expert outputs:

$$\tilde{\mathbf{h}}_i = \mathbf{h}_i^s + \sum_{e \in \tau} \alpha_{i,e}\, \mathcal{E}_{m_i,e}(\mathbf{h}_i) \tag{10}$$

By construction, MA-MoE disentangles modality-invariant alignment from modality-specific modeling, enabling effective joint reasoning over heterogeneous sign and text representations within a unified LLM architecture.

**Load Balancing Loss.** To prevent expert collapse and encourage balanced utilization of modality-specific experts, we adopt a standard load balancing objective for Top-$K$ routing. Since the shared expert is always activated, the balancing loss is applied only to modality-specific experts. For a modality $m \in \{\text{sign}, \text{text}\}$, let $\mathcal{E}_m$ denote its modality-specific experts and let $N_m$ be the number of modality-$m$ tokens in a mini-batch. Given the routing probability $\alpha_{i,e}$ for token $\mathbf{h}_i$, the average routing probability of expert $e$ is

$$P_e^{(m)} = \frac{1}{N_m} \sum_{i:m_i=m} \alpha_{i,e}. \tag{11}$$

The routing load of expert $e$ is defined as the fraction of tokens dispatched to it:

$$F_e^{(m)} = \frac{1}{N_m K} \sum_{i:m_i=m} \mathbf{1}\left(e \in \text{TopK}(\alpha_i)\right), \tag{12}$$

The $\text{TopK}(\alpha_i)$ denotes the set of the Top-$K$ selected experts for token $i$. The load balancing loss for modality $m$ is

$$\mathcal{L}_{\text{bal}}^{(m)} = |\mathcal{E}_m| \sum_{e \in \mathcal{E}_m} P_e^{(m)} F_e^{(m)}. \tag{13}$$

The final load balancing loss is computed as the weighted sum over modalities: $\mathcal{L}_{\text{bal}} = \lambda_{\text{sign}}\, \mathcal{L}_{\text{bal}}^{(\text{sign})} + \lambda_{\text{text}}\, \mathcal{L}_{\text{bal}}^{(\text{text})}$ where $\lambda_{\text{sign}}$ and $\lambda_{\text{text}}$ are hyperparameters controlling the strength of the balancing constraint for each modality. This loss encourages the router to distribute tokens evenly across modality-specific experts while still allowing experts to specialize in capturing modality-dependent patterns. As a result, the load balancing loss improves training stability and ensures effective capacity utilization within each modality.

Table 2: Effect of Multi-Head Classification Layer.

| Model | DEV | | TEST | |
|---|---|---|---|---|
| | WER | del/ins | WER | del/ins |
| OneSign♡ | 21.1 | 7.2/3.0 | 22.1 | 8.1/2.3 |
| OneSign♠ W/O MHCL | 23.5 | 11.8/2.1 | 24.9 | 11.7/2.5 |
| OneSign♠ | 20.0 | 5.2/3.1 | 20.1 | 5.7/2.9 |

Table 3: Impact of activated expert count on Dev set.

| $N_s$=$N_t$ | Top-$k$ | Params(M) | Activated(M) | Rouge | BLEU1 | BLEU4 |
|---|---|---|---|---|---|---|
| 0 | 0 | 1235.8 | 1235.8 | 49.45 | 49.36 | 22.67 |
| 2 | 1 | 4457.0 | 2041.1 | 52.34 | 53.41 | 26.98 |
| 4 | 1 | 7678.2 | 2041.1 | 52.74 | 53.65 | 26.78 |
| 4 | 2 | 7678.2 | 3651.7 | 52.99 | 53.94 | 26.97 |

Table 4: Ablation studies of MA-MoE configurations.

| Experts | | | DEV | | | | TEST | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Shared | Sign | Text | WER | Rouge | BLEU1 | BLEU4 | WER | Rouge | BLEU1 | BLEU4 |
| ✓ | | | 22.3 | 49.12 | 49.89 | 23.45 | 22.5 | 48.99 | 49.72 | 23.01 |
| ✓ | ✓ | | 21.0 | 50.12 | 49.65 | 23.65 | 21.4 | 50.01 | 49.41 | 23.47 |
| ✓ | | ✓ | 20.8 | 50.64 | 50.12 | 24.36 | 21.0 | 50.03 | 49.95 | 24.13 |
| ✓ | ✓ | ✓ | 20.0 | 52.34 | 53.41 | 26.98 | 20.1 | 52.01 | 53.36 | 26.45 |

Table 5: Effect of heterogeneous training datasets.

| Dataset | TEST | | | |
|---|---|---|---|---|
| | WER | Rouge | BLEU1 | BLEU4 |
| PHOENIX14T | 22.1 | 51.18 | 52.68 | 26.22 |
| +CSLR | 21.5 | 51.65 | 52.45 | 26.25 |
| +ISLR+CSLR | 21.0 | 52.18 | 52.65 | 26.24 |
| +ISLR+CSLR+SLT | 20.1 | 52.01 | 53.36 | 26.45 |

**Training.** The OneSign model is optimized with two objectives that combine recognition and translation losses:

$$\begin{aligned}\mathcal{J}(\Theta) = \; & \alpha\, \mathbb{E}_{(X,G)\sim\mathcal{D}}\left[\ell_{\mathrm{CTC}}(R(X;\Theta_R), G)\right] \\ & + \beta\, \mathbb{E}_{(\tilde{\mathbf{h}}_i,Y)\sim\mathcal{D}}\Big[\ell_{\mathrm{CE}}\Big(Y, T(\tilde{\mathbf{h}}_i)\Big) + \mathcal{L}_{\mathrm{bal}}^{(s)} + \mathcal{L}_{\mathrm{bal}}^{(t)}\Big].\end{aligned} \tag{14}$$

Here, $R$ and $T$ denote the recognition and translation modules, respectively. Following (Gan et al., 2026d), the CTC loss supervises the visual backbone with MHCL, while the translation module $T$ is optimized with a translation loss with MOE balancing losses. During training, we first warm up the visual backbone for 20 epochs by setting $\alpha = 1$,$\beta = 0$, and activate the translation objective by further setting $\beta = 1$ in the subsequent 30 epochs.

## 3 Ablation Study

Following previous work (Min et al., 2021), we perform ablation studies on the Phoenix14t dataset to verify the effectiveness of the proposed OneSign model. *Please refer to Appendix C for detailed implementation settings.*

**A1: Effect of the Multi-Head Classification Layer.** Using a single CTC-based classification head for heterogeneous SLU datasets leads to a large vocabulary and unstable training. Table 2 compares three settings: (1) **OneSign**♡ , trained on a single dataset; (2) **OneSign**♠ W/O MHCL, trained on multiple datasets with a shared head; and (3) **OneSign**♠ with an independent head per dataset. As can be seen, directly sharing a head across datasets dilutes the original distributions and yields limited gains, whereas MHCL improves both performance and training stability, highlighting its effectiveness for multi-dataset learning.

**A2: Effect of Heterogeneous Training Datasets.** Table 5 shows the impact of jointly training on heterogeneous datasets: *ISLR* (WLASL, ASLCitizen, MultiVSL), *CSLR* (CSL-Daily, Phoenix14t), and *SLT* (CSL-Daily, Phoenix14t, How2Sign, CSL-OpenWorld), evaluated on Phoenix14t. Adding datasets without task-specific finetuning yields only modest gains. The results show that naively adding heterogeneous datasets brings only modest improvements (*e.g.*, +CSLR to +ISLR+CSLR), as the lack of any downstream finetuning can dilute the original dataset distributions. Still, OneSign effectively leverages complementary supervision, demonstrating the benefit of cross-task knowledge transfer in unified multi-dataset training.

**A3: Effect of the Proposed MA-MoE.** Table 4 presents an ablation study on different expert configurations in the proposed MA-MoE. Using only the shared expert (*i.e.*, Llama3.2 1B model) results in suboptimal performance, indicating that modality-agnostic modeling alone is insufficient. Adding either a sign or text specific expert on top of the shared expert consistently improves performance, highlighting the value of modality-aware MOE design. The full configuration, which includes both sign and text specific experts, achieves the best results across all metrics, confirming that MA-MoE effectively balances shared semantic modeling with modality-specific representation learning.

**A4: Effect of Sparse MA-MoE Configurations.** Table 3 examines expert sparsity by varying the total number of experts ($N_s = N_t$) and activated experts (Top-$k$). Activating a single expert

Table 6: Effects of Dynamic Part Graph Fusion.

| | Dev | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|
| | WER | ROUGE | BLEU1 | BLEU4 | WER | ROUGE | BLEU1 | BLEU4 |
| W/O DPGF | 21.6 | 51.01 | 52.87 | 25.83 | 21.1 | 51.07 | 52.17 | 25.63 |
| W/ Fix | 20.6 | 51.87 | 52.36 | 26.36 | 20.7 | 51.46 | 52.48 | 25.87 |
| W/ DPGF | 20.0 | 52.35 | 53.41 | 26.98 | 20.1 | 52.01 | 53.36 | 26.45 |

Table 7: Training and inference speed.

| Model | Training Time | Inference Time (250 frames) |
|---|---|---|
| Baseline | ∼10.4 min/epoch | ∼12.5 samples/s |
| +MA-MOE | ∼11.2 min/epoch | ∼11.8 samples/s |

Table 8: ISLR results on various benchmarks. * denotes results reproduced by us using publicly available code, and ✓* indicates pretraining on large-scale SL datasets.

| Method | Extra | | | ASL Citizen | | WLASL2000 | | MultiVSL1000 | |
|---|---|---|---|---|---|---|---|---|---|
| | Pose | RGB | FT | Top@1 | Top@5 | P-I | P-C | Top@1 | Top@5 |
| I3D | | ✓ | ✓ | 63.10 | 86.09 | 37.91 | 35.90 | 40.60 | 73.03 |
| HMA (Hu et al., 2021) | | ✓ | ✓ | - | - | 37.91 | 35.90 | - | - |
| NLA-SLR (Zuo et al., 2023) | | ✓ | ✓ | - | - | 61.05 | 58.05 | - | - |
| Uni-Sign (Li et al., 2025) | ✓ | ✓ | ✓* | - | - | 63.52 | 61.32 | - | - |
| SHuBERT (Gueuwou et al., 2025b) | | ✓ | ✓* | 65.00 | 87.00 | 60.90 | 58.01 | - | - |
| Pose-based | | | | | | | | | |
| ST-GCN (Yan et al., 2018) | ✓ | | ✓ | 59.52 | 82.68 | 34.40 | 32.53 | 40.61* | 73.04* |
| BEST (Zhao et al., 2023) | ✓ | | ✓ | - | - | 46.25 | 43.52 | - | - |
| SignBERT+ (Hu et al., 2023a) | ✓ | ✓ | ✓ | - | - | 48.85 | 46.37 | - | - |
| MSLU (Zhou et al., 2025) | ✓ | | ✓* | - | - | 56.29 | 53.29 | - | - |
| Uni-Sign (Li et al., 2025) | ✓ | | ✓* | 60.92* | 83.41* | 63.13 | 60.90 | 70.92* | 89.90* |
| OneSign♡ | ✓ | | ✓ | 65.41 | 84.19 | 60.45 | 61.02 | 80.71 | 95.43 |
| OneSign♠ | ✓ | | | 68.41 | 89.24 | 64.15 | 65.34 | 82.34 | 96.43 |

Table 9: CSLR results on CSL-Daily and Phoenix14T. "FT" denotes fine-tuning on each dataset.

| Method | Extra | | | CSL-Daily | | Phoenix14T | |
|---|---|---|---|---|---|---|---|
| | Pose | RGB | FT | Dev | Test | Dev | Test |
| SignBT (Zhou et al., 2021a) | | ✓ | ✓ | 33.2 | 33.2 | - | - |
| CorrNet (Hu et al., 2023b) | | ✓ | ✓ | 30.6 | 30.1 | 18.9 | 20.5 |
| SEN (Hu et al., 2023c) | | ✓ | ✓ | 31.1 | 30.7 | 19.3 | 20.7 |
| C2ST (Zhang et al., 2023) | | ✓ | ✓ | 25.9 | 25.8 | 17.5 | 17.7 |
| SignGraph (Gan et al., 2024) | | ✓ | ✓ | 27.3 | 26.4 | 17.8 | 19.1 |
| Uni-Sign (Li et al., 2025) | ✓ | ✓ | ✓* | 26.7 | 26.0 | - | - |
| TwoStream (Chen et al., 2022) | ✓ | ✓ | ✓ | 25.4 | 25.3 | 17.7 | 19.3 |
| Pose-based | | | | | | | |
| CoSign1s (Jiao et al., 2023) | ✓ | | ✓ | 29.5 | 29.1 | 20.4 | 20.6 |
| MSLU (Zhou et al., 2025) | ✓ | | ✓ | 28.6 | 27.9 | 20.1 | 21.3 |
| Uni-Sign (Li et al., 2025) | ✓ | | ✓* | 28.2 | 27.4 | 21.4* | 22.0* |
| OneSign♡ | ✓ | | ✓ | 29.0 | 29.1 | 21.1 | 22.1 |
| OneSign♠ | ✓ | | | 25.2 | 25.7 | 20.0 | 20.1 |

with more total experts gives minor gains, while activating multiple experts allows aggregation of complementary information, but requires more computational resources. To balance performance and efficiency, we choose $N_s = N_t = 2$ with Top-$k = 1$ as our final setting.

**A5: Effect of Dynamic Part Graph Fusion.** Table 6 reports an ablation study on the proposed Dynamic Part Graph Fusion (DPGF) module. When DPGF is removed (*i.e.*, features from different parts are fused using a simple linear ), the model performs worse on all metrics, indicating that naive part fusion is insufficient to capture fine-grained part-level interactions. Applying a fixed fusion strategy (W/ Fix), where fusion graphs are predefined for each part, brings modest and consistent improvements, suggesting that part-aware aggregation is beneficial but limited by its static design. In contrast, our model with DPGF module achieves the best performance , reducing WER from 21.1 to 20.1 and improving BLEU4 from 25.63 to 26.45. These gains indicate that dynamically adjusting fusion weights according to input content enables more effective modeling of inter-part dependencies, which in turn leads to more accurate recognition and translation.

**A6: Training and Inference Speed.** We evaluate the training and testing time on the Phoenix14t using a single NVIDIA RTX 4090 GPU. Specifically, training time is measured on the Phoenix14t training set, while inference speed is evaluated on test samples with an average length of 250 frames, averaged over 100 runs. As observed, the model requires about 10–13 minutes per epoch during training. During inference, it achieves a throughput of 10–12 samples per second on a single GPU. These results show that the proposed MA-MOE achieves good efficiency in both training and inference, making it practical for large-scale SL tasks.

## 4 COMPARISONS

We compare **OneSign** with existing task-specific models on ISLR, CSLR, and SLT benchmarks. To systematically evaluate the effect of unified training and inference, we report results under two training and evaluation settings: (1) **OneSign**$^{\heartsuit}$ is trained and evaluated on a single dataset(*i.e.*, each corresponding dataset). (2) **OneSign**$^{\spadesuit}$ is trained jointly on the training splits of all available datasets and directly evaluated on each benchmark without any further fine-tuning, measuring the generalization capability of a single unified checkpoint. *Note that a direct comparison with RGB-based models is not entirely fair, as RGB inputs contain richer visual information.*

**Evaluation on ISLR Tasks.** Table 8 compares OneSign with existing ISLR methods. Prior approaches are mostly RGB- or pose-based and often require dataset-specific pretraining or fine-tuning. Under the dataset-specific setting, OneSign$^{\heartsuit}$ already show promising performance across all benchmarks (*e.g.*, 65.41 Top@1 on ASL Citizen vs. 59.52 for ST-GCN). On WLASL2000 and MultiVSL1000, it similarly achieves clear gains. OneSign$^{\spadesuit}$ further improves performance, establishing a new baseline under a unified protocol. These results show that OneSign effectively leverages heterogeneous supervision and achieves strong multi-dataset performance with a single checkpoint.

Table 10: Comparison of SLT performance on Phoenix14T and CSL-Daily. * denotes results reproduced by us using publicly available code. We underline pose-based results that outperform our baseline model for clarity.

| SLT | Extra | | | Phoenix14T | | | | | | CSL-Daily | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | DEV | | | TEST | | | DEV | | | TEST | | |
| | Pose | RGB | FT | ROUGE | BLEU1 | BLEU4 | ROUGE | BLEU1 | BLEU4 | ROUGE | BLEU1 | BLEU4 | ROUGE | BLEU1 | BLEU4 |
| PGG-SLT (Guo et al., 2025) | | ✓ | ✓ | 52.01 | 53.16 | 27.09 | 51.85 | 53.45 | 26.85 | - | - | - | - | - | - |
| GFSLT-VLP (Yin et al., 2023) | | ✓ | ✓ | 43.72 | 44.08 | 22.12 | 42.49 | 43.71 | 21.44 | 36.44 | 39.20 | 11.07 | 36.70 | 39.37 | 11.00 |
| SignCL (Ye et al., 2024) | | ✓ | ✓ | | | | 49.04 | 49.76 | 22.74 | - | - | - | 48.92 | 47.47 | 16.16 |
| Sign2GPT (Wong et al., 2024) | | ✓ | ✓ | - | - | | 48.90 | 49.54 | 22.52 | - | - | - | 42.36 | 41.75 | 15.40 |
| GFSLT-VLP-SignCL (Ye et al., 2024) | | ✓ | ✓ | - | - | - | 49.04 | 49.76 | 22.74 | - | - | - | 48.92 | 47.47 | 16.16 |
| SignLLM (Gong et al., 2024) | | ✓ | ✓ | 44.49 | 46.88 | 25.25 | 47.23 | 45.21 | 23.40 | 39.18 | 42.45 | 12.23 | 39.91 | 39.55 | 15.75 |
| LLaVA-SLT (Liang et al., 2024) | | ✓ | ✓ | - | - | - | 50.44 | 51.20 | 23.43 | - | - | - | 51.26 | 52.15 | 20.42 |
| FLa-LLM (Chen et al., 2024) | | ✓ | ✓ | - | - | - | 45.27 | 46.29 | 23.09 | - | - | - | 37.25 | 37.13 | 14.20 |
| C2RL (Chen et al., 2025) | | ✓ | ✓ | - | - | - | 50.96 | 52.81 | 26.75 | - | - | - | 48.21 | 49.32 | 21.61 |
| MixSignGraph (Gan et al., 2026d) | | ✓ | ✓ | 51.71 | 51.07 | 24.87 | 51.14 | 50.01 | 24.02 | 49.16 | 49.98 | 20.43 | 49.93 | 50.24 | 20.78 |
| SignLlama (Gan et al., 2026a) | | ✓ | ✓ | 52.53 | 52.13 | 25.73 | 52.81 | 52.54 | 26.74 | 51.62 | 52.12 | 24.83 | 51.43 | 52.38 | 24.76 |
| TwoStream (Chen et al., 2022) | ✓ | ✓ | ✓ | 54.08 | 54.32 | 28.66 | 53.48 | 54.90 | 28.95 | 55.10 | 55.21 | 25.76 | 55.72 | 55.44 | 25.79 |
| Uni-Sign (Li et al., 2025) | ✓ | ✓ | ✓* | - | - | - | - | - | - | 56.03 | 55.30 | 26.25 | 56.51 | 55.08 | 26.36 |
| Pose-based | | | | | | | | | | | | | | | |
| VAP (Jiao et al., 2024) | ✓ | | ✓ | 51.47 | 52.78 | 26.62 | 51.28 | 53.07 | 26.02 | 51.19 | 53.31 | 23.84 | 51.09 | 52.98 | 23.65 |
| Uni-Sign (Li et al., 2025) | ✓ | | ✓* | 48.14* | 49.14* | 22.47* | 48.10* | 49.26* | 22.78* | 54.34 | 53.24 | 25.27 | 53.86 | 54.92 | 25.61 |
| OneSign♡ | ✓ | | ✓ | 51.28 | 52.68 | 26.38 | 51.18 | 52.68 | 26.22 | 50.05 | 52.43 | 24.85 | 51.39 | 51.02 | 24.42 |
| OneSign♠ | ✓ | | | 52.34 | 53.41 | 26.98 | 52.01 | 53.36 | 26.45 | 54.31 | 53.45 | 25.28 | 53.39 | 54.45 | 26.15 |

Table 11: Comparison of gloss-free SLT performance on How2Sign and CSL-OpenWorld.

| Gloss-free SLT | Extra | | | How2Sign | | | | | | CSL-OpenWorld | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | DEV | | | TEST | | | DEV | | | TEST | | |
| | Pose | RGB | FT | ROUGE | BLEU1 | BLEU4 | ROUGE | BLEU1 | BLEU4 | ROUGE | BLEU1 | BLEU4 | ROUGE | BLEU1 | BLEU4 |
| PGG-SLT (mBART) (Guo et al., 2025) | ✓ | ✓ | ✓ | 32.6 | 41.8 | 15.9 | 31.50 | 38.90 | 13.10 | - | - | - | - | - | - |
| YouTube-SLT* (Uthus et al., 2023) | | ✓ | ✓ | - | - | - | - | 14.96 | 1.22 | - | - | - | - | - | - |
| $C^2$RL (Chen et al., 2025) | | ✓ | ✓ | - | - | - | 27.02 | 29.07 | 9.37 | - | - | - | - | - | - |
| FLa-LLM (Chen et al., 2024) | | ✓ | ✓ | - | - | - | 27.81 | 29.81 | 9.66 | - | - | - | - | - | - |
| GloFE-VN (Lin et al., 2023) | | ✓ | ✓ | 12.98 | 15.21 | 2.37 | 12.61 | 14.94 | 2.24 | 15.92* | 16.33* | 4.36* | 15.86* | 16.23* | 4.30* |
| SSVP-SLT (Rust et al., 2024) | | ✓ | ✓ | - | - | - | 25.70 | 30.20 | 7.00 | - | - | - | - | - | - |
| SignMusketeers (Gueuwou et al., 2025a) | | ✓ | ✓ | - | - | - | - | 18.80 | 4.20 | - | - | - | - | - | - |
| SLT-IV (Tarrés et al., 2023) | | ✓ | ✓ | | | | - | 34.01 | 8.03 | - | - | - | - | - | - |
| MixSignGraph (Gan et al., 2026d) | | ✓ | ✓ | 29.24 | 34.82 | 11.28 | 28.01 | 34.74 | 10.41 | 22.30* | 23.75* | 8.75* | 22.32* | 24.00* | 8.91* |
| SHuBERT (Gueuwou et al., 2025b) | | ✓ | ✓* | - | - | - | - | - | 16.20 | - | - | - | - | - | - |
| Pose-based | | | | | | | | | | | | | | | |
| VAP (Jiao et al., 2024) | ✓ | ✓ | ✓ | 30.27 | 42.34 | 14.91 | 27.77 | 39.22 | 12.87 | - | - | - | - | - | - |
| Uni-Sign (Li et al., 2025) | ✓ | | ✓* | - | - | - | 34.30 | 40.40 | 14.50 | 30.45* | 31.44* | 12.47* | 31.48* | 32.17* | 12.32* |
| OneSign♡ | ✓ | | | 35.19 | 39.46 | 14.92 | 36.46 | 39.47 | 14.51 | 31.65 | 34.19 | 13.88 | 31.57 | 33.60 | 13.48 |
| OneSign♠ | ✓ | | | 37.16 | 41.56 | 15.41 | 39.15 | 42.78 | 15.07 | 32.90 | 35.18 | 14.71 | 32.99 | 34.83 | 14.44 |

**Evaluation on CSLR Tasks.** Table 9 shows results on CSL-Daily and Phoenix14t. Under dataset-specific settings, OneSign♡ achieves strong performance using only pose inputs (29.1 WER on CSL-Daily, 22.1 on Phoenix14t), matching or surpassing several RGB-based methods. OneSign♠, trained jointly on heterogeneous datasets without fine-tuning, further reduces WER to 25.7 and 20.1, respectively, showing good performance. Compared with prior pose-based methods that require fine-tuning, OneSign♠ achieves comparable or better results under a unified setting.

**Evaluation on SLT.** Tables 10 and 11 report results on two gloss-based SLT benchmarks: Phoenix14t, CSL-Daily, and two gloss-free SLT benchmarks: How2Sign, and CSL-OpenWorld. Unlike most SLT methods that rely on dataset-specific fine-tuning, OneSign operates purely on pose and supports unified training without fine-tuning. On Phoenix14t and CSL-Daily (Table 10), **OneSign**♡ already achieves competitive performance using only pose features. Joint training further brings consistent gains: **OneSign**♠ improves BLEU4 from 26.22 to 26.45 on Phoenix14t and from 24.42 to 26.15 on CSL-Daily. On the more challenging How2Sign and CSL-OpenWorld(Table 11), OneSign shows strong generalization. Without fine-tuning, **OneSign**♠ achieves the best or highly competitive results among pose-based methods, outperforming prior approaches such as Uni-Sign and VAP on many metrics. Overall, OneSign achieves strong cross-dataset performance with a single unified checkpoint, without relying on RGB inputs or dataset-specific fine-tuning.

## 5 Conclusion

In this paper, we present a unified framework that integrates ISLR, CSLR and SLT within a single model. By reformulating all tasks under a unified paradigm, our approach enables joint training and inference across different SL tasks using a single model checkpoint, without task-specific fine-tuning. To support effective learning from heterogeneous datasets, we introduce the MHCL that reduces vocabulary conflicts while allowing the shared backbone to be jointly optimized. Furthermore, to accommodate the diverse characteristics of sign and text modalities within a shared decoder, we propose a MA-MoE architecture with both shared and modality-specific experts, dynamically activated by a modality-aware router. Extensive experiments on seven public benchmarks across ISLR, CSLR, and SLT tasks demonstrate that our unified framework consistently achieves strong performance, validating its effectiveness across tasks and datasets.

### AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing and proofreading, including improving the clarity and fluency of the manuscript and identifying grammatical and typographical errors. Generative AI tools were not used for developing research ideas or methodologies, designing or conducting experiments, implementing code, analyzing experimental results, or drawing scientific conclusions. All research ideas, methods, experimental designs, implementations, and analyses were developed and conducted by the authors.

All AI-assisted textual revisions were carefully reviewed and verified by the authors to ensure that they accurately reflect the intended technical content and do not introduce unsupported claims or alter the scientific meaning. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

### ETHICS STATEMENT

This work develops a unified modeling framework for multiple sign language understanding tasks and involves no direct interaction with human subjects. In addition to existing public sign language benchmarks, we construct a dataset from publicly available online resources following standard research practices. The collected data are used solely for research purposes, and we do not intentionally collect or introduce additional personally identifiable information beyond what is already present in the original public content.

The proposed framework is intended to advance sign language understanding research and may contribute to improved accessibility of information and communication for Deaf and hard-of-hearing communities. At the same time, we recognize that sign language technologies may exhibit performance disparities across signers, sign languages, and demographic groups due to biases and limited coverage in available training data. Such systems should therefore be carefully evaluated before deployment in real-world or user-facing applications.

We do not foresee significant ethical risks arising directly from the proposed methodology beyond those generally associated with machine learning research using publicly available data. All datasets and experimental resources used in this work are handled in accordance with their respective terms of use and are used exclusively for research purposes.

### REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the OneSign architecture and training objectives in the main paper. The Appendix further documents the datasets, data preprocessing procedures, implementation details, training configurations, and hyperparameter settings used in our experiments. The datasets introduced in this work will be made publicly available to facilitate future research and reproduction. Additional experimental results and ablation studies are also provided to support verification of our findings. Together, the main paper and Appendix provide the necessary information to reproduce the experimental setup and results reported in this work.

## REFERENCES

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 32897–32912. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d46662aa53e78a62afd980a29e0c37ed-Paper-Conference.pdf.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pp. 7784–7793, 2018.

Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–5. IEEE, 2021.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *NeurIPS*, 35:17043–17056, 2022.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pp. 13359–13368, 2021.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7071–7081, 2024.

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. C 2 rl: Content and context representation learning for gloss-free sign language translation and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *ECCV*. Springer, 2020.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *ACL*, pp. 1280–1297, 2024.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. *NeurIPS*, 36:76893–76907, 2023.

Nguyen Son Dinh, Tuan Dung Nguyen, Duc Tri Tran, Nguyen Dang Huy Pham, Thuan Hieu Tran, Ngoc Anh Tong, Quang Huy Hoang, and Phi Le Nguyen. Sign language recognition: A large-scale multi-view dataset and comprehensive evaluation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7887–7897. IEEE, 2025.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1932–1945, 2024.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2735–2744, 2021.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. Skeleton-aware neural sign language translation. In *MM*, pp. 4353–4361, 2021.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. Contrastive learning for sign language recognition and translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 763–772, 2023a.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. Towards real-time sign language recognition and translation on edge devices. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 4502–4512, 2023b.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Hongkai Wen, Lei Xie, and Sanglu Lu. Signgraph: A sign sequence is worth graphs of nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13470–13479, 2024.

Shiwei Gan, Xiao Liu, Yafeng Yin, Zhiwei Jiang, Bowen Guo, Lie Xie, Sanglu Lu, and Hongkai Wen. Signllama: Enhancing gloss-free sign language translation by prioritizing visual features for llms. *Proceedings of the ACM international conference on multimedia*, 2026a.

Shiwei Gan, Xiao Liu, Yafeng Yin, Nan Liu, Kuizhuang Liu, Desibieer Tuerdaken, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. Learning effective sign features without text for gloss-free sign language translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9827–9836, 2026b.

Shiwei Gan, Lichen Wang, Xiao Liu, Yafeng Yin, Kuizhuang Liu, Sanglu Lu, and Lei Xie. Sign language question answering: A new task, benchmark, and baseline for sign language understanding. *arXiv preprint arXiv:2607.27826*, 2026c.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, Sanglu Lu, and Hongkai Wen. Mixsigngraph: A sign sequence is worth mixed graphs of nodes. *Advances in Neural Information Processing Systems*, 38:139371–139395, 2026d.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18362–18372, 2024.

Shester Gueuwou, Kate Takyi, Mathias Müller, Marco Stanley Nyarko, Richard Adade, and Rose-Mary Owusuaa Mensah Gyening. Afrisign: Machine translation for african sign languages. In *4th Workshop on African Natural Language Processing*, 2023.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. Signmusketeers: An efficient multi-stream approach for sign language translation at scale. *Findings of the Association for Computational Linguistics: ACL 2025*, 2025a.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28792–28810, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.1397/.

Dan Guo, Wengang Zhou, Meng Wang, and Houqiang Li. Sign language recognition based on adaptive hmms with data augmentation. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2876–2880. IEEE, 2016.

Jianyuan Guo, Peike Li, and Trevor Cohn. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. *arXiv preprint arXiv:2505.15438*, 2025.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. Extending the public dgs corpus in size and depth. In *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives*, pp. 75–82, 2020.

Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *ICCV*, pp. 11303–11312, 2021.

Hezhen Hu, Wengang Zhou, Junfu Pu, and Houqiang Li. Global-local enhancement network for nmfs-aware sign language recognition. *arXiv preprint arXiv:2008.10428*, 2020.

Hezhen Hu, Wengang Zhou, and Houqiang Li. Hand-model-aware sign language recognition. In *AAAI*, volume 35, pp. 1558–1566, 2021.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Temporal lift pooling for continuous sign language recognition. In *European conference on computer vision*, pp. 511–527. Springer, 2022.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2529–2539, June 2023b.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Self-emphasizing network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 854–862, 2023c.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI*, 2018.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *ICCV*, pp. 20676–20686, 2023.

Peiqi Jiao, Yuecong Min, and Xilin Chen. Visual alignment pre-training for sign language translation. In *European Conference on Computer Vision*, pp. 349–367. Springer, 2024.

Tao Jin and Zhou Zhao. Contrastive disentangled meta-learning for signer-independent sign language translation. In *MM*, pp. 5065–5073, 2021.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. Sign language translation with hierarchical spatio-temporal graph neural network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3367–3376, 2022.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683, 2019.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1459–1469, 2020a.

Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6205–6214, 2020b.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-sign: Toward unified sign language understanding at scale. *arXiv preprint arXiv:2501.15187*, 2025.

Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juze Zhang, Xin Chen, Jingyi Yu, and Lan Xu. Llava-slt: Visual language tuning for sign language translation. *arXiv preprint arXiv:2412.16524*, 2024.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Yi Yang, et al. Gloss-free end-to-end sign language translation. In *ACL*, 2023.

Tao Liu, Wengang Zhou, and Houqiang Li. Sign language recognition with long short-term memory. In *ICIP*, pp. 2871–2875. IEEE, 2016.

Xiao Liu, Shiwei Gan, Yafeng Yin, Bowen Guo, Zhiwei Jiang, Shunmei Meng, Lei Xie, and Sanglu Lu. Signpr: A progressive vector-quantized diffusion framework for sign language production. pp. 2198–2208, 2026a.

Xiao Liu, Shiwei Gan, Yafeng Yin, Jiaxin Yin, Bowen Guo, Yaqi Sun, Zhiwei Jiang, and Lei Xie. Bridging the gap between semantics and reconstruction: Unifying sign language translation and production. *Proceedings of the 31st ACM international conference on multimedia*, 2026b.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Philip S Yu. Learning multiple tasks with multilinear relationship networks. *Advances in neural information processing systems*, 30, 2017.

Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV*, pp. 11542–11551, 2021.

Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *ECCV*, pp. 172–186. Springer, 2020.

Alptekin Orbay and Lale Akarun. Neural sign language translation by learning tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 222–228. IEEE, 2020.

Bowen Pan, Yikang Shen, Haokun Liu, Mayank Mishra, Gaoyuan Zhang, Aude Oliva, Colin Raffel, and Rameswar Panda. Dense training, sparse inference: Rethinking training of mixture-of-experts language models. *arXiv preprint arXiv:2404.05567*, 2024.

Maria Parelli, Katerina Papadimitriou, Gerasimos Potamianos, Georgios Pavlakos, and Petros Maragos. Spatio-temporal graph convolutional networks for continuous sign language recognition. In *ICASSP*. IEEE, 2022.

Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting continuous sign language recognition via cross modality augmentation. In *MM*, pp. 1497–1505, 2020.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8624–8641, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.467.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6365–6379, 2022.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12026–12035, 2019.

Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. Is context all you need? scaling neural sign language translation to large domains of discourse. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1955–1965, 2023.

Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. Graph-based multimodal sequential embedding for sign language translation. *TMM*, 2021.

Garrett Tanzer. Fleurs-asl: Including american sign language in massively multilingual multitask evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6167–6191, 2025.

Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5624–5634, 2023.

Dominique Uebersax, Juergen Gall, Michael Van den Bergh, and Luc Van Gool. Real-time sign language letter and word recognition from depth data. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pp. 383–390. IEEE, 2011.

Dave Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36: 29029–29047, 2023.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, pp. 19175–19186, 2023.

Chengcheng Wei, Jian Zhao, Wengang Zhou, and Houqiang Li. Semantic boundary detection with reinforcement learning for continuous sign language recognition. *TCSVT*, 31(3):1138–1149, 2020.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2gpt: leveraging large language models for gloss-free sign language translation. In *ICLR 2024: The Twelfth International Conference on Learning Representations*, 2024.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Signrep: Enhancing self-supervised sign representations. *arXiv preprint arXiv:2503.08529*, 2025.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Zhaoyang Yang, Zhenmei Shi, Xiaoyong Shen, and Yu-Wing Tai. Sf-net: Structured feature network for continuous sign language recognition. *arXiv preprint arXiv:1908.01341*, 2019.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. Improving gloss-free sign language translation by reducing representation density. In *NeurIPS*, 2024.

Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Simulslt: End-to-end simultaneous sign language translation. In *MM*, pp. 4118–4127, 2021.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5109–5119, 2022.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2551–2562, 2023.

Huaiwen Zhang, Zihang Guo, Yang Yang, Xin Liu, and De Hu. C2st: Cross-modal contextualized sequence transduction for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21053–21062, October 2023.

Zhihao Zhang, Junfu Pu, Liansheng Zhuang, Wengang Zhou, and Houqiang Li. Continuous sign language recognition via reinforcement learning. In *ICIP*, pp. 285–289. IEEE, 2019.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. Best: Bert pre-training for sign language recognition with coupling tokenization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3597–3605, 2023.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, pp. 1316–1325, 2021a.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for sign language recognition and translation. *TMM*, 2021b.

Wengang Zhou, Weichao Zhao, Hezhen Hu, Zecheng Li, and Houqiang Li. Scaling up multimodal pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Qidan Zhu, Jing Li, Fei Yuan, and Quan Gan. Multiscale temporal network for continuous sign language recognition. *Journal of Electronic Imaging*, 33(2):023059–023059, 2024.

Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *CVPR*, pp. 5131–5140, 2022.

Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *CVPR*, pp. 14890–14900, 2023.

# A Appendix

# B Related Work

**ISLR.** SL Recognition (SLR) aims to recognize linguistic units from sign inputs and contains two tasks: ISLR and CSLR. Although both tasks focus on understanding sign semantics from videos, they differ primarily in supervision granularity and output structure, and are therefore typically treated as distinct problems. ISLR (Zuo et al., 2023) considers short, trimmed video clips containing a single sign and formulates recognition as a classification task. The standard ISLR paradigm follows a task-specific pipeline, where a visual backbone extracts spatiotemporal features that are directly mapped to predefined categories via a classification layer.

**CSLR.** In contrast, CSLR (Gan et al., 2024) addresses the recognition of unsegmented sign sequences from continuous video streams, introducing the challenge of implicitly aligning visual inputs with gloss sequences. As a result, CSLR is commonly formulated as a sequence recognition problem with weak temporal supervision, where the input sign sequences and gloss sequences share the same order but differ in length. To handle this alignment, the majority of CSLR methods adopt Connectionist Temporal Classification (CTC)-based frameworks. These approaches typically employ a vision backbone to extract sign features, followed by temporal modules to capture long-range dependencies, and a CTC-based decoder to align visual sequences with gloss sequences.

**SLT.** Early SLT approaches are closely tied to advances in CSLR (Zuo & Mak, 2022) . The prevailing paradigm in gloss-based SLT (GBSLT) relies on a pretrained CSLR backbone to provide intermediate gloss-aligned representations, which are subsequently fed into translation modules. In this setting, CSLR effectively serves as a pretext task that functions as a "sign tokenizer," aligning visual features with gloss sequences and enabling downstream translation. Further, increasing attention has been directed toward Gloss-Free SLT (GFSLT), which aims to directly translate sign videos into spoken language without intermediate gloss supervision. A fundamental challenge in GFSLT lies in how to tokenize sign sequence into discriminative sign features in the absence of gloss guidance. To address this issue, existing GFSLT methods typically rely on text annotations during pretraining, adopting strategies such as contrastive learning (Liang et al., 2024), CTC-based pseudo-gloss objectives (Gan et al., 2026d), text-aligned vector quantization (Gong et al., 2024), and multi-stage or curriculum-based pretraining schemes (Guo et al., 2025; Gueuwou et al., 2025b;a). These approaches aim to compensate for the lack of explicit gloss supervision by leveraging auxiliary alignment signals from textual modalities.

Existing SLR and SLT methods, though effective, are largely task-specific, with models and training objectives differing across ISLR, CSLR, and SLT. This fragmentation limits representation sharing and cross-task knowledge transfer, motivating the need for a unified, task-agnostic framework. Recent works, such as UniSign (Li et al., 2025), MSLU (Zhou et al., 2025), and ShuBERT(Gueuwou et al., 2025b), attempt multi-task learning by jointly pretraining their visual backbones or translation modules on large SL datasets and fine-tuning on each specific dataset. However, these models still rely on task- or dataset-specific fine-tuning.

**Mixture of Experts (MOE).** Compared with standard LLMs that process all tokens using shared parameters, MoE (Jiang et al., 2024; Dai et al., 2024) introduces conditional computation by routing each input token to a subset of experts. This design enables a better trade-off between model capacity and computational efficiency, and has been shown to be particularly effective in handling heterogeneous data distributions and mitigating task interference (Lepikhin et al., 2020). Existing MoE architectures can be broadly categorized into dense MoE and sparse MoE. In dense MoE (Dou et al., 2024; Pan et al., 2024), all experts are activated during both training and inference, with different experts contributing via weighted combinations. In contrast, sparse MoE selects only Top-$k$ experts for each token, leading to significantly improved computational efficiency while maintaining strong model capacity. In addition, prior works such as VLMo (Bao et al., 2022) and BEiT-3 (Wang et al., 2023) introduce mixture-of-modality-experts, which include a vision expert, a language expert, and a vision-language expert to handle different data pairs during vision-language pretraining. During fine-tuning, these models typically encode image and text separately.

Inspired by VLMo and BEiT-3, we propose MA-MoE, a sparse MoE architecture consisting of a shared expert and modality-specific experts for sign and text tokens. Compared with existing multi-

Table 12: Details of datasets used in our paper.

| Dataset | Video Samples | | | Gloss Vocabulary | | | Token Vocabulary | | |
|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Validation | Train | Test | Validation | Train | Test | Validation |
| ASL Citizen (Desai et al., 2023) | 40154 | 32941 | 10304 | 2731 | 2731 | 2731 | - | - | - |
| WLASL2000 (Li et al., 2020a) | 14289 | 2878 | 6149 | 2000 | 1999 | 2000 | - | - | - |
| MultiVSL1000 (Dinh et al., 2025) | 20161 | 4538 | 3712 | 999 | 999 | 999 | - | - | - |
| PHOENIX14T (Camgoz et al., 2018) | 7096 | 642 | 519 | 1085 | 411 | 393 | 2143 | 976 | 927 |
| CSL-Daily (Zhou et al., 2021a) | 18401 | 1176 | 1077 | 2000 | 1345 | 1358 | 2342 | 1358 | 1344 |
| How2Sign (Duarte et al., 2021) | 30904 | 2322 | 1713 | - | - | - | 8816 | 3312 | 3046 |
| CSL-OpenWorld (Ours) | 86387 | 7038 | 7033 | - | - | - | 12382 | 9134 | 9153 |

modal MoE models, MA-MoE differs in two key aspects. (1) MA-MoE is specifically designed to address the modality heterogeneity in sign language understanding, which has been largely underexplored in prior SLU works. (2) MA-MoE maintains a consistent architecture during both training and inference, whereas models such as VLMo and BEiT-3 require task-specific adaptations to the inference structure for different downstream tasks.

**Pose estimation.** Moreover, pose-based representations enable significantly more efficient training and inference, making them particularly suitable for large-scale scenarios. Considering limited GPU resources and the trade-off between training efficiency and model performance, we adopt pose data as the primary input modality instead of raw RGB video. Prior SL models, such as UniSign, CoSign, TwoStream, and MSLU, have demonstrated that pose-based inputs remain effective and robust across multiple benchmarks, suggesting that pose estimation errors do not substantially hinder overall performance in SL tasks. Nevertheless, pose estimation methods (DWPose) may still introduce noise, which can potentially affect model performance.

# C Experiment

**Dataset.** We adopt seven publicly available datasets, including three ISLR datasets (ASL Citizen, WLASL2000, and MultiVSL1000), two CSLR datasets, which also serve as gloss-based SLT datasets (CSL-Daily and PHOENIX14T), and two gloss-free SLT datasets (How2Sign and our newly proposed CSL-OpenWorld). Table 12 summarizes detailed statistics of all datasets used in this paper, including training set size and vocabulary size.

- **WLASL2000** (Li et al., 2020a) is the largest public Word-Level ASL dataset. It contains 21,083 video clips performing 2,000 common ASL glosses, featuring 119 signers to ensure variety in signing styles and backgrounds.
- **ASL Citizen** (Desai et al., 2023) is a large-scale, crowdsourced dataset designed for isolated sign language recognition in real-world settings. It comprises approximately 84,000 video clips of 2,731 signs performed by 52 signers, capturing significant diversity in lighting, camera angles, and backgrounds.
- **Multi-VSL1000** (Dinh et al., 2025) is a multi-view dataset for isolated Vietnamese Sign Language recognition, covering 1,000 sign categories. It provides synchronized observations from multiple viewpoints, enabling research on robust sign recognition under varying viewing conditions.
- **PHOENIX14T** (Camgoz et al., 2018) is a German Sign Language (DGS) dataset annotated with both glosses and German translations. It consists of 7,096 training, 519 validation, and 642 test samples from 9 signers, with a vocabulary of 1,066 glosses and 2,877 German words.
- **CSL-Daily** (Zhou et al., 2021a) is a Chinese Sign Language dataset focusing on daily life topics. It contains 18,401 labeled videos for training from 10 signers, providing 2,000 gloss tokens and 2,343 Chinese words for translation tasks.
- **How2Sign** (Duarte et al., 2021) is a large-scale American Sign Language (ASL) dataset comprising over 80 hours of multi-view video and multimodal data. We utilize the RGB frontal-view videos, which include 31,128 training, 1,741 validation, and 2,322 test samples.

Table 13: Effect of Dataset and Task descriptor.

| Dataset | Task | Dev | | | | TEST | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | WER | Rouge | BLEU1 | BLEU4 | WER | Rouge | BLEU1 | BLEU4 |
| | | 20.1 | 51.58 | 52.12 | 26.01 | 20.2 | 51.59 | 52.01 | 25.64 |
| ✓ | | 20.0 | 52.24 | 53.14 | 26.76 | 20.2 | 51.87 | 53.16 | 26.31 |
| | ✓ | 20.0 | 52.13 | 53.35 | 26.65 | 20.1 | 51.64 | 53.07 | 26.14 |
| ✓ | ✓ | 20.0 | 52.34 | 53.41 | 26.98 | 20.1 | 52.01 | 53.36 | 26.45 |

- **CSL-OpenWorld** is our proposed large-scale CSL dataset, collected from diverse online platforms. It includes 86,386 training, 7,033 validation, and 7,038 test samples, featuring over 10K video instances that reflect complex real-world signing scenarios.

*About dataset, although MSASL is a widely used benchmark for ISLR, we were unable to include it in our experiments. During data acquisition, we found that a substantial portion of the YouTube links provided by MSASL were no longer valid or accessible, making it infeasible to reconstruct the dataset reliably. OpenASL contains samples with multiple signers in a single video, which makes it unsuitable for our pose-based modeling. Further processing is required; we plan to incorporate OpenASL results in future work. Therefore, MSASL and OpenASL is excluded from our experimental evaluation.*

**Although our experiments involve multiple sign language datasets, they are independently collected and constructed from distinct data sources. We confirm that there is no sample- or video-level overlap across the datasets used in our experiments. Consequently, joint training on these datasets does not introduce cross-dataset leakage through duplicated samples or videos.**

**Implementation Details.** Our architecture consists of following key components.

**(1)** ***Pose Extraction:*** We employ the open-source DWPose toolkit to extract pose features from videos, including 21 keypoints for each hand, 68 facial keypoints, and 14 body joint keypoints.

**(2)** ***MA-LLM MoE:*** We adopt LLaMA 3.2 (1B) as the backbone of our modality-adaptive LLM with a Mixture-of-Experts (MoE) design. As the default setting, we use two pose experts with one activated expert and two text experts with one activated expert. The dataset descriptor indicates the sign language video dataset, while the task descriptor specifies the corresponding task (*i.e.*, Isolated Sign Language Recognition, Continuous Sign Language Recognition, and Sign Language Translation).

**(3)** ***Unified Paradigm:*** Following prior work (Gan et al., 2026d), pseudo-gloss annotations are constructed by removing punctuation, performing lemmatization, and applying word-level tokenization to the text transcripts. In all experiments, the loss weights $\lambda_{\text{sign}}$ and $\lambda_{\text{text}}$ are both set to 0.5.

**(4)** ***Optimization:*** Our model is implemented in PyTorch 2.6 and trained on four NVIDIA RTX 4090 GPUs (24GB memory each) using half-precision training. The initial learning rate is set to $1 \times 10^{-4}$ for the recognition module and $1 \times 10^{-5}$ for the translation module, with a decay factor of 0.5 applied at epochs 20, 30, and 40. The batch size is set to 8. Notably, the model is trained without incorporating any additional external sign language datasets beyond those described above.

**Evaluation Metrics.** For ISLR, to ensure fair comparison with prior work and their reported baselines, we follow the evaluation protocols used in each dataset. Specifically, we report Top-1 and Top-5 accuracy (Top@1 and Top@5) on ASL Citizen and MultiVSL1000, and per-instance (P-I) and per-class (P-C) Top-1 accuracy on WLASL2000. For CSLR, we use Word Error Rate (WER) as the evaluation metric. For SLT, we adopt ROUGE-L F1 and BLEU-1/2/3/4 to evaluate translation quality. These metrics are widely used in existing CSLR and SLT literature, enabling direct and fair comparison with prior methods.

## D More Ablation Study

Table 14: Phoenix14T results with $\lambda$ in $\mathcal{L}_{\text{bal}}$ .

| $\lambda_{sign} = \lambda_{text}$ | 0.3 | 0.5 | 0.7 | 0.9 | 1 |
|---|---|---|---|---|---|
| WER↓ | 22.1 | 20.1 | 22.3 | 23.4 | 25.1 |
| BLUE4 ↑ | 25.75 | 26.45 | 25.74 | 25.15 | 24.78 |

Table 15: Comparison of different experts.

| Model | Parameter(M) | | Dev | | | | TEST | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Total | Activated | WER | Rouge | BLEU1 | BLEU4 | WER | Rouge | BLEU1 | BLEU4 |
| Dense FFN | 2041.1 | 2041.1 | 22.6 | 49.63 | 50.23 | 23.68 | 22.8 | 49.56 | 49.32 | 23.67 |
| Sparse MOE | 4457.0 | 2041.1 | 22.4 | 49.23 | 49.36 | 23.15 | 22.6 | 48.65 | 49.36 | 22.64 |
| shared+modality-specific FFN | 2041.1 | 2041.1 | 21.4 | 51.69 | 51.60 | 25.68 | 21.5 | 51.52 | 51.98 | 25.79 |
| MA-MOE | 4457.0 | 2041.1 | 20.0 | 52.34 | 53.41 | 26.98 | 20.1 | 52.01 | 53.36 | 26.45 |

Table 16: Effect of MA-MOE on different LLMs.

| Model | Test | | | | +MA-MOE | TEST | | | |
|---|---|---|---|---|---|---|---|---|---|
| | WER | Rouge | BLEU1 | BLEU4 | | WER | Rouge | BLEU1 | BLEU4 |
| Llama 3.2 1B | 22.5 | 48.99 | 49.72 | 23.01 | → | 20.1 | 52.01 | 53.36 | 26.45 |
| Gemma3 1b | 22.9 | 49.96 | 48.04 | 22.85 | → | 20.2 | 51.32 | 52.66 | 26.35 |
| Qwen2.5 1.5B | 23.7 | 49.06 | 48.81 | 22.93 | → | 20.3 | 51.17 | 52.69 | 26.34 |

**Effect of the Dataset and Task descriptor.** As illustrated in Fig. 5, we introduce dataset and task descriptors in the MA-MoE module to provide auxiliary conditioning during both training and inference. These descriptors are not used to create separate branches, but rather to guide the model toward the corresponding data distributions in a unified manner.

To evaluate their impact, we conduct an ablation study by selectively removing the dataset and/or task descriptors. The results are summarized in Table 13. We observe that removing either descriptor individually leads to only marginal performance degradation, indicating that the model remains robust and effective even without explicit conditioning signals. Nevertheless, incorporating both descriptors yields the best overall performance, suggesting that they provide complementary information that helps the model better capture dataset-specific and task-specific variations. Importantly, these results demonstrate that our framework does not rely on explicit descriptors to maintain unified modeling capability, while still benefiting from them when available.

**Hyperparameter Analysis of the Load Balancing Coefficient.** We further investigate the sensitivity of the load-balancing coefficient $\lambda$ in Eq. 13. As shown in Table 14, the best performance is achieved when $\lambda = 0.5$, yielding the lowest WER (21.8) and the highest BLEU-4 (26.45). When $\lambda$ is relatively small (e.g., 0.3), the load-balancing constraint is insufficient to effectively prevent expert imbalance, leading to slightly inferior performance. As $\lambda$ increases beyond 0.5, both WER and BLEU-4 gradually deteriorate. This suggests that an excessively strong balancing constraint forces the router to distribute tokens more uniformly across experts, limiting expert specialization and reducing the routing flexibility required for modeling modality-specific patterns. Overall, these results indicate that a moderate load-balancing coefficient provides the best trade-off between balanced expert utilization and expert specialization.

**Comparison of different expert designs** To validate the effectiveness of the proposed sparse MA-MoE, we investigate whether the performance gains mainly stem from the MoE routing mechanism or simply from increased parameter capacity. Specifically, we compare MA-MoE with two alternative settings: (1) **Shared + modality-specific FFN**, which consists of a shared expert along with one sign expert and one text expert, but without sparse routing; (2) **Dense FFN**, where three experts are always activated without modality-specific specialization. (3) **Sparse MoE**, which consists of five experts with Top-$k = 3$ activated for each token, but without modality-specific specialization.

As shown in Table 15, under comparable parameter scales, the Dense FFN achieves the worst performance. This indicates that merely increasing the number of parameters does not lead to performance improvement. In contrast, the proposed sparse modality-specific experts effectively capture modality heterogeneity, demonstrating that the performance gains primarily come from the adaptive routing mechanism rather than parameter scaling.

Table 17: Impact of activated expert count.

| $N_s = N_t$ | Top-$k$ | Params (M) | Activated (M) | Rouge | BLEU1 | BLEU4 | Rouge | BLEU1 | BLEU4 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1235.8 | 1235.8 | 49.45 | 49.36 | 22.67 | 48.99 | 49.72 | 23.01 |
| 2 | 1 | 4457.0 | 2041.1 | 52.34 | 53.41 | 26.98 | 52.01 | 53.36 | 26.45 |
| 4 | 1 | 7678.2 | 2041.1 | 52.74 | 53.65 | 26.78 | 52.14 | 53.24 | 26.53 |
| 4 | 2 | 7678.2 | 2846.4 | 52.99 | 53.94 | 26.97 | 52.45 | 53.47 | 26.56 |

Table 18: CSLR Qualitative results on Phoenix-2014T and CSL-Daily.

| example(a) | Phoenix14T dataset |
|---|---|
| Groundtruth | HEUTE NACHT REGEN DOCH MITTE NORD<br>(TODAY NIGHT RAIN HOWEVER MIDDLE NORTH) |
| OneSign | HEUTE NACHT REGEN DOCH MITTE NORD<br>TODAY NIGHT RAIN HOWEVER MIDDLE NORTH |
| W/O MA-MOE | HEUTE NACHT NEBEL DOCH MITTE NORD<br>( TODAY NIGHT FOG HOWEVER MIDDLE NORTH ) |
| example(b) | Phoenix14T dataset |
| Groundtruth | TAG OST VIER FLUSS SIEBEN GRAD<br>(DAY EAST FOUR RIVER SEVEN DEGREE) |
| OneSign | TAG DANN VIER FLUSS SIEBEN GRAD<br>(DAY THEN FOUR RIVER SEVEN DEGREE) |
| W/O MA-MOE | TAG DANN VIER FLUSS SIEBEN<br>( DAY THEN FOUR RIVER SEVEN) |
| example(c) | CSL-daily dataset |
| Groundtruth | 去 学校 我 踢足球<br>(Go School I Play soccer) |
| OneSign | 去 学校 我 踢足球<br>Go School I Play soccer |
| W/O MA-MOE | 来 学校 我 踢足球<br>Come School I Play soccer |
| example(d) | CSL-daily dataset |
| Groundtruth | 我 房子 在 看 电视 你 看 电视 喜欢<br>(I House At Wathc TV You Watch Like) |
| OneSign | 我 房子 看 电视 你 看 电视 喜欢<br>I House Wathc TV You Watch Like |
| W/O MA-MOE | 我 房子 在 看 电视 我 看 电视<br>I House At Watch TV I Watch TV |

**Effect of MA-MoE on different LLM models.** We further evaluate the generalization capability of MA-MoE across different backbone LLMs. Specifically, we conduct experiments on several representative open-source models, including LLaMA3-1B (our baseline), Gemma-3-1B, and Qwen-2.5-1.5B. As shown in Table 16, MA-MoE consistently improves performance across all evaluated models. These results demonstrate that our approach is model-agnostic and can be seamlessly integrated into a variety of LLM architectures. Moreover, the consistent gains across different backbones indicate that MA-MoE effectively addresses modality heterogeneity in a generalizable manner, rather than relying on model-specific inductive biases, leading to stable and significant improvements.

**Full ablation on the Impact of Activated Expert Count.** Due to space limitations in the main paper, we only reported the ablation results on the DEV set in Table 3 of main paper. Here, we provide the complete results on both DEV and TEST sets in Table 17.

## E Qualitative Analysis

As shown in Table 18 and Table 19, We provide qualitative analyses for both recognition and translation tasks. Specifically, we present recognition results on PHOENIX-2014T and CSL-Daily, and translation results on PHOENIX-2014T, How2Sign, CSL-Daily, and CSL-OpenWorld. As observed, the model without MA-MoE already achieves reasonable recognition and translation performance. With MA-MoE, OneSign consistently produces more accurate and semantically coherent outputs across datasets, further demonstrating the effectiveness of our approach.

Table 19: SLT Qualitative results on Phoenix-2014T, How2Sign , CSL-Daily and CSL-OpenWorld.

| example(a) | Phoenix14T dataset |
|---|---|
| Groundtruth | und nun die wettervorhersage für morgen sonntag den zwölften juli<br>(And now the weather forecast for tomorrow, Sunday, July 11th) |
| OneSign | und nun die wettervorhersage für morgen sonntag den zwölften juli<br>(And now the weather forecast for tomorrow, Sunday, July 11th) |
| W/O MA-MOE | und nun die wettervorhersage für morgen sonntag den elften juli<br>(And now the weather forecast for tomorrow, Sunday, July 12th) |
| example(b) | How2Sign dataset |
| Groundtruth | So, we've got to find a way to get to the exit. |
| OneSign | we've got to find a way to get to the exit. |
| W/O MA-MOE | we have found a way to get to the exit. |
| example(b) | CSL-daily dataset |
| Groundtruth | 人生旅途, 多一个人相伴, 是一种幸福。<br>(On the journey of life, having someone by your side is a form of happiness.) |
| OneSign | 人生旅途, 多个人相伴, 是一种幸福。<br>On the journey of life, having others by your side is a form of happiness. |
| W/O MA-MOE | 人生旅途, 相伴是一种幸福。<br>On the journey of life, companionship is a blessing. |
| example(d) | CSL-OpenWorld dataset |
| Groundtruth | 文化遗产是全人类的共同财富。<br>(Cultural heritage is the common wealth of all humanity.) |
| OneSign | 文化遗产是人类的共同财富。<br>Cultural heritage is the common wealth of humanity. |
| W/O MA-MOE | 文化是全人类的共同财富。<br>Culture is the common wealth of all humanity. |

## F Dataset Visualization

We provide visualizations of the formatted pose data used in our experiments across seven datasets. As shown in Figures 6 and 7, the pose representations exhibit noticeable variations in skeleton structure, motion patterns, and temporal dynamics across different datasets. In addition, we present representative video samples from the proposed CSL-OpenWorld dataset, as illustrated in Figure 8, to provide an intuitive understanding of its visual diversity and real-world complexity.

## G Limitations and Discussion

We outline potential directions for future improvement.

**Heterogeneous datasets training.** While OneSign explicitly models modality-level distribution differences, it does not yet account for distributional discrepancies across heterogeneous datasets. When multiple datasets are jointly used for training, the original data distribution of a specific dataset may be diluted, which can adversely affect performance on its corresponding test set. Consequently, how to effectively balance multi-task and multi-dataset aggregation while preserving dataset-specific characteristics remains an important direction for future research.

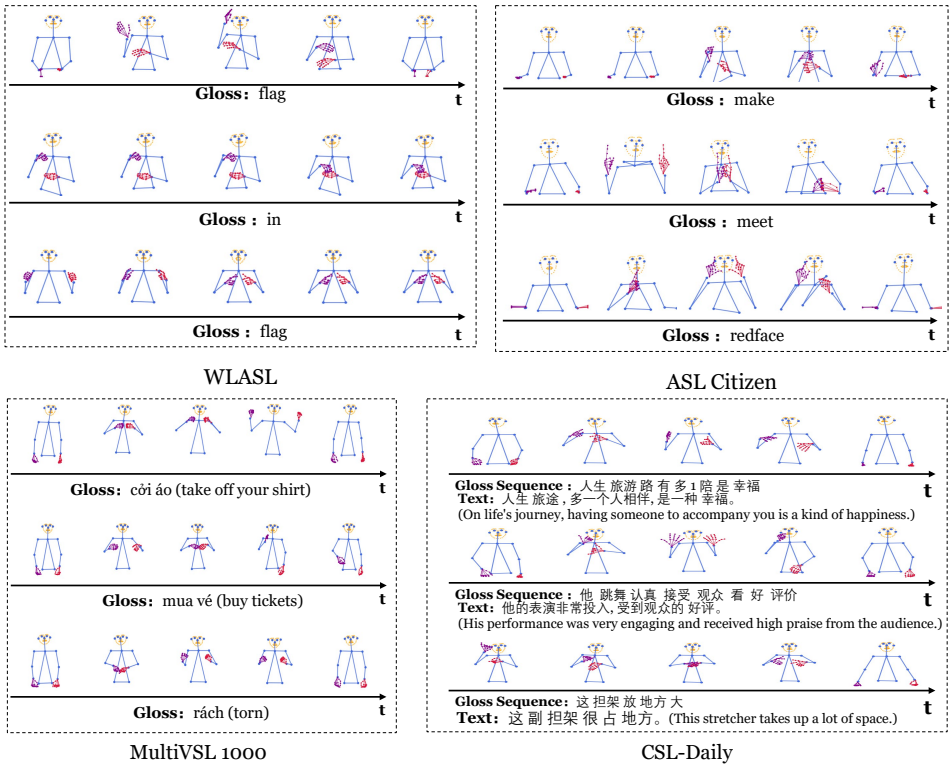


Figure 6: Visualization of used datasets.

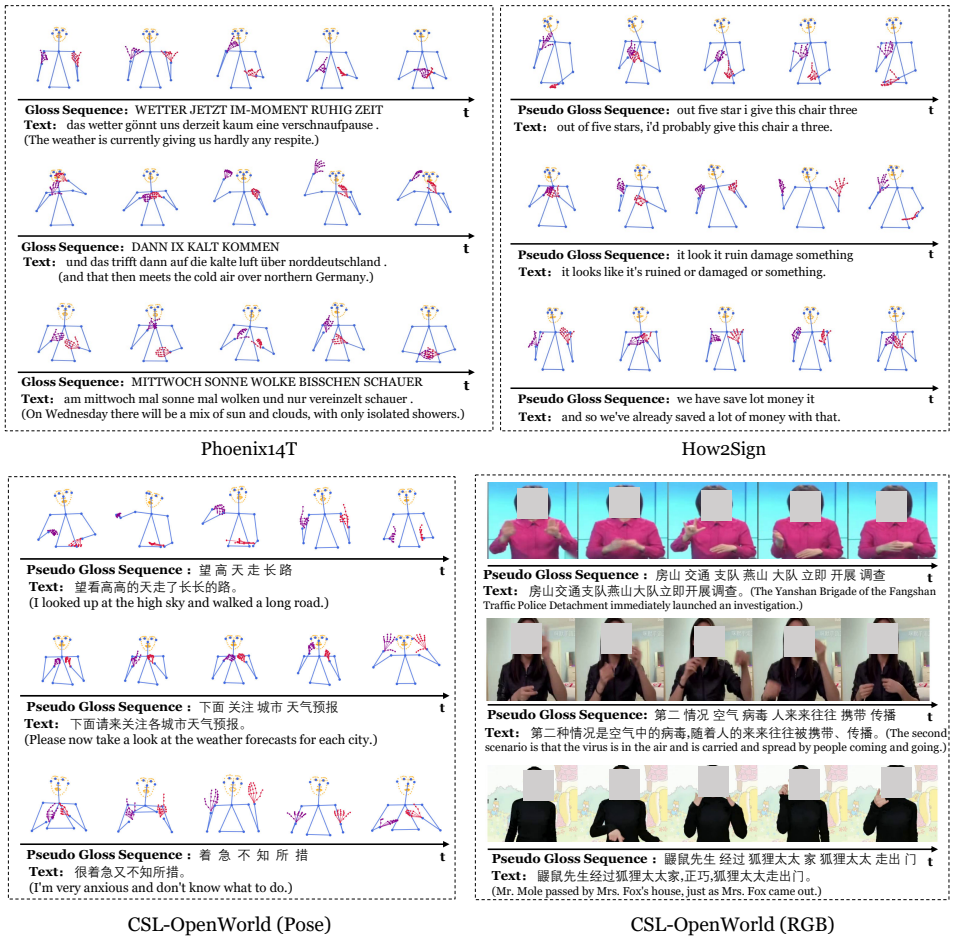


Figure 7: Visualization of used datasets. We blurred the facial areas for privacy.

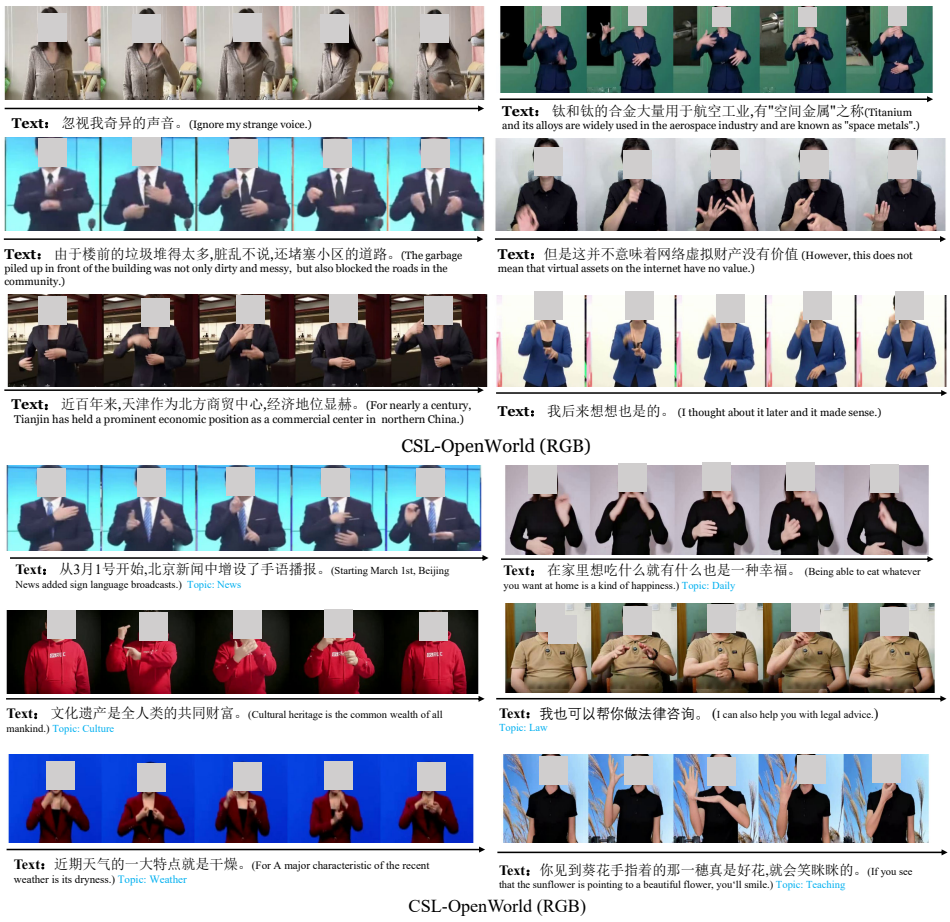


Figure 8: More visualizations of the proposed CSL-OpenWorld dataset, illustrating its diversity in backgrounds, signers, and topics. We blurred the facial areas for privacy.