Where Motion Matters: Conditioning the Pre-Contextualization Interface for CTC-Based Sequence Learning

Anonymous ACL submission

Abstract

In CTC-based sequence recognition, representations must transition from locally-encoded frame features to globally-contextualized sequences—yet the interface where this transformation occurs remains underexplored. We show that in CTC-based sequence learning, the pre-contextualization interface is a critical bottleneck, and that conditioning representations at this interface reduces alignment errors. We study this interface in Continuous Sign Language Recognition (CSLR), where we find that *conditioning* feature transformation on motion cues—rather than simply adding motion features—reduces alignment errors. We propose **MoRE** (Motion-conditioned Representation Enhancement), a lightweight module that uses motion-derived gates to interpolate between two learned projections of visual features before sequence modeling. Controlled ablations on PHOENIX-2014 isolate three key findings: (1) placement at the pre-contextualization interface is critical—post-contextualization placement *degrades* performance below baseline; (2) learned gating outperforms fixed alternatives; and (3) MoRE primarily reduces deletion errors, the dominant CTC failure mode. We show that where motion is applied—at the pre-contextualization interface—matters more than how it is incorporated under CTC supervision. We observe consistent improvements on PHOENIX-2014 and mixed results on CSL-Daily, suggesting dataset-specific factors influence effectiveness.

1 Introduction

Continuous Sign Language Recognition (CSLR) transcribes an untrimmed video of signing into a sequence of glosses without frame-level temporal annotations (Koller et al., 2015; Koller, 2020). Trained with Connectionist Temporal Classification (CTC) (Graves et al., 2006), CSLR systems must implicitly discover the mapping between visual input and linguistic output under weak supervision. This remains challenging due to signer variability, coarticulation between consecutive signs, and transitional motion that carries no semantic content (Adaloglou et al., 2022).

The dominant pipeline consists of four stages: a visual backbone for per-frame spatial features, temporal convolutions for local patterns, a bidirectional LSTM for sequence-level context, and CTC decoding (Cui et al., 2017; Min et al., 2021; Guo et al., 2023). Recent advances have pursued stronger representations through multi-cue fusion (Zhou et al., 2022; Hu et al., 2023; Jiao et al., 2023), cross-lingual transfer (Wei and Chen, 2023), and refined training strategies (Min et al., 2021; Hao et al., 2021; Guo et al., 2023). However, one design choice remains under-explored: how frame-level features are prepared *before* temporal contextualization.

We observe that the interface between local temporal encoding (TCN) and global sequence modeling (BiLSTM)—which we term the *pre-contextualization interface*—represents a critical bottleneck. Prior work focuses primarily on *what* features to extract or *how* to train. We focus instead on *where* representations are transformed and *how* that transformation is conditioned.

Motion patterns provide a natural conditioning signal: frames with different motion characteristics may benefit from different feature transformations. Yet motion's use in CSLR has been limited to auxiliary input streams (Simonyan and Zisserman, 2014) or additive feature injection (Zheng et al., 2023). We hypothesize that motion is most valuable when it determines *how* features are transformed, not merely *what* features are added.

We propose **MoRE** (**Mo**tion-conditioned **R**epresentation **E**nhancement), a lightweight module inserted at the pre-contextualization interface. MoRE conditions on local motion cues to produce per-dimension mixing coefficients that interpolate between two learned feature projections. Unlike

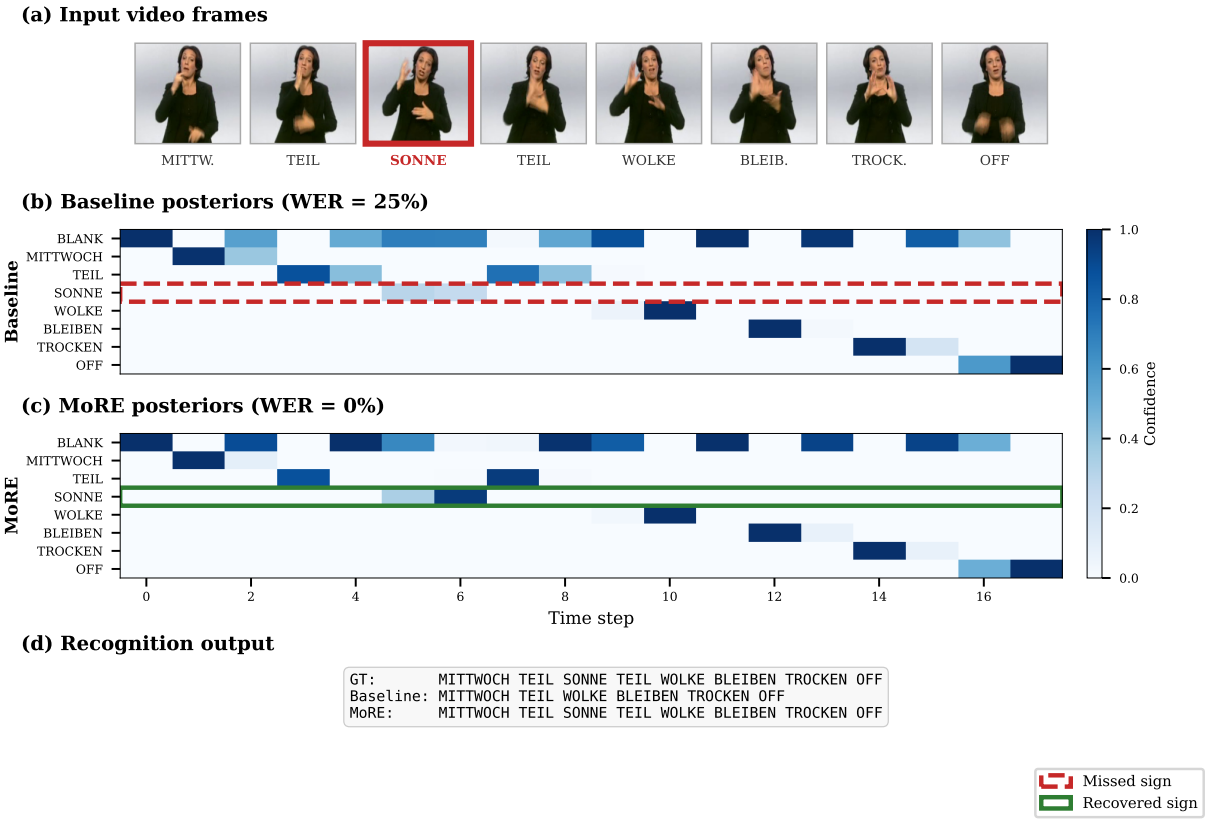

Figure 1: Motivating example from PHOENIX-2014 development set. **(a)** Representative video frames with corresponding sign glosses. The sign SONNE (highlighted in red) is correctly performed by the signer. **(b)** CTC posterior probabilities from our baseline model (ResNet-18 + TCN + BiLSTM; see Section 5.1) show weak activation for SONNE, resulting in its deletion during decoding (WER = 25%). **(c)** With MoRE, the model produces stronger, more localized posteriors for SONNE, enabling correct recognition (WER = 0%). **(d)** Recognition outputs confirm that the baseline skips "SONNE TEIL" while MoRE recovers the complete sequence.

prior work that injects motion additively (Zheng et al., 2023) or requires architectural duplication (Ahn et al., 2024), MoRE provides learned, selective enhancement as a drop-in module.

We evaluate MoRE on PHOENIX-2014 (Koller et al., 2015) and CSL-Daily (Zhou et al., 2021), focusing on diagnostic evidence rather than state-of-the-art claims. Controlled ablations isolate when and why motion-conditioned enhancement helps under pure CTC supervision.

Contributions.

- We identify the **pre-contextualization interface**—between local temporal modeling and global sequence modeling—as a critical bottleneck in CTC-based sequence learning.
- We show that **conditioning** representation transformation (not content) at this interface reduces deletion errors—the dominant CTC failure mode.
- We provide **diagnostic evidence** explaining where and why motion cues help, including controlled ablations that isolate placement, gating behavior, and motion conditioning.

2 Related Work

Continuous Sign Language Recognition. CSLR has evolved from HMM-based approaches (Koller et al., 2015) to end-to-end neural architectures trained with CTC (Graves et al., 2006; Cui et al., 2017; Pu et al., 2019). Recent work has systematically characterized deep learning approaches for sign language processing (Toshpulatov et al., 2025). Advances include multi-cue integration (Zhou et al., 2022; Hu et al., 2023; Jiao et al., 2023), knowledge distillation (Hao et al., 2021; Min et al., 2021; Guo et al., 2023), and cross-lingual transfer (Wei and Chen, 2023; Chen et al., 2022). Transformer-based architectures have also emerged (Camgöz et al., 2020; Müller et al., 2022), though hybrid designs remain competitive

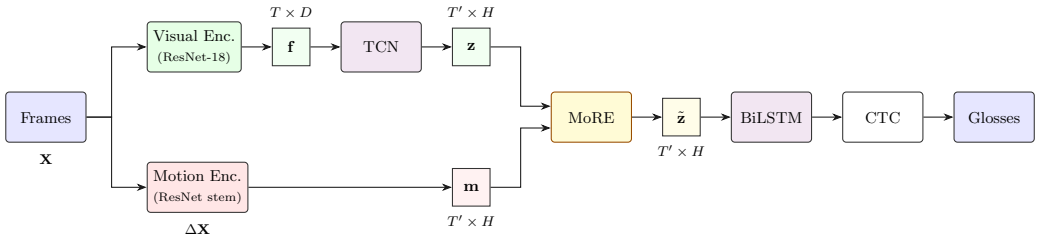

Figure 2: Our CSLR pipeline with MoRE. The ResNet-18 visual encoder extracts per-frame features $\mathbf{f}$ (dimension D=512), which are temporally downsampled and projected by the TCN to produce $\mathbf{z}$ (dimension H, reduced length T'). A parallel motion encoder processes frame differences $\Delta\mathbf{X}$ and produces motion features $\mathbf{m}$ at matching resolution. MoRE combines $\mathbf{z}$ and $\mathbf{m}$ at the pre-contextualization interface, producing enhanced representations $\tilde{\mathbf{z}}$ for the BiLSTM and CTC decoding.

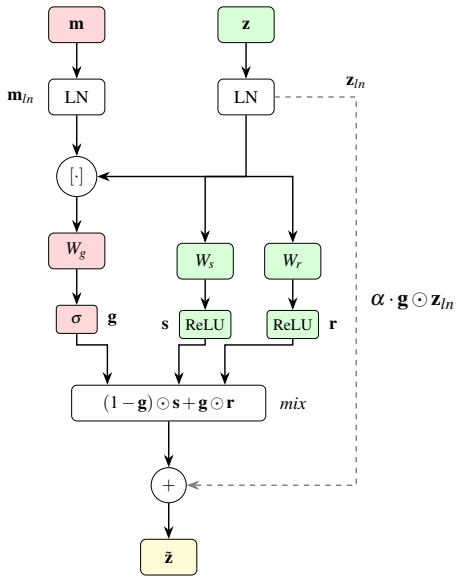

Figure 3: MoRE module. Visual features $\mathbf{z}$ and motion features $\mathbf{m}$ are normalized and concatenated to compute gate $\mathbf{g}$ (Eq. 1). The gate interpolates between projections $\mathbf{s}$, $\mathbf{r}$ (Eqs. 2–3). The dashed line indicates the residual connection (Eq. 4), modulated by the gate.

(Zuo et al., 2023). These approaches primarily target the backbone or training procedure rather than the interface between local and global modeling.

Motion Modeling. Two-stream networks (Simonyan and Zisserman, 2014) process RGB and optical flow in parallel, while 3D convolutions (Tran et al., 2015; Carreira and Zisserman, 2017) capture spatiotemporal patterns jointly. Motion carries particular significance for sign language, encoding hand trajectory and movement dynamics (Stokoe, 2005; Adaloglou et al., 2022). Temporal dynamics have been captured through multi-modal fusion (Toshpulatov et al., 2024), auxiliary flow streams (Cui et al., 2019), and SlowFast architectures (Feichtenhofer et al., 2019; Ahn et al., 2024). Our approach differs: we use motion as a *conditioning signal* rather than an additional input stream.

Feature Modulation. Modulating features based on auxiliary signals has proven effective across domains. LSTMs (Hochreiter and Schmidhuber, 1997) and GRUs (Cho et al., 2014) use learned coefficients for memory updates. Squeeze-and-Excitation networks (Hu et al., 2018) recalibrate channel responses, while FiLM (Perez et al., 2018) enables feature-wise transformations. Graph-based architectures have shown similar benefits (Safarov et al., 2025). MoRE targets a specific interface: the boundary between local extraction and global sequence modeling in CTC-based systems.

CTC Alignment. CTC enables training without frame-level supervision (Graves et al., 2006), but its conditional independence assumption creates challenges (Graves, 2012). Alignment behavior has been studied in speech recognition (Zeyer et al., 2017; Liu et al., 2018). Prior work addresses CTC limitations through auxiliary losses (Min et al., 2021; Hao et al., 2021) and iterative refinement (Pu et al., 2019). Our work is complementary: we enhance frame representations *before* sequence modeling.

3 Background

3.1 Problem Formulation

Given a video $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ of T frames, CSLR aims to predict a gloss sequence $\mathbf{Y} = (y_1, \ldots, y_N)$ where $N \ll T$ and no frame-level alignment is provided. Each gloss y_n belongs to a vocabulary $\mathcal{V}$.

Configuration	WER (%)		Δ Relative (%)	
	Dev	Test	Dev	Test
Main Results (PHOENIX-2014, mean ± std over 3 seeds)				
ResNet-18 Baseline	$19.17_{\pm.07}$	$20.21_{\pm.19}$	—	—
+ MoRE	$\mathbf{18.65}_{\pm.07}$	$\mathbf{19.58}_{\pm.22}$	−2.7	−3.1
ResNet-34 Baseline	18.84	19.81	—	—
+ MoRE	**18.50**	**19.50**	−1.8	−1.6
Main Results (CSL-Daily)				
ResNet-18 Baseline	28.30	27.10	—	—
+ MoRE	**27.66**	27.12	−2.3	+0.1
Ablation (a): Placement (PHOENIX-2014)				
Pre-BiLSTM (default)	**18.59**	**19.24**	−3.0	−4.8
Post-BiLSTM	20.41	20.95	+6.5	+3.7
Ablation (b): Gate Behavior (PHOENIX-2014)				
Learned gate (default)	**18.59**	**19.24**	−3.0	−4.8
Fixed $\mathbf{g}$=0 (always $\mathbf{s}$)	19.15	19.79	−0.1	−2.1
Fixed $\mathbf{g}$=1 (always $\mathbf{r}$)	19.37	20.14	+1.0	−0.3
Ablation (c): Motion Conditioning (PHOENIX-2014)				
With motion (default)	**18.59**	**19.24**	−3.0	−4.8
Without motion ($\mathbf{m}$=$\mathbf{0}$)	19.40	19.79	+1.2	−2.1
Ablation (d): Residual Warm-Start (PHOENIX-2014)				
With residual ($\alpha \to 0.2$, default)	**18.59**	**19.24**	−3.0	−4.8
Without residual (α=0)	19.28	19.98	+0.6	−1.1
Ablation (e): Motion Fusion Strategy (PHOENIX-2014)				
Concatenation (default, Eq. 1)	**18.59**	**19.24**	−3.0	−4.8
Additive fusion (Eq. 5)	18.37	19.68	−4.2	−2.6

Table 1: Main results and ablations. Top: main results on PHOENIX-2014 (3 seeds for ResNet-18) and CSL-Daily. Bottom: ablations on PHOENIX-2014 (single seed). Δ Relative shows percentage change vs. baseline; negative = improvement. Key findings: (a) post-BiLSTM *degrades* below baseline; (b) learned gating outperforms fixed; (c) motion conditioning helps beyond mixing alone; (d) residual warm-start stabilizes training; (e) concatenation generalizes better than additive fusion.

3.2 CSLR Pipeline

We adopt the dominant architecture processing video through four stages:

Visual Backbone. ResNet-18 extracts per-frame spatial features $\mathbf{f}_t = \text{ResNet-18}(\mathbf{x}_t) \in R^D$ where $D = 512$.

Temporal Convolutions. A TCN captures local temporal patterns with downsampling: $\mathbf{z} = \text{TCN}(\mathbf{f}) \in R^{T' \times H}$ where $T' = T/4$.

Sequence Model. A BiLSTM aggregates global context: $\mathbf{s}_{t'} = \text{BiLSTM}(\mathbf{z})_{t'} \in R^{2H'}$.

CTC Decoding. A linear classifier produces posteriors over the vocabulary augmented with a blank token.

3.3 The Pre-Contextualization Interface

We identify the TCN-to-BiLSTM transition as a critical interface. At this point, features $\mathbf{z}$ have captured local patterns but lack global context. When frame representations fail to capture sign-discriminative information, CTC posteriors become diffuse, leading to deletion errors.

This interface presents a unique opportunity for intervention. Before the BiLSTM, representations encode *what* is happening locally but not *how* it

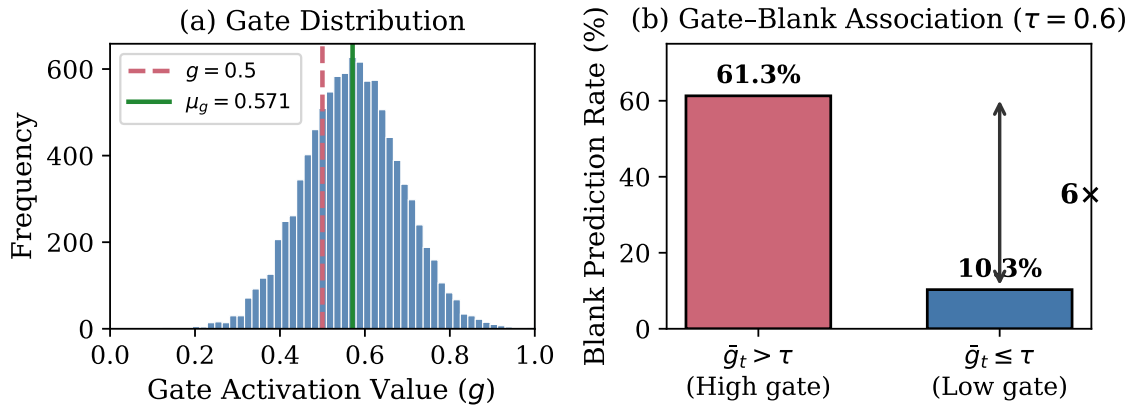

Figure 4: Learned gate diagnostics on PHOENIX-2014 test set. **(a)** Gate activation distribution centers around μ_g=0.571 with meaningful variance (σ_g=0.114), confirming non-trivial interpolation. **(b)** Frames with high gate activation ($\bar{g}_t > \tau$) produce blank predictions 6× more often than low-gate frames, suggesting the gate identifies transitional motion.

Method	Del	Ins	Sub	Total
PHOENIX-2014 Dev (mean over 3 seeds)				
ResNet-18	6.76	2.26	10.15	19.17
+ MoRE	6.25	2.20	10.19	18.65
Δ (rel.)	**−7.5%**	−2.7%	+0.4%	−2.7%
PHOENIX-2014 Test (mean over 3 seeds)				
ResNet-18	7.12	2.41	10.68	20.21
+ MoRE	6.58	2.35	10.65	19.58
Δ (rel.)	**−7.6%**	−2.5%	−0.3%	−3.1%

Table 2: Error breakdown (%) on PHOENIX-2014. MoRE primarily reduces deletion errors.

relates to the full sequence. After the BiLSTM, representations have been globally contextualized—local modifications at this stage risk conflicting with established sequential dependencies. The pre-contextualization interface thus represents the last opportunity to enhance frame-level representations before they are integrated into the global sequence model.

Motion information—captured through frame differences—provides a natural conditioning signal at this interface. Frames with high motion typically correspond to sign transitions or dynamic gesture phases, while low-motion frames may indicate holds or preparation phases. By conditioning enhancement on local motion, MoRE allows the model to learn when and how to modulate representations based on the local temporal dynamics.

4 Method

We introduce MoRE (**Mo**tion-conditioned **R**epresentation **E**nhancement), a lightweight module that conditions visual feature transformation on motion cues. Figure 2 illustrates the architecture.

4.1 Visual Feature Extraction

Given input video frames $\mathbf{X} \in R^{T \times H_{\text{img}} \times W_{\text{img}} \times 3}$, a ResNet-18 backbone extracts per-frame features $\mathbf{f} = \text{ResNet-18}(\mathbf{X}) \in R^{T \times D}$ where $D = 512$. A TCN with temporal downsampling produces $\mathbf{z} = \text{TCN}(\mathbf{f}) \in R^{T' \times H}$ where $T' = T/4$.

4.2 Motion Feature Extraction

We extract motion features from temporal differences: $\Delta\mathbf{X}_t = \mathbf{X}_t - \mathbf{X}_{t-1}$ for $t = 2, \ldots, T$, with $\Delta\mathbf{X}_1 = \mathbf{0}$. A motion encoder processes these through early convolutional layers with projection to match post-TCN resolution: $\mathbf{m} = \text{Resample}(\text{MotionEnc}(\Delta\mathbf{X})) \in R^{T' \times H}$.

4.3 MoRE: Motion-Conditioned Representation Enhancement

MoRE combines visual and motion features through learned gating (Figure 3).

Normalization. Both streams are normalized: $\mathbf{z}_n = \text{LayerNorm}(\mathbf{z})$, $\mathbf{m}_n = \text{LayerNorm}(\mathbf{m})$.

Gate Computation. A per-dimension gate is computed by concatenating features and applying a linear projection with sigmoid:

$$\mathbf{g} = \sigma\big(\mathbf{W}_g[\mathbf{z}_n; \mathbf{m}_n] + \mathbf{b}_g\big) \in (0,1)^{T' \times H} \quad (1)$$

where $\mathbf{W}_g \in R^{H \times 2H}$. Motion features contribute *only* to gate computation—they determine how visual features are transformed but do not directly enter the output.

Dual Projections. Two linear transformations with ReLU are applied to visual features:

$$\mathbf{s} = \text{ReLU}(\mathbf{W}_s\mathbf{z}_n + \mathbf{b}_s), \quad \mathbf{r} = \text{ReLU}(\mathbf{W}_r\mathbf{z}_n + \mathbf{b}_r) \tag{2}$$

where $\mathbf{W}_s, \mathbf{W}_r \in R^{H\times H}$. Our ablations confirm $\mathbf{s}$ learns generally useful representations while $\mathbf{r}$ provides context-specific modulation.

Gated Mixing. The gate interpolates between projections:

$$\text{mix} = (1 - \mathbf{g}) \odot \mathbf{s} + \mathbf{g} \odot \mathbf{r} \tag{3}$$

Residual Connection. Enhanced features combine mixing with a gated residual:

$$\tilde{\mathbf{z}} = \text{mix} + \alpha \cdot \mathbf{g} \odot \mathbf{z}_n \tag{4}$$

where α is annealed during training (Section 5.1). Starting with $\alpha = 1.0$ prevents the randomly initialized module from disrupting early gradients; α decays to 0.2, allowing gated mixing to increasingly dominate while retaining a modest residual contribution.

4.4 Design Rationale

Positioning. We place MoRE after TCN but before BiLSTM. At this stage, features have undergone local aggregation but lack global context. Motion conditioning allows the BiLSTM to receive representations adapted to local dynamics.

Motion as Conditioning. Unlike methods concatenating motion directly (Pu et al., 2019), MoRE uses motion exclusively for gate computation, modulating transformation without introducing motion-specific content.

Parameter Overhead. MoRE adds **4.4M** parameters (∼**13**% of baseline): two projections ($\mathbf{W}_s, \mathbf{W}_r \in R^{H\times H}$), gate projection ($\mathbf{W}_g \in R^{H\times 2H}$) with H=1024, plus early ResNet layers for motion encoding. The motion encoder contributes only 0.2M parameters by reusing the first ResNet block, keeping the overhead modest relative to the full 33M-parameter pipeline.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on: (1) **PHOENIX-2014** (Koller et al., 2015), 6,841 German Sign Language sentences with 1,295 glosses (5,672/540/629 train/dev/test), and (2) **CSL-Daily** (Zhou et al., 2021), 20,654 Chinese Sign Language sentences with 2,000 glosses (18,401/1,077/1,176 train/dev/test).

Metric. Word Error Rate (WER) = $(S + D + I)/N$. Lower is better. We use greedy CTC decoding without language model.

Architecture. Baseline: ResNet-18 (D=512) → 2-layer TCN (stride 4, H=1024) → 2-layer BiLSTM (512 per direction) → CTC decoder. MoRE is inserted between TCN and BiLSTM.

Training. 80 epochs, Adam with lr 10^{-4}, step decay at epochs 20, 30, 35. Batch size 4, gradient clipping 5.0. MoRE uses α-annealing: α decays from 1.0 to 0.2 over 8 epochs.

5.2 Main Results and Ablations

Table 1 presents results and ablations on PHOENIX-2014.

PHOENIX-2014. MoRE reduces WER by 0.63% absolute on test (3.1% relative), with consistent improvements across seeds. With ResNet-34 backbone, MoRE similarly improves from 19.81% to 19.50% (−1.6% relative), confirming gains transfer to stronger backbones.

CSL-Daily. MoRE improves dev WER by 0.64% but shows negligible test change (+0.02%). We explore this in Section 6.

(a) Placement. Pre-BiLSTM substantially outperforms post-BiLSTM (19.24% vs 20.95%). Post-BiLSTM *hurts* vs. baseline (20.95% vs 20.21%), confirming enhancement must occur before contextualization.

(b) Gate Behavior. Learned gate outperforms both fixed alternatives. Fixed $\mathbf{g}$=0 beats fixed $\mathbf{g}$=1, suggesting $\mathbf{s}$ captures generally useful representations.

(c) Motion Conditioning. Removing motion degrades test WER from 19.24% to 19.79%, confirming motion provides useful conditioning beyond visual features alone.

(d) Residual Warm-Start. Setting α=0 removes the gated residual term (Eq. 4), forcing the module to rely purely on gated mixing. This degrades test WER from 19.24% to 19.98%, indicating the residual warm-start stabilizes training and improves final alignment quality.

(e) Motion Fusion Strategy. We compare concatenation-based gating (Eq. 1) against an additive alternative:

$$\mathbf{g}_{\text{add}} = \sigma\big(\mathbf{W}'_g(\mathbf{z}_n + \beta \cdot \phi(\mathbf{m}_n)) + \mathbf{b}'_g\big) \quad (5)$$

where ϕ is a learned projection and β=1.0. While additive fusion achieves competitive dev WER (18.37%), it generalizes worse to test (19.68% vs 19.24%), suggesting concatenation provides more robust motion conditioning by preserving separate visual and motion representations.

5.3 Error Analysis

Table 2 decomposes errors by type. MoRE primarily reduces **deletion errors** (-7.5% on dev, -7.6% on test), with modest insertion reduction and negligible substitution change. This pattern is interpretable: deletions occur when CTC posteriors are too diffuse, defaulting to blank. MoRE produces sharper posteriors at frames that would otherwise be suppressed (Figure 1).

The selective reduction in deletions—without increasing substitutions—suggests MoRE enhances discriminability rather than simply amplifying all predictions. If MoRE merely increased posterior magnitudes uniformly, we would expect substitution errors to decrease proportionally or insertions to increase. Instead, the targeted deletion reduction indicates that motion conditioning specifically helps at frames where the baseline produces ambiguous posteriors, sharpening predictions toward the correct class rather than an incorrect alternative. Appendix C provides an additional qualitative example demonstrating this pattern.

5.4 Gate Behavior Analysis

To understand what the gate learns, we analyze activation statistics on PHOENIX-2014 test set at convergence. We report gate–blank association by thresholding frame-level mean gate activation $\bar{g}_t$ at τ=0.6 for analysis only; this threshold does not affect training.

Figure 4 reveals two key findings. First, the gate learns a non-trivial distribution (μ_g=0.571, σ_g=0.114) rather than collapsing to fixed values, indicating genuine interpolation between projections $\mathbf{s}$ and $\mathbf{r}$. Second, frames with high gate activation produce blank predictions at dramatically higher rates (61.3% vs 10.3%)—a 6$\times$ difference. This provides indirect evidence for the coarticulation hypothesis: high gate values appear at frames where the model is uncertain, often corresponding to transitional movements between signs. The gate learns this behavior without explicit supervision, emerging purely from CTC training.

6 Discussion

Why Does Placement Matter? The dramatic pre- vs. post-BiLSTM difference reveals where conditioning should occur. Post-contextualization, the BiLSTM has aggregated global information. Motion conditioning then introduces local signals conflicting with global context.

Design principle: Post-contextualization conditioning introduces local motion noise conflicting with established global semantic context. Conditioning signals should be applied *before* global context aggregation, not after.

Dataset Differences. MoRE shows consistent PHOENIX-2014 improvements but mixed CSL-Daily results. Possible factors include vocabulary size differences (1,295 vs 2,000), sequence length, or signing characteristics. In exploratory experiments combining MoRE with SignGraph (Gan et al., 2024) and auxiliary regularization (Appendix B), CSL-Daily dev improved (27.95% vs 28.13% baseline) though test remained comparable (27.49% vs 27.22%), reinforcing that dataset-specific factors influence effectiveness.

Relation to Linguistic Structure. The motion gate may relate to coarticulation—where consecutive signs blend—creating transitional movements that are visually present but linguistically ambiguous (Moryossef et al., 2023). Our gate analysis (Figure 4) provides indirect support: the 6$\times$ difference in blank prediction rates between high- and low-gate frames suggests the gate learns to distinguish transitional motion from stable signing, without explicit supervision.

This emergent behavior has implications for understanding CTC-based sequence learning more broadly. The gate appears to learn a soft segmentation that correlates with linguistic boundaries, despite receiving no explicit boundary supervision. This suggests that motion-conditioned gating may serve as an implicit attention mechanism, allowing the model to weight frame representations according to their likely informativeness for the sequence-level task. Whether similar patterns emerge in other CTC domains (e.g., speech recognition with acoustic features) remains an open question.

7 Conclusion

We studied representation enhancement at the pre-contextualization interface in CTC-based sign language recognition. Our analysis yields three findings:

Interface positioning is critical. MoRE helps before the sequence model but *hurts* after, supporting that enhancement should prepare features for contextualization.

Conditioning outperforms fixed alternatives. The learned gate provides consistent improvements over fixed values.

Enhancement reduces deletions. MoRE primarily reduces deletion errors by strengthening representations at frames that would otherwise produce diffuse posteriors.

These findings suggest the encoder-contextualizer boundary deserves attention in weakly-supervised sequence learning.

Limitations

We test on only two datasets with two backbones (ResNet-18, ResNet-34) and one sequence model (BiLSTM). Preliminary Transformer experiments (Appendix A) show modest gains but require further investigation. MoRE shows mixed CSL-Daily results we cannot fully explain. We use simple frame differences rather than optical flow. We have not validated on other CTC tasks (speech, OCR), so cross-domain relevance remains speculative. We have not explored integration with multi-cue fusion.

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th. Papadopoulos, Vassiliki Zacharaki, and Petros Daras. 2022. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia*, 24:1750–1762.

Junseok Ahn, Youngjoon Jang, and Joon Son Chung. 2024. SlowFast network for continuous sign language recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3920–3924.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5130.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7361–7369.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xie, and Sanglu Lu. 2024. SignGraph: A sign sequence is worth graphs of nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1741–1750.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376.

Leming Guo, Wanli Xue, Qing Guo, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Shengyong Chen. 2023. Distilling cross-temporal contexts for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10771–10780.

Aiming Hao, Yuecong Min, and Xilin Chen. 2021. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11303–11312.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

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

Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Hao Lei, Xiujuan Chai, and Xilin Chen. 2023. CoSign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20676–20686.

Oscar Koller. 2020. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Hu Liu, Sheng Jin, and Changshui Zhang. 2018. Connectionist temporal classification with maximum entropy regularization. In *Advances in Neural Information Processing Systems*, volume 31.

Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. 2021. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11542–11551.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12097–12111.

Mathias Müller, Elizabeth Salesky, Rico Sennrich, and Jan Niehues. 2022. Findings of the second WMT shared task on sign language translation (WMT-SLT 2022). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Junfu Pu, Wengang Zhou, and Houqiang Li. 2019. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4174.

Furkat Eshpulatovich Safarov, Mukhiddin Toshpulatov, Komoliddin Mamasabirovich Misirov, Akmalbek Abdusalomov, Azizbek Khojamurotov, and Wookey Lee. 2025. Hyperspectral anomaly detection with enhanced spectral graph transformer network. *IEEE Access*, 13:170554–170576.

Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, volume 27.

William C. Stokoe. 2005. *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*. Journal of Deaf Studies and Deaf Education.

Mukhiddin Toshpulatov, Wookey Lee, Jaesung Jun, and Suan Lee. 2025. Deep learning pathways for automatic sign language processing. *Pattern Recognition*, 164:111475.

Mukhiddin Toshpulatov, Wookey Lee, Suan Lee, Hoyoung Yoon, and U Kang. 2024. DDC3N: Doppler-driven convolutional 3D network for human action recognition. *IEEE Access*, 12:93546–93567.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497.

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Albert Zeyer, Eugen Beck, Ralf Schlüter, and Hermann Ney. 2017. CTC in the context of generalized full-sum HMM training. In *Proceedings of Interspeech*, pages 944–948.

Wenjie Zheng, Hui Chen, Hao Zhang, and Ziyang Zhou. 2023. Spatial-temporal enhanced network for continuous sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2022. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14890–14900.

A Transformer Sequence Model

We evaluate MoRE with a Transformer encoder replacing BiLSTM. Our implementation uses 4 layers, 8 attention heads, hidden dimension 512, and learned positional embeddings—a relatively simple configuration without techniques common in recent CSLR Transformers (e.g., relative position encodings, local attention windows).

Config	Dev	Test	Del	Ins
Trans. Baseline	23.34	23.95	9.64	3.80
+ MoRE	22.78	23.61	8.49	4.10
Δ (rel.)	-2.4	-1.4	-11.9	$+7.9$

Table 3: MoRE with Transformer on PHOENIX-2014. Δ: relative %.

MoRE reduces deletion errors (-11.9%) consistent with BiLSTM findings, though absolute WER is substantially higher than the BiLSTM baseline (23.61% vs 19.24%). We attribute this performance gap to our basic Transformer configuration: the 4-layer architecture with standard positional embeddings may be suboptimal for variable-length sign sequences compared to the recurrent inductive bias of BiLSTM. Proper Transformer integration with relative positional encodings or hybrid architectures remains important future work.

B Auxiliary Gate-Based Regularization

We explored auxiliary losses leveraging the learned gate to encourage alignment with motion patterns and CTC behavior.

Phase Alignment Loss. Encourages gate-motion correlation:

$$\mathcal{L}_{\text{phase}} = \frac{1}{T'} \sum_t (\bar{g}_t - E_t)^2 \tag{6}$$

where $\bar{g}_t$ is mean gate activation and E_t is normalized motion energy at frame t.

Gate Budget Loss. Regularizes gate activation toward a target prior ρ:

$$\mathcal{L}_{\text{budget}} = \left| \frac{1}{T'} \sum_t \bar{g}_t - \rho \right| \tag{7}$$

Experimental Setup. We combine MoRE with SignGraph (Gan et al., 2024) backbone and apply soft-thresholded versions of the auxiliary losses. Based on limited hyperparameter exploration, we set: threshold τ=0.6, softness 0.05, target prior ρ=0.25, λ_{phase}=1.0, λ_{budget}=0.05. These values were selected from a small grid search prioritizing dev set stability; systematic tuning may yield further improvements.

Configuration	Dev	Test
SignGraph Baseline	18.28	19.57
+ MoRE + Aux. (seed 1)	17.87	18.99
+ MoRE + Aux. (seed 2)	18.30	19.16
+ MoRE + Aux. (seed 3)	17.92	19.07
Mean $\pm$ std	18.03$\pm$.24	19.07$\pm$.09

Table 4: MoRE with auxiliary gate regularization on PHOENIX-2014 using SignGraph backbone.

Auxiliary regularization yields 19.07% mean test WER (vs. 19.57% baseline), a 2.6% relative improvement. The best single run achieves 18.99%, suggesting that explicit gate supervision may complement the learned conditioning. While these results are preliminary and the hyperparameter space remains underexplored, they indicate potential for further gains. With proper tuning and integration with recent advances in CSLR architectures, such auxiliary losses may offer a complementary path to improving sequence alignment.

C Additional Qualitative Example

Figure 5 presents an additional qualitative example from PHOENIX-2014, demonstrating MoRE's ability to recover multiple consecutive deleted signs.

In this example, the baseline model fails to recognize MEHR (more) and WOLKE (clouds)—two semantically important signs in a weather forecast context. The CTC posteriors reveal that the baseline produces diffuse activations across these frames, defaulting to blank predictions and incorrectly substituting KOMMEN (come). MoRE's motion-conditioned enhancement produces sharper, more localized posteriors that correctly identify both signs, reducing WER from 29% to 0% on this sequence. This example illustrates how MoRE addresses the consecutive deletion pattern that frequently occurs during coarticulated signing.

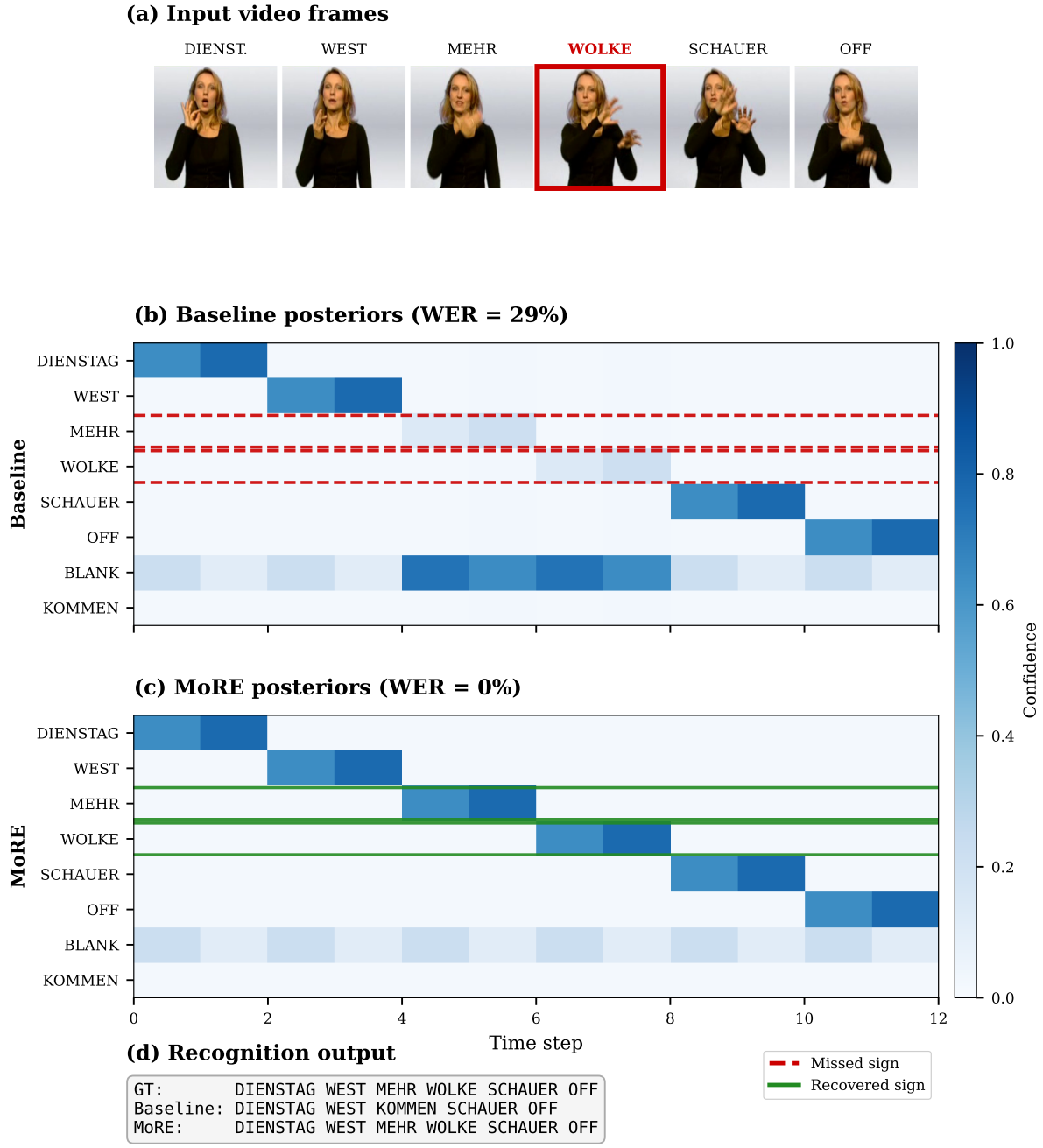

Figure 5: Additional qualitative example from PHOENIX-2014 development set. **(a)** Input video frames showing the sign sequence "DIENSTAG WEST MEHR WOLKE SCHAUER OFF" (Tuesday, west, more, clouds, showers, off). The sign WOLKE is highlighted. **(b)** Baseline CTC posteriors show weak activation for MEHR and WOLKE (dashed red lines), causing their deletion and an incorrect substitution with KOMMEN. **(c)** MoRE posteriors show recovered activation for the missed signs (solid green lines), producing correct recognition. **(d)** Recognition outputs confirm baseline achieves 29% WER while MoRE achieves 0% WER on this sequence.