Beyond CTC Spikes: LLM-Driven Gloss-Temporal Supervision for Continuous Sign Language Recognition

Junfeng Zhu and Cheng Jin and Yuchun Fang*
School of Computer Engineering and Science, Shanghai University
Shanghai, China
jzf458@shu.edu.cn, chengj@shu.edu.cn, ycfang@shu.edu.cn

Abstract

Continuous Sign Language Recognition aims to predict gloss sequences from unsegmented sign language videos. Existing methods typically rely on Connectionist Temporal Classification (CTC) loss for implicit alignment, which is the most dominant loss function for current sequence alignment tasks but causes low-entropy alignment and weak visual-semantic correspondence learning. To address this issue, we propose an explicit temporal-semantic supervision framework that provides dense signals beyond CTC. Specifically, we design the Gloss-Temporal Module (GTM) to construct sustained gloss-to-frame interval supervision by using an LLM to integrate three types of information from RGB videos. We further introduce a joint optimization method with frame-level and sequence-level objectives, coupled with a curriculum learning strategy, enabling a transition from local to global alignment and improving the model's learning of visual-semantic features. The results show that dense temporal supervision substantially alleviates alignment collapse and enables even simple RGB-only baseline models to achieve competitive performance among state-of-the-art methods without extra cues, while the recognizer converges significantly faster.

1 Introduction

Unlike Isolated Sign Language Recognition (Hu et al., 2021a,b; Sarhan and Frintrop, 2023), Continuous Sign Language Recognition (CSLR) deals with natural expressions, including abundant co-articulations and non-manual cues such as facial expressions and body posture (Dreuw et al., 2007; Ong and Ranganath, 2005). The core challenge lies in temporal alignment, where the correspondence between videos and glosses is unknown.

However, existing CSLR methods often overlook the alignment between motion segments and gloss labels. Instead, they typically rely solely on CTC loss (Graves et al., 2006), which is designed for sequence-level modeling (with formula in Appendix A), lacking explicit supervision to inform the model which motion features correspond to specific glosses and which correspond merely to transitional movements between glosses.

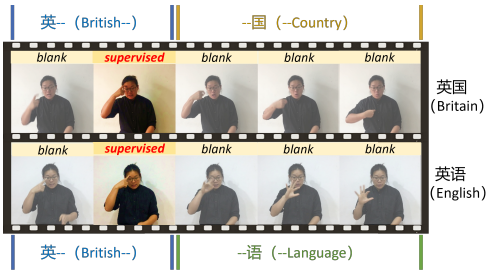


Figure 1: Example of supervised intervals generated by self-distillation. The top shows the complete movement for gloss "Britain", while the bottom shows the complete movement for gloss "English". "British–" is the shared-prefix for these two glosses in Chinese Sign Language.

Using only CTC loss to update the parameters of the entire model is insufficient to fully train the feature extractor. As a result, alignment predictions often present a spike phenomenon. That is, the final predictions typically consist of a small number of non-blank key frames and a large number of high-confidence blank frames. This also explains the source of errors in self-distillation methods in most cases, as illustrated in Figure 1, where the supervised intervals correspond to the same visual features despite representing different glosses.

From an information-theoretic perspective, the high uncertainty over the latent alignment variables is forcibly minimized, resulting in a low-entropy degenerate distribution that collapses to the "safest" class. In this situation, a large portion of visual-semantic information is lost. Our empirical analysis shows that on the PHOENIX14-T dataset (Camgoz et al., 2018), the average logarithmic probability of predicting blank is 18.2 higher than that

*Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 13237–13252
October 24-29, 2026 ©2026 Association for Computational Linguistics

of predicting non-blank labels. On the CSL-Daily dataset (Zhou et al., 2021), this gap reaches 20.7.

Inspired by Chain-of-Thought (Wei et al., 2022), we argue that providing neural networks with intermediate supervision is analogous to equipping the model with a learnable reasoning chain. Having the model identify individual glosses before outputting the final sequence also mirrors how humans recognize continuous sign language. Thus, we shift the focus to supervision design and propose the Gloss-Temporal Module (GTM) to construct sustained gloss-to-frame interval supervision signals. GTM is an information integration module that combines anchor estimation to obtain reliable positioning points, temporal variation signals to capture video structure, frame-level descriptions to inject semantic information gain, and leverages an LLM for structured reasoning. Building upon this, we further propose a dual-objective optimization framework that jointly optimizes frame-level and sequence-level objectives under a curriculum learning strategy. This strategy gradually expands the temporal receptive field, enabling a smooth transition from local discrimination to global alignment, ultimately restoring the entropy from a degenerate distribution to a more informative one. Our main contributions are summarized below:

- We rethink CTC alignment and reconstruct the paradigm to generate more sufficient and denser gloss-to-frame intervals as references for distribution calibration via KL divergence.
- We use window-based differences for boundary capture, frame descriptions as reasoning cues, and an LLM as the inference operator.
- We design a joint optimization method that combines frame-level and sequence-level objectives and introduce a curriculum learning strategy to improve stability and capability.
- The proposed method achieves competitive performance on two benchmark datasets only relying on RGB inputs and does not increase trainable parameters of the baseline recognition model, while demonstrating advantages in convergence speed and interpretability.

2 Related Work

Early approaches typically relied on handcrafted features (Gao et al., 2004) or HMM-based systems (Koller et al., 2017). More recently, CNNs and RNNs have brought significant progress to CSLR (Pu et al., 2019; Cheng et al., 2020; Niu and Mak, 2020; Min et al., 2021). These methods generally learn visual representations from videos and then model temporal dependencies, typically trained with CTC loss for sequence prediction.

However, these methods are limited by weak feature extraction, poor motion continuity modeling, and insufficient use of video information, ultimately resulting in inadequate supervision.

2.1 Visual Representation Learning

CTC loss encourages the model to implicitly align visual features with corresponding glosses during training, but it fails to accurately capture the precise temporal boundaries of each gloss. To address the issue of "CTC-dominated training with insufficient supervision for feature extractors", VAC (Min et al., 2021) distills long-term contextual knowledge back into the feature extractor. Similar approaches include self-mutual distillation (Hao et al., 2021) and pseudo-labeling methods (Cheng et al., 2020).

However, the teacher model itself may be inaccurate, causing the visual branch to inherit such erroneous biases, as illustrated in Figure 1.

2.2 Temporal Modeling

Sign language actions may consist of a sequence of continuous movements. In continuous sign language videos, there also exist transitional and pause motions. CorrNet (Hu et al., 2023a) models motion trajectories by computing pixel-wise correlations between adjacent frames using dot-product operations. SignGraph (Gan et al., 2024) constructs sparse temporal graphs across adjacent frames, enabling graph convolution to capture cross-frame changes. In the broader field of video understanding, recent works have explored multi-scale temporal modeling (Hu et al., 2022; Dai et al., 2022; Li et al., 2026; Zhang et al., 2026), where 1D CNNs with different kernel sizes are employed to capture motion patterns across varying temporal ranges.

However, these methods do not explicitly provide intermediate supervision to indicate which features correspond to specific glosses or not.

2.3 Information Augmentation

To enable models to better learn discriminative features, existing works have explored incorporating extra cues such as skeleton (Gan et al., 2021; Zuo and Mak, 2022), depth images (Tang et al., 2022), and local areas (Cihan Camgoz et al., 2017; Camgoz et al., 2020). Other approaches focus on leveraging more training data through techniques such

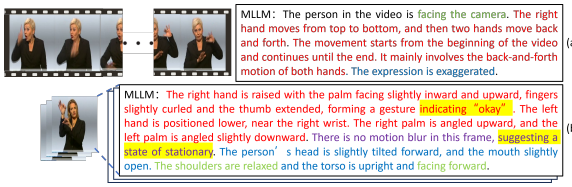


Figure 2: Examples of description outputs generated by Qwen3-VL-8B (Bai et al., 2025). (a) Video-level description. (b) Frame-level descriptions, where green indicates relatively redundant or highly consistent content, red denotes hand-related information, purple represents motion states, blue corresponds to facial and body cues, and yellow highlights speculative descriptions.

as back-translation (Zhou et al., 2021), cross-modal augmentation (Pu et al., 2020), contrastive learning (Gan et al., 2023), or pretraining on other datasets (Chen et al., 2022; Liang et al., 2024).

However, these methods require either increased model complexity or additional data. Moreover, for frames with severe motion blur or inherent data quality issues, extracting reliable hand skeletons is challenging (see Appendix B for examples).

In the related task of Continuous Sign Language Translation, an increasing number of works have begun to leverage Multimodal Large Language Models (MLLMs) (Gong et al., 2024; Wong et al., 2024; Hwang et al., 2025), including enhancing visual representations by concatenating feature vectors derived from frame-level textual descriptions (Kim et al., 2025). As illustrated in Figure 2, MLLMs-generated frame-level descriptions provide richer details than video-level, but often produce similar texts for semantically distinct yet visually similar signs (e.g., "okay", "grad", "three"). Additionally, these descriptions may include highly redundant content (e.g., "facing forward").

3 Method

3.1 Overall Architecture

As shown in the blue-based workflow on the right side of Figure 3, the CSLR baseline model consists of a feature extractor (2D CNN+1D CNN), a BiLSTM and a classifier. Given a sign language video with T input frames denoted as $X = \{x_t\}_{t=1}^{T} \in \mathbb{R}^{T\times 3\times H_0\times W_0}$, the model aims to translate the input video into a gloss sequence $Y = \{y_i\}_{i=1}^{N}$, where N denotes the length of the label sequence. The 2D CNN first encodes the input into frame-level features $F = \{\mathbf{f}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$. The 1D CNN and BiLSTM then perform short- and long-term temporal modeling based on these features. Finally, the classifier predicts the probability $p(Y \mid X)$ of the target gloss sequence using CTC loss.

The proposed GTM constructs prompts from positioning points, video structures, and semantic guidance, and leverages an online LLM to integrate these clues. Using the structured GTM outputs, we introduce Cross-Entropy loss as intermediate supervision in a dual-objective framework with a shared classifier, which strengthens feature extraction for individual glosses without extra parameters. Besides, GTM is applied only once during preprocessing, while the recognition model is trained and evaluated end-to-end using RGB inputs only.

3.2 Positioning Points

In CSLR, given an input sequence X and a label sequence Y, the true frame-level alignment path π is the latent variable to be inferred. Since CTC-based alignment does not significantly deviate from the correct temporal regions, we incorporate it as a preliminary component of the GTM and reinterpret it as a low-entropy yet high-confidence anchor estimator. Specifically, we first train the baseline model to fit on the training set and then apply "S^3 dynamic programming" (with state transition in Appendix C) to enforce alignment with the ground-truth gloss sequence, selecting the most probable path among all valid candidates:

$$\arg \max_{\pi \in \mathcal{B}^{-1}(Y)} \log p(\pi \mid X) \tag{1}$$

This path provides a set of high-confidence discrete alignment points along the temporal axis, so the goal here is not full alignment but a set of positioning points. This "anchor+uncertainty" representation transforms the original unstructured alignment problem into a segmented search problem with locally deterministic regions, thereby reducing the risk of errors in subsequent inference.

3.3 Video Structures

We further introduce temporal variation modeling to capture the boundaries between latent semantic units. For an input video, each frame x_t is mapped to a feature vector $\mathbf{f}_t$. We compute the first-order temporal difference between adjacent frames:

$$\mathbf{v}_t = \mathbf{f}_{t+1} - \mathbf{f}_t, \quad t = 1, \ldots, T-1 \tag{2}$$

Since the movement of a gloss is a pattern of changes, different semantic units cannot be distinguished based solely on differences between adjacent frames. Therefore, we consider defining a

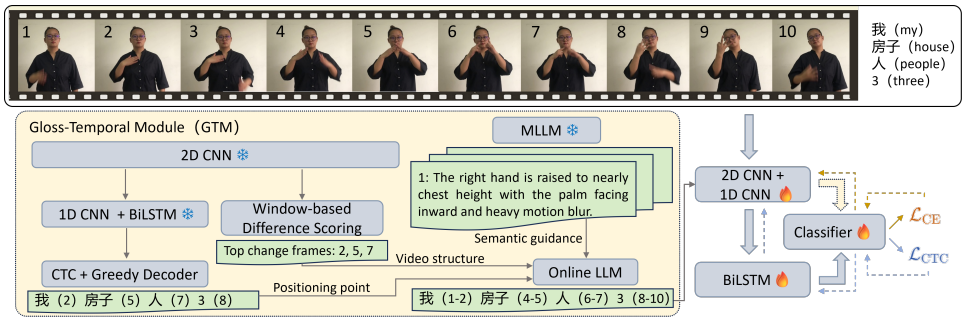


Figure 3: Overall flowchart. Yellow part: proposed method, and dashed arrows indicate gradient backpropagation.

symmetric temporal window of size w. For each time step t, we compute the average of the difference vectors within the left and right windows:

$$\mathbf{v}_t^{\text{left}} = \frac{1}{w} \sum_{i=t-w}^{t-1} \mathbf{v}_i, \quad \mathbf{v}_t^{\text{right}} = \frac{1}{w} \sum_{i=t}^{t+w-1} \mathbf{v}_i \quad (3)$$

The change score is defined as:

$$d_t = \left\| \mathbf{v}_t^{\text{right}} - \mathbf{v}_t^{\text{left}} \right\|_2 \quad (4)$$

This score essentially measures the degree of motion variation before and after the current time step. High-score points correspond to potential semantic boundaries, while low-score regions indicate stable segments. Finally, we select the Top-K highest-scoring points from all time steps and impose a minimum interval constraint to avoid redundant detections. Thus, this section introduces a label-independent structural prior derived from the video, which complements the label-driven signal of CTC.

3.4 Semantic Guidance

The first two steps provide anchor constraints and structural boundaries. However, both types of information are based on low-order statistical signals. By leveraging Prompt A in Figure 4 together with an MLLM, we obtain frame-level descriptions as illustrated in Figure 2(b). These descriptions contain information about motion states and grounded inferences, which are difficult to obtain from visual feature extractors. Therefore, we formulate the final alignment as a structured prediction problem that integrates multi-source information: it measures the consistency between segmentation and anchor points, evaluates the alignment with temporal change boundaries, and assesses the coherence of the descriptions within each segment. The third term involves high-level semantic content included in very long (about 8000 words per video) and partially inconsistent textual sequences that cannot be accurately modeled by simple heuristics. Rather than regarding it as a feature, what we aim to extract from it are the expressions of actions and the coherence. Thus, we leverage an LLM as a powerful approximate inference operator:

$$S^* \approx \mathcal{F}_{\text{LLM}}(Y, A, B, D) \quad (5)$$

where A, B, D denote the set of anchors, boundaries, and descriptions. The LLM acts as a global reasoner by using information in the input video X. We adopt DeepSeek-V3 (Liu et al., 2025a), and utilize Prompt B in Figure 4 to integrate multi-source information and infer segmentation. During

A: Describe this image. Pay special attention to the person's hands (e.g., hand gestures, relative hand positions, and palm orientations). You may infer the speed of hand movements based on the degree of motion blur in the image. Also pay attention to the non-manual parts of the person, including facial expressions, especially unusual eyebrow and mouth shapes, as well as body posture. Do not describe the person's attributes, portrait characteristics, or the background. Do not fabricate details. Base your description entirely on the image content. If something is genuinely unclear, you may write "unclear" or "infer the hand based on the arm position." You may also write sentences such as "there is no hand in the image".

B: You are an expert in sign action and temporal video analysis. I will provide you with a segment of a gloss sequence, coarse positioning points, frame indices with the largest motion-pattern changes, and frame-wise textual descriptions. Please perform fine-grained segmentation based on this information. Note: the provided coarse positioning points may not be completely accurate; the described actions may not necessarily be the most prototypical realizations of the corresponding glosses in your knowledge base; and the frame descriptions may also be imperfect. You should place greater emphasis on analyzing the motion continuity reflected in the frame descriptions. The frames with the largest motion-pattern changes are only provided as potential boundaries. The action corresponding to one gloss may consist of several changing sub-parts. Segments do not need to be continuous, but they must not overlap. Each gloss must correspond to exactly one interval, and the intervals must follow the order of the input gloss sequence. There is no need to pursue excessive precision down to a specific frame. The input is as follows: ...

Figure 4: A: Frame-level description prompt, colors corresponding to the meanings in Figure 2. B: Final reasoning prompt, different colors represent different info types and brown indicates precautions to take during reasoning.

our observations of the reasoning chain, we find the LLM tends to over-focus on assigning precise boundaries to individual frames, which may lead to discarding entire segments. Thus, we include an instruction that discourages over-precise alignment. Meanwhile, we observe that, apart from highly imageable words like "I", most actions do not match the actions in the LLM's knowledge base (an example is provided in Appendix D). This suggests that the gain is better explained by constrained integration of the provided clues, rather than by dataset memorization. To mitigate the impact of stochasticity in LLM output, we perform multiple independent runs on the same input and aggregate the results.

3.5 Curriculum Learning

With the above paradigm, we ultimately derive training-effective gloss-to-frame interval signals (several examples are provided in Appendix E). GTM can be further understood as an information augmentation and distribution calibration module:

$$D_{\mathrm{KL}}(q_t \,\|\, p_t) = \sum_z q_t(z) \log \frac{q_t(z)}{p_t(z)} \tag{6}$$

where $p_t(z)$ denotes the predicted distribution over gloss label z at time step t, while $q_t(z)$ represents the explicit supervision provided by GTM. By minimizing the KL divergence, the model is encouraged to form a more reasonable probability distribution rather than collapsing to the "safest" label.

CE is equivalent to the sum of KL divergence and the entropy of the target distribution:

$$\mathcal{L}_{\mathrm{CE}}(q_t, p_t) = D_{\mathrm{KL}}(q_t \,\|\, p_t) + H(q_t) \tag{7}$$

Thus, minimizing CE is equivalent to minimizing the KL divergence. In this way, we transfer the core idea of Chain-of-Thought to the alignment problem in CSLR. Specifically, CE explicitly injects local alignment constraints, enabling the model to strengthen visual–semantic correspondence and ultimately realize entropy recovery.

CE performs local optimization by supervising valuable temporal intervals, while CSLR essentially relies on holistic sequence modeling. To resolve the conflict between local optimization and sequence learning, we design a curriculum-based weighting scheme. This strategy constructs a learning paradigm that progresses from easy to hard. The overall objective function is defined as:

$$\mathcal{L} = \lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{CTC}}\mathcal{L}_{\mathrm{CTC}} \tag{8}$$

We formulate the training process as a progressive optimization procedure from low to high entropy, ensuring that appropriate supervision signals are provided at different training stages. In the early stage, we assign a higher weight to λ_{CE}, constraining the model within a locally stable solution space and preventing the collapse caused by CTC. As training proceeds, we gradually adjust the loss weights via linear interpolation, enabling a smooth transition from frame-level to sequence-level dominance. In the later stage, the model has acquired a certain level of alignment capability and can effectively leverage sequence-level supervision to further improve overall recognition performance. Meanwhile, a small weight on CE is retained to preserve local discriminative constraints.

4 Experiments

4.1 Datasets

We conduct experiments on two public datasets:

PHOENIX14-T. This German Sign Language dataset is an extension of PHOENIX14 (Koller et al., 2015). It contains 1,085 glosses from 9 signers. The training, validation and test sets include 7,096, 519, and 642 samples, respectively. Data are collected from a weather forecast broadcast.

CSL-Daily. This Chinese Sign Language dataset contains 2,000 glosses from 10 signers. The training, validation and test sets consist of 18,401, 1,077, and 1,176 samples, respectively. Data are collected in a lab environment.

4.2 Evaluation Metric

We adopt Word Error Rate (WER) as the evaluation metric for CSLR. It is defined as the minimum number of substitutions (sub), insertions (ins), and deletions (del) required to transform the predicted sequence into the reference sequence (ref):

$$\mathrm{WER} = \frac{\#\mathrm{sub} + \#\mathrm{ins} + \#\mathrm{del}}{\#\mathrm{ref}} \tag{9}$$

4.3 Experimental Setting

We follow the same settings as prior SOTA methods (Min et al., 2021; Hu et al., 2023a) to construct our CSLR baseline model and data preprocessing pipeline. Detailed configurations are given in Appendix F. All of our experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU.

Window size. We set w for the "video structure" as the total number of frames divided by the number of glosses plus 3. We then select the

top (#glosses + 3) positions as candidate boundary points. Considering the temporal downsampling factor of 4 introduced by the 1D CNN, we enforce a minimum temporal interval of 4 frames.

Loss weights. We conduct experiments on PHOENIX14-T, initializing the weighting ratio between CE and CTC as 4:1. As shown in Table 1, neither assigning a consistently dominant role to the CE loss nor to the CTC loss yields optimal performance. The proposed curriculum learning strategy outperforms fixed-weight schemes.

We observe that by the 7th epoch, the mean CE loss decreases from 17.80 at the first epoch to 1.96. This indicates that the model has learned discriminative visual features for classification. Accordingly, from epoch 8 to 20, we linearly decrease λ_{CE} from 1 to 0.20 and increase λ_{CTC} from 0.25 to 1. Meanwhile, the CTC loss decreases from 40.6 at the first epoch to 7.16 at the 10th, and eventually falls below 1. This trend confirms that the proposed weighting schedule works as expected, shifting optimization from CE-dominated local supervision to CTC-dominated sequence learning.

LLM output variability. We obtain three independent LLM outputs whose segment count matches the input glosses. We average the start/end boundaries of each segment, yielding an average boundary standard deviation of 3.82 frames, below the temporal downsampling factor of 4 and thus within the recognizer's temporal resolution. Larger deviations occur mainly at the ending boundaries of long-duration glosses, followed by their starting boundaries, which often correspond to motion continuations, holdings, or transitions. Averaging the three outputs further smooths these uncertainties.

4.4 Performance Comparison

Table 2 presents a comparison between our method and other state-of-the-art (SOTA) approaches. Entries marked with "*" indicate that these methods utilize additional cues such as facial or hand features. Our goal is not to set a marginal new SOTA improvement by heavy architecture design, but to validate the effectiveness of correcting CTC alignment and demonstrate that alignment supervision itself is the bottleneck. Our method achieves a WER of 19.7 and 20.7 on the validation and test sets of PHOENIX14-T, as well as 30.8 and 29.6 on the validation and test sets of CSL-Daily. Some methods with lower WERs, such as S2Net and CAP-SLR, exploit additional cues beyond RGB inputs, while SignGraph adopts a graph-based architecture with more trainable parameters. Under the same ResNet-18 backbone, our approach uses RGB inputs only and introduces no additional trainable parameters to the baseline recognizer. By leveraging dense temporal supervision signals with a simple CE loss, our method substantially closes the gap between a simple CTC baseline and stronger systems. This comparison supports our main claim that improving temporal supervision alone can significantly alleviate the limitations of CTC-only training, rather than claiming an architecture-level SOTA.

4.5 Ablation Study

Table 3 reports ablation results on PHOENIX14-T. When semantic guidance is removed, interval boundaries are obtained by simply extending each anchor to its nearest left and right action change points. All GTM components contribute to the final performance. Notably, boundary-based anchor extension reduces WER by 2.8% over baseline, which demonstrates that reasonably extending the supervision intervals itself consistently improves WER, while GTM further reduces WER by 2.3%.

We also try SoftCE for local supervision, where the target probability of each gloss is defined by its frame proportion within the current temporal convolution window. SoftCE yields 20.7%/21.5% WER on the validation/test sets, worse than hard CE, likely because assigning probability mass to neighboring glosses or blanks dilutes the corrective signal for dense visual-semantic supervision.

Besides, Focal-CTC(Feng et al., 2019) using γ = 0.5 and α = 1.0 achieves 23.4%/24.4% Dev/Test WER on PHOENIX14-T with higher runtime.

To show that GTM is not tied to specific MLLMs

Configurations	Dev (%)	Test (%)
$\lambda_{\text{CE}} = 1,\ \lambda_{\text{CTC}} = 0.25$	21.4	22.2
$\lambda_{\text{CE}} = 0.25,\ \lambda_{\text{CTC}} = 1$	21.0	21.6
Ours	**19.7**	**20.7**

Table 1: Performance comparison of different weighting strategies on PHOENIX14-T.

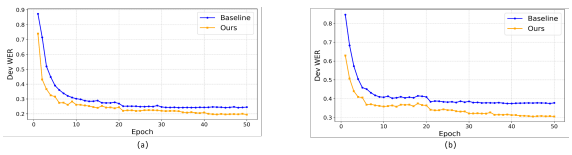


Figure 5: Validation WER curves over epochs during training. (a) PHOENIX14-T. (b) CSL-Daily.

Model	Trainable Parameters (– indicates not open-sourced)	Backbone	PHOENIX14-T		CSL-Daily	
			Dev (%)	Test (%)	Dev (%)	Test (%)
CoSign* (Jiao et al., 2023)	–	GCN	19.5	20.1	28.1	27.2
S^2Net* (Yang et al., 2025)	–	GCN	17.2	18.9	25.8	24.5
CAP-SLR* (Wei et al., 2025)	–	ResNet34	17.0	18.8	26.7	26.6
SEN (Hu et al., 2023b)	34,513,232	ResNet18	19.3	20.7	31.1	30.7
CTCA (Guo et al., 2023)	–	ResNet18	19.3	20.3	31.3	29.4
CorrNet (Hu et al., 2023a)	33,370,146	ResNet18	18.9	20.5	30.6	30.1
SignGraph (Gan et al., 2024)	36,631,733	GCN	**18.3**	**20.0**	**27.3**	**26.4**
STAgg (Hu et al., 2024)	–	ResNet34	19.0	20.4	30.6	30.1
VCL (Liu et al., 2025b)	67,692,400	ResNet18	18.6	20.2	28.6	28.2
Baseline	**32,974,672**	ResNet18	24.5	25.8	37.5	36.8
GTM + Baseline	**32,974,672**	ResNet18	19.7±0.2	20.7±0.2	30.8±0.1	29.6±0.2

Table 2: Comparison with state-of-the-art methods. * indicates methods that utilize additional cues.

Positioning	Structure	Semantic	Dev (%)	Test (%)
×	×	✓	22.9	23.9
✓	✓	×	21.9	23.0
✓	×	✓	20.3	21.2
✓	✓	✓	**19.7**	**20.7**

Table 3: Ablation study results on PHOENIX14-T.

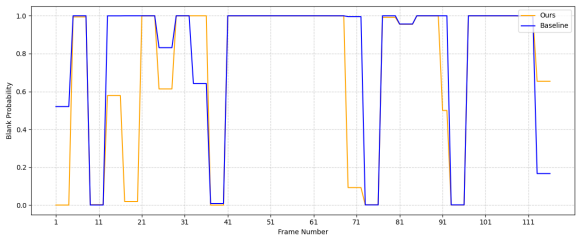


Figure 6: Frame-level probability of being predicted as blank for an example video.

MLLM	LLM	Dev (%)	Test (%)
InternVL3.5-8B	Qwen3.5-plus	20.1	21.1
Qwen3-VL-8B	Qwen3.5-plus	20.0	21.0
InternVL3.5-8B	DeepSeek-V3	19.8	20.8
Qwen3-VL-8B	DeepSeek-V3	19.7	20.7

Table 4: Performance comparison of different MLLM and LLM combinations on PHOENIX14-T.

Backbone	Dev (%)		Test (%)	
	Baseline	GTM	Baseline	GTM
ResNet18	24.5	19.7	25.8	20.7
MobileNetV3-S	27.7	20.3	29.1	21.6
DeiT-Tiny	26.0	19.3	26.7	20.3
EdgeNeXt-XX-S	25.4	19.9	26.2	21.5

Table 5: Performance comparison of different backbones on PHOENIX14-T.

or LLMs, we further test InternVL3.5-8B (Wang et al., 2025) and Qwen3.5-plus (Team, 2026) on PHOENIX14-T, as shown in Table 4. It turns out that GTM is robust across different combinations.

To demonstrate that GTM is not tied to any specific backbone architecture, we evaluate it on four distinct backbones: ResNet18, a traditional CNN-based backbone; MobileNetV3-Small, an efficient CNN-based backbone (Howard et al., 2019); DeiT-Tiny, a lightweight ViT-based backbone (Touvron et al., 2021); and EdgeNeXt-XX-Small, a lightweight CNN–Transformer hybrid backbone (Maaz et al., 2023). The results reported in Table 5 confirm that GTM consistently achieves effective performance across all backbone variants.

We conduct the full ablation suite mainly on PHOENIX14-T. Nevertheless, other experimental results on CSL-Daily show consistent improvements, suggesting that the proposed supervision paradigm generalizes beyond a single benchmark.

4.6 Effectiveness Analysis

Figure 5 shows the validation WER curves over epochs. With our method, the validation WER converges faster, allowing the model to reach the baseline performance with fewer training epochs.

In addition, on PHOENIX14-T, the average proportion of frames predicted as "blank" on the validation set decreases from 0.69 for the baseline to 0.65 with GTM, while the average frame-level posterior entropy increases from 0.16 to 0.19. On CSL-Daily, the blank-frame proportion decreases from 0.76 to 0.72, and the average frame-level posterior entropy increases from 0.18 to 0.22. Overall, GTM reduces the blank-frame proportion by 5.8% and 5.3% on the two datasets, while increasing the posterior entropy by 18.8% and 22.2%. These results indicate that GTM alleviates blank-dominated posterior con-

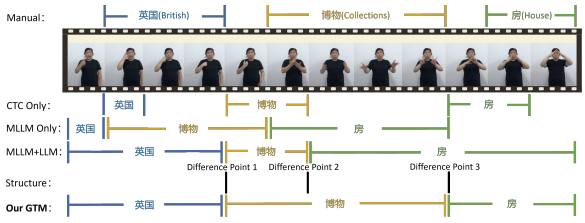


Figure 7: Qualitative comparison of gloss-to-frame segmentations produced by different methods.

centration and encourages more informative frame-level predictions, reducing the model's reliance on sparse CTC spike frames. Together with a decrease in WER, these findings support the increase in the model's effective information acquisition. Figure 6 provides a line chart of the blank prediction probability for each frame in an example video. For this sample, the average probability of predicting a frame as "blank" is 79.7% for baseline, whereas our method reduces it to 70.2%.

4.7 Visual Analysis

Segmentation quality. Figure 7 provides an intuitive comparison of different methods. The top row shows the manually annotated result, followed by the results of CTC-only (self-distillation), local 8B-MLLM-only, local MLLM with frame-level descriptions combined with an online LLM, top-k window differencing frames, and GTM. As illustrated, self-distillation yields sparse supervisory frames, although mostly lie within the correct temporal regions. The segments obtained by the local MLLM alone are largely inaccurate. When combining MLLM with an LLM, the LLM tends to rationalize errors in the descriptions, leading to misalignment. In this case, the latter part of "collections" is described as "hands moving upward in the front", which is incorrectly assigned to the segment "house". However, this approach provides longer supervision ranges. The windowed difference method can reasonably capture the temporal structure. By combining the strengths of CTC-only, MLLM+LLM, and temporal structure, GTM produces more training-effective interval supervision.

Human evaluation of interval quality. Since precise temporal boundary annotations are not available in the datasets, we conduct a blind human rating as a complementary evaluation. Seven undergraduate annotators from different majors are asked to score the interval quality of three anonymized methods from 0 to 100: CTC only, extending CTC spikes to the nearest video-structure points, and GTM. The rating considers action continuity, boundary reasonableness, and interval coverage. As shown in Table 6, GTM obtains the highest average score of 87.9, outperforming Structure Extension and CTC Only by 10.0 and 17.2 points. The annotators generally observe that GTM produces more coherent boundaries and more complete action intervals, while Structure Extension may still place boundaries inside high-motion phases or leave some actions uncovered, and CTC Only is often located in the correct region but too sparse to provide effective dense supervision. The detailed reasons for the scores are provided in Appendix G. This evaluation further supports the quality of GTM-generated interval supervision.

Annotator	CTC Only	Structure Extension	GTM
A1	80	80	90
A2	65	75	80
A3	70	80	90
A4	75	80	85
A5	75	70	90
A6	70	80	90
A7	60	80	90
Average	70.7	77.9	**87.9**

Table 6: Blind human rating of interval quality. Structure Extension denotes extending CTC spikes to the nearest video-structure points.

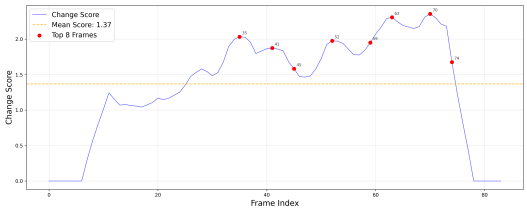


Figure 8: Windowed difference scores for each frame in an example video.

Windowed difference scores for each frame in an example video. Based on manual annotation, the four temporal boundaries should lie within the ranges of frames 31–33, 39–44, 58–61, and 70–73. As illustrated in Figure 8, among the top-8 frames with the highest difference scores, frames 35, 45, 58, and 70 are included. This observation indicates that the proposed windowed difference scoring provides reliable temporal structure cues. A format-adjusted version of the LLM reasoning process for this example is provided in Appendix H.

Grad-CAM (Selvaraju et al., 2017) **visualizations of the feature extractor**. It can be observed from Figure 9 that, after incorporating GTM and CE, the model focuses more on fine-grained re-

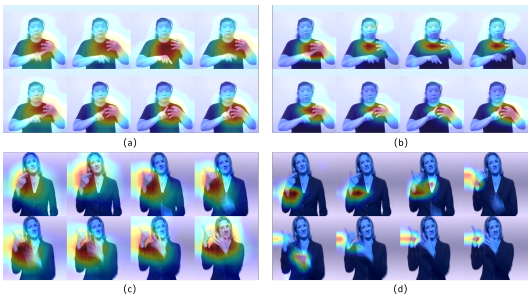


Figure 9: Grad-CAM visualizations. Left: baseline. Right: GTM + Baseline.

Reference:	小 程序 这 不用 下载 可以 直接 用 多 方便
	(small program this noneed download can directly use so convenient)
Ours:	程序 这 不 下载 可以 一直 用 多 方便
	(program this no download can always use so convenient)
Reference:	我 国 今天 年 进 口 越 少
	(I country today year enter port more few)
Ours:	我 今天 年 进 口 越
	(I today year enter port more)
Reference:	他 处罚 对 女 人 重
	(he punishment for female person heavy)
Ours:	他 处罚 不好 对 女 人 好了
	(he punishment bad for female person done)

Figure 10: Examples of errors.



Figure 11: Examples of error causes.

gions, such as the hands and face, rather than coarse central areas of the image. This suggests that GTM provides more informative supervision signals.

4.8 Error Analysis

Compared with the baseline, GTM reduces deletions and insertions on CSL-Daily by 9 cases (0.94%) and 25 cases (13.37%). On PHOENIX14-T, GTM slightly increases deletions by 6 cases (1.25%), but reduces insertions by 39 cases (27.08%). We further analyze the recognition results of our method on the test set and summarize four typical error causes: recognition errors due to similar actions, gloss deletion errors due to similar actions, gloss insertion errors due to action combination, and other recognition errors. Examples on CSL-Daily are shown in Figure 10. To make the analysis more systematic, we randomly sample 100 test-set prediction results and manually assign each case to its dominant error category. The proportions of the four categories are 45%, 20%, 15%, and 20%. The "other" category mainly consists of general misrecognition and missed-recognition cases that do not clearly fall into the previous three categories. These results indicate that action similarity remains the dominant source of residual errors, while GTM mainly reduces insertion errors and alleviates over-segmentation caused by CTC spike-like predictions. We also find four types of original dataset annotation issues, especially in CSL-Daily, including extra glosses, missing glosses, order swapping, and complete mismatch.

In the first example of Figure 10, "small" lasts for about 12 frames but is unrecognized, while "noneed" and "no" represent recognition errors. The substitution of "directly" for "always" is a typical case of recognition errors due to similar actions, as shown in Figure 11(a). In the second example, missing "country" and "few" are typical gloss deletion errors due to similar actions. The model recognizes both segments as the same gloss and the CTC decoder combines them together, resulting in gloss deletion. A visualization of the "more+few" example is shown in Figure 11(b). In the third example, spurious "bad" is a typical gloss insertion error due to action combination. The latter part of the action "punishment" coincides with "bad". This might be due to the long-tail distribution, which makes some stubborn biases difficult to correct.

5 Conclusions

In this work, we verify that replacing CTC spike-based supervision with dense temporal supervision effectively alleviates the low-entropy alignment collapse. Specifically, we propose GTM to construct gloss-to-frame interval annotations via multi-source integration and LLM-driven reasoning, which act as structured reasoning chains to calibrate alignment distributions toward more informative patterns. To strengthen visual-semantic correspondence, we introduce a dual-objective optimization framework, where frame-level supervision improves local discrimination and sequence-level CTC preserves global alignment ability, optimized via curriculum learning strategy for a smooth transition from local to holistic learning. Beyond performance improvements, we show that explicit intermediate supervision can effectively compensate for the limitations of purely data-driven implicit alignment without extra parameters. This suggests that improving supervision, rather than model complexity, is a more effective direction for advancing sequence alignment and may generalize to lip reading and other sequence tasks.

Limitations

Although the proposed framework can mitigate the spike-dominated CTC behavior in CSLR, several aspects can be further improved. First, GTM relies on external LLMs for annotation, which may introduce output variance and offline cost. The LLM is used as a constrained reasoning operator rather than a sign language recognizer, but our preliminary attempts with local Qwen3-8B-Instruct models found that smaller models may lack sufficient context length for long frame-level descriptions and reasoning ability over noisy multi-source clues. Future work may reduce the cost through distillation, key-frame descriptions, or algorithms for integrating noisy multi-source information without relying on large online LLMs. Second, even in the baseline model without GTM and CE loss, the WER on the training set can be reduced to nearly zero, which is a significant gap compared with the performance on the validation and test sets. Learning more generalizable features from limited data remains an open challenge.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No.62576202), Natural Science Foundation of Shanghai (Grant No.25ZR1401115), National Natural Science Foundation of China (Grant No.82430060), National Key R. and D. Program of China (Grant No.2024YFC3306703, 2024YFC3306700), the Shanghai Key Laboratory of Forensic Medicine, Academy of Forensic Science, Ministry of Justice (Grant No.KF202413), and by Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University. We would also like to thank all the participants involved in the annotation as well as all the reviewers for their valuable comments.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision Workshops*, pages 301–319.

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

Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. 2020. Fully convolutional networks for continuous sign language recognition. In *European Conference on Computer Vision*, pages 697–714.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3056–3065.

Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S Ryoo, and François Brémond. 2022. Ms-tct: Multi-scale temporal convtransformer for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20009–20019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. 2007. Speech recognition techniques for a sign language recognition system. In *Proc. Interspeech 2007*, pages 2513–2516.

Xinjie Feng, Hongxun Yao, and Shengping Zhang. 2019. Focal ctc loss for chinese optical character recognition on unbalanced datasets. *Complexity*, 2019(1):9345861.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Hongkai Wen, Lei 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 13470–13479.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. 2023. Contrastive learning for sign language recognition and translation. In *IJCAI*, volume 23, pages 763–772.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. 2021. Skeleton-aware neural sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4353–4361.

Wen Gao, Gaolin Fang, Debin Zhao, and Yiqiang Chen. 2004. A chinese sign language recognition system based on sofm/srn/hmm. *Pattern Recognition*, 37(12):2389–2402.

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

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 11283–11292.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021a. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11067–11076.

Hezhen Hu, Wengang Zhou, and Houqiang Li. 2021b. Hand-model-aware sign language recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1558–1566.

Huazhang Hu, Sixun Dong, Yiqun Zhao, Dongze Lian, Zhengxin Li, and Shenghua Gao. 2022. Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18991–19000.

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

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

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2024. Spatial temporal aggregation for efficient continuous sign language recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(6):3925–3935.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. 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)*, pages 3901–3920.

Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, 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 20619–20629.

Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. 2025. Leveraging the power of mllms for gloss-free sign language translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21048–21058.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

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.

Oscar Koller, Sepehr Zargaran, and Hermann Ney. 2017. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3416–3424.

Haiqin Li, Yuhan Yang, Jun Zeng, Min Gao, and Junhao Wen. 2026. Multi-scale transformers with dual attention and adaptive masking for sequential recommendation. *Information Processing & Management*, 63(1):104318.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, and 180 others. 2025a. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Kailin Liu, Yonghong Hou, Zihui Guo, Wenjie Yin, and Yi Ren. 2025b. Visual context learning based on cross-modal knowledge for continuous sign language recognition. *The Visual Computer*, 41(6):4017–4031.

Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. 2023. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European Conference on Computer Vision Workshops*, pages 3–20.

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 11522–11531.

Zhe Niu and Brian Mak. 2020. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *European Conference on Computer Vision*, pages 172–186. Springer.

Sylvie CW Ong and Surendra Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(06):873–891.

Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. 2020. Boosting continuous sign language recognition via cross modality augmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1497–1505.

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 4160–4169.

Noha Sarhan and Simone Frintrop. 2023. Unraveling a decade: A comprehensive survey on isolated sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 3202–3211.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. 2022. Graph-based multimodal sequential embedding for sign language translation. *IEEE Transactions on Multimedia*, 24:4433–4445.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International conference on Machine Learning*, pages 10347–10357.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Dong Wei, Xu-Hua Yang, Yiyang Weng, Xuanyu Lin, Hongxiang Hu, and Sheng Liu. 2025. Cross-modal adaptive prototype learning for continuous sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 35:7354–7367.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *International Conference on Learning Representations*, pages 18157–18174.

Yifan Yang, Yuecong Min, and Xilin Chen. 2025. S2net: Skeleton-aware slowfast network for efficient sign language recognition. In *Asian Conference on Computer Vision*, pages 355–372.

Qiming Zhang, Zhengping Hu, Yulu Wang, Hehao Zhang, and Jirui Di. 2026. Multi-scale pyramidformer network with multiple consistency constraints for semi-supervised video action detection. *Information Processing & Management*, 63(2, Part A):104421.

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.

Ronglai Zuo and Brian Mak. 2022. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5121–5130.

A CTC Loss Formula

CTC loss is designed for end-to-end temporal classification tasks that process unsegmented sequential data. It defines a many-to-one mapping function to align paths to label sequences, which transforms a path π into the final label sequence Y by removing repeated labels and blank symbols. Using this function, CTC can supervise the model parameters θ by summing the probabilities of all possible paths:

$$\begin{aligned}\mathcal{L}_{\mathrm{CTC}} &= -\log p(Y \mid X;\theta) \\ &= -\log\left(\sum_{\pi\in\mathcal{B}^{-1}(Y)} p(\pi \mid X;\theta)\right)\end{aligned} \tag{10}$$

The conditional probability of a path is computed under the conditional independence assumption as:

$$p(\pi \mid X) = \prod_{t=1}^{T'} p(\pi_t \mid X;\theta) \tag{11}$$

B Problematic Frames in Data

Figure 12(a) exhibits a frame that should correspond to part of the sign "high" in German Sign Language, where the index finger points upward with a rapid upward motion; however, severe motion blur caused by low sampling rate obscures the movement. Figure 12(b) exhibits a frame with inherent data quality issues, where the hand is occluded.



(a) (b)

Figure 12: Examples of problematic frames in the PHOENIX14-T dataset.

C Stay-Step-Skip Alignment

Algorithm 1 S^3 Alignment

Require: Log probabilities log_probs $\in \mathbb{R}^{T\times C}$, target sequence $y_1, \ldots, y_U$, blank symbol
Ensure: Frame-level alignment Y with length T

```
Construct extended sequence ext with blanks inserted
Initialize DP table and backpointers
for each time step t do
    for each state s in ext do
        Compute max probability from stay (s), step (s − 1), skip (s − 2) transitions
    end for
end for
Backtrack from optimal final state to obtain alignment Y
return Y
```

D Comparison of action descriptions with the LLM knowledge base

Gloss names are part of the available supervised training labels. The role of the LLM is to integrate gloss order, anchors, temporal structure, and frame descriptions, not to recognize signs solely from gloss names or knowledge base.

Figure 13 shows the PHOENIX14-T frames and frame-level descriptions for "SAMSTAG" (Saturday) in German Sign Language, while Figure 14 shows the web version of DeepSeek's typical description of this sign. In external knowledge bases, "SAMSTAG" is a two-component action, whereas in PHOENIX14-T it is a single-component movement unrelated to either "week" or "6". This suggests that GTM does not rely on the LLM's prototypical sign knowledge, but grounds its decisions in the provided clues.

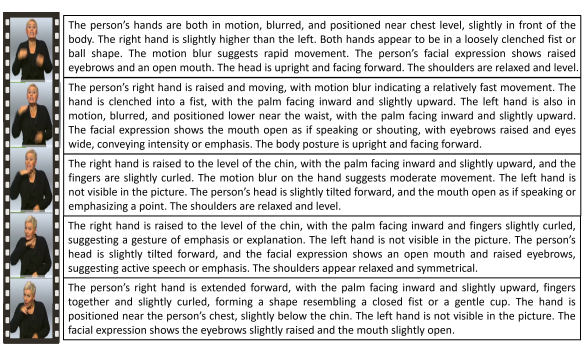


Figure 13: The action of "SAMSTAG" in PHOENIX14-T and the MLLM-generated frame-level descriptions.

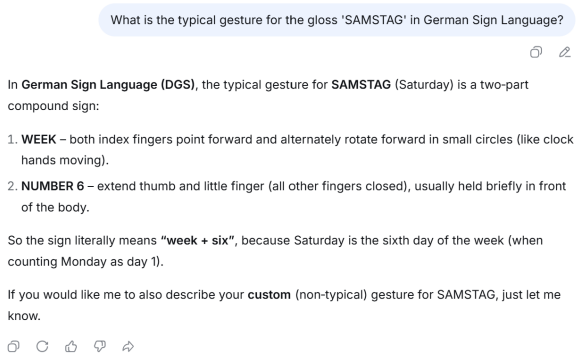


Figure 14: Description of the typical action for "SAMSTAG" in the online LLM knowledge base.

E Final Annotated Data Samples

Annotations are formatted as "index]gloss(start frame-end frame)". We annotate every training sample in CSL-Daily and PHOENIX14-T, preserving the original number and order. Our data is open-sourced at GTM, all annotations should be used only for research purposes, in compliance with the original dataset access conditions.

- 1]you all(6–30) 2]hello(33–51)
- 1]appendix(6–38) 2]inflammation(42–74) 3]often(77–104) 4]temptation(107–139) 5]other(142–152) 6]what(155–164) 7]illness(166–177)
- 1]weather(4–30) 2]winter(31–56) 3]son(57–88) 4]study(89–112) 5]more(113–135) 6]wear(137–144) 7]clothes(144–185)
- 1]NORD(1–22) 2]WOLKE(29–46) 3]IX(53–72) 4]ZWISCHEN(73–82) 5]SCHOTTLAND(85–122)
- 1]WIE-AUSSEHEN(1–12) 2]IN-KOMMEND(13–23) 3]DONNERSTAG(69–75) 4]AUCH(77–88) 5]REGEN(89–95) 6]IX(96–104) 7]SONST(105–111) 8]SUEDOST(112–136) 9]AUCH(138–142) 10]SONNE(145–156) 11]DABEI(159–170)
- 1]NACHT(2–22) 2]NEBEL(40–56) 3]HOCH(57–67) 4]NEBEL(69–92) 5]REGION(93–127)

F Implement Detail

All models are used as intended.

CSLR baseline. We follow the same settings as prior SOTA methods (Min et al., 2021; Hu et al., 2023a). We adopt ResNet-18 (He et al., 2016) with ImageNet (Deng et al., 2009) pretrained weights as the 2D CNN backbone. The 1D CNN is configured as a sequence of layers K5, P2, K5, P2, where Kσ and Pσ denote a 1D convolution layer and a pooling layer with kernel size σ. A two-layer BiLSTM with 512 units per direction is employed, followed by a fully connected layer with a 1024-dimensional input. The model is trained for 50 epochs with an initial learning rate of 1×10^{-4}, which is decayed by a factor of 5 at epochs 20, 30, and 40. The batch size is set to 2. We use the Adam optimizer (Kingma and Ba, 2017) with a weight decay of 1×10^{-4}. The PyTorch version is 2.5.0+cu121. During inference, we employ the widely used ctcdecoder[1] and set the beam width of beam search to 10, leaving other settings as default.

Data preprocessing. All input frames are first resized to 256×256, then random cropping to 224×224. During training, we apply random horizontal flipping with a probability of 50% and temporal scaling with a ratio of 20%. During inference, only a center crop of size 224×224 is used.

MLLM and LLM inference. We set the temperature to 0.3 and top$_p$ to 0.95. We call the MLLM once per frame and each LLM inference costs approximately 14k tokens per training example. For the LLM reasoning results, the proportion of single-sampling cases in which the number of generated segments is inconsistent with the number of glosses is about 10%. An LLM retry is triggered when the output cannot be parsed into the required JSON format or when the number of generated segments is inconsistent with the input gloss sequence. We allow at most nine retries for each training example; cases that still fail after these retries are manually inspected to determine the intervals. Since positioning points themselves are obtained from a baseline with a downsampling factor of 4 and do not overlap, and the intervals between time boundaries are larger than 4 frames, we did not observe any cases in which the LLM reasoning results produced "overly short intervals" or "overlapping intervals".

Runtime cost on a single RTX 4090. For CSL-Daily, generating frame-level descriptions takes approximately 40 days, LLM-based integration and reasoning takes approximately 20 days, training one recognition epoch takes about 30 minutes, and running inference on the test set once takes about 20 minutes. For PHOENIX14-T, generating frame-level descriptions takes approximately 25 days, LLM-based integration and reasoning takes

[1] https://github.com/parlance/ctcdecode

approximately 10 days, training one recognition epoch takes about 10 minutes, and running inference on the test set once takes about 10 minutes. Although the annotation stage is time-consuming, it is performed offline only once. The actual training and inference of the recognition model remain lightweight compared with SOTA models.

G Blind human rating of interval quality

The following are the specific scoring reasons provided by the 7 annotators. Among them, annotator A4 did not provide any reasons.

GTM:

- A1: Mostly accurate, with only minor start and end offsets within the same action.
- A2: Segment boundaries often align with gesture changes, but unsure if they truly correspond to gloss transitions (no sign language knowledge).
- A3: Not familiar with sign language, but boundaries show clearly better motion continuity than the other two.
- A5: The vast majority of deviations are very small; overall continuity is significantly better than the other two.
- A6: Intervals are longer than Group 3 and clearly outperform it.
- A7: Clearly the best overall quality.

Structure Extension:

- A1: Some segmentations have large deviations or missing/incorrect glosses.
- A2: Some actions fall outside intervals, though most are acceptable.
- A3: Some gloss boundaries fall at action peaks (rather than transitions).
- A5: 70% are good, but 30% have notable omissions.
- A6: Appears as a simple but somewhat rigid extension of Group 3.
- A7: Slightly worse than Group 1.

CTC Only:

- A1: Shorter intervals than Group 2, but overall similar in quality.
- A2: Accurate where present, but contributes little to final recognition.
- A3: Accurately placed but too peaky.
- A5: Accurate in placement, but intervals are too short, losing completeness — penalized for being "overly concise".
- A6: Baseline method.
- A7: Discontinuous; intervals are fragmented.

H LLM reasoning example

The following text is the thought process of the online version of DeepSeek, with the input Prompt B in Figure 4, which integrates positioning points, video structural points, and frame-level descriptions, to demonstrate the LLM's reasoning process. In this example, the LLM first uses the CTC-derived anchors to localize each gloss, then uses high-change frames only as structural references rather than mandatory boundaries, and finally relies on action continuity in the frame descriptions to decide whether a change point is an internal motion variation or a true gloss transition. For instance, it keeps frame 52 as an internal change within the first GRAD segment rather than switching immediately to SIEBZEHN, and leaves short transition gaps when the motion is ambiguous. This demonstrates that the generated intervals come from constrained temporal refinement instead of unconstrained sign recognition.

"In this fine-grained segmentation, I started with the coarse anchor points: MORGEN at frames 0-7, SECHS at frames 32-39, the first GRAD at frames 44-47, SIEBZEHN at frames 64-71, and the second GRAD at frames 76-79. I also received auxiliary reference information: the eight frame indices with the largest motion-pattern changes, specifically frames 35, 41, 45, 52, 59, 63, 70, and 74, which could serve as potential segmentation boundaries. However, the user explicitly noted that the coarse anchor points may not be completely accurate, the actions may not be the most prototypical realizations, and the frame descriptions may also be imperfect. Therefore, I placed greater emphasis on analyzing the motion continuity reflected in the frame descriptions, using the auxiliary boundary frames only as references. Analyzing the frame descriptions: from frame 1 to frame 31, the right hand gradually rises from the chest to the face, with the index finger eventually pointing to or touching the cheek and temple area, while the left hand never appears. This is a highly coherent motion sequence. The coarse anchor placed MORGEN at frames 0-7 is clearly too short, and the earliest auxiliary boundary frame at 35 occurs well after this motion ends, so I determined that MORGEN should cover frames 1 through 31. Next, frames 33 to 39 show a clear thumbs-up handshape with the thumb extended and the other four fingers curled, and the left hand begins to appear, which perfectly matches SECHS. The coarse anchor for SECHS is

precisely frames 32-39, while auxiliary boundary frame 35 falls inside this interval, which may be a misjudgment, so SECHS remains as frames 33-39. Then I addressed the first GRAD. The coarse anchor placed it at frames 44-47, but careful observation of frame descriptions reveals that from frame 41 onward, both hands appear simultaneously and open with palms facing forward or outward, performing an outward movement. Frame 45 is exactly one of the auxiliary boundary frames, indicating a significant motion-pattern change at this point, but this change is an internal peak within the GRAD action rather than a transition to another gloss. The difficulty arises in subsequent frames: from frame 49 onward, the hands gradually shift from an "opening outward" motion to a "moving inward toward each other" motion. Frame 52 is another auxiliary boundary frame, indicating another major motion-pattern change. At this point, I encountered a critical uncertainty—does frame 50 belong to the first GRAD or to SIEBZEHN? If I strictly analyze frame by frame, frame 49 still shows the right hand with palm forward, frame 51 shows no motion blur and the right hand still has palm forward, so the posture remains close to GRAD. Although frame 52 is a potential boundary, the handshape has not yet fully transitioned into the typical SIEBZEHN configuration of palms facing each other with reciprocal back-and-forth movement. The user explicitly stated that there is no need to pursue excessive precision down to a specific frame and that judgments should be based on motion continuity, and also emphasized that one gloss action may consist of several changing subparts. Therefore, instead of fixating on the exact frame within the transition zone, I made a holistic judgment based on motion continuity: from frame 41 to frame 60, the overall dominant feature is both hands open with palms outward or forward, and it is only after frame 60 that the motion clearly enters the reciprocal phase characteristic of SIEBZEHN. Even though the auxiliary boundary frames point to frame 52 as a change point, this change is merely an internal directional adjustment within the GRAD action rather than a signal to switch to the next gloss. Thus, I assigned frame 50 to GRAD and extended the first GRAD interval from the coarse anchor of frames 44-47 to frames 41 through 60, leaving frames 61 to 63 as an unassigned transition region. Next, I processed SIEBZEHN. The coarse anchor placed it at frames 64-71, and the frame descriptions show that during this period both hands have palms facing each other, fingers curled, moving back and forth. Frames 63 and 70 are both auxiliary boundary frames—frame 63 marks the formal entry from the transition into SIEBZEHN, and frame 70 is an internal motion change point within SIEBZEHN. Although this is not a typical action for SIEBZEHN, based on motion continuity, the SIEBZEHN interval aligns well with the anchor, so I kept it as frames 64-71. Finally, I processed the second GRAD. The coarse anchor placed it at frames 76-79, but frame descriptions show that from frame 73 onward, the hands are already in a posture with the right hand near the chin and the left hand extended outward. Frame 74 is an auxiliary boundary frame, indicating a clear motion-pattern change at this point, which marks the beginning of the second GRAD. This posture continues until frame 80, after which the motion begins to dissipate, so I extended the second GRAD from frame 73 to frame 80. All intervals are non-overlapping and strictly follow the order of the gloss sequence, yielding the final result: 1] MORGEN(1-31), 2] SECHS(33-39), 3] GRAD(41-60), 4] SIEBZEHN(64-71), 5] GRAD(73-80)."