WHEN FREQUENCY MEETS SIGN: EXPLOITING COMPLEMENTARY SIGNALS FOR VISION-ONLY SIGN LANGUAGE SENTENCE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing sign language translation (SLT) systems are developed for sentence-level video clips, limiting their ability to handle long videos containing multiple sentences in real-world scenarios. To bridge this gap, we introduce the *Vision-only Sign Language Sentence Segmentation (VSLSS)* setting, which partitions long multi-sentence videos into sentence-level clips solely from visual signals. As a preprocessing step, VSLSS allows existing sentence-level SLT models to process multi-sentence videos without retraining on multi-sentence data. However, sign language often lacks explicit motion pauses between sentences, making accurate sentence boundary localization from visual information alone challenging. To address this challenge, we propose **FVS**, a frequency-aware vision-only sign language sentence segmentation framework. To better exploit visual cues for sentence boundary localization, FVS introduces a Skeletal Motion Cepstral Coefficient (SMCC) extractor to capture additional high-frequency motion information beyond pose features. The pose and motion-frequency features are processed by two parallel branches and adaptively fused through a learnable Gated Fusion module. To alleviate under-segmentation, over-segmentation, and fragmented predictions, we further design Prior-Guided Losses (PGL). Extensive experiments demonstrate the effectiveness of FVS, and downstream experiments further validate its utility for sentence-level SLT.

1 INTRODUCTION

Sign languages, as the primary means of communication within the Deaf community, are naturally evolved and diversely structured systems in a rule-governed way. To bridge this communication gap, sign language translation (SLT) (Li et al., 2025; Camgoz et al., 2020; Zhou et al., 2022; Zhang et al., 2023; Chen et al., 2022; Zhou et al., 2023) has attracted much attention, which aims to convey the sign language video into the corresponding spoken language text. Historically, since benchmark datasets are built from isolated sentence clips, SLT has been framed as a sentence-level task. However, this framing does not align with the natural communication of deaf individuals, who express themselves through continuous, multi-sentence signing.

Facing this dilemma, there are two possible solutions, *i.e.*, collecting large-scale continuous signing data and scaling up SLT models, or segmenting long videos into sentence-level clips before translation. For the former, while large-scale models such as Google's SL2T [1] can process clips up to 60 seconds, they rely on massive expert annotation and computation. Due to limited computational resources, training a long-input-window SLT model from scratch is unaffordable for many researchers. In contrast, developing a segmentation model is a more practical and effective direction. Recently, Guo et al. (2025) first introduced the sentence-level sign language segmentation task. Their goal is to automatically expand sentence-level datasets from large-scale long sign language videos with captions. However, this setting is inconsistent with our goal of achieving multi-sentence SLT, where textual information is not accessible at inference.

To enable multi-sentence SLT, we introduce a new task named Vision-only Sign Language Sentence Segmentation (VSLSS), as illustrated in Fig 1. It aims to partition long multi-sentence sign language

[1] https://deepmind.google/blog/putting-sign-language-ai-into-users-hands/

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

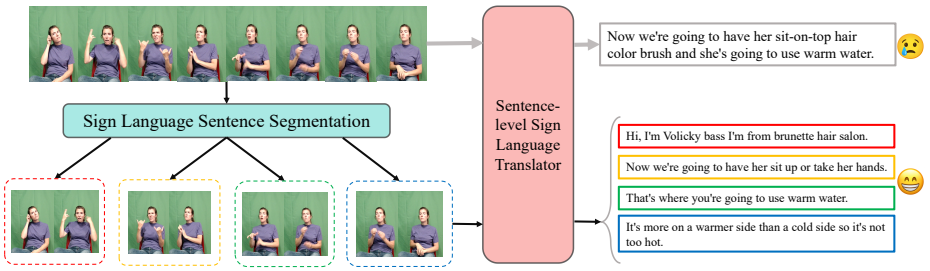

Figure 1: Illustration of vision-only Sign Language Sentence Segmentation (VSLSS). The gray box shows direct translation of the original multi-sentence video, while the colored boxes show translations obtained after VSLSS-based sentence segmentation.

videos into consecutive sentence-level clips without captions or subtitles at inference. Specifically, VSLSS is viewed as a frame-level classification task. VSLSS faces two critical challenges, *i.e.*, boundary ambiguity and inter-sentence confusion. The interweaving of signing movements and transitional movements causes boundary ambiguity. Adjacent sentences may be visually or semantically similar, leading to insufficiently discriminative features. The lack of textual cues makes them even harder to distinguish.

Effective segmentation requires capturing both the macroscopic semantic context ("what is signed") and the microscopic signal-level characteristics ("how it changes"). Sentence boundaries in sign language convey semantic completeness. Human perception of boundaries also relies on kinematic signal-level cues, such as abrupt transitions or velocity drops. Under the vision-only setting, selecting an effective video representation is crucial. Pose is a widely-used and effective representation for sign language, offering inherent advantages in computational efficiency and privacy preservation. However, existing temporal models such as TCNs and Transformers aggregate motion features indiscriminately. Since sign language signals are dominated by low-frequency macroscopic semantics, high-frequency boundary cues constitute only a minor fraction of the overall signal. Consequently, these models tend to smooth over these sparse high-frequency variations, failing to capture the signal-level characteristics necessary for precise boundary localization. This motivates us to develop a more effective representation and modeling strategy.

In this work, we propose the frequency-aware vision-only sign language sentence segmentation framework (FVS), a method based on complementary signals, *i.e.*, pose and derived frequency motion variations. The pose sequence captures macroscopic semantics, while the motion-frequency variations extract micro-kinematic high-frequency cues from the pose, serving as a critical complement by capturing how movement changes abruptly at sentence transitions. The proposed FVS consists of a pose-based branch, an SMCC-based motion-frequency branch and a Gated Fusion module. Specifically, the pose sequence is first processed to extract pose features. However, pose representations may insufficiently capture high-frequency motion patterns that are informative for sentence transitions. To capture motion-frequency information beyond the pose representation, we propose a Skeletal Motion Cepstral Coefficient (SMCC) Extractor that derives motion-frequency features from the pose sequence. The pose features and motion-frequency features are independently modeled to capture temporal dynamics and then adaptively integrated through a learnable *Gated Fusion* module. The resulting representation is finally used to predict frame-level inter-sentence transition probabilities for sentence boundary localization. To alleviate potential under-segmentation, over-segmentation, and fragmented predictions in VSLSS, we further propose Prior-Guided Losses (PGL) to provide stronger supervision during training.

Extensive experiments demonstrate the effectiveness of **FVS** for vision-only Sign Language Sentence Segmentation on three benchmark datasets, namely PHOENIX-2014T (Camgoz et al., 2018), How2Sign (Duarte et al., 2021), and OpenASL (Shi et al., 2022). We further evaluate its downstream utility for sentence-level sign language translation, showing that the proposed segmentation can effectively benefit downstream translation.

2 Related Works

2.1 Action Segmentation

Temporal action segmentation partitions long untrimmed videos into temporally coherent action segments and is typically formulated as frame-wise action classification (Lea et al., 2017). Fully supervised methods have evolved from temporal convolutional models such as TCN and MS-TCN (Lea et al., 2017; Farha & Gall, 2019) to multi-stage and Transformer-based architectures such as MS-TCN++, ASFormer, and FACT (Li et al., 2023; Yi et al., 2021; Lu & Elhamifar, 2024). More recently, diffusion models have also been explored for iterative action segmentation (Liu et al., 2023). Weakly supervised methods instead exploit ordered transcripts to infer latent temporal alignments without frame-level labels (Lu & Elhamifar, 2021; Kuehne et al., 2017). These methods provide important methodological foundations for modeling long-range temporal dependencies and reducing over-segmentation errors (Farha & Gall, 2019; Li et al., 2023). However, conventional action segmentation predicts an action identity for every frame, whereas VSLSS predicts the boundaries between sentence-level units and can therefore be formulated as binary boundary detection.

2.2 Sign Language Segmentation

Temporal segmentation has been studied in sign language at different linguistic granularities (Moryossef et al., 2023). Existing studies can be broadly distinguished into lexical-level segmentation, which aims to localize individual signs or glosses, and higher-level segmentation, which aims to identify boundaries between larger linguistic units such as phrases, utterances, or sentences.

Lexical-level and sign-level segmentation. Early studies mainly focused on segmenting continuous signing into individual signs or glosses. Yang et al. (Yang et al., 2010) investigated movement epenthesis and hand segmentation ambiguities in continuous sign language recognition and introduced a nested dynamic programming framework to handle transitional movements between signs. Farag and Brock (Farag & Brock, 2019) further explored frame-wise sign boundary detection from spatio-temporal motion features extracted from body-joint trajectories. Renz et al. (Renz et al., 2021a) introduced a temporal convolutional framework with iterative segment refinement to resolve ambiguous sign boundary cues, while their subsequent changepoint-modulated pseudo-labelling approach further exploited motion-sensitive boundary cues for adapting sign segmentation to unlabeled target domains (Renz et al., 2021b). More recently, graph- and Transformer-based methods have also been explored for boundary detection by explicitly modeling hand-joint relations and long-range temporal dependencies (Rastgoo et al., 2025). Although these approaches provide useful foundations for detecting temporal boundaries in continuous signing, they primarily operate at the lexical or sign level rather than segmenting complete sentences.

Sentence-level segmentation and temporal alignment. Higher-level temporal segmentation has increasingly focused on phrases and sentence-like units, often by incorporating linguistic or textual cues. Moryossef et al. (Moryossef et al., 2023) jointly modeled sign-level and phrase-level segmentation and investigated linguistic cues such as prosody and optical flow for locating phrase boundaries. In parallel, subtitle-video alignment methods have leveraged textual information as an additional semantic and temporal signal. Bull et al. (Bull et al., 2021) proposed a Transformer-based model that combines subtitle embeddings with visual features to predict frame-level alignment between subtitles and continuous signing. Rao et al. (Rao et al., 2025) further explored subtitle-level segmentation using aligned subtitles and optical flow. More recently, Guo et al. (Guo et al., 2025) explicitly formulated sentence-level sign language segmentation as the task of partitioning long sign language videos into consecutive sentence-level clips, and proposed a boundary-learning framework with caption-guided boundary optimization. Jiang et al. (Jiang et al., 2026) further investigated scalable subtitle-sign alignment by combining sign-level video segmentation, cross-modal embedding, and efficient temporal alignment across multiple datasets.

In contrast to these efforts, our work specifically focuses on the VSLSS setting, where sentence boundaries must be inferred from visual signing information without textual cues at inference.

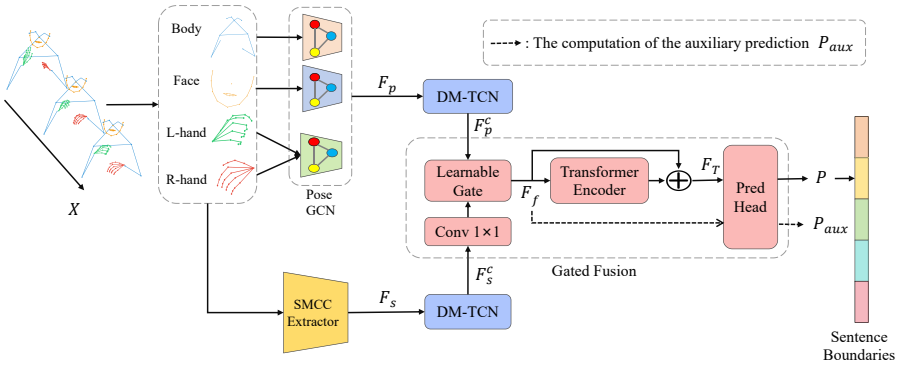

Figure 2: Framework of the proposed **FVS**. The input pose sequence $\mathbf{X}$ is processed by GCN and DM-TCN to obtain pose features $\mathbf{F}_p^c$, while the corresponding pose motion is transformed into SMCC sequence $\mathbf{F}_s$ and further modeled by DM-TCN to obtain SMCC features $\mathbf{F}_s^c$. The two representations are aligned and fused through a learnable gate, followed by a Transformer encoder for global temporal modeling. A shared prediction head produces auxiliary predictions $\mathbf{P}_{aux}$ and final transition predictions $\mathbf{P}$ under the proposed PGL.

3 Problem Definition

We introduce the Vision-only setting of the Sign Language Sentence Segmentation (VSLSS), which partitions a long, untrimmed sign language video into consecutive sentence-level clips solely from visual signing information. Given a video $\mathbf{V} = \{f_t\}_{t=1}^{T}$, the goal is to obtain N non-overlapping sentence-level clips $\{v_j\}_{j=1}^{N}$, where $v_j = \{f_{s_j}, \ldots, f_{e_j}\}$ denotes the j-th sentence segment, and s_j and e_j denote its start-frame and end-frame indices. Rather than representing a sentence boundary by a single frame, we model each boundary as a short inter-sentence transition region $\mathcal{R}_k = [l_k, r_k]$, since a transition may span multiple consecutive frames. These regions are converted into frame-level transition labels $\mathbf{Y} = \{y_t\}_{t=1}^{T}$, where $y_t = 1$ indicates that frame t belongs to an inter-sentence transition region and $y_t = 0$ denotes an intra-sentence frame. The model predicts a frame-level transition probability sequence $\mathbf{P} = \{p_t\}_{t=1}^{T}$, where p_t denotes the probability that frame t belongs to an inter-sentence transition region. From $\mathbf{P}$, we locate candidate transition regions and determine the final sentence boundaries.

Compared with the caption-assisted setting in Guo et al. (Guo et al., 2025), frame-level transition prediction can be formulated as $P_\theta(\mathbf{Y} \mid \mathbf{V}, \mathbf{Z})$, where $\mathbf{Z}$ denotes synchronized captions or subtitles. In contrast, our setting estimates the same transition labels solely from visual information, i.e., $P_\theta(\mathbf{Y} \mid \mathbf{V})$. At inference, neither textual cues nor the sentence count N are used.

4 Method

4.1 Overview

We propose **FVS**, a framework for VSLSS, as illustrated in Fig. 2. To eliminate background interference, we extract 2D pose sequence $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ and the corresponding joint confidence sequence $\mathbf{C} = \{\mathbf{c}_t\}_{t=1}^{T}$ from the input videos $\mathbf{V}$. Pose sequence $\mathbf{X}$ is decomposed into four regions, which are encoded by three GCN modules (Yan et al., 2018), with the hand regions sharing the same GCN parameters. The regional pose features $\mathbf{F}_p$ are modeled by DM-TCN to obtain the pose representation $\mathbf{F}_p^c$. Through a Skeletal Motion Cepstral Coefficient (SMCC) Extractor, the pose regions are converted into SMCC features $\mathbf{F}_s$ and processed by a DM-TCN to obtain the SMCC representation $\mathbf{F}_s^c$. After temporal alignment, $\mathbf{F}_p^c$ and $\mathbf{F}_s^c$ are fused by a learnable gate to obtain the fused representation $\mathbf{F}_f$, which is further processed by a Transformer encoder for global temporal modeling, yielding $\mathbf{F}_T$. A shared transition prediction head is applied to $\mathbf{F}_f$ and $\mathbf{F}_T$ to produce the auxiliary prediction $\mathbf{P}_{aux}$ and the final transition prediction $\mathbf{P}$. During training, Prior-Guided Losses (PGL) provide direct transition-region supervision together with temporal continuity and sentence-count consistency constraints.

4.2 SKELETAL MOTION CEPSTRAL REPRESENTATION

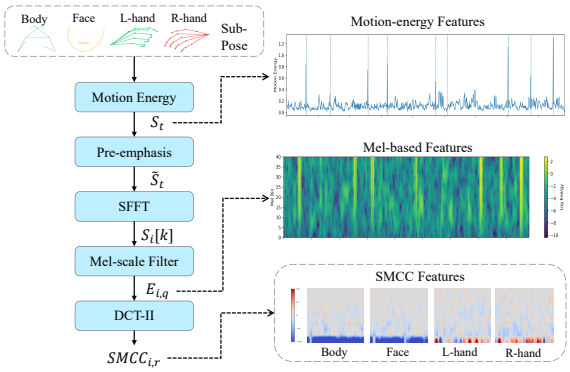

Figure 3: Overview of the proposed SMCC extraction pipeline. The right side shows visualizations of the corresponding motion-energy, Mel-based, and SMCC features.

To mitigate the limited temporal expressiveness of pose-based features, we derive a compact motion-frequency cepstral representation from confidence-weighted 2D skeletal motion. As illustrated in Fig. 3, the SMCC pipeline converts regional pose sequences into frequency features through motion-energy extraction, pre-emphasis, short-time spectral analysis, Mel filtering, logarithmic compression, and DCT-II. Given a skeleton sequence of T frames, let $\mathbf{x}_t \in \mathbb{R}^{N\times 2}$ denote the 2D coordinates of N joints and $\mathbf{c}_t \in \mathbb{R}^N$ their confidence scores. For consecutive frames, the confidence-weighted motion-energy signal is computed as $s_t = \sum_{n=1}^{N} \bar{c}_{t,n} \left\|\mathbf{x}_{t+1,n}\mathbf{x}_{t,n}\right\|_2$, where $c_{t,n} = \frac{c_{t,n}+c_{t+1,n}}{2}$. A zero value is prepended to preserve the original temporal length. To suppress the dominance of slowly varying components and relatively emphasize rapid motion changes that are informative for sentence transitions, we then apply first-order pre-emphasis, $\tilde{s}_t = s_t - \rho s_{t-1}$, where ρ is a pre-emphasis coefficient.

We then follow the standard MFCC-style signal processing pipeline (Davis & Mermelstein, 1980). The pre-emphasized motion signal is divided into overlapping short-time windows and multiplied by a Hamming window to reduce spectral leakage. For each window, we apply a Fast Fourier Transform (FFT), referred to as the Short-time Fast Fourier Transform (SFFT), and compute the corresponding power spectrum $\mathcal{S}_i[k]$. The resulting spectrum is then processed by a Mel filterbank, followed by logarithmic compression and DCT-II to obtain the SMCC representation. Finally, the resulting SMCC are:

$$\mathrm{SMCC}_{i,r} = \sum_{q=1}^{Q} \log(E_{i,q}+\epsilon)\cos\left(\frac{\pi r(q-\frac{1}{2})}{Q}\right), \qquad E_{i,q} = \sum_{k} \mathcal{S}_i[k]H_q(k), \tag{1}$$

where $H_q(k)$ denotes the q-th Mel filter, $\mathcal{S}_i[k]$ denotes the power spectrum obtained from the SFFT of the i-th short-time window, Q is the number of Mel filters, r indexes the SMCC coefficient, and ϵ is a small constant introduced for numerical stability. The resulting representation retains the computational structure of MFCCs, while replacing the speech waveform with a confidence-weighted skeletal motion signal, thereby adapting cepstral frequency modeling to sign language dynamics.

The skeleton is divided into four regions and SMCC is extracted independently for each region. The four regional streams are concatenated along the feature dimension to form the frequency sequence used by the SMCC branch.

4.3 DUAL-PATH MULTI-DILATION TEMPORAL CONVOLUTION MODELING

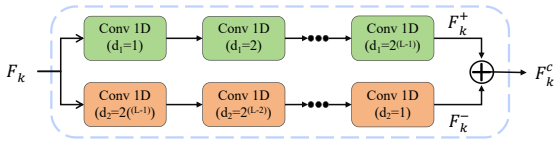

Figure 4: Detailed architecture of the proposed DM-TCN.

To capture complementary temporal dependencies at multiple receptive fields, we introduce a Dual-Path Multi-Dilation Temporal Convolution module (DM-TCN), as illustrated in Fig. 4. The DM-TCN employs two parallel convolutional paths with complementary dilation factors for encoding. Each path consists of L residual dilated temporal-convolutional blocks. In the first path, the dilation factor $d1$ increases exponentially with network depth, with $d1 \in \mathcal{D}^+ =$

$\{1, 2, 4, \ldots, 2^{L-1}\}$, enabling the capture of long-range dependencies. In contrast, the dilation factor in the second path $d2$ decays exponentially, with $d2 \in \mathcal{D}^- = \{2^{L-1}, \ldots, 4, 2, 1\}$, emphasizing local details. Let $\mathbf{F}_k$ denote the input feature sequence to the DM-TCN. The outputs of the two paths are denoted by $\mathbf{F}_k^+$ and $\mathbf{F}_k^-$, respectively, and are concatenated along the feature dimension to obtain $\mathbf{F}_k^c = [\mathbf{F}_k^+; \mathbf{F}_k^-]$, where the c denotes their concatenation.

The pose features $\mathbf{F}_p$ and SMCC features $\mathbf{F}_s$ are independently processed by the same DM-TCN architecture, yielding $\mathbf{F}_p^c$ and $\mathbf{F}_s^c$. We have $\mathbf{F}_p^c \in \mathbb{R}^{T \times D_p}$ and $\mathbf{F}_s^c \in \mathbb{R}^{T_s \times D_s}$, where $T_s < T$ because the SMCC branch is computed over short-time analysis windows.

4.4 Gated Fusion

This module consists of four stages: temporal and channel alignment of the pose and SMCC representations, learnable gated fusion, global temporal modeling with a Transformer encoder, and shared transition prediction. These components progressively integrate the pose and SMCC representations while preserving both local temporal information and global contextual dependencies.

To enable frame-wise fusion, we first transform the SMCC representation $\mathbf{F}_s^c$ to match the temporal resolution and channel dimensionality of the pose representation $\mathbf{F}_p^c$. Specifically, we apply temporal upsampling and a 1×1 convolution to $\mathbf{F}_s^c$, obtaining the transformed representation $\hat{\mathbf{F}}_s^c \in \mathbb{R}^{T \times D_p}$. The resulting $\mathbf{F}_p^c$ and $\hat{\mathbf{F}}_s^c$ are then fused using a learnable gate:

$$\mathbf{F}_f = \mathbf{g} \odot \mathbf{F}_p^c + (1 - \mathbf{g}) \odot \hat{\mathbf{F}}_s^c, \qquad \mathbf{g} = \sigma\Big(\mathrm{Conv}_{1\times1}\left([\mathbf{F}_p^c; \hat{\mathbf{F}}_s^c]\right)\Big). \tag{2}$$

As a result, the fused representation $\mathbf{F}_f$ adaptively balances pose representation and SMCC representation at each temporal position and channel.

The fused representation is then processed by a Transformer encoder (Vaswani et al., 2017) to model global temporal dependencies. We use a residual weight η to combine the fused representation and transformer representations to get the final representation $\mathbf{F}_T = (1-\eta)\mathbf{F}_f + \eta \cdot \mathrm{Transformer}(\mathbf{F}_f)$. We reuse the same transition prediction head for both prediction stages. The shared prediction head consists of a 1×1 convolution followed by a sigmoid function. The auxiliary and main predictions are given by:

$$\mathbf{P}_{aux} = \sigma\left(\mathrm{Conv}_{1\times1}(\mathbf{F}_f)\right), \qquad \mathbf{P} = \sigma\left(\mathrm{Conv}_{1\times1}(\mathbf{F}_T)\right), \tag{3}$$

where $\mathbf{P}_{aux}$ provides auxiliary supervision, while $\mathbf{P}$ is the final frame-level transition probability sequence.

During inference, the predicted transition probabilities are post-processed into sentence boundaries by thresholding and grouping consecutive transition frames, followed by selecting the maximum-probability frame within each candidate transition region.

4.5 Training Objective

The training objective is tailored to the VSLSS setting. Weighted binary cross-entropy (BCE) serves as the basic transition-region supervision. However, frame-wise classification alone does not explicitly enforce the agreement between the predicted transition region and the ground-truth region as a whole. Beyond BCE, we introduce Prior-Guided Losses (PGL), consisting of Dice loss, sentence-count loss, and temporal continuity loss. To alleviate overfitting, we further employ auxiliary supervision. For brevity, let p_t denote the transition probability at frame t from either of the two predictions.

BCE Loss. As the basic training objective, the weighted BCE directly encourages p_t to approach the target label y_t, while assigning a larger penalty to transition frames to compensate for their sparsity:

$$\mathcal{L}_{BCE} = -\frac{1}{T}\sum_{t=1}^{T}\left[w \cdot y_t \log p_t + (1 - y_t)\log(1 - p_t)\right], \tag{4}$$

where w denotes the positive-class weight for transition-region frames and is used to alleviate the class imbalance between transition and intra-sentence frames.

Dice Loss. To directly encourage a larger overlap between the predicted probability mass and the ground-truth transition region, we introduce the soft Dice loss:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{t=1}^{T} p_t y_t + \epsilon}{\sum_{t=1}^{T} p_t + \sum_{t=1}^{T} y_t + \epsilon}, \tag{5}$$

where ϵ is a small constant introduced for numerical stability and to avoid division by zero.

Sentence-Count Loss. To align the predicted sentence count with the ground truth, we introduce a sentence-count loss during training, which is computed as:

$$\mathcal{L}_{Count} = \left(\frac{\hat{N} - N}{N + 1}\right)^2, \qquad \hat{N} = (1 - p_1) + \sum_{t=1}^{T-1} p_t(1 - p_{t+1}), \tag{6}$$

where N is the ground-truth sentence count obtained by replacing the predicted probabilities p_t with the binary transition labels y_t in the same counting function, and $\hat{N}$ is the differentiable predicted sentence count. More details are provided in Appendix A.2.

Temporal Continuity Loss. We further impose a local temporal continuity loss to suppress fragmented transition responses within sentence interiors. Specifically, only adjacent frames that both belong to intra-sentence regions are regularized:

$$\mathcal{L}_{Temp} = \frac{\sum_{t=1}^{T-1} m_t \, |p_{t+1} - p_t|}{\sum_{t=1}^{T-1} m_t + \epsilon}, \qquad m_t = \mathbb{I}(y_t = 0 \wedge y_{t+1} = 0). \tag{7}$$

This constraint keeps neighboring intra-sentence predictions similar, reducing fragmented false transitions. Excluding transition regions prevents smoothing of true boundaries.

Auxiliary Supervision. We introduce auxiliary supervision at the intermediate feature level. The auxiliary prediction is supervised jointly with the final prediction using the same composite objective. For each prediction stage $k \in \{\text{aux}, \text{main}\}$, the composite loss is:

$$\mathcal{L}^{(k)} = \mathcal{L}^{(k)}_{BCE} + \alpha\mathcal{L}^{(k)}_{Dice} + \beta\mathcal{L}^{(k)}_{Temp} + \gamma\mathcal{L}^{(k)}_{Count}, \tag{8}$$

where $\mathcal{L}^{(\text{aux})}$ and $\mathcal{L}^{(\text{main})}$ are computed from the auxiliary prediction $\mathbf{P}_{aux}$ and the final prediction $\mathbf{P}$, respectively. The total training objective is:

$$\mathcal{L}_{Total} = \lambda\mathcal{L}^{(\text{aux})} + \mathcal{L}^{(\text{main})}, \tag{9}$$

where λ controls the contribution of the auxiliary supervision.

5 Experiments

5.1 Datasets

We evaluate VSLSS on PHOENIX-2014T (Camgoz et al., 2018), How2Sign (Duarte et al., 2021), and OpenASL (Shi et al., 2022). These datasets cover two different sources of long-form videos. How2Sign and OpenASL consist of naturally continuous sign language videos. In contrast, PHOENIX-2014T provides sentence-level sign language clips from German weather forecasts. We construct multi-sentence videos by grouping these clips according to their recording dates and concatenating them in their original sequence order. Following previous work (Guo et al., 2025), we adopt Variable-length Boundary Annotation (VBA) to convert sentence transitions into multi-frame transition regions for frame-level supervision. Details are provided in Appendix A.3.

5.2 Implementation Details

Experimental Setup We use the RTMW model (Jiang et al., 2024) from MMPose [2] as the 2D whole-body human pose extractor. All input videos are uniformly sampled at 25 FPS. We set the pose feature dimension to $D_{pose} = 1024$, the SMCC feature dimension to $D_{smcc} = 256$, and the

[2] https://github.com/open-mmlab/mmpose

fused feature dimension to $D_f = 1024$. The DM-TCN contains $L = 10$ residual dilated-convolution blocks in each temporal path. The Transformer encoder uses hidden dimension $D_T = 1024$, $H = 16$ self-attention heads, and 4 layers. We set the residual combination weight to $\eta = 0.5$. For the training objective, we set the transition-region weight to $w = 5$ and the loss weights to $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.0001$, and $\lambda = 0.2$. More details are provided in Appendix A.4.

Evaluation Metrics. Following established temporal action segmentation protocols (Lea et al., 2017; Farha & Gall, 2019), we evaluate sentence-level segmentation using segmental F1 at temporal overlap thresholds of 10%, 25%, and 50%, denoted as F1@10, F1@25, and F1@50, respectively. Higher F1 values indicate better segmentation performance. Additionally, we report the Segment Error Rate (SER) introduced for SLSS (Guo et al., 2025) to measure the discrepancy between the predicted and ground-truth numbers of sentence segments. Lower SER indicates a more accurate estimation of the number of sentence segments.

5.3 Overall Performance

To ensure a fair comparison, all methods are evaluated under the same dataset partitions, frame-level boundary annotations, input video setting, and evaluation protocol, with F1@10, F1@25, F1@50, and SER computed using the same criteria.

Method	PHOENIX-2014T				How2Sign				OpenASL			
	F1@10	F1@25	F1@50	SER	F1@10	F1@25	F1@50	SER	F1@10	F1@25	F1@50	SER
MS-TCN (Farha & Gall, 2019)	89.65	89.25	81.18	0.086	61.56	59.71	39.94	0.428	77.77	76.61	64.24	0.208
MS-TCN++ (Li et al., 2023)	90.77	90.37	81.82	0.086	69.06	67.37	48.65	0.428	76.23	74.96	62.22	0.245
ASFormer (Yi et al., 2021)	89.18	88.45	80.96	0.110	71.61	69.70	51.89	0.394	80.45	79.28	67.74	0.196
FACT (Lu & Elhamifar, 2024)	89.73	88.84	81.16	0.169	80.89	78.99	58.43	0.323	85.03	83.24	70.19	0.217
DiffAct (Liu et al., 2023)	–	–	–	–	82.76	80.41	60.35	0.294	81.25	78.90	59.92	0.317
SignBD w/o text (Guo et al., 2025)	91.85	90.33	84.62	0.121	80.31	78.02	59.85	0.252	81.90	80.25	67.33	0.103
FVS (ours)	**92.37**	**92.26**	**90.98**	**0.062**	**83.58**	**82.64**	**77.02**	**0.091**	**89.32**	**88.07**	**85.94**	**0.078**

Table 1: Performance comparison on the test sets of PHOENIX-2014T, How2Sign, and OpenASL. All methods use vision-only input.

Table 1 compares FVS with representative vision-only baselines, including temporal convolutional, Transformer-based, diffusion-based, and sign-language-specific approaches, on the test sets of PHOENIX-2014T, How2Sign, and OpenASL. Detailed development set results are provided in Appendix A.6. FVS achieves the best performance among the evaluated vision-only methods across all three datasets in F1@10, F1@25, F1@50, and SER.

The advantage is particularly pronounced at the strict F1@50 criterion. FVS achieves 90.98%, 77.02%, and 85.94% on PHOENIX-2014T, How2Sign, and OpenASL, respectively, outperforming the strongest vision-only baselines by 6.36, 16.67, and 15.75 percentage points. Moreover, FVS achieves the lowest SER on all three datasets, with values of 0.062, 0.091, and 0.078, indicating more accurate sentence-count estimation. Overall, FVS consistently achieves the best performance across F1@10, F1@25, F1@50, and SER on all three benchmarks, demonstrating its effectiveness for precise sentence-level boundary localization using visual signing information alone.

5.4 Ablation Study

We conduct ablation studies on the key components of FVS on the PHOENIX-2014T dev set.

Impact of Frequency Representation. To evaluate the effectiveness of the proposed frequency representation, we compare a series of representations with increasing levels of frequency-domain processing while keeping the remaining model components unchanged. Specifically, Motion Energy uses the confidence-weighted motion signal directly, SFFT-based further applies short-time Fourier analysis and logarithmic compression, Mel-based additionally introduces Mel-scale filtering, and SMCC further applies DCT to obtain a compact cepstral representation. As shown in Table 2, SMCC improves over the intermediate frequency representations, while combining pose and SMCC features achieves the best overall performance.

Impact of Dual-Path Multi-Dilation Temporal Convolution Module. To verify the effectiveness of the proposed dual-path multi-dilation temporal modeling, we compare the model without tem-

Representation	F1@10	F1@25	F1@50	SER
Pose-only	85.69	85.33	80.33	0.204
Motion Energy	84.29	83.28	72.55	0.188
SFFT-based	88.66	88.17	82.10	0.136
Mel-based	89.72	89.19	85.41	0.128
SMCC	91.67	91.44	86.12	0.125
Pose+SMCC	**91.90**	**91.56**	**90.05**	**0.088**

Table 2: Impact of frequency representation.

Cnt.	Dice	Temp	AUX	F1@10	F1@25	F1@50	SER
×	×	×	×	81.08	80.27	72.92	0.250
×				89.80	89.54	85.31	0.118
	×			88.09	87.98	82.30	0.154
		×		89.44	89.20	84.53	0.124
			×	85.98	84.85	80.15	0.197
				91.90	**91.56**	**90.05**	**0.088**

Table 3: Impact of the proposed Prior-Guided Losses (PGL). × indicates that the corresponding loss component is removed.

Variant	F1@10	F1@25	F1@50	SER
w/o TCN	71.28	70.03	62.21	0.336
Single-path TCN	86.72	86.48	81.87	0.156
Reverse-path TCN	84.93	84.80	78.50	0.214
DM-TCN	**91.90**	**91.56**	**90.05**	**0.088**

Table 4: Impact of multi-dilation temporal convolution module.

Fusion	F1@10	F1@25	F1@50	SER
Fixed $g = 0.5$	89.10	88.98	84.50	0.122
Concatenation + FC	86.76	86.64	82.33	0.145
Learnable Gate	**91.90**	**91.56**	**90.05**	**0.088**

Table 5: Impact of fusion strategy.

poral convolution, a single-path TCN with increasing dilation, a reverse-path TCN with decreasing dilation, and the proposed DM-TCN. All variants only change the temporal convolution structure. As shown in Table 4, DM-TCN achieves the best performance among the evaluated variants

Impact of Fusion Strategy. We compare fixed-weight fusion, concatenation with FC projection, and the proposed learnable gate. As shown in Table 5, the learnable gate achieves the best performance, while the fixed baseline uses $g = 0.5$ according to the sensitivity analysis in Appendix A.5.

Impact of Prior-Guided Losses. To verify the effectiveness of the proposed PGL, we individually evaluate the sentence-count loss, Dice loss, temporal continuity loss, and auxiliary supervision while retaining weighted BCE as the base transition-region supervision in all variants. Table 3 shows that the complete PGL configuration achieves the best performance among the evaluated loss combinations, demonstrating the effectiveness of jointly incorporating these supervision terms for VSLSS.

Downstream Impact on Sign Language Translation. Table 6 evaluates the downstream utility of VSLSS using the same pretrained Uni-Sign model (Li et al., 2025), with only the input video segmentation varied. GT-S uses ground-truth sentence clips, MS uses the original unsegmented long videos, and PFS-S uses clips produced by FVS. FVS improves Test BLEU-4 from 2.98 for MS to 13.35, approaching the 14.49 obtained with GT-S, demonstrating the utility of vision-only sentence segmentation for downstream SLT.

Input	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE
GT-S	39.38	26.44	19.21	14.49	34.51
MS	11.29	6.75	4.39	2.98	26.65
PFS-S	37.49	23.91	16.83	13.35	30.52

Table 6: Downstream SLT performance of Uni-Sign on the How2Sign test set.

6 Conclusion

In this work, we study the *vision-only* setting of Sign Language Sentence Segmentation (SLSS), which partitions long multi-sentence sign language videos into consecutive sentence-level clips without textual input. We propose **FVS**, a pose-based framework that combines the proposed Skeletal Motion Cepstral Coefficients (SMCC), dual-path multi-dilation temporal convolution module, and learnable gated temporal-frequency fusion. We further introduce Prior-Guided Losses (PGL) to incorporate temporal continuity and sentence-count constraints into boundary prediction. FVS achieves strong vision-only performance on PHOENIX-2014T, How2Sign, and OpenASL, especially at strict temporal-overlap thresholds. Downstream experiments further show that accurate sentence-level segmentation improves long-form SLT.

AI USE STATEMENT

In this work, generative AI tools were used solely to assist with language editing and refinement of the manuscript. All scientific content, including the research methodology, experimental design, theoretical claims, results, figures, and conclusions, was developed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses publicly released sign language datasets, including PHOENIX-2014T, How2Sign, and OpenASL, and does not involve new human participant recruitment, interviews, or additional personal data collection. We process the released video and annotation data solely for research purposes and do not attempt to identify individual signers. We acknowledge that the composition of existing sign language datasets may not fully represent the diversity of languages, signing communities, domains, and recording conditions, and therefore our results should not be interpreted as universally representative of sign language users.

REPRODUCIBILITY STATEMENT

We provide sufficient implementation and experimental details to support reproduction of our results. The main paper specifies the problem formulation, model architecture, SMCC construction, temporal modeling modules, fusion strategy, training configuration, loss functions, dataset construction protocols, and evaluation metrics. Additional implementation and experimental details are provided in the appendix. To further facilitate reproducibility, we plan to release the source code and preprocessing scripts following publication.

REFERENCES

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning subtitles in sign language videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 11552–11561, 2021.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7784–7793, 2018.

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

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5120–5130, 2022.

Steven B. Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, 28:357–366, 1980.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2735–2744, 2021.

Iva Farag and Heike Brock. Learning Motion Disfluencies for Automatic Sign Language Segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7360–7364, 2019.

Yazan Abu Farha and Juergen Gall. MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3575–3584, 2019.

Bowen Guo, Shiwei Gan, Yafeng Yin, Xiao Liu, Zhiwei Jiang, and Shunmei Meng. Sentence-level Segmentation for Long Sign Language Videos with Captions. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 3428–3437, 2025.

Tao Jiang, Xinchen Xie, and Yining Li. RTMW: Real-Time Multi-Person 2D and 3D Whole-body Pose Estimation. *arXiv preprint arXiv:2407.08634*, 2024.

Zifan Jiang, Youngjoon Jang, Liliane Momeni, Gül Varol, Sarah Ebling, and Andrew Zisserman. Segment, embed, and align: A universal recipe for aligning subtitles to signing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 30371–30384, 2026.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-Body Human Pose Estimation in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Hilde Kuehne, Alexander Richard, and Juergen Gall. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding (CVIU)*, 163:78–89, 2017.

Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal Convolutional Networks for Action Segmentation and Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 156–165, 2017.

Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45:6647–6658, 2023.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-Sign: Toward Unified Sign Language Understanding at Scale. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion Action Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 10139–10149, 2023.

Zijia Lu and Ehsan Elhamifar. Weakly-Supervised Action Segmentation and Alignment via Transcript-Aware Union-of-Subspaces Learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 12926–12936, 2021.

Zijia Lu and Ehsan Elhamifar. FACT: Frame-Action Cross-Attention Temporal Modeling for Efficient Action Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18175–18185, 2024.

Amit Moryossef, Zifan Jiang, Mathias Muller, Sarah Ebling, and Yoav Goldberg. Linguistically Motivated Sign Language Segmentation. In *Findings of the Association for Computational Linguistics (EMNLP Findings)*, pp. 12703–12724, 2023.

Kavu Maithri Rao, Yasser Hamidullah, and Eleftherios Avramidis. Sign Language Video Segmentation Using Temporal Boundary Identification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL Workshops)*, pp. 1213–1224, 2025.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. A non-anatomical graph structure for boundary detection in continuous sign language. *Scientific Reports*, 15:25683, 2025.

Katrin Renz, Nicolaj C. Stache, Samuel Albanie, and Gül Varol. Sign Language Segmentation with Temporal Convolutional Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2135–2139, 2021a.

Katrin Renz, Nicolaj C. Stache, Neil Fox, Gül Varol, and Samuel Albanie. Sign segmentation with changepoint-modulated pseudo-labelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3403–3412, 2021b.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Ruiduo Yang, Sudeep Sarkar, and Barbara Loeding. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:462–477, 2010.

Fangqiu Yi, Hongyu Wen, and Tingting Jiang. ASFormer: Transformer for Action Segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 236, 2021.

Biao Zhang, Mathias Müller, and Rico Sennrich. SLTUNET: A Simple Unified Model for Sign Language Translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 20871–20881, 2023.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation. *IEEE Transactions on Multimedia (TMM)*, pp. 768–779, 2022.

A Appendix

A.1 Framework Implementation

Keypoints extraction. We employ the RTMW (Jiang et al., 2024) from MMPose to extract whole-body keypoints. The visualization of whole-body keypoints is shown in Figure 5. As mentioned in Figure 2, we divide the pose into sub-poses (left hand, right hand, face, and body). We select the indices for the left hand ({92-112}), right hand ({113-133}), body ({1, 4-11}), and face ({24, 26, 28, 30, 32, 34, 36, 38, 40, 54, 84-91}) to represent each group. Additionally, we select 92, 113, and 54 as the root indices for the hands and face to normalize the keypoints, while the body is not normalized using root coordinates.

Pose GCN We detail the output dimensions of pose GCN in Table 7. The body, face, and hand groups use separate encoder parameters, while the left-hand and right-hand regions share the same parameters. Accordingly, we use three parameter sets for the body, face, and hands to model the three groups independently.

Layer	Dimensions	Temporal kernel
Linear	64	None
Pose encoder	[64, 128, 256]	None
Temporal encoder	[256, 256, 256]	5

Table 7: Output dimension of each layer in pose GCN.

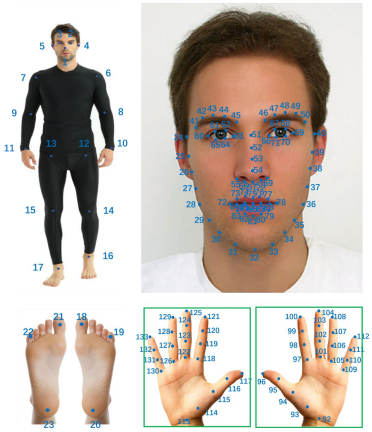

Figure 5: The visualization of the whole-body 133 keypoints, derived from (Jin et al., 2020).

A.2 SENTENCE-COUNT LOSS

In the vision-only setting, the predicted number of sentence segments is unconstrained, and the model tends to over-segment or under-segment. Transition-region supervision only optimizes local frame-level accuracy, so it cannot directly penalize global count errors. Therefore, we add a sentence-count loss during training, which uses the ground-truth sentence count as supervision. To align the predicted sentence count with the ground truth, we need the number of segments implied by the predictions. Directly counting them requires thresholding and grouping, which is non-differentiable. We instead compute the count from the same state-transition functional for both the binary labels and the continuous probabilities:

$$\mathcal{L}_{Count} = \left(\frac{\hat{N} - N}{N + 1} \right)^2, \qquad \hat{N} = (1-p_1) + \sum_{t=1}^{T-1} p_t(1-p_{t+1}), \qquad N = (1-y_1) + \sum_{t=1}^{T-1} y_t(1-y_{t+1}), \tag{10}$$

where $(1 - y_1)$ identifies a sentence beginning at the first frame, and each $y_t(1 - y_{t+1})$ identifies a transition from an inter-sentence region to the subsequent intra-sentence region. Specifically, the term $p_t(1 - p_{t+1})$ captures the transition from an inter-sentence region to an intra-sentence region. For intra-sentence frames ($y_t = 0$), p_t approaches 0; for inter-sentence frames ($y_t = 1$), p_t approaches 1. Thus, $p_t(1 - p_{t+1})$ is close to 1 only when $p_t \approx 1$ and $p_{t+1} \approx 0$, which corresponds to a transition from an inter-sentence region to the subsequent intra-sentence region. In contrast, the other combinations—intra-to-intra ($p_t \approx 0, p_{t+1} \approx 0$), inter-to-inter ($p_t \approx 1, p_{t+1} \approx 1$), and intra-to-inter ($p_t \approx 0, p_{t+1} \approx 1$)—all yield $p_t(1 - p_{t+1}) \approx 0$. Adding $(1 - p_1)$ accounts for a sentence that begins at the first frame when $p_1 \approx 0$. Therefore, replacing the binary labels y_t with continuous probabilities p_t yields the differentiable predicted sentence count $\hat{N}$. Then, the predicted sentence count $\hat{N}$ can be directly aligned with the ground-truth count N without additional sentence-count annotations.

Dataset	Train		Dev		Test	
	Samples	Avg. sentences	Samples	Avg. sentences	Samples	Avg. sentences
PHOENIX-2014T	822	8.07	102	8.27	104	7.52
How2Sign	3720	6.52	461	5.71	464	5.48
OpenASL	11023	6.68	1378	6.69	1379	6.79

Table 8: Dataset statistics

A.3 DATASET CONSTRUCTION

We construct long-form multi-sentence videos from the three datasets using dataset-specific procedures. For PHOENIX-2014T, which provides sentence-level sign language clips from German weather forecasts, we follow the original recording order and concatenate clips according to their

Configuration	Value
Optimizer	AdamW
Base learning rate	2×10^{-4}
Weight decay	1×10^{-4}
AdamW betas	$(0.9, 0.999)$
Learning rate schedule	Cosine decay
Training epochs	20
Batch size	4
Gradient accumulation steps	4

Table 9: Training configuration of FVS.

Configuration	Value
Pose feature dimension	1024
SMCC feature dimension	256
Fused feature dimension	1024
DM-TCN layers per path	10
DM-TCN dilation order	$1, 2, 4, \ldots, 2^{L-1}$ / reverse
Transformer encoder layers	4
Transformer attention heads	16
Transformer dropout	0.2
Transformer residual weight η	0.5
SMCC pre-emphasis ρ	0.97
SMCC SFFT size N_{SFFT}	256
Mel filters Q	40
Cepstral coefficients R	20
Window length	0.8 s
Hop size	0.2 s
ϵ	1e-6

Table 10: Model architecture and SMCC extraction configuration of FVS.

recording dates and continuous sequence indices. This ordering preserves the original temporal sequence and avoids introducing artificial discontinuities between adjacent clips.

How2Sign and OpenASL, in contrast, contain naturally continuous long-form videos, some of which are too long to be directly processed by the model in a single pass. We therefore partition each original video into contiguous 1–2 minute segments while preserving the original temporal order. The resulting segments are non-overlapping and gap-free, so that their union provides complete temporal coverage of the source video. Each segment contains multiple consecutive sentences.

Following Guo et al. (Guo et al., 2025), we adopt Variable-length Boundary Annotation (VBA) to construct frame-level transition labels. Instead of assigning a single frame to each sentence boundary, VBA expands the transition region around each sentence boundary by a number of frames proportional to the corresponding sentence duration, thereby providing multi-frame transition supervision and alleviating the severe class imbalance between transition and intra-sentence frames. We use the same annotation proportion $V = 0.2$ as in (Guo et al., 2025). Frames within the resulting transition regions are assigned label 1, while the remaining intra-sentence frames are assigned label 0. During inference, consecutive frames with transition probabilities above the threshold are grouped into candidate transition regions, from which the frame with the highest transition probability is selected as the final sentence boundary.

The resulting dataset statistics, including the number of samples and the average number of sentences per sample, are reported in Table 8.

A.4 Implementation Details and Configuration

We provide the main training, model, representation, and loss configurations to facilitate reproducibility. Unless otherwise specified, all experiments use the same optimization and model configuration. The training configuration is summarized in Table 9, the model architecture and SMCC

Configuration	Value
Boundary weight w	5
Dice weight α	0.4
Temporal weight β	0.3
Count weight γ	0.0001
Auxiliary loss weight λ	0.2
Inference threshold	0.5
ϵ	1e-6

Table 11: Hyperparameters of the Prior-Guided Losses and inference procedure.

extraction parameters are reported in Table 10, and the hyperparameters of the Prior-Guided Losses and inference procedure are provided in Table 11.

A.5 Sensitivity to Fixed Fusion Weights

We further analyze the sensitivity of fixed fusion weights by evaluating $g \in \{0.1, 0.2, \ldots, 0.9\}$ while keeping all other model components and training settings unchanged. As shown in Table 12, $g = 0.5$ achieves the best performance among all evaluated fixed weights, with 89.10%, 88.98%, and 84.50% F1 at 10%, 25%, and 50% IoU thresholds, respectively, together with the lowest SER of 0.122. The results also show that fixed fusion is sensitive to the choice of g, especially when the weight is biased toward either the pose or SMCC branch. We therefore use $g = 0.5$ as the fixed-weight baseline in the main ablation study, while the proposed learnable gate avoids manually selecting a fixed fusion weight.

Fixed g	F1@10	F1@25	F1@50	SER
0.1	83.58	83.33	78.49	0.190
0.2	82.83	82.16	76.41	0.227
0.3	86.84	86.21	81.93	0.146
0.4	83.96	83.58	79.33	0.183
0.5	**89.10**	**88.98**	**84.50**	**0.122**
0.6	74.24	73.56	64.24	0.277
0.7	77.93	77.25	70.46	0.253
0.8	70.98	70.00	59.43	0.325
0.9	71.50	70.25	60.75	0.294

Table 12: Sensitivity of fixed fusion weights on the PHOENIX-2014T development set.

A.6 Complete Results of Experiments

Due to page limitations, some experimental results have been omitted from the main paper. To facilitate future research in referencing the relevant results, the complete experimental results are provided in Table 13, Table 14 and Table 15.

Method	Dev				Test			
	F1@10	F1@25	F1@50	SER	F1@10	F1@25	F1@50	SER
MS-TCN (Farha & Gall, 2019)	89.07	88.61	80.66	0.133	89.65	89.25	81.18	0.086
MS-TCN++ (Li et al., 2023)	89.01	88.34	79.83	0.091	90.77	90.37	81.82	0.086
ASFormer (Yi et al., 2021)	88.62	87.73	79.67	0.132	89.18	88.45	80.96	0.110
FACT (Lu & Elhamifar, 2024)	90.74	89.36	80.78	0.131	89.73	88.84	81.16	0.169
SignBD w/o text (Guo et al., 2025)	90.76	89.81	82.47	0.124	91.85	90.33	84.62	0.121
FVS (ours)	**91.90**	**91.56**	**90.05**	**0.088**	**92.37**	**92.26**	**90.98**	**0.062**

Table 13: Performance comparison on PHOENIX-2014T.

Comparison with Caption-Assisted SLSS To further examine the differences between vision-only and caption-assisted SLSS, we compare FVS with the caption-assisted SignBD model on the

Method	Dev				Test			
	F1@10	F1@25	F1@50	SER	F1@10	F1@25	F1@50	SER
MS-TCN (Farha & Gall, 2019)	60.61	58.45	38.04	0.558	61.56	59.71	39.94	0.428
MS-TCN++ (Li et al., 2023)	69.29	67.62	48.03	0.406	69.06	67.37	48.65	0.428
ASFormer (Yi et al., 2021)	71.34	69.41	51.43	0.385	71.61	69.70	51.89	0.394
FACT (Lu & Elhamifar, 2024)	80.26	78.30	56.66	0.340	80.89	78.99	58.43	0.323
DiffAct (Liu et al., 2023)	83.12	80.94	60.87	0.299	82.76	80.41	60.35	0.294
SignBD w/o text (Guo et al., 2025)	79.43	76.51	58.77	0.263	80.31	78.02	59.85	0.252
FVS (ours)	**84.87**	**84.19**	**78.60**	**0.087**	**83.58**	**82.64**	**77.02**	**0.091**

Table 14: Performance comparison on How2Sign.

Method	Dev				Test			
	F1@10	F1@25	F1@50	SER	F1@10	F1@25	F1@50	SER
MS-TCN (Farha & Gall, 2019)	76.60	75.25	62.61	0.505	77.77	76.61	64.24	0.208
MS-TCN++ (Li et al., 2023)	74.65	73.23	60.72	0.257	76.23	74.96	62.22	0.245
ASFormer (Yi et al., 2021)	79.31	78.12	65.83	0.206	80.45	79.28	67.74	0.196
FACT (Lu & Elhamifar, 2024)	84.77	83.12	69.89	0.229	85.03	83.24	70.19	0.217
DiffAct (Liu et al., 2023)	81.68	79.39	60.40	0.321	81.25	78.90	59.92	0.317
SignBD w/o text (Guo et al., 2025)	81.47	80.03	66.42	0.107	81.90	80.25	67.33	0.103
SignBD w/ text† (Guo et al., 2025)	88.68	86.33	73.69	0.016	89.19	86.91	74.66	0.002
FVS (ours)	**89.39**	**88.26**	**86.17**	**0.055**	**89.32**	**88.07**	**85.94**	**0.078**

Table 15: Performance comparison on OpenASL.

development and test sets of PHOENIX-2014T and How2Sign. As shown in Table 16, SignBD benefits from the additional textual cues and achieves higher F1@10 and F1@25 on both datasets and splits, as well as lower SER on PHOENIX-2014T. In contrast, FVS consistently achieves higher F1@50 on both datasets, with gains of 5.50 and 3.40 percentage points on the development and test sets of PHOENIX-2014T, and 10.22 and 7.84 percentage points on How2Sign, respectively. FVS also achieves lower SER than SignBD on How2Sign, with values of 0.087 vs. 0.128 on the development set and 0.091 vs. 0.098 on the test set. These results show that the two approaches exhibit different strengths across evaluation criteria, while FVS achieves competitive sentence boundary localization using visual information alone.

Dataset	Method	Dev				Test			
		F1@10	F1@25	F1@50	SER	F1@10	F1@25	F1@50	SER
PHOENIX-2014T	SignBD w/ text†	**92.99**	**91.79**	84.55	**0.036**	**93.94**	**92.65**	87.58	**0.036**
	FVS (ours)	91.90	91.56	**90.05**	0.088	92.37	92.26	**90.98**	0.062
How2Sign	SignBD w/ text	**88.34**	**85.91**	68.38	0.128	**89.11**	**86.21**	69.18	0.098
	FVS (ours)	84.87	84.19	**78.60**	**0.087**	83.58	82.64	**77.02**	**0.091**

Table 16: Comparison between FVS and caption-assisted SignBD on the development and test sets of PHOENIX-2014T and How2Sign.

A.7 Qualitative Analysis of SMCC

To further examine whether the proposed SMCC captures motion variations related to sentence boundaries, we visualize the temporal changes of the SMCC features in Fig. 6. Specifically, for each body region, we first average the SMCC coefficients along the feature dimension and then compute the absolute first-order temporal difference. The horizontal axis corresponds to video time, while the red dashed lines indicate the ground-truth sentence boundaries. As shown in Fig. 6, prominent peaks consistently occur around most ground-truth boundaries across the body, face, and hand regions. This qualitative observation suggests that the proposed SMCC is sensitive to temporal motion changes near sentence transitions and provides boundary-related information beyond the original pose representation.

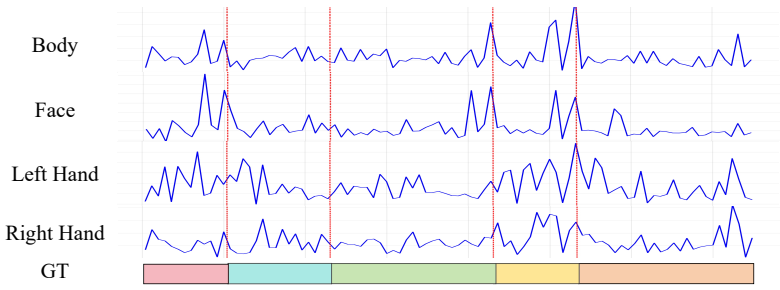

Figure 6: Qualitative visualization of temporal changes in the proposed SMCC features. For each body region, the SMCC coefficients are averaged along the feature dimension, followed by the absolute first-order temporal difference. The horizontal axis denotes video time, the vertical axis indicates the magnitude of temporal variation, and red dashed lines mark the ground-truth sentence boundaries. Peaks are generally concentrated around sentence boundaries, indicating that SMCC captures motion variations informative for boundary localization.