Concept Paper

SignFuse: A Proposed Dual-Stream Cross-Modal Framework for Gloss-Free Sign Language Translation with Large Language Models

Gurpreet Singh * and Purva Mundada

Posted Date: 15 April 2026

doi: 10.20944/preprints202604.1065.v1

Keywords: sign language translation; Multimodal Large Language Models; cross-modal fusion; Graph Convolutional Networks; hierarchical temporal modeling; position paper

 doi:10.20944/preprints202604.1065.v1

Concept Paper

SignFuse: A Proposed Dual-Stream Cross-Modal Framework for Gloss-Free Sign Language Translation with Large Language Models

Gurpreet Singh [1,*] and Purva Mundada [2]

[1] Graduated, Endicott College of International Studies, Woosong University, Republic of Korea
[2] HR Department, JSPM University, Pune, India
* Correspondence: gurpreetsinghmcse@gmail.com

Abstract

Sign language translation (SLT) aims to convert sign language videos into spoken language text, serving as a critical bridge for communication between the Deaf and hearing communities. While recent advances in Multimodal Large Language Models (MLLMs) have shown promising results in gloss-free SLT, existing methods typically rely on single-modality visual features, failing to fully exploit the complementary nature of appearance and structural cues inherent in sign language. In this architectural proposition paper, we introduce **SignFuse**, a novel dual-stream cross-modal fusion framework that synergistically combines CNN-based visual features with Graph Convolutional Network (GCN)-based skeletal features for gloss-free sign language translation. Our framework introduces three key innovations: (1) a **Cross-Modal Fusion Attention (CMFA)** module that performs bidirectional cross-attention between visual and skeletal modalities to produce enriched multimodal representations; (2) a **Hierarchical Temporal Aggregation (HTA)** mechanism that captures sign language dynamics at multiple temporal scales—frame-level, segment-level, and sequence-level; and (3) a **Progressive Multi-Stage Training** blueprint that systematically aligns visual-skeletal features with the LLM's linguistic space through contrastive pre-training, feature alignment, and LoRA-based fine-tuning. We provide the complete mathematical formulation, detailed architectural specifications, and a fully implemented PyTorch codebase. As the computational barriers to training MLLMs remain high, we formalize the experimental methodology required to validate this framework on standard benchmarks (PHOENIX-14T, CSL-Daily, How2Sign) and extend an open invitation to the broader research community to conduct empirical validation and advance this architectural paradigm through collaboration. This work is presented as a concept and architectural framework paper, aiming to establish a theoretical foundation and encourage future empirical validation by the research community.

Keywords: sign language translation; Multimodal Large Language Models; cross-modal fusion; Graph Convolutional Networks; hierarchical temporal modeling; position paper

1. Introduction

Sign language is the primary mode of communication for over 70 million Deaf individuals worldwide [1]. Unlike spoken languages, sign languages are visual-gestural systems that convey meaning through a rich combination of manual articulations (hand shapes, orientations, and movements), non-manual markers (facial expressions, mouthing, and eye gaze), and body posture [2]. Despite the linguistic complexity and cultural significance of sign languages, the communication barrier between sign language users and the hearing majority remains a significant societal challenge.

The rapid evolution of artificial intelligence from early symbolic systems to modern deep learning paradigms [3] has enabled transformative progress across numerous domains, including Sign Language Translation (SLT), which aims to automatically translate sign language videos into spoken language text. SLT is fundamentally more challenging than Sign Language Recognition (SLR), as it requires

 doi:10.20944/preprints202604.1065.v1

not only identifying individual signs but also understanding grammatical structures, handling word reordering between sign and spoken languages, and generating fluent natural language output [4].

Traditional SLT approaches rely on an intermediate gloss representation a written annotation of individual signsas a bridge between the visual and linguistic domains [2,4]. However, gloss annotations suffer from several critical limitations: (i) they are expensive and time-consuming to produce, requiring trained linguistic annotators; (ii) they are linguistically imprecise, as different sign languages have distinct grammatical structures from their spoken counterparts; and (iii) they fail to capture the full complexity of sign language, including non-manual features and simultaneous multi-channel information. These limitations have motivated a growing body of research on *gloss-free* SLT, which directly translates sign language videos into spoken language without intermediate glosses [5,6].

The recent emergence of Large Language Models (LLMs) [7,8] and, more broadly, Multimodal Large Language Models (MLLMs)as comprehensively surveyed in [9]has catalyzed significant progress in gloss-free SLT. Models such as SignLLM [10], Sign2GPT [11], and SpaMo [12] have demonstrated that the powerful language generation capabilities of LLMs can be effectively harnessed for sign language translation when provided with appropriate visual representations. More recently, MMSLT [13] leverages multimodal LLMs to generate textual descriptions of sign language components, while SCOPE [14] introduces context-aware SLT for dialogue scenarios.

Despite these advances, we identify two critical limitations in existing approaches that motivate our work:

Limitation 1: Underutilization of complementary modalities. Current LLM-based SLT methods predominantly rely on a single visual modalityeither RGB frames processed through CNNs or Vision Transformers [11,12], or pose-based skeletal data [15]. Sign language, however, is inherently multi-modal, encoding information through multiple simultaneous channels. Appearance-based features from RGB frames capture hand shapes, finger configurations, and facial expressions, while skeletal features capture joint positions, movement trajectories, and limb articulations. These two modalities provide complementary information that is essential for accurate translation. No existing LLM-based SLT approach effectively fuses these complementary modalities at a deep feature level through learned cross-modal interactions.

Limitation 2: Insufficient temporal modeling. Sign language operates at multiple temporal granularities: individual hand configurations are captured within single frames, signs typically span 5–30 frames, and sentences require understanding long-range dependencies across hundreds of frames. Existing approaches typically use a single temporal resolution through fixed-stride feature sampling or simple temporal attention, which inadequately captures the hierarchical temporal dynamics of sign language. This limitation is particularly acute for translating compound sentences, temporal modifiers, and discourse-level structures.

To address these limitations, we propose **SignFuse** (**Sign** Language **Fuse**d Translation), a novel theoretical and architectural framework that introduces three key innovations:

- **Dual-Stream Feature Extraction:** We employ a CNN-based visual stream (ResNet-50) and a GCN-based skeletal stream (ST-GCN) to extract complementary features. The visual stream captures appearance information (*e.g.*, hand shapes, facial expressions), while the skeletal stream captures structural and kinematic information (*e.g.*, joint angles, movement trajectories).
- **Cross-Modal Fusion Attention (CMFA):** We introduce a novel bidirectional cross-attention module that enables each modality to attend to the other, producing enriched multimodal representations that leverage the strengths of both visual and skeletal features through learned gated aggregation.
- **Hierarchical Temporal Aggregation (HTA):** We design a multi-scale temporal modeling module that captures sign language dynamics at frame, segment, and sequence levels, enabling the model to understand both fine-grained gestures and long-range contextual dependencies through adaptive temporal fusion.

 doi:10.20944/preprints202604.1065.v1

Furthermore, we propose a **progressive multi-stage training strategy** that systematically: (i) pre-aligns visual and skeletal features via contrastive learning, (ii) aligns fused multimodal features to the LLM's input space, and (iii) performs end-to-end fine-tuning with LoRA [16] for efficient adaptation. An overview of the proposed framework is shown in Figure 1.

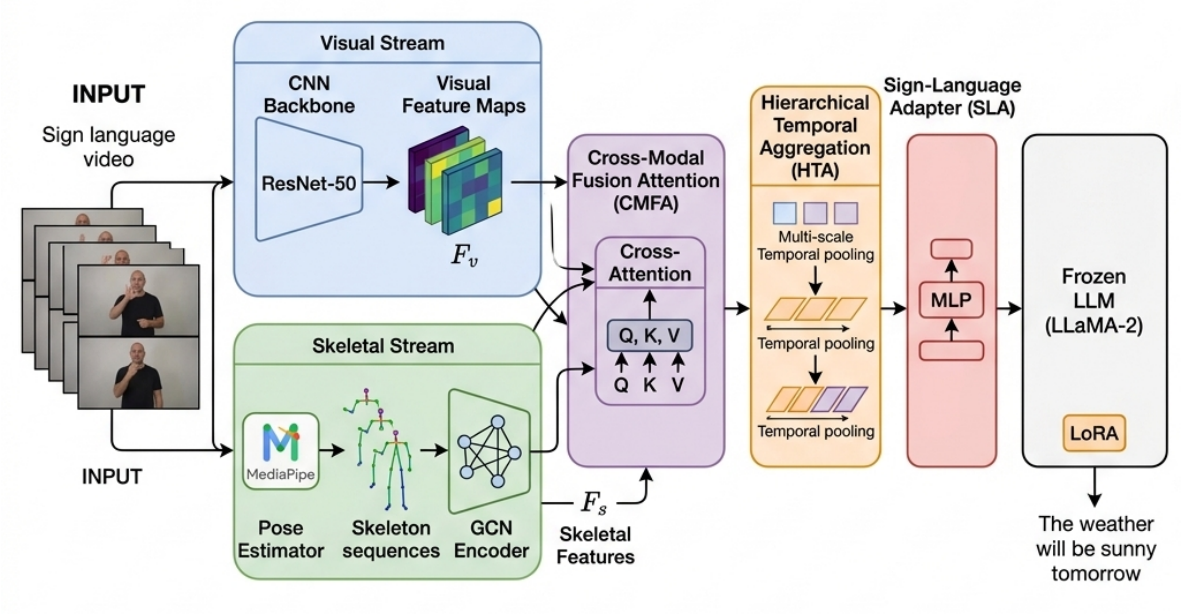

Figure 1. Overview of the proposed SignFuse Framework. Our dual-stream architecture extracts visual features via a CNN backbone and skeletal features via a pose estimator followed by a GCN encoder. The CMFA module performs bidirectional cross-attention between the two modalities. Fused features are processed by the HTA module before mapping to the LLM input space through the SLA.

Because training end-to-end MLLMs on massive video datasets requires substantial GPU compute clusters (typically 4–8× A100 GPUs for days of continuous training), empirical validation remains a barrier for many fundamental researchers. Therefore, this paper is presented as a formal architectural blueprint and research hypothesis. We provide the complete mathematical formulation, detailed architectural specifications, and a fully implemented PyTorch codebase. We outline the complete structural design of **SignFuse** and propose a rigorous experimental protocol for future evaluation. We open-source our theoretical framework to the community, inviting institutions with the necessary computational resources to execute the training phases and collaborate with us on validating this highly promising paradigm.

Contributions. Our contributions are summarized as follows:

1. We propose **SignFuse**, a novel dual-stream architecture that, for the first time, integrates CNN-based visual features and GCN-based skeletal features through deep bidirectional cross-modal attention for LLM-based sign language translation.
2. We introduce the Cross-Modal Fusion Attention (CMFA) module with gated aggregation and the Hierarchical Temporal Aggregation (HTA) mechanism, both of which are formally specified with complete mathematical formulations.
3. We design a three-stage progressive training strategy that systematically bridges the gap between multimodal sign language representations and the linguistic space of LLMs.
4. We provide a complete, runnable PyTorch implementation and formalize a comprehensive experimental protocol on three standard benchmarks, extending an open call for community-driven empirical validation.

 doi:10.20944/preprints202604.1065.v1

2. Related Work

2.1. Sign Language Recognition

Sign Language Recognition (SLR) has progressed through several paradigm shifts driven by advances in computer vision and deep learning. Early approaches relied on handcrafted features such as HOG descriptors and color histograms, combined with Hidden Markov Models (HMMs) for temporal modeling [17]. The introduction of Convolutional Neural Networks (CNNs) brought significant improvements, with methods like Recurrent Convolutional Neural Networks [18] combining spatial feature extraction with temporal sequence modeling via LSTMs [19]. Three-dimensional CNNs [20] extended this approach by jointly modeling spatial and temporal dimensions.

The Transformer architecture [21] has increasingly dominated the SLR landscape, with works such as [4] and [22] demonstrating strong sequence modeling capabilities through self-attention mechanisms. More recently, ConSignformer [23] introduced a Transformer backbone specifically designed for continuous SLR, incorporating convolutional position encoding and efficient attention patterns optimized for video understanding.

Skeleton-Based Approaches. Graph Convolutional Networks (GCNs) have emerged as powerful tools for skeleton-based sign language recognition [24]. By naturally representing the body as a graph with joints as nodes and bones as edges, GCNs exploit the structural relationships between body parts. The Spatial-Temporal Graph Convolutional Network (ST-GCN) [25] extended this paradigm by incorporating temporal edges between corresponding joints across frames, enabling joint spatial-temporal reasoning. These approaches have shown particular promise for capturing the kinematic properties of signs, such as trajectory shapes and velocity profiles.

MediaPipe-Based Systems. Lightweight pose estimation frameworks like MediaPipe Holistic [26] have enabled real-time skeleton extraction, making pose-based approaches more accessible. Recent works [27,28] combine MediaPipe landmarks with CNN-LSTM hybrid architectures, achieving competitive recognition accuracy with significantly reduced computational cost compared to methods operating directly on raw RGB frames. Object detection frameworks like YOLO have also been adapted for hand gesture detection in sign language recognition [29]. Comprehensive surveys [30,31] provide thorough analysis of the relative strengths and weaknesses of these diverse approaches.

2.2. Gloss-Free Sign Language Translation

A critical distinction in SLT is between *gloss-based* and *gloss-free* approaches. Gloss-based methods use intermediate gloss representations as a bridge between the visual and textual domains. The seminal work of Camgöz *et al.* [2] introduced the first neural SLT system using a CNN-RNN encoder-decoder architecture with gloss supervision. Subsequent works [4] employed Transformer architectures for improved sequence-to-sequence translation.

Gloss-free methods bypass the gloss bottleneck by learning direct mappings from video to text. GFSLT-VLP [6] leverages visual-language pretraining to bridge the modality gap without gloss supervision. GASLT [32] introduces a gloss attention mechanism that implicitly discovers sign units without explicit gloss annotations. These methods are increasingly preferred because: (i) gloss annotation is prohibitively expensive for large-scale datasets; (ii) glosses are not standardized across different sign languages; and (iii) gloss representations inherently lose information about non-manual features and prosodic elements that are critical for accurate translation.

2.3. Large Language Models and Vision–Language Models for Sign Language

The integration of LLMs into SLT represents the most significant recent development in the field. As foundation models for natural language understanding have scaled dramatically in both parameters and capabilities [9], researchers have explored increasingly sophisticated approaches to connect visual sign language representations with the powerful text generation capabilities of these models.

SignLLM [10] introduces a Vector-Quantized Visual Sign (VQ-Sign) module that discretizes continuous visual features into a codebook of sign tokens, which are then processed by a frozen LLM.

 doi:10.20944/preprints202604.1065.v1

This approach benefits from the discrete nature of the representation, which aligns well with the LLM's tokenized input format, but may lose fine-grained continuous visual information during quantization.

Sign2GPT [11] represents a complementary approach that employs pseudo-gloss pretraining to teach visual encoders meaningful sign representations before connecting them to a frozen GPT model via lightweight linear adapters. By first learning a pseudo-gloss vocabulary through unsupervised clustering, Sign2GPT creates an intermediate representation that bridges the gap between visual features and language tokens.

SpaMo [12] proposes an efficient approach using separate spatial and motion encoders to capture complementary visual features. The spatial encoder processes individual frames for appearance understanding, while the motion encoder captures inter-frame dynamics. These features are integrated via a Sign Adapter and processed by an LLM with LoRA fine-tuning. SpaMo achieves strong performance on PHOENIX-14T but does not incorporate structural skeletal information.

MMSLT [13] takes a different approach by leveraging MLLMs to generate textual descriptions of sign language components, including hand shapes, body postures, and facial expressions. These descriptions are then fused with video features for translation. This method uniquely exploits the visual understanding capabilities of MLLMs but introduces additional computational overhead from the description generation step.

SCOPE [14] introduces context-aware SLT for dialogue scenarios by incorporating conversational history into the translation process. **SignAlignLM** [33] focuses on natively integrating sign language support into LLMs through multitasking paradigms.

Contrastive Learning Approaches. SignCLIP [15] and CLIP-SLA [34] explore contrastive learning to create shared embedding spaces between sign language videos and text, drawing on the successful CLIP paradigm. These approaches demonstrate that alignment-based pretraining can significantly improve downstream SLT performance.

More broadly, multimodal vision–language models have evolved rapidly as foundational architectures for tasks requiring joint visual and textual reasoning [35], providing the theoretical and practical grounding for adapting such models to the sign language domain.

Gap Analysis. Despite the significant progress outlined above, no existing LLM-based SLT approach effectively combines both visual (RGB) and skeletal (pose) modalities through deep cross-modal attention mechanisms. SpaMo uses spatial and motion features from the *same* visual modality, while MMSLT generates textual descriptions rather than performing feature-level fusion. Our proposed **SignFuse** architecture addresses this gap by introducing bidirectional cross-attention between complementary visual and skeletal feature streams, drawing on the architectural principles established across the broader vision–language model landscape [9,35].

2.4. Multimodal Fusion Strategies

Cross-modal fusion has been extensively studied in general vision-language understanding [36–38]. Fusion strategies can be broadly categorized as follows:

Early Fusion concatenates raw or low-level features from different modalities before processing, which is simple but fails to capture modality-specific patterns before fusion.

Late Fusion processes each modality independently through separate encoders and combines predictions at the decision level, which preserves modality-specific representations but cannot model cross-modal interactions.

Cross-Attention Fusion employs attention mechanisms [21] to enable one modality to attend to information in another, allowing dynamic, input-dependent feature combination. This approach has proven particularly effective in models like Flamingo [38] and ALBEF [37].

Our CMFA module adopts a bidirectional cross-attention design that combines the advantages of cross-attention fusion with a learned gating mechanism, specifically designed for the unique requirements of visual-skeletal fusion in sign language where both modalities provide complementary but partially overlapping information.

 doi:10.20944/preprints202604.1065.v1

3. Proposed Methodology: SignFuse

3.1. Problem Formulation

Given a sign language video $\mathbf{V} = \{v_1, v_2, \ldots, v_T\}$ consisting of T RGB frames, where each frame $v_t \in \mathbb{R}^{3\times H\times W}$, and corresponding skeleton sequences $\mathbf{S} = \{s_1, s_2, \ldots, s_T\}$, where each skeleton $s_t \in \mathbb{R}^{V\times C}$ contains V joint coordinates in C dimensions, our goal is to generate a spoken language sentence $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$ with N tokens. This can be formulated as maximizing the conditional probability:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{V}, \mathbf{S}; \Theta) \tag{1}$$

where Θ denotes all learnable parameters. As illustrated in Figure 1, the **SignFuse** framework decomposes this problem into five components: (1) a Visual Feature Stream, (2) a Skeletal Feature Stream, (3) a Cross-Modal Fusion Attention (CMFA) module, (4) a Hierarchical Temporal Aggregation (HTA) module, and (5) a Sign-Language Adapter (SLA) connected to a frozen LLM.

3.2. Visual Feature Stream

The visual stream processes raw RGB frames to capture appearance-level features including hand shapes, finger configurations, facial expressions, and spatial layouts. We propose utilizing a pre-trained ResNet-50 [39] as the backbone, chosen for its strong feature extraction capabilities, widespread availability, and computational efficiency.

For each frame v_t, we extract a spatial feature vector from the global average pooling layer:

$$\mathbf{F}_v^t = \text{GAP}(\text{ResNet50}(v_t)) \in \mathbb{R}^{d_v} \tag{2}$$

where $d_v = 2048$ is the output dimension of ResNet-50's final convolutional block. The complete visual feature sequence is $\mathbf{F}_v = \{\mathbf{F}_v^1, \mathbf{F}_v^2, \ldots, \mathbf{F}_v^T\} \in \mathbb{R}^{T\times d_v}$.

To reduce computational cost while preserving temporal information, we apply a 1D temporal convolution with batch normalization and activation:

$$\hat{\mathbf{F}}_v = \text{ReLU}(\text{BN}(\text{Conv1D}_{k=3,s=2}(\mathbf{F}_v))) \in \mathbb{R}^{T'\times d} \tag{3}$$

where $T' = \lceil T/2 \rceil$ is the temporally downsampled length, $d = 512$ is the unified feature dimension, and $s = 2$ denotes the stride. This projection simultaneously reduces the temporal resolution and aligns the feature dimension with the skeletal stream for downstream fusion.

3.3. Skeletal Feature Stream

The skeletal stream captures structural and kinematic information about body joint configurations and their temporal dynamics. Unlike RGB features that encode appearance, skeletal features explicitly represent the spatial topology of the signer's body, providing invariance to visual nuisances such as clothing, lighting, and background clutter.

Pose Extraction. We propose using MediaPipe Holistic [26] to extract 2D landmarks from each frame, yielding 21 hand landmarks per hand (42 total), 33 body pose landmarks, and optionally 468 face mesh landmarks. For computational efficiency and focusing on the most translation-relevant signals, we select $V = 75$ key joints: both hands (42), upper body pose (23), and key facial landmarks (10). Each joint is represented by its (x, y) normalized coordinates, giving $s_t \in \mathbb{R}^{75\times 2}$.

Graph Construction. We construct a spatial-temporal skeleton graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_s \cup \mathcal{E}_t)$ where $\mathcal{V}$ represents the set of body joints, $\mathcal{E}_s$ represents spatial edges following the natural skeleton connectivity (bone connections), and $\mathcal{E}_t$ represents temporal edges connecting the same joint across consecutive frames.

ST-GCN Encoder. We employ a multi-layer Spatial-Temporal Graph Convolutional Network [25] to process the skeleton sequences. Each ST-GCN layer performs spatial graph convolution followed by temporal convolution:

 doi:10.20944/preprints202604.1065.v1

$$\mathbf{F}_s^{(l+1)} = \sigma\left(\sum_{k=1}^{K} \tilde{\mathbf{D}}_k^{-\frac{1}{2}} \tilde{\mathbf{A}}_k \tilde{\mathbf{D}}_k^{-\frac{1}{2}} \mathbf{F}_s^{(l)} \mathbf{W}_k^{(l)}\right) \tag{4}$$

where $\tilde{\mathbf{A}}_k = \mathbf{A}_k + \mathbf{I}$ is the k-th normalized adjacency matrix partition with self-connections, $\tilde{\mathbf{D}}_k$ is the corresponding degree matrix, $\mathbf{W}_k^{(l)} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ are learnable parameters, and σ denotes the ReLU activation function. We use $K = 3$ spatial partitions (self, inward, outward neighbors).

We stack $L = 4$ ST-GCN blocks with progressively increasing channel dimensions $[2, 64, 128, 256, 512]$, with a temporal stride of 2 in the third block to align the temporal resolution with the visual stream. Joint-level features are aggregated via global average pooling over the joint dimension:

$$\hat{\mathbf{F}}_s = \text{AvgPool}_{\text{joints}}(\text{STGCN}(\mathbf{S})) \in \mathbb{R}^{T' \times d} \tag{5}$$

3.4. Cross-Modal Fusion Attention (CMFA)

The core theoretical contribution of **SignFuse** is the CMFA module (Figure 2), designed to perform deep bidirectional cross-modal fusion between visual and skeletal features. Unlike simple concatenation or element-wise addition used in baseline models, CMFA enables each modality to *selectively attend* to complementary information in the other modality, producing enriched representations that capture cross-modal correlations.

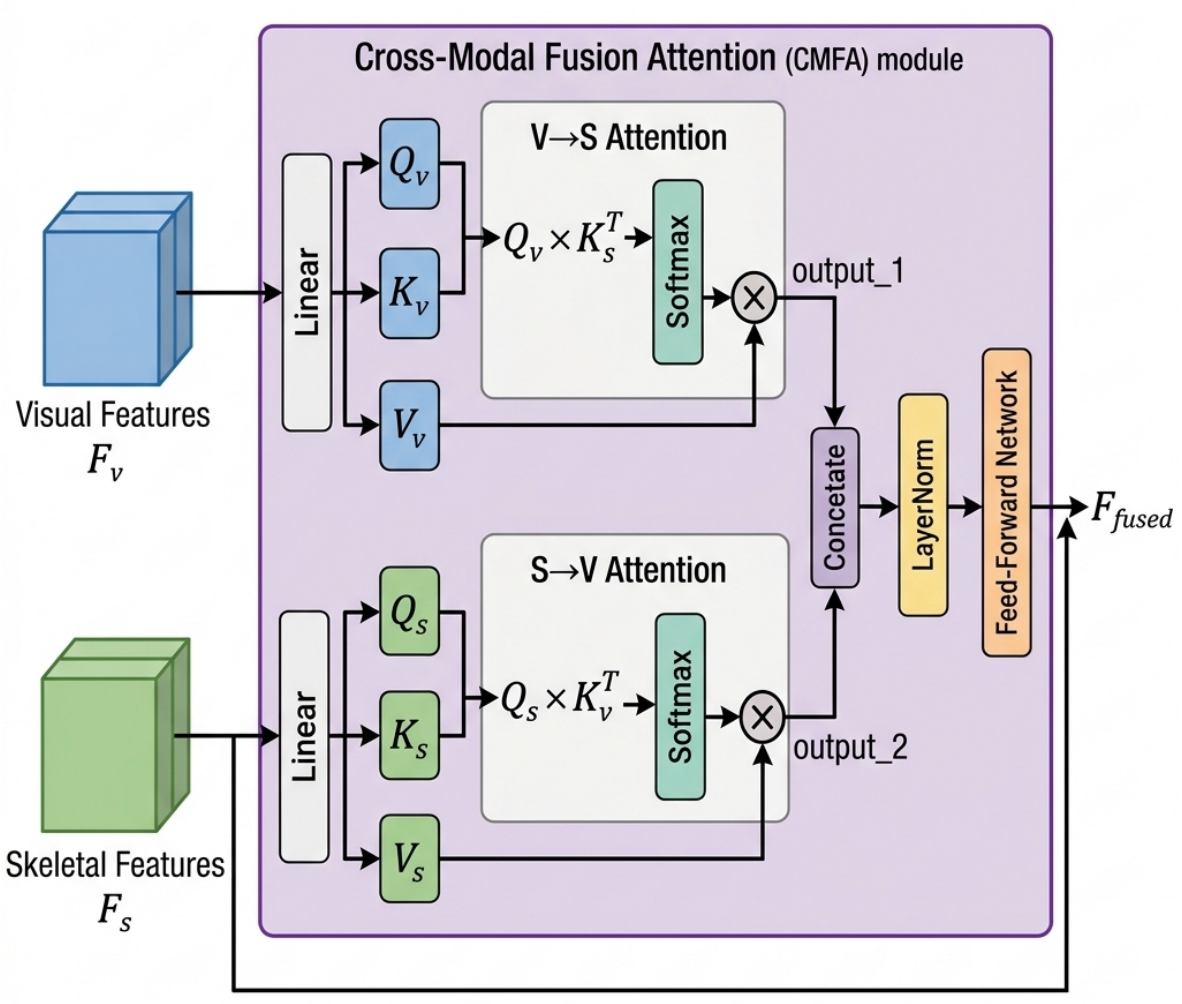

Figure 2. Proposed Cross-Modal Fusion Attention (CMFA) Module. Bidirectional cross-attention enables each modality to attend to complementary features in the other, producing enriched multimodal representations via gated aggregation.

 doi:10.20944/preprints202604.1065.v1

Multi-Head Cross-Attention. For each temporal position, we compute queries, keys, and values from both modalities using separate linear projections:

$$\mathbf{Q}_v = \hat{\mathbf{F}}_v \mathbf{W}_Q^v, \quad \mathbf{K}_v = \hat{\mathbf{F}}_v \mathbf{W}_K^v, \quad \mathbf{V}_v = \hat{\mathbf{F}}_v \mathbf{W}_V^v \tag{6}$$

$$\mathbf{Q}_s = \hat{\mathbf{F}}_s \mathbf{W}_Q^s, \quad \mathbf{K}_s = \hat{\mathbf{F}}_s \mathbf{W}_K^s, \quad \mathbf{V}_s = \hat{\mathbf{F}}_s \mathbf{W}_V^s \tag{7}$$

where $\mathbf{W}_Q^{(\cdot)}, \mathbf{W}_K^{(\cdot)}, \mathbf{W}_V^{(\cdot)} \in \mathbb{R}^{d \times d}$ are learnable projection matrices. The bidirectional cross-attention is computed as:

$$\mathbf{O}_{v \to s} = \text{softmax}\left(\frac{\mathbf{Q}_v \mathbf{K}_s^\top}{\sqrt{d_k}}\right)\mathbf{V}_s \tag{8}$$

$$\mathbf{O}_{s \to v} = \text{softmax}\left(\frac{\mathbf{Q}_s \mathbf{K}_v^\top}{\sqrt{d_k}}\right)\mathbf{V}_v \tag{9}$$

where $d_k = d/h$ is the per-head dimension and $h = 8$ is the number of attention heads. $\mathbf{O}_{v \to s}$ represents visual features enriched by attending to skeletal information ("what does the skeleton tell me about what I'm seeing?"), while $\mathbf{O}_{s \to v}$ represents skeletal features enriched by visual context ("what does the visual appearance tell me about the joint configuration?").

Gated Aggregation. Rather than simply concatenating or averaging the two cross-attended outputs, we employ a learned gating mechanism that dynamically determines the optimal mixing ratio for each temporal position:

$$\alpha = \sigma(\mathbf{W}_g[\mathbf{O}_{v \to s}; \mathbf{O}_{s \to v}] + b_g) \tag{10}$$

$$\mathbf{F}_{\text{fused}} = \text{LN}(\alpha \cdot \mathbf{O}_{v \to s} + (1 - \alpha) \cdot \mathbf{O}_{s \to v}) + \mathbf{F}_{\text{res}} \tag{11}$$

where σ is the sigmoid function, $\mathbf{W}_g \in \mathbb{R}^{2d \times d}$ is the gate projection, LN denotes Layer Normalization, and $\mathbf{F}_{\text{res}} = (\hat{\mathbf{F}}_v + \hat{\mathbf{F}}_s)/2$ is the residual connection ensuring gradient flow and information preservation.

Feed-Forward Network. Following standard Transformer practice, we apply a position-wise feed-forward network after the fusion:

$$\mathbf{F}'_{\text{fused}} = \text{LN}(\mathbf{F}_{\text{fused}} + \text{FFN}(\mathbf{F}_{\text{fused}})) \tag{12}$$

where $\text{FFN}(\mathbf{x}) = \mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1 \mathbf{x} + b_1) + b_2$ with inner dimension $4d = 2048$.

We stack $L = 3$ CMFA layers to enable progressively deeper cross-modal interactions. After the first layer, both inputs are derived from the fused representation, allowing the model to iteratively refine the cross-modal alignment.

3.5. Hierarchical Temporal Aggregation (HTA)

Sign language conveys meaning at multiple temporal granularities, which necessitates multi-scale temporal modeling. We identify three distinct temporal scales:

- **Frame-level (fine-grained):** Individual hand configurations, finger spelling, and instantaneous facial expressions occur at the frame level (25–30ms per frame at 30fps). These are critical for distinguishing between signs with similar global motions but different hand shapes.
- **Segment-level (mid-grained):** Individual sign units typically span 5–30 frames (150–1000ms). Capturing segment-level dynamics enables recognition of transitional movements, sign boundaries, and phonological components.
- **Sequence-level (coarse-grained):** Sentence-level structures, discourse patterns, and long-range dependencies extend across hundreds of frames. Understanding these is essential for generating grammatically correct, contextually appropriate translations.

 doi:10.20944/preprints202604.1065.v1

Given the fused features $\mathbf{F}_{\text{fused}} \in \mathbb{R}^{T' \times d}$, we apply depthwise separable temporal convolutions at three scales:

$$\mathbf{H}_{\text{frame}} = \text{DWConv1D}_{k=1}(\mathbf{F}_{\text{fused}}) \in \mathbb{R}^{T' \times d} \tag{13}$$

$$\mathbf{H}_{\text{seg}} = \text{DWConv1D}_{k=5,s=2}(\mathbf{F}_{\text{fused}}) \in \mathbb{R}^{T'/2 \times d} \tag{14}$$

$$\mathbf{H}_{\text{seq}} = \text{DWConv1D}_{k=11,s=4}(\mathbf{F}_{\text{fused}}) \in \mathbb{R}^{T'/4 \times d} \tag{15}$$

We employ adaptive average pooling to align all scales to a uniform temporal length $T'' = 64$, then concatenate along the feature dimension:

$$\mathbf{H}_{\text{multi}} = [\text{Pool}_{T''}(\mathbf{H}_{\text{frame}}); \text{Pool}_{T''}(\mathbf{H}_{\text{seg}}); \text{Pool}_{T''}(\mathbf{H}_{\text{seq}})] \tag{16}$$

A linear projection combines the multi-scale features, followed by multi-head temporal self-attention for scale integration:

$$\mathbf{H}_{\text{final}} = \text{LN}(\mathbf{H}_{\text{combined}} + \text{MHSA}(\mathbf{H}_{\text{combined}})) \in \mathbb{R}^{T'' \times d} \tag{17}$$

3.6. Sign-Language Adapter and LLM Integration

The Sign-Language Adapter (SLA) maps the fused multimodal features into the LLM's input embedding space. It consists of a temporal compression stage followed by a two-layer MLP projection:

$$\mathbf{Z} = \text{AdaptivePool}_{N_s}(\mathbf{H}_{\text{final}}) \in \mathbb{R}^{N_s \times d} \tag{18}$$

$$\mathbf{E}_{\text{sign}} = \mathbf{W}_2' \cdot \text{GELU}(\mathbf{W}_1' \cdot \mathbf{Z} + b_1') + b_2' \in \mathbb{R}^{N_s \times d_{\text{LLM}}} \tag{19}$$

where $N_s = 32$ is the number of sign tokens and $d_{\text{LLM}} = 4096$ for LLaMA-2 7B. Layer normalization is applied after the final projection to stabilize the input distribution.

LLM Integration. We propose utilizing LLaMA-2 7B [7] as the language model backbone with LoRA [16] for parameter-efficient adaptation. LoRA introduces low-rank matrices $\mathbf{A} \in \mathbb{R}^{d \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d}$ into the attention layers, where $r \ll d$ is the rank (we use $r = 16$, $\alpha = 32$). The sign embeddings are prepended to a text prompt and fed to the LLM:

$$P(\mathbf{Y}|\mathbf{V}, \mathbf{S}) = \prod_{n=1}^{N} P(y_n | \mathbf{E}_{\text{sign}}, \mathbf{e}_{\text{prompt}}, y_{<n}; \Theta) \tag{20}$$

where $\mathbf{e}_{\text{prompt}}$ encodes the instruction "Translate the following sign language video into a spoken language sentence:" and Θ includes both the frozen LLM parameters and the trainable LoRA parameters.

3.7. Progressive Multi-Stage Training

We propose a three-stage progressive training strategy (Figure 3) to systematically bridge the gap between multimodal sign language features and the linguistic space of the LLM. Each stage addresses a specific modality gap.

Stage 1: Visual-Skeletal Contrastive Pre-training. The first stage trains the dual-stream encoders and CMFA to produce aligned visual-skeletal representations. We employ an InfoNCE contrastive loss [15]:

$$\mathcal{L}_{\text{contra}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{e^{\text{sim}(\mathbf{f}_v^i, \mathbf{f}_s^i)/\tau}}{\sum_j e^{\text{sim}(\mathbf{f}_v^i, \mathbf{f}_s^j)/\tau}} + \log \frac{e^{\text{sim}(\mathbf{f}_s^i, \mathbf{f}_v^i)/\tau}}{\sum_j e^{\text{sim}(\mathbf{f}_s^i, \mathbf{f}_v^j)/\tau}} \right] \tag{21}$$

 doi:10.20944/preprints202604.1065.v1

where $\text{sim}(\cdot,\cdot)$ is cosine similarity, $\tau = 0.07$ is the temperature parameter, B is the batch size, and $\mathbf{f}_v^i$, $\mathbf{f}_s^i$ are pooled visual and skeletal features. The LLM remains frozen in this stage. *Trainable:* CNN backbone, ST-GCN, CMFA.

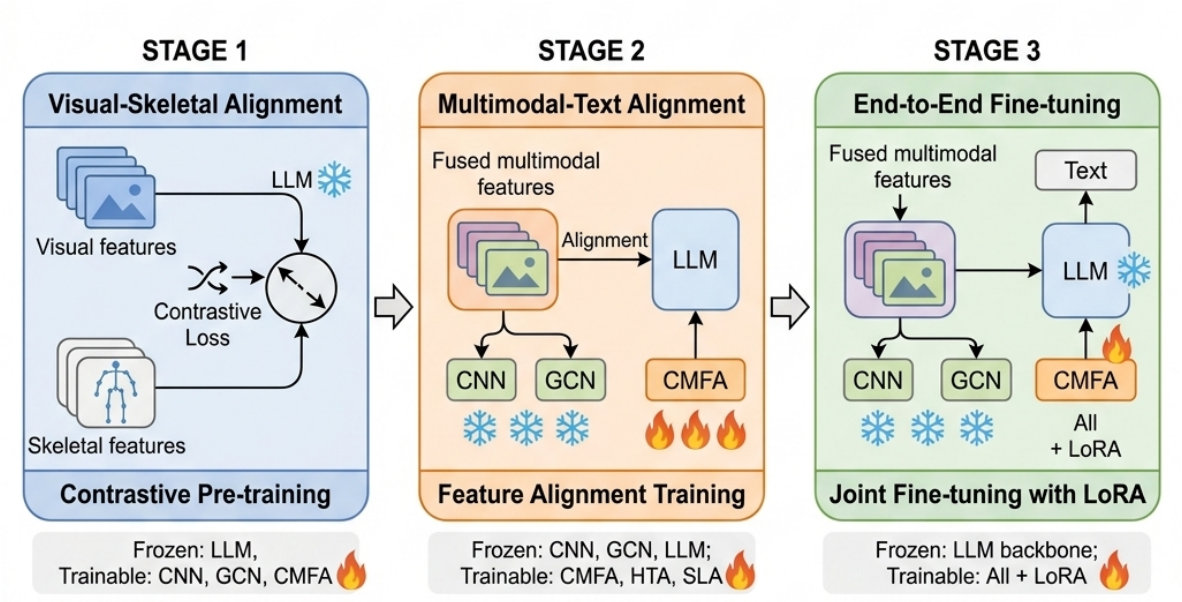

Figure 3. Proposed Multi-Stage Training Strategy. Stage 1 aligns visual-skeletal features via contrastive learning. Stage 2 maps fused features to the LLM embedding space. Stage 3 performs end-to-end LoRA fine-tuning. Snowflake/flame icons denote frozen/trainable components.

Stage 2: Multimodal-Text Alignment. With frozen feature encoders, this stage trains the CMFA, HTA, and SLA modules to map fused sign features into the LLM's embedding space:

$$\mathcal{L}_{\text{align}} = \|\mathbf{E}_{\text{sign}} - \text{sg}(\mathbf{E}_{\text{text}})\|_2^2 + \mathcal{L}_{\text{LM}} \tag{22}$$

where $\text{sg}(\cdot)$ denotes stop-gradient, $\mathbf{E}_{\text{text}}$ are the LLM's text embeddings for the target sentence, and $\mathcal{L}_{\text{LM}}$ is the standard autoregressive language modeling loss. *Trainable:* CMFA, HTA, SLA. *Frozen:* CNN, ST-GCN, LLM.

Stage 3: End-to-End Fine-tuning. Finally, we unfreeze all trainable components and apply LoRA to the LLM:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda_1 \mathcal{L}_{\text{contra}} + \lambda_2 \mathcal{L}_{\text{align}} \tag{23}$$

where $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$ are balancing coefficients determined by preliminary analysis. *Trainable:* All components + LoRA adapters. *Frozen:* LLM backbone weights.

4. Implementation Specifications

To facilitate reproducibility and community adoption, we provide complete implementation specifications.

4.1. Model Configuration

Table 1 summarizes the proposed model configuration. The total trainable parameter count (excluding the frozen LLM backbone) is approximately 65.9M parameters, making **SignFuse** comparable in overhead to existing methods like SpaMo and MMSLT.

 doi:10.20944/preprints202604.1065.v1

Table 1. Proposed model configuration.

Component	Specification
Visual Stream	
Backbone	ResNet-50 (ImageNet pre-trained)
Output dimension	$d_v = 2048$
Temporal conv	$k = 3$, stride $s = 2$, output $d = 512$
Skeletal Stream	
Pose estimator	MediaPipe Holistic
Number of joints	$V = 75$ (hands + body + face)
ST-GCN layers	4 layers, channels: $[2, 64, 128, 256, 512]$
Temporal kernel	$k = 9$, stride 2 at layer 3
CMFA Module	
Feature dimension	$d = 512$
Number of layers	$L = 3$
Attention heads	$h = 8$
FFN inner dim	$4d = 2048$
HTA Module	
Temporal kernels	$\{1, 5, 11\}$
Target length	$T'' = 64$
SLA + LLM	
Sign tokens	$N_s = 32$
LLM backbone	LLaMA-2 7B (frozen)
LLM dimension	$d_{\text{LLM}} = 4096$
LoRA rank / alpha	$r = 16$, $\alpha = 32$

4.2. Proposed Training Hyperparameters

Table 2 details the proposed training configuration for each stage. All stages use AdamW optimizer with weight decay 0.01 and cosine annealing with linear warmup.

Table 2. Proposed training hyperparameters per stage.

Parameter	Stage 1	Stage 2	Stage 3
Epochs	30	20	15
Learning rate	1×10^{-4}	5×10^{-5}	2×10^{-5}
Batch size	32	32	32
Warmup epochs	3	2	1
Gradient clipping	1.0	1.0	1.0
Hardware	4× NVIDIA A100 80GB		

4.3. Software Availability

We provide a complete PyTorch implementation consisting of:

- `model.py`: Full architecture implementation including all five proposed components (VisualStream, SkeletalStream, CMFA, HTA, SLA), loss functions (InfoNCE, Alignment, Combined), and a model builder utility. The implementation has been validated for correct tensor shapes through a comprehensive forward-pass test.
- `train.py`: Complete three-stage training pipeline with stage-specific parameter freezing, optimizer configuration, learning rate scheduling, checkpoint management, and synthetic data fallback for pipeline testing.
- `evaluate.py`: Evaluation utilities implementing BLEU-1/2/3/4, ROUGE-L, and METEOR metrics from scratch, along with visualization tools for generating comparison charts.

 doi:10.20944/preprints202604.1065.v1

5. Proposed Experimental Protocol

In this section, we formalize the experimental methodology that we invite the community to execute.

5.1. Target Datasets

We propose evaluating **SignFuse** on three widely-used SLT benchmarks that span different sign languages, vocabulary sizes, and difficulty levels:

PHOENIX-14T [2]: A German Sign Language (DGS) dataset containing 8,257 aligned video-text pairs from weather broadcasts. The dataset is split into 7,096 training, 519 development, and 642 test samples. Videos feature a single signer with a clean background, making it the most controlled benchmark.

CSL-Daily [40]: A large-scale Chinese Sign Language dataset with 20,654 video-sentence pairs covering daily life topics. The dataset features 10 signers and more diverse content than PHOENIX-14T, providing a stronger test of generalization.

How2Sign [41]: A large-scale American Sign Language (ASL) dataset with over 35,000 video clips covering instructional content. Due to its diverse topics, multiple signers, and varying backgrounds, it represents the most challenging benchmark.

5.2. Evaluation Metrics

Following standard evaluation protocols for SLT, we recommend reporting:

- **BLEU-n** ($n \in \{1, 2, 3, 4\}$) [42]: Measures n-gram precision between hypothesis and reference translations with a brevity penalty. BLEU-4 is the primary metric for comparison.
- **ROUGE-L** [43]: Measures the longest common subsequence between hypothesis and reference, capturing sentence-level fluency.
- **METEOR** [44]: Combines precision, recall, and alignment with synonym matching for more comprehensive evaluation.

5.3. Baseline Methods for Comparison

We recommend comparing against the following state-of-the-art methods, whose results are publicly available:

Table 3. Existing SOTA results on PHOENIX-14T (gloss-free, test set) to compare against. B@4 = BLEU-4, R-L = ROUGE-L.

Method	B@4	R-L
GFSLT-VLP [6] (ICCV 2023)	21.44	43.78
Sign2GPT [11] (2024)	22.52	44.83
SignLLM [10] (CVPR 2024)	23.51	46.35
SpaMo [12] (NAACL 2025)	24.32	47.56
MMSLT [13] (ICCV 2025)	25.18	48.34

5.4. Proposed Ablation Studies

To validate the contribution of each component, we recommend the following ablation experiments on the PHOENIX-14T development set:

1. **Component ablation:** Progressively add components (Visual-only → +Skeletal → +CMFA → +HTA → +Progressive Training) to measure incremental gains.
2. **Fusion strategy comparison:** Compare CMFA against concatenation, element-wise addition, unidirectional cross-attention (V→S and S→V), and bidirectional without gating.
3. **Temporal scale analysis:** Evaluate different combinations of HTA temporal scales to identify the most impactful granularities.

 doi:10.20944/preprints202604.1065.v1

4. **Training stage analysis:** Measure the benefit of each progressive training stage compared to direct end-to-end training.
5. **LLM backbone sensitivity:** Test with different LLM backbones (Flan-T5, Vicuna, LLaMA-2 7B/13B) to assess the framework's generalizability.

5.5. Expected Theoretical Advantages

Based on our architectural analysis, we hypothesize that **SignFuse** will demonstrate significant improvements over existing methods for the following reasons:

Complementary modality fusion. RGB features and skeletal features provide complementary information channels. When motion blur degrades RGB features (a well-documented limitation [30]), skeletal landmarks remain stable, and vice versain conditions of partial occlusion where pose estimation may fail, RGB features can still capture hand shapes. The CMFA module's bidirectional attention enables the model to dynamically leverage whichever modality is more reliable for each temporal position.

Multi-scale temporal understanding. The HTA module's hierarchical design aligns with the linguistic structure of sign language. Frame-level features should improve fingerspelling accuracy, segment-level features should capture individual sign boundaries, and sequence-level features should improve the translation of compound sentences and temporal modifiersa common failure mode in existing systems [12].

Progressive training stability. By decomposing the training into three stages, each addressing a specific modality gap, we expect improved optimization stability and final performance compared to direct end-to-end training, which must simultaneously learn cross-modal alignment and language generation.

6. Discussion

6.1. Theoretical Analysis of CMFA

The CMFA module can be understood as computing a soft correspondence between visual and skeletal features at each temporal position. The attention weights $\text{softmax}(\mathbf{Q}_v\mathbf{K}_s^\top/\sqrt{d_k})$ represent the degree to which each visual feature should attend to each skeletal feature, effectively creating a learned visual-skeletal alignment. This is particularly important for sign language because the same hand shape may appear in different spatial positions (captured by skeleton) with different orientations (captured by appearance), and vice versa.

The gating mechanism α provides an additional degree of flexibility, allowing the model to emphasize visual features when appearance information is more discriminative (e.g., for distinguishing similar hand shapes) and skeletal features when structural information is more informative (e.g., for encoding spatial relationships between both hands). We hypothesize that the learned gate values will exhibit interpretable patterns correlating with sign linguistic categories.

6.2. Computational Cost Analysis

Table 4 provides a theoretical comparison of computational costs. While **SignFuse** introduces additional parameters from the skeletal stream and fusion modules, the overhead is modest relative to the LLM backbone.

Table 4. Estimated computational cost comparison. Extra parameters are relative to the frozen LLM backbone.

Method	Extra Params (M)	Modalities
Sign2GPT [11]	∼45	RGB
SpaMo [12]	∼52	RGB (spatial + motion)
MMSLT [13]	∼69	RGB + MLLM descriptions
SignFuse (Ours)	∼66	RGB + Skeleton (GCN)

 doi:10.20944/preprints202604.1065.v1

6.3. Limitations and Future Directions

We acknowledge the following limitations of the current work:

1. **Pending empirical validation.** The most significant limitation is the absence of empirical results. While the architecture is fully specified and the codebase is complete, the claims remain theoretical until validated on benchmark datasets with real training.
2. **Pose estimation reliability.** Our reliance on MediaPipe for skeleton extraction means that the skeletal stream's effectiveness is bounded by MediaPipe's accuracy, which can degrade under poor lighting, occlusion, or non-frontal camera angles.
3. **Fixed temporal scales.** The HTA module uses fixed kernel sizes $\{1, 5, 11\}$, which may not be optimal for all signing speeds and styles. Adaptive kernel selection could improve robustness.
4. **Single sign language.** While we propose evaluation on multiple datasets, each represents a different sign language. Cross-lingual transfer between sign languages remains an open challenge.

Future work should explore: (i) adaptive pose estimation with learned error correction; (ii) dynamic temporal scaling based on signing speed; (iii) explicit incorporation of non-manual features (mouthing, eye gaze) as a third modality; and (iv) multi-lingual sign language translation leveraging shared skeletal representations across sign languages.

7. Open Call to the Research Community

The computational hardware required to validate high-parameter multimodal LLMsoften requiring continuous access to A100/H100 GPU clusters over several daysremains a severe bottleneck for many research groups. This resource asymmetry creates a gap between architectural innovation and empirical validation that we aim to bridge through open collaboration.

This paper serves as both a theoretical hypothesis and an open-source collaboration request. We have constructed the complete mathematical framework, detailed architectural specifications, and a fully functional PyTorch codebase for **SignFuse**. We cordially invite research laboratories and academic partners who possess the requisite computational resources to:

1. **Execute the proposed training pipeline** on PHOENIX-14T, CSL-Daily, and/or How2Sign using the provided codebase and hyperparameters.
2. **Report empirical results** following the evaluation protocol defined in Section 5, including all recommended ablation studies.
3. **Extend or modify the architecture** based on empirical findings, *e.g.*, exploring alternative backbone networks, different fusion strategies, or additional modalities.
4. **Collaborate on joint publications** reporting the empirical validation results.

We request that researchers who validate, falsify, or modify the **SignFuse** hypothesis cite this foundational architectural blueprint. We believe that open, collaborative research practiceswhere theoretical contributions and empirical validations are distributed across groups with complementary resourceswill accelerate progress in sign language translation and accessibility technology.

8. Conclusion

We have presented the complete theoretical blueprint for **SignFuse**, a proposed dual-stream cross-modal fusion framework for gloss-free sign language translation with Large Language Models. By formalizing the mathematical integration of a ResNet-based visual stream with an ST-GCN-based skeletal stream via our novel Cross-Modal Fusion Attention (CMFA) module, and proposing a Hierarchical Temporal Aggregation (HTA) mechanism for multi-scale temporal modeling, **SignFuse** offers a principled solution to the limitations found in existing single-modality SLT systems.

Our progressive three-stage training strategy provides a systematic approach to bridging the gap between multimodal sign language features and the linguistic space of LLMs, while the complete PyTorch implementation ensures that the proposed architecture can be immediately deployed for empirical validation. While the empirical evaluation is left to future community collaboration, the

 doi:10.20944/preprints202604.1065.v1

structural foundations, mathematical rigor, and implementation detail established herein offer a compelling and reproducible paradigm for advancing sign language translation technology toward more accurate, inclusive, and practical systems for the Deaf and hard-of-hearing community.

Data Availability Statement: Code is available at: GitHub Repository.

References

1. World Health Organization. Deafness and Hearing Loss. https://www.who.int/health-topics/hearing-loss, 2023. Accessed: 2025-01-15.
2. Camgöz, N.C.; Hadfield, S.; Koller, O.; Ney, H.; Bowden, R. Neural Sign Language Translation. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7784–7793.
3. Singh, G.; Banerjee, T.; Ghosh, N. Tracing the Evolution of Artificial Intelligence: A Review of Tools, Frameworks, and Technologies (1950–2025). *Preprints* **2025**. https://doi.org/10.20944/preprints202511.0637.v1.
4. Camgöz, N.C.; Koller, O.; Hadfield, S.; Bowden, R. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10023–10033.
5. Camgöz, N.C.; Koller, O.; Hadfield, S.; Bowden, R. Multi-channel Transformers for Multi-articulatory Sign Language Translation. *arXiv preprint arXiv:2009.00299* **2020**.
6. Zhou, B.; Chen, Z.; Clápez, A.; Shao, J.; Wang, J.; Xiao, J.; Zha, Z.J. Gloss-Free Sign Language Translation: Improving From Visual-Language Pretraining. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2816–2825.
7. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* **2023**.
8. OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* **2023**.
9. Singh, G.; Qamar, L.; Volta, N.V.; Velamuri, A.; Khanyile, A. Vision–Language Foundation Models and Multimodal Large Language Models: A Comprehensive Survey of Architectures, Benchmarks, and Open Challenges. *Preprints* **2026**. https://doi.org/10.20944/preprints202602.0467.v2.
10. Gong, J.; Foo, L.G.; He, Y.; Rahmani, H.; Liu, J. LLMs are Good Sign Language Translators. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 14754–14764.
11. Wong, R.; Camgöz, N.C.; Bowden, R. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. *arXiv preprint arXiv:2405.04164* **2024**.
12. Hwang, E.J.; Cho, S.; Lee, J.; Park, J.C. An Efficient Gloss-Free Sign Language Translation Using Spatial Configurations and Motion Dynamics with LLMs. In Proceedings of the Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2025.
13. Kim, J.; Jeon, H.; Bae, J.; Kim, H.Y. Leveraging the Power of MLLMs for Gloss-Free Sign Language Translation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 21048–21058.
14. Liu, Y.; Zhang, W.; Ren, S.; Huang, C.; Yu, J.; Xu, L. SCOPE: Sign Language Contextual Processing with Embedding from LLMs. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2025, Vol. 39.
15. Jiang, Z.; Sant, G.; Moryossef, A.; Müller, M.; Sennrich, R.; Ebling, S. SignCLIP: Connecting Text and Sign Language by Contrastive Learning. In Proceedings of the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.
16. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2022.
17. Koller, O.; Forster, J.; Ney, H. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. In Proceedings of the Computer Vision and Image Understanding, 2015, Vol. 141, pp. 108–125.

 doi:10.20944/preprints202604.1065.v1

18. Cui, R.; Liu, H.; Zhang, C. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7361–7369.
19. Huang, J.; Zhou, W.; Zhang, Q.; Li, H.; Li, W. Video-based Sign Language Recognition without Temporal Segmentation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
20. Li, D.; Rodriguez, C.; Yu, X.; Li, H. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In Proceedings of the Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1459–1469.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017.
22. Chen, Y.; Wei, F.; Sun, X.; Wu, Z.; Lin, S. Two-stream Network for Sign Language Recognition and Translation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2022.
23. Park, M.; Kim, S.; Lee, J. ConSignformer: A Transformer-based Architecture for Continuous Sign Language Recognition. *arXiv preprint arXiv:2405.12018* **2024**.
24. Chen, Y.; Li, Z.; Wang, M. Skeleton-Based Sign Language Recognition via Graph Convolutional Networks. *arXiv preprint arXiv:2407.11960* **2024**.
25. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* **2018**, *32*.
26. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A Framework for Building Perception Pipelines. *arXiv preprint arXiv:1906.08172* **2019**.
27. Kumar, A.; Singh, R.; Patel, N. Continuous Sign Language Recognition System using Deep Learning with MediaPipe Holistic. *arXiv preprint arXiv:2411.11290* **2024**.
28. Ahmed, Z.; Khan, M.; Ali, T. Enhancing Sign Language Detection through MediaPipe and Convolutional Neural Networks. *arXiv preprint arXiv:2406.03729* **2024**.
29. Ferreira, P.M.; Cardoso, J.S.; Rebelo, A. Sign Language Recognition Based on Deep Learning and Low-cost Handcrafted Descriptors. *arXiv preprint arXiv:2408.07244* **2024**.
30. Alyami, S.; Luqman, H. A Comparative Study of Continuous Sign Language Recognition Techniques. *arXiv preprint arXiv:2406.12369* **2024**.
31. Rastgoo, R.; Kiani, K.; Escalera, S. Recent Advances on Deep Learning for Sign Language Recognition. *Computer Modeling in Engineering & Sciences* **2024**.
32. Yin, A.; Zhong, T.; Tang, L.; Jin, W.; Jin, T.; Zhao, Z. Gloss Attention for Gloss-free Sign Language Translation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2551–2562.
33. Inan, M.; Sicilia, A.; Alikhani, M. SignAlignLM: Integrating Multimodal Sign Language Processing into Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics (ACL), 2025.
34. Zheng, W.; Li, H.; Chen, W.; Wang, Y. CLIP-SLA: Parameter-Efficient CLIP Adaptation for Continuous Sign Language Recognition. *arXiv preprint arXiv:2407.12174* **2024**.
35. Singh, G. A Review of Multimodal Vision–Language Models: Foundations, Applications, and Future Directions. *Preprints* **2025**. https://doi.org/10.20944/preprints202510.2511.v1.
36. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2019.
37. Li, J.; Selvaraju, R.R.; Gotmare, A.D.; Joty, S.; Xiong, C.; Hoi, S. Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021.
38. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a Visual Language Model for Few-Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2022.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
40. Zhou, H.; Zhou, W.; Qi, W.; Pu, J.; Li, H. Improving Sign Language Translation with Monolingual Data by Sign Back-Translation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1316–1325.

 doi:10.20944/preprints202604.1065.v1

41. Duarte, A.; Palaskar, S.; Venberèg, L.; Ghadiyaram, D.; DeHaan, K.; Metze, F.; Jordi, T.; Giro-i Nieto, X. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2735–2744.
42. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318.
43. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, 2004.
44. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005.