Dynamic Feature Fusion for Sign Language Translation Using HyperNetworks

Ruiquan Zhang[1,2,3,4,*], **Rui Zhao**[1,2,3,4,*], **Zhicong Wu**[1,2,3,4], **Liang Zhang**[1,2,3,4],
Haoqi Zhang[1,2,3,4], **Yidong Chen**[1,2,3,4,†]

[1]Department of Artificial Intelligence, School of Informatics, Xiamen University, China
[2]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China
[3]National Language Resources Monitoring and Research Center for Education and Teaching Media, Xiamen University, China
[4]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China

{rqzhang, zhsqzr}@stu.xmu.edu.cn, ydchen@xmu.edu.cn

Abstract

Sign language is a visual language that conveys information through gestures and facial expressions. Drawing inspiration from how the human brain simultaneously processes color, shape, and motion, this work presents an efficient dual-stream early fusion approach, which combines features from both RGB and keypoint streams at an early stage for improved sign language translation performance. A key challenge addressed is extracting complementary features from both streams while ensuring global semantic consistency to enhance generalization. To address this challenge, a hypernetwork-based fusion strategy is introduced to extract salient features from both modalities, along with a partial shortcut connection training method that reinforces the complementary relationship between the streams. Additionally, self-distillation and shared semantic space (SST) contrastive learning are employed to preserve feature advantages while aligning features in a shared semantic space for better consistency. Experimental results demonstrate that the proposed approach achieves state-of-the-art performance on two public sign language datasets, reducing model parameters by approximately two-thirds while improving translation accuracy. Codes and models are available at HyperSign.

1 Introduction

Sign language is the primary mode of communication within the deaf community. Unlike spoken languages, it has a distinct grammar and vocabulary, conveying meaning through static shapes and dynamic movements of the hands, face, and body. Sign Language Translation (SLT) plays a crucial role in bridging the communication gap between signers and non-signers by converting sign language into spoken or written text (Camgoz et al., 2018, 2020; Chen et al., 2022a; Zhang et al., 2025).

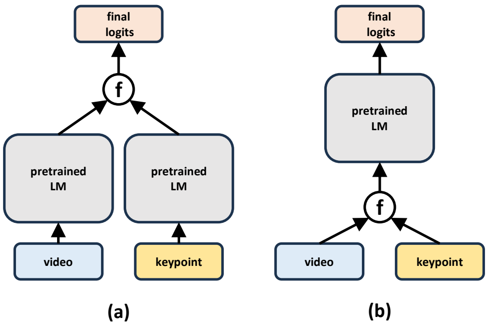

Figure 1: The overview of (a) SLT models with late fusion and (b) SLT models with early fusion (ours).

To achieve SLT, researchers typically use RGB streams, which capture rich visual details but are highly susceptible to noise caused by varying backgrounds, lighting conditions, and occlusions (Camgoz et al., 2020; Chen et al., 2022a; Guo et al., 2020). Alternatively, keypoint streams provide a more abstract representation by focusing on geometric aspects of gestures, but they may miss finer nuances of complex movements (Xiao et al., 2021; Cui et al., 2019). Therefore, combining both RGB streams for static features and keypoint streams for dynamic motion could enhance SLT accuracy by efficiently integrating both modalities (Xiao et al., 2021; Cui et al., 2019).

Current dual-stream SLT models mostly rely on late fusion or no fusion, as illustrated in Figure 1 (a). For example, SignBERT+ (Hu et al., 2023) applies a simple cross-attention mechanism for late fusion, while TS-Network (Chen et al., 2022b) averages the final outputs of separately trained translation networks. However, this late fusion approach leads to parameter redundancy and fails to synchronize static and dynamic visual cues effec-

[*]These authors contributed equally to this work.
[†]Corresponding author

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6242–6254
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

tively (Shankar et al., 2022; Ak et al., 2023). From a cognitive psychology perspective(Grossberg and Rudd, 1989; Wagemans et al., 2012), the brain integrates static (e.g., shape and color) and dynamic (e.g., motion) information early on to form a unified perception. Inspired by this, we propose early fusion to better capture the interdependencies between color, shape, and motion, as shown in Figure 1 (b). Early fusion synchronizes static and dynamic information, avoids parameter redundancy, and improves overall model performance.

However, training an early fusion dual-stream model presents two key challenges: (1) effectively extracting complementary information from RGB and keypoint streams, and (2) ensuring semantic consistency between the two to prevent misalignment, which could negatively impact the model's generalization ability.

To address these challenges, we propose the **HyperSign** SLT model, which uses hypernetworks (Ha et al., 2016) to dynamically integrate RGB and keypoint streams based on each SL sample, optimizing the fusion process. Unlike traditional static networks, hypernetworks dynamically adjust network parameters, enhancing adaptability to different SL actions. Additionally, we implement a partial shortcut connection strategy to progressively train the hypernetwork, further improving the feature extraction capabilities of both streams.

To ensure that the dual-streams are semantically aligned, we introduce self-distillation at both the encoder and decoder stages, maintaining consistency in their shared semantic space. For instance, in the sign for "drinking water," the RGB stream captures static hand shapes resembling a cup, while the keypoint stream focuses on the dynamic drinking motion. Ensuring alignment prevents misinterpretation, such as confusing the action with "brushing teeth." Additionally, we propose a contrastive learning approach using shared semantic tokens, reducing discrepancies between the streams and significantly enhancing translation accuracy.

In summary, our contributions include: (1) introducing a hypernetwork model with a partial shortcut connection strategy for dynamically integrating RGB and keypoint streams; (2) proposing a semantic alignment mechanism to harmonize static and dynamic visual features; and (3) conducting extensive experiments on the PHOENIX14T and CSL-Daily datasets, demonstrating that our model achieves state-of-the-art performance in SLT while significantly improving inference speed.

2 Methodology

In this section, to better illustrate our method, we first simply introduce our base model. Then, we describe the hypernetwork mechanism used for dynamic feature fusion and the partial shortcut connection strategy. Finally, we explain the semantic synergy mechanisms, including self-distillation and SST contrastive learning.

2.1 The Base Model

Given a sign language video $\mathcal{V} = (v_1, \ldots, v_T)$ containing T frames, our goal is to optimize the SLT model to predict a spoken sentence $\mathcal{S} = (s_1, \ldots, s_L)$ containing L words.

As shown in Figure 1(B), in this study, we adopt a dual-stream fusion SLT model based on early fusion. The current paper focuses primarily on optimizing the model's translation network; therefore, in terms of feature extraction, we follow the strategy of (Chen et al., 2022b) to extract and model the keypoint information, and train individual feature extractors and VL-mapper for each stream. The extracted visual features $\mathbf{F}_\mathrm{v}$ and keypoint features $\mathbf{F}_\mathrm{k}$ both have dimensions of T/4×1024.

Subsequently, we integrate $\mathbf{F}_\mathrm{v}$ and $\mathbf{F}_\mathrm{k}$ into a unified feature representation $\mathbf{F}_\mathrm{f}$ through a specific fusion strategy. The fused feature $\mathbf{F}_\mathrm{f}$ is fed into the multilingual mBART model, where the decoder generates tokens for the spoken sentence $\mathcal{S}$ one by one using a causal masking mechanism.

2.2 Hypernetworks for Dynamic Feature Fusion

Hypernetworks (Ha et al., 2016) can dynamically parameterize the weights of the fusion network based on the specific characteristics of the dual-stream sample itself, thereby dynamically combining effective information from the RGB and keypoint streams to produce high-quality fusion stream. The hypernetwork includes a generator network for dynamically generating weights and a dynamic MLP for deep semantic fusion. Additionally, we introduce an efficient partial shortcut connection strategy to optimize the training process and stability of the hypernetwork.

2.2.1 Generator Network

The generator network is a core module of the hypernetwork, receiving RGB features $\mathbf{F}_\mathrm{v}$ and keypoint features $\mathbf{F}_\mathrm{k}$, and dynamically generating the

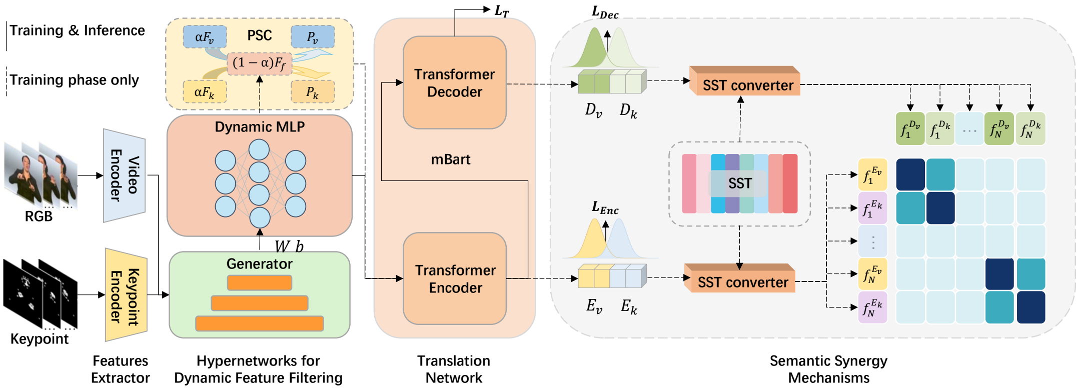

Figure 2: The framework of HyperSign: The SLT model is responsible for converting sign language videos into corresponding spoken text. Initially, the model encodes RGB and keypoint videos separately, obtaining visual features $\mathbf{F}_\text{v}$ and keypoint features $\mathbf{F}_\text{k}$. These features are jointly input into the hypernetwork, where they undergo dynamic fusion to produce the fused visual feature $\mathbf{F}_\text{f}$. In the inference phase, this stream is directly fed into the translation network for spoken text prediction. During the training phase, $\mathbf{F}_\text{f}$ generates $\mathbf{P}_\text{v}$ and $\mathbf{P}_\text{k}$ through partial shortcut connections. The dual-stream features are processed through a unified translation network, generating both visual ($\mathbf{E}_\text{v}$, $\mathbf{E}_\text{k}$) and textual ($\mathbf{D}_\text{v}$, $\mathbf{D}_\text{k}$) representations. During the training phase, two-stage self-distillation enforces KL divergence constraints on these dual-stream hidden states. Subsequently, the shared semantic space (SST) converter transforms these visual and textual outputs into the SST space, facilitating simultaneous intra-stream and inter-stream contrastive learning. These specific distillation and transformation processes are omitted during the inference phase to streamline model execution.

weight matrix W and bias matrix b based on the characteristics of the input samples. Unlike traditional static weights, the generator network uses subnetworks G_W and G_b to produce adaptive weights and biases, ensuring the network can tailor its processing to the specific features of each sample. The operations of the generator network are defined as:

$$W = G_W(\mathbf{F}_\text{v}, \mathbf{F}_\text{k}), \tag{1}$$

$$b = G_b(\mathbf{F}_\text{v}, \mathbf{F}_\text{k}). \tag{2}$$

2.2.2 Dynamic MLP

The dynamic MLP receives the weights W and biases b generated by the generator network, along with the input RGB features $\mathbf{F}_\text{v}$ and keypoint features $\mathbf{F}_\text{k}$, outputting the fusion features $\mathbf{F}_\text{f}$. The computation process is as:

$$\mathbf{F}_\text{f} = \sigma(W \odot \text{LayerNorm}(\mathbf{F}_\text{v} \odot \mathbf{F}_\text{k}) + b), \tag{3}$$

where σ represents the activation function, and $\odot$ denotes the element wise product.

2.2.3 Partial Shortcut Connection

In the initial stages of training the hypernetwork, directly using the fusion features may introduce noise, affecting the learning outcomes of the model. To mitigate these issues and enhance model performance, we propose a method of partial shortcut connection. This method adjusts the fusion coefficients gradually, allowing the model to focus on a single feature stream initially, reducing information conflicts, and progressively increasing the fusion level of different feature streams later. The fusion process is described as:

$$\mathbf{P}_i = \alpha \mathbf{F}_i + (1-\alpha)\mathbf{F}_\text{f}, \tag{4}$$

where i takes v for RGB features or k for keypoint features. The fusion coefficient α_t follows a truncated normal distribution with a mean of 0 and a standard deviation of σ_t, constrained within the interval $[0,1]$, i.e., $\alpha_t \sim \mathcal{N}_{[0,1]}(0, \sigma_t^2)$. As the number of training steps t increases, σ_t is defined as $\sigma_t = \max(1 - \frac{t}{T}, \epsilon)$, where T is the total number of steps, and ϵ is a very small positive number.

2.3 Semantic Synergy Mechanisms

Although hypernetworks can integrate RGB and keypoint features effectively, the significant representational differences between these modalities can lead to semantic inconsistencies. To ensure accurate capturing and alignment of semantic information from both feature types during the fusion process and to better perform dynamic feature fusion, we introduce semantic synergy mechanisms, including self-distillation and SST contrastive learning.

Methods	Dev					Test				
	R	**B1**	**B2**	**B3**	**B4**	**R**	**B1**	**B2**	**B3**	**B4**
					PHOENIX14T					
MMTLB-KP (2022)	50.21	51.13	38.43	30.59	25.28	50.79	52.01	39.50	31.52	26.14
MMTLB (2022)	53.10	53.95	41.12	33.14	27.61	52.65	53.97	41.75	33.84	28.39
CV-SLT (2024)	54.43	55.09	42.60	34.63	29.10	54.33	54.88	42.68	34.79	29.27
CorrNet+ (2024)	54.54	54.56	42.31	34.48	29.13	53.76	**55.32**	42.74	**34.86**	**29.42**
SignBERT+ (2023)	51.12	51.46	38.28	30.30	24.95	50.63	52.01	39.19	31.06	25.70
TS-SLT (2022)	54.08	54.32	41.99	34.15	28.66	53.48	54.90	42.43	34.46	28.95
HyperSign (ours)	**54.98**	**55.35**	**43.06**	**35.21**	**29.78**	**54.51**	55.20	**42.80**	34.84	**29.42**
					CSL-Daily					
MMTLB-KP (2022)	47.98	47.95	35.49	26.72	20.59	48.63	48.58	35.96	27.03	20.84
MMTLB (2022)	53.38	53.81	40.84	31.29	24.42	53.25	53.31	40.41	30.87	23.92
CV-SLT (2024)	56.36	58.05	44.73	35.14	28.24	57.06	58.29	45.15	35.77	28.94
CorrNet+ (2024)	55.52	55.64	42.78	33.13	26.14	55.84	55.82	42.96	33.26	26.14
TS-SLT (2022)	55.10	55.21	42.31	32.71	25.76	55.72	55.44	42.49	32.87	25.79
HyperSign (ours)	**57.32**	**58.68**	**45.54**	**36.08**	**29.28**	**57.89**	**58.96**	**45.93**	**36.53**	**29.55**

Table 1: Model performance on PHOENIX14T and CSL-Daily datasets.

2.3.1 Self-Distillation

During model training iterations, although both $\mathbf{P}_{\text{v}}$ and $\mathbf{P}_{\text{k}}$ possess dual-stream semantic characteristics, they emphasize different aspects. $\mathbf{P}_{\text{v}}$ highlights effective RGB-related features such as color, texture, and partial gestures, whereas $\mathbf{P}_{\text{k}}$ emphasizes keypoint-related features like motion trajectories and speeds. Using self-distillation at both encoder and decoder ends by minimizing the KL divergence between different feature streams, our approach maintains semantic consistency across modalities, enabling the model to discard irrelevant or noisy information and enhancing effective features that are continuously passed to the fused stream.

At the encoder and decoder ends of the mBART, the encoder encodes $\mathbf{P}_{\text{v}}$ into $\mathbf{E}_{\text{v}}$ and $\mathbf{P}_{\text{k}}$ into $\mathbf{E}_{\text{k}}$, respectively. The decoder then decodes these into $\mathbf{D}_{\text{v}}$ and $\mathbf{D}_{\text{k}}$. The Kullback-Leibler divergence is used for self-distillation at both the decoder and encoder ends:

$$\mathcal{L}_{\text{Enc}} = KL_{\text{enc}}(\mathbf{E}_{\text{v}} \parallel \mathbf{E}_{\text{k}}), \tag{5}$$

$$\mathcal{L}_{\text{Dec}} = KL_{\text{dec}}(\mathbf{D}_{\text{v}} \parallel \mathbf{D}_{\text{k}}). \tag{6}$$

2.3.2 SST Contrastive Learning

Despite the dual-stage KL divergence self-distillation ensuring semantic consistency at both encoder and decoder ends and enhancing the transfer of effective information between modalities, a direct alignment mechanism across modalities is missing. Such a mechanism is crucial for maintaining semantic consistency among RGB stream, keypoint stream, and textual information. We draw on contrastive learning to further strengthen the alignment between visual and textual semantics. However, directly aligning cross-modal information is challenging due to differences in semantic levels and granularity between visual and textual information. Therefore, we introduce Shared Semantic Tokens (SST), mapping different modal features into a shared semantic space and synchronizing them across time and space. This method ensures more accurate semantic interaction between RGB, keypoints, and text, significantly enhancing the overall performance of the model.

Let $\{c_i \mid i = 1, \ldots, C\}$ denote the shared semantic token space, where C is the number of tokens, each of size d. Let $\mathbf{F}_x$ represent the feature transformed by an MLP from any one of $\mathbf{E}_{\text{v}}$, $\mathbf{D}_{\text{v}}$, $\mathbf{E}_{\text{k}}$, or $\mathbf{D}_{\text{k}}$, of size $\mathbb{R}^{N\times d}$. The mapping process to the shared semantic token space can be described as follows: First, we measure the relevance between feature vectors and SST by calculating the maximum dot product between them, i.e.,

$r_i = \max_j \langle \mathbf{F}_{x,j}, c_i \rangle$, where $\langle \cdot, \cdot \rangle$ represents the dot product.

To reduce noise and improve model interpretability, following the method in (Chen et al., 2023), we employ the Sparsemax function (Martins and Astudillo, 2016) to enhance the sparsity of relevance. Specifically, Sparsemax calculates a threshold that sets weights below this threshold to zero, thus obtaining a sparsified weight vector $\mathbf{w}$. This process can be represented as $\text{Sparsemax}(\mathbf{r}) = \arg\min_{\mathbf{w}\in\Delta^{K-1}} \|\mathbf{w}-\mathbf{r}\|^2$, where Δ^{K-1} denotes the $(K-1)$-dimensional probability simplex. Finally, the mapped features $f = \sum_{i=1}^{C} \mathbf{w}_i \cdot c_i$ are obtained by summing the weighted SST. The algorithm for mapping features to the SST space is referred to as the SST Converter, detailed in Algorithm 1.

For any two features $\mathbf{E}_p$ and $\mathbf{D}_q$, the contrastive loss function is defined as:

$$L_{i,j} = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp\left(\text{sim}\left(f_i^{E_p}, f_j^{D_q}\right)/\tau\right)}{\sum_{m=1}^{N}\exp\left(\text{sim}\left(f_i^{E_p}, f_m^{D_q}\right)/\tau\right)}, \tag{7}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity function, and τ is the temperature parameter.

Considering all possible feature pair combinations, the total loss function $\mathcal{L}_{\text{SST}}$ is:

$$\mathcal{L}_{\text{SST}} = \frac{1}{4}(L_{E_v,D_v} + L_{E_k,D_k} + L_{E_v,D_k} + L_{E_k,D_v}), \tag{8}$$

These loss functions respectively represent the contrastive losses between the visual information from RGB features and keypoint features with textual information, as well as the cross-modal contrastive losses between visual and textual information.

2.3.3 Loss Function

The overall loss function comprises translation losses for two different paths ($\mathcal{L}_{T_v}$, $\mathcal{L}_{T_k}$), KL losses at the mBART encoder and decoder ends ($\mathcal{L}_{\text{Enc}}$, $\mathcal{L}_{\text{Dec}}$), and a discrete contrastive learning loss $\mathcal{L}_{\text{SST}}$:

$$\mathcal{L} = \mathcal{L}_{T_v} + \mathcal{L}_{T_k} + \lambda_1\mathcal{L}_{\text{Enc}} + \lambda_2\mathcal{L}_{\text{Dec}} + \mathcal{L}_{\text{SST}}. \tag{9}$$

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate our approach on SLT using the PHOENIX14T (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021) datasets. All ablation studies are conducted on the PHOENIX14T SLT task. PHOENIX14T is a German Sign Language dataset with a vocabulary of 1066 for annotated Glosses and 2887 for German text. It consists of 7096, 519, and 642 samples in the training, development, and test sets, respectively; CSL-Daily is a newly released large-scale Chinese Sign Language dataset, with a vocabulary of 2000 for annotations and 2343 for Chinese text, comprising 18401, 1077, and 1176 samples in the training, development and test sets, respectively. Following previous work (Chen et al., 2022a; Zhao et al., 2024), we use BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to assess the models of SLT. Higher BLEU and ROUGE scores indicate better translation performance.

Algorithm 1 SST converter

Input: Feature set $\mathbf{E}_{\text{v}}$ (similar operations for $\mathbf{D}_{\text{v}}$, $\mathbf{E}_{\text{k}}$, $\mathbf{D}_{\text{k}}$), semantic tokens $\{c_i\}_{i=1}^{C}$

Output: Mapped features $f^{E_{\text{v}}}$

```
$\mathbf{F}_x \leftarrow \text{MLP}(\mathbf{E}_{\text{v}})$
for $i = 1$ to $C$ do
    $r_i \leftarrow \max_j \langle \mathbf{F}_{x,j}, c_i \rangle$
end for
$\mathbf{w} \leftarrow \text{Sparsemax}(\mathbf{r})$        ▷ Invoke Sparsemax
procedure Sparsemax($\mathbf{r}$)
    Sort $\mathbf{r}$ to get $\mathbf{r}_{\text{sorted}}$
    Compute cumulative sums $s_k = \sum_{j=1}^{k} r_{\text{sorted},j}$
    Find $k^* = \max\left\{k \mid \frac{s_k - 1}{k} > r_{\text{sorted},k}\right\}$
    $\tau = \frac{s_{k^*} - 1}{k^*}$
    return $\max(\mathbf{r} - \tau, 0)$
end procedure
$f^{E_{\text{v}}} \leftarrow \sum_{i=1}^{C} \mathbf{w}_i c_i$        ▷ Weighted sum of semantic tokens
```

3.2 Implementation Details

In line with TS-SLT (Chen et al., 2022b), we employ the same configurations for independently pre-training the RGB encoder, keypoint encoder, and the pretrained mBART. However, we do not perform joint training on RGB and keypoint encoders nor integrate multiple mBARTs for logits, but instead, use a single shared mBART for text decoding. The model is trained for only 40 epochs. Additionally, the hypernetwork generator consists of a two-layer feed-forward neural network and employs dropout=0.5 for regularization. The number of tokens and dimensions in SST are both set to 128, with a learning rate of 1e-5 and the temperature hyperparameter τ set at 0.07. For PHOENIX14T, we set the partial shortcut connection fusion steps T to 40; for the CSL-Daily dataset, where the modal discrepancy between RGB and keypoint streams is smaller, T is set to 15. Based on preliminary experiments, the weight for KL divergence at the encoder end is 1, and at the decoder end, it is 5. We employ a KL annealing technique (Bowman et al., 2015) to prevent the disappearance of KL divergence in the first 4K training steps. During inference, we follow previous studies (Chen et al., 2022a,b), using a beam search with a beam size of 5 and a length penalty of 1. The batch size is set to 16, and due to computational constraints, AMP (Baboulin et al., 2009) is applied. HyperSign is implemented based on the open-source SLRT project[1]. All experiments are conducted on a single 3090 GPU.

3.3 Comparison with State-of-the-Art Methods

As shown in Table 1, MMTLB is currently the baseline model in the field of SLT. Building on this, TS-SLT incorporates a translation network integrating both RGB and Keypoint streams. Following the reference (Chen et al., 2022b), we replaced the input stream of MMTLB with a keypoint stream and showcased the results of MMTLB-KP. Other results are drawn from the original paper.

Our proposed HyperSign method outperforms all previous state-of-the-art SLT methods. Compared to the CV-SLT model, it achieves performance improvements of +0.68/+0.15 and +1.04/+0.61 on the development/test sets of PHOENIX14T and CSL-Daily, respectively. Notably, compared to MMTLB using only the RGB stream and MMTLB-KP using only the Keypoint stream, our method shows significant improvements on the PHOENIX14T development set, with increases in BLEU4 scores of +2.17 and +4.50, demonstrating the complementary and enhancing effects of dual stream information. Although CorrNet+ (Hu et al., 2024) slightly outperforms us in B1 and B3 on the test set of the PHOENIX14T dataset, our method demonstrates significant superiority in most metrics compared to theirs, particularly on the CSL-Daily dataset.

As shown in Figure 3, although the TS-SLT model employs a more complex dual-stream joint pre-training and multiple translation network integration strategy, our HyperSign model achieves a 1.12 BLEU4 score increase on the PHOENIX14T development set with only one-third the parameters of TS-SLT. This significant performance improvement is attributed to HyperSign's powerful capabilities in dynamic feature fusion.

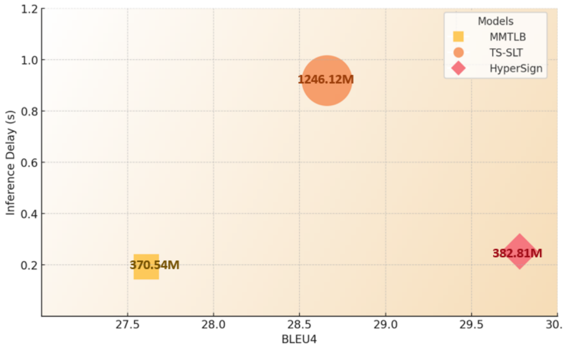

Figure 3: This scatter plot illustrates the performance of three translation models (MMTLB, TS-SLT, and HyperSign) in terms of BLEU4 score (horizontal axis) and inference latency (vertical axis), with the size of each marker representing the number of parameters in each model. Different markers are used to represent each model: squares for MMTLB, circles for TS-SLT, and diamonds for HyperSign. The color gradient, which transitions from the top-left to the bottom-right, indicates the trade-off between accuracy and inference speed, with models closer to the bottom-right exhibiting superior overall performance.

3.4 Ablation Studies

We conducted extensive ablation studies on our proposed feature dynamic fusion module based on hypernetworks and semantic synergy mechanisms using the PHOENIX14T dataset.

3.4.1 Influence of Feature Fusion Type

Experiment No. 2 in Table 2 shows that using a simple sample level mixup results in a 1.25 point decrease in BLEU4 score, indicating the need for carefully designed dual stream dynamic fusion networks. Experiment No. 3 utilizes a gated method, generating two sets of static weights for weighting RGB and Keypoint features. Experiment No. 4 employs the traditional Linear+ReLU approach, resulting in static model weights. Compared to these two experiments, the method proposed in this paper uses hypernetworks to dynamically generate fusion weights based on individual sample features. On the Dev set, relative to Experiments No. 3 and No. 4, our method achieved improvements of +2.84 and +2.02 on the BLEU4 and +1.32 and +0.69 on the ROUGE metrics, respectively, demonstrating the superiority of dynamic feature fusion via hypernetworks.

[1] https://github.com/FangyunWei/SLRT

No.	Method	BLEU4	ROUGE
1	**HyperSign**	**29.78**	**54.98**
Influence of Feature Fusion Type			
2	MixUp	28.53	54.09
3	Gated	26.94	52.96
4	Linear + ReLU	28.46	54.29
Influence of Partial Shortcut Connection Strategy			
5	R-Drop	29.05	54.67
6	Fixed Value 0.5	29.12	54.73
7	Random MixUp	28.87	54.52
Influence of Self-Distillation Strategy			
8	w/o Encoder KL	28.27	54.02
9	w/o Decoder KL	27.12	52.38
Influence of SST Contrastive Learning			
10	w/o Contrastive Learning	28.89	54.87
11	w/o SST	28.67	54.42
12	w/o Cross Stream	29.49	54.92

Table 2: Studies of contribution for each component on Dev set of PHOENIX14T.

Metric	DSG	DShG	PIG	CIG
BLEU4	29.78	28.89	29.06	28.98
ROUGE	54.98	54.61	54.54	54.62

Table 3: Performance of Different Hypernet Variants on BLEU4 and ROUGE Scores on the PHOENIX14T Dev Set.

3.4.2 Influence of Different Hypernetwork Variants

In this study, we designed and tested with four variants of hypernetworks to evaluate their performance on the SLT task, as illustrated in Figure 4 (see figure caption for detailed structures). The impact of these variants on the HyperSign model's performance is summarized in Table 3. **DSG** achieved the highest performance with a BLEU4 score of 29.78 and a ROUGE score of 54.98, as its separate generators allowed for distinct optimization of RGB and keypoint features. **DShG**, which used a single generator for both streams, resulted in a BLEU4 score decrease of 0.89 due to less precise adaptations. **PIG** processed RGB and keypoint features separately, but the lack of interaction during fusion led to a BLEU4 score of 29.72, slightly lower than DSG. **CIG**, with cross-stream processing, also saw a decline in performance, reducing the BLEU4 score by 0.80 compared to DSG.

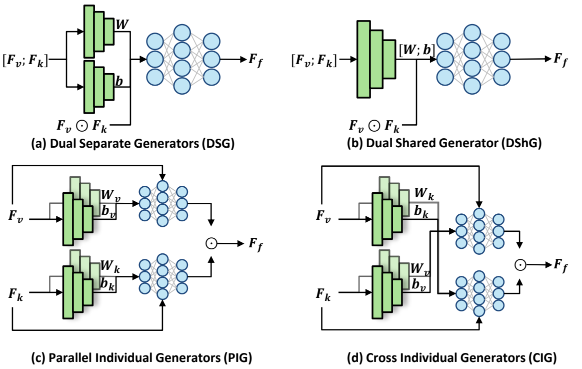

Figure 4: Illustration of hypernetwork variants for dynamic feature fusion. **(a) Dual Separate Generators (DSG)** employs two distinct generators to dynamically produce unified weights W and biases b, which are then applied to the combined RGB and keypoint features. **(b) Dual Shared Generator (DShG)** utilizes a single generator for both feature streams, outputting a vector that is segmented into W and b for the combined RGB and keypoint features. **(c) Parallel Individual Generators (PIG)** operates with individual generators for each stream, combining features after separate dynamic MLP processing. **(d) Cross Individual Generators (CIG)** incorporates cross-stream generator processing, where each stream's features are dynamically tailored by the generator of the alternate stream, with outputs combined following dynamic MLP processing.

3.4.3 Influence of Partial Shortcut Connection Strategy

In Table 2, Experiment No. 5 sets α to 0 to simulate R-Drop (Wu et al., 2021). Compared to Experiment No. 5, Experiment No. 6, with α set to 0.5, shows a slight performance improvement, suggesting that partial shortcut connections can further enhance the model's effectiveness. Relative to Experiment No. 7, which randomly samples α within the range of [0,1], our HyperSign method improved the BLEU4 score by 0.91 points, underscoring the efficacy of the progressive α sampling strategy introduced in this study.

3.4.4 Influence of Self-Distillation Strategy

In Experiment No. 8 and Experiment No. 9, we removed the feature self-distillation KL divergence strategy at both the encoder and decoder stages. The results showed a notable decline in model performance, with a decrease of 1.51 BLEU4 points in Experiment No. 8 and 2.66 BLEU4 points in Experiment No. 9. This decline substantiates the critical role that self-distillation plays in enhancing semantic fusion and alignment within the model, ensuring that the RGB and Keypoint streams can learn complementary information from each other.

3.4.5 Influence of SST Contrastive Learning

In Table 2, removing SST contrastive learning in Experiment No. 10 reduced the BLEU4 score by 0.89, underscoring its role in dual-stream feature fusion. In Experiment No. 11, substituting SST aggregation with average pooling lowered the score by 1.11, demonstrating the advantage of mapping visual and textual information into the same SST space to address feature granularity. Experiment No. 12 further showed that cross-stream contrastive learning improves dual-stream semantic alignment. Figure 5 illustrates that using 128 SSTs and token dimensions optimally balances computational complexity and feature granularity for SLT.

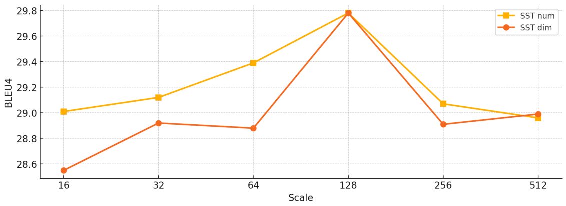

Figure 5: BLEU4 scores from an ablation study on token numbers and dimensions in HyperSign trained on the PHOENIX14T Dev set.

4 Related Work

4.1 Sign Language Translation

Sign Language Translation (SLT) involves translating sign language videos into spoken text. There are two mainstream translation paradigms: the traditional pipeline approach, which first involves sign language recognition to annotate sign language videos with text Glosses, followed by a text decoder to translate the Glosses into spoken text. Although this approach has provided substantial enhancements for SLT, it introduces a predefined information bottleneck that limits translation accuracy and typically fails to provide long-term dependencies and contextual information (Camgoz et al., 2020; Kapoor et al., 2021; Zhou et al., 2023). Consequently, more and more methods (Chen et al., 2022a,b; Zhao et al., 2024) use a visual encoder directly as a tokenizer to extract visual features and a text decoder to directly generate spoken text, breaking the limitations of traditional methods. The text decoders typically utilize sequence models such as RNNs (Camgoz et al., 2018), LSTMs (Guo et al., 2018), or Transformers (Zhang et al., 2023; Yu et al., 2023; Yao et al., 2023; Zhou et al., 2023).

In this study, we focus on optimizing this direct translation approach, particularly emphasizing the dynamic fusion of dual-stream visual features and semantic alignment, to achieve higher translation accuracy and model efficiency.

4.2 Multi Stream Fusion

Multi-stream networks have shown great potential in video understanding and sign language recognition as they can handle data from different sources and capture dynamic information and details across modalities effectively. Researchers often use attention mechanisms (Mo and Morgado, 2023; Shan et al., 2024; Lv et al., 2024) or gated networks (Ai and Wang, 2024; Yi et al., 2024) to fuse features from different streams. In sign language tasks, the design of multi-stream networks typically includes handling multiple independent data streams for video, such as a single RGB stream (Chen et al., 2022a), and adding keypoint streams (Chen et al., 2022b), optical flow (Cui et al., 2019; Chen et al., 2024), and cropped feature map streams (Zheng et al., 2021).

The keypoint streams have demonstrated substantial enhancement in model performance. However, previous methods combining RGB and keypoint streams (Hu et al., 2023; Chen et al., 2022b) typically used a late fusion strategy, i.e., fine-tuning two different pretrained language models separately to obtain their own translation networks and then averaging their translation logits to aggregate effects. This approach failed to effectively address distribution discrepancies between streams early on, thus necessitating separate fine-tuning of pretrained language models for each stream, leading to resource wastage.

5 Conclusion

We propose an efficient dual-stream fusion model for sign language translation using an early fusion strategy. Our model achieves dynamic fusion of visual stream features through a hypernetwork and partial shortcut connections. Additionally, our proposed semantic synergy alignment mechanism significantly reduces representational differences between RGB and keypoint feature streams, further strengthening semantic alignment between the streams. Experiments on two SLT datasets demonstrate that our approach surpasses all state-of-the-art methods. Our method can be applied to various sign languages. In future work, we hope to introduce more effective visual streams and explore more advanced fusion strategies to achieve better visual representations.

Limitations

The current sign language translation models have certain limitations when applied in multi-sign language translation environments, mainly reflected in the following aspects. First, the language processing capability of the model is typically limited to a single sign language or a specific language, which significantly restricts the model's performance in communication scenarios involving multiple sign language users.

Secondly, existing sign language datasets are often focused on specific domains or scenarios. The limitations of the data result in suboptimal performance of the model in complex and dynamic environments. Due to the homogeneity of the data samples, the model may struggle to fully meet the sign language expression needs across different domains, thus limiting its broader applicability.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62476232 and 62076211 and in part by the University-Industry Cooperation Programs of Fujian Province of China under Grant No. 2023H6001.

References

Hao Ai and Lin Wang. 2024. Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. *arXiv preprint arXiv:2403.16376*.

Kenan Emir Ak, Gwang-Gook Lee, Yan Xu, and Mingwei Shen. 2023. Leveraging efficient training and feature fusion in transformers for multimodal classification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1420–1424. IEEE.

Marc Baboulin, Alfredo Buttari, Jack Dongarra, Jakub Kurzak, Julie Langou, Julien Langou, Piotr Luszczek, and Stanimire Tomov. 2009. Accelerating scientific computations with mixed precision algorithms. *Computer Physics Communications*, 180(12):2526–2533.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Hao Chen, Jiaze Wang, Ziyu Guo, Jinpeng Li, Donghao Zhou, Bian Wu, Chenyong Guan, Guangyong Chen, and Pheng-Ann Heng. 2024. Signvtcl: Multi-modal continuous sign language recognition enhanced by visual-textual contrastive learning. *arXiv preprint arXiv:2401.11847*.

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

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

Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N Metaxas, and Hongxia Yang. 2023. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15095–15104.

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

Stephen Grossberg and Michael E Rudd. 1989. A neural architecture for visual motion perception: Group and element apparent motion. *Neural Networks*, 2(6):421–450.

Dan Guo, Wen gang Zhou, Anyang Li, Houqiang Li, and Meng Wang. 2020. Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation. *IEEE Transactions on Image Processing*, 29:1575–1590.

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2018. Hierarchical lstm for sign language translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

David Ha, Andrew Dai, and Quoc V Le. 2016. Hypernetworks. *arXiv preprint arXiv:1609.09106*.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239.

Lianyu Hu, Wei Feng, Liqing Gao, Zekang Liu, and Liang Wan. 2024. Corrnet+: Sign language recognition and translation via spatial-temporal correlation. *arXiv preprint arXiv:2404.11111*.

Parul Kapoor, Rudrabha Mukhopadhyay, Sindhu B Hegde, Vinay Namboodiri, and CV Jawahar. 2021. Towards automatic speech to sign language generation. *arXiv preprint arXiv:2106.12790*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhengyao Lv, Yuxiang Wei, Wangmeng Zuo, and Kwan-Yee K Wong. 2024. Place: Adaptive layout-semantic fusion for semantic image synthesis. *arXiv preprint arXiv:2403.01852*.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR.

Shentong Mo and Pedro Morgado. 2023. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. *arXiv preprint arXiv:2312.01017*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ziyu Shan, Yujie Zhang, Qi Yang, Haichen Yang, Yiling Xu, Jenq-Neng Hwang, Xiaozhong Xu, and Shan Liu. 2024. Contrastive pre-training with multi-view fusion for no-reference point cloud quality assessment. *arXiv preprint arXiv:2403.10066*.

Shiv Shankar, Laure Thompson, and Madalina Fiterau. 2022. Progressive fusion for multimodal integration. *arXiv preprint arXiv:2209.00302*.

Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R Pomerantz, Peter A Van der Helm, and Cees Van Leeuwen. 2012. A century of gestalt psychology in visual perception: Ii. conceptual and theoretical foundations. *Psychological bulletin*, 138(6):1218.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

F. Xiao, Cong Shen, Tiantian Yuan, and Shengyong Chen. 2021. Crb-net: A sign language recognition deep learning strategy based on multi-modal fusion with attention mechanism *. *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2562–2567.

Huijie Yao, Wengang Zhou, Hao Feng, Hezhen Hu, Hao Zhou, and Houqiang Li. 2023. Sign language translation with iterative prototype. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15592–15601.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. 2024. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. *arXiv preprint arXiv:2403.16387*.

Pei Yu, Liang Zhang, Biao Fu, and Yidong Chen. 2023. Efficient sign language translation with a curriculum-based non-autoregressive decoder. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5260–5268.

Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. Sltunet: A simple unified model for sign language translation. *arXiv preprint arXiv:2305.01778*.

Ruiquan Zhang, Cong Hu, Pei Yu, and Yidong Chen. 2025. Improving multilingual sign language translation with automatically clustered language family information. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3579–3588.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. 2024. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19643–19651.

Jiangbin Zheng, Yidong Chen, Chong Wu, Xiaodong Shi, and Suhail Muhammad Kamal. 2021. Enhancing neural sign language translation by highlighting the facial expression information. *Neurocomputing*, 464:462–472.

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

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

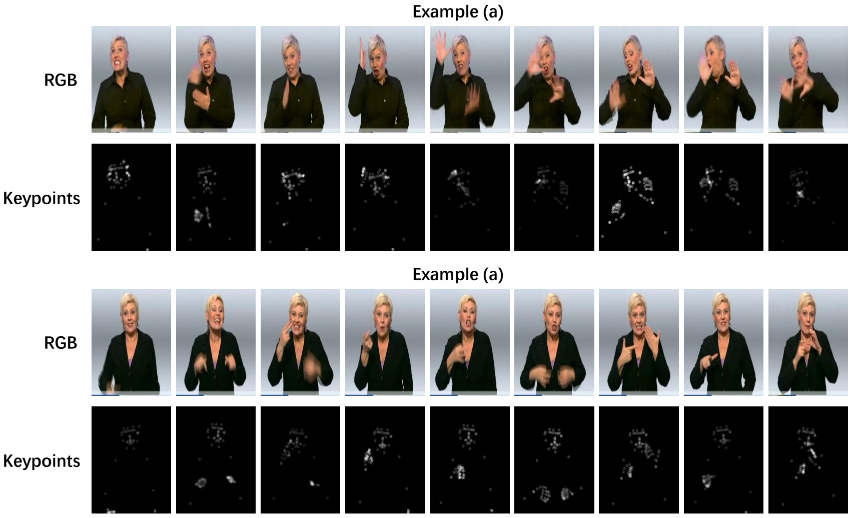

Figure 6: Visualization of Raw Videos and Keypoint Sequences from the PHOENIX14T dataset.

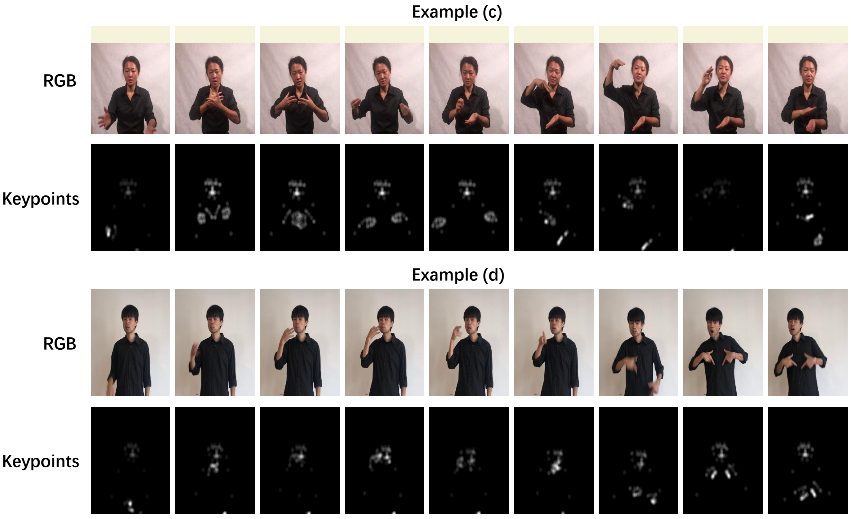

Figure 7: Visualization of Raw Videos and Keypoint Sequences from the CSL-Daily Dataset.

A Qualitative Results

A.1 Keypoint Modeling

We employ a methodology similar to the TS-Network, utilizing HRNet to extract keypoints from RGB videos and generate corresponding keypoint heatmaps. As illustrated in Figures 6 and 7, the heatmaps clearly and effectively capture the signer's appearance, hand positions, and palm orientations, also intuitively displaying the movement information of the signer's limb actions.

A.2 Results of HyperSign

As shown in Tables 4 and 5, we conducted a qualitative analysis of HyperSign, showcasing two samples from the PHOENIX14T and CSL-Daily development sets respectively. The experimental results indicate that HyperSign can dynamically fuse dual-stream features based on the respective strengths of RGB and keypoints, thereby enhancing the translation performance.

Example (a)	**Translation**
Gloss	SKANDINAVIEN / NORD / WOLKE / TIEF / AUCH / DEUTSCHLAND / REGION / KOMMEN (Scandinavia / north / cloud / low / also / Germany / region / come)
Groundtruth	die wolken des tiefs über nordskandinavien überqueren bis morgen auch die östlichen teile deutschlands. (The clouds of the low over northern Scandinavia will also cross the eastern parts of Germany by tomorrow.)
RGB only	die wolken eines tiefs über skandinavien überqueren heute nacht den norden deutschlands. (The clouds of a low over Scandinavia will cross the north of Germany tonight.)
Keypoint only	die wolken des tiefs über nordskandinavien überqueren bis morgen auch die östlichen teile deutschlands. (The clouds of the low over northern Scandinavia will also cross the eastern parts of Germany by tomorrow.)
HyperSign	die wolken des tiefs über nordskandinavien überqueren bis morgen auch die östlichen teile deutschlands. (The clouds of the low over northern Scandinavia will also cross the eastern parts of Germany by tomorrow.)
Example (b)	**Translation**
Gloss	JETZT / WETTER / WIE-AUSSEHEN / MORGEN / SONNTAG / SIEBEN / ZWANZIG / SEPTEMBER (Now / weather / appearance / tomorrow / Sunday / seven / twenty / September)
Groundtruth	und nun die wettervorhersage für morgen sonntag den siebenundzwanzigsten september. (And now the weather forecast for tomorrow, Sunday the twenty-seventh of September.)
RGB only	und nun die wettervorhersage für morgen sonntag den siebenundzwanzigsten september. (And now the weather forecast for tomorrow, Sunday the twenty-seventh of September.)
Keypoint only	und nun die wettervorhersage für morgen sonntag den sechsundzwanzigsten juli. (And now the weather forecast for tomorrow, Sunday the twenty-sixth of July.)
HyperSign	und nun die wettervorhersage für morgen sonntag den siebenundzwanzigsten september. (And now the weather forecast for tomorrow, Sunday the twenty-seventh of September.)

Table 4: Qualitative results on PHOENIX14T.

Example (c)	**Translation**
Gloss	爱 / 感情 / 跳 / 每 / 脉搏 (Love / emotion / beat / every / heartbeat)
Groundtruth	爱情跳动在每一次脉搏之中。 (Love beats with every heartbeat.)
RGB only	爱情是跳跃的, 每个人都有脉搏。 (Love is jumping, everyone has a heartbeat.)
Keypoint only	爱情跳动在每一次脉搏之中。 (Love beats with every heartbeat.)
HyperSign	爱情跳动在每一次脉搏之中。 (Love beats with every heartbeat.)
Example (d)	**Translation**
Gloss	猫 / 看 / 这 / 小 / 鸟 / 惊吓 (Cat / look / this / small / bird / frighten)
Groundtruth	猫被背后的小鸟惊吓。 (The cat is frightened by the small bird behind it.)
RGB only	猫被背后的鸟惊吓。 (The cat is frightened by the bird behind it.)
Keypoint only	猫前的那只小鸟被惊吓了。 (The small bird in front of the cat is frightened.)
HyperSign	猫被背后的小鸟惊吓。 (The cat is frightened by the small bird behind it.)

Table 5: Qualitative results on CSL-Daily.