

TransAlign-SLT: Complementary Temporal and Semantic Supervision for Gloss-Free Sign Language Translation

Pallapu Mohan Krishna¹*, Uppala Sanjay Kumar², M. K. Bhuyan^{1,3}, Debanga Raj Neog¹

¹Mehta Family School of Data Science and Artificial Intelligence

²Department of Mathematics ³Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Assam, India

{k.pallapu,k.uppala,mkb,dneog}@iitg.ac.in

Abstract

Gloss-free sign language translation (SLT) aims to translate sign videos directly into spoken-language sentences without relying on costly gloss annotations. However, when supervision is available only at the sentence level, the correspondence between visual dynamics and linguistic content remains weak. We introduce **TransAlign-SLT**, a lightweight fully gloss-free framework that derives complementary temporal and semantic supervision from the available translations. Auxiliary TranslationCTC provides coarse temporal regularization, while Masked Entropic Optimal Transport (MOT) aligns visual temporal features with contextualized word-level representations from a frozen, corpus-adapted BERT encoder. Since the semantic branch is used only during training, the inference pipeline remains lightweight and autoregressive. TransAlign-SLT achieves **23.54 BLEU-4 / 48.46 ROUGE-L** on PHOENIX14T and **16.39 BLEU-4 / 44.68 ROUGE-L** on CSL-Daily, with **45.6M** inference parameters and **123 GFLOPs**. Ablation studies show that temporal modeling, semantic alignment, and decayed TranslationCTC supervision provide complementary benefits. Overall, the results show that useful visual-language structure can be learned directly from translation supervision without glosses, pseudo-glosses, or a large pre-trained language decoder at inference.

1 Introduction

Sign languages express meaning through coordinated hand movements, facial expressions, body posture, and other non-manual cues. Sign Language Translation (SLT) aims to translate continuous sign videos into spoken-language sentences, supporting communication between signed and spoken languages.

Many existing SLT systems use gloss annotations as an intermediate form of supervision (Cam-

goz et al., 2018, 2020). Glosses provide useful temporal and lexical structure, but producing them requires linguistic expertise and substantial annotation effort, which makes them difficult to scale across languages and domains. This limitation has encouraged growing interest in *gloss-free* SLT, where models learn directly from paired sign videos and sentence-level translations (Zhou et al., 2023; Fu et al., 2023; Chen et al., 2025). In this setting, however, the model must infer both temporal structure and visual-language correspondence from the relatively weak supervision provided by sentence-level translations.

Recent gloss-free methods address this difficulty in several ways, including gloss-inspired attention (Yin et al., 2023), visual-language pretraining (Zhou et al., 2023; Jiao et al., 2024), cross-modal contrastive learning (Fu et al., 2023; Chen et al., 2025), and the use of large pretrained language or multimodal models (Wong et al., 2024; Gong et al., 2024; Chen et al., 2024; Kim et al., 2025). While these approaches have improved translation performance, they may involve additional pretraining stages, auxiliary recognition models, pseudo-label generation, or large language components. This raises a natural question: *can useful temporal and semantic structure be learned directly from raw video-translation pairs without gloss supervision or a large pretrained language decoder at inference?*

To investigate this question, we introduce **TransAlign-SLT**, a fully gloss-free framework that combines complementary temporal and semantic supervision. The visual stream couples multi-scale temporal modeling with bidirectional sequence refinement. During training, **TranslationCTC** provides a coarse temporal-linguistic prior, while **Masked Entropic Optimal Transport (MOT)** encourages soft semantic correspondence between visual temporal features and contextualized word-level representations obtained from a

*Corresponding author: k.pallapu@iitg.ac.in

frozen, corpus-adapted BERT encoder. To obtain word-level representations, WordPiece subwords belonging to the same lexical unit are pooled before transport alignment. Translation is then generated autoregressively using an attention-based GRU decoder, with optional post-hoc CTC rescoring at inference.

We evaluate TransAlign-SLT on PHOENIX14T and CSL-Daily and examine the contributions of temporal modeling, semantic supervision, TranslationCTC weighting, and decoding.

Our main contributions are:

- We introduce **TransAlign-SLT**, a fully gloss-free SLT framework trained directly from sign videos and spoken-language translations without gloss or pseudo-gloss supervision.
- We combine **TranslationCTC** for coarse temporal regularization with **MOT-based semantic alignment** using corpus-adapted contextualized word representations and subword-aware pooling.
- We evaluate the framework on PHOENIX14T and CSL-Daily and provide component-level analyses of temporal modeling, semantic supervision, TranslationCTC, and decoding, while retaining an inference pipeline that does not require a large pretrained language decoder.

The implementation is publicly available at <https://github.com/pallapumohankrishna/TransAlign-SLT>.

2 Related Work

2.1 Gloss-Supervised and Weakly Supervised SLT

Early SLT systems often relied on gloss annotations as an intermediate source of supervision (Camgoz et al., 2018, 2020). Later work improved temporal, multi-cue, and sequence modeling through stochastic transformers, spatial-temporal representations, cross-modal reranking, and stronger sequence architectures (Voskou et al., 2021; Zhou et al., 2022; Zhao et al., 2022; Chen et al., 2022a,b; Zhang et al., 2023; Zhou et al., 2021). While these approaches have been effective, many of them depend on gloss annotations, which require substantial linguistic expertise and are difficult to scale across languages and domains.

Weakly supervised methods aim to reduce this dependence by introducing auxiliary alignment or pseudo-label signals. TSPNet (Li et al., 2020) explored temporal-semantic correspondence, while GASLT (Yin et al., 2023) introduced gloss attention and transferred sentence-level similarity knowledge from a pretrained natural language model. ConSLT (Fu et al., 2023) introduced contextual cross-modal contrastive learning, and VAP (Jiao et al., 2024) employed visual alignment pretraining. Although these approaches reduce reliance on manual gloss supervision, they may still involve pseudo-glosses, pretrained recognition models, or additional pretraining stages.

2.2 Fully Gloss-Free SLT

Fully gloss-free SLT learns directly from paired sign videos and spoken-language translations without using intermediate gloss supervision. GFSLT-VLP (Zhou et al., 2023) introduced visual-language pretraining, while C²RL (Chen et al., 2025) used cross-modal contrastive representation learning to improve semantic consistency between visual and textual representations. More recent systems, including FLa-LLM (Chen et al., 2024), Sign-LLM (Gong et al., 2024), and MMSLT (Kim et al., 2025), increasingly make use of large pretrained language or multimodal models to strengthen translation generation.

TransAlign-SLT takes a different approach by deriving complementary supervision directly from the paired translation signal. TranslationCTC provides a coarse temporal prior, while masked optimal transport supplies training-time semantic correspondence with contextualized word representations. This design keeps the inference pipeline fully gloss-free and avoids both pseudo-gloss generation and the use of a large pretrained language decoder.

3 Methodology

3.1 Method Overview

Given a continuous sign video $\mathbf{X} = \{x_1, \dots, x_T\}$ and its corresponding spoken-language translation $\mathbf{Y} = \{y_1, \dots, y_N\}$, fully gloss-free SLT learns a mapping $f_\theta : \mathbf{X} \rightarrow \mathbf{Y}$ without relying on intermediate gloss or pseudo-gloss supervision.

As shown in Figure 1, TransAlign-SLT combines a visual-temporal encoding pipeline with complementary temporal and semantic training signals. The encoding process is defined as

$$\mathbf{X} \xrightarrow{\phi_{\text{CNN}}} \mathbf{F} \xrightarrow{\phi_{\text{HMTF}}} \mathbf{H}^{\text{conv}} \xrightarrow{\phi_{\text{BiLSTM}}} \mathbf{H}^{\text{seq}}, \quad (1)$$

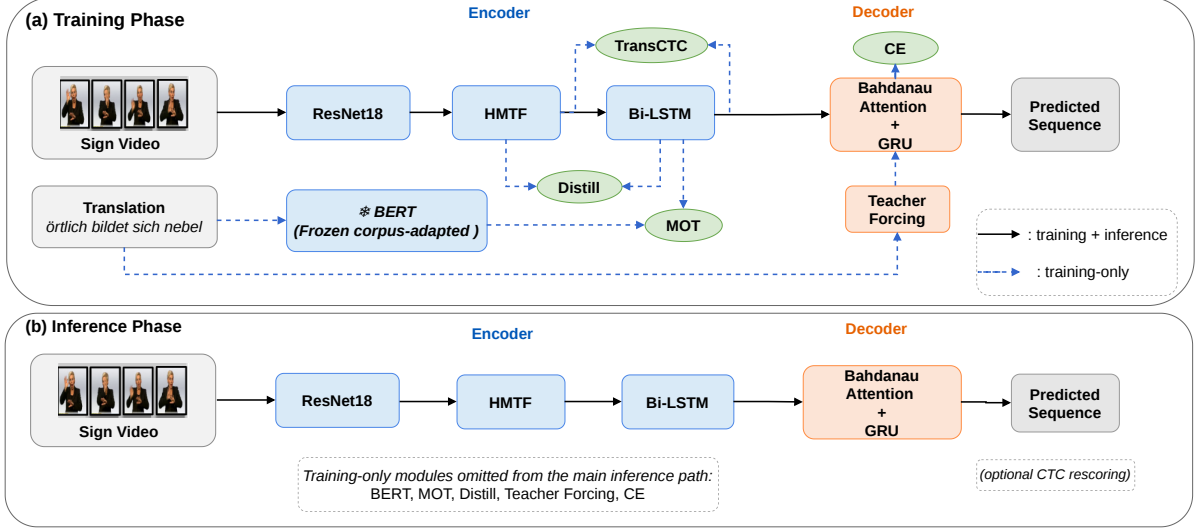


Figure 1: Overview of TransAlign-SLT during training and inference. ResNet18 features are processed by HMTF and BiLSTM before autoregressive translation with a Bahdanau-attention GRU. During training, TranslationCTC provides auxiliary temporal guidance, while frozen corpus-adapted BERT representations provide semantic supervision through MOT. BERT, MOT, distillation, teacher forcing, and training losses are excluded from the main inference path. TranslationCTC scores may optionally be used for post-hoc beam rescoring.

where ϕ_{CNN} is an ImageNet-pretrained ResNet18 (He et al., 2016; Deng et al., 2009), ϕ_{HMTF} denotes the Hierarchical Multi-scale Temporal Fusion encoder adapted from GPGN (Guo et al., 2024), and ϕ_{BiLSTM} provides bidirectional sequence refinement (Graves and Schmidhuber, 2005).

TranslationCTC encourages coarse temporal organization of the visual representations, while Masked Entropic Optimal Transport (MOT) provides soft visual-text semantic correspondence. Together, these signals improve representation learning from sentence-level translation supervision before an attention-based GRU generates the target sentence autoregressively.

3.2 Hierarchical Multi-scale Temporal Fusion

HMTF adapts the Parallel Densely-Connected Temporal Feature Extraction design of GPGN (Guo et al., 2024) to the translation setting. We adopt only the architectural design from GPGN and do not initialize HMTF using GPGN checkpoints or any gloss-supervised pretrained weights. Parallel dilated temporal convolutions capture information over multiple receptive fields, while cross-scale residual connections and hierarchical fusion integrate the resulting temporal representations. A detailed view of the HMTF architecture is shown in Figure 2.

Given frame-level features $\mathbf{F}$ and valid lengths $\mathbf{L}$,

HMTF produces temporal representations, TranslationCTC logits, and encoded sequence lengths:

$$(\mathbf{H}^{\text{conv}}, \mathbf{Z}^{\text{conv}}, \ell) = \phi_{\text{HMTF}}(\mathbf{F}, \mathbf{L}). \quad (2)$$

Here, $\mathbf{H}^{\text{conv}} \in \mathbb{R}^{\ell_{\text{max}} \times B \times d_h}$ and $\mathbf{Z}^{\text{conv}} \in \mathbb{R}^{\ell_{\text{max}} \times B \times V_{\text{ctc}}}$.

The fused temporal representations are then refined by a stacked BiLSTM:

$$\mathbf{H}^{\text{seq}} = \phi_{\text{BiLSTM}}(\mathbf{H}^{\text{conv}}, \ell), \quad (3)$$

which introduces bidirectional sequence context that complements the multi-scale convolutional modeling performed by HMTF.

3.3 Auxiliary TranslationCTC Supervision

Without gloss supervision, there is no explicit intermediate correspondence between the visual sequence and linguistic units. We therefore apply TranslationCTC, based on Connectionist Temporal Classification (CTC) (Graves et al., 2006), to both the HMTF and BiLSTM representations as an auxiliary temporal regularizer.

Let $r \in \{\text{conv}, \text{seq}\}$ denote the two branches, and let $\mathbf{Z}_b^r$ denote their logits produced by a shared classifier. For sample b , the branch-specific objective is

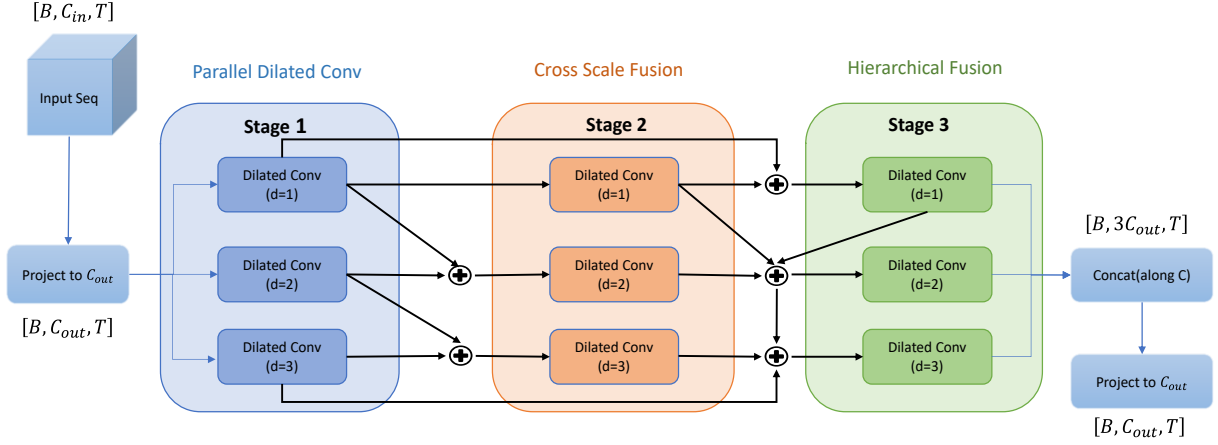


Figure 2: Hierarchical Multi-scale Temporal Fusion (HMTF). Parallel dilated convolutions capture multiple temporal scales, cross-scale residual connections exchange information across branches, and hierarchical fusion produces the final temporal representation.

$$\mathcal{L}_{\text{CTC}}^r = \frac{1}{B} \sum_{b=1}^B \text{CTC} \left(\log \text{softmax}(\mathbf{Z}_b^r), \mathbf{y}_b^{\text{ctc}}, \ell_b, N_b \right). \quad (4)$$

where $\mathbf{y}_b^{\text{ctc}}$ is the target spoken-language token sequence, ℓ_b is the valid visual sequence length, and N_b denotes the corresponding target sequence length.

TranslationCTC uses the same translation vocabulary as the decoder, excluding PAD, BOS, and EOS tokens, with an additional CTC blank class. We use PyTorch CTCLoss with zero_infinity=True to handle invalid alignments.

During the early stages of training, the visual representations are only weakly organized with respect to the target language. As a result, fine-grained visual-semantic correspondence is difficult to learn directly from sentence-level supervision. TranslationCTC is therefore assigned a relatively strong initial weight, providing a coarse token-level temporal-linguistic prior that helps organize the visual sequence.

This weight is smoothly reduced from 0.5 to 0.1 over the course of training. Reducing its influence relaxes the monotonic auxiliary constraint and increases the relative contribution of MOT-based semantic alignment and autoregressive translation learning. This is important because sign-to-spoken-language translation does not necessarily follow a strictly monotonic frame-to-word correspondence. TranslationCTC therefore acts primarily as an early

optimization prior rather than as the final alignment mechanism.

We further transfer sequence-level information from the BiLSTM branch to the convolutional branch through knowledge distillation:

$$\mathcal{L}_{\text{Dist}} = \frac{\tau^2}{N_{\text{valid}}} \sum_{b=1}^B \sum_{t=1}^{\ell_b} D_{\text{KL}} \left(\tilde{\mathbf{p}}_{t,b}^{\text{seq}} \parallel \tilde{\mathbf{p}}_{t,b}^{\text{conv}} \right), \quad (5)$$

where τ is the distillation temperature and $N_{\text{valid}} = \sum_b \ell_b$ is the total number of valid temporal positions. We set $\tau = 8$ in all experiments.

3.4 Masked Semantic Alignment with Frozen BERT

TranslationCTC provides temporal guidance, but it does not explicitly capture contextual sentence semantics. To complement this signal, we use **Masked Entropic Optimal Transport (MOT)** to establish soft semantic correspondence between visual temporal features and contextualized textual representations.

A language-specific BERT encoder (Devlin et al., 2019) is first adapted using only the training-set translations of the target corpus through masked language modeling and is then frozen during SLT training. Since WordPiece tokenization may divide a single lexical unit into multiple subwords, subwords associated with the same original word are mean-pooled before alignment, yielding $\mathbf{T}_b = \{\mathbf{t}_{b,j}\}_{j=1}^{T_b}$.

Each visual sequence feature is projected into the textual embedding space:

$$\mathbf{q}_{b,i} = \mathbf{W}_p \mathbf{h}_{b,i}^{\text{seq}} + \mathbf{b}_p, \quad (6)$$

and the corresponding visual-text transport cost is defined using cosine distance:

$$\mathbf{C}_{ij}^{(b)} = 1 - \hat{\mathbf{q}}_{b,i}^\top \hat{\mathbf{t}}_{b,j}. \quad (7)$$

To prevent padded positions from contributing to the alignment, uniform marginals are normalized only over valid visual and textual positions:

$$\mu_{b,i} = \frac{m_{b,i}^v}{\sum_k m_{b,k}^v}, \quad \nu_{b,j} = \frac{m_{b,j}^t}{\sum_k m_{b,k}^t}. \quad (8)$$

The entropically regularized transport plan is defined as

$$\pi^{(b)} = \arg \min_{\mathbf{T} \in \Pi(\mu_b, \nu_b)} \left[\langle \mathbf{T}, \mathbf{C}^{(b)} \rangle - \epsilon \mathcal{H}(\mathbf{T}) \right], \quad (9)$$

with the semantic alignment objective

$$\mathcal{L}_{\text{MOT}} = \frac{1}{B} \sum_{b=1}^B \langle \pi^{(b)}, \mathbf{C}^{(b)} \rangle. \quad (10)$$

We solve Eq. 9 using 50 log-space Sinkhorn iterations (Cuturi, 2013) with $\epsilon = 0.1$. In contrast to TranslationCTC, MOT does not explicitly enforce monotonic frame-to-word correspondence. It therefore provides a complementary form of semantic supervision. The frozen BERT/MOT branch is used only during training.

3.5 Translation Decoder and CTC Rescoring

The target translation is generated autoregressively using a single-layer GRU with Bahdanau additive attention (Bahdanau et al., 2015; Camgoz et al., 2018), conditioned on $\mathbf{H}^{\text{seq}}$. At decoding step u , the model predicts

$$P(y_u \mid y_{<u}, \mathbf{H}^{\text{seq}}) = \text{softmax} \left(\mathbf{W}_o[\mathbf{s}_u; \mathbf{c}_u; \mathbf{x}_u] + \mathbf{b}_o \right). \quad (11)$$

where $\mathbf{s}_u$ is the GRU hidden state and $\mathbf{c}_u$ denotes the Bahdanau attention context.

At inference, hypotheses are generated autoregressively using beam search and may optionally be reranked with TranslationCTC scores:

$$S_{\text{final}} = (1 - \beta_{\text{ctc}}) S_{\text{beam}} + \beta_{\text{ctc}} S_{\text{CTC}}. \quad (12)$$

The autoregressive beam score is normalized using a length penalty, while the TranslationCTC score is normalized by the target hypothesis length before interpolation. The CTC score is applied only

to completed hypotheses and does not participate in autoregressive beam expansion. Thus, it does not constrain the decoder’s generation path, although its sequence score retains the monotonic alignment assumption of CTC.

3.6 Overall Training Objective

The complete training objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ctc}}(e) (\mathcal{L}_{\text{CTC}}^{\text{conv}} + \mathcal{L}_{\text{CTC}}^{\text{seq}}) + 10 \mathcal{L}_{\text{Dist}} + 2 \mathcal{L}_{\text{MOT}} + \mathcal{L}_{\text{TransCE}}, \quad (13)$$

where $\lambda_{\text{ctc}}(e)$ is the epoch-dependent TranslationCTC coefficient, smoothly reduced from 0.5 to 0.1. TranslationCTC provides coarse temporal regularization, MOT contributes contextual semantic alignment, and translation cross-entropy directly optimizes the final autoregressive translation objective.

4 Experiments

4.1 Datasets and Evaluation

Datasets. We evaluate TransAlign-SLT on RWTH-PHOENIX-Weather 2014T (PHOENIX14T) (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021). PHOENIX14T consists of 7,096 training, 519 development, and 642 test video-translation pairs collected from German weather broadcasts. Its spoken-language vocabulary contains 2,887 words. CSL-Daily contains 18,401 training, 1,077 development, and 1,176 test examples covering daily-life communication, with a spoken-language vocabulary of 2,343 Chinese words. Although gloss annotations are available for both datasets, TransAlign-SLT uses only sign videos and spoken-language translations during training and inference.

Metrics.

Following standard SLT evaluation protocols, we report BLEU-1 through BLEU-4 (Papineni et al., 2002) together with ROUGE-L (Lin, 2004). BLEU scores are computed using NLTK’s corpus-level BLEU implementation with smoothing method 3. ROUGE-L is computed as the mean sentence-level F1 score based on the longest common subsequence of token sequences. The translation text is lowercased, and punctuation is removed before dictionary encoding.

4.2 Implementation Details

We use a ResNet18 backbone initialized with ImageNet-pretrained weights from torchvision.

Following TSN-style sampling (Wang et al., 2016), 50% of the frames are retained during training by partitioning each video into uniformly spaced temporal segments and randomly sampling one frame from each segment. Each sampled frame is resized to 256×256 , randomly cropped to 224×224 , and horizontally flipped. During inference, we sample the same number of frames at uniformly spaced temporal positions and apply center cropping.

The semantic branch uses a language-specific BERT encoder with a hidden dimension of 768. The encoder is adapted using only the training-set translations of the corresponding benchmark, with MLM masking ratios of 0.5 for PHOENIX14T and 0.3 for CSL-Daily. After this adaptation, BERT remains frozen throughout SLT training and is not used during inference.

All models are trained for 60 epochs using Adam. For PHOENIX14T, both the learning rate and weight decay are set to 10^{-4} . For CSL-Daily, the learning rate is 5×10^{-5} and the weight decay is 10^{-6} . The TranslationCTC coefficient is gradually reduced from 0.5 to 0.1. MOT is computed using 50 log-space Sinkhorn iterations with $\epsilon = 0.1$. All experiments were conducted on an NVIDIA RTX A6000 GPU.

At inference, we use beam search with a beam size of 10, followed by post-hoc TranslationCTC reranking with $\beta_{\text{ctc}} = 0.6$. CTC scores are applied only to completed hypotheses and therefore do not affect autoregressive beam expansion. The deployed model contains 45.6M parameters and requires 123 GFLOPs for 64 RGB frames at 224×224 . The frozen BERT/MOT branch is used only during training.

4.3 Main Results

Tables 1 and 2 summarize performance on both the development and test sets. All component-level ablations are conducted on the PHOENIX14T development split.

PHOENIX14T. TransAlign-SLT achieves 22.86 BLEU-4 / 47.63 ROUGE-L on the development set and 23.54 BLEU-4 / 48.46 ROUGE-L on the test set (Table 1). Among fully gloss-free systems, it outperforms GFSLT-VLP, FLA-LLM, and SignLLM in test BLEU-4, while STN-SLT, C²RL, and MMSLT achieve higher scores. These results show that competitive translation performance can be obtained directly from video-translation pairs without gloss or pseudo-gloss supervision and without

relying on a large pretrained language decoder at inference.

CSL-Daily. On CSL-Daily, TransAlign-SLT obtains 16.17 BLEU-4 / 44.13 ROUGE-L on the development set and 16.39 BLEU-4 / 44.68 ROUGE-L on the test set (Table 2). It again outperforms GFSLT-VLP, FLA-LLM, and SignLLM in test BLEU-4, while C²RL and MMSLT achieve higher scores. These results show that TransAlign-SLT remains competitive across two distinct SLT benchmarks, while component-level ablations are conducted on PHOENIX14T.

4.4 Qualitative Analysis

Table 3 presents representative examples of successful semantic preservation together with remaining translation errors. In the first example, TransAlign-SLT preserves most of the sentence-level content while producing a slightly shorter translation. In the second example, the model retains the geographic and weather context but fails to preserve the distinction between *Schnee* (snow) and *Schneeregen* (sleet). The third example illustrates a fine-grained numerical error, where *fünften* (fifth) is replaced by *zehnten* (tenth). Taken together, these examples suggest that the model captures broad semantic content more reliably than fine-grained lexical and numerical details.

5 Ablation and Analysis

We study the contribution of temporal modeling, semantic supervision, TranslationCTC, decoding strategy, and decoder choice on the PHOENIX14T development set.

5.1 Temporal Modeling Components

We begin by examining the temporal modeling components of TransAlign-SLT. Table 4 compares the full HMTF + BiLSTM encoder with two controlled architectural variants.

Replacing HMTF with a conventional temporal convolution module while retaining the BiLSTM lowers BLEU-4 from 22.86 to 18.18, corresponding to a drop of 4.68 points. Keeping HMTF but removing the BiLSTM reduces BLEU-4 to 20.14 and ROUGE-L to 43.77. These results indicate that the two stages play complementary roles. HMTF captures temporal patterns at multiple scales, while the BiLSTM further refines these representations using bidirectional sequence context.

Method	Dev					Test				
	B1	B2	B3	B4	R	B1	B2	B3	B4	R
Gloss-supervised SLT										
SLRT (Camgoz et al., 2020)	47.26	34.40	27.05	22.38	—	46.61	33.73	26.19	21.32	—
STMC-T (Zhou et al., 2022)	47.60	36.43	29.18	24.09	48.24	46.98	36.09	28.70	23.65	46.65
BN-TIN-Transf.+SignBT (Zhou et al., 2021)	51.11	37.90	29.80	24.45	50.29	50.80	37.75	29.72	24.32	49.54
MMTLB (Chen et al., 2022a)	53.95	41.12	33.14	27.61	53.10	53.97	41.75	33.84	28.39	52.65
TS-SLT (Chen et al., 2022b)	54.32	41.99	34.15	28.66	54.08	54.90	42.43	34.46	28.95	53.48
SLTUNET (Zhang et al., 2023)	—	—	—	27.87	52.23	52.92	41.76	33.99	28.47	52.11
Weakly Supervised Gloss-free SLT										
TSPNet (Li et al., 2020)	—	—	—	—	—	36.10	23.12	16.88	13.41	34.96
GASLT (Yin et al., 2023)	—	—	—	—	—	39.07	26.74	21.86	15.74	39.86
ConSLT (Fu et al., 2023)	—	—	—	21.11	47.74	—	—	—	21.59	47.69
Sign2GPT (Wong et al., 2024)	—	—	—	—	—	49.54	35.96	28.83	22.52	48.90
VAP (Jiao et al., 2024)	52.78	—	—	26.62	51.47	53.07	—	—	26.16	51.28
Fully Gloss-free SLT										
NSLT (Camgoz et al., 2018)	28.10	16.81	11.82	9.12	31.00	27.10	15.61	10.82	8.35	29.70
NSLT+Bahdanau (Camgoz et al., 2018)	31.87	19.11	13.16	9.94	31.80	32.24	19.03	12.83	9.58	31.80
NSLT+Luong (Camgoz et al., 2018)	31.58	18.98	13.22	10.00	32.60	29.86	17.52	11.96	9.00	30.70
STN-SLT (Voskou et al., 2021)	49.12	36.29	28.34	23.23	—	48.61	35.97	28.37	23.65	—
CSGCR (Zhao et al., 2022)	35.85	24.77	18.65	15.08	38.96	36.71	25.40	18.86	15.18	38.85
GFSLT-VLP (Zhou et al., 2023)	44.08	33.56	26.74	22.12	43.72	43.71	33.18	26.11	21.44	42.49
FLa-LLM (Chen et al., 2024)	—	—	—	—	—	46.29	35.33	28.03	23.09	45.27
SignLLM (Gong et al., 2024)	46.88	<u>36.59</u>	<u>29.91</u>	<u>25.25</u>	47.23	45.21	34.78	28.05	23.40	44.49
C ² RL (Chen et al., 2025)	—	—	—	—	—	52.81	40.20	32.20	26.75	50.96
MMSLT (Kim et al., 2025)	48.73	37.78	30.51	25.47	48.58	48.92	<u>38.12</u>	<u>30.79</u>	<u>25.73</u>	47.97
TransAlign-SLT (Ours)	48.13	36.04	28.55	22.86	<u>47.63</u>	48.53	36.97	29.39	23.54	<u>48.46</u>

Table 1: Comparison with prior methods on PHOENIX14T. Methods are grouped by supervision setting. B1 to B4 denote BLEU-1 to BLEU-4, and R denotes ROUGE-L. Best and second-best results within the fully gloss-free setting are shown in bold and underlined, respectively.

Method	Dev					Test				
	B1	B2	B3	B4	R	B1	B2	B3	B4	R
Gloss-supervised SLT										
SLRT (Camgoz et al., 2020)	37.47	24.67	16.86	11.88	37.96	37.38	24.36	16.55	11.79	36.74
BN-TIN-Transf.+SignBT (Zhou et al., 2021)	51.46	37.23	27.51	20.80	49.49	51.42	37.26	27.76	21.34	49.31
MMTLB (Chen et al., 2022a)	53.81	40.84	31.29	24.42	53.38	53.31	40.41	30.87	23.92	53.25
Twostream-SLT (Chen et al., 2022b)	55.21	42.31	32.71	25.76	55.10	55.44	42.59	32.87	25.79	55.72
SLTUNET (Zhang et al., 2023)	—	—	—	23.99	53.58	54.98	41.44	31.84	25.01	54.08
Weakly Supervised Gloss-free SLT										
TSPNet* (Li et al., 2020)	—	—	—	—	—	17.09	8.98	5.07	2.97	18.38
GASLT (Yin et al., 2023)	—	—	—	—	—	19.90	9.94	5.98	4.07	20.35
ConSLT (Fu et al., 2023)	—	—	—	14.80	41.46	—	—	—	14.53	40.98
Sign2GPT (Wong et al., 2024)	—	—	—	—	—	41.75	28.73	20.60	15.40	42.36
VAP [◊] (Jiao et al., 2024)	50.41	—	—	21.16	48.72	49.99	—	—	20.85	48.56
Fully Gloss-free SLT										
SLRT-GF [‡] (Camgoz et al., 2020; Zhou et al., 2023)	21.03	9.97	5.96	4.04	20.51	20.00	9.11	4.93	3.03	19.67
NSLT+Luong (Camgoz et al., 2018)	34.22	19.72	12.24	7.96	34.28	34.16	19.57	11.84	7.56	34.54
GFSLT-VLP (Zhou et al., 2023)	39.20	25.02	16.35	11.07	36.70	39.37	24.93	16.26	11.00	36.44
FLa-LLM (Chen et al., 2024)	—	—	—	—	—	37.13	25.12	18.38	14.20	37.25
SignLLM (Gong et al., 2024)	42.45	26.88	17.90	12.23	39.18	39.55	28.13	20.07	15.75	39.91
C ² RL (Chen et al., 2025)	—	—	—	—	—	<u>49.32</u>	<u>36.28</u>	27.54	21.61	48.21
MMSLT (Kim et al., 2025)	50.05	36.39	26.91	20.51	48.53	49.87	36.37	<u>27.29</u>	<u>21.11</u>	48.92
TransAlign-SLT (Ours)	<u>44.59</u>	<u>32.22</u>	<u>22.75</u>	<u>16.17</u>	<u>44.13</u>	44.59	32.21	22.89	16.39	44.68

Table 2: Comparison with prior methods on CSL-Daily (Zhou et al., 2021). Methods are grouped by supervision setting. B1 to B4 denote BLEU-1 to BLEU-4, and R denotes ROUGE-L. * denotes TSPNet results reproduced by prior work (Yin et al., 2023). ‡ denotes the SLRT architecture reproduced under the gloss-free setting by GFSLT-VLP (Zhou et al., 2023). For VAP (Jiao et al., 2024), marked with ◊, we report results without punctuation post-processing following (Min and Chen, 2023). Best and second-best results within the fully gloss-free setting are shown in bold and underlined, respectively.

Reference:	auch morgen erwartet uns eine ruhige herbstmischung aus hochnebel wolken und sonne (tomorrow too we can expect a calm autumnal mix of high fog, clouds and sunshine)
TransAlign-SLT:	morgen erwartet uns eine mischung aus hochnebel wolken und sonne (tomorrow we can expect a mix of high fog, clouds and sunshine)
Reference:	im osten und südosten auch schnee oder schneeregen (in the east and southeast also snow or sleet)
TransAlign-SLT:	im osten und südosten schnee oder schnee (in the east and southeast snow or snow)
Reference:	und nun die wettervorhersage für morgen samstag den fünften dezember (and now the weather forecast for tomorrow, Saturday, 5 December)
TransAlign-SLT:	und nun die wettervorhersage für morgen samstag den zehnten dezember (and now the weather forecast for tomorrow, Saturday, 10 December)

Table 3: Qualitative translations on the PHOENIX14T test set. Correctly preserved content is shown in blue, semantically related but inaccurate content is shown in green, and incorrect content is shown in red.

Temporal Encoder	B4	R
TemporalConv + BiLSTM	18.18	43.09
HMTF w/o BiLSTM	20.14	43.77
HMTF + BiLSTM	22.86	47.63

Table 4: Effect of temporal modeling components on PHOENIX14T Dev. B4 denotes BLEU-4 and R denotes ROUGE-L.

5.2 Semantic Supervision

Table 5 examines the individual components of the training-time semantic branch. The baseline removes BERT-based semantic guidance while keeping the remaining TransAlign-SLT components unchanged. Introducing frozen BERT representations through MOT increases BLEU-4 from 21.54 to 21.85 and ROUGE-L from 45.10 to 45.95.

Adapting BERT to the target translation corpus using masked language modeling provides a further improvement, reaching 22.54 BLEU-4 and 46.90 ROUGE-L. The best performance is obtained when WordPiece subwords are additionally pooled into word-level representations before transport alignment, yielding 22.86 BLEU-4 and 47.63 ROUGE-L. This result suggests that the semantic signal is more useful when the text representations are adapted to the target corpus and aligned at the word level rather than at fragmented subword positions.

Method	MLM Adapt	Word Pool	MOT Align	B1	B2	B3	B4	R
Baseline	✗	✗	✗	47.43	34.00	27.21	21.54	45.10
+ BERT	✗	✗	✓	47.50	34.23	27.88	21.85	45.95
+ MLM adaptation	✓	✗	✓	48.00	35.88	28.10	22.54	46.90
+ Word pooling	✓	✓	✓	48.13	36.04	28.55	22.86	47.63

Table 5: Effect of training-time semantic supervision on PHOENIX14T Dev. B1 to B4 denote BLEU-1 to BLEU-4, and R denotes ROUGE-L.

5.3 Effect of TranslationCTC

Table 6 studies both the strength and the scheduling of TranslationCTC supervision. Removing TranslationCTC ($\lambda = 0$) causes BLEU-4 to fall from 22.86 to 10.50 and ROUGE-L from 47.63 to 27.01, showing that this auxiliary signal is important during representation learning. At the same time, maintaining a strong constant weight is not optimal. A fixed weight of $\lambda = 0.6$ produces 21.50 BLEU-4, while moderate constant weights recover most of the performance.

The proposed $0.5 \rightarrow 0.1$ schedule achieves the best BLEU-2, BLEU-3, and BLEU-4 scores. This behavior supports the role of TranslationCTC as an early temporal-linguistic prior whose influence is gradually reduced as training progresses.

CTC Strategy	B1	B2	B3	B4	R
Constant ($\lambda = 0.0$)	16.58	13.56	11.20	10.50	27.01
Constant ($\lambda = 0.6$)	47.56	35.01	26.32	21.50	47.20
Constant ($\lambda = 0.5$)	48.32	35.93	28.24	22.42	47.46
Constant ($\lambda = 0.4$)	48.50	35.76	28.09	22.40	47.70
Decay ($0.5 \rightarrow 0.1$)	48.13	36.04	28.55	22.86	47.63

Table 6: Effect of TranslationCTC weighting on PHOENIX14T Dev.

Complementarity with MOT. TranslationCTC encourages coarse temporal organization, whereas MOT provides soft semantic correspondence without explicitly enforcing a monotonic frame-to-word alignment. The separate ablations show that both signals are useful. However, we do not interpret these results as a fully controlled factorial decomposition of their interaction.

5.4 Decoding with CTC Rescoring

Table 7 separates the contribution of autoregressive beam search from that of post-hoc TranslationCTC rescoring. Beam search accounts for most of the decoding improvement, raising BLEU-4 from 20.58 with greedy decoding to 22.40.

Applying TranslationCTC rescoring to completed beam hypotheses provides an additional improvement, reaching 22.86 BLEU-4 and 47.63 ROUGE-L. By comparison, CTC rescoring without beam search produces only a small gain over greedy decoding.

CTC scores are used only to rerank completed hypotheses and do not participate in beam expansion. TranslationCTC therefore does not impose a hard monotonic decoding path. Moreover, the relatively small gain from inference-time rescoring, compared with the large performance drop observed when TranslationCTC is removed during training, suggests that its main contribution comes from representation learning rather than from decoding-time reranking.

Decoding	Beam	CTC	B1	B2	B3	B4	R
Greedy	✗	✗	45.84	33.32	26.03	20.58	45.44
CTC rescore only	✗	✓	45.90	33.50	26.20	20.60	45.60
Beam	✓	✗	47.18	35.27	27.97	22.40	46.54
Beam + CTC	✓	✓	48.13	36.04	28.55	22.86	47.63

Table 7: Effect of beam search and post-hoc TranslationCTC rescoring on PHOENIX14T Dev.

5.5 Effect of Decoder Choice

We also examine whether the observed gains can be explained primarily by the choice of a larger pretrained language decoder. An mBART-based decoder variant achieves 20.43 BLEU-4 on PHOENIX14T Dev, compared with 22.86 BLEU-4 for the GRU-Bahdanau decoder used in TransAlign-SLT. This 2.43-point difference indicates that simply increasing decoder scale does not account for the improvements. Strong visual-temporal representations and training-time alignment remain important even when translation is generated with a lightweight recurrent decoder.

Translation quality versus model size. TransAlign-SLT achieves 23.54 test BLEU-4 with 45.6M inference parameters, compared with 25.73 BLEU-4 using 640.94M parameters for MMSLT (Kim et al., 2025) and 22.52 BLEU-4 using approximately 1.77B parameters for Sign2GPT (Wong et al., 2024). This indicates a

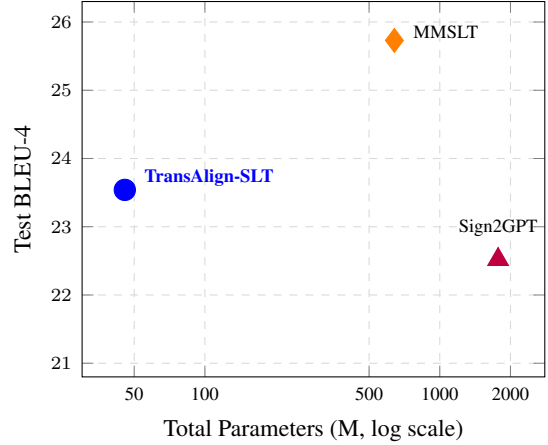


Figure 3: Translation quality versus model size on PHOENIX14T. The x-axis reports total model parameters on a logarithmic scale, and the y-axis reports test BLEU-4. Only methods with directly reported or measured complete model parameter counts are included.

favorable trade-off between translation quality and model size rather than establishing a direct runtime efficiency comparison. Figure 3 visualizes this trade-off on PHOENIX14T.

Limitations and future work. Evaluation is limited to PHOENIX14T and CSL-Daily, so generalization beyond these benchmarks remains untested. We report model size and computational cost but not hardware-matched latency, memory, or training-time comparisons. Future work will consider broader sign languages and domains and controlled efficiency benchmarking.

6 Conclusion

We presented **TransAlign-SLT**, a fully gloss-free SLT framework that learns directly from paired sign videos and spoken-language translations using complementary temporal and semantic supervision. TranslationCTC provides coarse temporal regularization, while MOT encourages semantic alignment between visual features and contextualized word-level representations from a frozen, corpus-adapted BERT encoder. Experiments on PHOENIX14T and CSL-Daily, together with component-level ablations, show that these training signals improve learning from raw translation supervision while preserving a compact GRU-based inference pipeline that does not rely on glosses, pseudo-glosses, or a large pretrained language decoder. The results suggest a practical direction for scalable gloss-free SLT with compact inference pipelines.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.
- 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 (CVPR)*, 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 (CVPR)*, pages 10023–10033.
- 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 (CVPR)*, 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.
- Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. 2025. [C²RL: Content and context representation learning for gloss-free sign language translation and retrieval](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):8533–8544.
- Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. [Factorized learning assisted with large language model for gloss-free sign language translation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7071–7081, Torino, Italia. ELRA and ICCL.
- Marco Cuturi. 2013. [Sinkhorn distances: Lightspeed computation of optimal transport](#). In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300. Curran Associates, Inc.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. [ImageNet: A large-scale hierarchical image database](#). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Biao Fu, Peigen Ye, Liang Zhang, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2023. [A token-level contrastive framework for sign language translation](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372.
- Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. [Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks](#). In *International Conference on Machine Learning*, pages 369–376.
- Alex Graves and Jürgen Schmidhuber. 2005. Frameworkwise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.
- Leming Guo, Wanli Xue, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Dimitris Metaxas. 2024. [Gloss prior guided visual feature learning for continuous sign language recognition](#). *IEEE Transactions on Image Processing*, 33:3486–3495.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Peiqi Jiao, Yuecong Min, and Xilin Chen. 2024. [Visual Alignment Pre-Training for Sign Language Translation](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–367.
- Jungeun Kim, Hyeonwoo Jeon, Jongseong Bae, and Ha Young Kim. 2025. [Leveraging the power of MLLMs for gloss-free sign language translation](#). In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21048–21058.
- Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Ben Swift, Hanna Suominen, and Hongdong Li. 2020. TSPNet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Yuecong Min and Xilin Chen. 2023. [Towards faithful sign language translation](#). Submitted to NeurIPS 2023.

- 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.
- Andreas Voskou, Konstantinos P. Panousis, Dimitrios Kosmopoulos, Dimitris N. Metaxas, and Sotirios Chatzis. 2021. Stochastic transformer networks with linear competing units: Application to end-to-end SL translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955.
- Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, pages 20–36.
- Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. [Sign2GPT: Leveraging large language models for gloss-free sign language translation](#). In *International Conference on Learning Representations*, volume 2024, pages 18157–18174.
- Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2551–2562.
- Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. [SLTUNET: A simple unified model for sign language translation](#). In *The Eleventh International Conference on Learning Representations*.
- Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. 2022. [Conditional sentence generation and cross-modal reranking for sign language translation](#). *IEEE Transactions on Multimedia*, 24:2662–2672.
- 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 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 (CVPR)*, pages 1316–1325.
- Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2022. [Spatial-temporal multi-cue network for sign language recognition and translation](#). *IEEE Transactions on Multimedia*, 24:768–779.