SONAR-SLT: Multilingual Sign Language Translation via Language-Agnostic Sentence Embedding Supervision

Yasser Hamidullah[1] **Shakib Yazdani**[1] **Cennet Oguz**[1]
Josef van Genabith[1] **Cristina España-Bonet**[1,2]

[1]German Research Center for Artificial Intelligence (DFKI GmbH),
Saarland Informatics Campus, Saarbrücken, Germany
[2]Barcelona Supercomputing Center (BSC-CNS), Barcelona, Catalonia, Spain

{yasser.hamidullah,shakib.yazdani,cennet.oguz,josef.van_genabith,cristinae}@dfki.de

Abstract

Sign language translation (SLT) is typically trained with text in a single spoken language, which limits scalability and cross-language generalization. Earlier approaches have replaced gloss supervision with text-based sentence embeddings, but up to now, these remain tied to a specific language and modality. In contrast, here we employ language-agnostic, multimodal embeddings trained on text and speech from multiple languages to supervise SLT, enabling direct multilingual translation. To address data scarcity, we propose a coupled augmentation method that combines multilingual target augmentations (i.e. translations into many languages) with video-level perturbations, improving model robustness. Experiments show consistent BLEURT gains over text-only sentence embedding supervision, with larger improvements in low-resource settings. Our results demonstrate that language-agnostic embedding supervision, combined with coupled augmentation, provides a scalable and semantically robust alternative to traditional SLT training.[1]

1 Introduction

Sign languages (SLs) are inherently visual and culturally embedded. Each SL has evolved independently and is closely tied to the communities and spoken languages of its region. As a result, most sign language translation (SLT) datasets are built around a *single* sign–spoken language pair (e.g., DGS→German), which makes it difficult to scale models across languages or to combine datasets. Training a system for a new target language typically requires a separate model and fresh parallel data collection.

Historically, SLT systems have relied on manually provided *gloss supervision* (Camgoz et al., 2018), discrete word-like labels whose design and availability are language-, culture-, and region-specific. Even *gloss-free* SLT approaches assume that sign inputs should be supervised by text from the co-occurring spoken language, keeping the learning signal tied to a single language (Gong et al., 2024; Wong et al., 2024; Chen et al., 2024; Hamidullah et al., 2022) and limiting cross-dataset reuse and generalization.

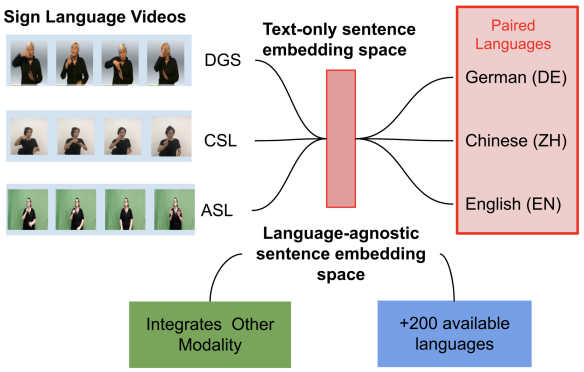

Figure 1: Text-only vs. language-agnostic sentence embedding supervision.

Recent work by Hamidullah et al. (2024) has reduced the reliance on glosses by supervising SLT with *text-based sentence embeddings*. This yields better semantic alignment, but the embeddings remain modality-specific and typically require dataset-specific fine-tuning. Furthermore, compared to large pre-trained models that exploit vast text corpora, these text-only embeddings show limited cross-lingual transfer and reduced robustness. This raises the key question: *Can language-agnostic, multimodal sentence embedding supervision replace text-only alignment in SLT?* We hypothesize that **language-agnostic, multimodal sentence embeddings** can reduce the residual dependence on text. Concretely, we build on SONAR (Duquenne et al., 2023), a pretrained multilingual and multimodal embedding space that jointly rep-

[1]We release the code, models, and features to facilitate further research. Github repository: https://github.com/DFKI-SignLanguage/sonar-slt.git; Huggingface: https://huggingface.co/mtmlt

Proceedings of the Tenth Conference on Machine Translation, pages 301–313

resents text and speech. SONAR embeddings are claimed to be language-agnostic. Our approach aligns sign representations directly with language-agnostic semantic vectors, thereby decoupling supervision from any specific spoken language and removing the need for glosses. Our model integrates multiple modalities and supports direct supervision across all 200 languages covered by SONAR (see Figure 1). In contrast to prior systems that relied on additional stages or separate models for multi-target translation, our method enables *direct* translation into multiple languages within a single model.

A major obstacle for SLT is the scarcity of annotated data. Recent work on self-supervised pre-training from unannotated or anonymized data (Rust et al., 2024) has shown promise in addressing this challenge. This motivates our second question: *Can target-language augmentation further alleviate data scarcity and enhance robustness, particularly when combined with video augmentation?*

Our **coupled multiple target language and video perturbation augmentation** strategy addresses these challenges by combining (i) *target-language augmentation*, which pairs each sign sample with parallel sentences in multiple languages, and (ii) *video augmentation*, which perturbs the visual stream through spatial, temporal, and photometric transformations. These augmentations are complementary: multiple target-language augmentation strengthens semantic supervision without requiring new sign recordings, while video augmentation improves the invariance of the sign encoder. Together, they yield a more robust SLT model and provide a scalable, semantically grounded alternative to traditional training, unifying supervision across languages and modalities while reducing dependence on language-, culture-, and region-specific annotations. In all, our contributions can be summarized as:

- **Language-agnostic supervision.** We align signs to a multilingual, multimodal embedding space, removing reliance on language-specific text or glosses.
- **Coupled augmentation.** We jointly apply multilingual target augmentation and video perturbations to improve robustness and reduce data scarcity.
- **Direct multilingual decoding.** Our model translates into multiple spoken languages in a single step, without pivots or extra fine-tuning.
- **Open-source resources.** We release a Hugging Face–compatible *visual extension of* SONAR and model port to enable reproducibility and further work.

2 Related Work

2.1 Sign Language Representation

Traditional SLT systems rely on glosses —textual labels that represent signs— as an intermediate representation. MSKA-SLT (Guan et al., 2025) remains a strong baseline using glosses, reporting ∼29 BLEU on PHOENIX-2014T (Camgoz et al., 2018). However, glosses are neither universal nor standardized: they are tightly coupled to specific languages, cultures, and regions. Moreover, producing gloss annotations is highly time-consuming, requiring expert linguistic knowledge (Müller et al., 2023b).

In parallel, gloss-free SLT has emerged, enabling training on weakly annotated datasets exceeding 1,000 hours for some sign languages (Uthus et al., 2023).[2] Hamidullah et al. (2024) aligns sign language videos with sentence-level text embeddings. This supervision avoids feeding long, fine-grained frame sequences to the decoder, thereby reducing redundancy in video features, lowering the need for aggressive masking, and encouraging learning at the sentence-semantic level. While intermediate supervision of visual blocks is common in multimodal models, compressing video into a sentence-level embedding before decoding improves semantic grounding and flexibility in target text generation. Nevertheless, current approaches (Hamidullah et al., 2024; Gueuwou et al., 2025b) remain limited by their reliance on *text-only* embedding spaces with restricted language coverage, constraining augmentation and cross-sign transfer.

2.2 Large Language Models in SLT

Complementary approaches leverage large language models (LLMs). SignLLM (Gong et al., 2024) discretizes videos into tokens and prompts a frozen LLM; Sign2GPT (Wong et al., 2024) feeds pseudo-glosses to XGLM, reporting ∼22 BLEU on

[2] A *weakly annotated dataset* provides only coarse or noisy supervision. For instance, YouTube-ASL datasets are collected from online videos where annotations rely solely on automatically generated or the provided subtitles, without manual realignment, leading to potential inaccuracies and temporal misalignments.

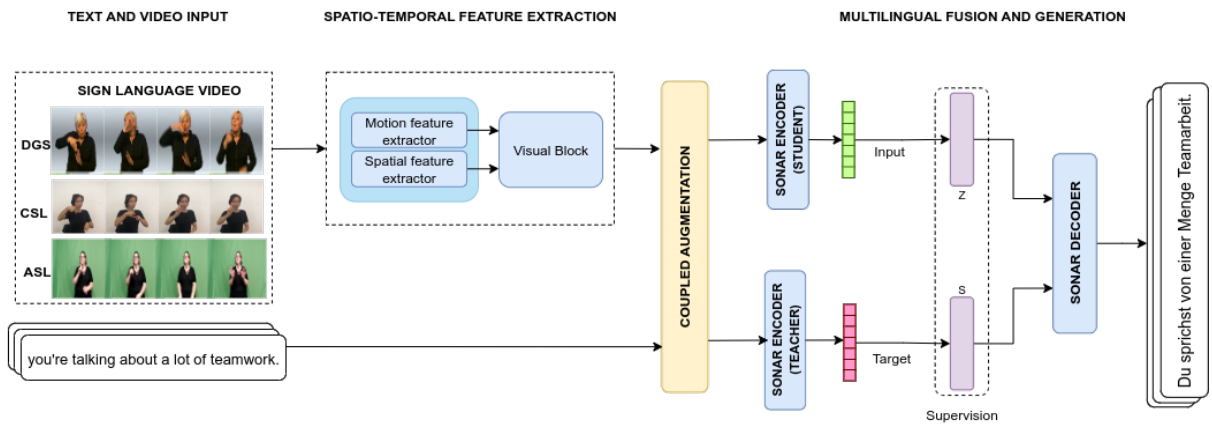

Figure 2: Overall architecture of our SONAR-SLT model. Visual inputs are processed through spatial and spatio-temporal encoders, fused using (Hwang et al., 2025) and encoded into a semantic vector aligned with multilingual sentence embeddings.

PHOENIX-2014T and ∼15 BLEU on CSL-Daily. SpaMo (Hwang et al., 2025) employs a straightforward approach that extracts spatial and motion features from sign language videos and utilizes a low-rank adapter to fine-tune an LLM for sign language translation. Chen et al. (2024) introduced FLa-LLM, a two-stage, gloss-free framework that first pre-trains the visual encoder and then fine-tunes a pre-trained LLM for the downstream SLT task. These methods inherit LLM fluency but are largely monolingual and require substantial tokenization and training overhead. In contrast, our PEFT-based SONAR adapters maintain multilinguality without retraining a large decoder on discretized video tokens. More recent work has explored large-scale pre-training to improve sign language understanding, with Uni-Sign (Li et al., 2025) proposing a unified generative framework that treats downstream tasks as SLT and incorporates prior-guided fusion.

2.3 Multilingual SLT Datasets and Models

Despite these advances, large-scale multilingual datasets (Uthus et al., 2023; Yazdani et al., 2025b) remain scarce and noisy. Crawled web data increases coverage but introduces label and alignment errors that current models struggle to absorb, leading many studies to focus on a single language or a small set of cleaner corpora. Additionally, performance often varies widely even within the same language due to differences in feature pipelines and recording conditions.

Multilingual SLT models also remain in their early stages. MLSLT (Yin et al., 2022) covers ten European sign languages via a routing mechanism, while JWSign (Gueuwou et al., 2023) scales to 98 languages with language-ID tokens. More recently, Sign2(LID+Text) (Tan et al., 2025) incorporated token-level language identification with a CTC loss, achieving competitive results. In addition, Yazdani et al. (2025a) explored continual learning for multilingual SLT. Recent work applies heavy pre-processing (Gueuwou et al., 2025b,a), sometimes obscuring whether improvements arise from better SLT modeling or dataset-specific engineering. Both gloss-based and gloss-free methods perform best when signer distance, camera setup, and motion characteristics closely match training conditions.

3 Methodology

3.1 System Overview

We propose **SONAR-SLT**, a modular SLT framework that decouples *semantic understanding* from *text generation*. As illustrated in Figure 2, the system first maps an input sign language video into a multilingual, multimodal semantic space, and then (optionally) decodes from this space into a chosen spoken language. This design allows training on heterogeneous sign language datasets, supports multilingual supervision, and removes the need for gloss annotations. A detailed architecture is presented in Appendix A.1 and summarized in the next subsections.

3.2 Visual Feature Extraction and Encoding

The first stage maps raw video frames into a compact visual embedding. Let $x = (f_1, \ldots, f_T)$ denote a sign language video of T frames. We extract per-frame spatial features $\mathbf{s}_t$ with ViT (Dosovitskiy

et al., 2020) and spatio-temporal motion features $\mathbf{m}_t$ with VideoMAE (Tong et al., 2022). These are fused through a lightweight block (1D Conv followed by a multi-layer perceptron) $\mathcal{F}$ (Hwang et al., 2025):

$$\mathbf{h}_t = \mathcal{F}(\mathbf{s}_t, \mathbf{m}_t), \qquad t = 1, \ldots, T. \tag{1}$$

A Transformer-based encoder $\mathcal{E}_v$ contextualizes the sequence:

$$\mathbf{z}_{1:T} = \mathcal{E}_v(\mathbf{h}_{1:T}). \tag{2}$$

Finally, temporal pooling (mean or attention) produces a global visual embedding $\mathbf{z} \in \mathbb{R}^d$:

$$\mathbf{z} = \mathrm{Pool}(\mathbf{z}_{1:T}). \tag{3}$$

3.3 Semantic Alignment

Next, we align sign-derived embeddings with multilingual textual embeddings. We adopt a pretrained multilingual, multimodal sentence encoder $\mathcal{E}$ (i.e., SONAR). Given a reference sentence y, we obtain its semantic embedding:

$$\mathbf{s} = \mathcal{E}_{txt}(y) \in \mathbb{R}^d. \tag{4}$$

The visual encoder is trained to align $\mathbf{z}$ with $\mathbf{s}$. Alignment can be done via a squared ℓ_2 loss as per (Duquenne et al., 2023; Hamidullah et al., 2024):

$$\mathcal{L}_{\text{sem}} = \|\mathbf{z} - \mathbf{s}\|_2^2. \tag{5}$$

We also consider a cosine similarity loss,

$$\mathcal{L}_{\cos} = 1 - \frac{\langle \mathbf{z}, \mathbf{s} \rangle}{\|\mathbf{z}\|_2 \, \|\mathbf{s}\|_2}, \tag{6}$$

used either alone ($\mathcal{L}_{\text{sem}} = \mathcal{L}_{\cos}$) or combined with the MSE above:

$$\mathcal{L}_{\text{sem}} = \alpha \|\mathbf{z} - \mathbf{s}\|_2^2 + \beta \mathcal{L}_{\cos}, \qquad \alpha, \beta \geq 0. \tag{7}$$

Target-language augmentation. To enforce language-agnostic supervision, each reference sentence is paired with K translations $\{y^{(k)}\}_{k=1}^K$ (from the embedding decoder). At each iteration, one translation $y^{(k)}$ is sampled and encoded as $\mathbf{s} = \mathcal{E}_{txt}(y^{(k)})$.

3.4 Multilingual Generation from the Semantic Vector

We then decode into natural language from the semantic embedding. A pretrained decoder $\mathcal{D}$ from SONAR generates text from a semantic vector and a target language token ℓ. Conditioned on the sign-derived and semantically text-aligned (Section 3.3) embedding $\mathbf{z}$, the decoder is trained with teacher forcing:

$$\mathcal{L}_{\text{ce}} = -\sum_{t=1}^{T_y} \log p_\theta(y_t \mid y_{<t}, \mathbf{z}, \ell). \tag{8}$$

3.5 Auto-Encoding (Decoder Anchoring)

To keep the decoder aligned to the pretrained semantic space, we introduce an auto-encoding step. Specifically, the decoder reconstructs the target sentence directly from its text-derived embedding $\mathbf{s}$:

$$\mathcal{L}_{\text{ae}} = -\sum_{t=1}^{T_y} \log p_\theta(y_t \mid y_{<t}, \mathbf{s}, \ell). \tag{9}$$

This mirrors SONAR's original training and prevents drift, while the visual encoder learns to project videos into the same space.

3.6 Optional Contrastive Alignment

We optionally strengthen alignment through a symmetric InfoNCE loss (van den Oord et al., 2018). For a batch $\{(\mathbf{z}_i, \mathbf{s}_i)\}_{i=1}^N$, we define similarity as $\mathrm{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\tau}$ (temperature $\tau > 0$, optional ℓ_2 normalization). The corresponding loss is:

$$\mathcal{L}_{\text{nce}} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{\exp(\hat{\mathbf{z}}_i^\top \hat{\mathbf{s}}_i / \tau)}{\sum_{j=1}^N \exp(\hat{\mathbf{z}}_i^\top \hat{\mathbf{s}}_j / \tau)} - \log \frac{\exp(\hat{\mathbf{s}}_i^\top \hat{\mathbf{z}}_i / \tau)}{\sum_{j=1}^N \exp(\hat{\mathbf{s}}_i^\top \hat{\mathbf{z}}_j / \tau)} \right]. \tag{10}$$

3.7 Joint Training Objective

The final loss combines all components:

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{sem}} \mathcal{L}_{\text{sem}} + \lambda_{\text{ce}} \mathcal{L}_{\text{ce}} + \lambda_{\text{ae}} \mathcal{L}_{\text{ae}} + \lambda_{\text{nce}} \mathcal{L}_{\text{nce}}, \tag{11}$$

with non-negative weights λ_i (setting $\lambda_{\text{nce}} = 0$ disables the contrastive term).

3.8 Cross-Lingual and Multi-Sign Dataset Fusion

Finally, we leverage the language-agnostic semantic space for dataset fusion. Because supervision is defined independently of any specific spoken language, videos from different sign languages can be trained jointly with textual supervision in *any* available language. For example, German sign language

Dataset	Language	Domain	#Videos	#Sent.	Vocab.	Split (train/dev/test)
PHOENIX-2014T	DGS → German	Weather Forecast	∼7k	∼8k	∼3k	7,096 / 519 / 642
CSL-Daily	CSL → Chinese	Daily Communication	∼20k	∼25k	∼5k	18,401 / 1,078 / 1,057

Table 1: Characteristics of the datasets used in our experiments.

videos annotated in German can be re-aligned with English, French, or Chinese translations via $\mathcal{E}$, allowing unified training across datasets such as PHOENIX-2014T and CSL-Daily. This enables direct multi-target translation without glosses and facilitates fusion of heterogeneous sign-language corpora.

4 Experiments

4.1 Datasets

We evaluate our approach on the following datasets:

- **PHOENIX-2014T** (Camgoz et al., 2018): German Sign Language (DGS) weather forecast videos with parallel German text.
- **CSL-Daily** (Zhou et al., 2021): A Chinese Sign Language (CSL) corpus tailored for sign-to-Chinese SLT, emphasizing interactions in daily communication contexts.

Statistics of both datasets are summarizes in Table 1.

4.2 Evaluation Metrics

We evaluate our method following (Müller et al., 2022; Müller et al., 2023a), using BLEU[3] (via SacreBLEU (Post, 2018)) for lexical overlap, ROUGE (Lin, 2004)[4] for recall-oriented n-gram overlap, and BLEURT (Sellam et al., 2020)[5] for semantic quality.

4.3 State-of-the-art Systems

We evaluate our method against several strong recent state-of-the-art systems within the gloss-free paradigm. CSGCR (Zhao et al., 2021) improves SLT accuracy and fluency through three modules: word existence verification, conditional sentence generation, and cross-modal re-ranking for richer grammatical representations. GFSLT-VLP (Zhou et al., 2023) leverages vision–language pretraining, while FLa-LLM (Chen et al., 2024) adopts a two-stage gloss-free pipeline that first pre-trains the visual encoder and then fine-tunes a pre-trained LLM for SLT. Sign2GPT (Wong et al., 2024) maps visual inputs to pseudo-gloss sequences and decodes them with GPT-style language modeling, whereas SignLLM (Gong et al., 2024) discretizes sign features into visual tokens to prompt a frozen LLM. SEM-SLT (Hamidullah et al., 2024) aligns sign language videos with sentence embeddings and serves as the foundation of our work. For multilingual settings, Sign2(LID+Text) (Tan et al., 2025) combines token-level sign language identification with a CTC objective to generate spoken text.

4.4 Implementation Details

• **Feature Extraction.** We begin by processing each sign language video $x = \{f_1, f_2, \ldots, f_T\}$ as a sequence of T RGB frames. From each frame, we extract:

- **Spatial Features (s_t):** Using a Vision Transformer (ViT (Dosovitskiy et al., 2020)) pretrained on ImageNet.
- **Motion Features (m_t):** Using VideoMAE (Tong et al., 2022).

These features are then fused via the visual fusion block $\mathcal{F}$ from SpaMo to yield a joint representation h_t for each timestep.

• **Training the visual block (LoRA).** We train the visual block using LoRA with:

- **LoRA:** $r = 16,\ \alpha = 32$
- **Batching:** batch size 4, gradient accumulation 2, on 8 GPUs in parallel
- **Loss weights:**
  - $\lambda_{\text{ce}} = 0.1$ (auxiliary soft translation signal)
  - $\lambda_{\text{sem}} = 1.0$ (primary objective)
  - $\lambda_{\cos} = 2.7$ (stabilizes angular alignment)
  - $\lambda_{\text{nce}} = 0.0$
  - $\lambda_{\text{mse}} = 7000.0$ (strong magnitude regularizer)

[3] `BLEU|nrefs:1|bs:1000|seed:16|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0`

[4] `ROUGE|L|nrefs:1|tok:13a|case:mixed|version:1.5.5`

[5] BLEURT v0.0.2 using checkpoint BLEURT-20.

Because our model operates on embedding vectors with small magnitudes, the MSE loss can rapidly fall to $\sim 10^{-5}$ even when cosine similarity remains suboptimal. Empirically, we observed that **cosine and MSE only begin to correlate at** $\sim 10^{-6}$. Optimizing cosine alone often stalls, as MSE ceases to decrease, while optimizing MSE alone improves fidelity but does not guarantee angular alignment. To address this, we up-weight MSE to maintain shrinkage and retain a non-negligible cosine term to enforce directional consistency. We also experimented with InfoNCE, but under our effective batch size (with few hard negatives) it led to slower convergence and negligible improvements and we do not use it in our final experiments.

• **Sentence embedding pooling.** The original SONAR pools by running a shallow decoder: it feeds a special token (the EOS id in `M200M100`) as input and uses the encoder outputs as hidden states; the first decoder output is taken as the sentence embedding. During the Visual Block training, we adopt this approach with a shallow decoder initialized from the first three SONAR decoder layers and train it only for pooling. This supplies language context during pooling, while the incoming features themselves are language-agnostic (from another modality). Text generation is then conditioned on the target language.

• **Visual representation.** We adopt the best-performing visual representation strategies reported in prior work, noting that optimal choices vary across datasets. To ensure comparability in our multi–sign language experiments, we restrict evaluation to datasets with similar video settings and select the strongest corresponding model. The SpaMo Visual Block performs best with global, high-quality cues e.g., high-resolution videos with a moderate signer–camera distance (CSL-Daily), or lower-resolution videos where the signer is close and centered (PHOENIX-2014T). Consequently, we conduct multilingual experiments on CSL-Daily and PHOENIX-2014T.

• **Training the translation model with visual features.** We train the end-to-end translation system (with the Visual Block or the fused spatial+motion features) using the **same LoRA configuration** as above.

- **Batching & schedule:** batch size 8 on a single GPU
  - **CSL-Daily:** cosine learning-rate schedule with a peak LR of 3×10^{-4}
  - **PHOENIX-2014-T (monolingual):** constant LR (we found it more stable)
- **Text augmentation.** To expand the datasets using NLLB (NLLB Team, 2024), we machine-translate the target texts into three high-resource languages (English, French and Spanish) using the `facebook/nllb-200-distilled-600M` model.
- **Video augmentation.** Coupled with the target-language augmentation, we also perturb the input videos so that each training instance is presented with both linguistic and visual variability. At each iteration, one augmented variant is sampled. In this work we restrict ourselves to:
  - `frame_mask_ratio` $= 0.2$
  - `frame_dropout_prob` $= 0.2$
  - `add_noise_std` $= 0.04$
  - `shuffle_window` $= 3$

5 Results and Analysis

5.1 Comparative Analysis

We compare our approach with other gloss-free methods on both PHOENIX-2014T and CSL-Daily datasets in Table 2. Our method shows a clear advantage on the semantics-oriented BLEURT metric. It reaches a **BLEURT of 0.545**, outperforming the sentence-based supervision model using text-only sentence embedding (SEM-SLT). BLEURT uses a BERT-based scorer and is designed to capture meaning and fluency, unlike BLEU and ROUGE, which primarily measure n-gram overlap. Moreover, our model outperforms previous monolingual and multilingual systems on CSL-Daily in terms of BLEU and achieves comparable results on PHOENIX-2014T.

• **Observed gaps.** We observe a decrease in BLEU compared to the SEM-SLT system, which is expected since our model is not fine-tuned on sign-language text. Our language-agnostic, sentence embedding-based supervision preserves semantics without requiring fine-tuning on specific dataset: it goes beyond surface n-gram matching to produce translations that are contextually accurate, grammatically correct, and cross-lingually robust. Part

Method	PHOENIX-2014T			CSL-Daily		
	BLEU	BLEURT	RG	BLEU	BLEURT	RG
Monolingual						
CSGCR (Zhao et al., 2021)	15.18	–	38.85	–	–	–
GFSLT-VLP (Zhou et al., 2023)	21.44	–	42.29	11.00	–	36.44
FLa-LLM (Chen et al., 2024)	23.09	–	45.27	14.20	–	37.25
Sign2GPT (Wong et al., 2024)	22.52	–	48.90	15.40	–	**42.36**
SignLLM (Gong et al., 2024)	23.40	–	44.49	15.75	–	39.91
SEM-SLT (Hamidullah et al., 2024)	24.10	0.481	–	–	–	–
Multilingual						
Sign2(LID+Text) (Tan et al., 2025)	**24.23**	–	**50.60**	14.18	–	40.00
SONAR-SLT (Ours)	22.01	**0.545**	41.44	16.23	**0.561**	42.29

Table 2: Comparison of SONAR-SLT with other gloss-free models on PHOENIX-2014T and CSL-Daily (metrics: BLEU, BLEURT, ROUGE (RG)). Unreported metrics are left blank; SONAR-SLT sets the best reported BLEURT on PHOENIX-2014T and remains strongly competitive with several LLM-based baselines on both datasets.

Resource	Language	BLEU
High	Spanish (es)	22.3
	French (fr)	22.6
	English (en)	21.6
Low	Turkish (tr)	13.1
	Malagasy (mg)	11.8
	Persian (fa)	8.7

Table 3: SONAR-SLT performance across target languages in both high- and low-resource settings on PHOENIX-2014T, reported using BLEU scores.

of the remaining gap stems from dataset capture conditions. Our feature extractor (Hwang et al., 2025) is tuned for global cues and can be less accurate in cases where fine-grained articulations, such as facial expressions and finger movements, are critical. Recent top systems address this with keypoint-based representations and extensive preprocessing (Gueuwou et al., 2025b), which help preserve these fine-grained details.

• **Multilingual and multi–sign language.** We evaluate target-side augmentation, where language translations are included in training. Results for both low- and high-resource languages on PHOENIX-2014T are presented in Table 3. In our experiments, we augmented the target set in training with three high-resource languages—French, Spanish, and English—while the model was evaluated on other unseen languages. Using this augmented target set yields a modest improvement over training with a single target language. However, we observe a gap in performance between high- and low-resource languages, which primarily stems from lower reference translation quality in the low-resource languages.[6] The narrow domain of PHOENIX-2014T can also introduce dataset-specific idiosyncrasies, complicating fair comparisons.

Table 4 shows that pre-training on concatenated multi-sign corpora followed by monolingual fine-tuning proves most effective. In contrast, joint multi-sign fine-tuning risks resembling another full training run without yielding substantial gains. In our experiments, we first pre-train on the combined data and then fine-tune monolingually, consistent with (Hamidullah et al., 2024); post-fine-tuning performance remains largely unchanged (see Table 4, mono vs. multilingual setup). Differences in dataset capture conditions still matter—for example, methods that rely solely on global visual features can underperform when fine-grained articulations, such as hand or facial details, are crucial. Pipelines that integrate keypoints with extensive preprocessing (Gueuwou et al., 2025b) help mitigate such losses and achieve stronger results.

5.2 Multitask Learning Effect on the Visual Block

The effect of sentence-embedding supervision is strongest when the Visual Block is still learning feature representations. Once the block has converged—or is pretrained—the additional impact of cosine or MSE objectives diminishes. This occurs because cross-entropy loss often remains rel-

[6] Reference translations were obtained using `facebook/nllb-200-distilled-600M` model.

Type	Variant	PHOENIX-2014T			CSL-Daily		
		BLEURT	BLEU	RG	BLEURT	BLEU	RG
Multi	VB pretrained	**0.523**	21.52	41.10	**0.561**	**16.23**	**42.29**
	VB scratch	0.508	21.38	**42.03**	0.472	14.68	42.12
	VB frozen	0.516	**21.56**	41.39	0.549	16.06	41.95
Mono	VB pretrained	**0.545**	**22.01**	40.52	**0.558**	**16.07**	**42.13**
	VB scratch	0.490	19.79	39.95	0.447	14.14	40.59
	VB frozen	0.520	21.56	**41.44**	0.529	15.70	41.79

Table 4: SONAR-SLT results for the Visual Block (VB) variants under Multilingual and Monolingual settings on PHOENIX-2014T and CSL-Daily. Metrics include BLEURT, BLEU, and ROUGE (RG); best scores per dataset/metric are in bold.

atively high (above 2–3), while MSE rapidly falls to $\mathcal{O}(10^{-5})$ and cosine similarity saturates around ~ 0.3.

In contrast, introducing the auto-encoding loss provides a second cross-entropy signal, which exerts a stronger influence on the Visual Block. Here, intermediate supervision continues to be beneficial, and the auto-encoding objective itself accelerates convergence. We consistently observed this effect in CSL-Daily and in the augmented translation setup on PHOENIX-2014T.

5.3 Qualitative and Semantic Error Analysis

• **Qualitative analysis.** Table 5 shows examples of two contrasting outcomes: cases where the model accurately captures the intended meaning and cases where it fails. When contextual understanding is incomplete, the decoder frequently compensates by generating fluent continuations via next-token prediction. This behavior is characteristic of SLT systems that rely on pretrained language models as decoders: they can mask weaknesses in semantic grounding by producing outputs that are coherent but only partially faithful to the source. As a result, improvements in BLEU may reflect the decoder's ability to recover plausible sentences rather than true gains in sign-to-text comprehension. Therefore, exact sequence matching metrics such as BLEU are insufficient and in some cases misleading for evaluating translation quality in SLT.

Language-specific tendencies. We deep into the analysis of two languages: German, a language trained with original data and French, a language trained via machine translation augmentation.

- **German:** Errors often arise from compound nouns, flexible word order, and embedded clauses, leading to partial omissions, attribute reordering, or unnatural compounds. When alignment is uncertain, the model may insert generic stock phrases or repetitions.

- **French:** Our analysis shows more frequent noun substitutions, agreement mismatches, and text modality shifts (e.g., hedging with *"sont possibles"*). Register differences from determiners or prepositions are also common. Incorrect date and numeric substitutions occur more frequently than in German, likely due to segmentation differences in temporal expressions.

• **Semantic analysis.** Surface-form scores vs. meaning preservation. We observe a systematic mismatch between surface-form metrics (e.g., BLEU) and semantic adequacy (BLEURT) across both German and French. Outputs with only moderate n-gram overlap can still be semantically faithful, while some high-scoring predictions contain factual errors.

Semantically near correct and correct paraphrases (German). As illustrated by the green-highlighted examples in Table 5, incorrect lexical or numeric substitutions leave most of the remaining meaning intact (e.g., date shifts: *"Sonntag, den neunzehnten Dezember"* → *"Sonntag, den siebzehnten August"*; temperature adjustments: *"sechs Grad an den Alpen"* → *"neun Grad am Alpenrand"*). We also observe benign stylistic reformulations (*"es gelten entsprechende Warnungen"* → *"es bestehen Unwetterwarnungen"*) and word-order changes without semantic effect (*"aus Südwest bis West"* → *"aus West bis Südost"*).

Semantically near correct and correct paraphrases (French). Similarly, the first green-highlighted examples show structural or modality shifts that preserve much of the remaining meaning,

German (DGS weather domain)
Ref (DE): Sonntag, den neunzehnten Dezember. *EN: Sunday, the nineteenth of December.* **Pred (DE):** Sonntag, den siebzehnten August. *EN: Sunday, the seventeenth of August.*
Ref (DE): sechs Grad an den Alpen. *EN: Six degrees in the Alps.* **Pred (DE):** neun Grad am Alpenrand. *EN: Nine degrees on the edge of the Alps.*
Ref (DE): Höhenlagen Süddeutschlands. *EN: High-altitude areas of southern Germany.* **Pred (DE):** Küsten. *EN: Coasts.*
French (weather domain)
Ref (FR): vingt-huit août. *EN: Twenty-eighth of August.* **Pred (FR):** vingt-cinq novembre. *EN: Twenty-fifth of November.*
Ref (FR): Des rafales orageuses de l'ouest. *EN: Stormy gusts from the west.* **Pred (FR):** Des rafales orageuses sont possibles. *EN: Stormy gusts are possible.*
Ref (FR): Risque d'inondation. *EN: Risk of flooding.* **Pred (FR):** Avertissements météorologiques violents. *EN: Severe weather warnings.*

Table 5: German and French examples —two semantically near correct paraphrases (green) and one semantically incorrect output (red), with English translations.

such as date substitutions (*"vingt-huit août"* → *"vingt-cinq novembre"*), modality changes (*"Des rafales orageuses de l'ouest"* → *"Des rafales orageuses sont possibles"*), or expanded phrasing (*"Également orages sur la mer du Nord"* → *"Il y a également des orages sur la mer du Nord"*).

Semantically incorrect outputs, true errors (French and German). The red-highlighted rows in Table 5 illustrate errors such as topic drift (predicting wind instead of temperature), incorrect locations (*"Höhenlagen Süddeutschlands"* → *"Küsten"*; *"sud-est"* → *"nord"*), system inversions (*"Hoch"* ↔ *"Tief"*; *"haut"* ↔ *"profonde"*), hallucinated entities, or incorrect hazard categories (*"Risque d'inondation"* → *"Avertissements météorologiques violents"*).

Overall, as in other machine translation tasks, n-gram metrics penalize near or even fully legitimate paraphrases and sometimes fail to capture serious factual errors. Robust SLT evaluation requires *semantic* metrics that explicitly reward meaning preservation while penalizing distortions or hallucinations.

Implications. Evaluation and model development for multilingual SLT should be language-aware. In practice, one should combine semantics-focused metrics with targeted, language-specific checks (e.g., temporals and agreement in French; word order and compounding in German) to obtain fair comparisons and actionable diagnostics.

6 Conclusion

We presented a scalable SLT framework that breaks the traditional close dependency between sign and spoken languages in training data and system development. By aligning sign language videos with multilingual, multimodal sentence embeddings from SONAR, our approach yields a language-agnostic semantic representation that generalizes across both sign languages and spoken targets. This reduces reliance on language-model priors and prioritizes visual grounding and SLT-specific grammar over surface-level text patterns.

Experiments show that language-agnostic supervision enables robust translation even under sign–target mismatches. Multilingual text augmentations, combined with visual augmentation, improves performance on PHOENIX-2014T despite limited data. Ablations further confirm the advantages of this approach in preserving semantic adequacy.

Current evaluation practices often emphasize surface overlap rather than meaning. Future work should develop metrics aligned with semantic similarity and extend supervision to low-resource sign languages and continuous signing in the wild.

Limitations

Our main limitation lies in the visual feature extractor rather than the model architecture itself. We used a pre-existing visual block to avoid evaluation bias, which restricted us to datasets with compatible video settings (CSL-Daily and PHOENIX-2014T) and excluded larger corpora such as How2Sign or YouTube-ASL. As a result, our approach focuses on preserving semantics rather than maximizing exact sentence matches.

Machine translation for data augmentation might induce unintended cultural mistakes that go beyond literal translation. The evaluation on non-human translated datasets also limits the strength of the conclusions for low-resourced languages.

Acknowledgments

This work was partially funded by the German ministry for education and research (BMBF) through projects BIGEKO (grant number 16SV9093) and TRAILS (grant number 01IW24005).

References

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

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *ArXiv*.

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.

Mo Guan, Yan Wang, Guangkun Ma, Jiarui Liu, and Mingzu Sun. 2025. Mska: Multi-stream keypoint attention network for sign language recognition and translation. *Pattern Recogn.*, 165(C).

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025a. SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521, Vienna, Austria. Association for Computational Linguistics.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025b. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria. Association for Computational Linguistics.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9907–9927, Singapore. Association for Computational Linguistics.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2022. Spatio-temporal sign language representation and translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 977–982.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. *Preprint*, arXiv:2501.15187.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023a. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023b. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden,

Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2022. Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 744–772, Abu Dhabi. Association for Computational Linguistics.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sihan Tan, Taro Miyazaki, and Kazuhiro Nakadai. 2025. Multilingual gloss-free sign language translation: Towards building a sign language foundation model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 553–561, Vienna, Austria. Association for Computational Linguistics.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-ASL: A large-scale, open-domain american sign language-english parallel corpus. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*.

Shakib Yazdani, Josef Van Genabith, and Cristina España-Bonet. 2025a. Continual learning in multilingual sign language translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10923–10938, Albuquerque, New Mexico. Association for Computational Linguistics.

Shakib Yazdani, Yasser Hamidullah, Cristina España-Bonet, and Josef van Genabith. 2025b. Seeing, signing, and saying: A vision-language model-assisted pipeline for sign language data acquisition and curation from social media. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 1374–1384, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5109.

Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. 2021. 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 *Proceedings of the 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.

A Appendix

A.1 Detailed Architecture

Figure 3 shows the details of our system architecture as explained in Section 3.

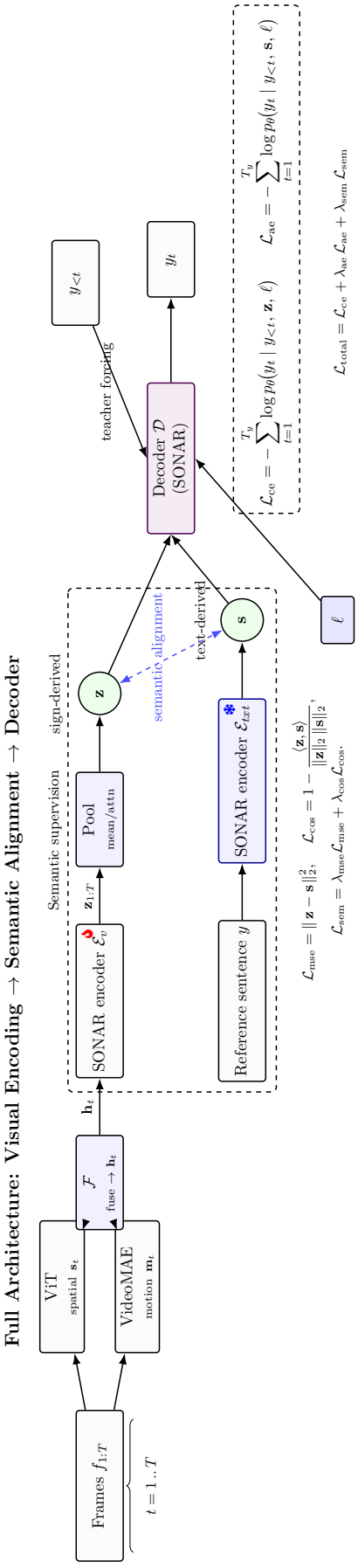

Figure 3: Detailed architecture without the contrastive term (NCE loss).

A.2 Porting SONAR from NLLB Fairseq to Huggingface.

SONAR is officially supported in fairseq, but only its text encoder is available on Hugging Face. To enable full conditional generation, we ported both the encoder and decoder weights from the original SONAR checkpoints into `M200M100`, extending the earlier encoder-only port provided by the NLLB team. In particular, we transferred the decoder weights directly from fairseq, validated their functionality, and released the complete model for public use. The resulting `M200M100ForConditionalGeneration` can now be loaded end-to-end and fine-tuned directly.

A.3 Additional Qualitative Results

Additional translation examples for CSL-Daily and PHOENIX-2014T are provided in Tables 6 and 7, respectively.

Text
Ref (ZH): 我们下午三点见面。 *EN: We will meet at three in the afternoon.* **Pred (ZH):** 我们三点钟下午见。 *EN: We meet at three o'clock in the afternoon.*
Ref (ZH): 我早上吃面包和牛奶。 *EN: I eat bread and milk in the morning.* **Pred (ZH):** 我早上吃了牛奶和面包。 *EN: I had milk and bread in the morning.*
Ref (ZH): 我们乘坐飞机去旅游，今天在酒店住宿。 *EN: We took a plane to travel, and are staying in a hotel today.* **Pred (ZH):** 我们飞机去上海，今天喝酒睡觉。 *EN: We took a plane to Shanghai, today we drink alcohol and sleep.*

Table 6: CSL-Daily examples —good translations (green) and one bad translation (red), showing reference and prediction in Chinese, with English translations for clarity.

Text
Ref (DE): ich wünsche ihnen noch einen schönen abend. *EN: I wish you a pleasant evening.* **Pred (DE):** und jetzt wünsche ich ihnen noch einen schönen abend. *EN: And now, I wish you a pleasant evening.* **Pred (FR):** Et maintenant, je vous souhaite une bonne soirée. *EN: And now, I wish you a good evening.*
Ref (DE): der wind aus süd bis west weht schwach bis mäßig. *EN: The wind from the south to west blows weakly to moderately.* **Pred (DE):** der wind weht meist schwach aus süd bis west. *EN: The wind generally blows weakly from south to west.* **Pred (FR):** Le vent souffle généralement faiblement du sud-ouest. *EN: The wind generally blows weakly from the southwest.*
Ref (DE): in deutschland gibt es nur schwache luftdruckunterschiede. *EN: In Germany, there are only slight air pressure differences.* **Pred (DE):** im nordosten deutschlands sorgt das hoch für wenig unbeständiges wetter. *EN: In northeastern Germany, the high pressure system causes little unsettled weather.* **Pred (FR):** Dans certaines régions de l'Allemagne, la pression atmosphérique élevée n'est toujours pas atteinte. *EN: In some regions of Germany, high atmospheric pressure has still not been reached.*

Table 7: PHOENIX-2014T examples —two good translations (green) and one bad translation (red), showing reference (German), predictions (German and French), and English translations.