

FiMA: Redundancy-Aware Description Enhancement for Gloss-Free Sign Language Translation

Peidong Liu¹ Lianyu Hu^{2*} Wei Feng¹

¹ College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

² College of Computing and Data Science,
Nanyang Technological University, Singapore 639798

peidongliu_2023@tju.edu.cn lianyu.hu@ntu.edu.sg wfeng@ieee.org

Abstract

Gloss-free Sign Language Translation (SLT) avoids costly gloss annotations but remains limited by weak visual-text alignment. Recent MLLM-based methods use frame descriptions as auxiliary supervision, yet these descriptions are often repetitive, producing nearly identical text features; meanwhile, global pooling washes out sparse but meaningful gestures. We propose **FiMA** (Fine-grained Multimodal Alignment), a pre-training framework for redundancy-aware description enhancement. FiMA combines a Description Enhancement Module, which removes shared template redundancy from description features, with a Multimodal Alignment Module, which aligns each text token to its most relevant visual frame through Max-Sim matching. On PHOENIX14T and CSL-Daily, FiMA achieves state-of-the-art gloss-free SLT performance and visualizations show clearer feature separation and temporal grounding. The code will be released at <https://github.com/Ryeliu/FiMA>.

1 Introduction

Sign Language Translation (SLT) aims to automatically translate continuous sign videos into natural spoken sentences, serving as a vital bridge between the Deaf community and the hearing majority. Because sign languages possess unique grammatical structures and word orders that differ significantly from spoken languages (Stokoe Jr, 2005; Bragg et al., 2019), direct translation is highly challenging. To bridge this gap, early approaches (Camgoz et al., 2020; Voskou et al., 2021; Hu et al., 2023; Chen et al., 2022b; Fu et al., 2023; Hu et al., 2022) traditionally rely on “glosses” – atomic textual labels that directly correspond to individual sign gestures in their original chronological order. Consequently, these methods predominantly adopt

a Sign-to-Gloss-to-Text pipeline, where glosses serve as an intermediate supervision signal. While effective, acquiring frame-level gloss annotations is labor-intensive, expensive, and often unavailable for low-resource languages (Yin and Read, 2020; Müller et al., 2023). Consequently, Gloss-free SLT (Camgoz et al., 2018; Bahdanau et al., 2014; Luong et al., 2015; Zhao et al., 2021; Zhou et al., 2023; Wong et al., 2024; Chen et al., 2024; Gong et al., 2024; Hwang et al., 2025; Kim et al., 2025), which learns translation directly from video-text pairs, has attracted increasing attention as a more scalable alternative.

To compensate for the lack of fine-grained gloss supervision, recent advancements (Zhou et al., 2023; Chen et al., 2024; Kim et al., 2025) typically adopt a “pre-training and fine-tuning” two-stage paradigm. In the pre-training stage, most methods have pivoted towards Vision-Language Pre-training (VLP), which utilizes contrastive learning to achieve video-sentence feature alignment. A representative paradigm (Kim et al., 2025) involves leveraging Multimodal Large Language Models (MLLMs, e.g., LLaVA (Liu et al., 2023) or Video-LLaVA (Lin et al., 2024)) to generate frame-wise textual descriptions, and then adds a mapper to predict the description from the visual feature of a frame. Finally, concatenating this predicted description and the visual features can effectively enrich feature representations. However, we argue that current methods have two key limitations that hinder alignment.

The first bottleneck is the high feature similarity caused by homogenized descriptions. Our statistical analysis (Table 1) reveals that over 95% of the generated descriptions share identical template prefixes (e.g., “The person in the image...”). This extreme redundancy causes the resulting text embeddings to become highly homogenized. Lacking discriminative power, these features fail to provide the high-quality supervision required for the visual

*Corresponding author.

Most Common Words			Description Templates			Feature Similarity	
Word	Count	Ratio (%)	Template Prefix	Count	Ratio (%)	Intra-Video (<i>Same Sentence</i>)	
the	221,645	3.84	the person in the image	54,938	95.3%	Mean	Std
their	186,725	3.24	the person is making a	904	1.6%	0.9984	0.0008
with	123,630	2.14	the person has their hands	348	0.6%	Inter-Video (<i>Different Sentence</i>)	
a	95,028	1.65	the person has their right	271	0.5%		
is	90,418	1.57	the person has their mouth	220	0.4%	Mean	Std
and	89,395	1.55	Others	955	1.6%	0.9985	0.0012
hand	71,435	1.24	Total Descriptions	57,636	100.0%		

Table 1: **Linguistic collapse in generated descriptions.** Among 57,636 frame descriptions from 500 videos, 95.3% share the same prefix, and inter-video feature similarity remains as high as intra-video similarity.

encoder to accurately capture the corresponding semantic differences between frames.

The second bottleneck is the loss of fine-grained information due to coarse temporal aggregation. Sign language videos are typically long, but the meaningful gestures only occur in short, sparse segments, while the majority of the video consists of meaningless transitions. (Pu et al., 2019) However, mainstream methods use Global Average Pooling (GAP) to aggregate features across the entire video. This approach forces the few informative frames to be averaged with the dominant background noise, causing critical motion cues to fade away. Consequently, the model struggles to capture the fine-grained features necessary to localize key gestures and establish precise associations between text words and video frames.

To resolve these challenges, we propose FiMA, a redundancy-aware description-enhancement framework for gloss-free SLT. First, to address the description redundancy, we introduce the **Description Enhancement Module (DEM)**. Instead of learning from the raw, homogenized descriptions, DEM explicitly calculates and removes the shared common patterns from the embedding space. This process enhances the distinctiveness of the supervision signal, ensuring the model learns meaningful semantic differences. Second, to tackle the loss of fine-grained details, we propose the **Multimodal Alignment Module (MAM)**. Departing from coarse global pooling, MAM utilizes a Max-Sim mechanism to pinpoint the most relevant visual frame for each text token. This enables precise frame-word alignment without requiring manual gloss annotations.

Figure 1 further supports this hypothesis. The baseline produces diffuse activations over broad temporal regions, which suggests that global aggregation smears semantic evidence across the whole

video. In contrast, FiMA yields sharper peaks for content words, indicating that Max-Sim can localize the most informative visual moments while suppressing irrelevant transition frames.

Our contributions are:

- We identify two key bottlenecks in MLLM-based gloss-free SLT: feature collapse caused by repetitive generated descriptions and detail loss caused by global temporal pooling.
- We propose DEM, a feature-level rectification module that removes both dataset-level template redundancy and instance-level static bias from frame description supervision.
- We introduce MAM with Max-Sim matching, which aligns each text token to its most relevant visual frame and enables sparse temporal grounding without gloss annotations.
- FiMA achieves state-of-the-art results on PHOENIX14T and CSL-Daily, including +3.11 BLEU-4 on CSL-Daily over prior gloss-free methods, with visual evidence of clearer feature separation and frame-word grounding.

2 Related Work

Sign Language Translation. SLT aims to translate continuous sign videos into natural spoken sentences. Because sign and spoken languages differ in grammar, word order, and non-manual cues (Stokoe Jr, 2005; Bragg et al., 2019), early systems commonly follow a Sign-to-Gloss-to-Text pipeline (Camgoz et al., 2020; Voskou et al., 2021; Hu et al., 2023; Chen et al., 2022b; Fu et al., 2023; Hu et al., 2022). Gloss labels provide a sparse and roughly monotonic intermediate sequence, so these methods often use CTC (Graves et al., 2006) to connect visual spikes with gloss tokens. This supervision is effective, but obtaining gloss annotations

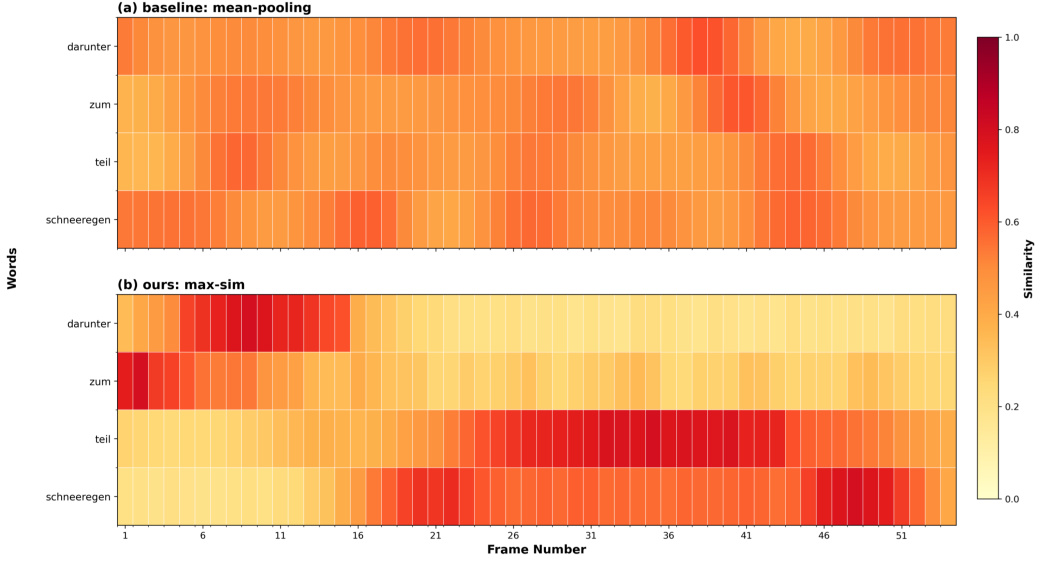


Figure 1: **Fine-grained semantic alignment.** Compared to the diffuse attention patterns in the baseline (a), our approach (b) produces sparse activations, indicating precise temporal localization.

is costly and the gloss sequence can discard facial expressions, body posture, and other non-manual information (Yin and Read, 2020).

Gloss-free SLT. Gloss-free methods reduce the annotation burden by learning directly from video-text pairs. Early approaches bridge the modality gap with attention, cross-modal reconstruction, keyword spotting (Wong et al., 2024), fixed or learned segment matching (Gong et al., 2024), or pseudo-labeling. More recent systems incorporate VLP or large language models (Zhou et al., 2023; Chen et al., 2022a, 2024; Kim et al., 2025) to provide richer supervision. However, most of them still aggregate a full video into one global embedding, which blends discriminative gestures with long transition regions. Dense alignment methods such as VAP (Jiao et al., 2024) attempt to recover local detail, but forcing every frame to match language can introduce noise because many frames correspond to preparation, hold, or background motion rather than lexical content. FiMA differs from both directions: it improves the quality of textual supervision before distillation and then selects sparse visual peaks for frame-word alignment.

3 Method

3.1 Overview of the Baseline Framework

Given a raw video $\mathcal{V}$ with T frames and a target sentence $\mathcal{S} = \{y_l\}_{l=1}^L$, the visual encoder yields frame-wise features $F_v = \{v_t\}_{t=1}^T \in \mathbb{R}^{T \times C}$, and the text encoder produces token embeddings $Z_s \in$

$\mathbb{R}^{L \times d}$. The goal is to learn a translation model $\mathcal{F} : \mathcal{V} \rightarrow \mathcal{S}$ without gloss supervision.

Recent gloss-free SLT systems often use two stages (Kim et al., 2025). In pre-training, an MLLM generates frame-wise descriptions $\mathcal{D}$, and a frozen text encoder such as BERT (Devlin et al., 2019) converts them into description targets $W = \{w_t\}_{t=1}^T$. A visual projector is trained to predict these targets from F_v , after which predicted text features are concatenated with visual features and processed by a temporal encoder. The resulting video sequence and the target sentence embedding are usually reduced by global average pooling and optimized with a contrastive objective (Oord et al., 2018). In fine-tuning, the pre-trained encoder is attached to a translation decoder such as mBART (Liu et al., 2020) and trained end-to-end. Fine-grained alignment remains underexplored.

3.2 Proposed Framework: FiMA

3.2.1 Motivation

Despite the success of the standard baseline, we identify two key limitations in its pipeline that hinder representation learning:

High Feature Similarity of Frame Descriptions:

The descriptions generated by frozen MLLMs suffer from severe homogenization and redundancy. This leads to abnormally high similarity among the supervision targets (W), preventing the visual encoder from learning discriminative representations.

Loss of Fine-grained Information: Existing meth-

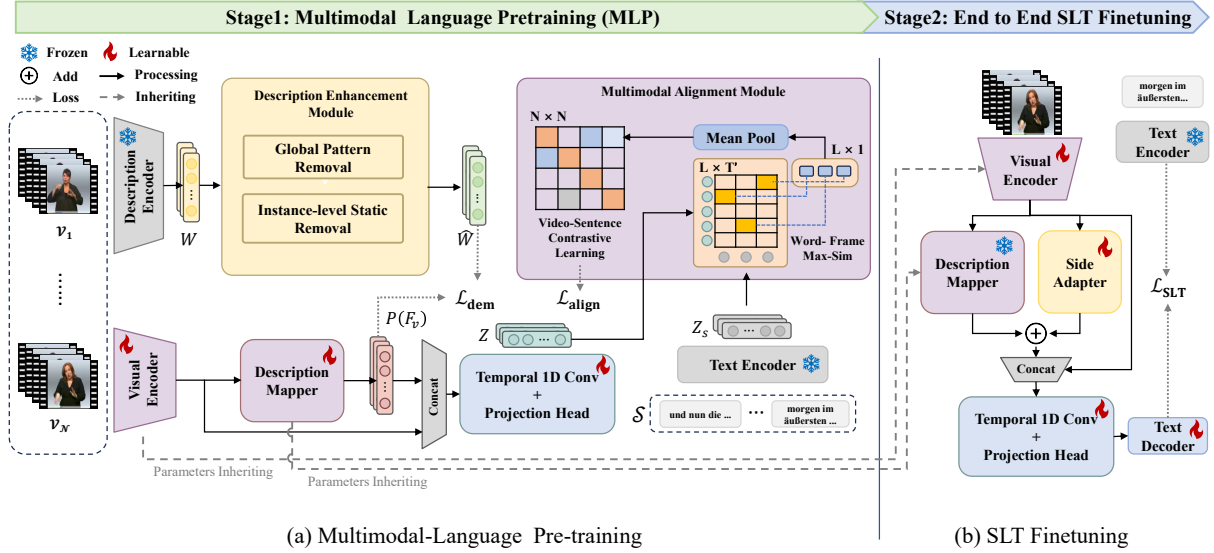


Figure 2: **The overall framework of FiMA.** In Stage 1 (Multimodal Pre-training), the **Description Enhancement Module (DEM)** mitigates feature homogenization by dynamically recalibrating the description targets ($\mathcal{L}_{dem}$). Concurrently, the **Multimodal Alignment Module (MAM)** employs a fine-grained **Max-Sim** strategy to align visual (Z) and text (Z_s) features, ensuring precise temporal grounding without global pooling ($\mathcal{L}_{align}$). In Stage 2 (End-to-End Fine-tuning), the pre-trained visual encoder and a translation decoder are optimized for the SLT task ($\mathcal{L}_{SLT}$).

ods predominantly rely on GAP to aggregate temporal features before alignment. This coarse-grained operation washes away fine-grained visual details, as sparse, meaningful sign gestures are indiscriminately averaged with dominant transitional noise (Pu et al., 2019).

3.2.2 Our Solution

To address the aforementioned issues, our proposed FiMA framework introduces targeted rectification modules specifically designed for the two training stages.

Stage 1: Multimodal-Language Pre-training (MLP). To overcome the limitations of the baseline’s pre-training, we redesign the supervision and alignment mechanisms:

Description Enhancement Module (DEM) (Sec. 3.3): To tackle the high feature similarity of the frame descriptions (W), we introduce DEM. Instead of passively distilling from homogenized targets, DEM explicitly identifies and removes the redundant common patterns from W to generate enhanced, discriminative targets $\hat{W}$. The visual projector is then trained to predict these high-quality targets $\hat{W}$ from the visual features F_v .

Multimodal Alignment Module (MAM) (Sec. 3.4): After concatenating F_v with the predicted text embeddings and passing them through the temporal encoder, we obtain the

enriched visual features $Z = \{z_t\}_{t=1}^{T'}$. We input Z and the sentence embeddings Z_s into MAM to perform fine-grained alignment. Departing from global pooling, MAM utilizes a fine-grained Max-Sim mechanism to automatically pinpoint and align the most relevant visual frames for each text token, ensuring precise alignment.

Stage 2: End-to-End Fine-tuning (SLT). We equip the pre-trained encoder with an mBART decoder. To facilitate the transition from alignment to translation, we introduce a lightweight Side Adapter (Sec. 3.5). Functioning as a residual pathway, it allows for task-specific calibration without disrupting the fine-grained pre-trained representations. This design helps bridge the objective gap between semantic alignment and autoregressive translation generation. The entire framework is then optimized end-to-end using the standard sequence-to-sequence translation objective.

3.3 Description Enhancement Module (DEM)

Raw description embeddings are highly clustered due to repeated templates. DEM therefore removes shared and instance-specific static components before using them as supervision. Let $w_{i,t}$ be the raw text feature of frame t in video i within a mini-batch of size B .

1. Global Pattern Removal. We first estimate a dataset-level template feature μ_{global} with a mo-

mentum buffer (He et al., 2020):

$$\mu_{\text{global}} \leftarrow m \cdot \mu_{\text{global}} + (1 - m) \cdot \frac{1}{B} \sum_{i=1}^B \bar{w}_i \quad (1)$$

where m is the momentum and $\bar{w}_i = \frac{1}{T} \sum_{t=1}^T w_{i,t}$. We subtract this prior as $w'_{i,t} = w_{i,t} - \mu_{\text{global}}$.

2. Instance-level Static Removal. We further suppress video-specific static bias by subtracting the sequence mean of the original description embeddings:

$$\hat{w}_{i,t} = w'_{i,t} - \frac{1}{T} \sum_{k=1}^T w_{i,k} \quad (2)$$

Finally, a projection head $P(\cdot)$ predicts the enhanced target from visual input $v_{i,t}$:

$$\mathcal{L}_{\text{dem}} = \frac{1}{B \times T} \sum_{i=1}^B \sum_{t=1}^T \|P(v_{i,t}) - \hat{w}_{i,t}\|_2^2 \quad (3)$$

This encourages the visual model to capture semantic variation rather than template artifacts. This dynamically recalibrates the supervision targets using dataset-level and video-level statistics. DEM is not intended as an exact semantic disentanglement mechanism: the subtracted statistics may retain useful semantic information, but the transformation empirically reduces feature concentration and places greater relative emphasis on frame-specific variation.

3.4 Multimodal Alignment Module (MAM)

MAM avoids GAP by matching each word to its most relevant video frame. This design follows the observation that lexical signs are sparse: a content word is often expressed by a short visual segment, while surrounding frames may contain transitions, preparation, or holds.

1. Feature Preparation. Visual features are processed by a 1D convolution into $Z = \{z_t\}_{t=1}^{T'} \in \mathbb{R}^{T' \times d}$, and the target sentence is encoded as $Z_s = \{u_l\}_{l=1}^L \in \mathbb{R}^{L \times d}$.

2. Max-Sim Alignment. We construct a frame-word similarity matrix $\mathcal{M} \in \mathbb{R}^{L \times T'}$ by comparing every sentence token with every temporally processed visual frame:

$$\mathcal{M}_{l,t} = \text{sim}(u_l, z_t), \quad S_l = \max_{1 \leq t \leq T'} \mathcal{M}_{l,t}, \quad (4)$$

where $1 \leq l \leq L$ and $1 \leq t \leq T'$. Each row of $\mathcal{M}$ contains the similarities between one sentence

token and all video frames, and S_l denotes the strongest frame-level evidence for token l .

This Max-Sim operator acts as a sparse temporal selector (Yao et al., 2021). Unlike dense frame-word matching, it does not force every background frame to explain a token; unlike global pooling, it preserves localized peaks. Function words may not have a unique visual anchor, but content words with clear visual evidence dominate the video-sentence score.

3. Contrastive Learning. The video-sentence score is $S(Z, Z_s) = \frac{1}{L} \sum_{l=1}^L S_l$ and is optimized with symmetric InfoNCE:

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B} \sum_{i=1}^B \left(\log \frac{e^{S(Z^{(i)}, Z_s^{(i)})/\tau}}{\sum_j e^{S(Z^{(i)}, Z_s^{(j)})/\tau}} + \log \frac{e^{S(Z^{(i)}, Z_s^{(i)})/\tau}}{\sum_j e^{S(Z^{(j)}, Z_s^{(i)})/\tau}} \right) \quad (5)$$

where τ controls the sharpness of the similarity distribution. The Stage 1 loss is $\mathcal{L}_{\text{mlp}} = \mathcal{L}_{\text{align}} + \lambda \mathcal{L}_{\text{dem}}$.

3.5 Sign Language Translation (Stage 2)

In Stage 2, we construct the end-to-end model by attaching an mBART decoder. However, features optimized for alignment (Stage 1) might suppress generation-relevant details. To bridge this gap, we introduce a Side Adapter to calibrate the features.

Task-Specific Side Adapter. We instantiate the Side Adapter $\mathcal{A}(\cdot)$ as a lightweight MLP attached in parallel with the visual projector. We fuse the streams via a residual connection where α is a learnable scalar:

$$d_t^{\text{final}} = P(v_t) + \alpha \cdot \mathcal{A}(v_t) \quad (6)$$

Zero-Initialized Residual Fusion. Crucially, we use Zero-Initialization for the adapter’s last layer, ensuring $\mathcal{A}(v_t) = 0$ at the start. This allows the model to inherit pure aligned representations from Stage 1 while gradually injecting task-specific cues during fine-tuning.

End-to-End Fine-tuning. The fused features $D^{\text{final}} = \{d_t^{\text{final}}\}_{t=1}^T$ are concatenated with the original visual features and fed into the mBART decoder (Liu et al., 2020). The entire model is fine-tuned using the standard cross-entropy loss:

$$\mathcal{L}_{\text{SLT}} = -\sum_{l=1}^L \log p_{\theta}(y_l | y_{<l}, \mathcal{V}; \theta) \quad (7)$$

where y_l is the target token and θ represents the model parameters.

4 Experiments

4.1 Datasets and Metrics

Datasets. We evaluate on two standard gloss-free SLT benchmarks:

PHOENIX14T (Camgoz et al., 2018): German weather-domain signing, with 7,096/519/642 train/dev/test pairs and a 2,887-word vocabulary.

CSL-Daily (Zhou et al., 2021a): Chinese daily-life signing, with 18,401/1,077/1,176 train/dev/test pairs and a 2,343-word vocabulary.

Evaluation Metrics. Following prior SLT work (Kim et al., 2025; Zhou et al., 2023), we report BLEU-1 to BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004).

Implementation Details. We use ResNet-18 (He et al., 2016) as the visual backbone, frozen BERT (Devlin et al., 2019) for description encoding, and mBART (Liu et al., 2020) as the translation decoder. Frame-level descriptions are pre-extracted with a frozen LLaVA-OneVision-7B model (Li et al., 2024). Stage 1 is trained for 80 epochs with AdamW (Loshchilov and Hutter, 2019), batch size 4, base learning rate 10^{-4} , $\tau = 0.07$, $\lambda = 0.1$, and DEM momentum $m = 0.9$. Stage 2 fine-tunes the pre-trained encoder, zero-initialized adapter, and decoder for 200 epochs, and inference uses beam search with beam size 4. Additional details and efficiency results are reported in the supplementary material.

4.2 Comparison with State-of-the-Art

Results on PHOENIX14T and CSL-Daily. As shown in Table 2, FiMA achieves state-of-the-art gloss-free performance on both datasets. On CSL-Daily, it improves over MMSLT (Kim et al., 2025) by +3.11 BLEU-4 and +4.34 ROUGE-L; on PHOENIX14T, it also obtains the best gloss-free BLEU-4 (27.28), demonstrating robust gains without gloss supervision. The gains are consistent across both a controlled weather-domain benchmark and a larger daily-life benchmark. This is important because FiMA does not add gloss supervision or external manual labels; it improves the alignment signal extracted from the same video-text pairs. The larger improvement on CSL-Daily suggests that removing redundant generated descriptions is especially useful when the visual and linguistic variation is broader.

4.3 Ablation Studies

We conduct ablations on PHOENIX14T (Camgoz et al., 2018) to isolate the contribution of each design choice.

Effectiveness of Main Components. As shown in Table 3, the baseline is limited by noisy description targets and coarse alignment. DEM improves BLEU-4 from 20.95 to 23.73 by making the frame-level supervision more discriminative. MAM alone brings a larger gain, reaching 26.09 BLEU-4, because sparse frame-word alignment directly addresses temporal detail loss. Combining DEM, MAM, and the side adapter achieves 27.28 BLEU-4 and 50.58 ROUGE-L, confirming that cleaner supervision and fine-grained alignment are complementary.

Impact of Alignment Strategies. Table 4 compares MAM with global pooling and dense alignment (Jiao et al., 2024). Global average pooling dilutes short gestures with long transitional regions, while global max pooling ignores token-level structure. Dense alignment improves over pooling but still forces many non-informative frames to participate in matching. Max-Sim achieves the highest BLEU-4 and remains competitive on ROUGE-L because it allows each word to select the strongest visual evidence and naturally filters out background frames.

Internal Mechanisms of DEM. Table 5 shows that both DEM operations are necessary. Global pattern removal subtracts common templates shared across the dataset, improving BLEU-4 from 26.09 to 26.82. Instance-level static removal targets video-specific constant bias and reaches 27.18. Combining them yields the best score, indicating that description collapse comes from both corpus-level phrasing and sequence-level static information.

Stage 2 Adapter Design. Table 6 studies the transition from alignment pre-training to translation fine-tuning. Direct fine-tuning is competitive but can overwrite aligned features. A randomly initialized adapter introduces additional capacity but perturbs the representation at the start of training. The zero-initialized adapter (Zhang et al., 2023b, 2024) performs best because it initially preserves the pre-trained pathway and then gradually learns task-specific corrections.

Method	CSL-Daily					PHOENIX14T				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Gloss-based										
SLRT (Camgoz et al., 2020)(CVPR20)	37.38	24.36	16.55	11.79	36.74	46.61	33.73	26.19	21.32	-
STN-SLT (Voskou et al., 2021)(ICCV21)	-	-	-	-	-	48.61	35.97	28.37	23.65	-
STMC-T (Zhou et al., 2021b)(TMM21)	-	-	-	-	-	46.98	36.09	28.70	23.65	46.65
BN-TIN-Transf.+SignBT (Zhou et al., 2021a)(CVPR21)	51.42	37.26	27.76	21.34	49.31	50.80	37.75	29.72	24.32	49.54
MMTLB (Chen et al., 2022a)(CVPR22)	53.31	40.41	30.87	23.92	53.25	53.97	41.75	33.84	28.39	52.65
TS-SLT (Chen et al., 2022b)(NeurIPS22)	55.44	42.59	32.87	25.79	55.72	54.90	42.43	34.46	28.95	53.48
SLTUNET (Zhang et al., 2023a)(ICLR23)	54.98	41.44	31.84	25.01	54.08	52.92	41.76	33.99	28.47	52.11
Weakly supervised gloss-free										
TSPNet (Li et al., 2020)(NeurIPS20)	17.09	8.98	5.07	2.97	18.38	36.10	23.12	16.88	13.41	34.96
GASLT (Yin et al., 2023)(CVPR23)	19.90	9.94	5.98	4.07	20.35	39.07	26.74	21.86	15.74	39.86
ConSLT (Fu et al., 2023)(ICASSP23)	-	-	-	14.53	40.98	-	-	-	21.59	47.69
VAP (Jiao et al., 2024)(ECCV24)	49.99	-	-	20.85	48.56	53.07	-	-	26.16	51.28
Gloss-free										
NSLT (Camgoz et al., 2018)(CVPR18)	-	-	-	-	-	27.10	15.61	10.82	8.35	29.70
NSLT+Bahdanau (Bahdanau et al., 2014)(CVPR18)	-	-	-	-	-	32.24	19.03	12.83	9.58	31.80
NSLT+Luong (Luong et al., 2015)(CVPR18)	34.16	19.57	11.84	7.56	34.54	29.86	17.52	11.96	9.00	30.70
CSGCR (Zhao et al., 2021)(TMM21)	-	-	-	-	-	36.71	25.40	18.86	15.18	38.85
GFSLT-VLP (Zhou et al., 2023)(ICCV23)	39.37	24.93	16.26	11.00	36.44	43.71	33.18	26.11	21.44	42.49
Sign2GPT (Wong et al., 2024)(ICLR24)	41.75	28.73	20.60	15.40	42.36	49.54	35.96	28.83	22.52	48.90
FLa-LLM (Chen et al., 2024)(LREC-COLING24)	37.13	25.12	18.38	14.20	37.25	46.29	35.33	28.03	23.09	45.27
SignLLM (Gong et al., 2024)(CVPR24)	39.55	28.13	20.07	15.75	39.91	45.21	34.78	28.05	23.40	44.49
SpaMo (Hwang et al., 2025)(NAACL25)	48.90	36.90	26.78	20.55	47.46	49.80	37.32	29.50	24.32	46.57
MMSLT (Kim et al., 2025) (ICCV25)	49.87	36.37	27.29	21.11	48.92	48.92	38.12	30.79	25.73	47.97
FiMA (Ours)	53.91	41.27	31.12	24.22	53.26	51.14	40.24	32.69	27.28	50.58

Table 2: Performance on CSL-Daily and PHOENIX14T.

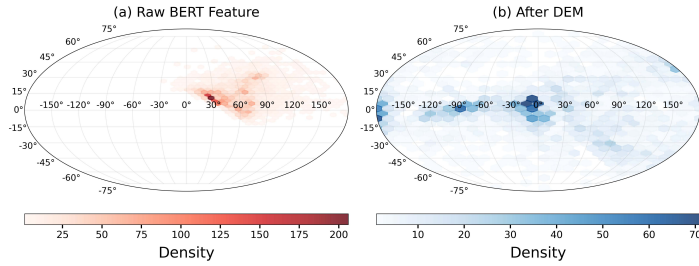


Figure 3: **Feature Distribution Visualization.** (a) Baseline features are highly clustered in a narrow space due to repetitive templates, making different video frames indistinguishable. (b) By explicitly removing shared redundancy, our method produces a widely dispersed distribution. This geometric expansion forces the features apart, enabling the model to accurately capture diverse and fine-grained semantic gestures.

DEM	MAM	Adapt.	B-4	R-L
-	-	-	20.95	41.76
✓	-	-	23.73	45.96
-	✓	✓	26.09	49.81
✓	✓	✓	27.28	50.58

Table 3: **Main components.** DEM and MAM provide complementary gains.

4.4 Visualization

4.4.1 Visualization of DEM's Effect

Feature Distribution Topology. To verify the impact of our Description Enhancement Module (DEM), we visualize the distribution of BERT feature embeddings in Figure 3. The baseline (a) shows that features are tightly clustered in a narrow direction, which is caused by the homogenized

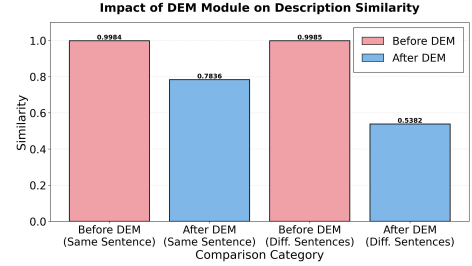


Figure 4: **Effect of DEM on feature similarity.** DEM reduces the pathological similarity of baseline features (0.998). The massive drop in inter-sentence similarity (-46.1%) compared to intra-sentence (-21.5%) shows that DEM restores the discriminative power of descriptions

Alignment Strategy	B-4	R-L
Global Avg Pooling	25.73	47.97
Global Max Pooling	24.59	46.72
Dense Alignment	26.64	50.76
Max-Sim (Ours)	27.28	50.58

Table 4: **Alignment strategies.** Max-Sim achieves the highest BLEU-4 and the second-highest ROUGE-L.

descriptions (e.g., identical prefixes). This means the features are almost identical and lack discriminative power. In contrast, our method (b) successfully disperses the features, resulting in a more uniform distribution. To further quantify the impact of DEM, we compare the cosine similarity of text features before and after applying the module (Figure 4). Baseline features suffer from severe

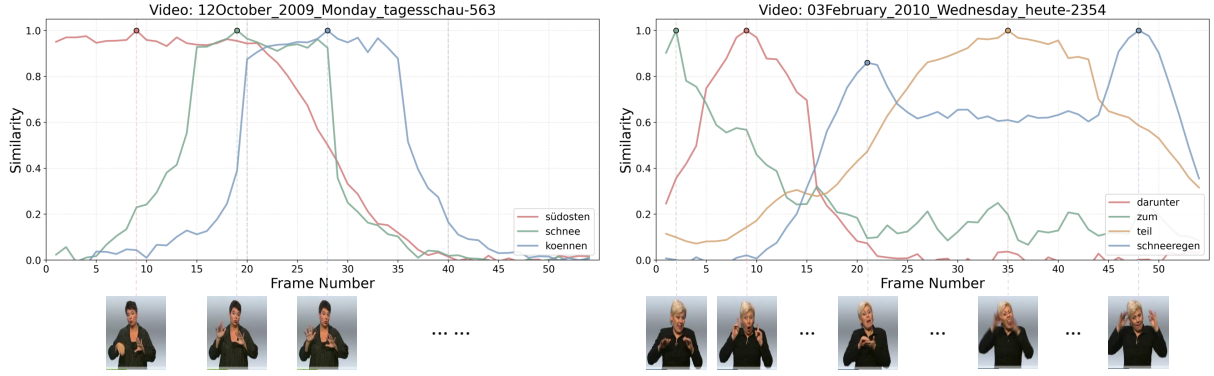


Figure 5: **Temporal Activation Curves.** We visualize the similarity scores between a specific word (e.g., “schneeregen”/ sleet) and video frames. The curve shows a sharp peak that aligns with the actual signing gesture. This demonstrates that our MAM module can pinpoint the most relevant visual moment for a semantic concept.

Global	Instance	B-4	R-L
-	-	26.09	49.81
✓	-	26.82	49.93
-	✓	27.18	50.22
✓	✓	27.28	50.58

Table 5: **DEM variants.** Both removals are useful.

Transfer Strategy	B-4	R-L
Direct Fine-tuning	26.81	50.09
Adapter w/o Init	26.56	49.80
Adapter w/ Zero-Init	27.28	50.58

Table 6: **Adanter design.** Zero-init transfers best.

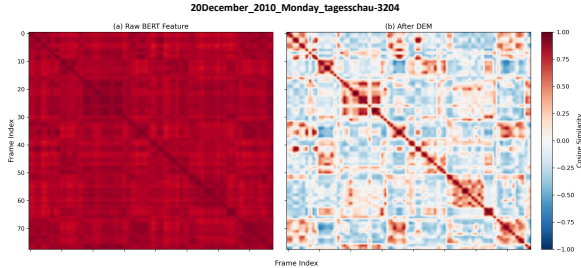


Figure 6: **Comparison of Temporal Self-Similarity Matrices (SSM).** (a) The baseline shows uniformly high similarity across the entire video (all red), indicating that the features are dominated by repetitive templates and fail to distinguish between frames. (b) Our method effectively removes these shared patterns, revealing a clear structure where frames are only similar to their neighbors, proving that the model now captures distinct motion changes.

homogenization, showing near-identical similarity (0.998) across all frames. By explicitly removing redundancy, DEM decreases intra-sentence similarity to 0.7836 (-21.5%) and drastically drops inter-sentence similarity to 0.5382 (-46.1%). This significant margin proves that DEM successfully restores the discriminative power of the text representations.

Temporal Self-Similarity. Figure 6 visualizes the

self-similarity matrix of a video feature sequence. The baseline (a) exhibits near-uniform high similarity across the entire matrix. This implies that the features are dominated by the repetitive template words (e.g., “The person...” rather than the actual sign gestures, making it impossible to distinguish one frame from another. In contrast, our method (b) suppresses these common patterns. The resulting matrix shows a clear diagonal structure, demonstrating that the features have evolved to capture dynamic motion changes, effectively distinguishing different temporal segments.

4.4.2 Visualization of MAM’s Effect

Analysis on Semantic Alignment. Figure 5 illustrates the frame-to-word similarity curves produced by our Multimodal Alignment Module (MAM). The results show that MAM encourages the model to discover sharp alignments for meaningful semantic concepts, verifying its strong capability to capture fine-grained visual features. Unlike global pooling, which smears signals across the entire video, MAM isolates semantic spikes. For concrete content words (e.g., nouns and verbs like ‘schneeregen’), the model produces distinct peaks. This implies that the visual encoder has successfully learned to filter out meaningless transitional trajectories (Pu et al., 2019), achieving accurate implicit temporal grounding. Conversely, for functional words (e.g., connectives like ‘zum’ and ‘teil’) which naturally lack corresponding visual gestures in sparse sign language grammar (Stokoe Jr, 2005), the model may generate spurious alignments as it attempts to find a match.

However, the quantitative results in Table 4 confirm that the benefit of precise grounding for content words significantly outweighs this noise. By

explicitly focusing on these fine-grained, discriminative cues rather than pooled global representations, our approach effectively bridges the modality gap and leads to better translation performance.

5 Conclusions

This work addresses two critical challenges in MLLM-enhanced gloss-free SLT: the high feature similarity caused by homogenized frame descriptions and the loss of fine-grained temporal information caused by coarse pooling. FiMA tackles these issues with DEM, which removes redundant description patterns before distillation, and MAM, which performs sparse Max-Sim frame-word alignment without gloss supervision. Experiments on PHOENIX14T and CSL-Daily show new state-of-the-art gloss-free results.

Limitations

FiMA relies on frame-level descriptions generated by a frozen MLLM during pre-training. Although DEM reduces template redundancy in the feature space, the framework can still inherit visual biases or omissions from the captioning model. In addition, Max-Sim is intentionally sparse: it is effective for content words with localized visual evidence, but function words or discourse markers may not correspond to a unique frame. We therefore view FiMA as a step toward fine-grained gloss-free alignment rather than a replacement for linguistically grounded sign annotations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62572349 and the National Key R&D Program of China under Grants 2023YFF0906200.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.
- Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.
- Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5120–5130.
- Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022b. Two-stream network for sign language recognition and translation. *Advances in neural information processing systems*, 35:17043–17056.
- 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.
- 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, volume 1 (long and short papers)*, pages 4171–4186.
- 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. IEEE.
- 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*, pages 18362–18372.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceed-*

- ings of the *IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2022. Temporal lift pooling for continuous sign language recognition. In *European conference on computer vision*, pages 511–527. Springer.
- Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. Self-emphasizing network for continuous sign language recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 854–862.
- 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.
- Peiqi Jiao, Yuecong Min, and Xilin Chen. 2024. Visual alignment pre-training for sign language translation. In *European Conference on Computer Vision*, pages 349–367. Springer.
- Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. 2025. Leveraging the power of mllms for gloss-free sign language translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21048–21058.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. *Llava-onevision: Easy visual task transfer*. Preprint, arXiv:2408.03326.
- Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. 2020. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045.
- Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Ilya Loshchilov and Frank Hutter. 2019. *Decoupled weight decay regularization*. Preprint, arXiv:1711.05101.
- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1412–1421.
- Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios Gonzales, and Sarah Ebling. 2023. 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.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- 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.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *Pytorch: An imperative style, high-performance deep learning library*. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Junfu Pu, Wengang Zhou, and Houqiang Li. 2019. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4165–4174.
- William C Stokoe Jr. 2005. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37.
- 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.

- Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2gpt: Leveraging large language models for gloss-free sign language translation. *arXiv preprint arXiv:2405.04164*.
- Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.
- 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*, pages 2551–2562.
- Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. In *Proceedings of the 28th international conference on computational linguistics*, pages 5975–5989.
- Biao Zhang, Mathias Müller, and Rico Sennrich. 2023a. Situnet: A simple unified model for sign language translation. *arXiv preprint arXiv:2305.01778*.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847.
- Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2024. *Llama-adapter: Efficient fine-tuning of language models with zero-init attention*. Preprint, arXiv:2303.16199.
- 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*, pages 20871–20881.
- Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021a. 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. 2021b. *Spatial-temporal multi-cue network for sign language recognition and translation*. *IEEE Transactions on Multimedia*, 24:768–779.

A Implementation Details

FiMA is implemented in PyTorch (Paszke et al., 2019) and trained on four NVIDIA RTX 3090 GPUs. We use an ImageNet-pretrained ResNet-18 (He et al., 2016) as the visual encoder. Its classification layer is removed, and the 512-dimensional output after layer4 and global average pooling is used as the frame representation. Descriptions are encoded using the final-layer [CLS] representation of a frozen bert-base-cased model, producing 768-dimensional targets. We use the facebook/mbart-large-50-many-to-many-mmt checkpoint for the sentence encoder and translation decoder. Frame-level descriptions are generated offline using a frozen LLaVA-OneVision-7B model (Li et al., 2024); this model is not required during inference. Videos are augmented and cropped to 224×224 .

For Stage 1 pre-training, we train for 80 epochs with batch size 4, AdamW (Loshchilov and Hutter, 2019), base learning rate 10^{-4} with cosine decay, contrastive temperature $\tau = 0.07$, DEM weight $\lambda = 0.1$, and DEM momentum $m = 0.9$. For Stage 2 fine-tuning, the pre-trained encoder, zero-initialized side adapter, and mBART decoder are trained for 200 epochs. We use beam search with beam size 4 at inference. Unless otherwise stated, all backbones, data splits, and baseline settings follow MMSLT (Kim et al., 2025) for a fair comparison.

B Robustness to Description Generation

A natural concern is whether DEM can be replaced by prompt engineering or simple text preprocessing. Prior work (Kim et al., 2025) reports that forcing frozen MLLMs to produce highly specific sign descriptions can destabilize generation and introduce hallucinations. We therefore compare the original prompt, a hard prompt that asks the MLLM to avoid common prefixes, and a rule-based strategy that strips frequent prefixes after generation.

Table 7 shows that text-level fixes are insufficient. Hard prompting lowers feature similarity but hurts BLEU-4, indicating that lexical diversity obtained through prompt pressure can damage supervision quality. Rule-based stripping improves the baseline modestly, but inter-sentence similarity remains high. DEM consistently improves all variants, which supports our claim that the key redundancy is a latent feature-space problem rather than only a surface-form problem.

Text Generation Strategy	Method	Intra-Sim ↓	Inter-Sim ↓	BLEU-4 ↑	ROUGE-L ↑
Original Prompt	Baseline (w/o DEM)	0.998	0.998	26.09	49.81
	+ DEM (Ours)	0.783	0.538	27.28	50.58
Hard Prompt (No prefix)	Baseline (w/o DEM)	0.930	0.892	24.10	45.10
	+ DEM (Ours)	0.826	0.784	26.42	48.90
Rule-based (Strip prefix)	Baseline (w/o DEM)	0.880	0.866	25.90	48.21
	+ DEM (Ours)	0.780	0.762	27.20	50.04

Table 7: Robustness to description generation strategies on PHOENIX14T. DEM consistently improves over prompt-based and rule-based text preprocessing.

C Max-Sim Variants

Comparison with CTC. Connectionist Temporal Classification (CTC) (Graves et al., 2006) is widely used when gloss annotations are available, because glosses provide a monotonic intermediate label sequence. In gloss-free SLT, however, the target spoken sentence may differ in order and granularity from the signed sequence. Max-Sim does not assume a monotonic gloss path. It computes a global frame-word similarity matrix and independently selects the most relevant frame for each token, making it more suitable for weakly supervised frame-word grounding.

Alternative Aggregations. We compare strict Max-Sim with Top- K Mean-Sim and Peak-Neighborhood Averaging. Top- K Mean-Sim averages the top K scores over the sequence. Peak-Neighborhood Averaging first normalizes similarities to $[0, 1]$, then averages the peak frame with adjacent frames whose similarity is above 0.9. Table 8 shows that strict Max-Sim performs best. Averaging additional frames reintroduces temporally disconnected noise or onset/offset transitions, weakening the sharp gesture-level signal.

MAM Variant	BLEU-4	ROUGE-L
Max-Sim (Default)	27.28	50.58
Top- K Mean-Sim ($K = 3$)	25.90	48.20
Top- K Mean-Sim ($K = 5$)	25.12	46.20
Peak-Neighborhood Averaging	25.26	47.56

Table 8: Ablation on similarity aggregation in MAM.

D Hyperparameter Sensitivity

DEM uses a momentum-updated global pattern estimate. Table 9 evaluates the interaction between momentum m and batch size B . FiMA is robust to small batch sizes, while very low momentum values underperform because the global estimate is

overwritten by local mini-batch noise. The default setting $m = 0.9$, $B = 4$ gives the best BLEU-4.

Momentum	$B = 2$	$B = 4$	$B = 8$
$m = 0.1$	25.96	26.12	25.80
$m = 0.5$	26.64	26.58	26.24
$m = 0.9$	27.02	27.28	26.81
$m = 0.99$	26.92	27.16	27.01

Table 9: DEM sensitivity on PHOENIX14T, reported as BLEU-4.

E Additional Visualizations

Figure 7 reports fine-tuning dynamics. FiMA converges to a stronger translation model than the baseline, supporting that the pre-training alignment is useful for the downstream decoder rather than only for contrastive retrieval.

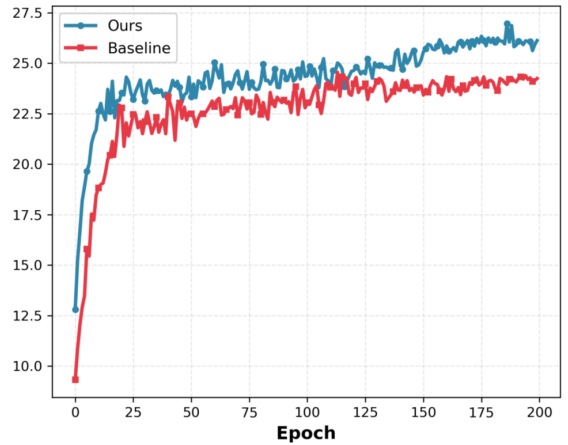


Figure 7: **Fine-tuning dynamics.** FiMA maintains better validation BLEU during end-to-end SLT fine-tuning.

F Qualitative Translation Examples

Table 10 presents a qualitative comparison between MMSLT and FiMA on representative translation

Ground Truth	tiefer luftdruck bestimmt in den nächsten tagen unser wetter
<i>English</i>	<i>Low pressure determines our weather in the next few days.</i>
MMSLT (Kim et al., 2025)	das tief das nach mitteleuropa zieht bestimmt morgen unser wetter [[in den nächsten tagen]]
<i>English</i>	<i>The low that moves to Central Europe determines our weather tomorrow [[in the next few days]].</i>
FiMA	tiefer luftdruck bestimmt in den nächsten tagen unser wetter
<i>English</i>	<i>Low pressure determines our weather in the next few days.</i>
Ground Truth	erst richtung süden so alpenrand werden die wolken dichter und zwischen donau und alpenrand kann es auch immer noch ein bisschen schneien wird aber auch im laufe des tages weniger
<i>English</i>	<i>Only towards the south, around the edge of the Alps, the clouds become thicker, and between the Danube and the edge of the Alps it can still snow a bit, but it also becomes less during the day.</i>
MMSLT (Kim et al., 2025)	in der südhälfte dagegen mehr wolken als sonne und vor allem an der donau und am alpen kann es gebietsweise noch schneien aber allmählich ändert sich das wetter [[wird aber auch im laufe des tages weniger]]
<i>English</i>	<i>In the southern half on the other hand, more clouds than sun, and especially on the Danube and the Alps it can still snow in places, but gradually the weather changes [[but it also becomes less during the day]].</i>
FiMA	auch im süden gibt es mal mehr mal weniger wolken die vor allem in der donau und in alpennähe gebietsweise schneefall bringen aber der schneefall lässt im laufe des tages allmählich nach
<i>English</i>	<i>Also in the south there are sometimes more, sometimes less clouds, which bring localized snowfall especially in the Danube and near the Alps, but the snowfall gradually subsides during the day.</i>

Table 10: Qualitative comparison. **Green** marks correct semantic matches, **red** marks hallucinated or incorrect content, and **orange brackets** mark missing fragments.

examples. In the first case, MMSLT captures the core topic of the sentence (“our weather” and “low pressure”) but introduces hallucinated details such as “moving to Central Europe” and incorrectly narrows the temporal scope to “tomorrow”, while also omitting the important phrase “in the next few days.” In contrast, FiMA reproduces the full semantic content accurately without introducing spurious information. In the second example, MMSLT again preserves part of the general weather semantics but exhibits noticeable semantic drift, replacing the original description with loosely related content such as “more clouds than sun” and “the weather gradually changes,” while failing to retain the key detail that snowfall decreases throughout the day. FiMA, by comparison, generates a semantically faithful and coherent translation that preserves the essential meteorological events and temporal progression. These examples suggest that FiMA is more effective at maintaining global semantic consistency and temporal completeness, while significantly reducing hallucination and omission errors compared with the baseline.

G Training and Inference Efficiency

We evaluate efficiency on the same 4-GPU machine used for training. Table 11 shows that DEM and MAM introduce a small pre-training overhead: 80 epochs take 9h 15m, only 29 minutes longer than

the baseline. The overhead comes from the moving average update and the frame-word similarity matrix.

Method	Time / Epoch	Total Time
Baseline	~6.58 min	8h 46m
FiMA	~6.94 min	9h 15m

Table 11: Pre-training time for 80 epochs on 4 RTX 3090 GPUs.

During inference, DEM and contrastive MAM are not used. Only the lightweight side adapter remains active. Table 12 shows that FiMA adds 0.329M trainable parameters and about 17 ms latency per video on the PHOENIX14T test set.

Method	Params (M)	Total Time	Latency
Baseline	29.797	161s	~251 ms/video
FiMA	30.126	172s	~268 ms/video

Table 12: Inference efficiency on the PHOENIX14T test set of 642 videos.