Attention-Steered Vision-Language Models for Sign Language Translation

Meibo Hu[1], **Guohao Sun**[1], **Annemarie D. Ross**[3], **Sheng Li**[2], **Zhiqiang Tao**[1]
[1]Rochester Institute of Technology, [2]University of Virginia
[3]National Technical Institute for the Deaf / Rochester Institute of Technology

Abstract

Vision-language models (VLMs) have emerged as a powerful framework for multimodal video understanding. However, they remain limited in the sign language translation task, where we identify a key failure mode of existing VLM-based translators: *poor spatial-temporal visual grounding*. In particular, we find that standard next-token cross-entropy does not directly provide signal for where and when the model should attend, causing models to overlook sign-relevant regions and frames. To address this challenge, we propose **AttnSign**, a VLM-based spatial-temporal attention steering framework for sign language translation. AttnSign first introduces spatial attention supervision for sign-relevant regions, such as face and hands, in each frame; then develops an RL-based motion-cadence steering method that encourages the model to explore and focus on sign-level keyframes. Experimental results on How2Sign and OpenASL benchmarks show that our proposed AttnSign consistently outperforms existing methods. Code will be released.

1 Introduction

Sign language translation (SLT) converts continuous sign language videos into natural-language sentences. Early systems rely on gloss-based supervision (Camgoz et al., 2020; Zhang et al., 2023), while more recent methods bypass it entirely (Wong et al., 2024; Fish and Bowden, 2026; Kim et al., 2025; Jiang et al., 2026). These gloss-free methods typically adopt a modular design: a visual encoder is pretrained with pseudo-gloss supervision (Guo et al., 2026), contrastive video-text alignment (Zhou et al., 2023), or self-supervised learning (Hu et al., 2023; Gueuwou et al., 2025b), and then fine-tuned jointly with a language decoder.

Meanwhile, vision-language models (VLMs) such as Gemini (Team et al., 2023), GPT-4 (Achiam et al., 2023), and Qwen-VL (Bai et al., 2023) have rapidly advanced, enabling strong general-purpose multimodal perception and generation in a single unified backbone. This naturally raises the question: *can general VLMs serve as effective SLT translators?* However, we empirically observe that these VLMs show limited performance on SLT. We thereby investigate how to close this gap and propose **AttnSign**, an end-to-end, gloss-free training framework that adapts a general VLM for sign language translation.

To adapt VLMs for SLT, we start with supervised fine-tuning (SFT) (Liu et al., 2023) on visual instructional data, which teaches the VLM to translate by following instructions. After training, we observe that the model can output fluent translations while assigning high visual attention to background and clothing regions, resulting in poor interpretability. Yet, sign language conveys linguistic content through manual features such as handshape, movement, and location, as well as non-manual facial markers (Brentari and Eccarius, 2010; Pfau et al., 2010), whereas these non-signing regions carry no linguistic information. In this work, we hypothesize that cross-entropy applied only to text tokens cannot provide explicit signals to regularize misaligned attention. We address this problem by adding a spatial attention supervision, enabling the model to generate translations while aligning visual attention with an annotated prior.

SFT optimizes a token-level objective, while translation quality is judged at the sequence level. To complement SFT, we refine the model as a translation policy via reinforcement learning (RL), following existing works (Feng et al., 2026).

Standard RL algorithms such as GRPO (Shao et al., 2024) are task-agnostic, treating sign language as any other text-generation task without considering its specific properties. Meanwhile, we observe that its KL penalty term offers limited benefit and can restrict policy updates on our task, consistent with recent RL studies on LLM (Yu et al., 2025; Hu et al., 2025). To inject sign-specific

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 30596–30611
October 24-29, 2026 ©2026 Association for Computational Linguistics

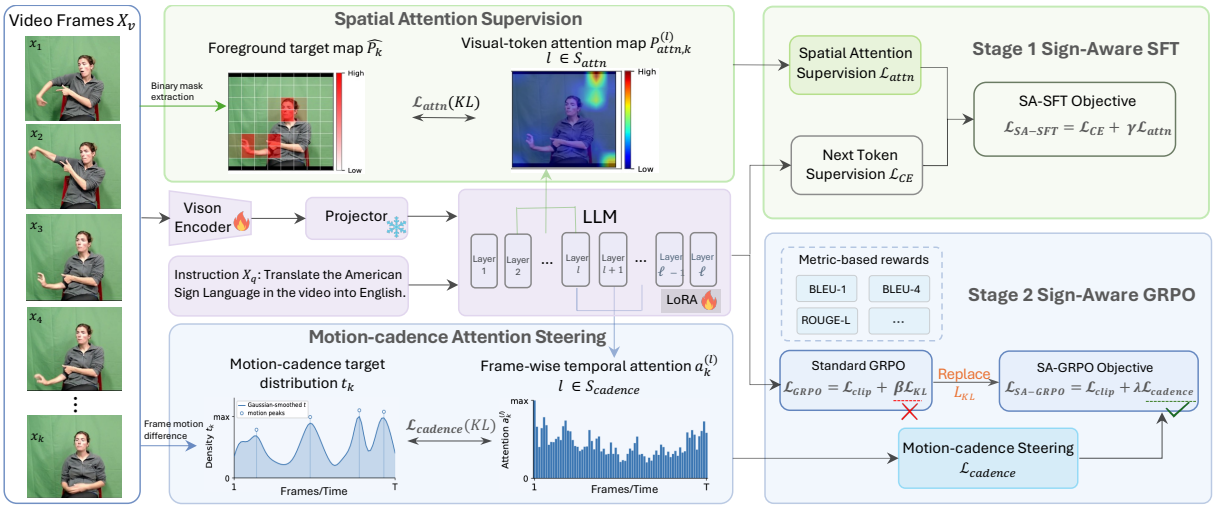


Figure 1: Overview of the proposed training framework. The general VLM takes video frames X_v and instruction X_q as inputs and generates the translation. In Sign-Aware SFT stage, we introduce a spatial attention $\mathcal{L}_{attn}$ at selected LLM layers with frame-level foreground targets, jointly with the next-token prediction loss $\mathcal{L}_{CE}$. In Sign-Aware GRPO stage, we add a motion-cadence steering objective $\mathcal{L}_{cadence}$ in place of the KL penalty term $\mathcal{L}_{KL}$, which aligns per-frame temporal attention $a^{(l)}_k$ with a Gaussian-smoothed motion-peak target t_k. We adopt the translation quality-based rewards.

signal, we leverage a temporal property of sign language: signs are organized through structured movements and holds that produce changes in visual motion energy over time (Brookshire et al., 2017), with motion peaks marking visually salient keyframes within each sign. Given that the KL penalty can offer limited benefit, we then replace it with a motion-cadence steering mechanism that guides temporal attention toward these keyframes during RL training.

We evaluate AttnSign on two large-scale American Sign Language benchmarks, How2Sign and OpenASL. Our proposed AttnSign reaches BLEU-4 of 15.54 on How2Sign and 20.75 on OpenASL, outperforming leading specialized SLT methods and general-purpose VLM baselines. We summarize the contributions as follows:

- We propose AttnSign, an end-to-end training framework that adapts a general VLM for SLT, achieving better interpretability and accuracy.
- In the SFT stage, we introduce a spatial attention supervision that guides the model to focus on sign-relevant regions in each frame.
- In the RL stage, we propose a motion-cadence steering mechanism in place of the KL penalty in standard GRPO, guiding the model's temporal attention toward keyframes of each sign.
- Experiments on benchmark datasets show the leading performance over both task-specific methods and general-purpose VLMs.

2 Related Work

Sign Language Translation. Prior SLT methods use gloss annotations to bridge visual and linguistic modalities (Chen et al., 2022; Zhang et al., 2023), which bring promising performance but can be hard to scale due to the annotation cost. More recently, gloss-free methods (Zhou et al., 2023; Gong et al., 2024; Rust et al., 2024; Pu et al., 2026; Hamidullah et al., 2024) learn from video-text pairs with auxiliary alignment or pretraining objectives, being more flexible and seeing further performance boost. Given the success of foundation models (Liu et al., 2020; Touvron et al., 2023; Bai et al., 2023), existing works (Zhang et al., 2024; Wong et al., 2024; Hwang et al., 2024; Rao et al., 2025; Wang et al., 2025b; Jang et al., 2025) adopt LLMs for decoding. However, these works generally need to tailor the model structure by integrating customized vision encoders (He et al., 2016; Oquab et al., 2023). How to directly tune a unified VLM model for SLT remains underexplored.

Training Strategies for SLT. These SLT systems are typically trained in two stages: visual-encoder pretraining followed by joint fine-tuning. Pretraining objectives include contrastive alignment between video and text (Kim et al., 2025), masked self-supervised modeling (Rust et al., 2024; Gueuwou et al., 2025b), and pseudo-gloss prediction (Wong et al., 2024). Most works rely on SFT for the fine-tuning stage (Liang et al., 2024;

Gong et al., 2024; Sun et al., 2024), while recent works (Rao et al., 2025; Pu et al., 2026) further explore reinforcement learning after SFT. Meanwhile, general-purpose VLMs—including LLaVA-OneVision (Li et al., 2024b) and Qwen-VL (Bai et al., 2025) offer strong general-purpose visual-language capabilities (Zhao et al., 2026; Inan et al., 2025; Sun et al., 2025). Instead of relying on a separate pretraining stage, we adapt a general VLM end-to-end through SFT and RL with sign-specific domain knowledge introduced at both stages.

Attention Supervision. Attention supervision has been studied in vision and multimodal learning to improve grounding, including attention transfer (Li et al., 2024a), human attention or gaze guidance (Selvaraju et al., 2019; Yan et al., 2024; Wang et al., 2025a), and region-level grounding for vision-language models (Wan et al., 2024). These works demonstrate that attention supervision can effectively shape model behavior, yet its potential remains underexplored in sign language translation. To address this gap, we develop attention steering methods tailored to SLT by guiding attention in the LLM decoder along both spatial and temporal dimensions.

3 Method

Section 3.1 introduces SFT on a general VLM, and Section 3.2 extends it with spatial attention supervision. Section 3.3 describes GRPO for sequence-level optimization, and Section 3.4 enhances it with motion-cadence attention steering.

3.1 Instructional SFT on a VLM

Given a sign language video $X_v = \{x_k\}_{k=1}^{K}$ of K frames, language instruction X_q, and a target translation Y, the goal is to learn a translator π_θ parameterized by θ, such that $Y = \pi_\theta(X_v, X_q)$.

We use a general VLM that models perception and generation of each output token $y_n \in \{y_1 \cdots y_N\} = Y$ jointly as:

$$P_\theta(Y \mid X_v, X_q) = \prod_{n=1}^{N} P_\theta(y_n \mid X_v, X_q, y_{<n}) \ . \tag{1}$$

where the input frames X_v are encoded into visual tokens using a vision encoder and projected to language space by a projector, followed by the instruction tokens.

Instructional Fine-tuning. Pretrained VLMs provide three capabilities directly useful for SLT: general visual perception, fluent language generation, and alignment between visual tokens and language. However, they can lack SLT-specific knowledge due to the scale and quality of the sign language data in the real world (Duarte et al., 2021). We empirically find that representative pretrained VLMs such as GPT-5 and Qwen2.5-VL-7B (Bai et al., 2025) can show limited performance on both How2Sign and OpenASL (Table 1 and Table 2).

We cast SLT as instruction-following generation task. Each training instance pairs a video X_v with a natural-language instruction X_q (*e.g.*, "*Translate the American Sign Language in this video to English*") and uses the target translation Y as the response. The model is trained with the standard cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{n=1}^{N} \log P_\theta(y_n \mid X_v, X_q, y_{<n}) \,, \tag{2}$$

where θ is optimized to maximize the likelihood of the translation conditioned on the input video and instruction. The SFT stage transfers the VLM's general capabilities to the SLT task.

Observation on Visual Attention. Despite the fact that sign-linguistic content is primarily conveyed by the signer's hands and face, We empirically find that after instructional SFT, the model can assign high visual attention to the less informative areas, including background and clothing regions (Figure 2, middle column). Considering that cross-entropy on text tokens does not penalize misaligned attention, we hypothesize that applying explicit spatial constraint to attend more on sign-relevant regions can benefit translation.

3.2 SA-SFT: Sign-Aware SFT

To achieve the goal, we propose a spatial supervision loss during SFT. Unlike approaches that rely on cropped regions or keypoints as inputs (Gan et al., 2023; Gueuwou et al., 2025a; Li et al., 2025), we keep the raw video as input and impose the spatial constraint through the loss. This preserves global visual context while guiding attention toward hand and face regions that carry primary sign-linguistic cues.

Foreground target. For each frame x_k of a training video, we use MediaPipe (Lugaresi et al., 2019) to extract a binary mask $\mathbf{M}_k$ covering the hands and face. We then resize the mask via average pooling to the 8×8 visual-token grid, yielding a soft foreground intensity vector $\mathbf{M}_{k,\text{flat}} \in [0,1]^{64}$, and con-

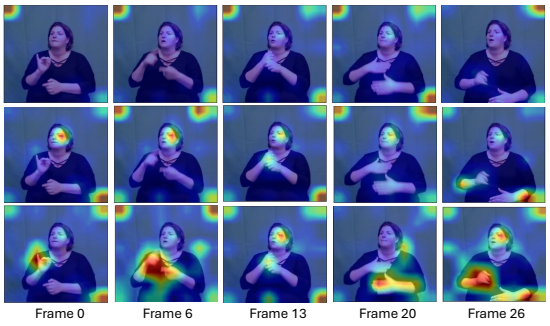


Figure 2: Visual-token attention maps at LLM layer 16 across five frames (columns) of the same test video. Rows show the three training stages: pre-trained VLM (top), standard SFT (middle), and Sign-Aware SFT with our spatial attention supervision (bottom).

vert it into a soft spatial target using a temperature-controlled softmax:

$$\hat{P}_k = \mathrm{softmax}(\mathbf{M}_{k,\mathrm{flat}}/T). \tag{3}$$

Spatial Attention loss. Let $P^{(l)}_{\mathrm{attn},k} \in \mathbb{R}^{64}$ denote the per-frame attention distribution over the 64 visual positions of frame k at LLM decoder layer l. We obtain this distribution by extracting the attention weights assigned to visual tokens and renormalizing them within each frame, so that the 64 spatial positions of frame k form a valid probability distribution. We supervise a set of decoder layers $\mathcal{S}_{\mathrm{attn}}$ with:

$$\mathcal{L}_{\mathrm{attn}} = \frac{1}{K\,|\mathcal{S}_{\mathrm{attn}}|} \sum_{k=1}^{K} \sum_{l \in \mathcal{S}_{\mathrm{attn}}} \mathrm{KL}\Big(P^{(l)}_{\mathrm{attn},k} \,\|\, \hat{P}_k\Big), \tag{4}$$

averaging over frames and supervised layers.

This objective penalizes attention mass assigned to regions with low target probability, thereby encouraging the model to concentrate attention on the hand and face regions. The SA-SFT objective combines the standard next-token cross-entropy loss with spatial attention supervision:

$$\mathcal{L}_{\text{SA-SFT}} = \mathcal{L}_{\mathrm{CE}} + \gamma\,\mathcal{L}_{\mathrm{attn}}, \tag{5}$$

where γ controls the strength of attention supervision. We supervise mid-depth LLM decoder layers, which we empirically find to be the most effective for spatial attention supervision (Section 4.5).

SFT optimizes token-level cross-entropy under teacher forcing, whereas translation quality is evaluated with sequence-level metrics such as BLEU, ROUGE, and BLEURT. This creates an objective mismatch: cross-entropy encourages high-probability next-token predictions, but does not directly optimize full-sentence quality or enhance the semantic score. Following existing works (Feng et al., 2026), we therefore further perform post-training stage upon SFT model, *e.g.*, policy optimization.

3.3 GRPO for Sequence-Level Optimization

We first study the canonical GRPO (Shao et al., 2024) to further fine-tune the SFT model for the SLT, considering GRPO can also support different rule-based rewards for optimizing non-differentiable translation metrics.

Formulation. For each input video X_v, we sample G candidate translations $\{Y^{(g)}\}_{g=1}^{G}$ from the current policy and score each with a reward $r(X_v, Y^{(g)})$. We define the reward as a reference-based function:

$$r(X_v, Y^{(g)}) = R(Y^{(g)}, Y^*), \tag{6}$$

where Y^* is the reference translation and $R(\cdot,\cdot)$ measures translation quality. This formulation allows us to instantiate R with different rule-based reward designs, which we compare in Section 4.

We compute group-relative advantages by comparing each candidate with the group mean reward:

$$A^{(g)} = r(X_v, Y^{(g)}) - \frac{1}{G}\sum_{g'=1}^{G} r(X_v, Y^{(g')}). \tag{7}$$

The advantage $A^{(g)}$ is assigned to all generated tokens of $Y^{(g)}$. Let $c = (X_v, X_q)$. The clipped GRPO surrogate is defined per-token as:

$$\mathcal{L}_{\mathrm{clip}} = -\mathbb{E}\Bigg[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|Y^{(g)}|}\sum_{t=1}^{|Y^{(g)}|}\min\Big(\rho_t^{(g)}A^{(g)}, \mathrm{clip}(\rho_t^{(g)}, 1-\epsilon, 1+\epsilon)\,A^{(g)}\Big)\Bigg], \tag{8}$$

where $\rho_t^{(g)} = \pi_\theta(y_t^{(g)} \mid c, y_{<t}^{(g)})/\pi_{\theta_{\mathrm{old}}}(y_t^{(g)} \mid c, y_{<t}^{(g)})$ is the token-level probability ratio, $\pi_{\theta_{\mathrm{old}}}$ denotes the policy before the update, and ϵ is the clipping bound. The standard GRPO objective combines this with a KL penalty:

$$\mathcal{L}_{\mathrm{GRPO}} = \mathcal{L}_{\mathrm{clip}} + \beta\,\mathcal{L}_{\mathrm{KL}}, \tag{9}$$

where β controls the KL strength, and $\mathcal{L}_{\mathrm{KL}}$ keeps the policy close to a fixed reference π_{ref}. The KL penalty term is:

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{c\in D}\left[D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c)\|\pi_{\mathrm{ref}}(\cdot \mid c))\right], \tag{10}$$

where D_{KL} is the KL divergence.

Limitations of standard GRPO. Standard GRPO can have limitations for SLT. First, it is task-agnostic and carries no explicit SLT-specific knowledge. Second, we observe that its generic $\mathcal{L}_{\text{KL}}$, which only anchors the policy to the reference π_{ref}, can bring limited benefit and even can restrict the potential of the policy model for SLT (Table 3). This is consistent with recent RL studies on LLM (Yu et al., 2025; Hu et al., 2025). Accordingly we replace $\mathcal{L}_{\text{KL}}$ with a proposed sign-specific steering mechanism that addresses both issues.

3.4 SA-GRPO: Sign-Aware GRPO

Sign language exhibits structured temporal dynamics, with signs organized through movements and holds that give rise to changes in visual motion energy over time (Liddell and Johnson, 1989; Brookshire et al., 2017). This makes motion peaks a natural target for guiding optimization. However, this prior alone does not tell us whether focusing attention on these peaks actually helps translation quality, or in which LLM layers this effect holds. Accordingly, we uncover the attention-performance relationship through a statistical analysis, connecting the model's per-layer attention concentration with translation quality.

Statistical Analysis. For each video and LLM layers of the SA-SFT model, we obtain a per-frame attention vector by aggregating over visual tokens within each frame and averaging over text-query positions. We compute the concentration score as $C = 1 - H/\log K$, where H is the entropy over K frames. Figure 3 shows that there exists a positive correlation relationship between C and BLEU-1, computed across both training and test videos, especially peaks at the upper-middle 17th and 19th layers ($r \approx +0.32$), indicating that enhancing such an attention concentration at these layers *during optimization* can potentially improve the translation quality *during inference*.

Motion-cadence Steering Loss. While the analysis identifies at which layers concentration helps, the linguistic prior that informative moments cluster around motion peaks tells us which frames to concentrate on. We therefore construct a frame-level target around motion peaks at these layers. For each training video, we compute a per-frame motion signal using grayscale frame differences:

$$m_k = \frac{1}{HW} \sum_{h,w} \left| x_k^{\text{gray}} - x_{k-1}^{\text{gray}} \right|_{h,w}, \qquad (11)$$

where x_k^{gray} denotes the grayscale frame at time k.

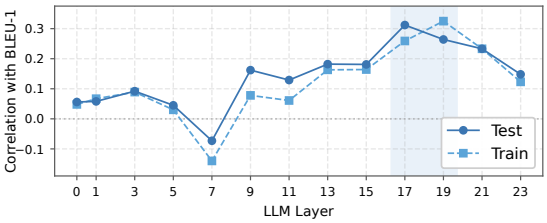


Figure 3: Correlation between ***temporal attention score*** and ***translation performance*** (measured by BLEU-1). For example, the strongest positive correlation emerges at the 17th and 19th layers for the SA-SFT model, motivating us to steer the temporal attention for the performance boost during optimization.

We detect local maxima $\mathcal{P}$ in m_k with a minimum inter-peak distance of 3 frames, and convert them into a normalized temporal target:

$$t_k = \frac{1}{Z} \sum_{p \in \mathcal{P}} \exp\left(-\frac{(k-p)^2}{2\sigma^2} \right), \qquad \sum_k t_k = 1, \qquad (12)$$

where the Gaussian bandwidth is set to $\sigma = 2.5$ frames, spreading each peak to neighboring frames to yield a soft cadence target. For videos where no local maxima are detected ($\mathcal{P} = \emptyset$), we skip $\mathcal{L}_{\text{cadence}}$ for that sample and retain only $\mathcal{L}_{\text{clip}}$.

Let $a_k^{(l)}$ denote the attention mass on frame k at LLM layer l. We obtain it by summing attention over the 64 visual tokens of frame k and normalizing across all frames. We supervise a set of layers $\mathcal{S}_{\text{cadence}}$ toward this target with:

$$\mathcal{L}_{\text{cadence}} = \frac{1}{|\mathcal{S}_{\text{cadence}}|} \sum_{l \in \mathcal{S}_{\text{cadence}}} \text{KL}\left(a^{(l)} \parallel t \right). \qquad (13)$$

This loss keeps temporal attention aligned with the soft cadence target t centered on motion peak.

SA-GRPO. We replace the $\mathcal{L}_{\text{KL}}$ in standard GRPO with the sign-specific cadence term $\mathcal{L}_{\text{cadence}}$. The SA-GRPO objective is:

$$\mathcal{L}_{\text{SA-GRPO}} = \mathcal{L}_{\text{clip}} + \lambda \mathcal{L}_{\text{cadence}}, \qquad (14)$$

where λ controls the strength of the cadence steering. Unlike generic $\mathcal{L}_{\text{KL}}$, $\mathcal{L}_{\text{cadence}}$ is conditioned on the input video and ties the regularization to sign-relevant motion structure. Interestingly, we observe a continued performance boost tendency by reducing the KL penalty in canonical GRPO or increasing our proposed motion-cadence term. See more discussion in Section 4.5.

Method	B-1	B-2	B-3	B-4	ROUGE-L	BLEURT
Closed-source VLMs						
Gemini-2.5-Pro (Comanici et al., 2025)	15.41	5.57	2.61	1.41	13.79	30.90
GPT-5	6.64	1.02	0.68	0.10	5.72	24.62
Open-source VLMs						
LLaVA-OV-7B (Li et al., 2024b)	11.21	3.56	1.45	0.16	4.10	20.15
Qwen2.5-VL-7B (Bai et al., 2025)	12.42	4.12	1.13	0.11	7.24	24.10
InternVL-2.5-8B (Chen et al., 2024a)	11.81	2.35	0.86	0.06	7.56	25.12
Trained on H2S only						
GloFE-VN (Lin et al., 2023)	14.94	7.27	3.93	2.24	12.61	31.65
Tarrés et al. (Tarrés et al., 2023)	34.01	19.30	12.18	8.03	—	—
SSVP-SLT (Rust et al., 2024)	30.20	16.70	10.50	7.00	25.70	39.30
SignMusketeers (Gueuwou et al., 2025a)	18.80	8.10	4.20	2.40	—	—
C^2RL (Chen et al., 2025)	29.07	18.56	12.92	9.37	27.02	—
SpaMo (Hwang et al., 2024)	33.41	20.28	**13.96**	**10.11**	**30.56**	42.23
AttnSign (Ours)	**35.82**	**20.68**	13.52	9.45	30.45	**42.73**
Trained on H2S + YT						
Uthus et al. (Uthus et al., 2023)	37.82	24.13	16.92	12.39	—	46.63
SSVP-SLT-LSP (Rust et al., 2024)	43.20	28.8	20.80	15.50	38.40	49.60
SignMusketeers (Gueuwou et al., 2025a)	41.50	27.20	19.30	14.30	—	—
Uni-Sign (Li et al., 2025)	40.20	—	—	14.90	36.00	49.40
SHuBERT (Gueuwou et al., 2025b)	—	—	—	**16.20**	—	49.90
Geo-Sign (Fish and Bowden, 2026)	40.80	—	—	15.10	35.40	—
AttnSign (Ours)	**43.45**	**28.82**	**20.94**	15.54	**39.02**	**50.44**

Table 1: Performance comparison on the How2Sign benchmark. Methods are grouped into four blocks: zero-shot closed-source VLMs, zero-shot open-source VLMs, specialized SLT methods using only How2Sign (H2S) dataset, and specialized SLT methods that additionally use YouTube-ASL (YT) dataset. B-n denotes BLEU-n. Our method uses InternVL-2.5-1B as the backbone.

Method	B-1	B-4	ROUGE-L	BLEURT
Closed-source VLMs				
Gemini-2.5-Pro (Comanici et al., 2025)	22.23	3.82	19.78	36.45
GPT-5	14.18	1.63	10.23	28.36
Open-source VLMs				
Qwen2.5-VL-7B (Bai et al., 2025)	11.38	0.55	6.30	24.71
InternVL-2.5-8B (Chen et al., 2024a)	12.96	0.38	6.09	21.15
Specialized SLT methods				
I3D-transformer (Shi et al., 2022)	18.31	5.66	18.64	28.82
OpenASL (Shi et al., 2022)	20.92	6.72	21.02	31.09
GloFE-VN (Lin et al., 2023)	21.56	7.06	21.75	36.35
C^2RL (Chen et al., 2025)	31.46	13.21	31.36	—
SHuBERT (Gueuwou et al., 2025b)†	—	**23.20**	—	60.60
Uni-Sign (Li et al., 2025)†	**49.35**	23.14	43.22	60.40
AttnSign (Ours)	**37.52**	**13.45**	**34.23**	**51.42**
AttnSign (Ours)†	47.74	20.75	**45.62**	**61.68**

Table 2: Performance comparison on the OpenASL benchmark. Methods are grouped into three blocks: zero-shot closed-source VLMs, zero-shot open-source VLMs, and specialized SLT methods. †: additionally uses YouTube-ASL data. B-n denotes BLEU-n. Our method uses InternVL-2.5-1B as the backbone.

4 Experiments

4.1 Datasets and Metrics

Datasets. We train and evaluate on two large-scale American Sign Language (ASL) benchmarks: How2Sign (H2S) (Duarte et al., 2021), about 80 hours of studio-recorded ASL, and OpenASL (Shi et al., 2022), about 288 hours ASL videos. For SFT scaling, we additionally incorporate YouTube-ASL (YT) (Uthus et al., 2023), a 984-hour web-scale corpus with weakly aligned English text translations. When augmenting OpenASL with YouTube-ASL, we remove all YouTube-ASL clips overlapping with any OpenASL split by video ID to prevent test leakage, excluding 620 test clips and 70,100 training clips. For GRPO training, we use only the in-domain training split of the target benchmark.

Evaluation metrics. Following previous work (Rust et al., 2024; Lin et al., 2023) we report BLEU (Papineni et al., 2002) computed with SacreBLEU (Post, 2018), ROUGE-L (Lin, 2004), and BLEURT (Sellam et al., 2020).

4.2 Implementation Details

Our backbone is InternVL-2.5-1B (Chen et al., 2024b). Each video frame is resized to 224×224 and frames are uniformly sampled over time at 16 FPS. Throughout training and inference, the model takes raw video and instructions as input.

SFT stage. We fully fine-tune the ViT and adapt the LLM with LoRA (Hu et al., 2021) at rank r=16. We use AdamW with a learning rate of 5×10^{-5} and cosine decay, training for up to 5 epochs with a

batch size of 64 on 8× NVIDIA A100 GPUs, taking approximately 12 hours in the H2S-only setting and 6 days in the full H2S+YT setting. For spatial attention supervision, we use weight γ=0.15, softmax temperature T=0.5, supervise layers 13 and 16 of LLM. Pose-based foreground masks are extracted only for How2Sign and OpenASL samples; YouTube-ASL samples are included in SFT with the cross-entropy loss only, and attention supervision is skipped on these samples.

RL stage. We start from the SA-SFT model, with group size G=4, clipping ϵ=0.2, and sampling temperature 0.7. As in the SFT stage, we fully fine-tune the ViT and adapt the LLM with LoRA at rank r=16. Training runs for 2 epochs on the 10K in-domain samples take approximately 8 hours. For How2Sign evaluation, we use the How2Sign training split, and for OpenASL evaluation, we use the OpenASL training split. We use BLEU-1 and ROUGE-L as the rule-based rewards with equal weight. The motion-cadence attention constraint uses λ=0.001 and is applied at LLM layers 17 and 19.

4.3 Main results

Tables 1 and 2 report our main results on How2Sign and OpenASL. AttnSign achieves competitive performance under both single-source and +YT training settings, and consistently leads on the semantic metrics ROUGE-L and BLEURT. Compared with SSVP-SLT (Rust et al., 2024), which uses the same training settings but adds a self-supervised pretraining stage on YT, AttnSign exceeds it without that stage, reaching BLEU-4 of 9.45 on H2S-only compared to 7.00, and 15.54 on H2S+YT compared to 15.50. On the semantic metrics ROUGE-L and BLEURT, which better reflect translation quality beyond exact n-gram overlap, AttnSign leads in every evaluation block, suggesting that a raw-video, attention-supervised design can generalize semantically without external feature pipelines.

4.4 Ablation Study

We validate the contribution of each pipeline component. Throughout this section, SFT is trained on How2Sign and YouTube-ASL, GRPO is trained on 10K How2Sign in-domain samples, and evaluation is on the How2Sign. Appendix J reports the same ablation on OpenASL with matching gains.

Effect of each component. Table 3 reports the contribution of each pipeline component. Instructional SFT raises performance over the zero-shot baseline, and spatial attention supervision in SA-SFT further improves BLEU-4 from 13.89 to 14.55 over plain SFT. Within the RL stage, removing the KL term of GRPO improves BLEU-4 from 15.04 to 15.38, indicating that anchoring the policy to the SFT reference limits beneficial updates. Combining motion-cadence steering with the KL penalty (w/ KL + cadence) reaches 15.13. SA-GRPO, which replaces the KL penalty with motion-cadence steering, reaches 15.54. This further outperforms GRPO w/o KL by +0.16 BLEU-4, indicating that cadence steering provides task-aligned guidance beyond what removing KL alone offers. This gain is consistent across 3 random seeds (SA-GRPO 15.54 ± 0.03 vs. GRPO w/o KL 15.38 ± 0.03), well above the per-seed variability. This supports replacing the KL penalty with cadence steering rather than combining the two.

Method	B-1	B-4	ROUGE-L	BLEURT
Zero-shot baseline	10.79	0.08	6.98	23.87
SFT stage				
SFT	41.23	13.89	37.13	49.02
SA-SFT (Ours)	41.94	14.55	37.67	49.55
RL stage				
GRPO (w/ KL)	42.74	15.04	38.47	50.01
GRPO (w/o KL)	43.22	15.38	38.86	50.30
GRPO (w/ KL + cadence)	42.87	15.13	38.62	50.13
SA-GRPO (Ours)	**43.45**	**15.54**	**39.02**	**50.44**

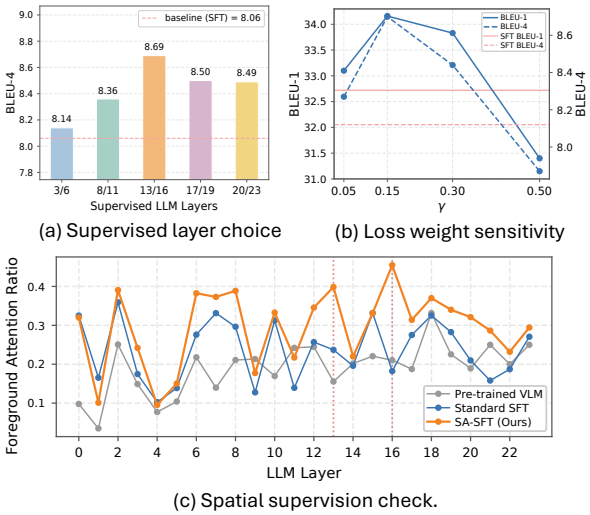
Table 3: Component ablation on How2Sign. We isolate the contribution of each stage in our pipeline: the SFT stage block compares standard SFT with our SA-SFT, and the RL stage block compares variants of GRPO. Combining SA-SFT and SA-GRPO yields the best results across all metrics.

4.5 Model Discussion

All analyses in this section are evaluated on How2Sign. Unless otherwise stated, the training data is also How2Sign only.

Layer choice for spatial supervision. We vary the supervised layer set $\mathcal{S}_{\text{attn}}$ while fixing γ at its default. Figure 4(a) sweeps supervised layer pairs spanning the LLM depth. Mid-depth supervision at the 13th and 16th layers produces the strongest gains, raising BLEU-4 from the baseline 8.06 to 8.69, while shallow, upper-middle, and final supervision yield smaller or no improvement. Mid-depth supervision consistently outperforms shallow or final layers across all metrics, supporting our choice of the 13th and 16th layers as default.

Spatial supervision verification and sensitivity. Fixing the supervised layers at the 13th and 16th,



(a) Supervised layer choice (b) Loss weight sensitivity

(c) Spatial supervision check.

Figure 4: SA-SFT analysis. (a) Layer choice: sweep over supervised layer pairs spanning the LLM depth; mid-depth performs best in BLEU-4. (b) Spatial attention loss weight γ sensitivity. (c) Per-layer foreground attention ratio: SA-SFT raises mass on foreground tokens at the supervised 13th and 16th layers over standard SFT and the pre-trained VLM.

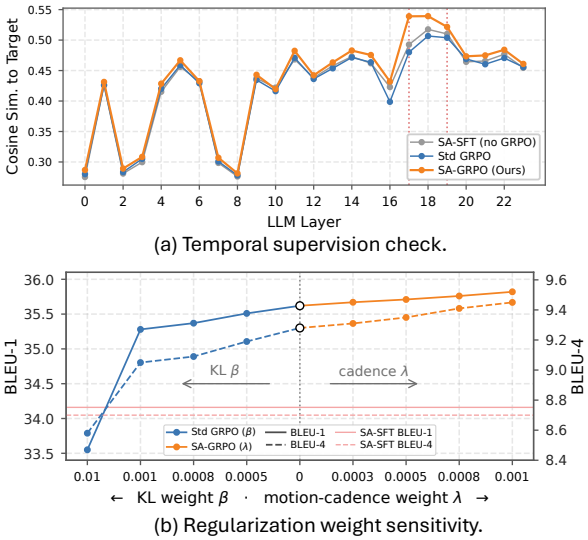
Reward	B-1	B-4	ROUGE-L	BLEURT
SA-SFT (baseline)	34.16	8.69	29.31	42.21
BLEU-1	35.03	9.18	30.21	42.42
BLEU-1 + BLEU-4	34.78	8.98	29.89	42.33
BLEU-1 + ROUGE-L	**35.68**	**9.28**	**30.33**	**42.62**
BLEU-1 + BLEU-4 + ROUGE-L	35.47	9.22	30.25	42.53

Table 4: Rule-based reward combinations for standard GRPO. The selected reward is reused in SA-GRPO for our main experiments.

Figure 4(b) sweeps the loss weight γ: the curve shows a clear interior optimum at γ=0.15, with smaller values weakening supervision and larger values over-regularizing.

Figure 4(c) reports the foreground attention ratio—the fraction of per-frame attention mass falling inside the pose-derived foreground mask—across LLM layers. SA-SFT raises this ratio at the supervised 13th and 16th layers compared with standard SFT and the pre-trained VLM, confirming that the supervision actually shifts attention toward the foreground regions.

GRPO reward design. Table 4 compares four rule-based reward combinations under standard GRPO, isolating reward design from the motion-cadence steering. BLEU-1 alone improves over the SA-SFT baseline. Adding BLEU-4 consistently degrades performance, suggesting that the sparse 4-gram signal is noisy at this scale. BLEU-1+ROUGE-L performs best across all metrics, and is adopted as the default reward in both standard GRPO and SA-GRPO for our main experiments.



(a) Temporal supervision check.

(b) Regularization weight sensitivity.

Figure 5: (a) Per-layer cosine similarity between attention and the motion-cadence target: SA-GRPO raises alignment at the supervised layers over the SA-SFT model and the standard GRPO model. (b) Sensitivity to the regularization weight; the shared center is the no-regularization baseline. Left: GRPO with KL weight β. Right: SA-GRPO with motion-cadence weight λ.

SA-GRPO analysis. Figure 5(a) verifies SA-GRPO's mechanism: cosine alignment between attention and the motion-cadence target rises at the supervised 17th and 19th layers, which are the layers identified by the statistical analysis, while GRPO w/o KL remains close to the SA-SFT model. This indicates that the alignment is a specific effect of motion-cadence steering, not a generic outcome of RL training.

Figure 5(b) contrasts the two forms of attention control from a shared center–GRPO without KL penalty or cadence term. For standard GRPO, any nonzero KL weight β reduces BLEU-4, and the largest weight β=0.01 drops the BLEU-4 by nearly a full point below the center baseline; this confirms that anchoring the policy to π_{ref} limits how much the policy can improve under our setting. Replacing the KL penalty with our motion-cadence steering reverses the pattern: BLEU-4 increases monotonically with λ across the swept range, peaking at λ=0.001, indicating that sign-specific steering carries useful signal.

SFT Data Scaling. Figure 6 shows that increasing the SFT training data, starting from the 80h H2S-only setting and progressively adding YT data, yields substantial gains in the early scale-up regime, while further scaling at the larger end yields only marginal additional improvement.

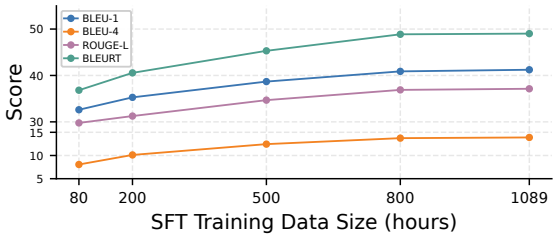


Figure 6: Effect of SFT training data size on the H2S, measured by total video duration. The 80h setting uses H2S only; larger sizes are obtained by adding YT data.

Method	BLEU-4	ROUGE-L	BLEURT
OOD transfer on How2Sign			
Uni-Sign (Li et al., 2025)	6.13	21.28	35.18
AttnSign (Ours)	**9.29**	**29.19**	**41.74**
In-the-wild evaluation on 100 YouTube videos			
Uni-Sign (Li et al., 2025)	2.13	16.28	31.18
AttnSign (Ours)	**5.01**	**19.19**	**35.74**

Table 5: Generalization comparison against Uni-Sign (Li et al., 2025) under two evaluation settings: (i) out-of-domain transfer, training on OpenASL+YT and testing on the H2S test set; (ii) in-the-wild evaluation on 100 randomly collected ASL clips from YouTube that are not present in H2S, OpenASL, or YT dataset, using their original English captions as ground truth.

Generalization. Beyond in-distribution performance, we assess whether AttnSign generalizes to unseen data. Table 5 evaluates AttnSign against Uni-Sign (Li et al., 2025) under two settings: (i) out-of-domain (OOD) transfer, where both methods are trained on OpenASL+YT and evaluated on the H2S test set; and (ii) in-the-wild evaluation on 100 randomly collected ASL clips from YouTube, with no overlap against H2S, OpenASL, or the YouTube-ASL training corpus. AttnSign outperforms Uni-Sign on all three metrics under both settings, suggesting stronger generalization to both unseen distributions and in-the-wild videos.

5 Conclusion

We presented AttnSign, an end-to-end training framework that adapts a general-purpose VLM to SLT in two stages. In the SFT stage, we introduce a spatial attention supervision that guides the model to focus on sign-relevant regions in each frame; in the RL stage, we add a motion-cadence steering mechanism in place of the KL penalty in standard GRPO, guiding the model's temporal attention toward keyframes of each sign. AttnSign achieves competitive performance on both How2Sign and OpenASL, showing that explicitly steering visual attention toward *where* and *when* sign-linguistic content occurs is a useful strategy for adapting a general-purpose VLM to SLT. We hope these findings encourage exploring attention-level supervision for other video-language tasks.

Limitations. Our experiments focus on American Sign Language (ASL); extending the proposed framework to other sign languages is left for future work. In addition, while attention steering yields strong empirical gains, a theoretical derivation of why this form of supervision benefits translation quality remains an open question that warrants further study.

Acknowledgements

This work was supported in part by the National Science Foundation under Grant No. 2502050, the National Institutes of Health under Award No. R16GM159146, and the NVIDIA Academic Grant Program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Diane Brentari and Petra Eccarius. 2010. *Handshape contrasts in sign language phonology*. na.

Geoffrey Brookshire, Jenny Lu, Howard C. Nusbaum, Susan Goldin-Meadow, and Daniel Casasanto. 2017. Visual cortex entrains to sign language. *Proceedings of the National Academy of Sciences*, 114(24):6352–6357.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*.

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. 2025. C 2 rl: Content and context representation learning for gloss-free sign language translation and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *CVPR*.

Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2026. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137.

Edward Fish and Richard Bowden. 2026. Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation. *Advances in Neural Information Processing Systems*, 38:99293–99330.

Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. 2023. Towards real-time sign language recognition and translation on edge devices. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4502–4512.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *CVPR*.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025a. Signmusketeers: An efficient multi-stream approach for sign language translation at scale. In *ACL*.

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 *ACL*.

Jianyuan Guo, Peike Li, and Trevor Cohn. 2026. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation. *Advances in Neural Information Processing Systems*, 38:77471–77499.

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.

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*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C Park. 2024. An efficient sign language translation using spatial configuration and motion dynamics with llms. *arXiv preprint arXiv:2408.10593*.

Mert Inan, Anthony Sicilia, and Malihe Alikhani. 2025. Signalignlm: Integrating multimodal sign language processing into large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3691–3706.

Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025. Lost in translation, found in context: Sign language translation with contextual cues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8742–8752.

Yiyang Jiang, Li Zhang, Xiao-Yong Wei, and Li Qing. 2026. Think in latent thoughts: A new paradigm for gloss-free sign language translation. *arXiv preprint arXiv:2604.15301*.

Jungeun Kim, Hyeongwoo Jeon, Jongseong Bae, and Ha Young Kim. 2025. Leveraging the power of mllms for gloss-free sign language translation. In *ICCV*.

Alexander C Li, Yuandong Tian, Beidi Chen, Deepak Pathak, and Xinlei Chen. 2024a. On the surprising effectiveness of attention transfer for vision transformers. *Advances in Neural Information Processing Systems*, 37:113963–113990.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024b. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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

Han Liang, Chengyu Huang, Yuecheng Xu, Cheng Tang, Weicai Ye, Juze Zhang, Xin Chen, Jingyi Yu, and Lan Xu. 2024. Llava-slt: Visual language tuning for sign language translation. *arXiv preprint arXiv:2412.16524*.

Scott K. Liddell and Robert E. Johnson. 1989. American sign language: The phonological base. *Sign Language Studies*, 64(1):195–277.

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

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. *arXiv preprint arXiv:2305.12876*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Neurips*, 36.

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.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, and 1 others. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Roland Pfau, Josep Quer, and 1 others. 2010. *Nonmanuals: Their grammatical and prosodic roles*. na.

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.

Muxin Pu, Xiao-Ming Wu, Mei Kuan Lim, Chun Yong Chong, Wei Li, and Chen Change Loy. 2026. Signdpo: Multi-level direct preference optimisation for skeleton-based gloss-free sign language translation. *arXiv preprint arXiv:2604.18034*.

Zhi Rao, Yucheng Zhou, Benjia Zhou, Yiqing Huang, Sergio Escalera, and Jun Wan. 2025. Rvlf: A reinforcing vision-language framework for gloss-free sign language translation. *arXiv preprint arXiv:2512.07273*.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. *arXiv preprint arXiv:2402.09611*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *ACL*.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. In *EMNLP*.

Guohao Sun, Can Qin, Yihao Feng, Zeyuan Chen, Ran Xu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2025. Structured policy optimization: Enhance large vision-language model via self-referenced dialogue. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 741–751. IEEE.

Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. 2024. Sq-llava: Self-questioning for large vision-language assistant. In *European Conference on Computer Vision*, pages 156–172. Springer.

Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. 2023. Sign language translation from instructional videos. In *CVPR*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Neurips*, 36.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. 2024. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pages 198–215. Springer.

Jiamian Wang, Kunpeng Li, Yulun Zhang, Xin Yuan, and Zhiqiang Tao. 2025a. s^2-transformer for mask-aware hyperspectral image reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4299–4316.

Jiamian Wang, Ziqi Zhou, Chaithanya Kumar Mummadi, Sohail Dianat, Majid Rabbani, Raghuveer Rao, Chen Qiu, and Zhiqiang Tao. 2025b. Visual self-refinement for autoregressive models. *arXiv preprint arXiv:2510.00993*.

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*.

Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. 2024. Voila-a: Aligning vision-language models with user's gaze attention. *Advances in neural information processing systems*, 37:1890–1918.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Xuechi Ma, Dawei Sheng, Yuxuan Tong, Chi Yuan, Hao Zhou, Minging Wang, Weinan Li, and 5 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

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

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. *Neurips*, 37.

Rui Zhao, Xuewen Zhong, Xiaoyun Zheng, Jinsong Su, and Yidong Chen. 2026. Cnsl-bench: Benchmarking the sign language understanding capabilities of mllms on chinese national sign language. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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 *ICCV*.

A Training Instructions

We use diverse instruction phrasings during instruction tuning to improve robustness to prompt variation. Examples include:

(1) "Translate the American Sign Language in this video to English."
(2) "What is being signed in this video? Provide the English translation."
(3) "Watch the sign language video and write the English sentence."

B Further Discussion on the Vision Encoder

ViT Setting	B-1	B-4	ROUGE-L	BLEURT
Freeze	20.45	1.79	16.89	30.11
Half (top-6)	22.85	3.87	18.97	31.20
Full (Ours)	**32.61**	**8.06**	**28.47**	**36.82**

Table 6: ViT freezing strategy on H2S. Fully tuning the ViT is critical for adapting to the sign-language visual distribution.

During the SFT stage, we study how much of the ViT should be trained in our task, comparing three strategies: keeping the entire ViT frozen, unfreezing only its top six transformer layers, and fully fine-tuning all ViT parameters jointly with the LLM. All variants share identical training data and hyperparameters; only the ViT trainable parameters differ. Table 6 shows that full fine-tuning of the ViT is essential for SLT. A fully frozen ViT yields only 1.79 BLEU-4, indicating that the pretrained general-image features are far from sufficient for the fine-grained hand and facial cues in signing videos. Partially unfreezing the top six layers improves BLEU-4 to 3.87 but still falls short of the fully fine-tuned setting with BLEU-4 8.06. We therefore fully fine-tune the ViT alongside the LLM throughout all of our experiments.

C Further Discussion on Spatial Attention Loss

Attention loss	B-1	B-4	ROUGE-L	BLEURT
SFT (baseline)	32.61	8.06	28.47	36.82
+ Background Penalty	32.35	7.81	27.42	36.47
+ KL Loss (Ours)	**34.16**	**8.69**	**29.31**	**42.21**

Table 7: Comparison of two attention supervision losses.

In the SA-SFT stage, our spatial attention supervision uses a KL loss via a soft foreground target. An alternative is a direct background penalty that directly suppresses attention mass outside the foreground mask:

$$\mathcal{L}_{\text{bg}} = \sum_{i \notin \text{FG}} P^{(l)}_{\text{attn},i}. \tag{15}$$

The background penalty has a degenerate solution: the loss can be minimized by placing all attention on few foreground tokens, which discards information from all other foreground regions. We observe this collapse at the supervised layers, which end up receiving little visual signal. As a result, the background penalty underperforms even plain SFT, dropping BLEU-4 from 8.06 to 7.81. In contrast, the KL loss matches the entire attention distribution against the soft target $\hat{P}_k$, which prevents collapse onto a single token and keeps attention distributed across the hands and face, yielding better performance.

D Foreground Mask Generation Details

We use MediaPipe (Lugaresi et al., 2019) to construct per-frame foreground masks for both How2Sign and OpenASL training sets. MediaPipe has been widely adopted in recent SLT works (Gueuwou et al., 2025b,a). MediaPipe's FaceDetection and Hands modules are widely used for their accurate, real-time detection of face and hand regions in videos, with reliable performance across diverse signers and camera setups. For each frame, we extract bounding boxes for the face and both hands and build a binary mask at the original frame resolution by marking all pixels inside these boxes as foreground. The mask is then average-pooled to the 8×8 visual-token grid and stored on disk before training. When MediaPipe fails to produce a valid mask for a video, we skip the spatial attention loss for that sample and retain only the cross-entropy term.

E Further Discussion on Softmax Temperature

In the SA-SFT stage, the softmax temperature T controls how sharply the spatial target $\hat{P}_k$ concentrates on foreground tokens: smaller T produces a sharper target focused on the hand and face regions, while larger T flattens the target toward a uniform distribution over all tokens. We sweep different temperature values on H2S, with the loss weight fixed at γ=0.15 (Figure 7). T=0.5 achieves the

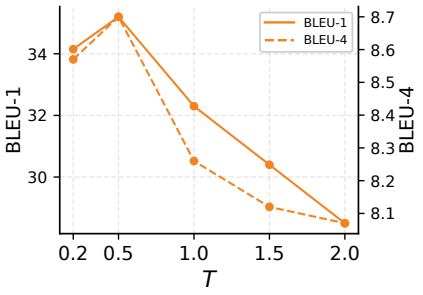


Figure 7: Effect of varying the softmax temperature T in the spatial attention target on H2S during SA-SFT, with γ=0.15 fixed.

best BLEU-4 of 8.69, with T=0.2 close behind, indicating that a relatively sharp foreground target is effective. As T increases beyond 0.7, performance drops back toward the baseline level (8.06): the target becomes too flat to provide useful guidance and the attention supervision becomes ineffective. We therefore use T=0.5 in all main experiments.

F Spatial Attention Visualization on OpenASL

To verify that the spatial attention supervision generalizes beyond H2S, we visualize attention maps on an OpenASL test video. Figure 8 mirrors the layout of Figure 2. As on H2S, SA-SFT concentrates attention on the signer's hands and face, while the pre-trained VLM and standard SFT model fail to focus, indicating that the spatial steering effect generalizes to OpenASL.

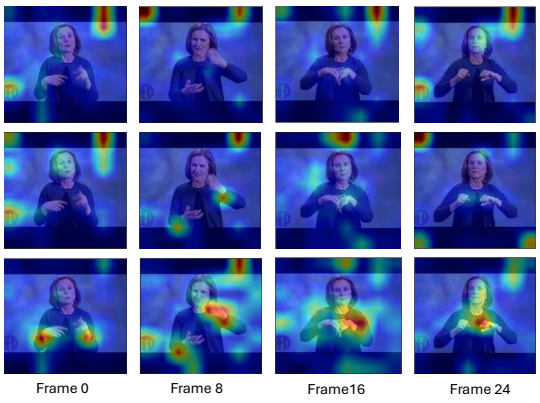


Figure 8: Visual-token attention maps at LLM layer 16 across four frames (columns) of an OpenASL test video. Rows show the three training stages: pre-trained VLM (top), standard SFT (middle), and Sign-Aware SFT with our spatial attention supervision (bottom).

G Further Discussion on RL Training Data

In the RL stage, we sweep the number of in-domain training samples across 5K, 10K, 15K, and 30K under standard GRPO (without our motion-cadence steering), to isolate the effect of data size from other design choices. Figure 9 shows that performance improves from 5K to 10K and then plateaus, with 15K and 30K offering no additional gain over 10K. We therefore use 10K samples as our default GRPO setting; additional samples beyond this size yield no improvement and only increase training cost.

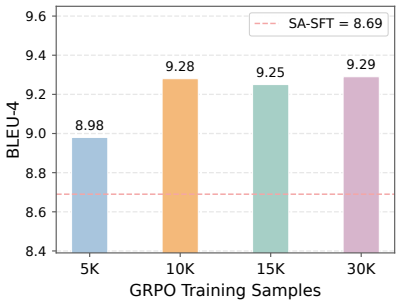


Figure 9: Effect of GRPO training set size on H2S, evaluated under standard GRPO. Performance plateaus beyond 10K samples.

H Motion-Cadence Target Ablation

In the SA-GRPO stage, the motion-cadence target t_k is built from detected motion peaks plus Gaussian smoothing. We conduct two checks: (i) whether the gain comes from content-aligned peak positions rather than an arbitrary target, and (ii) whether the model is sensitive to the specific motion-peak extractor. For (i), we replace the detected peaks with random peak positions; for (ii), we compare our default frame-difference extractor with optical flow (Farnebäck, 2003). The rest of the cadence loss (KL form, Gaussian smoothing, peak count, weight λ) is kept unchanged.

Target source	B-1	B-4	ROUGE-L	BLEURT
GRPO w/o KL	35.62	9.28	30.07	42.62
random peaks	35.59	9.21	29.84	42.59
optical flow	—	9.43	30.41	42.68
frame difference (Ours)	**35.82**	**9.45**	**30.45**	**42.73**

Table 8: Effect of the motion-cadence target source on H2S. Random peaks share the same number and Gaussian smoothing as the detected peaks; only the alignment with video content is removed. Optical flow and frame difference are two alternative motion-peak extractors.

Random-peak target shows no gain over the GRPO w/o KL baseline (BLEU-4 9.21 vs. 9.28), indicating that adding a cadence loss with an arbitrary target brings no benefit on its own. This shows that the gain comes from content-aligned peak positions specifically, not from adding any attention regularization in the RL stage. Switching between frame difference and optical flow yields near-identical results across all metrics (BLEU-4 9.45 vs. 9.43), showing that the method is not sensitive to the specific motion-peak extractor. Consistent with this, the peaks selected by the two extractors are highly overlapping, with a consensus rate of 91% within ± 3 frames, 79% at ± 2 frames, and 65% at ± 1 frames.

I Further Discussion on Supervised Layers for SA-GRPO

In the SA-GRPO stage, the statistical analysis selects the 17th and 19th layers based on their positive concentration-quality correlation. To check that the probe genuinely predicts the best layers, we move the motion-cadence steering to alternative layer groups spanning the network depth, where the analysis shows weak or lower correlation.

Layer	Corr. r	B-4	Δ vs. w/o KL
5/7	~ 0	9.19	-0.09
11/13	$+0.15$	9.30	$+0.02$
17/19 (Ours)	$+0.32$	**9.45**	$\mathbf{+0.17}$
21/23	$+0.20$	9.36	$+0.08$

Table 9: Layer-selection falsification for SA-GRPO. Corr. r denotes the concentration–quality correlation from Fig. 4; Δ vs. w/o KL is the BLEU-4 gain over the GRPO w/o KL baseline.

BLEU-4 gains monotonically track the concentration–quality correlation r: the 5/7 layers ($r \approx 0$) even underperform the baseline, whereas the 17/19 layers ($r = +0.32$) yield the largest gain ($+0.17$). This confirms the probe is a predictive signal for layer selection, and the causal relationship does not appear at randomly chosen layer combinations.

J Ablation Study on OpenASL

To verify that the component-wise gains observed on H2S are not specific to that dataset, we run the same ablation on OpenASL. SFT is trained on OpenASL and YT, GRPO is trained on 10K OpenASL in-domain samples, and evaluation is on the OpenASL test set. We reuse the same supervised layers as on H2S to test cross-dataset transferability.

Method	B-1	B-4	ROUGE-L	BLEURT
Zero-shot baseline	11.23	0.35	5.67	19.76
SFT stage				
SFT	45.67	18.92	41.85	58.76
SA-SFT	46.89	19.76	43.53	60.23
GRPO stage				
GRPO (w/ KL)	47.03	20.04	44.75	60.98
GRPO (w/o KL)	47.51	20.55	45.34	61.56
GRPO (w/ KL + cadence)	47.12	20.22	45.26	61.31
SA-GRPO (Ours)	**47.74**	**20.75**	**45.62**	**61.68**

Table 10: Component ablation on OpenASL. The pattern of gains mirrors that on H2S (Table 3): SA-SFT improves over plain SFT, removing the KL penalty further improves over standard GRPO, and SA-GRPO yields the best results across all metrics.

The component-wise gain pattern on OpenASL matches that on How2Sign, confirming that the contributions of SA-SFT and SA-GRPO generalize across both benchmarks rather than reflecting How2Sign-specific tuning.

K Qualitative Results

We show qualitative translation examples on How2Sign (Table 11) and OpenASL (Table 12), with each example reporting the ground-truth reference alongside the output of AttnSign. AttnSign successfully captures the semantic information in most examples and generates sentences close to the ground-truth references, while still struggling on more complex cases, as shown in the last example of each table.

Reference:	I'm CarolAnn with Studio Group X and just do those push ups every day and you'll improve.
AttnSign:	I'm Carol Ann with Studio Group X and we just did those pushups every day and you will improve.
Reference:	It has a bright color due to the fruit punch.
AttnSign:	It has a bright color because of the pineapple.
Reference:	I practice with the Barton Oaks Dental Group.
AttnSign:	I practice with the Barton Oaks Dental Groups.
Reference:	To come out of it, soften the knees, take the hands away from your feet, and slowly roll it up, tuck in the tail bone and rolling up one vertebra at a time and come straight up.
AttnSign:	And then she comes out of the tension soft hands to the floor and slowly roll her knee and she loosens her bones as she rolls up into a lunge.

Table 11: Qualitative results on the How2Sign dataset.

Reference:	We have a lot of respect for our Canadian neighbors.
AttnSign:	We have a lot of respect for our neighbors.
Reference:	His family were skilled woodcarvers and sold totems around Seattle since 1909, passing down their carving skills.
AttnSign:	His family was skilled woodcarvers and sold to totem poles around Seattle area since 1909, as an art skills.
Reference:	Biden had 150,000 more votes than Bernie.
AttnSign:	Biden is ahead by 150,000 votes than Bernie.
Reference:	You need to invest your time and passion to assist delegates to understand why this topic is important to ensure your proposed priority successfully ranks.
AttnSign:	We want you to spend your time and make sure you understand your passion, your desires, and your work to be successful in proposing an idea and getting it passed.

Table 12: Qualitative results on the OpenASL dataset.