arXiv:2607.27614v1 [cs.CL] 30 Jul 2026

DualAnchor: Preserving Language Priors and Improving Lexical Fidelity in Gloss-Free Sign Language Translation

Hongbin Zhang[1,2*], **Junhao Liu**[1*], **Xuefeng Bai**[1], **Youcheng Pan**[2], **Yang Xiang**[2], **Kehai Chen**[1,2]

[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China

Abstract

Recent advances in large language models (LLMs) have led sign language translation (SLT)—the task of converting sign-language videos into spoken-language text—to increasingly adopt LLMs as textual backbones. However, despite LLMs' strong language modeling capabilities, existing LLM-based SLT methods often undermine rather than exploit this language prior, producing disfluent translations—a failure we term *language-prior degradation*. Meanwhile, existing methods typically align videos and text at the sentence level; such alignment does not ensure accurate lexical details, creating a *lexical fidelity gap*. To address both issues, we propose **DualAnchor**, a gloss-free LLM-based SLT training framework that couples two complementary anchors for linguistically fluent and visually faithful generation: (i) **Token-level Prior Anchoring (TPA)** preserves the LLM's language prior by regularizing the multimodal decoder, at each decoding step, toward the next-token distribution of a frozen LLM conditioned on the same autoregressive prefix. (ii) **Optimal Transport Alignment (OTA)** improves lexical fidelity by formulating visual–textual matching as entropy-regularized partial optimal transport, with Sinkhorn optimization inducing a soft alignment between visual tokens and textual content tokens under a cosine cost. DualAnchor achieves strong overall performance on both PHOENIX-2014T and CSL-Daily. Targeted analyses link the gains to TPA's fluency improvement and OTA's reduction of fine-grained lexical errors.

Introduction

Sign Language Translation (SLT) translates sign-language videos into spoken-language text by recovering linguistic content from continuous visual signals and expressing it in target-language grammar (Camgoz et al. 2018; Camgöz et al. 2020; Yin and Read 2020). Recent advances in LLMs have led gloss-free SLT methods to adopt a generation framework in which the model encodes the sign video, maps the resulting visual representation into the LLM's language space, and uses the pretrained LLM as the textual backbone for translation (Wong, Camgoz, and Bowden 2024; Gong et al. 2024; Chen et al. 2024; Hwang et al. 2025). Within this framework, cross-modal alignment is usually learned at the sentence level through contrastive objectives over paired video and text representations or transferred semantic similarity (Zhou et al. 2023; Yin et al. 2023; Jiang et al. 2024).

However, this paradigm leaves two complementary issues unresolved. Despite the strong language modeling capabilities of their LLM backbones, existing methods often undermine rather than exploit the pretrained language prior during visual adaptation. Multimodal adaptation can shift the decoder's next-token distribution away from the linguistic structure acquired during text pretraining and reduce output fluency; we term this issue *language-prior degradation*. Sentence-level objectives align an entire sign video with its paired sentence, but they do not directly tie individual content words to local sign evidence. A model can therefore achieve strong global video–text alignment while mistranslating actions, entities, or attributes; we call this discrepancy the *lexical fidelity gap*. The two issues affect complementary dimensions of translation quality: language-prior degradation weakens linguistic form, while the lexical fidelity gap compromises fine-grained content accuracy. Figure 1 provides an intuitive illustration. The upper output is ungrammatical despite its LLM decoder. The lower example shows strong global semantic alignment, yet replaces *refrigerator* and *apple* with visually unsupported *cabinet* and *orange*.

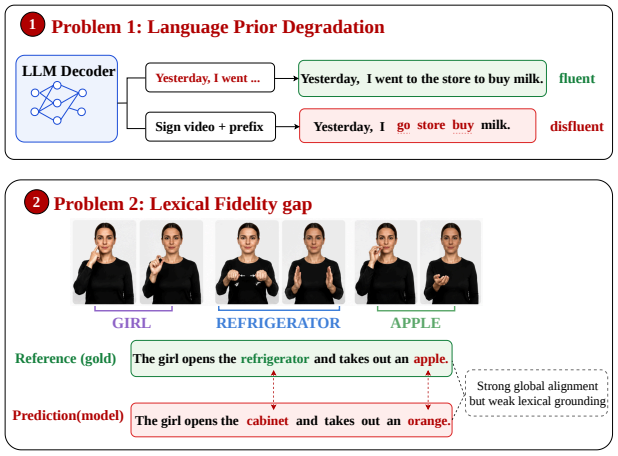


Figure 1: Motivating examples of the two issues in LLM-based SLT. Top: *language-prior degradation* produces disfluent text despite an LLM decoder. Bottom: strong global alignment can still coexist with weak lexical grounding, illustrating the *lexical fidelity gap*.

To address these issues, we propose DualAnchor, a gloss-free LLM-based SLT framework with two anchors:

*These authors contributed equally.

token-level prior anchoring and optimal transport alignment. **Token-level Prior Anchoring (TPA)** works in distribution space. Under the same autoregressive prefix, it regularizes the multimodal next-token distribution toward a frozen LLM prior and scales the constraint with the prior's confidence. This design preserves pretrained decoding behavior while still letting visual evidence influence uncertain positions. **Optimal Transport Alignment (OTA)** works in representation space. Entropy-regularized partial transport matches visual tokens to content tokens, while an unmatched sink absorbs unreliable correspondences. The partial formulation handles visual–lexical granularity mismatch without forcing every token to align. TPA preserves linguistic form, while OTA grounds lexical choice in fine-grained visual evidence.

We evaluate DualAnchor on PHOENIX-2014T (Camgoz et al. 2018) and CSL-Daily (Zhou et al. 2021). Across the two benchmarks, DualAnchor attains the best BLEU-4 among the compared gloss-free methods and the highest arithmetic mean across the ten reported metrics. Targeted analyses associate the gains with more fluent generation and fewer fine-grained lexical errors, supporting the intended roles of TPA and OTA over generic model capacity.

Our contributions are threefold:

- We identify two complementary LLM-based SLT issues: *language-prior degradation*, which hurts fluency during multimodal adaptation, and the *lexical fidelity gap*, where sentence-level alignment still leaves content-word errors.
- We introduce DualAnchor, coupling confidence-aware token-level prior anchoring with entropy-regularized partial transport to preserve the pretrained language prior and ground lexical choices in visual evidence.
- On PHOENIX-2014T and CSL-Daily, DualAnchor achieves the best BLEU-4 among the compared gloss-free methods, and targeted analyses show better fluency and fewer fine-grained lexical errors.

Preliminary Analysis

To characterize the two issues that motivate our method, we conduct two preliminary diagnostics. We first examine whether multimodal adaptation is accompanied by an output-level fluency gap and increased divergence from the pretrained language prior. We next ask whether high global video–reference compatibility can coexist with lexical errors.

Diagnostic I: Does multimodal adaptation coincide with an output-level fluency gap and greater prior divergence? Given a sign-language video $V_i = \{v_{i,t}\}_{t=1}^{T_i}$ and its spoken-language reference $Y_i = \{y_{i,n}\}_{n=1}^{N_i}$, the controlled baseline generates $\widehat{Y}_i \sim p_\theta(Y \mid V_i)$ autoregressively. We evaluate its hypotheses on the held-out sets of CSL-Daily (Chinese) (Zhou et al. 2021) and PHOENIX-2014T (German) (Camgoz et al. 2018). We measure fluency with a frozen external language model q_ϕ, using Baichuan2-7B (Yang et al. 2023) for Chinese and Qwen2.5-7B (Yang et al. 2024) for German. For a tokenized sentence $S = \{s_n\}_{n=1}^{|S|}$, we compute its perplexity and paired log-space fluency gap as

$$\mathrm{PPL}_\phi(S) = \exp\left(-\frac{1}{|S|}\sum_{n=1}^{|S|}\log q_\phi(s_n \mid s_{<n})\right), \quad (1)$$

$$\Delta_i^{\mathrm{flu}} = \log \mathrm{PPL}_\phi(\widehat{Y}_i) - \log \mathrm{PPL}_\phi(Y_i). \quad (2)$$

Lower PPL indicates higher target-language likelihood, while $\Delta_i^{\mathrm{flu}} > 0$ means that the hypothesis is less probable than its paired reference under the same evaluator.

Figure 2(a) reveals a pronounced output-level fluency gap on CSL-Daily: 81.9% of the adapted baseline's hypotheses have higher PPL than their paired references, and the mean PPL differs by $3.4\times$ (335.8 versus 98.3).

We next examine whether this output-level fluency gap co-occurs with increased divergence from the pretrained language prior by tracing the student-to-prior KL divergence throughout multimodal training. At epoch e, let $p^s_{e,i,n}$ denote the baseline's next-token distribution and $p^p_{i,n}$ the frozen prior distribution under the same gold autoregressive prefix. We average reverse KL over all valid target positions Ω:

$$D_e^{\mathrm{prior}} = \frac{1}{|\Omega|}\sum_{(i,n)\in\Omega} D_{\mathrm{KL}}\big(p^s_{e,i,n} \,\|\, p^p_{i,n}\big). \quad (3)$$

Larger values indicate greater deviation from the frozen language prior under synchronized decoding contexts.

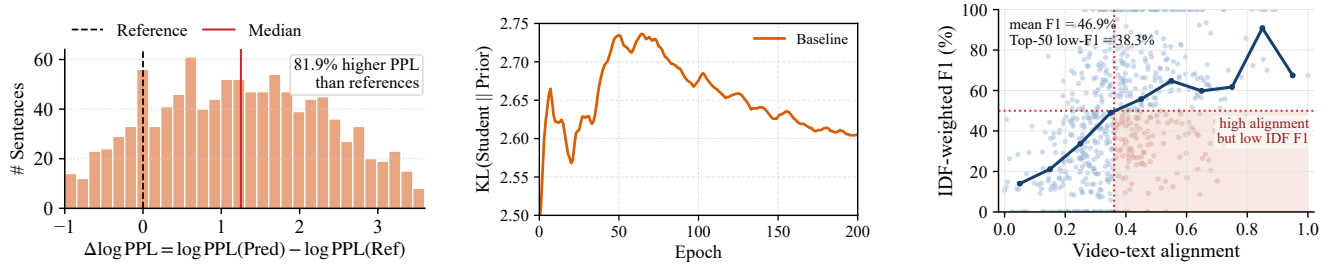
Figure 2(b) shows that the KL divergence increases from 2.50 to a peak of 2.77 near epoch 65 and stabilizes around 2.63 after epoch 220, remaining above its initial level.

Finding I. Multimodal adaptation is associated with an output-level fluency gap and increased prior divergence: the adapted baseline's hypotheses receive substantially higher external-LM PPL than their paired references, while student-to-prior KL increases during training and remains above its initial level. This coupled pattern motivates preserving the pretrained language prior during visual adaptation.

Diagnostic II: Can high global video–reference compatibility coexist with lexical translation errors? Using the same baseline outputs and held-out samples as in Diagnostic I, we examine whether gold video–reference pairs that are well matched at the sentence level can still have corresponding hypotheses with weak content-word fidelity. Let g_v and g_t denote the video and text encoders used by the sentence-level retrieval diagnostic. For each gold video–reference pair, we measure global compatibility as $s_i^{\mathrm{sent}} = \mathrm{sim}(g_v(V_i), g_t(Y_i))$, where larger values indicate stronger sentence-level compatibility between the video and its reference translation in the diagnostic representation space. To measure fine-grained lexical fidelity, let $K_i^h = K(\widehat{Y}_i)$ and $K_i^r = K(Y_i)$ denote the normalized content-word sets of the hypothesis and reference, respectively, and define $W(A) = \sum_{w\in A}\mathrm{idf}(w)$. The IDF-weighted content-word F1 is

$$F_{1,i}^{\mathrm{idf}} = \frac{2W(K_i^h \cap K_i^r)}{W(K_i^h) + W(K_i^r)}. \quad (4)$$

The resulting score measures content-word agreement between the baseline hypothesis and its reference. IDF assigns



(a) CSL-Daily: fluency gap **(b)** CSL-Daily: prior divergence **(c)** PHOENIX-2014T: global alignment vs. lexical fidelity

Figure 2: Preliminary diagnostics of prior drift and lexical fidelity. (a) On CSL-Daily, 81.9% of the adapted baseline's hypotheses have higher frozen-LM PPL than their paired references. (b) The student-to-prior KL divergence increases during multimodal adaptation and remains above its initial level. (c) On PHOENIX-2014T, 31.7% of baseline hypotheses in the highest video–reference alignment quartile have IDF-weighted content-word F1 below 50%, showing that high global compatibility can coexist with lexical translation errors.

greater weight to rare and informative words, while F1 captures both missing reference content and unmatched content introduced by the hypothesis. Content-word-aware translation likewise treats lexical units as unequally important to sentence meaning (Chen et al. 2020a).

Figure 2(c) shows that lexical errors remain common even among globally well-matched gold video–reference pairs: within the highest alignment quartile, 31.7% of baseline hypotheses have IDF-weighted content-word F1 below 50%. Thus, fine-grained lexical errors can persist despite high global video–reference compatibility.

Finding II. High global video–reference compatibility can coexist with substantial lexical translation errors: even within the highest alignment quartile, 31.7% of baseline hypotheses have IDF-weighted content-word F1 below 50%. This coexistence motivates visual–textual alignment at a finer lexical granularity than sentence-level matching.

Methodology

Problem Formulation

Given a sign video $V = \{v_\ell\}_{\ell=1}^{T}$ and its spoken-language translation $Y = \{y_t\}_{t=1}^{N}$, an LLM-based SLT model maps V to decoder-compatible visual tokens,

$$\mathbf{Z}^v = F_\psi(V) = \{\mathbf{z}_i^v\}_{i=1}^{n}, \qquad \mathbf{z}_i^v \in \mathbb{R}^d, \tag{5}$$

where F_ψ comprises the backbone's visual encoder and modality projection. The multimodal student factorizes the translation probability autoregressively as

$$p_{\theta_s}(Y \mid V) = \prod_{t=1}^{N} p_{\theta_s}(y_t \mid \mathbf{Z}^v, y_{<t}), \tag{6}$$

and learns from the teacher-forced translation loss

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\Omega|} \sum_{(b,t)\in\Omega} \log p_{\theta_s}(y_{b,t} \mid \mathbf{Z}_b^v, y_{b,<t}), \tag{7}$$

where Ω contains the non-padding target positions in a mini-batch. Cross-entropy supervises the observed target but does not explicitly preserve the pretrained trajectory of next-token distributions or impose local geometry between visual evidence and lexical units. DualAnchor therefore adds two complementary training constraints (Figure 3): TPA anchors linguistic form in distribution space, and OTA anchors lexical content in representation space. Together, TPA and OTA target a translation's linguistic form and visual content.

Token-Level Prior Anchoring

Knowledge distillation transfers teacher behavior to sequence models and multilingual translation systems (Kim and Rush 2016; Sun et al. 2020). TPA adapts this principle to token-level preservation during multimodal training.

Prefix-synchronized language prior. TPA retains a frozen copy of the language model, p_{θ_p}, to preserve the generative structure acquired during pretraining. Let s_b be a context prompt and $\pi_{b,t} = [s_b; y_{b,<t}]$ the context-augmented gold autoregressive prefix. The student and prior share $\pi_{b,t}$ and the vocabulary; only the student observes the sign video:

$$p_{b,t}^{s}(k) = p_{\theta_s}(y_t = k \mid \mathbf{Z}_b^v, \pi_{b,t}), \tag{8}$$

$$p_{b,t}^{p}(k) = p_{\theta_p}(y_t = k \mid \pi_{b,t}). \tag{9}$$

English Context prepends an English task instruction; *Target Context*, used by default, prepends its target-language rendering. In both settings, s_b is fixed before decoding and the target-dependent part of $\pi_{b,t}$ is exactly $y_{b,<t}$; the prefix-only variant sets $s_b = \varnothing$. We freeze θ_p and detach its probabilities from gradient computation. The synchronized prefix isolates multimodal adaptation's effect at each next-token position.

Confidence-adaptive distribution anchoring. The frozen prior is not equally reliable at every position: strict imitation at uncertain positions can suppress useful visual evidence, whereas a sharp prior provides a stronger linguistic constraint. TPA measures this reliability with token entropy

$$\begin{aligned} H_{b,t} &= -\sum_{k\in\mathcal{V}} p_{b,t}^{p}(k) \log p_{b,t}^{p}(k), \\ \bar{H} &= \frac{1}{|\Omega|} \sum_{(b,t)\in\Omega} H_{b,t}, \end{aligned} \tag{10}$$

where $\mathcal{V}$ is the target vocabulary. TPA then applies the confidence gate $\omega_{b,t} = \lambda_0 \sigma\big(\gamma(\bar{H} - H_{b,t})\big)$, where λ_0 sets the

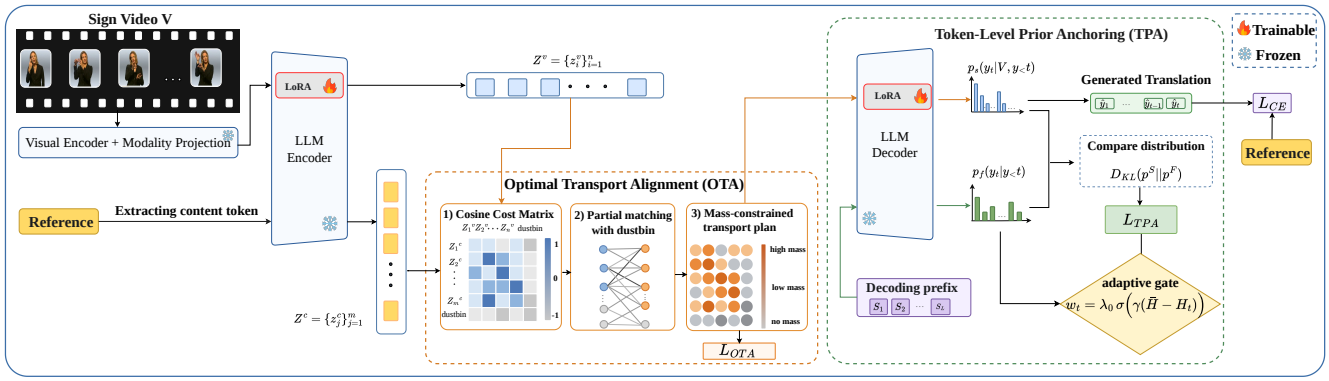


Figure 3: Overview of DualAnchor. TPA anchors the student next-token distribution to a frozen language prior under the same autoregressive prefix, while OTA learns partial correspondences between visual tokens and textual content tokens. Both modules operate only during training.

maximum anchoring scale, γ controls gate sharpness, and $\sigma(\cdot)$ is the sigmoid function. Low-entropy positions receive stronger prior guidance; high-entropy positions leave the student more freedom to follow the video.

TPA averages reverse KL over valid target positions:

$$\mathcal{L}_{\mathrm{TPA}} = \frac{1}{|\Omega|} \sum_{(b,t)\in\Omega} \omega_{b,t} D_{\mathrm{KL}}\left(p^{s}_{b,t} \,\|\, p^{p}_{b,t}\right). \tag{11}$$

This direction penalizes probability mass that the adapted student assigns to tokens unsupported by the frozen prior. TPA regularizes the predictive distribution rather than prescribing a decoded sentence, preserving linguistic structure while retaining visual conditioning.

Optimal Transport Alignment

Content-focused cross-modal geometry. Sentence-level alignment captures global video-text correspondence but leaves individual actions, entities, attributes, and temporal expressions ungrounded. OTA addresses this gap by aligning the visual tokens in Equation (5) with content-token representations $\mathbf{Z}^c = \{\mathbf{z}^c_j\}_{j=1}^m$, where $\mathbf{z}^c_j \in \mathbb{R}^d$. The content selector excludes padding, special symbols, punctuation, and weakly semantic function tokens so that non-lexical units do not dominate the alignment. After ℓ_2 normalization, the pairwise cosine cost is

$$D_{ij} = 1 - \frac{(\mathbf{z}^v_i)^\top \mathbf{z}^c_j}{\|\mathbf{z}^v_i\|_2 \, \|\mathbf{z}^c_j\|_2}, \qquad \mathbf{D} \in \mathbb{R}^{n\times m}. \tag{12}$$

The soft cost retains plausible correspondences because sign duration and text tokenization differ in granularity.

Entropy-regularized partial transport. Entropy regularization makes optimal transport differentiable and efficiently solvable by Sinkhorn scaling (Cuturi 2013), while partial transport permits unmatched mass when the two supports contain unreliable correspondences (Chapel, Alaya, and Gasso 2020). Let $\mathbf{a} = \frac{1}{n}\mathbf{1}_n$ and $\mathbf{b} = \frac{1}{m}\mathbf{1}_m$ denote uniform visual and textual masses, and let $\rho \in (0, 1]$ be the mass reserved for genuine cross-modal matching. We define the partial transport polytope

$$\Pi_\rho(\mathbf{a},\mathbf{b}) = \big\{\mathbf{P} \in \mathbb{R}^{n\times m}_+ : \mathbf{P}\mathbf{1}_m \le \mathbf{a},\ \mathbf{P}^\top \mathbf{1}_n \le \mathbf{b},\ \mathbf{1}_n^\top \mathbf{P} \mathbf{1}_m = \rho\big\}. \tag{13}$$

Introduce slack masses $\mathbf{s}^v = \mathbf{a} - \mathbf{P}\mathbf{1}_m$ and $\mathbf{s}^c = \mathbf{b} - \mathbf{P}^\top \mathbf{1}_n$, and set $\bar{\mathbf{P}} = \left[\begin{smallmatrix} \mathbf{P} & \mathbf{s}^v \\ (\mathbf{s}^c)^\top & 0 \end{smallmatrix}\right]$, $\bar{\mathbf{a}} = [\mathbf{a}; 1-\rho]$, and $\bar{\mathbf{b}} = [\mathbf{b}; 1-\rho]$. Let $\bar{\Pi}_\rho(\bar{\mathbf{a}}, \bar{\mathbf{b}})$ contain balanced couplings with these marginals and $\bar{P}_{n+1,m+1} = 0$. With $\bar{\mathbf{D}}$ equal to $\mathbf{D}$ on the real block and constant on dustbin edges, we solve

$$\bar{\mathbf{P}}^\star = \arg\min_{\bar{\mathbf{P}}\in\bar{\Pi}_\rho(\bar{\mathbf{a}},\bar{\mathbf{b}})} \langle \bar{\mathbf{P}}, \bar{\mathbf{D}} \rangle - \tau\, \mathcal{H}(\bar{\mathbf{P}}), \tag{14}$$

where $\tau > 0$ controls transport smoothness and $\mathcal{H}(\bar{\mathbf{P}}) = -\sum_{i,j} \bar{P}_{ij}(\log \bar{P}_{ij} - 1)$, with $0 \log 0 = 0$. Thus $\mathcal{H}(\bar{\mathbf{P}}) = \mathcal{H}(\mathbf{P}) + \mathcal{H}(\mathbf{s}^v) + \mathcal{H}(\mathbf{s}^c)$, explicitly regularizing unmatched capacity, while the real block transports exactly ρ mass. Let $\mathbf{M}$ be one except for $M_{n+1,m+1} = 0$; masked Sinkhorn scaling gives

$$\mathbf{K} = \mathbf{M} \odot \exp(-\bar{\mathbf{D}}/\tau), \qquad \bar{\mathbf{P}}^\star = \mathrm{diag}(\mathbf{u})\, \mathbf{K}\, \mathrm{diag}(\mathbf{r}), \tag{15}$$

where Sinkhorn iterations alternately rescale $\mathbf{u}$ and $\mathbf{r}$ to match $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$. The real block of $\bar{\mathbf{P}}^\star$ gives $\mathbf{P}^\star$, while the dustbins absorb the remaining $1 - \rho$ mass. OTA then minimizes the average cost carried by genuine matches:

$$\mathcal{L}_{\mathrm{OTA}} = \frac{\langle \mathbf{P}^\star, \mathbf{D} \rangle}{\sum_{i=1}^n \sum_{j=1}^m P^\star_{ij}}. \tag{16}$$

The transport plan supplies token-level structure, while the dustbin prevents ambiguous or semantically empty units from creating spurious supervision.

Joint Optimization and Inference

The two anchors regularize complementary model objects and are combined with translation supervision as

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \alpha_{\mathrm{TPA}} \mathcal{L}_{\mathrm{TPA}} + \alpha_{\mathrm{OTA}} \mathcal{L}_{\mathrm{OTA}}, \tag{17}$$

where α_{TPA} and α_{OTA} balance linguistic stability and visual grounding. At inference, the frozen prior, content selector, and Sinkhorn solver are removed, and the student follows the original autoregressive decoding path. DualAnchor therefore changes training supervision without adding test-time teacher access or optimal-transport computation.

PHOENIX-2014T					
Method	B1	B2	B3	B4	R-L
GFSLT-VLP	43.71	33.18	26.11	21.44	42.49
FLa-LLM	46.29	35.33	28.03	23.09	45.27
Sign2GPT	49.54	35.96	28.83	22.52	48.90
SignLLM	45.21	34.78	28.05	23.40	44.49
BeyondGloss	52.38	38.57	30.74	25.49	**52.89**
MMSLT	48.92	38.12	30.79	25.73	47.97
SpaMo	49.80	37.32	29.50	24.32	46.57
SCL-SLT	48.72	38.19	31.04	26.00	47.02
DualAnchor (ours)	**53.93**	**41.28**	**33.18**	**27.60**	50.62
CSL-Daily					
Method	B1	B2	B3	B4	R-L
GFSLT-VLP	39.37	24.93	16.26	11.00	36.44
FLa-LLM	37.13	25.12	18.38	14.20	37.25
Sign2GPT	41.75	28.73	20.60	15.40	42.36
SignLLM	39.55	28.13	20.07	15.75	39.91
BeyondGloss	53.12	38.63	27.82	21.53	**53.46**
MMSLT	49.87	36.37	27.29	21.11	48.92
SpaMo	48.90	36.90	26.78	20.55	47.46
SCL-SLT	52.81	39.28	29.82	23.25	51.08
DualAnchor (ours)	**53.15**	**39.84**	**30.65**	**24.21**	49.46

Table 1: Gloss-free SLT results on the official test sets. Bn and R-L denote BLEU-n and ROUGE-L; **bold** and underlining mark the best and second-best scores.

Experiments

Datasets and metrics. We evaluate on PHOENIX-2014T and CSL-Daily. PHOENIX-2014T contains 8,257 German Sign Language weather videos with German translations and uses the standard 7,096/519/642 train/development/test split (Camgoz et al. 2018); CSL-Daily contains 20,654 Chinese Sign Language videos on daily-life topics with Chinese translations and uses the standard 18,401/1,077/1,176 split (Zhou et al. 2021). We report BLEU-1–BLEU-4 (Papineni et al. 2002) and ROUGE-L (Lin 2004).

Baselines. We compare eight published gloss-free systems spanning visual-language pretraining (GFSLT-VLP (Zhou et al. 2023)), LLM adaptation (FLa-LLM (Chen et al. 2024), Sign2GPT (Wong, Camgoz, and Bowden 2024), SignLLM (Gong et al. 2024), and SpaMo (Hwang et al. 2025)), enriched visual or MLLM supervision (BeyondGloss (Asasi et al. 2025) and MMSLT (Kim et al. 2025)), and contrastive alignment (SCL-SLT (Lai et al. 2026)).

Implementation details. We fine-tune the language backbone with LoRA (Hu et al. 2022) using AdamW (Loshchilov and Hutter 2019). Unless otherwise specified, experiments use 4 GPUs, at most 500 epochs, and an early-stopping patience of 50 validation rounds. We use a learning rate of 3×10^{-4} with cosine decay, select checkpoints by the best validation BLEU-4, and decode with beam size 6 and length penalty 1.0; the auxiliary objectives are training-only.

Main Results. DualAnchor outperforms SpaMo on all ten metrics and gives the strongest higher-order BLEU results on both datasets (Table 1). BLEU-4 rises from 24.32 to 27.60 on PHOENIX-2014T and from 20.55 to 24.21 on CSL-Daily, while the remaining BLEU and ROUGE-L scores improve consistently. Across all compared methods, DualAnchor achieves the highest arithmetic mean over the ten reported metrics. Gains span sign languages and domains and are strongest in multi-token correspondence.

Backbone	Method	B4	R-L
NLLB	SpaMo	17.89	43.29
	MMSLT	17.96	41.17
	DualAnchor (ours)	**24.21**	**49.46**
mBART	SpaMo	19.70	46.17
	MMSLT	20.21	**50.58**
	DualAnchor (ours)	**22.35**	49.75
mT0	SpaMo	19.59	42.12
	MMSLT	19.92	43.32
	DualAnchor (ours)	**22.56**	**47.75**

Table 2: CSL-Daily cross-backbone results. **Bold** marks the best completed score for each backbone. Rows share cached visual features, decoding settings, and the evaluation script.

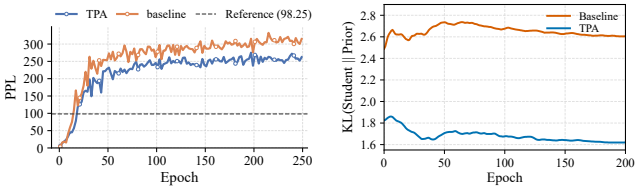
Cross-backbone generalization. Table 2 evaluates TPA and OTA with NLLB (NLLB Team 2024), mBART (Liu et al. 2020), and mT0 (Muennighoff et al. 2023) under the same protocol. Across completed runs, DualAnchor leads BLEU-4 on all three backbones and improves both metrics over SpaMo in every matched comparison. The gains are largest with NLLB and remain positive on the other two backbones. On mBART, MMSLT leads ROUGE-L but DualAnchor leads BLEU-4, making the ranking metric-dependent.

Analysis

This section asks four questions: (i) **RQ1:** How does TPA affect fluency and linguistic structure, and do these effects align with prior preservation? (ii) **RQ2:** Does OTA improve fine-grained lexical grounding and use local visual evidence as intended? (iii) **RQ3:** How sensitive are TPA and OTA to their loss weights? (iv) **RQ4:** Which TPA and OTA design choices govern translation quality and module behavior?

RQ1: How does TPA affect fluency and linguistic structure, and do these effects align with prior preservation? Using SpaMo on CSL-Daily, we compare otherwise identical models with and without TPA to test how prior preservation relates to fluency and linguistic form.

PPL and KL dynamics. At matched checkpoints, lower frozen-Baichuan2-7B corpus PPL indicates better external fluency, while lower mean student–prior KL under shared gold prefixes indicates less drift from the frozen prior.



(a) External PPL (b) Student–prior KL

Figure 4: CSL-Daily TPA diagnostics: (a) frozen-Baichuan2-7B PPL; (b) mean student–prior KL.

Syntactic metrics (baseline vs. TPA)			
Metric	Base	TPA	Δ
Mean dependency JSD ↓	0.4337	**0.3955**	↓**8.81%**
JSD > 0.3 ↓	73.89%	**66.07%**	↓**7.82 pp**
JSD > 0.5 ↓	35.12%	**28.66%**	↓**6.46 pp**
JSD > 0.7 ↓	11.48%	**7.14%**	↓**4.34 pp**
Dependency edit ↓	0.6551	**0.6139**	↓**6.29%**
Grammatical metrics (baseline vs. TPA)			
Metric	Base	TPA	Δ
Function-marker omissions ↓	62.1	**56.9**	↓**8.37%**
Question-particle precision ↑	39.6%	**46.8%**	↑**7.2 pp**
Structural-particle recall ↑	50.9%	**53.7%**	↑**2.8 pp**

Table 3: CSL-Daily TPA linguistic diagnostics (epoch 200; omissions per 100 sentences; pp: percentage points).

Across training, lower PPL accompanies lower student–prior KL with TPA (Figure 4). Their agreement links the external fluency gain to reduced prior drift.

Syntactic and grammatical analysis. Table 3 tests whether the pattern extends to linguistic structure. Stanza parses omit punctuation and the `root` relation. Dependency JSD compares hypothesis and reference relation histograms; its 0.3/0.5/0.7 exceedance rates capture mismatch severity, while normalized dependency edit divides relation-sequence Levenshtein distance by the longer sequence. Grammar metrics cover function-marker omissions per 100 sentences, question-particle precision, and structural-particle recall.

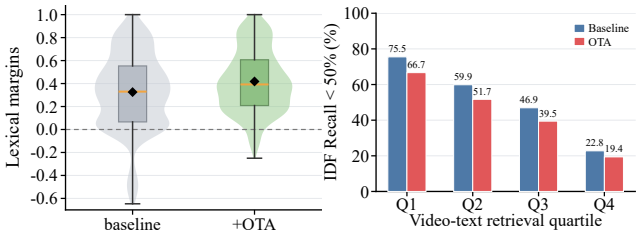
The distributional, sequence, and marker-level diagnostics in Table 3 agree: TPA reduces dependency mismatch and improves particle use. Together with Figure 4, they answer RQ1: prior preservation coincides with better-formed output in the matched comparison.

RQ2: Does OTA improve grounding and evidence use? With SpaMo on CSL-Daily, Figures 5 and 6 compare matched OTA/no-OTA models on lexical discrimination, retrieval strata, and category/error behavior. Figure 7 evaluates the baseline and full DualAnchor with matched occlusion.

Hard-negative separation and retrieval-quality analysis. Figure 5a compares sample-level lexical margins against semantically similar hard negatives that differ in action, object, attribute, time, or quantity. For each of 348 eligible examples, $M_i = S_{\text{lex}}(\widehat{Y}_i, D_i^+) - S_{\text{lex}}(\widehat{Y}_i, D_i^-)$, where S_{lex} is content-token overlap and D_i^+ and D_i^- are the visually sensitive token sets of the correct and negative references. Positive margins favor the video-specific reference. Figure 5b partitions the 1,176 test examples into quartiles of sentence-level video–reference cosine similarity (Q1 lowest; Q4 highest) and reports the share with IDF-weighted content-token recall below 50%. Recall is the recovered share of reference content-token IDF mass; lower low-recall rates are better.

The two panels separate local grounding from coarse retrieval quality. OTA better distinguishes the correct content from matched confounders, and its low-recall advantage persists within every sentence-level similarity quartile. This advantage spans the observed similarity range.

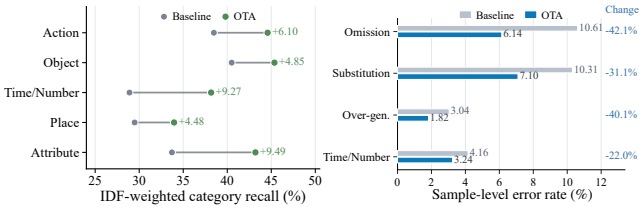
Token-level lexical analysis. At the analyzed checkpoint,



(a) Hard-negative margins (b) Low-recall rate by quartile

Figure 5: OTA grounding diagnostics on CSL-Daily: (a) lexical margins against semantically similar hard negatives; (b) low-recall rates across video–text similarity quartiles.

Figure 6a groups reference content tokens into actions, objects, times/numbers, places, and attributes and reports IDF-weighted recall; annotations give absolute OTA–baseline differences in percentage points. Across all 1,176 test examples, Figure 6b reports the percentage exhibiting omission, substitution, over-generalization, or time/number errors; annotations give relative changes from the baseline. Higher recall and lower error rates are better.



(a) Category-wise IDF recall (b) Sample-level error rates

Figure 6: CSL-Daily OTA diagnostics: (a) category-wise IDF-weighted recall; (b) sample-level lexical-error rates.

The category and error views agree: the recall advantage spans distinct content types and coincides with fewer examples exhibiting each measured lexical error. Together, the two views support broad lexical recovery.

OTA visual occlusion analysis. Figure 7 tests whether OTA transport mass identifies visual evidence used during generation. With decoding fixed, the baseline and OTA are evaluated on the original video and with equal-length masks over random or highest-mass spans. We report absolute IDF-weighted recall and BLEU-4, not precomputed drops; larger changes indicate greater masking sensitivity.

Equal-length high-transport masks disrupt content recovery and sentence-level quality more than random masks for both models. The analyses show that OTA improves fine-grained grounding and locates generation-sensitive visual evidence via transport mass.

RQ3: How sensitive are TPA and OTA to their respective loss weights? We vary one auxiliary-loss coefficient while fixing all other training and decoding settings. Figure 8 reports each run's maximum corpus BLEU-4 over saved checkpoints; zero denotes the corresponding no-loss baseline.

The sweeps distinguish robustness from insensitivity: each

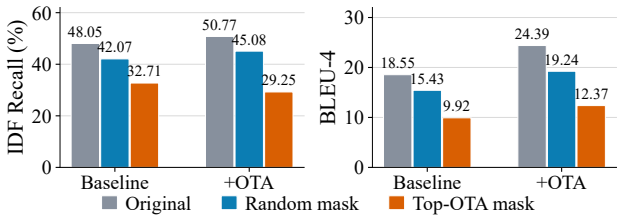
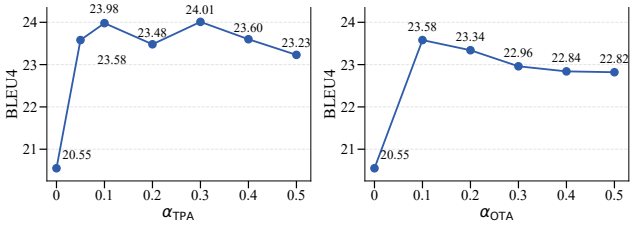


Figure 7: Absolute CSL-Daily IDF-weighted recall (left) and BLEU-4 (right) under original, random-mask, and Top-OTA-mask conditions for the baseline and OTA.



(a) TPA strength (b) OTA strength

Figure 8: CSL-Daily best-checkpoint BLEU-4 versus (a) α_{TPA} and (b) α_{OTA}; zero denotes no auxiliary loss.

objective remains effective across adjacent weights, but translation quality weakens when either term grows too dominant, especially OTA. Thus, the objectives tolerate moderate retuning but should remain auxiliary to translation loss.

RQ4: How do core TPA and OTA design choices affect translation quality and module behavior? Table 4(a) compares TPA KL directions and distillation prefixes against no TPA, pairing BLEU-4 with frozen-model PPL. Panel (b) compares sentence-level, attention-based, hard, and OTA alignment plus two OTA subsettings, pairing BLEU-4 with IDF-weighted recall. TPA remains advantageous across the tested directions and prefixes despite a BLEU-4–PPL tradeoff. Full OTA provides the strongest joint translation and recall result; removing the dustbin or transporting all tokens weakens this balance. Thus, TPA is robust across the tested implementations, whereas OTA benefits from selective partial matching that avoids unreliable correspondences.

(a) TPA Design

Setting	B4 ↑	PPL ↓
Baseline	20.55	317.90
Direction of KL Optimization		
TPA w/ Forward KL	23.37	265.92
TPA w/ Reverse KL	**23.75**	250.84
Context Prefix		
TPA w/ English Context	22.10	**222.90**
TPA w/ Target Context	23.56	249.46

(b) OTA Design

Design	B4 ↑	IDF Recall ↑
Alignment Design		
Sentence-level	20.55	44.32
Attention alignment	22.27	47.88
Hard alignment	22.38	52.58
OTA	**24.07**	**54.20**
OTA Subsettings		
No dustbin	22.93	53.05
All tokens	23.50	52.32

Table 4: TPA and OTA design ablations on CSL-Daily: (a) KL direction and context prefix, with Target Context as the default; (b) alignment design, dustbin use, and token scope.

Related Work

Gloss-Free and LLM-Based Sign Language Translation. Gloss-based, unified, and iterative-prototype systems bridge video and text via intermediate recognition or shared representations (Camgöz et al. 2020; Yin and Read 2020; Zhang, Müller, and Sennrich 2023; Yao et al. 2023). Gloss-free models use visual-language pretraining, boundary-aware features, concept queries, or contrastive sign–text learning (Zhou et al. 2023; Yin et al. 2023; Lin et al. 2023; Jiang et al. 2024); unsupervised and multilingual models expand supervision (Guo et al. 2024; Tan, Miyazaki, and Nakadai 2025). LLM systems condition pretrained decoders via adapters, discrete sign tokens, factorized adaptation, motion prompts, generated descriptions, or latent plans (Wong, Camgoz, and Bowden 2024; Gong et al. 2024; Chen et al. 2024; Hwang et al. 2025; Kim et al. 2025; Jiang et al. 2026).
Preserving Language Priors. Knowledge distillation transfers sequence behavior across models and multilingual translation tasks (Kim and Rush 2016; Sun et al. 2020), while InstructGPT constrains policy updates through per-token KL and pretraining gradients (Ouyang et al. 2022). Multimodal models preserve language competence via visual experts, replay, or selective cross-modal distillation (Wang et al. 2024; Lin et al. 2024; Irawan et al. 2026). TPA uses a shared gold prefix and confidence-weighted token-level KL to regularize conditional translation without an inference-time prior.
Fine-Grained Sign-Text Alignment. Content words carry disproportionate semantic weight in translation (Chen et al. 2020a), motivating lexical-unit alignment rather than a single sentence target. Prior work spans global video–text objectives (Zhou et al. 2023; Yin et al. 2023; Jiang et al. 2024), sign–word retrieval, dense contrastive separation, segment supervision, selective negatives (Cheng et al. 2023; Ye et al. 2024; Low, Sincan, and Bowden 2025; Lai et al. 2026), and optimal-transport sign codebooks (Gong et al. 2024). Entropy-regularized partial transport is differentiable and robust to unmatched units, including in cross-domain alignment (Cuturi 2013; Chapel, Alaya, and Gasso 2020; Chen et al. 2020b). OTA directly learns soft many-to-many grounding between continuous visual and content-token representations without an inference-time alignment module.

Conclusion

LLM-based gloss-free SLT must preserve fluent generation and ground lexical choices in visual evidence. DualAnchor meets both requirements through complementary training-only objectives: TPA anchors next-token predictions to a frozen language prior, while OTA uses partial optimal transport to align visual and content tokens without forcing unreliable matches. On PHOENIX-2014T and CSL-Daily, DualAnchor achieves the best BLEU-4 among the compared gloss-free methods and remains effective across three language backbones. Analyses associate TPA with lower prior drift, stronger fluency, and better grammar; OTA recovers more content tokens, reduces lexical errors, and identifies generation-sensitive visual evidence. Thus, preserving linguistic form and fine-grained grounding jointly enable more faithful sign language translation.

References

Asasi, S.; Lakhal, M. I.; Sincan, O. M.; and Bowden, R. 2025. Beyond Gloss: A Hand-Centric Framework for Gloss-Free Sign Language Translation. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA.

Camgoz, N. C.; Hadfield, S.; Koller, O.; Ney, H.; and Bowden, R. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7784–7793.

Camgöz, N. C.; Koller, O.; Hadfield, S.; and Bowden, R. 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*, 10020–10030.

Chapel, L.; Alaya, M. Z.; and Gasso, G. 2020. Partial Optimal Transport with Applications on Positive-Unlabeled Learning. In *Advances in Neural Information Processing Systems*, volume 33.

Chen, K.; Wang, R.; Utiyama, M.; and Sumita, E. 2020a. Content Word Aware Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 358–364.

Chen, L.; Gan, Z.; Cheng, Y.; Li, L.; Carin, L.; and Liu, J. 2020b. Graph Optimal Transport for Cross-Domain Alignment. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1542–1553.

Chen, Z.; Zhou, B.; Li, J.; Wan, J.; Lei, Z.; Jiang, N.; Lu, Q.; and Zhao, G. 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)*, 7071–7081. Torino, Italia: ELRA and ICCL.

Cheng, Y.; Wei, F.; Bao, J.; Chen, D.; and Zhang, W. 2023. CiCo: Domain-Aware Sign Language Retrieval via Cross-Lingual Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19016–19026.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26, 2292–2300. Curran Associates, Inc.

Gong, J.; Foo, L. G.; He, Y.; Rahmani, H.; and Liu, J. 2024. LLMs are Good Sign Language Translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18362–18372.

Guo, Z.; He, Z.; Jiao, W.; Wang, X.; Wang, R.; Chen, K.; Tu, Z.; Xu, Y.; and Zhang, M. 2024. Unsupervised Sign Language Translation and Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 14041–14055. Bangkok, Thailand: Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hwang, E. J.; Cho, S.; Lee, J.; and Park, J. C. 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)*, 3901–3920. Albuquerque, New Mexico: Association for Computational Linguistics.

Irawan, P. A.; Fuadi, E. H.; Kumar, S.; Aji, A. F.; and Kementchedjhieva, Y. 2026. LinguDistill: Recovering Linguistic Ability in Vision-Language Models via Selective Cross-Modal Distillation. arXiv:2604.00829.

Jiang, Y.; Zhang, L.; Wei, X.-Y.; and Qing, L. 2026. Think in Latent Thoughts: A New Paradigm for Gloss-Free Sign Language Translation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9993–10012. San Diego, California, United States: Association for Computational Linguistics.

Jiang, Z.; Sant, G.; Moryossef, A.; Müller, M.; Sennrich, R.; and Ebling, S. 2024. SignCLIP: Connecting Text and Sign Language by Contrastive Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9171–9193.

Kim, J.; Jeon, H.; Bae, J.; and Kim, H. Y. 2025. Leveraging the Power of MLLMs for Gloss-Free Sign Language Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21048–21058.

Kim, Y.; and Rush, A. M. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1317–1327.

Lai, C. H.; Zhao, R.; Zhong, X.; Su, J.; and Chen, Y. 2026. Selective Contrastive Learning For Gloss Free Sign Language Translation. In Liakata, M.; Moreira, V. P.; Zhang, J.; and Jurgens, D., eds., *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 45648–45660. San Diego, California, United States: Association for Computational Linguistics. ISBN 979-8-89176-390-6.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. VILA: On Pre-training for Visual Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26689–26699.

Lin, K.; Wang, X.; Zhu, L.; Sun, K.; Zhang, B.; and Yang, Y. 2023. Gloss-Free End-to-End Sign Language Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12904–12916.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Low, J. H.; Sincan, O. M.; and Bowden, R. 2025. SAGE: Segment-Aware Gloss-Free Encoding for Token-Efficient Sign Language Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 5011–5020.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Le Scao, T.; Bari, M. S.; Shen, S.; Yong, Z. X.; Schoelkopf, H.; Tang, X.; Radev, D.; Aji, A. F.; Almubarak, K.; Albanie, S.; Alyafeai, Z.; Webson, A.; Raff, E.; and Raffel, C. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15991–16111.

NLLB Team. 2024. Scaling Neural Machine Translation to 200 Languages. *Nature*, 630: 841–846.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744. Curran Associates, Inc.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Sun, H.; Wang, R.; Chen, K.; Utiyama, M.; Sumita, E.; and Zhao, T. 2020. Knowledge Distillation for Multilingual Unsupervised Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3525–3535.

Tan, S.; Miyazaki, T.; and Nakadai, K. 2025. Multilingual Gloss-free Sign Language Translation: Towards Building a Sign Language Foundation Model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 553–561.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; Xu, J.; Chen, K.; Xu, B.; Li, J.; Dong, Y.; Ding, M.; and Tang, J. 2024. CogVLM: Visual Expert for Pretrained Language Models. In *Advances in Neural Information Processing Systems*, volume 37, 121475–121499. Curran Associates, Inc.

Wong, R.; Camgoz, N. C.; and Bowden, R. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation. In *The Twelfth International Conference on Learning Representations*.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; Yang, F.; Deng, F.; Wang, F.; Liu, F.; Ai, G.; Dong, G.; Zhao, H.; Xu, H.; Sun, H.; Zhang, H.; Liu, H.; Ji, J.; Xie, J.; Dai, J.; Fang, K.; Su, L.; Song, L.; Liu, L.; Ru, L.; Ma, L.; Wang, M.; Liu, M.; Lin, M.; Nie, N.; Guo, P.; Sun, R.; Zhang, T.; Li, T.; Li, T.; Cheng, W.; Chen, W.; Zeng, X.; Wang, X.; Chen, X.; Men, X.; Yu, X.; Pan, X.; Shen, Y.; Wang, Y.; Li, Y.; Jiang, Y.; Gao, Y.; Zhang, Y.; Zhou, Z.; and Wu, Z. 2023. Baichuan 2: Open Large-scale Language Models. arXiv:2309.10305.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. arXiv:2412.15115.

Yao, H.; Zhou, W.; Feng, H.; Hu, H.; Zhou, H.; and Li, H. 2023. Sign Language Translation with Iterative Prototype. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15592–15601.

Ye, J.; Wang, X.; Jiao, W.; Liang, J.; and Xiong, H. 2024. Improving Gloss-free Sign Language Translation by Reducing Representation Density. In *Advances in Neural Information Processing Systems*, volume 37, 107379–107402. Curran Associates, Inc.

Yin, A.; Zhong, T.; Tang, L.; Jin, W.; Jin, T.; and Zhao, Z. 2023. Gloss Attention for Gloss-Free Sign Language Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2551–2562.

Yin, K.; and Read, J. 2020. Better Sign Language Translation with STMC-Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5975–5989.

Zhang, B.; Müller, M.; and Sennrich, R. 2023. SLTUNET: A Simple Unified Model for Sign Language Translation. In *International Conference on Learning Representations*.

Zhou, B.; Chen, Z.; Clapés, A.; Wan, J.; Liang, Y.; Escalera, S.; Lei, Z.; and Zhang, D. 2023. Gloss-Free Sign Language Translation: Improving from Visual-Language Pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 20871–20881.

Zhou, H.; Zhou, W.; Qi, W.; Pu, J.; and Li, H. 2021. Improving Sign Language Translation With Monolingual Data by Sign Back-Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1316–1325.