VTaMo: Video-Text Alignment Model for Sign Language Translation

Junyi Hu[1], Zhewen He[1], Haomian Huang[1], Aoxiang Yang[1], and Yi Fang[1,2⋆]

[1] New York University Abu Dhabi, UAE
[2] ChatSign Technology
jh10472@nyu.edu, yf23@nyu.edu

Abstract. Sign language translation (SLT) converts continuous sign videos into spoken language text. Gloss-free approaches leverage pretrained visual encoders and language models but rely on implicit cross-modal alignment from translation supervision alone. We present VTaMo, a framework that introduces explicit multi-granularity alignment at three levels: (1) local alignment via entropy-regularized optimal transport with a learnable null token for fine-grained frame-to-token correspondences; (2) global alignment via a learnable orthogonal transformation that calibrates embedding space geometry through Earth Mover's Distance; and (3) position-aligned contrastive learning for discriminative token-level representations. Experiments on Phoenix-2014T, CSL-Daily, How2Sign, and OpenASL demonstrate consistent state-of-the-art performance, with ablations confirming the complementary contributions of each component. Code is available at https://github.com/junyi2005/vtamo.

Keywords: Sign language translation · Cross-modal alignment · Optimal transport

1 Introduction

Sign language translation (SLT) aims to convert continuous sign language videos into spoken language sentences, enabling more accessible communication between Deaf and hard-of-hearing signers and hearing communities [1, 48]. Recent gloss-free systems decode text directly from visual features using large pretrained sequence-to-sequence language models [7,13,40], avoiding the costly gloss annotation required by gloss-based pipelines. However, most gloss-free benchmarks provide only natural language sentences as supervision, without gloss annotations or temporal segmentation [11, 36]. The core difficulty is therefore twofold: the semantic gap between vision and text, and the unknown correspondence between the temporal order of visual signs and the token order of the target text.

Sign languages often do not follow the word order of the corresponding spoken language: a signer may express key content words first and add grammatical

⋆ Corresponding author.

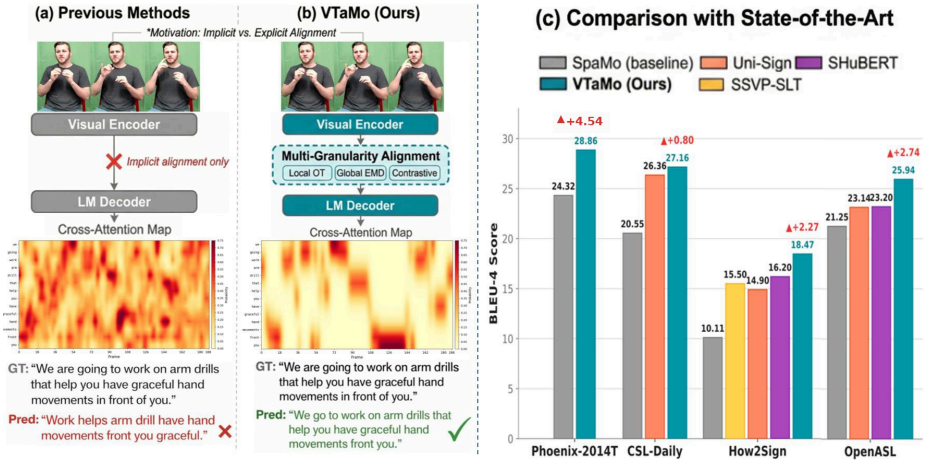


Fig. 1: Motivation and performance of VTaMo. (a) Prior gloss-free sign language translation methods typically rely on *implicit* cross-modal alignment learned inside the decoder attention, which can yield diffuse or mismatched cross-attention and translation errors. **(b)** VTaMo introduces *explicit* multi-granularity vision–text alignment—including local OT-based token-to-frame matching, global distribution alignment, and position-aligned contrastive learning—to sharpen correspondences and improve decoding. **(c)** VTaMo achieves state-of-the-art sign language translation quality across four benchmarks (Phoenix-2014T, CSL-Daily, How2Sign, and OpenASL), compared against SpaMo [16], Uni-Sign [21], SHuBERT [14], and SSVP-SLT [33]. Video frames in (a) and (b) are from the How2Sign dataset [11].

relations later, so the temporal progression of gestures can differ from the target word order. As a result, visual features extracted along the video timeline are often misaligned with the text tokens that an autoregressive decoder is trained to predict. Left unresolved, the decoder must simultaneously learn translation and implicitly discover a latent cross-modal permutation, which increases optimization difficulty and degrades accuracy, especially on large-scale benchmarks such as How2Sign and OpenASL. Fig. 1 illustrates this gap: when alignment is left implicit, the cross-attention map can be noisy and the decoder may produce incorrect word ordering, whereas our approach makes alignment explicit.

Existing methods largely treat the visual stream as an ordered sequence and rely on decoder attention to bridge the modality gap. ShuBERT [14] and Uni-Sign [21] strengthen representations through large-scale pretraining but do not explicitly model frame-to-token correspondence within a sentence, while SpaMo [16] uses batch-level contrastive learning yet leaves fine-grained, sentence-internal alignment unresolved. These limitations motivate a model that learns correspondence between visual segments and text tokens and uses it to present the decoder with a semantically ordered visual sequence.

We propose **VTaMo**, a vision–text alignment model for gloss-free SLT that explicitly aligns and reorders visual features before text generation. Because

many spoken words such as articles and prepositions have no sign-level counterpart, we align against and decode a content-word *pseudo-gloss* of the target rather than the raw sentence. Given a sign video, we build token-level pseudo-gloss embeddings from a frozen language model embedding layer and temporal visual embeddings from a video encoder, then estimate a soft alignment between visual positions and pseudo-gloss tokens by solving an entropy-regularized optimal transport problem [10] with a learnable null token. The null token absorbs transitional gestures and co-articulation that correspond to no explicit word, stabilizing alignment on continuous signing. Using the resulting transport plan, we reorder the visual features into a sequence that follows the target token order and feed it to the language model decoder. At inference the target token order is unknown, so the visual features are not reordered; the decoder generates pseudo-gloss directly from the signing-order features, and a lightweight text-only recovery model restores spoken word order and the omitted function words.

VTaMo is trained end to end with three complementary alignment objectives beyond the standard translation loss. First, a *local* optimal transport loss encourages fine-grained correspondence between temporal visual segments and text tokens while allowing unaligned frames to map to the null token. Second, a *global* loss calibrates the sentence-level geometry of the visual and textual embedding spaces through a learnable orthogonal transformation, supported by a memory queue that diversifies the sentence pairs used for alignment. Third, a *position-aligned contrastive* loss [29] sharpens token-level discriminability between reordered visual features and their text token embeddings without altering the language model embedding space. Together, these losses make the reordered visual features semantically coherent, simplifying decoder learning and improving robustness across datasets.

We evaluate on four widely used benchmarks spanning three languages and diverse data regimes: Phoenix-2014T (German), CSL-Daily (Chinese), How2Sign and OpenASL (English). Across all benchmarks, VTaMo achieves state-of-the-art performance under the gloss-free setting and yields consistent gains over strong baselines that decode without explicit alignment. Ablation studies verify that each alignment component contributes complementary improvements, and qualitative analysis of the learned transport plans reveals interpretable correspondences between signing segments and target tokens.

2 Related Work

2.1 Gloss-free SLT, Alignment, and Large-Scale Pretraining

Gloss-free sign language translation (SLT) translates continuous sign videos into spoken-language sentences without gloss supervision, but it typically underperforms gloss-based pipelines due to the absence of explicit intermediate structure [2,4,5,18,42,44,48,49]. Prior work improves gloss-free SLT by strengthening temporal modeling [20], introducing cross-modal alignment objectives [12,17,23,46], incorporating large language models [7,13,40], scaling training data [33,39], and large-scale sign pretraining such as ShuBERT [14] and Uni-Sign [21]; SpaMo [16]

further emphasizes the non-trivial sign–text correspondence via contrastive objectives. These approaches—including VAP [17], which applies text-derived constraints only as visual pre-training—keep sentence-internal correspondence implicit and rely on coarse decoder attention to resolve word order. Our method instead estimates token-level alignment and reorders visual features before decoding via Sinkhorn OT with a learnable null assignment and complementary local and global objectives.

2.2 Cross-Modal and Video–Text Alignment

Our objectives build on general cross-modal alignment tools but adapt them to the structure of sign language. Entropy-regularized optimal transport with the Sinkhorn algorithm [10] is a standard device for matching features across modalities, e.g., aligning visual features with text prompts in vision–language models [3]; learnable orthogonal transformations relate two embedding spaces without distorting their geometry, as in cross-lingual word-embedding alignment [19]; and contrastive video–text frameworks such as CLIP4Clip [25], VideoCLIP [41], and X-CLIP [26] learn joint embeddings for retrieval. Closer in spirit, procedure-learning and step-grounding methods align video frames with an ordered sequence of procedure steps, e.g., grounding instructional-article steps via narrations [27] and optimal-transport-guided procedure learning [8]. These methods, however, target global or clip-level correspondence, or assume a largely fixed, monotonic step order. Sign language translation differs: the mapping is *non-monotonic* at the lexical level and *partial*, since many frames are transitional and words have no manual counterpart. VTaMo addresses both by coupling OT with a null assignment, sentence-level orthogonal calibration, and position-aligned contrastive learning to yield a reordered visual sequence.

3 Method

We present **VTaMo**, a framework for gloss-free sign language translation that explicitly aligns and reorders visual features before text generation. As illustrated in Fig. 2, VTaMo introduces three complementary alignment objectives: (1) *local alignment* via entropy-regularized optimal transport for fine-grained frame-to-token correspondences; (2) *global alignment* via a learnable orthogonal transformation to calibrate the embedding spaces at the sentence level; and (3) *position-aligned contrastive learning* for discriminative representations.

3.1 Problem Formulation

Given a sign language video, we extract frame-level features $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_L\}$ using a pre-trained CLIP-ViT encoder, where $\mathbf{v}_i \in \mathbb{R}^{d_v}$. The goal is to produce the spoken language sentence corresponding to the video.

Spoken sentences contain many function words that have no direct sign-level realization, which makes a strict frame-to-word alignment ill-posed. We therefore

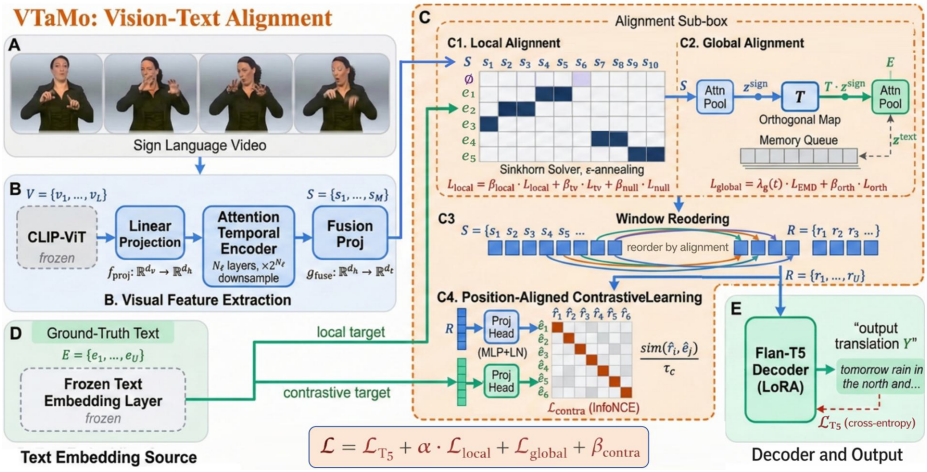


Fig. 2: VTaMo pipeline. A sign-language video is encoded into visual tokens using a frozen CLIP-ViT backbone followed by a lightweight temporal encoder and fusion projection (**A,B**). Given the ground-truth text embeddings (**D**), VTaMo performs **local alignment** with an entropy-regularized OT solver (Sinkhorn) to obtain soft token–frame matches, and **global alignment** by mapping sign features into the text space with an orthogonal transform and a memory queue (**C1,C2**). The resulting correspondence is used for **window reordering** (**C3**) and **position-aligned contrastive learning** (**C4**). A LoRA-adapted Flan-T5 decoder generates the translation (**E**). Video frames are from the Phoenix-2014T dataset [1].

align against, and decode, a *pseudo-gloss* sequence $\mathbf{Y} = \{y_1, \ldots, y_U\}$ of U tokens derived from the target sentence by a fixed offline filter. Concretely, we apply a frozen spaCy part-of-speech tagger that retains sign-relevant content tokens (`NOUN`, `VERB`, `ADJ`, `ADV`, `NUM`, `PRON`, `PROPN`) and removes tokens that usually lack a direct sign counterpart (`DET`, `ADP`, `AUX`, `PART`, `PUNCT`, `CCONJ`); the same filtering rule is applied across all datasets using language-specific spaCy models. Both the text tokens used for alignment and the decoder targets are pseudo-gloss, and a lightweight text-only recovery model restores fluent sentences at inference.

The visual features are projected to $\mathbb{R}^{d_h}$ via a linear layer, temporally downsampled by an attention-based encoder to produce $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_M\}$ with $M \ll L$, and mapped to the language model dimension d_t by a fusion projector. A LoRA-adapted Flan-T5 model then autoregressively generates the pseudo-gloss sequence. The central challenge is that sign language does not follow the word order of spoken language, and the temporal granularity of visual frames far exceeds that of linguistic tokens.

3.2 Attention-Based Temporal Encoding

We use an attention-based temporal convolution module with N_ℓ cascaded layers, each performing $2\times$ downsampling via learnable attention-weighted aggregation

over local windows of size W. For each output position i, a local window $\mathcal{W}_i$ centered at position $2i$ is extracted, and a query is formed from the mean-pooled local context plus a learnable positional embedding $\mathbf{p}$:

$$\mathbf{q}_i = \frac{1}{|\mathcal{W}_i|} \sum_{j \in \mathcal{W}_i} \mathbf{x}_j + \mathbf{p}. \tag{1}$$

Multi-head attention with relative position bias then aggregates the local context:

$$\alpha_{i,j} = \text{softmax}\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_h / N_{\text{head}}}} + r_{j - c_i} \right), \qquad \mathbf{o}_i = \sum_{j \in \mathcal{W}_i} \alpha_{i,j}\, \mathbf{v}_j, \tag{2}$$

where c_i is the window center, r_{j-c_i} is a learnable relative position bias, and N_{head} is the number of attention heads. This ensures gradient flow to all frames, yielding more expressive temporal representations than max-pooling.

3.3 Local Alignment via Optimal Transport

Local alignment is the core component of VTaMo, establishing fine-grained correspondences between temporal visual segments and text tokens, formulated as an entropy-regularized optimal transport problem. Let $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_M\} \in \mathbb{R}^{M \times d_t}$ denote the visual features after fusion projection, and $\mathbf{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_U\} \in \mathbb{R}^{U \times d_t}$ the pseudo-gloss token embeddings from the frozen encoder embedding layer. Since $\mathbf{S}$ is optimized against the fixed text embeddings $\mathbf{E}$ under the alignment and translation losses, their cosine similarity reflects sign-level semantic correspondence, making it a meaningful transport cost. We introduce a single learnable null token $\mathbf{e}_\varnothing \in \mathbb{R}^{d_t}$, prepended at position 0 of the text sequence, forming an augmented sequence $\tilde{\mathbf{E}} \in \mathbb{R}^{K \times d_t}$ with $K = U + 1$. The assignment is deliberately not one-to-one: frames that carry no explicit sign-language semantics—such as background, transition, or reset frames—are all allowed to align to this same null token, so the model is never forced to map every frame onto a content token. The pairwise cost matrix uses cosine distance, with a learnable bias $b_\varnothing$ controlling null affinity:

$$C_{m,k} = \begin{cases} 1 - \bar{\mathbf{s}}_m^\top \bar{\mathbf{e}}_k - b_\varnothing, & \text{if } k = 0 \text{ (null)}, \\ 1 - \bar{\mathbf{s}}_m^\top \bar{\mathbf{e}}_k, & \text{otherwise}, \end{cases} \tag{3}$$

where $\bar{\mathbf{s}}_m$ and $\bar{\mathbf{e}}_k$ are ℓ_2-normalized vectors. We solve the entropy-regularized OT problem via the Sinkhorn algorithm in log-domain:

$$\mathbf{A}^* = \arg \min_{\mathbf{A} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{A}, \mathbf{C} \rangle - \varepsilon\, H(\mathbf{A}), \tag{4}$$

where $\Pi(\mathbf{a}, \mathbf{b})$ is the set of transport plans with uniform marginals over valid frame and token positions, $H(\mathbf{A}) = -\sum_{m,k} A_{m,k} \log A_{m,k}$ is the entropy term,

and ε controls the sharpness of assignments. For numerical stability we iterate the dual potentials in the log domain for N_{iter} steps,

$$\begin{aligned} \log u_m &\leftarrow \log a_m - \log \textstyle\sum_k \exp(-C_{m,k}/\varepsilon + \log v_k)\,, \\ \log v_k &\leftarrow \log b_k - \log \textstyle\sum_m \exp(-C_{m,k}/\varepsilon + \log u_m)\,, \end{aligned} \tag{5}$$

and recover the plan as $A_{m,k} = \exp(\log u_m - C_{m,k}/\varepsilon + \log v_k)$. We adopt a multi-phase annealing schedule that progressively decreases ε from a high initial value to encourage exploration before converging to sharp, discrete assignments.

The local alignment loss combines three terms. The first is the transport cost $\mathcal{L}_{\text{trans}} = \frac{1}{B}\sum_i \sum_{m,k} A^{(i)}_{m,k} C^{(i)}_{m,k}$. The second is a temporal variation regularization that penalizes adjacent frames switching tokens, computed on the row-normalized plan $\hat{\mathbf{A}}$ restricted to real text tokens:

$$\mathcal{L}_{\text{tv}} = \frac{1}{(M-1)(K-1)} \sum_{m=1}^{M-1} \sum_{k=1}^{K-1} \big|\hat{A}_{m+1,k} - \hat{A}_{m,k}\big|. \tag{6}$$

The third is a null cost regularization that keeps the mean cost-based null affinity $\bar{p}_\varnothing$—computed via a row-wise softmax on the cost matrix independently of the transport plan—below a target ratio ρ_{target}, preventing the learnable bias $b_\varnothing$ from making the null column uniformly cheaper than all real tokens:

$$\mathcal{L}_{\text{null}} = \big[\bar{p}_\varnothing - \rho_{\text{target}}\big]_+^2, \qquad \bar{p}_\varnothing = \frac{1}{M}\sum_{m=1}^{M} \frac{\exp(-C_{m,0}/\tau)}{\sum_{k=0}^{K-1}\exp(-C_{m,k}/\tau)}. \tag{7}$$

The complete local objective is their weighted sum:

$$\mathcal{L}_{\text{local}} = \beta_{\text{local}}\,\mathcal{L}_{\text{trans}} + \beta_{\text{tv}}\,\mathcal{L}_{\text{tv}} + \beta_{\text{null}}\,\mathcal{L}_{\text{null}}. \tag{8}$$

3.4 Global Alignment via Orthogonal Transformation

Because the visual and textual encoders are pre-trained on different modalities, paired sign and text sentence embeddings can suffer a coordinate (orientation) mismatch even when they are semantically equivalent. We therefore apply a learnable *orthogonal transformation*—a rotation of the feature space that preserves vector lengths and angular relationships—to the visual sentence embeddings before comparing the two modalities at the sentence level. We constrain the transformation to be orthogonal precisely because the goal is to correct orientation without reshaping the space: this preserves norms and angular structure for the cosine-based transport cost and avoids the spurious low-cost matches that an unconstrained projection could introduce.

We aggregate frame-level features into one sentence-level vector per video and per sentence via attention pooling, using a mean query $\mathbf{q}$ over valid positions and the normalized weighted sum

$$\mathbf{z} = \frac{\sum_l \alpha_l \mathbf{x}_l}{\big\|\sum_l \alpha_l \mathbf{x}_l\big\|_2}, \qquad \alpha_l = \mathrm{softmax}_l\big(\mathbf{x}_l^\top \mathbf{q}/\sqrt{d_t}\big), \tag{9}$$

yielding normalized sentence-level vectors $\mathbf{z}^{\text{sign}}, \mathbf{z}^{\text{text}} \in \mathbb{R}^{d_t}$. A FIFO memory queue stores recent sentence-level vector pairs so that the sentence-level transport is computed over a sufficiently diverse batch rather than a few in-batch pairs. A learnable matrix $\mathbf{T} \in \mathbb{R}^{d_t \times d_t}$, initialized near identity, transforms the visual sentence vectors:

$$\tilde{\mathbf{z}}_n^{\text{sign}} = \frac{\mathbf{z}_n^{\text{sign}} \mathbf{T}}{\|\mathbf{z}_n^{\text{sign}} \mathbf{T}\|_2}. \tag{10}$$

We then form the pairwise cosine cost $C^g_{n,n'} = 1 - (\tilde{\mathbf{z}}_n^{\text{sign}})^\top \mathbf{z}_{n'}^{\text{text}}$ between the transformed visual vectors and the textual vectors and solve a Sinkhorn OT problem with uniform marginals to obtain a sentence-level transport plan $\mathbf{P}$; the resulting Earth Mover's Distance loss is

$$\mathcal{L}_{\text{EMD}} = \sum_{n,n'} P_{n,n'} \, (1 - (\tilde{\mathbf{z}}_n^{\text{sign}})^\top \mathbf{z}_{n'}^{\text{text}}). \tag{11}$$

An orthogonality constraint $\mathcal{L}_{\text{orth}} = \|\mathbf{T}^\top \mathbf{T} - \mathbf{I}\|_F^2$ keeps $\mathbf{T}$ close to a pure rotation and preserves pairwise distances. The global loss is activated only after local alignment stabilizes, using a scheduled ramp-up:

$$\mathcal{L}_{\text{global}} = \lambda_g(t) \cdot \mathcal{L}_{\text{EMD}} + \beta_{\text{orth}} \, \mathcal{L}_{\text{orth}}. \tag{12}$$

3.5 Window-Based Feature Reordering

Once local alignment has produced a reliable transport plan $\mathbf{A}$, we reorder the visual features to match the target token order. This is necessary because the decoder's autoregressive cross-entropy loss expects the input to follow the target text order; otherwise the visual features stay in temporal order and the cross-entropy gradient counteracts the alignment objectives.

Concretely, we partition the M visual frames into $N_w = U + 2$ overlapping temporal windows, where the two extra windows absorb boundary effects; the w-th window spans $[s_w, e_w)$ with $s_w = \lfloor w\,(M-1)/N_w + 0.5 \rfloor$ and $e_w = \lfloor (w+1)(M-1)/N_w + 0.5 \rfloor + 1$. For each window we accumulate the alignment mass $\boldsymbol{\pi}_w = \sum_{m=s_w}^{e_w-1} \mathbf{A}_{m,:} \in \mathbb{R}^K$ and assign it to the real text token with the largest aggregated mass, $\phi(w) = \arg\max_{u \in \{1,\dots,U\}} \pi_{w,u}$; windows dominated by the null token are excluded. Collecting the windows assigned to each token u into $\Omega_u = \{w \mid \phi(w) = u\}$ and traversing the token order left to right yields the reordered sequence

$$\mathbf{R} = \text{Concat}\big(\{\mathbf{S}_{s_w:e_w-1} \mid w \in \Omega_1^{\uparrow}\}, \dots, \{\mathbf{S}_{s_w:e_w-1} \mid w \in \Omega_U^{\uparrow}\}\big) \in \mathbb{R}^{U' \times d_t}, \tag{13}$$

where $\Omega_u^{\uparrow}$ orders the windows in Ω_u by their original temporal index. This preserves local temporal continuity, and the reordered sequence is fed to the decoder. For the contrastive objective, frames assigned to each token u are mean-pooled to produce a single token-level representation $\mathbf{r}_u$, yielding $\mathbf{R} = \{\mathbf{r}_1, \dots, \mathbf{r}_U\}$ aligned one-to-one with $\mathbf{E}$.

3.6 Position-Aligned Contrastive Learning

Because reordering is used only during training, the model must also function at inference when the target order is unknown. We therefore impose a token-level contrastive objective that binds each reordered visual feature to its text token embedding, so that even when tokens arrive out of order each is paired with its correct counterpart and the decoder can recover its lexical item; restoring fluent word order is then left to the recovery model.

We apply learnable projection heads to both modalities:

$$\hat{\mathbf{r}}_u = \mathrm{LN}\big(\mathbf{W}_2^{\mathrm{v}}\,\sigma(\mathbf{W}_1^{\mathrm{v}}\,\mathbf{r}_u)\big), \quad \hat{\mathbf{e}}_u = \mathrm{LN}\big(\mathbf{W}_2^{\mathrm{t}}\,\sigma(\mathbf{W}_1^{\mathrm{t}}\,\mathbf{e}_u)\big), \tag{14}$$

where σ is GELU activation and LN is layer normalization. After ℓ_2-normalization, the InfoNCE loss is:

$$\mathcal{L}_{\mathrm{contra}} = -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \log \frac{\exp\big(\mathrm{sim}(\hat{\mathbf{r}}_i, \hat{\mathbf{e}}_i)/\tau_c\big)}{\sum_{j\in\mathcal{V}} \exp\big(\mathrm{sim}(\hat{\mathbf{r}}_i, \hat{\mathbf{e}}_j)/\tau_c\big)}, \tag{15}$$

where $\mathcal{V}$ denotes valid token positions and τ_c is a temperature hyperparameter. Gradients flow through the visual branch but are blocked from the text branch to preserve the pre-trained language model representations.

Inference and pseudo-gloss recovery. At test time the target order is unknown, so no reordering is applied: the decoder reads the visual tokens in their original temporal order and emits a pseudo-gloss sequence in visual (signing) order rather than spoken-language order. The token-level grounding above makes this feasible, since each visual token is decoded into its correct lexical item even when the tokens arrive out of order. To recover a fluent translation, we pass this video-order pseudo-gloss through a lightweight recovery model that restores spoken word order and re-inserts the function words dropped by the pseudo-gloss filter (Sec. 3.1). Crucially, this recovery model is trained purely on text and never sees sign videos: we take plain sentences, apply the same pseudo-gloss extractor, permute the content words, and train the model to reconstruct the original sentence from this shuffled input.

The overall training objective combines the translation loss with the three alignment losses:

$$\mathcal{L} = \mathcal{L}_{\mathrm{T5}} + \lambda_{\mathrm{local}}\,\mathcal{L}_{\mathrm{local}} + \lambda_g(t)\,\mathcal{L}_{\mathrm{EMD}} + \beta_{\mathrm{orth}}\,\mathcal{L}_{\mathrm{orth}} + \beta_{\mathrm{contra}}\,\mathcal{L}_{\mathrm{contra}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{T5}}$ is the cross-entropy loss for autoregressive generation. Training proceeds in two phases: in the first phase, only alignment losses are active while the language model is frozen, allowing the visual encoder to establish meaningful correspondences; in the second phase, the full objective is optimized jointly with the LoRA-adapted language model.

4 Experiments

4.1 Datasets and Metrics

We evaluate our method on four publicly available sign language translation benchmarks that span different languages, vocabulary sizes, and recording conditions: Phoenix-2014T [1] (German), CSL-Daily [48] (Chinese), How2Sign [11] and OpenASL [36] (English). We report BLEU-n scores [30] for $n \in \{1, 2, 3, 4\}$ and ROUGE-L F1 [22] as the primary evaluation metrics. For English SLT datasets (How2Sign and OpenASL), we additionally report BLEURT [34] scores following established protocol. For every benchmark, each model trains only on its official split without additional sign-language data or task-specific pre-training.

4.2 Implementation Details

Visual features. We extract frame-level spatial features using a pre-trained CLIP-ViT-Large model [31] with S^2-Wrapper [35] for multi-scale feature extraction at scales 1 and 2, yielding 2048-dimensional feature vectors per frame.

Language model. We employ Flan-T5-XL [9] as the language model backbone and adapt it using LoRA [15] with rank $r = 16$, scaling factor $\alpha = 32$, and dropout rate 0.1. The LoRA adapters are applied to the query and value projection matrices across all transformer layers.

Architecture. The visual features are projected from $d_v = 2048$ to an intermediate dimension $d_h = 768$ through a linear layer. The attention-based temporal encoder consists of $N_\ell = 2$ cascaded layers with 4 attention heads and a window size of 8, yielding a $4\times$ temporal downsampling. The fusion projector maps the features from d_h to the T5 embedding dimension d_t. The contrastive projection heads are two-layer MLPs with GELU activation and layer normalization, projecting to 768 dimensions.

Training. We train all models using AdamW [24] with a peak learning rate of 6×10^{-4}, cosine learning rate schedule, and linear warm-up over the first 2,000 steps. The batch size is 6 with gradient accumulation over 2 steps, and gradient clipping is applied with a maximum norm of 1.0. All experiments were conducted on a single NVIDIA A100 GPU with mixed-precision (bf16) training.

Alignment hyperparameters. For the local alignment, the Sinkhorn algorithm runs for $N_{\text{iter}} = 10$ iterations. The epsilon annealing schedule proceeds from $\varepsilon_{\text{high}} = 0.12$ to $\varepsilon_{\text{mid}} = 0.10$ over the first 10 epochs and further to $\varepsilon_{\text{low}} = 0.03$ over the following 80 epochs. The null ratio target is $\rho_{\text{target}} = 0.2$, and the loss weights are $\beta_{\text{local}} = 2.0$, $\beta_{\text{tv}} = 0.1$, $\beta_{\text{null}} = 0.1$, with the overall local-alignment weight $\lambda_{\text{local}} = 1.0$. For the global alignment, the EMD solver uses $\varepsilon_g = 0.05$ with 20 Sinkhorn iterations, the memory queue capacity is 256, and the weight ramps

Table 1: Performance comparison on the Phoenix-2014T [1] and CSL-Daily datasets [48]; all numbers are reported on the official test sets. "Vis.Ft." denotes visual fine-tuning on sign language datasets. Bold marks the best result.

Setting	Methods	Modality	Vis.Ft.	Phoenix-2014T					CSL-Daily				
				B1	B2	B3	B4	RG	B1	B2	B3	B4	RG
Gloss-based	SLRT [2]	RGB	✓	46.61	33.73	26.19	21.32	49.54	37.38	24.36	16.55	11.79	36.74
	MMTLB [4]	RGB	✗	53.97	41.75	33.84	28.39	52.65	53.31	40.41	30.87	23.92	53.25
	TS-SLT [5]	RGB	✓	54.90	42.43	34.46	**28.95**	53.48	55.44	42.59	32.87	25.79	55.72
	SLTUNET [44]	RGB	✓	52.92	41.76	33.99	28.47	52.11	54.98	41.44	31.84	25.01	54.08
Weakly Gloss-free	TSPNet [20]	RGB	✓	36.10	23.12	16.88	13.41	34.96	17.09	8.98	5.07	2.97	18.38
	GASLT [43]	RGB	✓	39.07	26.74	21.86	15.74	39.86	19.90	9.94	5.98	4.07	20.35
	CSGCR [46]	RGB	✗	36.71	25.40	18.86	15.18	38.85	—	—	—	—	—
	GFSLT-VLP [47]	RGB	✓	43.71	33.18	26.11	21.44	42.49	39.37	24.93	16.26	11.00	36.44
	FLa-LLM [7]	RGB	✗	46.29	35.33	28.03	23.09	45.27	37.15	25.12	18.38	14.20	37.25
Gloss-free	Sign2GPT [40]	RGB	✓	49.54	35.96	28.83	22.52	48.90	41.75	28.73	20.60	15.40	42.36
	SignLLM [13]	RGB	✓	45.21	34.78	28.05	23.40	44.49	39.55	28.13	20.07	15.75	39.91
	C^2RL [6]	RGB	✓	52.81	40.20	32.20	26.75	50.96	49.32	36.28	27.54	21.61	48.21
	Uni-Sign [21]	Pose+RGB	✓	—	—	—	—	—	55.08	42.14	32.98	26.36	56.51
	SpaMo [16]	RGB	✗	49.80	37.32	29.50	24.32	46.57	48.90	36.90	26.78	20.55	47.46
	VTaMo (Ours)	RGB	✗	**61.32**	**46.27**	**35.15**	28.86	**60.24**	**57.64**	**44.12**	**33.78**	**27.16**	**57.28**

from 0 to $\lambda_g^{\max} = 0.1$ over 4,000 steps after warm-up. The orthogonality penalty coefficient is $\beta_{\text{orth}} = 0.05$. The contrastive learning temperature is $\tau_c = 0.1$, and the contrastive loss weight is $\beta_{\text{contra}} = 1.0$.

Training cost. On a single A100, end-to-end training takes 14 h on Phoenix-2014T, 32 h on CSL-Daily, and 50 h on How2Sign. The optimal transport solvers add negligible overhead: the local Sinkhorn step takes 54 ms (<3.2% of the per-step time), and the global OT over the 256×256 memory queue takes <38 ms.

4.3 Comparison with State-of-the-Art

Results on Phoenix-2014T and CSL-Daily. As shown in Table 1, most prior methods rely on visual-encoder fine-tuning or gloss supervision, yet VTaMo reaches state-of-the-art on both benchmarks without either. On Phoenix-2014T it attains 28.86 BLEU-4, surpassing the strongest gloss-free method SpaMo [16] by 4.54 points and coming within 0.09 of the best gloss-based result (TS-SLT [5], 28.95). On CSL-Daily it reaches 27.16 BLEU-4, a 6.61-point gain over SpaMo and 0.80 over Uni-Sign, showing that explicit multi-granularity alignment especially helps on topically diverse data.

Results on How2Sign. How2Sign is more challenging, with open-domain topics and a much larger vocabulary. As shown in Table 2, VTaMo achieves the best BLEU-1, BLEU-4, and BLEURT on the official test set (BLEU-4 18.47, BLEURT 50.26) using only RGB and no large-scale pre-training—an 8.36 BLEU-4 gain over SpaMo [16]—and ranks second on ROUGE-L behind SSVP-SLT [33] (38.40 vs. 37.14). It also surpasses SHuBERT [14] (16.2 / 49.9) and Uni-Sign [21] (14.9 / 49.4), both using pose+RGB and large-scale pre-training.

Results on OpenASL. OpenASL is the hardest setting, with a vocabulary exceeding 10,000 words. As shown in Table 3, VTaMo attains 25.94 BLEU-4 and

Table 2: Results on How2Sign [11]. Bold marks the best result.

Methods	Mod.	B1	B4	RG	BLEURT
GloFE-VN [23]	Lmk	14.9	2.2	12.6	31.7
YT-ASL [39]	Lmk	14.96	1.22	25.70	29.98
SSVP-SLT [33]	RGB	43.20	15.50	**38.40**	49.60
FLa-LLM [7]	RGB	29.81	9.66	27.81	—
Uni-Sign [21]	P+R	40.2	14.9	36.0	49.4
SHuBERT [14]	P+R	—	16.2	—	49.9
SpaMo [16]	RGB	33.41	10.11	30.56	42.23
VTaMo (Ours)	RGB	**46.67**	**18.47**	37.14	**50.26**

Table 3: Results on OpenASL [36]. Bold marks the best result.

Methods	Mod.	B1	B4	RG	BLEURT
GloFE-VN [23]	Pose	21.56	7.06	21.75	36.35
Conv-GRU [1]	RGB	16.11	4.58	16.10	25.65
OpenASL [36]	RGB	20.92	6.72	21.02	31.09
C^2RL [6]	RGB	31.46	13.21	31.36	—
Uni-Sign [21]	P+R	49.35	23.14	43.22	60.40
SHuBERT [14]	P+R	—	23.2	—	60.60
SpaMo [16]	RGB	46.28	21.25	41.04	57.82
VTaMo (Ours)	RGB	**53.58**	**25.94**	**46.85**	**62.48**

62.48 BLEURT on the official test set, outperforming Uni-Sign [21] and SHuBERT [14] by 2.80 and 2.74 BLEU-4 despite their pose+RGB inputs and large-scale pre-training. Gains over SpaMo grow on the larger-vocabulary How2Sign and OpenASL, consistent with alignment helping more as complexity increases.

4.4 Ablation Study

All ablations are conducted on the Phoenix-2014T test set.

Alignment components. Removing each alignment loss in turn (Table 4) shows that all three are necessary and non-substitutable. The position-aligned contrastive loss $\mathcal{L}_{\text{contra}}$ matters most (-6.30 BLEU-4), providing token-level discriminative grounding; the local OT loss $\mathcal{L}_{\text{local}}$ is second (-5.09), as it underpins the window-based reordering that feeds ordered inputs to the decoder; and the global loss $\mathcal{L}_{\text{EMD}}$ (-2.65) adds complementary embedding-space calibration.

Language-model backbone. To test whether the gains stem from a favorable language model rather than our alignment design, we vary the backbone following SpaMo's protocol (Table 5). The ranking matches SpaMo—Flan-T5-XL is best, mT0-XL is close behind, and Llama-2 underperforms despite its larger size—confirming that scaling the LM alone does not help under the limited-data SLT regime. Critically, under *every* fixed backbone our method beats SpaMo by +2.49 to +4.54 BLEU-4, so the improvement comes from the alignment block.

Local alignment quality. Three transport-plan metrics (Table 6)—peak assignment probability, normalized entropy, and segment change rate—characterize alignment sharpness. The full model is the sharpest and most confident; removing epsilon annealing yields diffuse assignments, removing the TV loss fragments them, and removing the null token forces transitional frames onto real tokens.

Global alignment components. Table 7 isolates the global module and reports the epoch of peak development BLEU-4 as an efficiency proxy. The orthogonality constraint prevents distance distortion, scheduled activation avoids premature gradients before local plans stabilize, and the memory queue is the largest contributor to faster convergence (36 vs. 68 epochs without it).

Table 4: Alignment-component ablation on the Phoenix-2014T [1] test set. Each row removes one loss from the full model.

Configuration	B1	B4	RG
Full model	**61.32**	**28.86**	**60.24**
w/o $\mathcal{L}_{\text{local}}$	52.62	23.77	51.29
w/o $\mathcal{L}_{\text{EMD}}$	57.46	26.21	57.92
w/o $\mathcal{L}_{\text{contra}}$	51.23	22.56	50.14

Table 5: Language-model backbone ablation on Phoenix-2014T [1] (BLEU-4), following SpaMo [16]'s protocol. Δ: our gain over SpaMo at the same backbone.

Backbone	#Param	SpaMo	Ours	Δ
Flan-T5-large [9]	0.78B	19.25	22.07	+2.82
mBART-50 [37]	0.61B	10.94	13.43	+2.49
mT0-XL [28]	3.5B	24.23	27.90	+3.67
Llama-2 [38]	7B	13.86	16.82	+2.96
Flan-T5-XL [9]	3.0B	24.32	**28.86**	+4.54

Table 6: Alignment quality metrics on the Phoenix-2014T [1] development set.

Configuration	Peak ↑	Entropy ↓	Change ↓
w/o annealing	0.48	0.55	0.32
w/o TV loss	0.72	0.67	0.54
w/o null token	0.61	0.46	0.33
Full model	**0.76**	**0.18**	**0.12**

Table 7: Ablation of global alignment components on Phoenix-2014T [1]. Ep. ↓: best epoch.

Configuration	B4	Ep. ↓
Full model	**28.86**	36
w/o orthogonality	28.13	52
w/o mem. queue	27.35	68
w/o sched. act.	26.42	43

4.5 Noise Sensitivity Analysis

Since VTaMo operates on RGB input without visual-encoder fine-tuning, we assess its robustness to appearance shifts. Using Stable Diffusion [32] with ControlNet [45] for pose-conditioned synthesis, we build four controlled conditions on the entire How2Sign test set while preserving each signer's hand and body kinematics (Fig. 3): two replace the background (BG-1, outdoor park; BG-2, indoor kitchen) and two alter signer appearance (AP-1, full body shape and identity transfer to a synthetic signer; AP-2, clothing). All appearance modifications are entirely synthetic—the transferred identity in AP-1 does not represent any real individual—and are used solely for robustness evaluation. Table 8 reports single-run performance across all conditions: background replacement causes only marginal degradation ($\leq$0.42 BLEU-4) and appearance edits even smaller drops ($\leq$0.33), showing that VTaMo attends to gesture semantics rather than scene context or signer identity. Additional examples are in the supplementary material.

4.6 Qualitative Analysis

Fig. 4(a) visualizes the learned transport plans on How2Sign: they exhibit clear block-diagonal structure—contiguous frame groups aligning to individual tokens—with the null token absorbing transitional frames, and the ordering is largely monotonic with local reorderings where signing order diverges from the target.

Fig. 3: Representative augmented samples used in the robustness analysis on How2Sign [11]. (a) and (c) are background-replacement conditions generated via text-prompt editing; (b) is a full body shape and identity transfer to a synthetic signer generated via text prompt. Hand and body kinematics are preserved via ControlNet pose conditioning. The synthetic signer in (b) does not represent any real individual. All modifications are used solely for research evaluation.

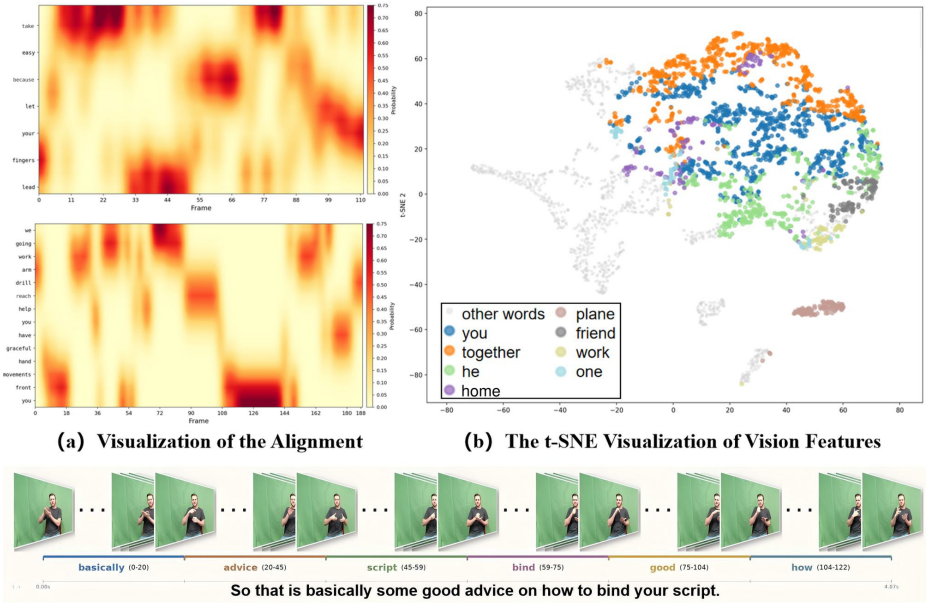
Table 8: Noise sensitivity analysis on the entire How2Sign [11] test set. BG: background replacement, AP: appearance modification. All modifications generated with Stable Diffusion and ControlNet pose conditioning.

Condition	B1	B2	B3	B4	RG
Original	46.67	32.47	24.11	18.47	37.14
BG-1 (outdoor)	46.24	32.12	23.98	18.26	37.02
BG-2 (kitchen)	45.89	31.48	23.62	18.05	36.67
AP-1 (identity)	46.15	32.07	23.91	18.14	36.92
AP-2 (clothing)	46.52	32.44	23.86	18.42	37.08

Fig. 4(c) marks the video segment assigned to each predicted content token along the timeline, and Fig. 4(b) shows a t-SNE with distinct semantic clusters, indicating alignment that is both temporally coherent at the frame level and semantically organized at the sentence level.

5 Conclusion

VTaMo achieves state-of-the-art gloss-free SLT through explicit multi-granularity cross-modal alignment: local OT with a null token, global orthogonal calibration, and position-aligned contrastive learning. Experiments on four benchmarks confirm consistent gains, and ablations verify each component's contribution, showing that explicit alignment is an effective, transferable direction for gloss-free SLT with modest overhead.



(a) Visualization of the Alignment **(b) The t-SNE Visualization of Vision Features**

(c) Explicit frame-to-pseudo-gloss alignment on a How2Sign test sentence

Fig. 4: Qualitative analysis on How2Sign [11] test samples. (a) OT transport plans from VTaMo's Sinkhorn solver: rows are visual frames and columns are pseudo-gloss tokens, with darker cells indicating higher transport mass; the block-diagonal structure reflects temporally coherent frame-to-token correspondences while the null token absorbs transitional frames. (b) t-SNE of visual features, colored by semantic category, forming distinct clusters. (c) Explicit frame-to-pseudo-gloss alignment on a test sentence: each colored interval marks the contiguous video segment assigned to a predicted content token, with representative frames shown above the timeline.

6 Limitations

Our evaluation targets offline accuracy rather than systems-level concerns such as real-time or on-device deployment; the spoken-language order is also recovered by a text-only model rather than observed directly.

Acknowledgements

This work was partially supported by ChatSign Technology, Ltd.; and the NYUAD Center for AI and Robotics (CAIR), funded by Tamkeen under the NYUAD Research Institute Award CG010. The generous computational support was provided by the HPC resources at NYU Abu Dhabi and NYU New York.

References

1. Camgöz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: CVPR (2018)
2. Camgöz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: CVPR (2020)
3. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: PLOT: Prompt learning with optimal transport for vision-language models. In: ICLR (2023)
4. Chen, Y., Wei, F., Sun, X., Wu, Z., Lin, S.: A simple multi-modality transfer learning baseline for sign language translation. In: CVPR. pp. 5120–5130 (2022)
5. Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., Mak, B.: Two-stream network for sign language recognition and translation. In: NeurIPS. vol. 35, pp. 17043–17056 (2022)
6. Chen, Z., Zhou, B., Huang, Y., Wan, J., Hu, Y., Shi, H., Liang, Y., Lei, Z., Zhang, D.: C2RL: Content and context representation learning for gloss-free sign language translation and retrieval. IEEE Transactions on Circuits and Systems for Video Technology (2025)
7. Chen, Z., Zhou, B., Li, J., Wan, J., Lei, Z., Jiang, N., Lu, Q., Zhao, G.: 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). pp. 7071–7081. ELRA and ICCL (2024)
8. Chowdhury, S.S., Chandra, S., Roy, K.: OPEL: Optimal transport guided procedure learning. In: NeurIPS (2024)
9. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. JMLR **25**(70), 1–53 (2024)
10. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: NeurIPS (2013)
11. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: A large-scale multimodal dataset for continuous american sign language. In: CVPR (2021)
12. Fu, B., Ye, P., Zhang, L., Yu, P., Hu, C., Shi, X., Chen, Y.: A token-level contrastive framework for sign language translation. In: ICASSP. pp. 1–5 (2023)
13. Gong, J., Foo, L.G., He, Y., Rahmani, H., Liu, J.: Llms are good sign language translators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18362–18372 (2024)
14. Gueuwou, S., Du, X., Shakhnarovich, G., Livescu, K., Liu, A.H.: SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In: ACL. pp. 28792–28810 (2025)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
16. Hwang, E.J., Cho, S., Lee, J., Park, J.C.: An efficient gloss-free sign language translation using spatial configurations and motion dynamics with llms. In: NAACL (2025)
17. Jiao, P., Min, Y., Chen, X.: Visual alignment pre-training for sign language translation. In: ECCV (2024)
18. Jing, L., Song, X., Zu, X., Zheng, N., Zhao, Z., Nie, L.: Vk-g2t: Vision and context knowledge enhanced gloss2text. In: ICASSP. pp. 7860–7864 (2024)
19. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: ICLR (2018)

20. Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., Li, H.: Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In: NeurIPS. vol. 33, pp. 12034–12045 (2020)
21. Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H.: Uni-sign: Toward unified sign language understanding at scale. In: ICLR (2025)
22. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
23. Lin, K., Wang, X., Zhu, L., Sun, K., Zhang, B., Yang, Y.: Gloss-free end-to-end sign language translation. In: ACL. pp. 12904–12916 (2023)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
25. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
26. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In: ACM MM. pp. 638–647 (2022)
27. Mavroudi, E., Afouras, T., Torresani, L.: Learning to ground instructional articles in videos through narrations. In: ICCV. pp. 15201–15213 (2023)
28. 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., Raffel, C.: Crosslingual generalization through multitask finetuning. In: ACL. pp. 15991–16111 (2023)
29. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: ACL (2002)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
33. Rust, P., Shi, B., Wang, S., Camgöz, N.C., Maillard, J.: Towards privacy-aware sign language translation at scale. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2024)
34. Sellam, T., Das, D., Parikh, A.: BLEURT: Learning robust metrics for text generation. In: ACL (2020)
35. Shi, B., Wu, Z., Mao, M., Wang, X., Darrell, T.: When do we not need larger vision models? In: ECCV (2024)
36. Shi, B., Brentari, D., Shakhnarovich, G., Livescu, K.: Open-domain sign language translation learned from online video. In: EMNLP (2022)
37. Tang, Y., Tran, C., Li, X., Chen, P.J., Goyal, N., Chaudhary, V., Gu, J., Fan, A.: Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401 (2020)
38. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
39. Uthus, D., Tanzer, G., Georg, M.: Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. In: NeurIPS. vol. 36 (2023)
40. Wong, R., Camgoz, N.C., Bowden, R.: Sign2gpt: Leveraging large language models for gloss-free sign language translation. In: ICLR (2024)

41. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In: EMNLP. pp. 6787–6800 (2021)
42. Yin, A., Zhao, Z., Liu, J., Jin, W., Zhang, M., Zeng, X., He, X.: Simul-slt: End-to-end simultaneous sign language translation. In: ACM MM. pp. 4118–4127 (2021)
43. Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., Zhao, Z.: Gloss attention for gloss-free sign language translation. In: CVPR. pp. 2551–2562 (2023)
44. Zhang, B., Müller, M., Sennrich, R.: Sltunet: A simple unified model for sign language translation. In: ICLR (2023)
45. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
46. Zhao, J., Qi, W., Zhou, W., Duan, N., Zhou, M., Li, H.: Conditional sentence generation and cross-modal reranking for sign language translation. IEEE TMM **24**, 2662–2672 (2021)
47. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., Zhang, D.: Gloss-free sign language translation: Improving from visual-language pretraining. In: ICCV. pp. 20871–20881 (2023)
48. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: CVPR. pp. 1316–1325 (2021)
49. Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for sign language recognition and translation. IEEE TMM **24**, 768–779 (2022)