

Shared-Encoder Contrastive Grounding for Pose-Based Indian Sign Language Translation

Gyanendra Shrestha
Florida State University
gshrestha@fsu.edu

Liting Zhang
University of West Florida
lzhang@uwf.edu

Abstract

The WSLP 2026 shared task studies sentence-level translation from Indian Sign Language pose sequences to English, with training on iSign and evaluation on a disjoint collection of source videos. We address two challenges in this setting: unstable pose scale caused by landmark errors and weak grounding between pose inputs and generated text. We introduce robust clip-level normalization and shared-encoder contrastive grounding, which aligns pose and target-text representations in a pretrained T5 encoder during training while retaining standard autoregressive decoding at inference. On a video-held-out iSign development split, contrastive grounding improves chrF from 0.102 to 0.121, while clip-level normalization further improves it to 0.128; the normalization gain is consistent across three matched seeds. A retrieval probe provides complementary evidence of improved pose-text alignment. Our final shared-task submission achieves chrF 0.16 on the official test set. However, input-independent English controls reach chrF 0.15, showing that small differences in the official score are difficult to interpret at the current level of translation accuracy. The source code is available at <https://github.com/gyan-shrestha/WSLP-2026-SLT>.

1 Introduction

The WSLP 2026 shared task studies sentence-level translation from Indian Sign Language (ISL) pose sequences to English. Training supervision comes from the public iSign corpus (Joshi et al., 2024), while the shared-task validation and test clips are drawn from a separate set of source videos with no overlap with iSign. Models must therefore transfer from iSign to a different collection of recordings and pose estimates without labeled examples from the evaluation collection.

Two problems became apparent in this setting. First, a pretrained encoder-decoder can produce

fluent English without learning a strong connection between the pose sequence and the sentence content. In preliminary experiments, unrelated clips often received the same generic predictions, indicating that the language model was easier to learn than the pose-to-text mapping. Second, frame-wise pose normalization was sensitive to unreliable shoulder estimates. Near-zero shoulder widths occasionally produced extreme coordinate values, and these failures were more common in the shared-task clips than in iSign.

We address these problems with robust clip-level normalization and shared-encoder contrastive grounding. Pose sequences are projected into the embedding space of a pretrained T5 encoder-decoder. During training, the corresponding target sentence is encoded by the same T5 encoder, and a contrastive objective aligns the pooled pose and text representations alongside the translation loss. The text branch is removed at inference. For pose preprocessing, scale is estimated from the median shoulder width over the clip rather than independently in each frame, reducing the effect of individual landmark failures.

Controlled experiments on held-out iSign videos show that contrastive grounding provides the largest improvement over cross-entropy training, with clip-level normalization providing a further gain. Repeating the normalization comparison across three matched seeds yields higher chrF with normalization in all three runs. A retrieval-based grounding probe provides complementary evidence that the contrastive objective affects pose-text alignment directly: the cross-entropy-only checkpoint performs near chance, while the contrastive checkpoints achieve substantially higher retrieval recall. Several other changes, including velocity features, augmentation, a larger T5 model, and masked-pose pretraining, do not show a similarly clear improvement.

The shared-task evaluation also exposes a limi-

tation of interpreting the official score in isolation. The same trained checkpoint receives different chrF scores under different output-length constraints, while input-independent English can recover a substantial portion of the score. We therefore report the controlled method evaluation separately from the analysis of the official test output.

The main contributions are:

- An analysis of the WSLP evaluation showing that, at the current level of translation accuracy, chrF is strongly affected by decoding choices and English surface form in addition to the signed input. A fixed English sentence selected using only the output register, without using pose, scores within 0.02 chrF of the top official submission (Section 4.6).
- A pose-to-text formulation that jointly trains autoregressive translation and pose-text contrastive alignment through the same pretrained T5 encoder, without requiring a separate text encoder at inference.
- A clip-level normalization procedure that reduces the effect of unstable pose scale across recording collections and provides an additional improvement over contrastive training alone.

The model combines established components: pose projection into a pretrained T5 encoder-decoder (Uthus et al., 2023), contrastive visual-text alignment (Zhou et al., 2023), and an InfoNCE objective (van den Oord et al., 2018; Radford et al., 2021). Their combination is computationally inexpensive, adds no inference-time parameters, and allows the alignment objective to act directly on the encoder states used by the translation decoder.

2 Related Work

Pose-to-text translation. The iSign benchmark introduced a large ISL-English corpus together with pose-to-text baselines for open-domain translation, with its strongest reported pose baseline reaching BLEU-4 1.47 on the conventional 0-100 scale (Joshi et al., 2024). A related pose-to-text formulation projects pose features into the embedding space of a pretrained T5 model and trains the encoder-decoder directly for translation (Uthus et al., 2023). Our backbone follows this general design, while the WSLP task introduces a different evaluation setting in which the test clips come from

a disjoint collection and no labeled target-domain development set is available.

Visual-language alignment. Early neural SLT systems were developed on constrained corpora such as PHOENIX14T (Camgoz et al., 2018), where jointly modeling recognition and translation is effective (Camgoz et al., 2020). Gloss-free SLT work has used visual-language pretraining to strengthen the connection between visual input and target text. GFSLT-VLP (Zhou et al., 2023), for example, learns aligned visual and textual representations before translation. We use the same broad contrastive principle, but place it inside end-to-end translation training: pose and target text share the T5 encoder, and the alignment loss regularizes the encoder states attended to by the decoder.

Motion and ISL modeling. Motion-aware modeling has shown stronger performance on the iSign test split, with BLEU-4 reaching 8.58 when optical flow is combined with multi-scale spatial encoding and a language-model decoder (Chowdhury and Sanyal, 2025). Although this setting differs from WSLP in both input representation and evaluation data, the result indicates that video-derived motion can provide information that sparse pose coordinates do not capture. For isolated ISL recognition, MediaPipe keypoints combined with a temporal convolutional network achieve 54% top-1 accuracy across 4,361 classes (Reddy and Kamakshi, 2025). Our work considers a different setting: sentence-level, open-domain translation from the pose sequences provided for WSLP Task 1.

3 Methodology

The proposed model translates ISL pose sequences directly into English using a pretrained T5 encoder-decoder. Figure 1 shows the overall pipeline. A temporal frontend first converts the pose sequence into T5-compatible embeddings. These embeddings are processed by the T5 encoder and passed to the decoder for autoregressive translation. During training, the target sentence is also encoded by the same T5 encoder, and a contrastive objective encourages paired pose and text representations to remain close while separating other sentences in the batch. The final training objective combines translation and contrastive losses:

$$\mathcal{L} = \mathcal{L}_{\text{MT}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} \quad (1)$$

where $\mathcal{L}_{\text{MT}}$ is token-level cross-entropy and

$\lambda_{\text{con}} = 0.5$. The target-text branch is used only during training. Register conditioning and multi-seed ensemble decoding complete the full system.

3.1 Pose representation and normalization

The released pose features contain 576 MediaPipe Holistic (Lugaresi et al., 2019) landmarks per frame at 25 fps. We retain 103 landmarks covering both hands, the upper-body joints, and 40 facial points around the lips, eyes, and brows. Each frame is represented by the corresponding two-dimensional coordinates.

Pose scale varies with camera distance and with the quality of landmark detection. A frame-wise normalization based on shoulder width is particularly unstable when one shoulder is missing or poorly localized, since a very small estimated width can produce extremely large coordinates. To reduce this sensitivity, each frame is centered on its shoulder midpoint, while scale is estimated from the median shoulder width over the full clip. The sequence is then standardized to zero mean and unit variance. This retains relative articulatory motion while reducing large frame-to-frame changes caused by pose-estimation errors.

3.2 Pose-to-text architecture

The temporal frontend contains a linear projection followed by two stride-2 convolutions, LayerNorm, and a learned scaling parameter. The two convolutions reduce the temporal resolution by a factor of four and give each output position a receptive field of approximately nine input frames. The resulting sequence is projected to the dimensionality of fln-t5-base (Raffel et al., 2020; Chung et al., 2024) and supplied through inputs_embeds. T5 then performs sequence encoding and autoregressive generation in the same way as for ordinary token embeddings.

Direct pose embeddings also need to match the scale of the representations seen during T5 pre-training. In the checkpoint used here, normal token embeddings have a mean norm near 300, whereas the LayerNorm output from the pose frontend has norm $\sqrt{d_{\text{model}}} = 27.7$. Feeding the latter without rescaling produced an initial loss near 609. A learned scalar is therefore initialized from the measured mean norm of the checkpoint embedding table before joint training.

Split	Clips	Videos	References
iSign, raw	124,194	3,726	yes
Task validation	5,256	1,804	no
Task test	5,333	1,813	no
<i>Video-id overlap, iSign $\cap$ evaluation: 0</i>			

Table 1: Data used for WSLP Task 1. The evaluation clips come from source videos that do not occur in iSign.

3.3 Shared-encoder contrastive grounding

Cross-entropy alone optimizes the probability of the target sentence but does not directly require the pose representation to distinguish one sentence from another. In preliminary runs, this often led to similar generic English predictions for unrelated pose sequences. The contrastive objective adds sentence-level discrimination to the same encoder used for translation.

Let H_i^p and H_i^t denote the shared T5 encoder states for pose sequence x_i and target sentence y_i . Mean pooling followed by ℓ_2 normalization gives z_i^p and z_i^t . For a batch of size B , the contrastive loss is

$$\mathcal{L}_{\text{con}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{sim}(z_i^p, z_i^t)/\tau)}{\sum_{j=1}^B \exp(\text{sim}(z_i^p, z_j^t)/\tau)} \quad (2)$$

where sim is cosine similarity and $\tau = 0.07$. This is the InfoNCE objective of (van den Oord et al., 2018), applied across modalities as in (Radford et al., 2021). Other target sentences in the batch serve as negatives. Since the two branches share the T5 encoder, the additional objective acts on the representation used by the translation decoder and introduces no separate text encoder at inference.

3.4 Dataset and preprocessing

Training uses the public iSign corpus (Joshi et al., 2024), while validation and test evaluation follow WSLP 2026 Task 1. The raw iSign file contains 124,194 sentence-level pairs from 3,726 source videos. The WSLP release contains 5,256 validation clips from 1,804 videos and 5,333 test clips from 1,813 videos. Comparing the source-video identifiers shows no overlap between iSign and the shared-task evaluation videos.

The WSLP validation release does not contain reference translations, so development experiments use an iSign holdout. A total of 1,940 clips are

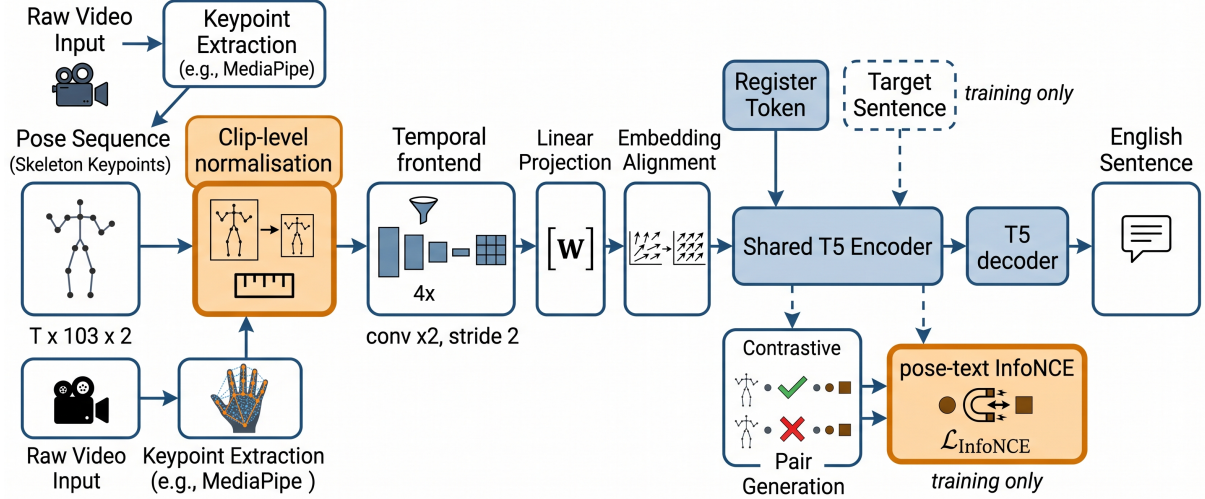


Figure 1: Overview of the proposed pose-to-text translation model. Orange boxes are the two changes introduced in this work: clip-level pose normalization and shared-encoder contrastive grounding. The pose sequence is normalized, reduced in time by a convolutional frontend, and projected into the embedding space of a pretrained T5 encoder–decoder together with a register control token. During training the target sentence is encoded by the *same* T5 encoder and an InfoNCE objective aligns the pooled pose and text representations, so the contrastive term regularizes exactly the states the decoder attends to and adds no parameters at inference. Dashed boxes and arrows are used only during training and are removed at inference.

held out by source video rather than by clip, avoiding nearby segments from the same recording on both sides of the split. The resulting development set remains in-domain with respect to iSign and is therefore used for controlled model comparison rather than as a direct estimate of WSLP test performance.

Before splitting the data, 3,408 iSign targets consisting of a single token, bare digits, or punctuation-like text are removed, leaving 120,786 usable pairs. The most frequent target in the raw file is 2. These examples are excluded because they strongly favor degenerate short predictions.

Two uid patterns in the evaluation data correspond to different target styles: capitalized, punctuated prose and lowercase fragments. Since the uid identifies this distinction at inference time, a control token is prepended to the pose embeddings to specify the target register. Training labels for this token are derived from the target surface form.

3.5 Implementation details

Training uses AdamW with separate learning rates for the randomly initialized pose frontend and the pretrained T5 backbone. The frontend uses 1×10^{-4} and T5 uses 3×10^{-5} , with cosine scheduling and 8% linear warmup. A single learning rate of 1×10^{-4} for both components caused the loss to rise from 5.8 to 15.1 near the peak of the schedule.

Decoder	f1an-t5-base (223M)
Pose frontend	6.1M parameters
Input	103 keypoints $\times$ 2, 25 fps
Max input / target	512 frames / 48 tokens
Optimizer	AdamW, $\beta = (0.9, 0.98)$
LR, frontend / T5	$10^{-4} / 3 \times 10^{-5}$
Schedule	cosine, 8% linear warmup
Weight decay / clipping	0.01 / 1.0
Batch size / epochs	64 / 14
Precision	bfloat16
Contrastive weight	0.5
Contrastive temperature	0.07
Development set	1,940 clips, held out by video
Hardware / runtime	$1 \times$ B200 / ≈ 60 min per model
Ensemble	5 models (chrF 0.15)
Submitted output	single model + fixed suffix
Submitted decoding	beam 8, length penalty 2.0 no-repeat 3-grams, min. 36 tokens

Table 2: Training and decoding configuration.

Models are trained for 14 epochs with batch size 64, weight decay 0.01, gradient clipping at 1.0, and bfloat16 precision. The contrastive weight is 0.5 and the temperature is 0.07.

The submitted model averages token log-probabilities from five runs that differ only in random seed and data order. Decoding uses beam size 8, length penalty 2.0, no-repeat 3-grams, and a minimum output length of 36 tokens. The local evaluation follows the shared-task metric stack: unsmoothed BLEU (Papineni et al., 2002) with the NLTK punkt tokenizer, ROUGE (Lin, 2004)

Configuration	chrF	ROUGE-L	BLEU
Cross-entropy only	0.102	0.076	0.0089
+ shared-encoder contrastive grounding	0.121	0.089	0.0113
+ robust clip-level normalization	0.128	0.094	0.0115

Table 3: Ablation results on the video-held-out iSign development split.

via $\text{rouge} = 1.0$, and chrF (Popović, 2015) via sacreBLEU (Post, 2018). As a scorer check, a constant hello prediction receives chrF 0.0317 locally and 0.03 from the official scorer.

4 Results

4.1 Shared-task performance

On the official WSLP 2026 Task 1 test set, our submitted system reaches chrF 0.16, ROUGE 0.08, and BLEU 0.00, ranking second among five submitted systems. Official chrF scores range from 0.13 to 0.17, while all five systems obtain BLEU at or below 0.02. Under the same decoding constraints, both the five-model ensemble and the single strongest model reach chrF 0.15; the remaining 0.01 comes from a fixed suffix appended to every prediction, analyzed in Section 4.5. Ensembling therefore improves the development scores reported below without producing a measurable gain on the evaluation collection. Because the official test references are not released, the effect of individual model components is evaluated on the video-held-out iSign development split.

4.2 Ablation study

Because the official test references are not released, component-level comparisons use the video-held-out iSign development split. Section 4.4 shows that performance on this split does not always predict official chrF, so the ablation results characterize the method on held-out iSign videos rather than establish the same component ordering on the WSLP collection.

The controlled development experiments in Table 3 isolate the two main changes in the proposed method. Cross-entropy training alone reaches chrF 0.102, ROUGE-L 0.076, and BLEU 0.0089. Adding shared-encoder contrastive grounding increases all three metrics and raises chrF to 0.121. Replacing frame-wise scaling with clip-level normalization further raises chrF to 0.128.

The normalization change also addresses a measurable difference in input scale. Under frame-wise

Seed	chrF	ROUGE-L	BLEU
0	0.1240	0.0953	0.0142
1	0.1305	0.0849	0.0090
2	0.1301	0.1024	0.0113
mean $\pm$ sd	0.128 ± 0.003	0.094 ± 0.007	0.0115 ± 0.0021
range	0.0065	0.0175	0.0052

Table 4: Three runs of the final configuration, with contrastive grounding and clip-level normalization, using different random seeds.

Seed	chrF	ROUGE-L	BLEU
0	0.1212	0.0893	0.0113
1	0.1166	0.0952	0.0123
2	0.1024	0.0911	0.0071
mean $\pm$ sd	0.113 ± 0.008	0.092 ± 0.003	0.0102 ± 0.0022
range	0.0187	0.0059	0.0052

Table 5: Three runs of the contrastive configuration *without* normalization, same seeds as Table 4.

shoulder scaling, training features have standard deviation 1.99 and a 99.9th percentile of 3.29, while the corresponding test statistics are 3.04 and 13.92. After clip-level standardization, both collections have unit variance. Together with the ablation result, this indicates that clip-level normalization reduces a measurable scale mismatch between the two collections.

The contrastive objective also learns a measurable pairing signal. Its loss decreases from 4.45 to 2.66, compared with $\ln 64 = 4.16$ at the batch size used here. This shows that the shared encoder learns to distinguish the paired sentence from the other sentences in its batch, but batch-level discrimination does not by itself establish semantic grounding. The rank-based retrieval probe in Section 4.6 provides a more direct test: the correct sentence ranks substantially above chance for the contrastive checkpoints and near chance for the cross-entropy-only checkpoint. Translation performance remains the primary evidence for the method; the retrieval probe provides complementary evidence of alignment beyond batch discrimination.

4.3 Run-to-run variation

Table 4 reports three runs of the final configuration, which combines contrastive grounding with clip-level normalization. chrF ranges from 0.1240 to 0.1305, with a mean of 0.128 ± 0.003 .

To assess the normalization gain across random seeds, we repeated the contrastive configuration without clip-level normalization using the same three seeds as the final configuration (Table 5).

Normalization improves chrF in all three

Configuration	chrF	ROUGE-L	BLEU
Contrastive (reference)	0.121	0.089	0.0113
+ velocity features	0.121	0.081	0.0099
+ pose augmentation	0.116	0.083	0.0095
+ flan-t5-large	0.106	0.074	0.0083
+ masked-pose pretraining	0.118	0.090	0.0078
+ all of the above	0.110	0.079	0.0084

Table 6: Single-run modifications on the video-held-out iSign development split, each varying one component from the contrastive configuration.

matched comparisons, by +0.0028, +0.0139, and +0.0277, with an average improvement of +0.015. The magnitude varies across runs, but the direction is consistent across all three seeds. ROUGE-L and BLEU vary more across runs, so small differences on those metrics should be interpreted cautiously.

Other single-run modifications, including velocity features, pose augmentation, flan-t5-large, and masked-pose pretraining, do not improve beyond the variation observed here. Table 6 reports each, varying one component from the contrastive (no-normalization) configuration; none of these differences is larger than that configuration’s own seed range (0.0187, Table 5), so we report them as unresolved rather than as evidence that the component is harmful.

We discuss two of these results in more detail below.

Velocity features. Appending first-order frame differences to the keypoint stream did not help. This differs from prior work showing substantial gains from optical flow computed directly on video (Chowdhury and Sanyal, 2025). First differences of 103 keypoint coordinates provide a much coarser motion signal than dense optical flow, so this negative result should not be interpreted as evidence against motion representations more broadly.

Masked-pose pretraining. We pretrained the frontend by masked reconstruction over iSign together with the unlabeled WSLP validation and test poses, as a form of unsupervised adaptation to the evaluation collection. It reduced development loss while producing lower translation scores. Reconstructing masked pose frames rewards features that predict neighboring coordinates, which is not the same as features that carry sentence-level content.

4.4 Cross-collection development

Performance on the iSign development split does not consistently predict performance on the WSLP collection. Across three submitted configurations, development BLEU rises from 0.0073 to 0.0188, whereas official chrF changes from 0.12 to 0.13 and then falls to 0.11. The configuration with the strongest development result is therefore the weakest of the three under official chrF.

This difference is consistent with the collection split described in Section 3: the development set contains held-out iSign videos, while the shared-task clips come from a separate set of source videos. Holding out complete videos reduces within-recording leakage, but it does not reproduce the target distribution.

4.5 Effect of decoding on the official score

The ablation above measures changes to the trained model, while the final WSLP score also depends on how that model is decoded. Table 7 evaluates one fixed checkpoint under six different output-length constraints. Increasing the mean output length from 11.6 to 25.9 words raises chrF from 0.13 to 0.15 even though the model itself is unchanged, and lengthening further to 60.9 words brings it back down to 0.14. Over the same sweep ROUGE-L decreases monotonically from 0.09 to 0.06 and BLEU remains 0.00, so the two metrics move in opposite directions for a model that never changed. For scale, that 0.02 swing from decoding alone is half the 0.04 spread separating all five submitted systems.

The final row of Table 7 appends the same English sentence to every 25.9-word prediction. This text is identical for all 5,333 clips and does not use the pose, yet it increases chrF from 0.15 to 0.16. The final 0.01 in the submitted score therefore comes from decoding rather than a change in the learned translation model.

4.6 Input-independent controls

The decoding results suggest that part of the chrF score may come from target-language surface form rather than the signed input. To estimate this effect directly, Table 8 evaluates several controls that do not use the pose sequence.

A fixed sentence repeated for every test clip reaches chrF 0.14, while selecting a fixed sentence using only the known output register raises the score to 0.15. The random-sentence control also

Mean output length	chrF	ROUGE	BLEU
11.6 words	0.13	0.09	0.00
19.6 words	0.14	0.09	0.00
25.9 words	0.15	0.08	0.00
32.6 words	0.15	0.08	0.00
42.1 words	0.15	0.07	0.00
60.9 words	0.14	0.06	0.00
25.9 words + fixed suffix (43.9)	0.16	0.08	0.00

Table 7: Official test scores for one fixed checkpoint under different output-length constraints. The final row appends the same fixed suffix to every prediction.

Prediction rule	chrF	ROUGE	BLEU
constant hello	0.03	0.00	0.00
one fixed sentence	0.14	0.09	0.00
fixed sentence, register-matched	0.15	0.08	0.00
random training sentence	0.13	0.04	0.00

Table 8: Sanity checks on the official test set. None of these controls use the pose sequence.

reaches 0.13. None of these predictions uses information from the signed content. All four controls were evaluated with the official scorer and are therefore directly comparable with the leaderboard scores in Section 4. Three fall within the 0.13–0.17 chrF range spanned by the five submitted systems, and the best input-independent control is only 0.02 below the top submission. In contrast, the constant hello control reaches only 0.03 chrF, indicating that the higher scores depend on plausible English surface form rather than arbitrary output.

Because the contrastive objective aligns pose and text in a shared space, the trained encoder can also be used as a retrieval index, giving a pose-conditioned control that the controls above lack. We embedded all training clips, embedded each development clip, and returned the nearest training sentence. On the development split, this reaches chrF 0.154, ROUGE-L 0.067, and BLEU 0.0069, compared with chrF 0.121 for generated output from the same checkpoint on the same split. We do not report this result on the official test set because no target-domain retrieval index is available for the WSLP evaluation collection. Retrieved sentences are fluent and human-written but, on inspection, unrelated in meaning to the query clip: the retrieval control separates two properties that chrF does not distinguish, since a retrieved sentence has ideal surface fluency and essentially no content accuracy, and still scores above generated output that is partially correct.

Checkpoint	R@1	R@5	R@10
Cross-entropy only	0.0005	0.0026	0.0057
+ contrastive	0.0072	0.0294	0.0567
+ normalization	0.0351	0.1191	0.1964
chance ($N=1,940$)	0.0005	0.0026	0.0052

Table 9: Pose-to-text retrieval recall on the 1,940-clip development split. R@1/5/10 reports the fraction of pose queries for which the paired reference sentence appears among the top 1, 5, or 10 candidates. Chance recall is shown for reference.

Retrieval recall as a grounding probe. The contrastive loss measures discrimination within a training batch, but does not show how well pose and text representations align on held-out examples. We therefore evaluate pose-to-text retrieval on the 1,940-clip development split. For each clip, we encode its pose and reference sentence separately with the shared T5 encoder, mean-pool the representations, and rank all 1,940 reference sentences by cosine similarity to the pose query. We report recall at 1, 5, and 10 (R@1/5/10), together with chance performance for a candidate pool of this size. The cross-entropy-only checkpoint provides a control because it was not trained with the contrastive objective.

The cross-entropy-only checkpoint performs near chance at each k , while both contrastive checkpoints perform substantially above chance. Contrastive grounding reaches R@10 of 0.057, compared with a chance level of 0.0052, and adding normalization raises R@10 to 0.196. The median rank of the correct sentence also improves from 158 to 49 out of 1,940 candidates after normalization. These results provide additional evidence that contrastive training produces pose–text alignment beyond the decrease in its training loss, while normalization further improves retrieval on this probe. Absolute retrieval performance nevertheless remains limited: at R@10 of 0.196, the correct sentence falls outside the top 10 for about four fifths of the queries.

Taken together, these results show that substantial character-level overlap can be obtained from English surface form alone in this evaluation setting, whether or not the prediction is conditioned on the pose. They do not imply that chrF is generally unsuitable for sign language translation; rather, they show that chrF is difficult to interpret on its own when translation accuracy is this low.

Reference: <i>Do you know whose idea it was to set up Clin d’Oeil? It was him!</i>	
CE only	This is why they have a lot of people.
+ contrastive	It is a very important importance of the world.
Reference: <i>Hello! Hello! Hello!</i>	
CE only	This is why you will be very happy.
+ contrastive	Let me tell you.
Reference: <i>It is all about Sign Language. Which is very important.</i>	
CE only	This is why it was very happy.
+ contrastive	This is why he was very happy.

Table 10: Development examples from cross-entropy and contrastive training.

4.7 Qualitative analysis

The generated sentences are often fluent even when their content is poor. Under cross-entropy alone, unrelated development clips frequently receive similar sentences beginning with “This is why”. Contrastive training reduces this repetitive pattern and increases output variation, although the examples in Table 10 remain far from their references. Before decoding calibration, one sentence accounts for 9.2% of the 5,333 official test predictions and only 23% of predictions are distinct.

The qualitative results are consistent with the quantitative ablation but also limit the strength of the grounding claim from generation alone. Shared-encoder contrastive training improves the measured translation scores and reduces an obvious repetition mode, but the generated sentences remain weakly aligned with the reference content; the retrieval probe in Section 4.6 is what supports a grounding claim beyond generation quality, and even there the absolute recall remains low.

Taken together, these experiments separate two sources of improvement. The controlled ablation shows that contrastive grounding and clip-level normalization improve the trained model on the available development data. The decoding and input-independent controls show that changes of a similar magnitude in official chrF can also arise without changing the model itself. This makes small chrF differences difficult to interpret on their own in the current low-accuracy setting.

5 Conclusion

This paper presents a pose-to-text approach for the WSLP 2026 Indian Sign Language translation task. The method combines robust clip-level pose normalization with shared-encoder contrastive ground-

ing, using the same T5 encoder for pose and target text during training. Controlled experiments on held-out iSign videos show improved translation scores after contrastive grounding and clip-level normalization. Across three matched seeds, normalization improves chrF in every comparison. A rank-based retrieval probe provides complementary evidence that contrastive training improves pose-text alignment, with a further improvement after normalization.

The shared-task evaluation reveals a separate difficulty. Performance on held-out iSign does not consistently predict performance on the WSLP collection, and the official chrF score is sensitive to output length and input-independent English content. These results support the proposed training method while also showing that, at the current level of translation accuracy, small changes in chrF do not necessarily correspond to better use of the signed input.

Limitations

The method is evaluated on a single sign language and a single training corpus (iSign); whether shared-encoder contrastive grounding transfers to other sign languages or pose-extraction pipelines is untested. Register conditioning relies on a uid pattern specific to the WSLP evaluation data, and would need a learned or predicted register signal in a setting where this side information is unavailable at inference. The contrastive objective improves translation scores and reduces a specific repetition failure mode, and a rank-based retrieval probe (Section 4.6) shows the resulting pose representation is above-chance predictive of its paired sentence, but the qualitative examples in Table 10 show generation itself does not establish accurate content grounding: generated sentences often remain unrelated to the reference in meaning even when fluent, and even the retrieval probe’s absolute recall is low (R@10 of 0.20 on a 1,940-clip pool). Because official test references are not released for this shared task, all component-level comparisons rely on a held-out development split, which Section 4.4 shows can rank configurations differently than the official evaluation; the true ranking of the tested components on the WSLP test collection therefore remains uncertain. The retrieval probe is subject to the same limitation.

Absolute translation quality also remains low: our system obtains BLEU 0.00, and every submit-

ted system obtains BLEU at or below 0.02. At this level of performance, small chrF differences cannot by themselves establish better use of the signed input.

Acknowledgments

We thank anonymous reviewers for their valuable feedback.

References

- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Anal Roy Chowdhury and Debarshi Kumar Sanyal. 2025. Enhancing indian sign language translation via motion-aware modeling. In *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pages 39–46.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Abhinav Joshi, Romit Mohanty, Mounika Kanakanti, Andesha Mangla, Sudeep Choudhary, Monali Barbate, and Ashutosh Modi. 2024. isign: A benchmark for indian sign language processing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10827–10844.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Tatigunta Bhavi Teja Reddy and Vidhya Kamakshi. 2025. Pose-based temporal convolutional networks for isolated indian sign language word recognition. In *Proceedings of the Workshop on Sign Language Processing (WSLP)*, pages 47–50.
- David Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A large-scale, open-domain American sign language-English parallel corpus. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.