What's the Point? Spatial Grammar & Index Resolution for Sign Language Recognition

Anonymous ACL submission

Abstract

Sign language models are predominantly trained with gloss-sequence or text supervision, thereby under-modeling non-lexical and productive constructions. One comparatively tractable instance is spatial *indexing*: pointing gestures that assign discourse entities to spatial loci for subsequent co-reference, which lexicon-centric objectives largely fail to capture. We present the first targeted evaluation of indexing in Sign Language Recognition, showing that despite comprising 10-15% of signing content, indexing is poorly recovered. We introduce a framework for training and evaluating indexing experts, establishing a baseline for index-aware sign language modeling. Our approach decomposes spatial reference resolution into index detection and discourse entity linking. The resulting mention representations enable automatic annotation and non-lexical structure modeling, and serve as an auxiliary indexing expert that augments a frozen SLR model at inference time.

1 Introduction

Most Sign Language Recognition (SLR) and Translation (SLT) pipelines map sign language to gloss-label sequences (Zhou et al., 2021a,b; Yin et al., 2021a; Chen et al., 2022; Zhang et al., 2023; Bragg et al., 2019; De Coster et al., 2024) or text (Zhou et al., 2023; Wong et al., 2024; Chen et al., 2024; Asasi et al., 2025; Chen et al., 2025; Kim et al., 2025; Sincan et al., 2025). This reflects supervision in current benchmarks, where signed phrases are aligned with lexical or coarse lexicon-like tokens, often under an assumption of one-to-one correspondence. Implicitly, this framework assumes that language in both modalities can be represented as a linearly ordered sequence drawn from a finite discrete lexicon (Brown et al., 2026; De Coster et al., 2024). This assumption captures only part of the linguistic structure of sign languages (Baker, 2016; Bragg et al., 2019).

In spontaneous signed discourse, approximately 40% of signs are non-lexical, consisting of productive constructions that exploit three-dimensional space, iconicity, and context (Fenlon et al., 2014). A particularly frequent and tractable instance of non-lexical sign classes is *spatial indexing* (pointing), whereby signers establish entities and roles by associating them with spatial loci and subsequently refer back to them through pointing signs (Liddell, 1980; Kegl, 1987; Cormier et al., 2010).

Although visually simple at the local level, indexing depends on discourse state and productive use of space: similar manual forms can serve pronominal, locative, or determiner-like functions, and referents may be reassigned as narratives evolve (Van Hoek, 1992; Emmorey and Lillo-Martin, 1995; Ozyurek et al., 2010; Cormier et al., 2013). Systems trained primarily under lexical objectives are therefore not optimized to (i) reliably distinguish indexing from lexically similar signs and (ii) maintain referential consistency across phrases.

We curate a training and evaluation framework for index-awareness, using linguistic corpora and an established benchmark with manually annotated pointing. We evaluate a state-of-the-art SLR model, which performs poorly in recovering pointing-related glosses. This motivates an explicit *indexing expert* that complements lexically oriented SLR models with targeted mechanisms for detecting indexing events and tracking discourse referents.

In this work, we present a modular pipeline for **index detection** and **discourse entity linking**. We decompose spatial reference into two stages: index detection and entity linking, separating local recognition from discourse-level tracking. Concretely, we implement (i) an Index Proposal Network (IPN) that identifies indexing segments, and (ii) a differentiable online Entity Linking Module (ELM) that incrementally clusters detected mentions into entity representations. The resulting state supports both corpus-scale annotation and downstream tasks.

Our contributions are as follows.

1. We show that spatial indexing is under-modeled by lexicon-based objectives.
2. We curate a training and evaluation framework for index-awareness in SLR.
3. We propose a two-stage index modeling approach using an Index Proposal Network and an online memory module for discourse-level entity tracking.
4. We introduce a targeted evaluation protocol for index-awareness and show that the resulting system improves indexing recovery and reduces WER as an auxiliary module for a frozen SLR model at inference time.

1.1 Task Definition

We model spatial reference in signed discourse at three levels: (i) sign-level *index detection*, (ii) document-level *entity linking* over detected mentions, and (iii) optional *expert integration* into a downstream SLR model.

Input Representation. A sign language document is represented as an ordered sequence of T temporally localized segments, $D = (x_1, \ldots, x_T)$, where each segment x_t corresponds to a short clip aligned to a gloss-level annotation boundary and is represented by pose features extracted from the underlying video frames. The segment granularity remains fixed throughout; the processing mode differs across stages as described below.

Index Detection. For each segment x_t, the IPN predicts whether the segment realizes an indexing sign. Segments predicted as indexing are treated as index mentions and passed to the linking stage.

Entity Linking. Detected index mentions are processed sequentially in document order. The model maintains a memory of discourse entities and assigns each mention either to an existing entity or to a newly created one, inducing a clustering over mentions within the document.

Index-Supported CSLR. At inference time, we incorporate index proposals and predicted entity assignments as an auxiliary bias for a frozen SLR model. Segments align with the evaluation units of the target gloss-based benchmark and are processed sequentially in document order. The index confidence score increases the relative weight of pointing-related tokens when indexing is detected, and the entity memory state promotes consistency across repeated mentions.

2 Background and Related Works

2.1 Non-Lexical Signs and Discourse

Sign languages consist of both established and productive lexicons. Lexical signs are relatively conventionalized forms with stable phonological parameters and meanings (Baker, 2016). While language-dependent, they account for approximately 60% of signs in discourse (Fenlon et al., 2014). The remaining content consists largely of non-lexical and productive constructions, including pointing signs, depicting constructions, and enactment (Brown et al., 2026), which exploit the three-dimensional signing space around the signer (Baker, 2016; De Coster et al., 2024). Through spatial location, movement, iconicity, and embodiment, they encode reference, structure events, and represent participants or actions. Unlike fully lexicalized signs, their interpretation is more context-dependent and emerges from the interaction between articulatory form and discourse structure.

2.2 Spatial Indexing and Annotations

Indexing signs establish reference by assigning discourse entities to spatial loci in signing space (Liddell, 1980; Kegl, 1987), a mechanism observed across many sign languages (Aronoff et al., 2003; Mathur and Rathmann, 2010). Beyond their pronominal use, where they refer to previously introduced entities (Van Hoek, 1992; Emmorey and Lillo-Martin, 1995; Cormier et al., 2010), they also serve locative or determiner-like functions (e.g., "here", "that") (Ozyurek et al., 2010; Cormier et al., 2013). Interpretation depends on discourse context rather than local appearance. Referents may be grounded in present entities or assigned to abstract loci for absent ones, and once introduced, loci can be reused via pointing, gaze, or agreement (Emmorey and Lillo-Martin, 1995; Schlenker, 2018). Unlike spoken pronouns, indexing signs lack fixed grammatical features (e.g., gender, number); instead, meaning derives from spatial loci and evolving discourse, which may also be reassigned over time (Cormier et al., 2010; Yin et al., 2021b). Despite their prevalence (10–16% of tokens; Table 1), indexing signs are inconsistently and coarsely annotated in SLR and SLT benchmarks, often collapsing functions into gloss labels and omitting referential structure. This limits discourse-level evaluation and motivates treating indexing as a structured prediction task: detecting indexing events and resolving their associated entities.

2.3 Sign Language Recognition

SLR maps signed input, typically RGB video or pose representations, to gloss labels (Bragg et al., 2019; De Coster et al., 2024). The task is commonly divided into isolated SLR (ISR), where each clip contains a single sign (Jiang et al., 2024; Wong et al., 2025; Ranum et al., 2024b; Wu et al., 2025; Hu et al., 2021; Zhao et al., 2024; Jiang et al., 2021; Ranum et al., 2024a), and SLR, where unsegmented signing streams must be transcribed into ordered gloss sequences (Camgoz et al., 2018; Guo et al., 2025; Raude et al., 2024; Camgoz et al., 2017; Hao et al., 2021; Koller et al., 2016; Pu et al., 2019; Cui et al., 2019). Progress has been driven by stronger visual backbones and transformer-based sequence models, as well as retrieval-style objectives that align signs and text. However, available benchmarks largely rely on subtitle-aligned interpreted broadcasts or gloss-annotated datasets from controlled laboratory environments (Brown et al., 2026; De Coster et al., 2024), thereby encouraging lexical sequence modeling rather than explicit discourse-aware modeling. Spatially grounded phenomena such as indexing are therefore rarely treated as structured reference mechanisms and are typically collapsed into gloss tokens.

2.4 Coreference Resolution and Memory-Based Entity Modeling

Coreference resolution in NLP clusters textual mentions into entity representations (Lee et al., 2017; Joshi et al., 2020). Memory-augmented models maintain bounded, online-updated entity representations to handle long documents (Liu et al., 2019; Xia et al., 2020). Similar to our approach, but designed for spoken language, Manikantan et al. (2024) decouples mention detection from entity assignment and updates a working memory via learned ranking, a decomposition we adapt to visually grounded discourse. Yin et al. (2021b) explored sign-language coreference over gloss annotations; we instead operate on the visual modality and generate supervision automatically with an LLM. We use the term *linking* rather than coreference resolution, as indexing encompasses not only anaphoric reference, but also locative and determiner-like functions realized through spatial loci.

3 Method

Our methodology models indexing as a two-stage process: (i) index detection from pose-based segments, and (ii) discourse entity linking via an online memory mechanism. The resulting entity state can optionally be integrated to refine SLP models. Figure 1 illustrates the pipeline.

3.1 Pose Representation and Encoder

Each segment is represented by poses extracted from video. Body pose is obtained with SMPL (Pavlakos et al., 2019) (8 upper-body 3D joints; Appendix F) and hand pose via WiLoR (Potamias et al., 2025) (21 joints per hand), forming a unified skeleton that captures handshape, motion, and spatial configuration. Poses are normalized for scale and viewpoint to support cross-dataset generalization. Features are extracted using the SL-GCN (Jiang et al., 2021), which models frames as spatial graphs and captures temporal dynamics via spatio-temporal convolutions. The encoder outputs a segment embedding $h_g \in \mathbb{R}^d$ via global pooling, where g indexes gloss segments and d is the embedding dimension.

3.2 Index Proposal Network (IPN)

Given a segment embedding h_g, the IPN predicts whether the segment corresponds to an index sign. A lightweight classification head maps h_g to an index confidence score $p_g \in [0, 1]$. Segments exceeding a decision threshold τ are treated as index mentions and passed to the entity linking stage. The encoder and classification head are trained using gloss-level indexing annotations from linguistically curated corpora.

3.3 Entity-Linking Memory (ELM)

Detected index mentions are processed in gloss-segment order. We maintain a dynamically growing entity memory $\{e_k\}_{k=1}^{K}$, where k indexes entities, building on memory-augmented coreference models (Liu et al., 2019; Manikantan et al., 2024).

Each segment embedding h_g is projected into a dedicated linking space to decouple entity resolution from index classification, yielding a mention embedding $m_g \in \mathbb{R}^{d'}$. Since the encoder is trained with a binary objective, its representations capture coarse pointing cues and may suppress fine-grained spatial information required to distinguish individual referents. We therefore augment m_g with geometric features f_g derived from pointing direction,

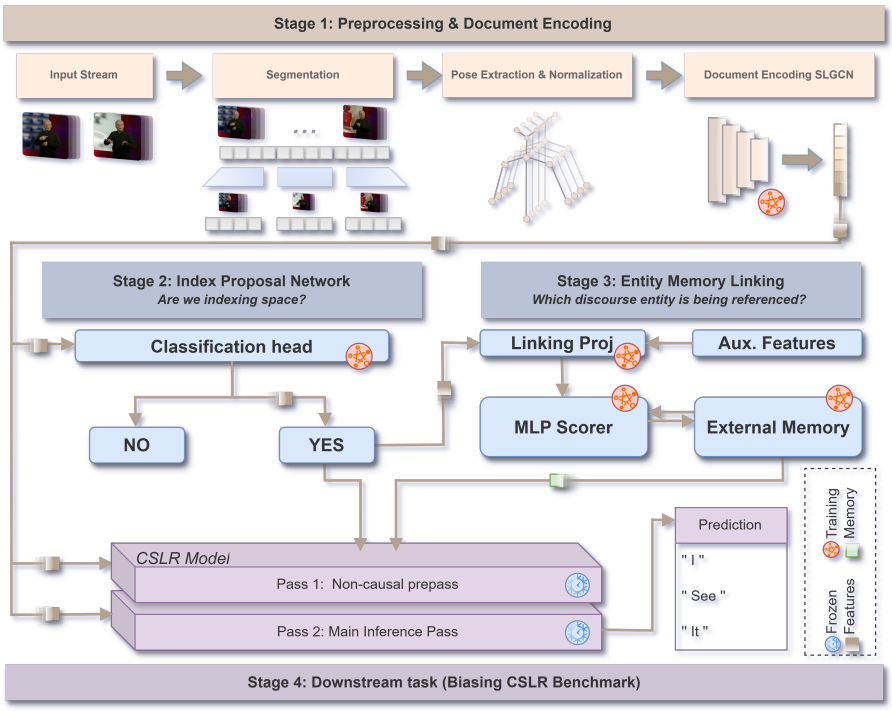

Figure 1: Overview of the proposed pipeline. Gloss-level pose segments are encoded with the SLGCN, then processed by the IPN. Detected mentions are passed to the ELM module, which maintains an external entity memory. The resulting signals can be injected as inference-time biases into a frozen SLR model.

spatial locus, and trajectory (Appendix F.1).

Given the current memory, we compute compatibility scores with each entity e_k:

$$s_{g,k} = \mathrm{MLP}([m_g;\, e_k;\, m_g \odot e_k;\, f_g]), \quad (1)$$

together with a fixed null score $s_{g,\varnothing} = 0$ for introducing a new entity. The assignment is obtained via $\arg\max$ over $\{s_{g,k}\}_{k=1}^{K} \cup \{s_{g,\varnothing}\}$ and trained using a softmax cross-entropy objective with LLM-derived labels.

If a mention links to an existing entity e_k, we update its representation via gated interpolation:

$$\begin{aligned} e_k &\leftarrow \alpha_g\, e_k + (1-\alpha_g)\, m_g, \\ \alpha_g &= \sigma(\mathrm{MLP}([e_k;\, m_g]))\,, \end{aligned} \quad (2)$$

where $\alpha_g \in (0,1)$ depends on both the entity state and the incoming mention. Otherwise, m_g initializes a new memory slot. During training, we use teacher forcing to mitigate error propagation.

3.4 Inference-Time SLR Integration

We integrate IPN and ELM into a frozen SLR model, CSLR2 (Raude et al., 2024). Each video is processed by CSLR2 to produce frame-level logits $\ell \in \mathbb{R}^{F\times V}$, where F is the number of frames and V the vocabulary size. Time-aligned gloss annotations partition each pose sequence into segments g. We introduce two inference-time interventions: (i) an IPN boost injecting index confidence into pointing-token logits, and (ii) an ELM prior biasing predictions via cluster-level entity assignments.

Detection boost w_{IPN}. Each gloss segment g with temporal extent $[t^g_{\text{start}}, t^g_{\text{end}}]$ is uniformly sampled (12 frames) and processed by the IPN to obtain a confidence score $p_g \in [0,1]$. This score is broadcast to frames within the segment, forming a segment mask $\mathbf{q}^{(g)} \in \{0,1\}^F$. Let $\mathbf{v}_{\mathcal{P}}$ denote the indicator over pointing tokens. The logits are updated as:

$$\ell \leftarrow \ell + w_{\text{IPN}} \cdot \mathbf{q}^{(g)} p_g \mathbf{v}_{\mathcal{P}}^{\top} \quad (3)$$

Entity Linking prior w_{ELM}. Segments with IPN score $p_g \geq \tau$ are assigned to ELM clusters. For each cluster k, we compute a consensus pointing token $v_k^{\star} \in \mathcal{P}$ using a non-causal full-episode prepass over CSLR2 predictions, aggregating votes across all segments in the cluster (including synonym-group expansion; Appendix I). Let $\mathbf{e}_{v_k^{\star}} \in \{0,1\}^V$ denote its one-hot representation. Each segment g in cluster k receives:

$$\ell \leftarrow \ell + w_{\text{ELM}} \cdot \mathbf{q}^{(g)} \mathbf{e}_{v_k^{\star}}^{\top} \quad (4)$$

Data	# Doc / Sent	V_T / V_G	# T_G	# T_{Idx}
BSLCP	409 / 8K	5K / 7K	74.8K	11.4K
BOBSL	35 / 4.4K	35K / 5K	32K	4.0K
MDGS	318 / 63.5K	22K / 9K	357K	57K

Table 1: Statistics from relevant corpora and benchmark. Alongside the number of documents and sentences, we report the vocabulary size of all text tokens V_T and glosses V_G, and the number of gloss tokens T_G and index-related tokens T_{idx}.

4 Experiments

4.1 Datasets and Benchmark

We use two annotated corpora (BSLCP (Schembri et al., 2014), MDGS (Konrad et al., 2020)) to train the modules. BSLCP mark indexing instances with a `"PT:"` prefix, and MDGS maps signs to coarse pronominal and non-specified indexing categories (`"Ich"`, `"Du"`, `"INDEX"`).

For SLR evaluation, we use CSLR-TEST (Raude et al., 2024), a 6-hour subset of the subtitle-aligned SENT-TEST split of BOBSL (Albanie et al., 2021) spanning 35 episodes. Annotators verified automatically segmented clips and assigned each a free-form English label with optional sign-type tags for non-lexical phenomena (e.g., `*D`: depictions, `*P`: pointing). The `*P` tag is broadly defined and may include non-referential signs with pointing handshapes (e.g., *what*, *why*) (Raude et al., 2024). The CSLR2 baseline excludes most `*P`-tagged segments when evaluating, we retain them to analysing spatial indexing.

General indexing set (GIS). CSLR-TEST contains 4.7K `*P`-tagged segments. Of these, 3.2K indexing instances (approximately 10% of all tokens) are not associated with any lexical gloss: $g_{\text{GIS}} = \{g \mid \text{Type}(g) = \texttt{*P} \wedge \text{Lex}(g) = \emptyset\}$, where Type($g$) is the non-lexical sign class and Lex(g) is the lexical gloss label of segment g.

Indexing proxy subset (IPS). To enable entity-specific evaluation, we isolate `*P`-labelled segments that overlap temporally with referential glosses, yielding 856 instances. Since `*P` conflates referential and non-referential cases, we define a filtered subset using a curated index-candidate vocabulary spanning pronouns, reflexives, and deictic forms: $g_{\text{IPS}} = \{g \mid \text{Type}(g) = \texttt{*P} \wedge \text{Lex}(g) \in \mathcal{I}\}$, where $\mathcal{I}$ = {me, my, mine, you, your, he, her, our, they, their, myself, yourself, this, that, here, there, *i, yours, him, his, she, it, its, we, us, ours, them, himself, herself, itself, themselves, these, those*}. This defines the subset g_{IPS} with identifiable referents. Only the first 16 items in $\mathcal{I}$ are present in the benchmark vocabulary. Although most `*P` segments are excluded from CSLR2 evaluation, g_{IPS} is partially retained in the lexically oriented evaluation when corresponding tokens exist in the pretrained CSLR2 vocabulary.

4.2 Automatic Entity Label Generation

To obtain supervision for entity linking at scale, we generate cluster labels via a two-stage pipeline. Candidate mentions are extracted from transcripts using POS tagging and rule-based span selection, after which an LLM assigns entity cluster IDs at document level under structural constraints. The resulting clusters are refined by a rule-based post-processing step that merges fragments sharing unambiguous person-pointing labels per signer and document (Appendix E).

4.3 Baseline: Index Recovery in CSLR

As described in Sec. 4.1, the CSLR2 evaluation protocol excludes most non-lexical signs, including instances from g_{GIS}. To analyse indexing behavior, we augment the evaluation sequences by reinserting these segments and re-evaluate using indexing-specific metrics. We consider two conditions: $\mathbf{BL_{Lex}}$, which follows the official CSLR2 protocol (WER including g_{IPS}), and $\mathbf{BL_{IR}}$, which reinserts the excluded g_{GIS} segments. We report WER_{All}, $\text{WER}_{\text{Index}}$, and WER_{Lex} as defined in Appendix H. For g_{GIS}, correctness is defined as a prediction $l_p \in \mathcal{I}$, reflecting token-level index recovery rather than referential accuracy (Sec. 6).

Table 3 shows that WER_{All} increases from 69.5% to 70.5% after index restoration, while $\text{WER}_{\text{Index}}$ reaches 98.4 % and Index IoU drops to 6.1%. These results indicate that indexing is strongly under-modelled by current SLR objectives and motivate an explicit modeling component. Our reproduced WER (69.5%) differs from the 65.5% reported in (Raude et al., 2024) but matches the performance reported for the released model weights.

4.4 Evaluation of Index Detection

We train the IPN on index-related labels from BSLCP and MDGS and evaluate within- and cross-dataset performance (Table 2). The model generalises well across datasets, suggesting that aspects of indexing signs transfer across sign languages despite signer and lexical variation. Joint training (BM) further improves robustness.

Exp.	IoU$_I$ (%)	WER$_{All}$	WER$_{Index}$	WER$_{Lex}$
BL$_{Lex}$	15.8	69.5	91.3	77.5
BL$_{IR}$	6.1	70.5	98.4	77.5

Table 3: CSLR2 baseline and index-restored performance on BOBSL-CSLR test set. BL$_{\text{Lex}}$ follows the standard CSLR2 protocol, while BL$_{\text{IR}}$ reinserts excluded index segments. IoU$_I$ measures index segment overlap; WER$_{\text{All}}$, WER$_{\text{Lex}}$, and WER$_{\text{Index}}$ denote overall, lexical, and index-specific error rates.

IPN/ELM	ECA	F1	WER$_{\text{All}}$	WER$_{\text{Lex}}$	WER$_{\text{Index}}$
BM$_{cw4}$/B	0.62	0.44	70.1	78.4	64.3
BM$_{cw4}$/BM	0.65	0.43	70.2	78.4	64.7

Table 4: Entity linking and downstream SLR performance on BSLCP and BOBSL. BSLCP metrics are averaged over three seeds; BOBSL WER is reported using the checkpoint at $w_{\text{IPN}} = 10$, $w_{\text{ELM}} = 60$. ECA denotes entity cluster accuracy; WER$_{\text{All}}$, WER$_{\text{Lex}}$, and WER$_{\text{Index}}$ denote overall, lexical, and index-related word error rates, respectively. Standard deviations are reported where applicable.

To reduce false positives that propagate into entity linking (*lexical bleed*), we prioritise precision over recall. We introduce a class-weighted variant (cw4), upweighting non-index errors by a factor of 4 in the cross-entropy loss to penalise spurious index predictions. Combined with a high-confidence threshold ($\tau = 0.90$; Appendix G), this yields a more conservative model. The resulting BM-cw4 model is used for downstream SLR experiments.

4.5 Evaluation of Entity Linking

We train the ELM on automatically generated entity clusters (Sec. 4.2) using indexed mentions from the frozen IPN. Table 4 reports clustering performance on a test set built from automatically labelled BSLCP data. All models use the BM cw4 encoder, and we compare BSLCP-only training with joint training on BM.

Joint training on MDGS and BSLCP slightly improves Entity Cluster Accuracy (ECA $= 0.65$ vs. 0.62) but yields similar F1 (0.43 vs. 0.44 ± 0.01), indicating better cluster structure but no gain in per-mention precision. This improvement does not transfer to downstream CSLR: BSLCP-only training achieves lower WER on BOBSL under the setup in the next section. We therefore use the BSLCP-trained linker for SLR integration due to better label alignment and reduced variability.

4.6 Index-Supported CSLR

We integrate the IPN and ELM as inference-time biases into a frozen CSLR2 model. We evaluated using a standard word error rate (WER$_{\text{All}}$) and two diagnostic subgroup metrics isolating indexing behavior. Across all evaluations, out-of-vocabulary IPS glosses are mapped to in-vocabulary pointing-form equivalents (Appendix I).

Trade-off between index recovery and lexical bleed. Figure 2 shows a grid search over the detection boost w_{IPN} and entity linking prior w_{ELM}. The heatmaps report (i) WER$_{All}$, (ii) WER$_{Index}$ (indexing sub-vocabulary), and (iii) WER$_{Lex}$ (lexical tokens). Increasing w_{IPN} is the primary driver of WER$_{Index}$ reduction, while w_{ELM} provides a further consistent gain, as shown in Table 5.However, large w_{IPN} causes *lexical bleed*, as misclassified non-index segments are biased toward pointing tokens, raising WER$_{Lex}$, creating a trade-off between index recovery and lexical accuracy that produces a visible valley in WER$_{All}$ around $w_{\text{IPN}} \approx 8$.

We select the operating point $w_{\text{IPN}} = 8$, $w_{\text{ELM}} = 10$ at the floor of this valley. Compared to the

Train	Eval	BalAcc	Macro F1	Prec$_{\text{Lex}}$	Rec$_{\text{Lex}}$	Prec$_{\text{Index}}$	Rec$_{\text{Index}}$
B	B	0.85	0.85	$0.85_{\pm 0.01}$	$0.86_{\pm 0.01}$	0.85	$0.85_{\pm 0.01}$
B	M	0.76	0.76	$0.83_{\pm 0.01}$	0.67	0.72	$0.86_{\pm 0.01}$
B	BOB	$0.77_{\pm 0.01}$	$0.68_{\pm 0.01}$	0.96	$0.87_{\pm 0.01}$	$0.35_{\pm 0.02}$	$0.66_{\pm 0.02}$
M	B	$0.75_{\pm 0.01}$	$0.74_{\pm 0.01}$	$0.69_{\pm 0.01}$	$0.92_{\pm 0.01}$	$0.87_{\pm 0.01}$	$0.58_{\pm 0.02}$
M	M	0.87	0.87	0.88	0.86	0.87	0.88
M	BOB	0.72	$0.70_{\pm 0.01}$	0.95	$0.92_{\pm 0.02}$	$0.41_{\pm 0.04}$	$0.52_{\pm 0.02}$
BM-cw4	B	0.82	0.81	0.76	0.93	$0.91_{\pm 0.01}$	$0.70_{\pm 0.01}$
BM-cw4	M	0.85	0.85	0.81	0.91	0.90	0.79
BM-cw4	BOB	$0.73_{\pm 0.01}$	$0.72_{\pm 0.01}$	0.95	0.94	$0.47_{\pm 0.01}$	$0.52_{\pm 0.01}$
BM	B	0.85	0.85	$0.83_{\pm 0.01}$	$0.89_{\pm 0.01}$	$0.88_{\pm 0.01}$	$0.82_{\pm 0.01}$
BM	M	0.87	0.87	0.87	$0.86_{\pm 0.01}$	$0.86_{\pm 0.01}$	0.88
BM	BOB	$0.78_{\pm 0.01}$	$0.69_{\pm 0.01}$	0.96	$0.87_{\pm 0.02}$	$0.37_{\pm 0.03}$	$0.68_{\pm 0.04}$

Table 2: IPN classifier performance (mean $\pm$ std over 3 seeds). B = BSLCP, M = MDGS, and BOB = BOBSL. L and I denote lexical and indexing classes, respectively. Decision threshold is set to 0.5; zero standard deviations are omitted for readability.

Ablation	WER_{All}	$\text{WER}_{\text{Index}}$	WER_{Lex}
Baseline	70.5	96.3	77.4
No-memory (IPN)	69.1	66.5	76.9
No-proposal (ELM)	70.1	89.5	77.4
IPN + ELM	68.3	59.6	77.1

Table 5: Ablation of IPN and ELM components in CSLR2. "No-memory (IPN)" removes entity linking memory while retaining index proposal detection, "No-proposal (ELM)" removes index proposal while retaining entity linking, and "IPN + ELM" corresponds to the full system with $w_{\text{IPN}} = 8$ and $w_{\text{ELM}} = 10$.

baseline ($\text{WER}_{Index} = 96.2$, $\text{WER}_{All} = 70.5$, $\text{WER}_{Lex} = 77.5$), this configuration reduces WER_{Index} to 59.6 and WER_{All} to 68.3. WER_{Lex} is slightly improved to 77.1 due to the removal of false-positive lexical segments, with these gains outweighing the introduction of new false-positive index predictions, confirming that substantial improvements in indexing performance are achieved without degrading lexical performance.

Fig. 3 presents a Sankey flow diagram for the full system ($w_{\text{IPN}} = 8$, $w_{\text{ELM}} = 10$). Each flow traces a ground-truth pointing token to the coarse referential category of the predicted token, grouped into pronoun and deictic classes (Appendix I). The overview panel shows all GIS and IPS flows, while the detail panel isolates IPS tokens to highlight category distinctions. Under the baseline, only 6.1% of ground-truth index tokens are recovered; the IPN detection boost alone raises this to 61.5%, and the full system reaches 71.2%, confirming that entity linking provides a consistent gain on top of detection.

Figure 4 shows an example of a qualitative evaluation for the ELM, with additional results in Appendix C. The model resolves clusters across sentences and captures subtle pointing distinctions, including second- and third-person forms. Failure cases include missed mentions and occasional over-clustering. Overall, the model identifies substantially more identity-related references than the baseline. We observe strong grouping of visually similar signs, with clusters often remaining consistent across sentences. However, errors include over-merging of spatially proximate entities and under-merging across long temporal ranges.

5 Conclusion

We presented a modular framework for spatial indexing in signed discourse, addressing a structural gap in lexicon-centric sign language modeling pipelines. The framework establishes index-aware modeling as a structured task over sign streams, rather than a by-product of SLR supervision.

We first showed that a strong CSLR2 model fails to recover indexing tokens despite competitive overall performance, motivating explicit modeling of index structure. Our approach decomposes the problem into index detection and discourse entity linking: a pose-based IPN generalises across data distributions and can be calibrated toward high precision, while a differentiable online ELM tracks discourse

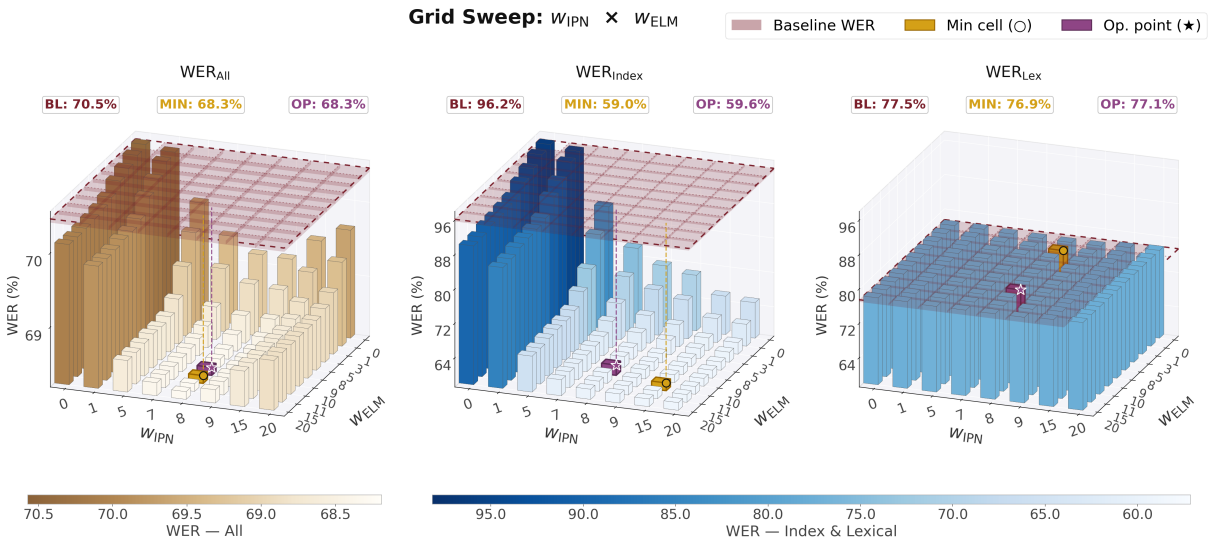

Figure 2: Coreference and downstream SLR performance on BSLCP and BOBSL. BSLCP metrics are averaged over three seeds; BOBSL results use the checkpoint at $w_{\text{IPN}} = 10$, $w_{\text{ELM}} = 60$. Standard deviations are shown where applicable.

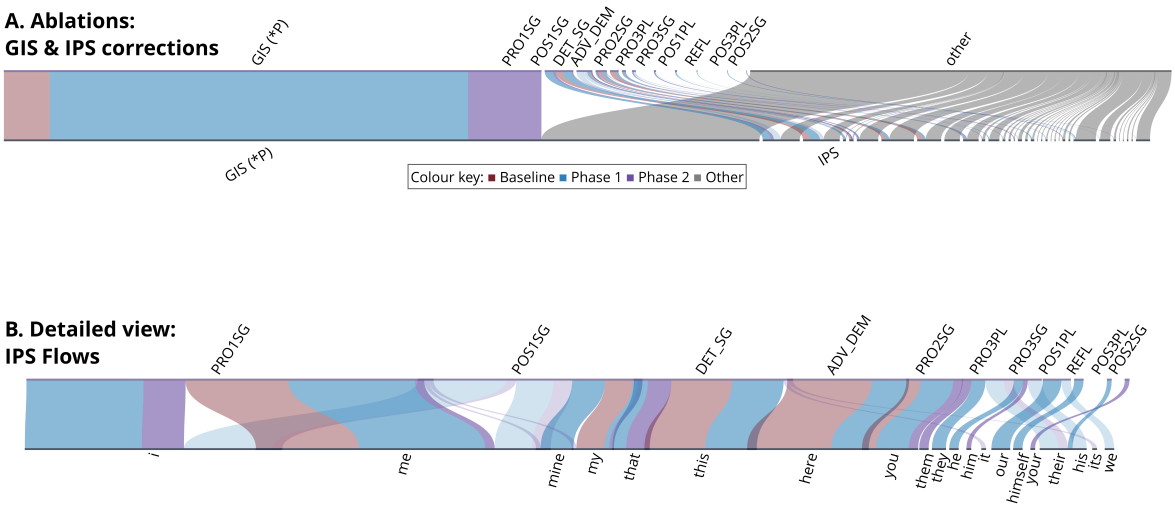

Figure 3: Flow-based analysis of pointing-token predictions under the full system configuration ($w_{\text{IPN}} = 8$, $w_{\text{ELM}} = 10$). Each flow traces a ground-truth pointing token to the coarse referential class of the predicted token, grouped into pronoun and deictic categories (Appendix I). The top panel shows all g_{GIS} and g_{IPS} instances, while the bottom panel isolates g_{IPS} tokens to highlight entity-specific behavior. Flows terminating in the correct category indicate successful recovery of referential structure; remaining flows correspond to lexical predictions outside the pointing vocabulary. Compared to the baseline, a larger fraction of flows terminate in the correct referential classes, indicating that the external indexing expert improves recovery of pointing expressions.

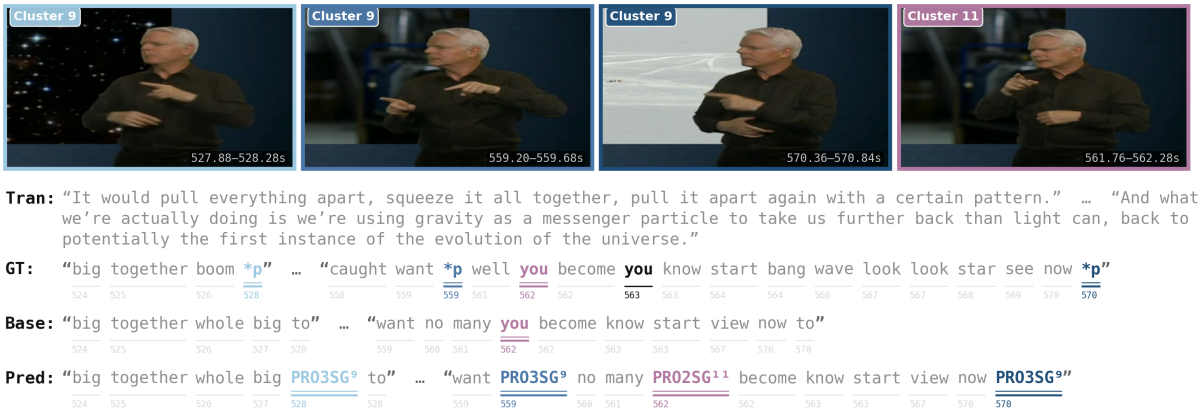

Figure 4: Samples from episode 6164207930460576679. The sample showcases both cross-sentence cluster recognition and in-sentence cluster separation between a third person singular entity and a second person 'you' being referenced. The base CSLR2 only picked up the 'you' index sign, but was not able to pick up any of the third-person references. The second reference to the second person is not detected.

entities via geometric and learned features. Automatic cluster supervision from an LLM pipeline reduces the need for large-scale coreference annotations. When integrated as inference-time biases into a frozen CSLR2 backbone, the system reduces $\text{WER}_{\text{Index}}$ from 96.2 to 59.6 while leaving WER_{Lex} stable, demonstrating that discourse-aware components can complement existing recognition systems without retraining. Future work will extend this treatment of spatial grammar to other productive constructions such as depicting signs and role shift, improve discourse-level supervision, for both recognition and translation models.

6 Limitations

Distribution Shift Across Signing Contexts. Training on conversational corpora from Deaf signers and evaluating on interpreted broadcast introduces substantial domain shift in factors including signer style and identity, recording conditions, and linguistic scope. While the training data consist of conversational signing by Deaf participants in controlled settings, BOBSL contains translated content produced by professional (often hearing) interpreters. This can lead to shifts in patterns of language use and references. For example, referents such as *I* and *you* may change dynamically over time, and pointing direction may reflect the source broadcast rather than the signer. Interpretation further introduces variability due to rapid decision-making under temporal constraints, which can alter referential structure and spatial grounding.

Labels and evaluation metrics. BOBSL annotations rely on automatic temporal segmentation with human verification, so ground-truth labels may be incomplete. For example, we observe some apparent false positives correspond to valid indexing signs absent from the annotations. This limits the reliability of quantitative evaluation and motivates complementary qualitative analysis.

In the BOBSL evaluation, the primary recovery metric for g_{GIS} measures the proportion of ground-truth `*P` segments for which any token in $\mathcal{I}$ is predicted. This definition is necessary because `*P` annotations do not encode explicit referential categories, preventing exact matching at the token level. As a consequence, semantically incorrect substitutions (e.g., predicting *they* for a self-reference *me*) are counted as correct under the g_{GIS} evaluation. Therefore, $\text{WER}_{\text{Index}}$ captures index token presence rather than true referential accuracy and should be interpreted as a lower bound on semantic error. More precise evaluation would require referential annotations aligned with individual `*P` instances, which are currently only available for g_{IPS}.

IPN architectural limitations. Detection errors primarily arise from lexical signs with index-like handshape configurations or transient pointing-like movements (e.g. `"what"`, `"one"`, `"why"` in BSL). Missed detections are caused by occlusion, coarticulation, and imprecise temporal boundaries, and can propagate into broken entity-linking chains. A common failure mode is confusion between personal and possessive forms (*I* vs. *my*): frontal occlusion obscures index-finger extension, while singular–plural distinctions, encoded in subtle 3D trajectory variations, are confounded by coarticulation and are not reliably captured. Cross-linguistic variation (e.g. between DGS and BSL) in the handshape realisation of possessive forms introduces additional challenges for multilingual training. Several informative cues not yet integrated include gaze direction and torso shifts; viewpoint normalisation may further suppress the latter.

ELM architectural limitations. Cluster supervision derived from the LLM pipeline is generated without a visual signal, so coreference chains that are visually grounded but linguistically ambiguous introduce noise, particularly in longer documents. The $s_{t,\varnothing}$ parameter, which controls the threshold for opening new entity slots, impacts cluster granularity: more negative values produce fewer, broader clusters, while less negative values fragment the entity space. We explored both fixed $s_{t,\varnothing}$ offsets and a learned $s_{t,\varnothing}$ head, but the absence of ground-truth coreference chains on BOBSL makes rigorous calibration intractable. In the downstream SLR experiment, $s_{t,\varnothing}$ primarily regulates cluster grouping sensitivity; WER converges across $s_{t,\varnothing}$ settings at moderate cluster weight, and a principled calibration method remains an open problem.

Qualitative samples and evaluation. Cluster formation is further constrained by the underlying recognition pipeline, and entity clusters are not assigned explicit contextual names. Automatic entity naming remains an open challenge, and is expected to improve as sign language recognition models become more robust.

SLR integration operates purely at inference time by biasing token logits, without joint training of the recognition backbone. While this enables plug-and-play deployment, it limits the ability to learn direct index-aware representations. Consequently, improvements in index recovery translate only modestly to WER_{All}, partly due to the class imbalance between indexing and lexical tokens.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset.

Mark Aronoff, Irit Meir, Carol Padden, and Wendy San-

dler. 2003. Classifier constructions and morphology in two sign languages. In *Perspectives on Classifier Constructions in Sign Languages*.

S. Asasi, M. I. Lakhal, O. M. Sincan, and R. Bowden. 2025. Beyond gloss: A hand-centric framework for gloss-free sign language translation. In *Proceedings of the 36th British Machine Vision Conference (BMVC)*.

Anne Baker. 2016. *The Linguistics of Sign Languages: An Introduction*. John Benjamins Publishing Company.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*.

Matt Brown, Oline Ranum, Edward Fish, Heidi Proctor, Bencie Woll, Richard Bowden, and Kearsy Cormier. 2026. Signgpt and the visual language toolkit. In *Proceedings of the LREC 2026 12th Workshop on the Representation and Processing of Sign Languages: Language in Motion*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yutong Chen, Ronglai Zuo, Fuchun Wei, Yiming Wu, Shiguang Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. In *Advances in Neural Information Processing Systems*.

Zhe Chen, Benjia Zhou, Yifan Huang, Jun Wan, Yuxin Hu, Hongan Shi, Yijun Liang, Zhen Lei, and David Zhang. 2025. C2rl: Content and context representation learning for gloss-free sign language translation and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhe Chen, Benjia Zhou, Jie Li, Jun Wan, Zhen Lei, Ning Jiang, Qi Lu, and Guoying Zhao. 2024. Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of LREC-COLING 2024*.

Kearsy Cormier, Adam Schembri, and Bencie Woll. 2010. Diversity across sign languages and spoken languages: Implications for language universals. *Lingua*.

Kearsy Cormier, Adam Schembri, and Bencie Woll. 2013. Pronouns and pointing in sign languages. *Lingua*.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2024. Machine translation from signed to spoken languages: state of the art and challenges. *Universal Access in the Information Society*.

Karen Emmorey and Diane Lillo-Martin. 1995. Processing spatial anaphora: Referent reactivation with overt and null pronouns in american sign language. *Language and Cognitive Processes*.

Jordan Fenlon, Adam Schembri, Ramas Rentelis, David Vinson, and Kearsy Cormier. 2014. Using conversational data to determine lexical frequency in british sign language: The influence of text type. *Lingua*.

Jianyuan Guo, Peike Li, and Trevor Cohn. 2025. Bridging sign and spoken languages: Pseudo gloss generation for sign language translation.

Aiming Hao, Yuecong Min, and Xilin Chen. 2021. Self-mutual distillation learning for continuous sign language recognition. In *Proc. of the IEEE/CVF international conference on computer vision*.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton Aware Multi-modal Sign Language Recognition . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*.

Judy Kegl. 1987. Coreference relations in american sign language. In *Studies in the Acquisition of Anaphora*.

J. Kim, H. Jeon, J. Bae, and H. Y. Kim. 2025. Leveraging the power of mllms for gloss-free sign language translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

O Koller, O Zargaran, H Ney, and R Bowden. 2016. Hybrid cnn-hmm for continuous sign language recognition.

R. Konrad, T. Hanke, G. Langer, D. Blanck, J. Bleicken, I. Hofmann, O. Jeziorski, L. Konig, S. Konig, R. Nishio, A. Regen, U. Salden, S. Wagner, S. Worseck, O. Böse, E. Jahn, and M. Schulder. 2020. Meine dgs annotiert: Public corpus of german sign language.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Scott K. Liddell. 1980. *American Sign Language Syntax*. De Gruyter Mouton.

Fei Liu, Luke Zettlemoyer, and Jacob Eisenstein. 2019. The referential reader: A recurrent entity network for anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Kawshik Manikantan, Shubham Toshniwal, Makarand Tapaswi, and Vineet Gandhi. 2024. Major entity identification: A generalizable alternative to coreference resolution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Gaurav Mathur and Christian Rathmann. 2010. Verb agreement in sign language. In *Sign Languages: A Cambridge Language Survey*.

Asli Ozyurek, Inge Zwitserlood, and Pamela Perniss. 2010. Locative expressions in signed languages: A view from turkish sign language (tid). *Linguistics*.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 2025. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junfu Pu, Wengang Zhou, and Houqiang Li. 2019. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Oline Ranum, Gomèr Otterspeer, Jari I. Andersen, Robert G. Belleman, and Floris Roelofsen. 2024a. 3D-LEX v1.0 – 3D lexicons for American Sign Language and Sign Language of the Netherlands. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*.

Oline Ranum, David Wessels, Gomèr Otterspeer, Erik J Bekkers, Floris Roelofsen, and Jari I. Andersen. 2024b. The NGT200 dataset - geometric multi-view isolated sign recognition. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*.

Charles Raude, K R Prajwal, Liliane Momeni, Hannah Bull, Samuel Albanie, Andrew Zisserman, and Gül Varol. 2024. A tale of two languages: Large-vocabulary continuous sign language recognition from spoken language supervision.

Adam Schembri, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. 2014. British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008–2014 (Second Edition).

Philippe Schlenker. 2018. Visible meaning: Sign language and the foundations of semantics. *Theoretical Linguistics*.

O. M. Sincan, J. H. Low, S. Asasi, and R. Bowden. 2025. Gloss-free sign language translation: An unbiased evaluation of progress in the field. *Computer Vision and Image Understanding*.

Karen Van Hoek. 1992. Conceptual spaces and pronominal reference in american sign language. *Nordic Journal of Linguistics*.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2gpt: Leveraging large language models for gloss-free sign language translation. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2025. Signrep: Enhancing self-supervised sign representations. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Kepeng Wu, Zecheng Li, Weichao Zhao, Hezhen Hu, Wengang Zhou, and Houqiang Li. 2025. Cross-modal consistency learning for sign language recognition. In *Proc. of the Computer Vision and Pattern Recognition Conference*.

Patrick Xia, João Sedoc, and Benjamin Van Durme. 2020. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2021a. Simulslt: End-to-end simultaneous sign language translation. In *Proc. of the 29th ACM International Conference on Multimedia*.

Kayo Yin, Kenneth DeHaan, and Malihe Alikhani. 2021b. Signed coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

B. Zhang, M. Müller, and R. Sennrich. 2023. Sltunet: A simple unified model for sign language translation. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Yunyao Mao, Min Wang, and Houqiang Li. 2024. Masa: Motion-aware masked autoencoder with semantic alignment for sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

B. Zhou, Z. Chen, A. Clapés, J. Wan, Y. Liang, S. Escalera, Z. Lei, and D. Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li. 2021a. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

H. Zhou, W. Zhou, Y. Zhou, and H. Li. 2021b. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*.

Supplementary Material

Appendix Overview

A Conceptual Illustration of Method

B Error Analysis: IPN Index Proposal

C Qualitative Results: ELM BOBSL

D Numerical Values of SLR Grid Sweeps

E Automatic Entity Label Generation

F Input Feature Layout and ELM Auxiliary Features

G IPN Threshold Ablation

H WER_{All}, WER_{Index} and WER_{Lex}

I Synonym-Based Evaluation Mapping

J SLR Integration: Inference-Time Guardrails

K Hyperparameters

L Model Size and Budget

M Disclaimer: Use of AI-Assisted Development

N License and Code

A Conceptual Illustration of Method

Figure 5 showcases a conceptual illustration of the proposed method.

B Error Analysis: IPN

We analyse the failure cases of the IPN (trained on BSLCP and MDGS, evaluated on BOBSL with strict timestamp matching) to assess whether misclassifications are systematic or incidental. Figure 6 shows representative failure cases; each panel displays the centre frame of the segment, with border colour indicating error type: red = FP (non-pointing sign predicted as index), blue = FN (pointing sign missed).

The false positives are largely systematic: the mispredicted signs frequently exhibit index-like properties, such as an extended or body-directed handshape oriented toward a location where referential pointing typically occurs. This is unsurprising, since the IPN operates on skeletal motion alone and has no access to the discourse context that would disambiguate a pointing gesture from a visually similar lexical sign. The false negatives reveal a complementary failure mode: many missed pointing signs are co-articulated with a co-occurring lexical sign, producing a composite articulation whose skeletal trajectory is pulled away from a canonical pointing gesture. Without access to wider contextual features or an explicit model of co-articulation, the network fails to isolate the indexical component within these compound productions.

C Qualitative Results of ELM on BOBSL

Figures 7–9 show predicted entity clusters on BOBSL, where only sparse referential labels are available. We show the predicted cluster membership alongside the subtitles for each segment, demonstrating that the cluster structure is reflected in coherent lexical output and visual similarity across instances assigned to the same entity.

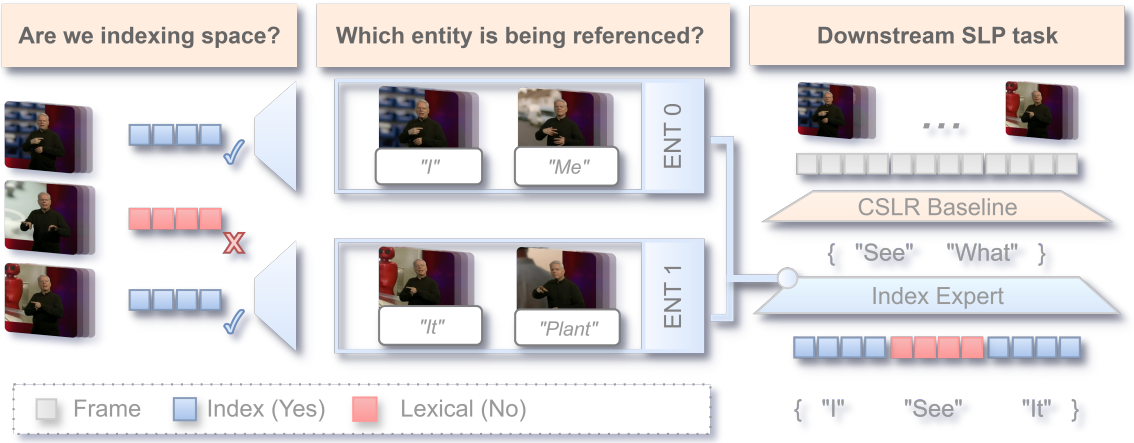

Figure 5: Indexing is a grammatical function where discourse entities are associated with spatial loci in the signing space and subsequently re-referenced through pointing.

Figure 6: Examples of false positives (two left columns) and false negatives (two right columns).

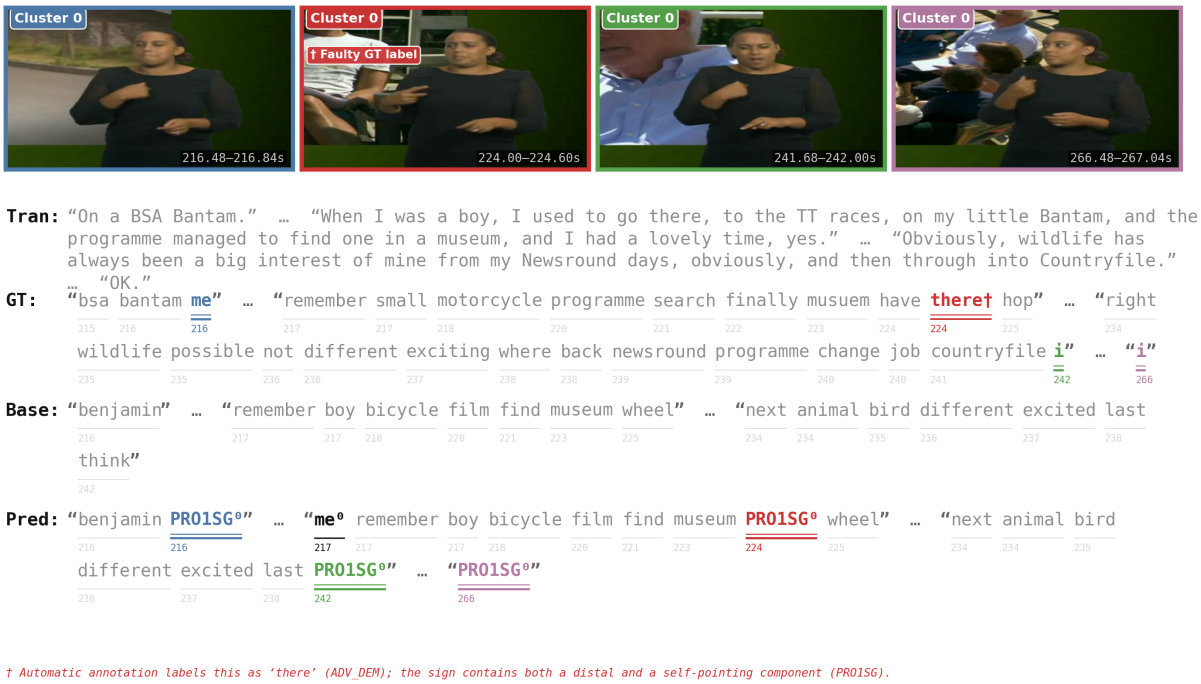

Figure 7: Instances of PRO1SG (*me*/*I*) across four examples in episode 6040895553921856506. Three instances are correctly detected and grouped. The second panel (red border) highlights an apparent ground-truth annotation or segmentation error: the sign contains both a locative (*there*) and a self-pointing (*me*) component, but the assigned label is only reflecting the former. The system nevertheless predicts PRO1SG consistently, suggesting that the pose features capture the self-referential component despite annotation noise.

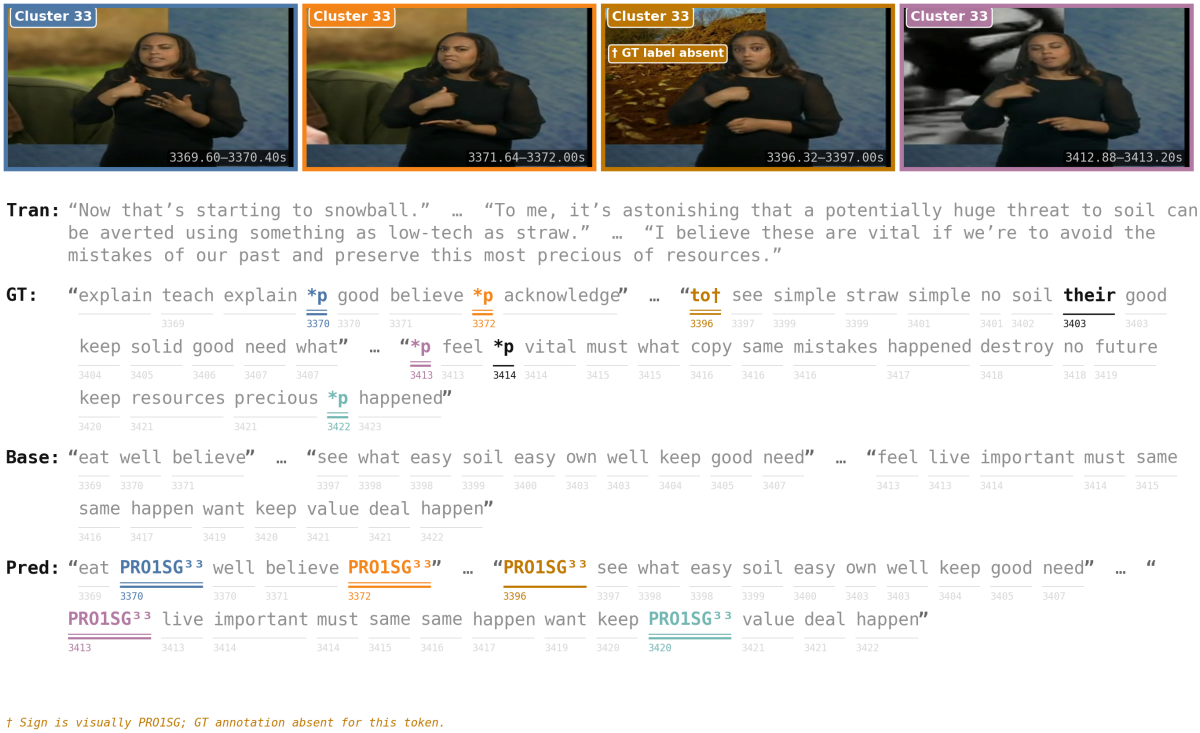

Figure 8: Instances of PRO1SG (*me*) across four examples in episode 6003446875091426252. The third panel highlights a missing ground-truth annotation: the video contains a self-pointing sign but is glossed as the preposition *to*, reflecting incomplete coverage of indexing signs. The model groups all four instances consistently and predicts PRO1SG throughout. The ground truth also reflects that two signs are not detected by either the baseline or the index expert.

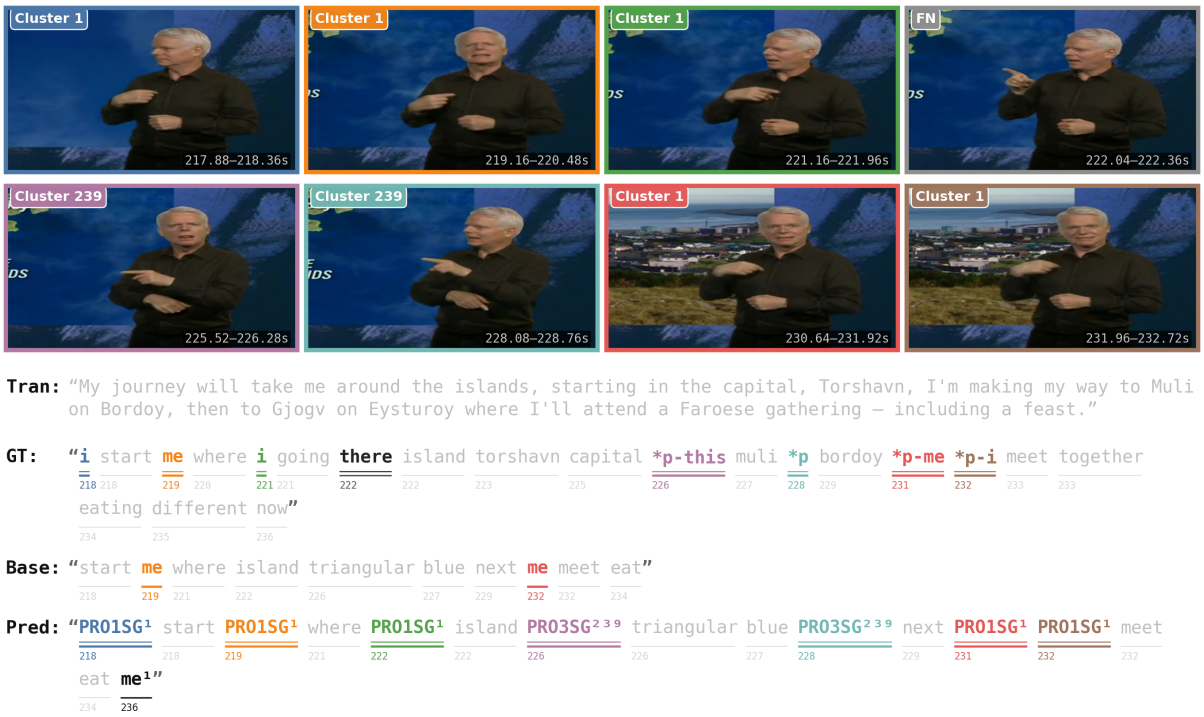

Figure 9: Two co-occurring entity clusters in episode 6177195911563690266. There are 5 instances of Cluster 1 (PRO1SG, *me/I*): all correctly detected and grouped. The fourth panel (grey border) is a false negative where a deictic *there* token is missed. There are two instances of a PRO3SG cluster, a distinct third-person referent introduced in the same sentence window. This demonstrates that the system maintains both entities simultaneously and continues to recover the self-reference across a substantial temporal gap. The baseline predicts only two of the eight pointing tokens.

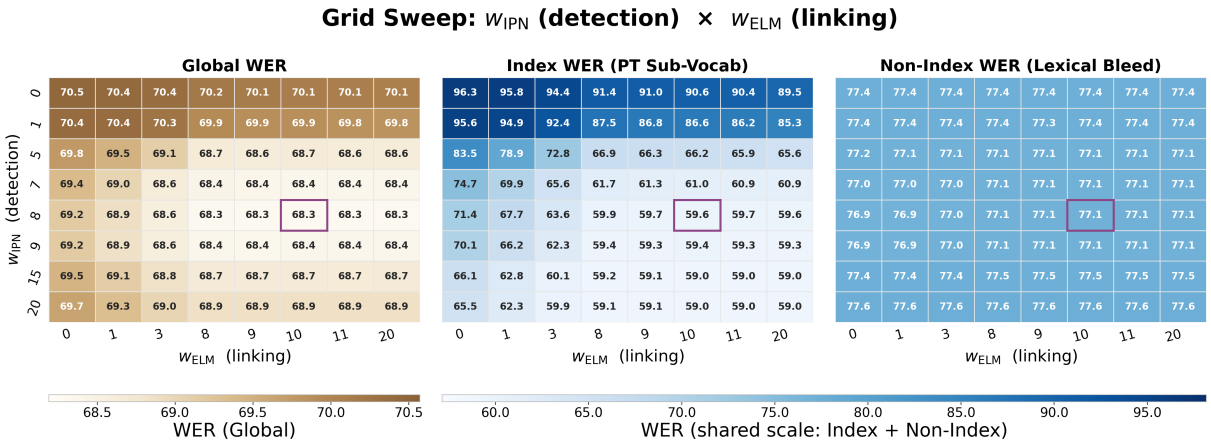

Figure 10: 2D projection of the 3D heatmap showing numerical WER values per grid cell.

D Numerical Values of SLR Grid Sweeps

Figure 10 displays the numerical WER values of each cell in the 3D heatmap. The cell chosen for optimal performance is highlighted with a purple border, corresponding to $w_{\text{IPN}} = 8$ and $w_{\text{ELM}} = 10$. The distribution reflects the underlying temperature parameter of CSLR2, which is set at $t = 20$.

E Automatic Entity Label Generation

This section describes the two-stage pipeline used to generate entity labels for the coreference training data (Sec. 5.2 of the main paper): mention extraction, LLM cluster assignment, validation and retry, and the post-processing step that resolves pointing-gesture glosses.

E.0.1 Stage 1: Rule-Based Mention Candidate Extraction

For each document, mention candidates are extracted from English free translations of MDGS and BSLCP using a spaCy POS-based span selector with `en_core_web_lg`. Four span patterns are applied and merged:

1. **Proper-noun sequences:** one or more consecutive `PROPN` tokens (e.g. "London", "British Sign Language").
2. **Nominal phrases:** optional `DET/PRON/NUM`, followed by zero or more `ADJ` and one or more `NOUN/PROPN` tokens (e.g. "the deaf community").
3. **Pronominal mentions:** single-token `PRON` (e.g. I, you, she, it).
4. **Standalone demonstratives:** `DET` tokens from {this, that, these, those} not followed by a noun.

Nested spans are pruned to maximal spans. The output is one line per dialogue turn with its extracted mentions:

```
[turn=3 speaker=BF25F29WHN] "So I started school in London",
    ["I", "school", "London"]
```

E.0.2 Stage 2: LLM Coreference Cluster Assignment

Model setup. We use a 120B-parameter open-weights model (`gpt-oss:120b`) via Ollama (`/api/chat`, `temperature=0`, 8K-token context, max 2K generation tokens; extended to 16K on `done_reason=length`).

Prompt design. The system prompt specifies task constraints and output format; the user message provides the document identifier and formatted dialogue.

System Prompt

You are a coreference clustering engine.
INPUT Each line is: [turn=T speaker=S] "TURN TEXT", ["mention1", "mention2", ...]
TASK For each mention in the list, assign a coreference cluster ID (an integer).
CORE RULES (PROCESSING ORDER)
1) Process the dialogue strictly in order, from the first turn to the last.
2) When a mention refers to an entity already seen, reuse the SAME cluster ID.
3) When a mention is new, assign the NEXT unused integer (starting at 0).
4) Never renumber, reuse, or reshuffle existing cluster IDs.
5) If a mention is not an entity mention or you are unsure, output "-".
REFERENCE RULES - Same entity => same integer. - I / me / my => current speaker. - you / your => most likely addressee.
OUTPUT FORMAT - Return ONLY one bracketed list per input line. - Item count MUST equal mention count. Examples: [0] [1, 2, 3] [4, 3, -]

User Message Format

doc_id: BF25F29WHN
Dialogue:
[turn=1 speaker=BF25F29WHN] "Well I was born deaf", ["I"]
[turn=2 speaker=BF25F29WHN] "Right and where did you grow up?", ["you"]
...

Document chunking. Documents are split into non-overlapping **100-turn chunks** processed sequentially. IDs from chunk k are offset by $\max(\text{IDs in chunks } 1 \ldots k-1) + 1$ before concatenation to ensure globally unique cluster IDs.

Validation and retry. Each LLM response is checked for (i) correct line count and (ii) correct per-line item count. Failures trigger up to **three retries**; if all fail and the chunk has $\geq$20 turns, it is recursively **halved** (minimum 10 turns) and each half retried. Successful chunks are cached for resume.

E.0.3 Stage 3: Resolving Pointing-Gesture Glosses

The LLM assigns clusters to lexical mentions in the translation. Sign-language corpora additionally contain pointing-gesture glosses absent from the translation; these are resolved in post-processing:

1. **Explicit mapping.** Fixed-referent glosses are resolved deterministically (`PT:PRO1SG`→first-person cluster; `PT:PRO2SG`→second-person cluster; etc.).
2. **Generic pointers.** All bare `PT:` tokens are first collapsed into a single shared cluster. `PT:` and `PT:DET` tokens are then grouped into runs; temporally proximate occurrences (within 25 entries) are merged into the same cluster.
3. **Soft third-person lookup.** Before co-occurrence scoring, `PT:PRO3SG`, `PT:POSS3SG`, `PT:PRO3PL`, and `PT:POSS3PL` labels search `entities_raw` for clusters containing a third-person pronoun mention (*he, she, her, his, it, they, them, their*), providing a lightweight resolution path that avoids spurious co-occurrence matches.
4. **Cross-sentence co-occurrence.** Remaining PT labels are linked to the LLM cluster with which they most frequently co-occur (processed by decreasing cluster size).
5. **BUOY grouping.** Nearby buoy tokens (`PT:BUOY`, `PT:LBUOY`, `PT:FBUOY`) within a 15-entry window share a cluster.
6. **Fallback.** Remaining instances attach to the

nearest assigned cluster of the same label type, or open a new cluster.

E.0.4 Stage 4: Rule-Based Post-Processing Refinements

Raw LLM clusters often contain systematic errors that degrade training label quality. Three targeted refinements are applied after Stage 3:

Inanimate co-referent suppression. A regex pattern blocks inanimate temporal, weather, and event nouns (e.g. *a day*, *the snow*, *the wedding*) from being assigned to person-pointing clusters. The pattern is applied at two levels: (i) cluster-level, where any cluster whose representative mention matches the pattern is excluded from matching against person-pointing PT labels; and (ii) mention-level, where per-sentence inanimate mentions are skipped when resolving person-pointing labels. This corrects a systematic LLM error where temporal or locative noun phrases are grouped into the same cluster as a person due to surface proximity in the text.

Person-label cluster unification. After co-occurrence-based assignment, the same person-pointing label (e.g. `PT:PRO3SG`) may be fragmented across several small clusters within a document. A second pass identifies all clusters assigned the same person-pointing label and merges them toward the cluster with the strongest evidence, updating all mention-level assignments accordingly.

Possessive cluster merging. Clusters where all mentions match a possessive relation pattern (e.g. *my wife*, *my girlfriend*) and all are assigned the same person-pointing label are merged under the dominant entity. This addresses cases where possessively-introduced referents are fragmented across multiple clusters due to variation in surface form.

E.0.5 Constructing a Small Evaluation Suite

To obtain a coarse estimate of the performance of the LLM-based coreference pipeline, we constructed a small manually annotated evaluation set. For three documents from BSLCP, we created ground-truth entity-link annotations by hand.

Table 7: Accuracy of LLM-based coreference annotations on three manually annotated documents.

Model	Acc. Doc 1	Acc. Doc 2	Acc. Doc 3
ChatGPT 5.2	0.82	0.88	0.85
gpt-oss:120b	0.76	0.93	0.85
gpt-oss:20b	0.63	0.83	0.78

While these annotations were produced by the authors themselves and therefore cannot be considered perfect, they provide a reasonable approximation for comparing different model configurations. In practice, the annotation process scales reasonably well, requiring approximately five minutes per document.

Using this setup, we evaluated several LLM-based coreference pipelines running on the DGS Spark cluster via Ollama. Preliminary experiments showed that most models failed to produce stable or well-structured cluster outputs. In practice, `gpt-oss:120b` was the only free-to-use model that consistently produced usable results, and it was therefore selected for the large-scale annotation pipeline. Table 7 reports the accuracy of several tested models on the three annotated documents.

F Input Feature Layout and ELM Auxiliary Features

Our model operates on pose representation from two 3D keypoint sources, combined into a single skeletal graph processed by the SL-GCN backbone.

WiLoR Hand Keypoints. WiLoR produces 21 MANO-compatible 3D keypoints per hand (wrist + 5×4 finger joints). Both hands are included, giving $2\times21 = 42$ additional nodes (right hand: nodes 8–28; left hand: nodes 29–49).

SMPL-X Body Keypoints. Whole-body poses are estimated from RGB video using SMPL-X, which outputs 55 joints per detected person. We retain the 8 joints that are informative towards the relative wrist position to body and motion trajectories. The map between the graph nodes and the SMPL-X joints are given in Table 6.

Table 6: Mapping between graph nodes and SMPL-X body joints.

Node	0	1	2	3	4	5	6	7
SMPL-X idx	21	19	17	16	18	20	0	22
Joint (side)	Right wrist	Right elbow	Right shoulder	Left shoulder	Left elbow	Left wrist	Pelvis	Upper spine

Combined Graph. The full graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ has $|\mathcal{V}| = 50$ nodes and $|\mathcal{E}| = 53$ undirected edges as depicted in figure 11.

Normalisation Pipeline. Each node carries raw 3D coordinates (x, y, z); the input tensor per temporal window is $(T \times 50 \times 3)$ with T=12 frames. Body-joint coordinates lie approximately in $[-1, 1]$; hand joints are scaled proportionally. Before use, coordinates undergo:

1. **Scale alignment.** Body skeleton auto-scaled so that the mean body–hand limb length matches a fixed target of 0.075 in WiLoR metric units, preventing the larger SMPL-X range from dominating hand features.
2. **Wrist stitching.** Body wrist nodes (0, 5) are translated to coincide with their WiLoR counterparts, enforcing a consistent hand origin.
3. **Root centring.** Skeleton translated so that the mean wrist position over the window lies at the origin.
4. **Scale normalisation.** All coordinates divided by the inter-shoulder distance (nodes 2–3), giving viewpoint invariance.
5. **Canonical orientation.** Rotation aligns the shoulder–shoulder axis with the y-axis, removing in-plane torso rotation while preserving relative hand pose.

F.1 ELM Auxiliary Geometric Features

As described in Sec. 4.3 of the main paper, the ELM scoring incorporates auxiliary feature streams that recover spatial and kinematic information compressed away by the IPN binary classification.

Engineered pointing scalar features. Six scalar features are computed from the dominant pointing hand (selected as the hand with greater mean wrist–elbow displacement over the clip):

1. **Elevation.** Vertical angle of the mean pointing direction $\arcsin(d_y)$, where $\mathbf{d}$ is the normalised sum of the finger vector (tip−wrist) and arm vector (wrist−elbow).
2. **Target-y.** Mean fingertip height relative to the shoulder midpoint, normalised by shoulder width. Distinguishes upward (addressee, abstract loci) from body-level (self, nearby referents) pointing.
3. **Target-z.** Mean fingertip depth relative to the shoulder midpoint, normalised by shoulder width. Distinguishes forward-directed (third-person, distal) from body-proximal pointing.
4. **Arm reach.** Normalised elbow-to-fingertip distance, capturing how fully extended the arm is.
5. **Index selectivity.** Difference between the normalised index fingertip extension and the mean extension of the middle, ring, and pinky fingertips, measured relative to palm scale. High values indicate a prototypical pointing handshape.
6. **Trajectory length.** Cumulative fingertip displacement over the clip, normalised by shoulder width. Distinguishes static holds from sweeping or arc-shaped pointing gestures.

For scoring compatibility with entity e_k, the six values for the current mention are concatenated with the stored per-cluster running mean and their element-wise difference (18 values total), then pro-

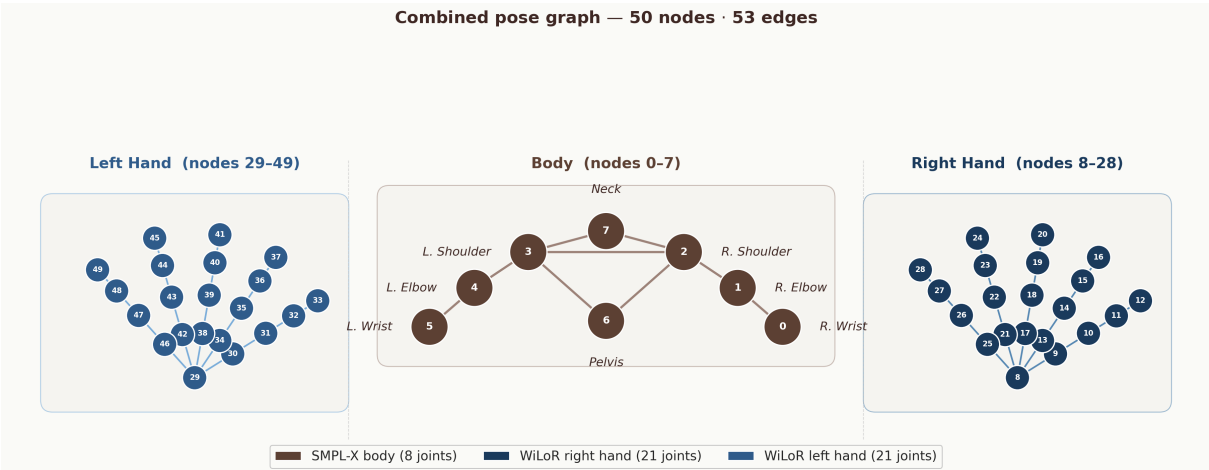

Figure 11: Combined pose graph. Brown: 8 SMPL-X body joints (nodes 0–7). Dark blue / steel blue: WiLoR right (8–28) and left (29–49) hands, 21 joints each: $\mathcal{V} = \{V_{\text{body}}^{0\ldots7},\ V_{\text{R}}^{8\ldots28},\ V_{\text{L}}^{29\ldots49}\}$.

jected to a fixed-size embedding via a two-layer MLP with LayerNorm.

G IPN Threshold Ablation

Table 8 reports IPN performance as a function of the decision threshold τ (B+M joint model, 3 seeds). The operating point τ=0.70 maximises F1 on the index class. For SLR integration we use τ=0.90 to further suppress false positives at the cost of recall, trading lexical bleed for precision in the downstream logit boost.

τ	F%	Prec-I	Rec-I	**F1-I**	BalAcc
0.30	28	$0.50_{\pm 0.04}$	$0.89_{\pm 0.02}$	$0.64_{\pm 0.03}$	$0.86_{\pm 0.01}$
0.35	26	$0.53_{\pm 0.04}$	$0.88_{\pm 0.03}$	$0.66_{\pm 0.03}$	$0.87_{\pm 0.01}$
0.40	25	$0.56_{\pm 0.04}$	$0.86_{\pm 0.03}$	$0.68_{\pm 0.03}$	$0.87_{\pm 0.01}$
0.45	23	$0.58_{\pm 0.04}$	$0.84_{\pm 0.03}$	$0.68_{\pm 0.02}$	$0.86_{\pm 0.01}$
0.50	22	$0.60_{\pm 0.03}$	$0.82_{\pm 0.03}$	$0.69_{\pm 0.02}$	$0.86_{\pm 0.01}$
0.55	21	$0.62_{\pm 0.03}$	$0.80_{\pm 0.02}$	$0.70_{\pm 0.02}$	$0.85_{\pm 0.01}$
0.60	19	$0.64_{\pm 0.03}$	$0.78_{\pm 0.02}$	$0.70_{\pm 0.01}$	$0.85_{\pm 0.01}$
0.65	18	$0.67_{\pm 0.03}$	$0.75_{\pm 0.02}$	$0.71_{\pm 0.01}$	$0.84_{\pm 0.01}$
0.70	16	$0.69_{\pm 0.02}$	$0.72_{\pm 0.03}$	$\mathbf{0.71}_{\pm 0.01}$	$0.83_{\pm 0.01}$
0.75	15	$0.73_{\pm 0.02}$	$0.68_{\pm 0.02}$	$0.70_{\pm 0.01}$	$0.82_{\pm 0.01}$
0.80	13	$0.76_{\pm 0.02}$	$0.63_{\pm 0.04}$	$0.69_{\pm 0.02}$	$0.80_{\pm 0.02}$
0.85	11	$0.80_{\pm 0.02}$	$0.56_{\pm 0.04}$	$0.66_{\pm 0.02}$	$0.77_{\pm 0.02}$
0.90	9	$0.84_{\pm 0.02}$	$0.45_{\pm 0.07}$	$0.58_{\pm 0.05}$	$0.72_{\pm 0.03}$
0.95	5	$0.89_{\pm 0.03}$	$0.28_{\pm 0.08}$	$0.42_{\pm 0.10}$	$0.64_{\pm 0.04}$

Table 8: Threshold sweep (B+M joint model, mean $\pm$ std, 3 seeds). Best F1-I bolded. F % Denotes the percent of indexing clips fiered.

Table 9 further reports the cw4-r1 model across individual evaluation datasets at representative thresholds.

H WER$_{all}$, WER$_{Index}$ and WER$_{Lex}$

Word Error Rate (WER). WER is the Levenshtein edit distance between predicted sequence $\hat{Y}$ and reference Y, normalised by reference length:

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100\%, \tag{5}$$

where S, D, I are substitutions, deletions, and insertions, and $N = |Y|$. Evaluation is case-insensitive and synonym-aware (Sec. I).

WER$_{Index}$. Both reference and hypothesis are filtered to the PT sub-vocabulary $\mathcal{V}_{\mathrm{PT}}$ (pronouns, demonstratives, reflexives) before applying Eq. (5):

$$\mathrm{WER}_{\mathrm{Index}} = \frac{\mathrm{EditDist}\left(\hat{Y}|_{\mathcal{V}_{\mathrm{PT}}},\ Y|_{\mathcal{V}_{\mathrm{PT}}}\right)}{\left|Y|_{\mathcal{V}_{\mathrm{PT}}}\right|} \times 100\%. \tag{6}$$

WER$_{Lex}$. All $\mathcal{V}_{\mathrm{PT}}$ tokens are *excluded* prior to alignment:

$$\mathrm{WER}_{\mathrm{Lex}} = \frac{\mathrm{EditDist}\left(\hat{Y}|_{\overline{\mathcal{V}_{\mathrm{PT}}}},\ Y|_{\overline{\mathcal{V}_{\mathrm{PT}}}}\right)}{\left|Y|_{\overline{\mathcal{V}_{\mathrm{PT}}}}\right|} \times 100\%. \tag{7}$$

This metric is unaffected by pointing-sign improvements and serves as a sanity check that the base model is not degraded due to lexical bleed.

I Synonym-Based Evaluation Mapping

Several pronoun and demonstrative forms share the same BSL sign (eg. *I* and *me* are both realised as a chest-point). To avoid penalising semantically correct but lexically variant predictions, synonymous forms within each group are treated as equivalent during alignment. A different challenge is that several out-of-vocabulary (OOV) forms appear in the test set but not in the pre-trained model vocabulary. To include these in the evaluation, we map these forms to their in-vocabulary closest visual and semantic neighbor before scoring (Table 10).

J SLR Integration: Inference-Time Guardrails

The IPN and ELM boosts are applied additively at inference time. Two additional guardrails are employed to prevent false or inconsistent boosts from degrading non-indexing predictions.

Directional consistency filter. At linking time, a candidate entity assignment is rejected if the cosine similarity between the current mention's pointing direction and the stored mean direction of the candidate cluster falls below a threshold δ_{dir}. In this case the mention is either assigned to a different cluster or opens a new one. This prevents the cluster prior of a left-side referent from being applied to a right-directed pointing sign, which would add incorrect vocabulary bias. The threshold $\delta_{\text{dir}} = 0.20$ is used in all reported experiments.

Cluster group partitioning. BSLCP PT gloss labels span multiple grammatical categories: person-pointing pronouns (`PRO1SG`, `PRO2SG`, `PRO3SG`, etc.), possessives (`POSS*`), locatives (`LOC`), determiners (`DET`), and spatial buoys (`BUOY*`). A single entity cluster accumulating evidence from both person-pointing and locative pointing would produce a confused vocabulary prior. With cluster group partitioning enabled, each cluster is restricted to one

Table 9: Threshold sweep for cw4-r1 across evaluation datasets (B = BSLCP, M = MDGS, Bob_{fs} = BOBSL full-sign).

Train	Eval	τ	BalAcc	Macro F1	Prec-L	Rec-L	Prec-I	Rec-I
cw4-r1	B	0.60	$0.80_{\pm 0.00}$	$0.80_{\pm 0.00}$	$0.74_{\pm 0.00}$	$0.94_{\pm 0.00}$	$0.92_{\pm 0.00}$	$0.66_{\pm 0.00}$
		0.70	$0.79_{\pm 0.00}$	$0.79_{\pm 0.00}$	$0.73_{\pm 0.00}$	$0.95_{\pm 0.00}$	$0.92_{\pm 0.00}$	$0.64_{\pm 0.00}$
		0.80	$0.78_{\pm 0.00}$	$0.77_{\pm 0.00}$	$0.71_{\pm 0.00}$	$0.96_{\pm 0.00}$	$0.94_{\pm 0.00}$	$0.60_{\pm 0.00}$
		0.90	$0.75_{\pm 0.00}$	$0.74_{\pm 0.00}$	$0.68_{\pm 0.00}$	$0.97_{\pm 0.00}$	$0.95_{\pm 0.00}$	$0.54_{\pm 0.00}$
cw4-r1	M	0.60	$0.85_{\pm 0.00}$	$0.84_{\pm 0.00}$	$0.80_{\pm 0.00}$	$0.92_{\pm 0.00}$	$0.91_{\pm 0.00}$	$0.77_{\pm 0.00}$
		0.70	$0.84_{\pm 0.00}$	$0.83_{\pm 0.00}$	$0.78_{\pm 0.00}$	$0.93_{\pm 0.00}$	$0.92_{\pm 0.00}$	$0.74_{\pm 0.00}$
		0.80	$0.82_{\pm 0.00}$	$0.81_{\pm 0.00}$	$0.75_{\pm 0.00}$	$0.94_{\pm 0.00}$	$0.93_{\pm 0.00}$	$0.69_{\pm 0.00}$
		0.90	$0.78_{\pm 0.00}$	$0.78_{\pm 0.00}$	$0.71_{\pm 0.00}$	$0.96_{\pm 0.00}$	$0.94_{\pm 0.00}$	$0.61_{\pm 0.00}$
cw4-r1	Bob_{fs}	0.60	$0.68_{\pm 0.00}$	$0.71_{\pm 0.00}$	$0.90_{\pm 0.00}$	$0.96_{\pm 0.00}$	$0.63_{\pm 0.00}$	$0.41_{\pm 0.00}$
		0.70	$0.68_{\pm 0.00}$	$0.71_{\pm 0.00}$	$0.90_{\pm 0.00}$	$0.96_{\pm 0.00}$	$0.65_{\pm 0.00}$	$0.39_{\pm 0.00}$
		0.80	$0.67_{\pm 0.00}$	$0.70_{\pm 0.00}$	$0.89_{\pm 0.00}$	$0.97_{\pm 0.00}$	$0.67_{\pm 0.00}$	$0.37_{\pm 0.00}$
		0.90	$0.66_{\pm 0.00}$	$0.69_{\pm 0.00}$	$0.89_{\pm 0.00}$	$0.97_{\pm 0.00}$	$0.69_{\pm 0.00}$	$0.34_{\pm 0.00}$

Table 10: BSL synonym groups. OOV forms are mapped to the in-vocabulary synonym before WER alignment; "—" = no OOV forms.

Group	**In-vocabulary**	**OOV → mapped**
PRO1SG	me	I
PRO2SG	you	—
PRO3SG	he, her	him, she, it
PRO1PL	our	we, us
PRO3PL	they	them
POS1SG	my	mine
POS2SG	your	yours
POS3SG	her	his, its
POS1PL	our	ours
POS3PL	their	—
DET_SG	this, that	—
ADV_DEM	here, there	these, those
REFL1SG	myself	—
REFL	yourself	himself, herself, itself, themselves

PT grammatical category: a mention of a different category cannot be linked to an existing cluster of a different group, forcing a new cluster to be opened instead. This ensures that person-pointing and locative-pointing referents are tracked separately, and that their respective vocabulary priors remain category-consistent.

K Hyperparameters

This section lists the hyperparameters used for training in IPN (Table 11), ELM (Table 12), and SLR integration (Table 13).

Table 11: IPN hyperparameters (Optuna best)

Hyperparameter	**Value**	**Notes**
Backbone	SL-GCN	Spatial-GCN on graph
Clip length	12 frames	
Body keypoints	8 (SMPL-X)	
Hand keypoints	42 (WiLoR, 21/hand)	
Batch size	96	
Epochs	25	Early stop patience: 7
Learning rate	1.10×10^{-3}	Adam
Weight decay	3.02×10^{-4}	
Non-PT CW	4.0	CE loss upweight
Focal γ	0.0	CE
Batch balance	`WeightedRandomSampler`	1:1 PT:non-PT

Table 12: ELM hyperparameters. Encoder and classification head are frozen.

Hyperparameter	**Value**
Encoder	Frozen IPN SL-GCN
Memory type	`learned_avg`, Hadamard similarity
Memory MLP size	300
Memory MLP depth	2
Entity slots	20
Entity dim	10
Linking proj. hidden	256 , 2-layer MLP, GELU, LN
Dropout	0.3
Batch size	32
Epochs	10
Learning rate	2.53×10^{-4}, Adam
Weight decay	8.07×10^{-6}
λ_{cls}	0.1 , Aux. classification weight

Table 13: CSLR2 inference parameters (no training; selected by sweep).

Hyperparameter	**Value, note**
Expert threshold τ	0.9 , IPN fire threshold
Scoring mode	Window-level , Single score per subtitle
Boost mode	Soft (additive) , Non-PT logits unchanged
BOBSL frame rate	25 fps
Pointing token	`"point"`, Vocabulary anchor

L Model Size and Budget

Model Parameters IPN and ELM share the same frozen SLGCN backbone weights; ELM adds approximately 714K trainable parameters on top.

Compute Budget All experiments were conducted on a single NVIDIA RTX 5090 GPU (24GB VRAM, Blackwell architecture; see Table 15. **Total reported compute:** $\sim$32 GPU-hours on a single RTX 5090.

M Disclaimer: Use of AI-Assisted Development

Parts of the implementation were developed with assistance from Claude Code (Anthropic), an AI-based coding assistant. AI assistance was used during the development and integration of the IPN and ELM inference pipeline with the CSLR2 baseline. All experimental design, modelling decisions, and scientific conclusions are the authors own.

N License and code

We will release the code upon publication under a standard research license. The use of existing datasets and models in this work is consistent with their intended use for academic research as specified in their original licenses. We do not repurpose any artifacts beyond their permitted research scope.

Component	Architecture	Parameters
IPN (Index Pointing Network, Phase 1)	ST-GCN backbone + linear classification head	7.62M (7,623,802)
		backbone: 7,622,776 head: 1,026
ELM (Entity Linking Model, Phase 2)	Shared ST-GCN backbone + memory module + linking MLP	8.34M (8,336,617)
		backbone: 7,622,776 memory: 513,903 linking: 198,912 head: 1,026
CSLR2 backbone (frozen, Phase 3)	Full CSLR2 architecture	412.37M (412,371,449)

Table 14: Parameter counts for all system components. IPN and ELM share a ST-GCN backbone; CSLR2 is kept frozen during integration.

Phase	Compute usage
IPN	35 runs (Optuna search + 5-seed final sweep), with runtimes ranging from 4–32 minutes per run. Total compute: $\sim$4.3 GPU-hours. Final model (5 seeds): $\sim$0.7 GPU-hours.
ELM	81 runs (Optuna search + ablation table configurations), with runtimes from 4 minutes to 3.5 hours. Total compute: $\sim$27.9 GPU-hours. Final reported configurations (6 settings $\times$ $\sim$40 min): $\sim$4 GPU-hours.
CSLR2 inference	Frozen backbone used only for evaluation (no fine-tuning). Evaluation over 55 BOBSL test episodes completes in under 5 minutes.

Table 15: Compute budget for IPN training, ELM training, and CSLR2 inference. We report the number of runs, runtime ranges, and total GPU-hours per phase. CSLR2 is used in a frozen setting with no fine-tuning.