

# Sanket: On-Device, Open-Vocabulary Indian Sign Language Recognition in the Browser

*Privacy-preserving, open-vocabulary ISL recognition on commodity devices  
using pose-based metric learning and training-free exemplar matching*

|                  |                                                                                                                        |
|------------------|------------------------------------------------------------------------------------------------------------------------|
| Candidate        | <b>Mithunish Prabhakaran</b>                                                                                           |
| Registration No. | <b>RA2512049015065</b>                                                                                                 |
| Programme        | M.Tech Artificial Intelligence, SRMIST (2025–2027)                                                                     |
| Main faculty     | Dr. Pitchaimanikam · pitchaib@srmist.edu.in ( <i>final review</i> )                                                    |
| Support faculty  | Dr. S Saravanan · saravans2@srmist.edu.in ( <i>Reviews 0–2</i> )                                                       |
| Review           | Review 0 · Sunday, 9 August 2026 · 2:00–6:00 PM · online, 1:1 with the support faculty                                 |
| SDG alignment    | SDG 10 — Reduced Inequalities                                                                                          |
| Living documents | <a href="https://sanket.kadal.cc/docs/review-0">sanket.kadal.cc/docs/review-0</a> — this PDF is the printable snapshot |

Deliverables per programme-office instructions: Title Identification · One-Page Abstract ·  
Literature Review · Novelty Proposal · Proposed System with Timeline

Written as the seed of the semester-2 research papers: original prose throughout, citations  
verified against publishing venues, ethics and licensing owned from Review 0.

# 1 Title identification

**Proposed title.** *Sanket: On-Device, Open-Vocabulary Indian Sign Language Recognition in the Browser* — with the formal long-form variant for the report cover: *Privacy-Preserving, Open-Vocabulary Indian Sign Language Recognition on Commodity Devices using Pose-Based Metric Learning and Training-Free Exemplar Matching*.

*Sanket* (सङ्केत) is Sanskrit-derived and shared across Hindi, Marathi and Nepali: a sign, a signal, a gesture that carries meaning — a name from the languages the product serves.

The long-form variant doubles as the **working title for the semester-2 research papers**: a title whose every phrase is a defensible claim survives the trip from case study to paper without renaming the work midway.

## 1.1 What each phrase commits the work to

- **On-device / in the browser** — pose extraction, recognition and sentence assembly all execute in the user’s own tab (WebAssembly) or desktop app. No video or landmark stream leaves the device; privacy is an architectural property, not a policy.
- **Open-vocabulary** — recognition retrieves over a dictionary-supervised vocabulary of 8,165 words (one reference performance for most), and the vocabulary grows without retraining.
- **Indian Sign Language** — datasets, references and evaluation are ISL throughout (INCLUDE, CISLR, ISLRTC dictionary material).
- **Pose-based metric learning** — landmark windows are embedded with margin-based losses and recognised by distance to class prototypes: the formulation that makes one-shot vocabulary growth possible.
- **Training-free exemplar matching** — a parallel pipeline matches live trajectories directly against dictionary references; it doubles as an instrument separating what is hard because of *learning* from what is hard because of *data*.

## 1.2 Out of scope for Review 0

Continuous signing / full sentence translation (application layer only, with an opt-in LLM polish step); finger-spelling as a dedicated modality; dialect coverage beyond the source dictionaries.

## 1.3 Alternates considered

“Indian Sign Language Recognition with Pose-Based Spatio-Temporal Networks” (accurate but silent on open vocabulary and deployment); “Real-Time ISL-to-Text Translation in the Browser” (overclaims translation); “A Shazam for Indian Sign Language” (the honest inspiration, kept as an explanatory device, too informal as a title); “Few-Shot Sign Language Recognition for Low-Resource Vocabularies” (generic — loses the language, the target, and the product).

## 2 One-page abstract

India has an estimated 63 million deaf and hard-of-hearing citizens and only a few hundred certified Indian Sign Language interpreters, so most everyday interactions — a bank counter, a clinic, a classroom — happen without one. Automatic word-level ISL recognition could close part of that gap, but the systems the literature offers are constrained in two ways that matter in India: they recognise a small closed set of classes fixed at training time, and they run server-side, imposing connectivity, cost, and a camera feed leaving the room.

This case study builds and evaluates **Sanket**, a recognition system that runs *entirely on the user's own device* in a web browser or desktop app. Video frames are reduced immediately to an 85-point pose-and-hand landmark skeleton (MediaPipe); the frames are discarded, and every later stage operates on landmarks alone. Recognition is formulated as **open-vocabulary retrieval** rather than closed-set classification, through two complementary pipelines. The *learned* pipeline embeds 48-frame landmark windows into a 256-dimensional space with a margin-loss encoder (ArcFace family) and retrieves over an int8-quantised prototype bank of **8,165 dictionary words** — combined with a calibrated closed-set head for the 263 well-resourced INCLUDE classes. The *training-free* pipeline — inspired by audio fingerprinting — matches live landmark trajectories directly against the dictionary references with an orientation-aware, dwell-weighted alignment, serving both as a fallback recogniser and as a diagnostic instrument that separates model limitations from data limitations.

At review 0 the prototype measured: **0.958 top-1** on the INCLUDE closed set; **0.243 top-1 / 0.411 top-20** on a held-out 2,285-query CISLR protocol over all 8,165 words; ~30 FPS pose tracking with ~14 ms per-window inference on a consumer laptop, fully offline. Analysis so far identifies *reference data quality* — not model capacity — as the binding constraint, and the proposed work therefore prioritises dictionary expansion (~7,000 further ISLRTC entries), reference hygiene, per-user enrolment, and a signer study, targeting the final report by **end of November 2026**. The system ships as a working accessibility suite — recognition studio, 3D-mannequin dictionary, and practice mode — aligned with **SDG 10, Reduced Inequalities**.

**Keywords:** Indian Sign Language · open-vocabulary recognition · on-device inference · metric learning · exemplar matching · accessibility

## 3 Literature review on relevant topics

### 3.1 Word-level SLR datasets, and where ISL stands

Word-level recognition matured on ASL corpora — WLASL (2,000 glosses, 21,000+ videos) [1] and MS-ASL (1,000 classes, 200+ signers) [2]. For ISL, INCLUDE [3] provides 4,287 studio recordings over **263 word classes** in 15 semantic categories — supervised-classification scale, but two orders of magnitude below the language; its best reported baseline reaches 85.6% (94.5% on the 50-class subset). CISLR [4] inverted the trade-off: 7,050 videos across roughly **4,700 words**, most with a *single* reference, framing large-vocabulary ISL as **one-shot retrieval**; its baseline is a prototype-based one-shot learner borrowing features from resource-rich ASL. iSign [5] added 118K+ video-sentence pairs for sentence-level ISL; the ISLRTC dictionary programme publishes thousands of official per-word demonstrations — underused supervision this project treats as first-class. *The realistic ISL regime is thousands of words × one or two references — methods that need dozens of examples per class do not serve the language.*

### 3.2 Appearance-based recognition

Video CNNs — I3D transfer [6] — set the closed-set baselines on WLASL and INCLUDE, but ingest raw RGB: heavy for in-browser inference, bound to studio appearance, and requiring the camera feed to enter the model — the exact privacy surface this work removes.

### 3.3 Pose-based recognition

ST-GCN introduced graph convolutions over joint sequences [7]; SAM-SLR carried skeleton ensembles to the top of SLR challenges [8]; SPOTER showed a compact pose transformer is competitive [9]. OpenHands [10] demonstrated pose-based pretraining across sign languages *including ISL* and argued pose is the accessibility-friendly substrate. MediaPipe [11], [12] supplies on-device holistic landmarks at camera rate — what makes the browser a viable target. *Pose-based work still mostly targets closed sets; the open-vocabulary regime needs retrieval.*

### 3.4 The low-resource, one-shot regime

With one reference per word, classification collapses into matching: learn an embedding where performances of a sign cluster, recognise by nearest prototype [13], with margin losses — ArcFace [14], CurricularFace [15] — providing discriminativeness. CISLR’s baseline applies exactly this framing. Reported open-vocabulary numbers stay low in absolute terms (top-1 ~0.1–0.3 over thousands of classes), attributed to signer variation against single references — *a data problem more than an architecture problem, a diagnosis this project’s experiments independently reproduce.*

### 3.5 Template matching and fingerprinting

Pre-deep-learning gesture recognition was alignment: DTW distance to stored exemplars [16]. Shazam [17] is the industrial proof that retrieval against a reference library can beat open-ended learning when reference and query are performances of the same object. This project revives that lineage as its training-free pipeline — dictionary demonstrations become landmark fingerprints matched by orientation-aware, dwell-weighted alignment — as much diagnostic instrument as recogniser.

### 3.6 On-device inference

ONNX Runtime Web executes quantised models in WebAssembly with SIMD and threads [18]; int8 quantisation keeps the full vocabulary bank shippable (<10 MB). Prior SLR literature rarely treats deployment as part of the research object; here latency, memory and privacy are first-class evaluation axes.

### 3.7 Scoping note

Continuous sign language *translation* is its own field [19]; this case study stops at word-level recognition plus rule-based gloss assembly and cites SLT only to mark the boundary.

### 3.8 Citation and originality discipline

This review is written to seed the research papers this case study intends to publish next semester, and is held to publication standards from the start: all prose is original to this project (nothing copied or closely paraphrased); every factual claim about prior work carries a citation, with dataset statistics verified against the publishing venues’ records (ACL Anthology, ACM DL, project pages; August 2026) rather than reproduced from memory; quotation is avoided entirely. These review documents are non-archival, so future papers may build on them without dual-publication conflict — target-venue self-overlap policies still apply per submission.

### 3.9 The gap this work occupies

| Prior work gives us                                          | Prior work does not give us                                                                                     |
|--------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Closed-set ISL classification at 263 classes (INCLUDE)       | Recognition over the <i>actual dictionary</i> of the language (8k+ words)                                       |
| One-shot retrieval framing for ISL (CISLR)                   | That framing <i>deployed</i> — running on end-user devices at camera rate                                       |
| Pose-based architectures and pretraining (OpenHands, SPOTER) | Pose pipelines engineered for the browser with privacy as a construction property                               |
| Metric-learning tools (ArcFace family, prototypes)           | An account of <i>why</i> open-vocabulary accuracy saturates — model vs data attribution                         |
| Classical exemplar alignment (DTW, Shazam)                   | An alignment designed for signs: orientation-aware, dwell-weighted, chirality-invariant, with reference hygiene |

#### Position of this case study

Every ingredient exists somewhere in the literature; no prior system composes them. The contribution is the composed system — open-vocabulary ISL retrieval running entirely on-device — plus a measured account of what limits it, backed by a training-free counterfactual no learned-only study can provide.

## 4 Novelty proposal

1. **Open-vocabulary ISL recognition shipped on-device.** Prior ISL systems classify closed sets offline or server-side; the one-shot framing exists only as a benchmark. Here: retrieval over an 8,165-word int8 prototype bank in the browser, composed with a calibrated 263-class head; vocabulary grows by adding references, not retraining. *Evidence: live system, ~30 FPS pose, ~14 ms/window; 0.958 top-1 closed-set; 0.243 / 0.411 top-1/top-20 open-vocabulary.*
2. **A training-free counterfactual pipeline as a scientific instrument.** Learned-only studies confound model, training and data. A parallel exemplar matcher over the same references makes failure attribution measurable: shared failures indict the data, avoided failures indict the learning. *Evidence: three fine-tuning tracks each gained  $\leq 0.3pp$  while the matcher traced the shared failures to reference-side defects — data is the binding constraint.*
3. **A sign-specific exemplar alignment.** Each property forced by a measured failure: holds are first-class (trajectory condensation); coverage is the vote (skipping a reference costs its duration); orientation is part of the sign (world-frame index and forearm direction — a “one” is the pointer directed upward); chirality is not (mirrored and opposite-handed performances match equally); idle hands are one shared state; position is a tiebreaker, not the signal. *Evidence: a committed regression harness with synthetic independent signers — natural one/zero/gun rank top-1 with 4–28 $\times$  margins; an inverted one is rejected by every word.*
4. **Reference hygiene as a first-class method.** Dictionary supervision is audited, not assumed: signing-space cropping removed approach/retreat footage from 4,132 of 8,187 clips; extraction failures are separated from genuinely static signs; per-word geometry checks expose anomalous references. Every finding converts into accuracy for both pipelines.
5. **Privacy and access as construction properties.** No server exists in the recognition path: frames become landmarks in-tab and are discarded; models and banks are cached static assets; the system works offline — the version of this technology that actually reaches its population, and the same property that lets the full suite (studio, 3D-mannequin dictionary, practice mode) deploy to any modern browser.

### The one-sentence novelty statement

A deployed, privacy-preserving, open-vocabulary ISL recogniser — and, through its training-free twin, the first measured account of *what actually limits* large-vocabulary ISL recognition: the references, not the model.

### 4.1 Publication path

The claims are scoped to decompose into two publishable papers if the semester’s results hold: a **systems & accessibility paper** (claims 1 and 5 plus the signer study — ACM ASSETS, W4A, CHI late-breaking) and a **method & analysis paper** (claims 2–4 plus the frozen-protocol ablations — LREC-COLING resource track, ACL/EMNLP findings, WACV workshops, ICVGIP). Three commitments follow: papers inherit numbers only by re-running the committed harnesses at submission time; the dataset licence audit (§5) completes before any submission; and these non-archival review documents may be built upon freely, subject to each venue’s originality policy.

## 5 Proposed system with timeline

### 5.1 Architecture at review 0 (all stages on-device — see the Addendum for the system as deployed)

|                              |                                                                                                                                                                                                                          |
|------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Capture</b>               | Webcam frames in the user's own tab; never stored or transmitted.                                                                                                                                                        |
| <b>Pose worker</b>           | MediaPipe extracts an 85-point skeleton (2×21 hand points, upper-body pose, face reference) at ~30 FPS; the frame is discarded immediately.                                                                              |
| <b>Canonical space</b>       | Landmarks re-anchored to the mid-shoulder origin, scaled by shoulder width — removing camera distance, framing and signer height. 48-frame windows.                                                                      |
| <b>Recognition worker</b>    | Signing-space and motion gates; then both pipelines: the 256-d int8 encoder (ONNX WASM) against the prototype bank + calibrated 263-class head, and the training-free fingerprint matcher against dictionary references. |
| <b>Decode &amp; assemble</b> | Decode cascade with temporal voting → gloss stream → rule-based sentence assembly; LLM polish only as explicit opt-in.                                                                                                   |

**Artifacts.** Encoder (int8 ONNX, ~14 ms/window) · prototype bank (8,165 × 256-d int8, ~2 MB) · fingerprint pack (8,055 orientation-aware trajectories, ~7 MB) · teach pack (8,187 compact landmark reference clips driving the 3D-mannequin dictionary and practice mode) · committed evaluation harnesses (frozen CISLR retrieval protocol, 2,285 held-out queries; natural-signing regression suite with synthetic independent signers).

**Evaluation rules.** The CISLR test split is sacred — never trained on, never tuned against. Open-vocabulary results reported as enrolled top-1/5/20 over all 8,165 words; closed-set on INCLUDE's standard split; on-device metrics from a consumer laptop; live behaviour guarded by the natural-signing harness, which fails the build if a change breaks naturally-performed signs.

### 5.2 Ethics, licensing and reproducibility

Obligations a publication venue would audit are owned from Review 0, not retrofitted. **Dataset terms** — INCLUDE and CISLR are used strictly for research under their release terms; ISLRTC dictionary videos are official public reference material processed into landmark skeletons, with no source video redistributed; a formal licence re-audit of every source is scheduled before any paper submission or public dataset artifact. **Human-participant ethics** — the signer study collects no video (recognition is on-device by construction): only recognition outcomes and structured feedback are recorded, under written informed consent with the right to withdraw, after whatever institutional approval SRMIST norms require. **Privacy** — no camera frame or landmark stream leaves the participant's device, in the study or in normal use. **Reproducibility** — every number in this handout runs from committed harnesses (frozen CISLR retrieval protocol; natural-signing regression suite); packs and models are versioned; papers inherit numbers only by re-running the harnesses. **Originality** — all prose in this document set is written for this case study; prior work is cited, never borrowed.

### 5.3 Timeline — 9 August to 30 November 2026

| Span      | Milestone                         | Work                                                                                                                                                                     |
|-----------|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Aug 9     | <b>Review 0</b> · support faculty | Proposal freeze: scope and evaluation protocol locked with the support faculty.                                                                                          |
| Aug 10–30 |                                   | Reference expansion (complete ~7,000 ISLRTC entries, quota-aware) · hygiene audit across all sources · benchmark harnesses frozen · dataset licence re-audit documented. |

|               |                                              |                                                                                                                                                                  |
|---------------|----------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Aug 24–Sep 13 |                                              | Personal enrolment (record-your-own references) · multi-reference matching · model-matcher ensemble explored.                                                    |
| Sep 7–27      | <b>Review 1 (≈Sep 20)</b> · support faculty  | Encoder retrain on expanded, cleaned references (Colab Pro) · quantisation · ablations: data-cleaning vs loss vs architecture.                                   |
| Sep 28–Oct 18 |                                              | Temporal voting and emission calibration · consent protocol and institutional approvals, then the structured signer study · Learn/Practice polish from findings. |
| Oct 12–31     | <b>Review 2 (≈mid-Oct)</b> · support faculty | Evaluation freeze: final measured results across accuracy, latency, footprint · feature freeze · report skeleton.                                                |
| Nov 1–21      |                                              | Final case-study report (these documents are its living draft) · demo hardening · reproducibility pass on every quoted number.                                   |
| Nov 22–30     | <b>Final review</b> · main faculty           | Buffer for review feedback and corrections; the final review/viva is conducted by the main faculty.                                                              |

## 5.4 Risks and mitigations

| Risk                                                 | Mitigation                                                                                                           |
|------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| ISLRTC download quotas throttle ingestion            | Quota-aware resumable fetching already built; spread over weeks 1–3, checkpointed.                                   |
| Single-reference variance keeps open-vocab top-1 low | Multi-reference + personal enrolment change the regime; top-5/top-20 reported as the assistive-use metrics they are. |
| Training compute budget (Colab Pro only)             | Data-centric plan front-loads GPU-free work; re-trains are few, late, well-targeted.                                 |
| Webcam domain shift vs studio references             | Canonical-space normalisation + the natural-signing harness catch this class of failure before users do.             |
| Solo-developer timeline                              | Phases overlap and each ships an increment — a slip degrades scope, not the deliverable.                             |

## 6 Addendum — where the proposal stands (4 September 2026)

This handout is the approval-record snapshot of 9 August; the living pages at [sanket.kadal.cc/docs](https://sanket.kadal.cc/docs) are kept current. This addendum carries the numbers that have moved since review 0, so the printed record cannot mislead.

Review 1 (30 August) measured the supervised model falling from 0.82 in-corpus to **0.33** on a signer it had never seen, and redirected training to latent self-supervision (ISL-JEPA) over a 119,064-unit unlabelled ISL pool. Since then that encoder, fine-tuned with gloss-matched dictionary clips, 3D-re-posed synthetic signers and a recover-and-resample augmentation, ensembled over four seeds and distilled, reaches **0.4818** on the same signer in a **4.57 MB** model — from 0.3318 in a 9.90 MB one.

### 6.1 The system as deployed

| Component           | At review 0                                                   | Now (4 Sep)                                                                                                                                       |
|---------------------|---------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| Landmarks           | 85-point skeleton, 48-frame windows                           | 107 canonical points (33 pose, 21+21 hands, face subset); 64-frame ring with capture epochs                                                       |
| Dictionary lane     | calibrated 263-class head                                     | 260-class JEPA encoder Cut@4 + attentive head, ONNX int8, 4.57 MB; preprocessing inside the graph                                                 |
| Retrieval lane      | 8,165 × 256-d int8 prototype bank                             | 256-d embedding vs a 40,439-row bank — 4,806 words / 8,188 entries; cosine; tier A 0.6 · floor 0.65 · shortlist 0.55 · gap 0.05; ArcFace scale 30 |
| Fingerprint matcher | 8,055 trajectory fingerprints                                 | runs beside the retrieval lane as a training-free witness                                                                                         |
| Evaluation          | frozen CISLR protocol (2,285 queries) + natural-signing suite | the two-corpus gate (RKMVU-220 + CISLR test-role 379 = 599 clips) plus the earlier harnesses                                                      |

### 6.2 Model lineage (cross-signer RKMVU top-1 unless stated)

| Model                                                                 | When  | RKMVU                       | Gate           | Notes                                                                                   |
|-----------------------------------------------------------------------|-------|-----------------------------|----------------|-----------------------------------------------------------------------------------------|
| Supervised v3.5.2 — the deployed retrieval-lane encoder               | Aug   | 0.3318                      | —              | 9.90 MB; the baseline                                                                   |
| Phase B JEPA final100@4 (jepa-ship4)                                  | 1 Sep | 0.3273 fp32 / 0.3318 int8   | —              | 4.57 MB; last-layer target collapsed (target-std 0.98 → 0.52); only layers ≤ 4 transfer |
| 1. gloss infusion + synthetic signers all+synth:full@2 (jepa-infuse4) | 2 Sep | 0.4318 (int8 0.4364)        | 0.4124         | 5-seed mean cross-val 0.3927, from 0.3018                                               |
| 4-seed ensemble → distilled student (jepa-distil4)                    | 3 Sep | 0.4545                      | 0.4424         | held-out 0.5385 · CISLR 0.4354                                                          |
| 1. recover-and-resample rr → distilled (jepa-rr4)                     | 3 Sep | <b>0.4818</b> (int8 0.4864) | <b>0.4591</b>  | held-out 0.5481 · CISLR 0.4459; 4-seed ensemble 0.4574, +TTA 0.4641                     |
| Phase C proper — all-layer teacher, 100 ep, full pool (jepa_c4)       | 4 Sep | <i>pending</i>              | <i>pending</i> | pilot: +4.8 pp gate at equal budget                                                     |

**The gate.** Clip-weighted top-1 over RKMVU-220 (one unseen signer) + CISLR test-role 379 (never trained on) — 599 clips, 1 clip = 0.17 pp — reported with RKMVU top-1/top-5, the 104-clip held-out odd classes, mirrored input, mirror-TTA average and CISLR-test top-1. In-corpus (INCLUDE held-out, ~0.80–0.83) is a guardrail only. Frozen probes rank encoders in the wrong order on this problem; decisions are taken on fine-tuned gates with at least two seeds.

### 6.3 Proposed work — status

Reference expansion: 4,928 ISLRTC units in the pool, ingestion continuing. Reference hygiene: applied, extended to ~2k audited gloss-matched clips in fine-tuning. Personal enrolment: *pending*. Encoder retrain: became the JEPA program (0.3318 → 0.4818; Phase C in training). Decode: per-epoch window vote and mirror TTA shipped. Signer study: *pending*, after Phase C.

## 7 Approval record

Per the programme-office instructions for Review 0, faculty approval is recorded below before the session. Reviews 0 through 2 are conducted by the support faculty; the final review is with the main faculty.

Candidate signature

---

Mithunish Prabhakaran · RA2512049015065

Support faculty (Reviews 0–2), signature & date

---

Dr. S Saravanan · saravans2@srmist.edu.in

Main faculty (final review), signature & date

---

Dr. Pitchaimanikam · pitchaib@srmist.edu.in

## 8 Glossary

Every technical, sign-language, dataset, and institutional term the documents use, each explained in a sentence or two. The same entries power the docs site’s glossary and its search.

### 8.1 Sign language & ISL

**Indian Sign Language (ISL)** The sign language used by the Deaf community across India — a complete natural language with its own grammar and word order, not a signed encoding of Hindi or English.

**Gloss** The written label for a single sign, conventionally in capitals (YESTERDAY, SCHOOL). A gloss names the sign; it is not a translation of the sentence the sign appears in.

**Word-level (isolated) sign language recognition** Recognising one sign at a time from a video clip or a live window — the task this case study addresses. Contrast with continuous recognition.

**Continuous sign language recognition** Recognising an unsegmented stream of signing, where sign boundaries must be discovered. Harder than word-level recognition and out of this case study’s scope.

**Sign language translation (SLT)** Producing grammatical spoken-language sentences from signing — a sequence-to-sequence task beyond recognition. This project stops at gloss streams plus rule-based assembly.

**Fingerspelling** Spelling out words letter by letter using a manual alphabet, typically for names and loanwords. Scoped out of this case study as a dedicated modality.

**Signing space** The region in front of the signer’s torso and face where signs are articulated. This project uses it operationally: hands below signing space are treated as idle.

**Dominant hand** The hand a signer leads with; left- and right-dominant signers produce mirrored performances of the same sign, which recognition must treat as identical.

**Chirality** Handedness of a performance — whether it is produced (or captured) left- or right-handed, including mirror flips introduced by selfie cameras. The matcher is chirality-invariant because handedness is not part of a sign’s identity.

**Static hold (held sign)** A sign whose content is a sustained handshape and orientation rather than movement — most number signs, for example. Holds are first-class citizens in this project’s matcher.

**Non-manual markers** Facial expression, mouth shape, head and torso movement that carry grammatical meaning alongside the hands. Tracked as face reference points here, but not yet modelled as grammar.

### 8.2 Institutions & programmes

**ISLRTC** The Indian Sign Language Research and Training Centre, a Government of India institute under the Ministry of Social Justice and Empowerment. Its official ISL dictionary videos are one of this project’s reference sources.

**SRMIST** SRM Institute of Science and Technology — the university this M.Tech (Artificial Intelligence) case study is conducted at.

**AI4Bharat** An IIT Madras research initiative building open AI resources for Indian languages; the group behind the INCLUDE dataset and the OpenHands sign-language toolkit.

**SDG 10 — Reduced Inequalities** The UN Sustainable Development Goal this case study aligns with: technology access should not depend on hearing, connectivity, or the availability of an interpreter.

**Review 0 / 1 / 2 / final review** The SRM case-study review cadence: proposal (Review 0), mid-term progress (Reviews 1–2) with the support faculty, and the final review conducted by the main faculty.

### 8.3 Datasets & benchmarks

**INCLUDE** The standard ISL word-level benchmark (AI4Bharat, ACM Multimedia 2020): 4,287 studio recordings over 263 word classes in 15 categories. This project’s closed-set head is trained and evaluated on it.

**CISLR** Corpus for Indian Sign Language Recognition (EMNLP 2022): 7,050 videos across roughly 4,700 words, mostly one performance per word — the dataset that frames large-vocabulary ISL as one-shot retrieval, and the source of this project’s open-vocabulary protocol.

**WLASL** Word-Level American Sign Language dataset (WACV 2020): 2,000 glosses across 21,000+ videos — the scale reference for word-level recognition in a resource-rich sign language.

**MS-ASL** A Microsoft ASL benchmark (BMVC 2019) of 1,000 classes from over 200 signers in unconstrained recording conditions.

**iSign** An ISL processing benchmark (ACL Findings 2024) of 118K+ video-sentence pairs with a suite of sentence-level tasks. Cited here to mark the boundary between recognition and translation.

**ISLRTC dictionary** The official public ISL dictionary programme: thousands of per-word demonstration videos by certified signers. This project processes them into landmark reference clips — the video itself is never redistributed.

## 8.4 Machine learning

**Pose estimation / landmarks** Detecting body and hand keypoints (landmarks) in each video frame, reducing appearance to a skeleton. This project’s entire pipeline operates on an 85-point landmark skeleton, never on pixels.

**Embedding** A learned fixed-size vector representation (here, 256 dimensions per landmark window) in which distance means dissimilarity — the substrate that makes recognition-by-retrieval possible.

**Metric learning** Training an embedding so that performances of the same sign land close together and different signs land far apart — recognition then reduces to nearest-neighbour lookup instead of a fixed classifier.

**ArcFace / CurricularFace (margin losses)** Training losses from face recognition that enforce an angular margin between classes, producing highly discriminative embeddings; CurricularFace adds a curriculum that emphasises hard examples over time.

**Prototype** A single representative embedding standing for a class — here, the embedding of a word’s dictionary reference. Recognising by distance to prototypes is what lets one example per word suffice.

**One-shot learning** The regime where a class must be recognised from a single labelled example. Large-vocabulary ISL is effectively one-shot: most words have exactly one dictionary performance.

**Closed-set recognition** Classification over a fixed set of classes decided at training time (INCLUDE’s 263). Accurate but unable to grow with the language.

**Open-vocabulary recognition** Recognition over a vocabulary that can grow without retraining — formulated here as retrieval against a bank of per-word reference embeddings covering 8,165 words.

**Retrieval** Answering “which word was signed?” by ranking stored references by similarity to the query, instead of running a fixed classifier — the formulation that scales to dictionary-sized vocabularies.

**Calibration** Making a model’s confidence scores mean what they say — a calibrated 0.9 is right about nine times in ten — so that emission thresholds behave predictably.

**Quantisation (int8)** Storing weights and reference vectors as 8-bit integers instead of 32-bit floats, shrinking artifacts roughly fourfold so an 8,165-word bank ships to a browser.

**Fine-tuning** Continuing training of an existing model on new data or a new objective. Three fine-tuning tracks in this project produced  $\leq 0.3pp$  — the measurement behind the data-not-model diagnosis.

**Ablation** Removing or swapping one component at a time to measure what each part actually contributes — the experimental discipline the semester’s retrain phase commits to.

**Top-k accuracy / enrolled top-1** A prediction counts if the right word appears in the model’s k best guesses. “Enrolled” means every one of the 8,165 vocabulary words is a candidate — nothing is excluded to flatter the number.

**Frozen protocol / held-out test** An evaluation set and procedure fixed in advance and never trained or tuned on. This project treats its 2,285-query CISLR test as sacred for exactly this reason.

**Domain shift** The gap between the data a system was built on (studio dictionary recordings) and the data it meets in use (webcams, home lighting, untrained signers) — a named risk in the proposal.

## 8.5 System & project terms

**On-device inference** Running the entire recognition pipeline on the user’s own hardware, inside the browser tab — no server round-trip, no video upload, and it works offline. In this project a construction property, not a policy.

**WebAssembly (WASM)** A portable binary format browsers execute at near-native speed; with SIMD and threads it is what makes running a neural network inside a tab practical.

**ONNX Runtime Web** Microsoft’s inference engine for the ONNX model format compiled to WebAssembly — the runtime this project’s quantised encoder executes on, at ~14 ms per window.

**MediaPipe** Google’s on-device perception framework; its holistic pose and hand models supply the per-frame landmarks this project consumes at camera rate.

**Canonical space** The normalised coordinate frame all landmarks are mapped into — origin at the mid-shoulder point, scale set by shoulder width — so camera distance, framing, and signer height stop mattering.

**Recognition window** A 48-frame slice of the landmark stream — the unit the encoder scores. Overlapping windows are voted over so a sign is recognised once, not once per window.

**Prototype bank** The int8 matrix of one 256-dimensional embedding per vocabulary word ( $8,165 \times 256$ , about 2 MB) that open-vocabulary retrieval scans on every window.

**Tier A / Tier B** The learned pipeline’s two heads: a calibrated closed-set classifier over the 263 INCLUDE classes (tier A) and cosine retrieval over the full prototype bank (tier B), composed in a decode cascade.

**Direct match** This project’s training-free pipeline: live landmark trajectories matched straight against dictionary reference fingerprints, with no model in the loop. Both a fallback recogniser and the instrument behind the model-vs-data attribution.

**Fingerprint (trajectory)** A compact orientation-aware feature track distilled from a reference clip or live window — wrist paths, forearm and finger directions, handshape ratios — the unit the direct-match aligner compares.

**Dynamic time warping (DTW)** A classic alignment algorithm that matches two sequences that unfold at different speeds. This project’s aligner descends from it, with sign-specific changes: condensed poses, dwell weighting, coverage penalties.

**Dwell weighting** Weighting each pose by how long it is held, so a sustained handshape carries its full duration of evidence while one-frame transitions barely count — how holds and movements coexist in one matcher.

**Signing-space & motion gates** Cheap checks that decide whether a window is worth scoring at all: are the hands raised into signing space, and is anything actually happening? They keep idle video from generating words.

**Audio fingerprinting (Shazam)** Identifying a recording by matching a compact signature against a reference library rather than by open-ended learning — the industrial precedent the direct-match pipeline is modelled on.

**Gloss stream / sentence assembly** The sequence of recognised glosses, and the rule layer that reorders them into a readable sentence. An optional LLM polish exists strictly as user opt-in.

**Temporal voting** Aggregating predictions across overlapping windows before emitting a word, trading a few hundred milliseconds of latency for large stability gains.

**Teach pack / .isl format** The project’s 8,187 per-word landmark reference clips in a compact binary format (.isl). It drives the 3D-mannequin dictionary, the practice mode, and the direct-match fingerprints alike.

**3D mannequin (Learn / Practice)** A rigged 3D avatar that performs any dictionary word by replaying its landmark clip — the dictionary you can watch from any angle, and the reference the practice mode scores you against.

**Natural-signing harness** A committed regression suite that matches synthetic independently-constructed signers — not replayed references — against the pack, and fails the build if naturally-performed signs stop ranking first.

## 8.6 Research & publishing

**Non-archival** A document or venue whose publications do not count as prior publication — these review documents are non-archival, so future papers can build on their text without dual-publication conflict.

**Self-overlap / dual publication** Reusing one’s own published text or results across submissions. Venues restrict it; this project’s rule is to check the target venue’s policy at every submission.

**Reproducibility** The property that every reported number can be regenerated by running the committed evaluation harnesses — the standard this document set holds itself to, and papers inherit.

**Informed consent** Written agreement from study participants who understand what is collected and why, with the right to withdraw. The signer study proceeds only under it — and collects no video at all.

**Dataset licence audit** Re-reading every source dataset’s release terms and documenting what use they permit — scheduled in this project before any paper submission or public artifact.

**Venues referenced** CVPR, WACV, BMVC (computer vision); ACL, EMNLP, LREC-COLING (language); NeurIPS, AAAI (machine learning); ACM Multimedia, ISMIR (media); ASSETS, W4A, CHI (accessibility and HCI); ICVGIP (Indian vision & graphics). The publication path targets the accessibility and language groups.

## References

- [1] D. Li, C. Rodriguez-Opazo, X. Yu, and H. Li, “Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison,” in *WACV*, 2020.
- [2] H. R. Vaezi Joze and O. Koller, “MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language,” in *BMVC*, 2019.
- [3] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, “INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition,” in *ACM Multimedia*, 2020.
- [4] A. Joshi, A. Bhat, and P. S. Basu, “CISLR: Corpus for Indian Sign Language Recognition in a Low-Resource Setting,” in *EMNLP*, 2022.
- [5] A. Joshi, “iSign: A Benchmark for Indian Sign Language Processing,” in *Findings of ACL*, 2024.
- [6] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” in *CVPR*, 2017.
- [7] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” in *AAAI*, 2018.
- [8] S. Jiang, “Skeleton Aware Multi-modal Sign Language Recognition,” in *CVPR Workshops*, 2021.
- [9] M. Boháček and M. Hruš, “Sign Pose-based Transformer for Word-level Sign Language Recognition,” in *WACV Workshops*, 2022.
- [10] P. Selvaraj, G. NC, P. Kumar, and M. Khapra, “OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages,” in *ACL*, 2022.
- [11] C. Lugaresi, “MediaPipe: A Framework for Building Perception Pipelines,” 2019.
- [12] F. Zhang, “MediaPipe Hands: On-device Real-time Hand Tracking,” in *CVPR Workshop on Computer Vision for AR/VR*, 2020.
- [13] J. Snell, K. Swersky, and R. Zemel, “Prototypical Networks for Few-shot Learning,” in *NeurIPS*, 2017.
- [14] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” in *CVPR*, 2019.
- [15] Y. Huang, “CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition,” in *CVPR*, 2020.
- [16] H. Sakoe and S. Chiba, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978.
- [17] A. Wang, “An Industrial-Strength Audio Search Algorithm,” in *ISMIR*, 2003.
- [18] Microsoft, “ONNX Runtime Web – WebAssembly execution provider.” [Online]. Available: <https://onnxruntime.ai/>
- [19] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, “Neural Sign Language Translation,” in *CVPR*, 2018.