

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

*Review 1 progress record: the measured real-world gap, the shipped fixes,
and the self-supervised training redirection it justified*

Candidate	Mithunish Prabhakaran
Registration No.	RA2512049015065
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 1 · Sunday, 30 August 2026 · 2:00–6:00 PM · SRM KTR campus, in person
SDG alignment	SDG 10 — Reduced Inequalities
Living documents	sanket.kadal.cc/docs/review-1 — this PDF is the printable snapshot

Deliverables per programme-office instructions: Literature Survey · Novelty Justification ·
Architectural Design for the Proposed System · Expected Outcomes

Continues the Review-0 posture: original prose, venue-verified citations,
pre-registered experimental gates, ethics and licensing owned in the design.

What changed since Review 0 — the honest headline

Review 0 proposed the system with preliminary in-corpus numbers. Since then the project **measured the number that matters** — accuracy on a real signer the model has never seen, through the exact bytes the deployed browser computes — and found the supervised model falls from **82% in-corpus to 33% cross-corpus** (12% for a left-hand-dominant signer). This handout presents that measurement, its decomposition, the fixes already shipped, and the training redirection it forced: latent-predictive self-supervision (ISL-JEPA) over a 118,168-unit unlabeled ISL pool, validated by two pre-registered pilot runs. Every number is measured unless explicitly marked as a target.

Progress since Review 0 — at a glance

Thread	At Review 0 (9 Aug)	At Review 1 (30 Aug)
Recognition models	Open-vocabulary router (8,188 words) live in the browser	1. 260-sign dictionary classifier shipped (9.9 MB int8, calibrated posterior, ~2 Hz)
Real-world evidence	In-corpus numbers only	Cross-corpus harness built; 0.82 → 0.33 measured and decomposed; mirror 0.12
Fixes shipped	—	Always-on mirror TTA (0.12 → 0.33 for left-dominant signers); vocabulary-conditional posteriors
Training approach	Supervised + augmentation plan	ISL-JEPA designed; A1/A2 pilots run against pre-registered gates — mirror cliff eliminated, diversity slope positive
Data	INCLUDE + CISLR + RKMVU key-points	118,168-unit unlabeled pool across 6 sources (~100 h), all in the deployed canonical space
Verification	Manual testing	Scripted live harness drives real footage through every studio surface before a bundle ships
Spend	—	Pilots ≈ \$2 GPU · corpus ≈ \$12 CPU — measured, not estimated

Since review 1 — status at 4 September 2026

Phase B ran on the full pool and **missed its pre-registered 0.50 gate** at 0.3273 — informatively: the last-layer JEPA target collapsed (target-std 0.98 → 0.52) while layers ≤ 4 still transferred, so the shipped encoder is cut at layer 4 (int8 0.3318, 4.57 MB, a tie with the 9.90 MB supervised model). Fine-tuning that encoder with gloss-matched dictionary clips, synthetic signers and a recover-and-resample augmentation, ensembling four seeds and distilling reached **0.4818** on the unseen signer (gate **0.4591**). A pilot of the Phase C objective (an all-layer teacher) adds +4.8 pp at equal budget; the full run is in training.

0.1 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	0.4818 (int8 0.4864)	0.4591	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.

1 Literature survey

The Review-0 survey (datasets; appearance- and pose-based recognition; one-shot retrieval; template matching; on-device inference) stands as submitted. This survey covers the threads that became load-bearing **after** Review 0, once the cross-corpus measurement (§2) showed the supervised approach is insufficient for real-world signers.

1.1 Latent-predictive self-supervision: the JEPA line

Joint-Embedding Predictive Architectures learn by predicting the *latent representation* of hidden content from visible context — never reconstructing pixels or coordinates. I-JEPA established the claim for images [1]; V-JEPA carried it to video with temporal span masking [2]; V-JEPA 2 scaled it to 1.2B parameters over ~1M hours of video, reaching state-of-the-art motion understanding and zero-shot robot planning [3]. V-JEPA 2.1 added a dense predictive loss, deep self-supervision, and released distilled checkpoints down to an 80M ViT-B — small enough to serve as a *teacher* for compact students [4]. The relevance is structural: a predictor can only succeed on what is predictable across the corpus, so signer-specific style — exactly what our measurement shows the supervised model memorises — is unpredictable noise that the objective discards.

1.2 Self-supervision on skeletons

S-JEPA is the blueprint for landmark data: predict the latents of masked joints, with motion-aware masking, an EMA target encoder and target centering against collapse; a vanilla transformer pretrained this way set state of the art on NTU-60/120, with the largest margins in the low-label regime — our regime [5]. Skeleton2vec [6] and the IJCV benchmark of skeleton SSL [7] replicate the core finding: latent targets beat coordinate reconstruction for skeletal action features.

1.3 Sign-language-specific self-supervision

SignBERT+ pioneered hand-model-aware masked pretraining [8]. SHuBERT is the strongest recent result: masked *multi-stream* cluster prediction — hand, face and body streams located by MediaPipe — over ~1,000 hours of ASL, transferring to state of the art across translation, isolated recognition and fingerspelling [9]. SignMAE shows masking should follow the *articulators*, not uniform patches [10]. SSL-SLR adds a caution this project adopts: contrastive negatives are ill-posed for sign because different signs share sub-movements [11]. Uni-Sign demonstrates large-scale unified sign pretraining [12]. On the data side, YouTube-SL-25 released video IDs for 3,207 hours across 25+ sign languages — including 3,023 Indian Sign Language videos, the largest open ISL source we know of [13].

1.4 The signer-independent turn

Evaluation methodology has converged on this project’s central concern: the MSLR workshop (ICCV 2025, CVPR 2026) runs an explicit signer-independent track — train on some signers, evaluate on unseen ones [14] — and continuous-SLR work treats each signer as a domain, penalising signer identity in the features [15]. This validates the protocol this project reached empirically: only a held-out *corpus* (different signer, studio, framing) measures what a deployed webcam will see.

1.5 What the threads leave unsolved

The literature gives us	It does not give us
Latent prediction as the anti-texture objective	Any JEPA on sign-language landmarks, or any ISL application
Skeleton SSL recipes with collapse rails	Skeleton SSL confronted with cross-corpus sign evaluation
Multi-stream sign SSL at ASL scale	The same for ISL, or on-device deployability
A signer-independent evaluation ethos	A byte-exact deployed-pipeline evaluation contract

3,000+ h of open ISL video IDs	That corpus turned into training-ready landmark data
--------------------------------	--

2 Novelty justification

2.1 The measurement: a byte-exact, deployed-pipeline cross-corpus harness

Published ISL numbers are computed in training frameworks. This project evaluates through the **exact numeric contract the deployed browser executes** — same canonicalisation, same aspect correction, same upper-median shoulder clamp, same int8 graph — replayed offline over 220 dictionary recordings by a signer, studio and framing the model has never seen.

Condition (220 real-signer clips)	Top-1
In-corpus, unseen signers (group-disjoint INCLUDE)	0.82
Cross-corpus, deployed contract	0.33
Cross-corpus, translated +0.15 / scaled $\times 1.25$	0.33 — <i>bit-identical</i>
Cross-corpus, mirrored (left-hand-dominant signer)	0.12

The decomposition is the contribution: position, height and framing are *provably solved* (the perturbed evaluations return the same bits); hand dominance is catastrophic and unhandled; and the residual gap has flat confusions — a distribution-wide style shift, not broken classes. Within-corpus validation, even signer-disjoint, structurally cannot see any of this.

2.2 Shipped inference novelties (measured, in production)

- **Always-on mirror test-time augmentation** — every dictionary-mode window is scored in both orientations and the more confident side answers. Measured: a left-hand-dominant signer recovers from 0.12 to the full 0.33 at ~ 0.5 pp cost to the majority orientation. The mirror map is exact in canonical space because the normalisation is mirror-symmetric.
- **Vocabulary-conditional posteriors** — a user’s vocabulary restriction renormalises the closed-set posterior to the honest conditional instead of masking scores.

2.3 ISL-JEPA: invariance as the pretext task

To our knowledge the first JEPA formulated on sign-language landmarks. Three masks over one EMA target encoder: **M1** hides an articulator stream (hand or body) and predicts its latents from the remaining streams; **M2** hides motion-aware temporal spans; **M3** — the genuinely new element — feeds a *deliberately corrupted view* (mirrored, non-uniformly tempo-warped, morphology-rescaled, tracker-noised) as context that must predict the **clean clip’s** latents. The measured real-world nuisances stop being augmentation regularisers and become the objective: same sign $\Rightarrow$ same latents, by construction.

Pilot evidence, pre-registered. Both pilots ran against gates written before launch:

Model (single seed, 150 epochs)	In-corpus	Cross-corpus	Mirrored
Supervised baseline	0.82	0.33	0.12
A1 · JEPA, INCLUDE-only, fine-tuned	0.75	0.23	0.21
A2 · JEPA + 2.5 $\times$ unlabeled, fine-tuned	0.76	0.28	0.28

A1’s gate passed on the mirror criterion — the dominance cliff is *eliminated by the objective*. A2’s gate (“a clearly positive slope from added unlabeled diversity”) passed on every metric, including a held-out class half whose clips never entered the pool even unlabeled (+2.9 pp). At 10.8k unlabeled clips, JEPA sits 3.4 pp under the supervised model cross-corpus *while being dominance-invariant* — and 11 $\times$ more prepared data is already banked.

2.4 The corpus machine

A 118,168-unit unlabeled ISL pool in the deployed canonical space: the Review-0 corpora joined by the re-extracted ISLRTC dictionary, 29.5k iSign sentence clips (streamed clip-by-clip out of a 54 GB gated archive via ranged reads of its central directory — the archive itself is never downloaded), and hand-presence-filtered windows from YouTube-SL-25's ISL subset. All of it passes through the browser's own extraction contract, so pretraining, evaluation and deployment share one numeric space.

Claims we are careful not to make

Pilot numbers are single-seed ($\pm 2-3$ pp) at a short schedule; the JEPA path has not yet beaten the supervised model's raw cross-corpus top-1; and JEPA cannot repair tracker failure. The Phase B run (2 seeds, 300 epochs, full pool, optional teacher distillation) exists to convert the measured slope into a crossing; its gate — cross-corpus ≥ 0.50 with in-corpus ≥ 0.80 — is already written down. *Since review 1*: that gate was missed (0.3273) and reported; the fine-tuning program has since reached 0.4818, 15 pp above the supervised model — on a 599-clip gate where one clip is 0.17 pp, which is why every number is a multi-seed mean.

3 Architectural design for the proposed system

Three architectures bound by one numeric contract.

3.1 The deployed inference architecture (running today)

Everything below the camera executes in the user's browser tab; no video, frame or landmark leaves the device.

1. **Worker A — pose.** MediaPipe hand + pose landmarks (GPU delegate, WASM fallback) produce the schema-driven 85-point canonical frame with per-group presence at 24–30 FPS; frames are discarded immediately; landmarks stream over a SharedArrayBuffer ring (postMessage relay where isolation is unavailable).
2. **Worker B — open-vocabulary router.** A 6.3M-parameter int8 encoder scores 64-frame windows against the 8,188-word prototype bank; a tiered cascade emits confident words, dashed prototype matches, tappable shortlists, or first-class silence. A training-free DTW exemplar matcher runs beside it as witness and counterfactual.
3. **Worker C — hand-shape expert.** A lazily-downloaded second opinion on RGB hand crops; may revise or veto uncertain words; structurally absent until first escalation.
4. **Worker D — dictionary classifier.** The 260-sign closed-set model (9.9 MB int8) answers $\sim 2\times/s$ with a calibrated posterior; always-on mirror TTA; vocabulary filters renormalise. *Since 3 September: a 4.57 MB int8 JEPA encoder Cut@4 + attentive head, preprocessing inside the graph.*
5. **Main thread.** Windowed voting, rule-based sentence assembly; cloud polish only as explicit per-press option. Practice and Learn score against the same heads the Studio runs.

Measured latency: router p50 $\approx 10\text{--}16$ ms/window; dictionary + mirror TTA $\approx 75\text{--}92$ ms at 2 Hz. Every threshold and landmark index comes from the model bundle at runtime, so a retrained model is a bundle drop, not an app release.

3.2 The proposed training architecture (ISL-JEPA)

1. **Corpus machine.** Cloud task queues re-extract every source through the browser-parity contract (idempotent blocks, quota-aware feeders, resumable): output, 118,168 canonical units (~ 100 h).
2. **Pretraining.** An articulator-token transformer (~ 6 M parameters, deploy-class) with an EMA target encoder, target centering and a variance rail; masks M1–M3; per-epoch stratified sampling keeps the sentence corpus from drowning the dictionary corpora.
3. **Teacher distillation (Phase B option).** V-JEPA 2.1's distilled ViT-B embeds raw video where held; a distillation head regresses the landmark student onto the video teacher.
4. **Heads.** Attentive probe $\rightarrow$ light fine-tune for the 260-sign dictionary (selected by cross-corpus accuracy, never in-corpus); re-embedding of the 8,188-word bank; an episodic head for few-shot per-signer adaptation.
5. **Export & verify.** int8 ONNX through the existing export path; a scripted live harness replays real dictionary footage through every studio surface before anything ships.

3.3 The evaluation architecture (what decides)

The decision metric is cross-corpus top-1 through the deployed contract. Its discipline: a held-out class half that never enters the pool (anti-transductive); an in-corpus guardrail (≥ 0.80 group-disjoint); pre-registered gates (Phase B ships only at cross-corpus ≥ 0.50 with the guardrail held and mirror parity retained); and equal-compute augmentation-arm controls so “self-supervision helped” is a comparison, not a claim.

One contract, three systems

The same canonicalisation — mid-shoulder anchor, shoulder-width scale with the deployed upper-median clamp, aspect-corrected, absent groups zeroed — is executed by the browser at inference, by the corpus machine at extraction, and by the harness at evaluation. A schema change propagates to all three or none.

4 Expected outcomes

4.1 Primary outcome: a real-world-credible recogniser

Metric	At review 1	Now (4 Sep)	Phase B target	Stretch
Cross-signer accuracy (RK-MVU-220)	0.33 supervised · 0.28 JEP A-A2	0.4818 jepa-rr4 (int8 0.4864)	≥ 0.50	≥ 0.65
Two-corpus gate (+ CISLR test 379)	—	0.4591	tracks	tracks
Same, mirrored signer	0.12 supervised · 0.28 JEP A-A2	parity by construction; ensemble + mirror TTA 0.4641	parity	parity
In-corpus guardrail	0.82	~0.80–0.83	≥ 0.80	≥ 0.80
Held-out class half	0.30 JEP A-A2	0.5481	tracks	tracks
Model size (dictionary lane)	9.9 MB int8	4.57 MB int8	≤ deployed	≤ deployed
Latency in-browser	10–16 ms · 75–92 ms	budget unchanged; int8 JEP A lane re-measurement <i>pending</i>	unchanged	unchanged

The decision rule was pre-registered, and it was applied: Phase B alone missed the 0.50 target (0.3273) and the miss was reported against the gate. The equal-compute fallback did not ship, because the depth finding and the fine-tuning program lifted the same encoder to 0.4818 with the guardrail held — 1.8 pp from the target at less than half the deployed size. Phase C (in training; +4.8 pp gate in pilot) is the candidate to cross it.

4.2 System outcomes (by final review, end of November)

A bundle drop into the deployed app, verified by the scripted live harness across every surface; the 8,188-word bank re-embedded in the invariant space and re-scored on the frozen retrieval protocol; the 118k-unit corpus, its builders and the harness released with the report (licences permitting per source); and an honest model card stating cross-corpus, mirrored and in-corpus numbers side by side.

4.3 Research outcomes: semester-2 paper seeds

Seed	Working claim	Fed by
P1 · Systems	On-device, open-vocabulary ISL in the browser: architecture + measured cost model	Review 0 §system; §3 here
P2 · Measurement	The deployed-contract cross-corpus audit: within-corpus evaluation overstates ISL readiness	§2; the harness + decomposition
P3 · Method	ISL-JEP A: nuisance-view latent prediction for signer-invariant landmark encoders	§1–4; Phase B results

4.4 Risks, stated with mitigations

Risk	Mitigation in place
Phase B misses the 0.50 gate	Equal-compute augmentation baseline already built; the gate decides the shipping path mechanically
Single-seed pilot noise (±2–3 pp)	Phase B runs 2 seeds; both reported

Tracker failure bounds any landmark model	Measured tracker-noise channel inside M3; teacher distillation carries appearance knowledge landmarks lack
Corpus licence constraints	Every source research-licensed; per-source terms recorded; gated sources never redistributed as video
Compute budget	Pilots cost $\approx$ \$2 (measured); Phase B budgeted $\approx$ \$15–40 within reserved credits

4.5 Societal outcome (SDG 10)

The outcome this project is accountable to: an ISL recogniser that works for the signer in front of the webcam — either hand dominant, any body type, any camera framing — with their video never leaving their device. The dominance result is the concrete case: a left-handed signer went from a 12%-accuracy user to a full-accuracy one, first by a shipped inference fix, now by an objective that never learns the bias at all. The remaining gap is quantified, its closure is funded and scheduled, and the honest number will ship in the product’s own model card.

References

- [1] M. Assran *et al.*, “Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture,” in *CVPR*, 2023.
- [2] A. Bardes *et al.*, “Revisiting Feature Prediction for Learning Visual Representations from Video (V-JEPA),” *TMLR*, 2024.
- [3] M. Assran, A. Bardes, D. Fan, Q. Garrido, and N. Ballas, “V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning,” 2025.
- [4] L. Mur-Labadia, M. Muckley, A. Bar, M. Assran, N. Ballas, and A. Bardes, “V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning,” 2026.
- [5] M. Abdelfattah and A. Alahi, “S-JEPA: A Joint Embedding Predictive Architecture for Skeletal Action Recognition,” in *ECCV*, 2024.
- [6] R. Xu, L. Huang, M. Wang, J. Hu, and W. Deng, “Skeleton2vec: A Self-supervised Learning Framework with Contextualized Target Representations for Skeleton Sequences,” 2024.
- [7] J. Zhang, L. Lin, S. Wang, and J. Liu, “Self-Supervised Skeleton-Based Action Representation Learning: A Benchmark and Beyond,” *IJCV*, 2026.
- [8] H. Hu, W. Zhao, W. Zhou, and H. Li, “SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding,” *IEEE TPAMI*, 2023.
- [9] S. Gueuwou, X. Du, G. Shakhnarovich, K. Livescu, and A. H. Liu, “SHuBERT: Self-Supervised Sign Language Representation Learning via Multi-Stream Cluster Prediction,” in *ACL*, 2025.
- [10] K. Xie, Z. Cai, and K. Stefanov, “SignMAE: Segmentation-Driven Self-Supervised Learning for Sign Language Recognition,” in *ICPR*, 2026.
- [11] A. Basso Madjoukeng, J. Fink, P. Poitier, E. B. Kenmogne, and B. Frenay, “SSL-SLR: Self-Supervised Representation Learning for Sign Language Recognition,” 2025.
- [12] Z. Li, W. Zhou, W. Zhao, K. Wu, H. Hu, and H. Li, “Uni-Sign: Toward Unified Sign Language Understanding at Scale,” in *ICLR*, 2025.
- [13] G. Tanzer and B. Zhang, “YouTube-SL-25: A Large-Scale, Open-Domain Multilingual Sign Language Parallel Corpus,” in *ICLR*, 2025.
- [14] MSLR Workshop Organisers, “Multimodal Sign Language Recognition Workshop — signer-independent track.” [Online]. Available: <https://m-slrt.github.io/MSLR2026/>
- [15] R. Zuo and B. Mak, “Improving Continuous Sign Language Recognition with Consistency Constraints and Signer Removal,” *ACM TOMM*, 2024.