# Towards Scalable Sign Production Leveraging Co-Articulated Gloss Dictionary for Fluid Sign Synthesis

Aparna Agrawal[1], Seshadri Mazumder[1], C.V. Jawahar[1], and Vinay Namboodri[2]
[1]IIIT Hyderabad [2]University of Bath

# Presentation roadmap

1. Motivation and problem setup
2. Background and related work
3. Method
4. Dataset: SanketVaani-1K
5. Experiments and results
6. Critical discussion

# Why does this problem matter?

- Sign Language Production (SLP) aims to convert spoken or written language into understandable sign language video.
- For Indian Sign Language (ISL), this is especially important because accessibility gaps persist in education, healthcare, legal access, banking, and public services.
- The paper frames SLP not just as a vision problem, but as an **accessibility infrastructure problem**: how can we generate fluent sign content without requiring prohibitively large continuous sign corpora?
- The authors emphasize two practical realities:
  - ISL is under-resourced relative to ASL/BSL-style benchmarks.
  - Certified interpreters are limited, so scalable automated support is valuable.
- This motivates a system that can reuse a relatively small but carefully annotated dataset and still synthesize **new, previously unseen sentences**.

# Core challenge in sign language production

**Why naive video generation is hard**

- Sign languages are not simple word-by-word gesture sequences.
- Meaning depends on manual signs, co-articulation, timing, facial cues, mouthing, pointing, and context.
- Abrupt concatenation of isolated signs creates jerky, unnatural transitions.
- Full sentence generation models require large aligned corpora that are costly to collect.

**Why ISL is even harder**

- Limited public gloss-level annotated datasets.
- Existing ISL resources are often designed for recognition or translation, not production.
- Missing gloss boundaries and sign-type labels make compositional synthesis difficult.
- Standardization is still evolving, so labeling and retrieval must tolerate variation.

## Paper's main bet

Instead of learning full-sentence generation end-to-end, retrieve gloss-level sign segments and interpolate across boundaries using a strong modern video interpolation model.

# What is the paper's central idea?

## One-line summary

Build a **co-articulated gloss dictionary** from a carefully annotated ISL dataset, retrieve context-appropriate sign clips for each gloss, and stitch them into fluent sign video using **FramePack**-based interpolation.

1. Translate English text into an ISL gloss sequence.
2. Encode each gloss in context and search a vector database.
3. Retrieve the most contextually appropriate sign segment.
4. Interpolate between neighboring sign segments to reduce discontinuities.
5. Produce a final continuous sign video with preserved articulation and identity.

## Why this is interesting

It avoids the data hunger of end-to-end video generation while preserving the photorealism of real signer footage.

# Main contributions claimed by the paper

1. A **retrieval + interpolation** framework for sign production grounded in gloss-level alignment.
2. A new ISL dataset, **SanketVaani-1K**, with 1,000 interpreter-produced sentences and gloss boundary annotations.
3. Use of **contextual gloss embeddings** and vector retrieval to handle sense disambiguation.
4. Use of a strong video interpolation model, **FramePack**, to create visually smoother transitions across gloss boundaries.
5. Quantitative and user-study evidence suggesting better fluency and naturalness than several baseline methods.

# Where does this paper sit in the literature?

## Three neighboring research areas

1. **Sign Language Recognition (SLR):** map video to glosses/labels.
2. **Sign Language Translation (SLT):** map sign video to spoken language text, or vice versa.
3. **Sign Language Production (SLP):** generate sign video from spoken/written language.

- Earlier SLP work used 3D avatars and notation systems.
- Later work used modular pipelines such as text-to-gloss, gloss-to-pose, and pose-to-video.
- GAN-based and diffusion-based methods improved realism, but still face artifacts, temporal inconsistency, or high data requirements.
- This paper positions itself as a **modular, retrieval-based alternative** that is practical for low-resource settings.

# Why existing approaches are insufficient

**Avatar / notation-based systems**

- Good controllability.
- Can model grammar and non-manual markers explicitly.
- But often suffer from the *uncanny valley*.
- Visually less realistic than true video.

**Pose / GAN pipelines**

- Modular and learnable.
- But transitions are often awkward.
- Lower visual fidelity and temporal artifacts remain a problem.

**End-to-end generative video models**

- Potentially expressive.
- But data hungry and difficult to train in low-resource sign languages.
- Need large paired text-gloss-video resources.

**What this paper changes**

- Reuses real signer footage.
- Requires only gloss-level segmentation.
- Moves complexity to retrieval and interpolation rather than full generative synthesis.

# Dataset gap highlighted by the authors

| Dataset | Lang. | Gloss ann. | Segmented | Controlled env. | Key limitation for SLP |
|---|---|---|---|---|---|
| How2Sign | ASL | ✓ | ✗ | ✓ | Large resource, but not designed for compositional gloss-wise synthesis |
| Phoenix14T | DGS | ✓ | ✗ | ✗ | Useful benchmark, but weather-domain and not production-oriented |
| ISLTranslate | ISL | ✗ | ✗ | ✗ | Translation-focused, lacks boundary annotations |
| iSign | ISL | ✓ | ✗ | ✗ | Sentence-level glosses, but limited for gloss-segment retrieval |
| **SanketVaani-1K** | **ISL** | ✓ | ✓ | ✓ | **Designed to support retrieval and interpolation** |

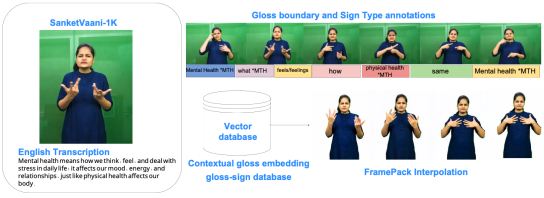
**Takeaway:** the paper argues that **annotation structure**, not raw dataset size, is the key requirement for retrieval-based ISL production.

# System overview

1. Input English sentence
2. LLM-based English-to-ISL gloss conversion
3. Contextual gloss embedding generation
4. Vector database search for matching sign clips
5. Stitching and interpolation across neighboring clips
6. Final continuous sign video

## Design philosophy

Keep the signer real, keep the units interpretable, and only synthesize the transition regions.



## Step 1: Gloss-level vector database construction

- Each sentence is first represented as a sequence of gloss tokens:

$$G = [g_1, g_2, \ldots, g_n]$$

- A contextual encoder (the paper mentions a pretrained language model such as BERT) produces a token embedding for each gloss using surrounding context.
- So the same gloss can map to different embeddings depending on neighboring glosses and sentence meaning.
- Every stored entry includes:
  - gloss label,
  - sign type (for example pointing or mouthing),
  - gloss position,
  - surrounding context,
  - pointer to the corresponding motion/video segment.

### Why contextual embeddings matter

A word like *light*, *bank*, or *fair* can correspond to different signs. The embedding helps retrieve the right sign **for the sentence context**, not just the lexical token.

# Retrieval objective and semantic role of the database

**Inference procedure**

1. Convert input sentence to glosses.
2. Compute contextual embedding for each query gloss.
3. Search nearest neighbors in the vector database.
4. Retrieve the motion segment with best contextual match.

$$\hat{s} = \arg\min_{s_j \in D} \mathrm{dist}(e_q, e_j)$$

where the paper uses cosine-style distance over stored contextual embeddings.

**Practical effect**

- Retrieval is modular and interpretable.
- New sentences can be synthesized compositionally.
- Ranking can consider sign type and gloss position in addition to embedding similarity.
- If confidence is low, the system can fall back to fingerspelling using isolated alphabet signs.

## Step 2: English-to-ISL gloss generation

- The paper argues that hand-written rule systems are brittle for unconstrained text.
- Instead, the authors use an LLM supplied with a compact **ISL rule card**.
- Reported rule examples include:
  - topic-comment word order,
  - article omission,
  - sentence-final WH words,
  - clause-initial time adverbials.
- The LLM emits an ISL gloss sequence whose length is determined dynamically by semantic segmentation.
- This is important because production quality depends on the quality of the intermediate gloss representation.

### Important modeling decision

The paper does **not** train a dedicated text-to-gloss model for ISL here; it relies on LLM prompting plus linguistic rules, which is pragmatic but also introduces dependence on prompt quality.

# Step 3: Why stitching is difficult

**If we simply concatenate clips**

- sudden hand-position jumps,
- broken motion continuity,
- unnatural changes in body pose,
- visible boundary artifacts,
- reduced intelligibility.

**Why sign is harder than generic motion**

- fine finger articulation,
- fast non-linear trajectories,
- both hands must stay synchronized,
- self-occlusions are common,
- facial detail matters.

**Why pose-space interpolation is insufficient**

- Linear interpolation in pose or latent space tends to create mechanical transitions.
- It does not faithfully model the non-linear dynamics of natural signing.
- Warping or avatar rendering often loses realism and expressive subtlety.

## Paper's answer

Interpolate **directly in video** between the last frame of one gloss and the first frame of the next gloss.

# Why the paper chooses FramePack

- FramePack was designed for next-frame prediction and long-video generation using efficient context packing.
- The key adaptation in this paper is to use it for **boundary interpolation** between neighboring gloss clips.
- For each boundary:
  - the last frame of segment $i$ and first frame of segment $i + 1$ are used as anchors,
  - intermediate frames are generated bidirectionally,
  - the anchors remain fixed, helping avoid drift.
- The authors argue this is superior to unidirectional generation, which can forget anchors or drift temporally.
- FramePack sits on top of a frozen Hunyuan-DiT backbone, giving strong generative capacity with lower overhead than re-training a full custom model.

# Method summary as a modular pipeline

| Stage | Input → output | Role in the system |
|---|---|---|
| Text to gloss | English sentence → ISL gloss sequence | Converts spoken-language syntax into an ISL-oriented intermediate representation |
| Gloss embedding | Gloss tokens → contextual vectors | Encodes lexical meaning plus sentence context |
| Vector retrieval | Query vectors → sign segments | Finds semantically appropriate signer video units |
| Interpolation | Neighboring segments → smooth transitions | Repairs discontinuities at boundaries while preserving realism |
| Composition | Retrieved clips + interpolated bridges → full sentence video | Produces the final sign output |

## Why this modularization is powerful

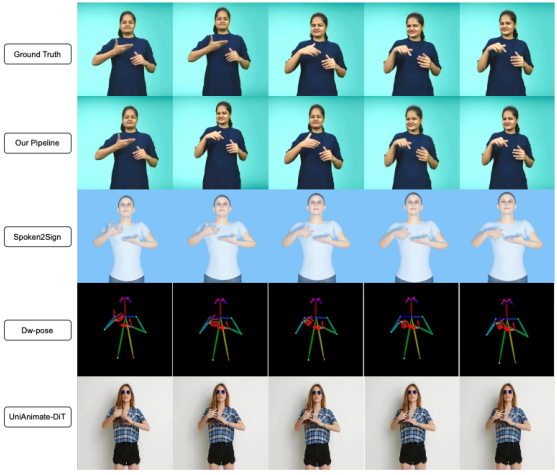
Each stage can be improved independently: better glossing, better retrieval, better interpolation, or better annotation immediately improves the final system.

# SanketVaani-1K: what was collected?

- 1,000 spoken sentences interpreted into ISL.
- Around 6k text vocabulary according to the paper's summary table.
- More than 40 practical domains.
- Single signer, controlled recording setup.
- Gloss annotations, sign-type tags, and temporal boundaries.

## Why the single-signer choice is deliberate

It improves consistency of articulation, appearance, and retrieval quality, which is beneficial for a first production-focused dataset.



# Why the domain design matters

- The dataset is intentionally skewed toward **high-utility communication scenarios** rather than academic benchmark convenience.
- Domains include healthcare, banking, finance, legal rights, government services, education, transportation, daily living, and mental health.
- This choice supports a realistic accessibility objective: the system should help in daily life interactions, not just narrow benchmark settings.

## Strong design choice

The paper emphasizes usefulness for Deaf users in the Indian socio-cultural context. That makes this dataset design more application-driven than many generic sign datasets.

## What the category chart suggests

Coverage is broad rather than dominated by one domain, which is useful for testing semantic retrieval across diverse contexts.

# Recording and preprocessing pipeline

1. Record with a front-facing GoPro10 at 1920×1080 and 30 fps.
2. Use clap cues at the beginning and end of each sentence to facilitate temporal segmentation.
3. Detect and track the signer with YOLOv9 + ByteTrack.
4. Expand and normalize the crop to maintain stable framing and scale.
5. Extract a foreground mask and perform video matting with MatAnyone.
6. Composite onto a neutral background for consistency across clips.

## Why this matters for retrieval-based synthesis

If signer scale, crop, and background vary too much, interpolation becomes harder and retrieval results become visually inconsistent. The preprocessing directly supports downstream stitching quality.

# Annotation granularity is one of the paper's strongest points

**What is annotated**

- free-form gloss labels,
- temporal boundaries for each gloss,
- sign types such as mouthing, pointing, and no-sign,
- cases where one spoken word maps to multiple signs.

**Why this is valuable**

- Enables segment-level retrieval.
- Supports contextual disambiguation.
- Makes the data useful beyond SLP: translation, gloss spotting, and retrieval.
- Preserves interpreter knowledge during annotation.

# Experimental agenda in the paper

1. Qualitative validation of the preprocessing pipeline.
2. Text-to-gloss evaluation, using RWTH-PHOENIX-Weather 2014T as a proxy benchmark because parallel ISL gloss data are unavailable.
3. Vector database retrieval analysis using ambiguous/polysemous words.
4. Interpolation comparison against DreamMover, BiM-VFI, and VFI-Mamba.
5. Comparison with other SLP styles such as Spoken2Sign and UniAnimate.
6. User study with Deaf raters to evaluate naturalness and understandability.

## Interpretation tip for the audience

The paper evaluates different modules separately because this is a modular system. That is appropriate, but it also means there is no single monolithic benchmark score that summarizes everything.

# Text-to-gloss evaluation: what do the numbers say?

| Metric | B1 | B2 | B3 | B4 |
|---|---|---|---|---|
| Score | 0.350 | 0.182 | 0.114 | 0.086 |

| Metric | Score |
|---|---|
| ROUGE-L | 0.393 |
| METEOR | 0.294 |
| CHrF | 0.473 |
| BERTScore F1 | 0.784 |

| Variant | TSOV | BERT F1 | Prec. |
|---|---|---|---|
| PHOENIX glosses | 0.72 | 0.723 | 0.704 |
| ChatGPT 4o glosses | 0.76 | 0.791 | 0.764 |

**Interpretation**

- Generated glosses are semantically competitive on the proxy benchmark.
- The LLM-generated glosses appear more aligned with canonical sign-language-oriented structure than the benchmark glosses themselves.

## Caution

This is not a direct ISL benchmark. It is evidence that the gloss-generation strategy is plausible, not definitive proof of ISL gloss correctness.

# Vector retrieval: can the database disambiguate meaning?

| Query sense | Rank | Top retrieved sentence theme |
|---|---|---|
| light (lamp) | 1 | Electrical devices / lights / household electricity |
| light (sun) | 1 | Photosynthesis, sunlight, plant growth |
| fair (carnival) | 1 | Rann Utsav festival in Gujarat |
| fair (just) | 1 | Shivaji's fair rule / justice-related sense |
| bank (finance) | 1 | Home loan, banks, documents, income, repayment |

## Why this matters

This is one of the strongest demonstrations in the paper. It shows the system is not retrieving signs solely by token identity; it is retrieving by **contextual sense**. That is critical for practical sign synthesis.

# Interpolation experiment design

- The paper focuses on gloss boundaries, where co-articulation errors are most visible.
- For each annotated boundary, the authors identify a midpoint and extract nine frames around it.
- The two outer frames are anchors; the seven middle frames are treated as ground truth targets.
- This produces 3,963 anchor-target pairs in the test set.
- Competing methods generate the seven missing intermediate frames.

## Baselines compared

DreamMover, BiM-VFI, VFI-Mamba, and the proposed FramePack-based interpolation.

## Metrics reported

PSNR, SSIM, LPIPS, FID, NIQE, MS-SSIM, GMSD, and KID.

# Interpolation results: quantitative table

| Method | PSNR | SSIM | LPIPS | FID | NIQE | MS-SSIM | GMSD | KID |
|---|---|---|---|---|---|---|---|---|
| DreamMover | 26.02 | 0.915 | 0.117 | 56.82 | 46.85 | 0.947 | 0.103 | 0.065 |
| VFI-Mamba | 27.81 | 0.950 | 0.066 | 52.96 | 70.12 | 0.961 | 0.094 | 0.079 |
| BiM-VFI | 27.88 | 0.956 | 0.103 | 39.77 | 58.84 | 0.956 | 0.097 | 0.031 |
| **FramePack** | 23.96 | 0.909 | 0.163 | 49.80 | 69.48 | 0.925 | 0.132 | 0.057 |

## Interesting outcome

The proposed method is **not best on many distortion metrics**. The paper explicitly argues that these metrics reward blur or pixel alignment and therefore can undervalue perceptually sharp, semantically faithful hand and face details.

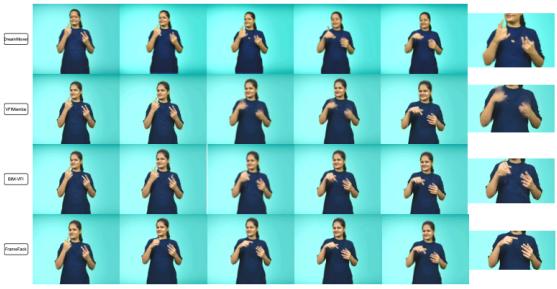
## Presentation note

This is a classic case where numerical image-quality metrics and human perception do not perfectly agree.

# Interpolation results: how the authors interpret them

**Reported qualitative behavior**

- VFI-Mamba: increasing blur, especially on hands and face.
- DreamMover: preserves global pose but introduces facial distortion and hand artifacts.
- BiM-VFI: visible artifact on one hand in an example.
- FramePack: better preservation of facial detail and hand articulation across the sequence.




## Key message

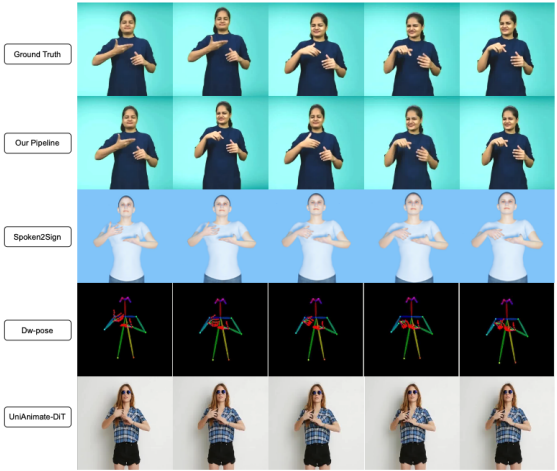
For sign language, preserving fine articulatory detail may matter more than maximizing generic image similarity metrics.

# Comparison with broader SLP methods

**Methods contrasted in the paper**

- **FramePack pipeline (ours):** interpolation-based, real video anchors.
- **Spoken2Sign:** avatar-based synthesis using SMPL-X.
- **UniAnimate-DiT:** diffusion-based animation from 2D keypoints.

**Paper's qualitative claim** The proposed pipeline better preserves signer identity and photorealism because it operates directly on real video segments.




Critical distinction

# User study design

- 12 participants contributed responses in the subjective evaluation.
- The study has two parts:
  1. compare interpolation quality in slowed-down clips containing anchors plus generated middle frames,
  2. compare full-sentence outputs for understandability, fluency, naturalness, and overall mean opinion score (MOS).
- The paper reports mean rank, number of first-place votes, Borda score, and MOS-style ratings.

## Why the user study is essential here

Automated metrics can miss the most important property: whether Deaf viewers actually find the output understandable and natural.

# User study results: interpolation preference

| Method | Only interpolation | | | In full video | | |
|---|---|---|---|---|---|---|
| | MR↓ | #1↑ | B↑ | MR↓ | #1↑ | B↑ |
| **Our pipeline** | 2.47 | 13 | 1.53 | 2.06 | 12 | 1.94 |
| DreamMover | 2.25 | 6 | 1.75 | 2.89 | 7 | 1.11 |
| VFI-Mamba | 2.61 | 9 | 1.39 | 2.36 | 7 | 1.64 |
| BiM-VFI | 2.67 | 8 | 1.33 | 2.69 | 10 | 1.31 |

## How to read this

The results are mixed in the *pure interpolation* view, but the proposed method looks strongest once embedded in the **full sign production pipeline**, which is the more meaningful end-use case.

# User study results: full SLP quality

| Method | Understandability | Fluency | Natural | Overall MOS |
|---|---|---|---|---|
| **Our Pipeline** | **3.12** | **3.00** | **3.33** | **3.15** |
| Spoken2Sign | 2.62 | 2.42 | 2.38 | 2.47 |
| UniAnimate | 2.08 | 2.08 | 2.00 | 2.06 |

## Most important empirical takeaway

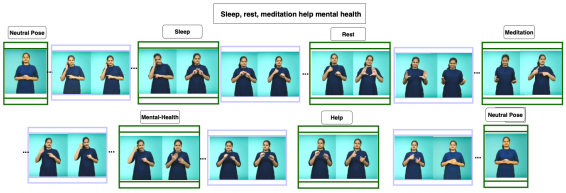
Human raters consistently preferred the proposed retrieval + interpolation pipeline over the compared avatar-based and diffusion-based baselines.

## Interpretation

This supports the paper's central thesis: for low-resource sign production, **reusing real signer segments plus strong interpolation** can outperform fully synthetic alternatives in perceived quality.

# Illustrative application example

- The paper includes a medical-use-case example.
- Example sentence: “Sleep, rest, meditation help mental health.”
- Retrieved sign clips are combined with interpolated segments, and neutral poses are inserted at the sequence boundaries.
- This is a good demonstration of how the system could support domain-specific accessible communication.



# What are the strongest aspects of this paper?

1. **Pragmatic system design:** it avoids unrealistic end-to-end data requirements.
2. **Annotation quality:** gloss boundaries and sign-type labels directly support generation.
3. **Contextual retrieval:** the vector database addresses lexical ambiguity.
4. **Human-centric evaluation:** Deaf user feedback is included rather than relying only on image metrics.
5. **Application relevance:** dataset domains are grounded in real communication needs.

## Overall impression

This is a strong systems paper for low-resource SLP because the dataset, retrieval design, and evaluation all support the same practical goal.

## What are the limitations?

- **Single signer**: improves consistency but limits signer diversity and broader generalization.
- **Approximate manual boundaries**: small alignment errors may propagate into interpolation artifacts.
- **Fixed-length interpolation**: real sign durations are context dependent.
- **Dependence on gloss quality**: errors in text-to-gloss conversion affect everything downstream.
- **Limited anonymity**: facial and articulatory detail are integral to the approach.
- **Standardization drift**: ISL lexemes and conventions continue to evolve.

### Important scientific point

The proposed method is highly useful, but it is not yet a universal sign video generation solution. It is a careful, practical step toward scalable ISL production.

# Future work proposed by the authors

- Refine onset and offset boundaries using motion and phonological cues.
- Add a duration predictor so interpolation span adapts to the specific context.
- Update labels as ISLRTC standardization evolves.
- Strengthen Deaf community representation in advisory input, annotation review, and evaluation.

## Natural next technical extensions

- multi-signer retrieval with identity-conditioned interpolation,
- explicit modeling of facial expression and non-manual markers,
- confidence-aware retrieval and compositional planning,
- downstream integration in public-service accessibility tools.

# Big-picture takeaway

## Main conclusion

The paper shows that **scalable sign production in a low-resource language does not necessarily require end-to-end generation from scratch**. A carefully designed gloss-annotated dataset, contextual retrieval, and high-quality transition interpolation can produce more useful and natural sign video.

- Real data + modular composition is a strong baseline.
- Human judgment matters more than generic image metrics alone.
- Dataset design is central to system design.
- This work opens a practical path for inclusive ISL content generation.

# Thank you

Questions?

Primary source: Agrawal et al., ICVGIP 2025