

Surveying Representation Collapse in Pretrained Models of American Sign Language

Lee Kezar[○]

Laurel Aichler[○]

Athena Willis[□]

Saki Imai[△]

Mert Inan[△]

Alicia Wooten[○]

Malihe Alikhani[△]

Lorna Quandt[○]

[○] Gallaudet University

[□] Independent Researcher

[△] Northeastern University

Abstract

The American Sign Language (ASL) lexicon contains heterogeneous sign types, including established signed words, neologisms, ad-hoc depictions of real-world phenomena, and fingerspelled words. Productions within each type exhibit unique phonotactics and substantial variability, potentially confounding models pretrained on downstream ASL tasks. To survey the capacity for representation collapse, we isolate and annotate 14 types of ASL productions as they appear in science, technology, engineering, and math (STEM) educational content. We then probe seven pretrained models' latent representations of the annotated clips, reporting collapse metrics with respect to domain, sign type, and meaning. Our findings show that out-of-distribution signs are more frequent in STEM content, and that pretrained models lose information about their visual and semantic contents in varying ways. Among the tested models, those which encoded the phonological properties of signs generalized to STEM data more effectively, while English- and gesture-based models can exhibit more brittleness. Future work can design interventions to mitigate the kinds of collapse documented here or extend this survey by evaluating additional models and datasets.

1 Introduction

Computational models of sign languages have the potential to make tasks like web searching or translation more accessible to deaf and hard-of-hearing (DHH) signers (Yin et al., 2021). However, that potential depends on the extent to which models accommodate the linguistic variation in signed utterances (Bragg et al., 2019; Desai et al., 2024). The American Sign Language (ASL) lexicon is typologically diverse, containing established signed words *depicting signs* that reproduce a referent's visual properties, *neologisms* for domain-specific or new concepts, and *fingerspelled* versions of English words (Neidle et al., 2022).

Although collectively frequent, most productions in the latter three sets are *out-of-distribution* (OOD) for pretrained ASL models because of their relatively high *variability* in form paired with chronic data scarcity in ASL. One level of variation is phonological: between sign types, contrasts in meaning are expressed with a unique distribution of phonological features and are structured with a distinct set of phonotactic regularities (Sandler and Lillo-Martin, 2006). Additionally, sign type usage varies with respect to dialect, register, and other contextual variables. These factors of variation may drastically increase the sparsity and reduce the frequency of entire sign types, while established signed words remain more frequent and less variable in form. Prior work has shown that sparsity in visual and language data can undermine representational quality and generalization (Joshi et al., 2020; Desai et al., 2024), motivating our central research question: *How well do pretrained ASL models generalize to phonologically distinct OOD productions?*

To examine the potential for *representation collapse*, this work evaluates seven pretrained encoders' basic receptive and comprehension skills on newly-annotated data from science, technology, engineering, and math (STEM) content (§3.1, 3.3). We show that these domains use more OOD terms as compared to open-ended narratives, potentially due to the absence of established terms for specific concepts and the practicality of depiction in describing shape, movement, and position (Sortino et al., 2026; Lualdi et al., 2023). We also selected these domains in support of making educational content more accessible to students whose primary language is ASL. Where there might be potentially several synonymous expressions of one scientific concept, search and translation technologies could facilitate the retrieval of ASL content and offer in-class lexical support.

This work’s data collection and analysis are organized around three central research questions:

Q1. *Which articulatory signals improve tokenization on OOD signing, and how sensitive are pretrained models to segmentation error?*

We use sign language tokenization (or segmentation) both to sample individual sign clips for annotation and as a predictor of representation collapse. In §4.1 and §5.1, we compare boundary judgments for tokenizers based on external movement, internal movement, their combination, and combined movement augmented with ISR uncertainty. We additionally explore the extent to which segmentation error predicts OOD transfer in §5.2.

Q2. *How do decisions about training data, architectures, and objective functions shape the loss of articulatory information under OOD transfer?*

Prior work on sign-input modeling has progressed on tasks like translation, isolated sign recognition (ISR), contrastive retrieval, and self-supervised reconstruction via training on multilingual sign languages, gesture, and English text embeddings. In §3.2 and §4.2, we probe seven pretrained representations using sign-type classification, ASLLRP-to-STEM transfer, English sentence-context prediction, and collapse diagnostics (§4.3; §5.2). Here OOD primarily denotes phonologically distinct or productive ASL forms.

Q3. *To what extent is representation collapse influenced by sign type, corpus, tokenization, and annotation artifacts?*

The cause of representation collapse can be spread across corpus differences, segmentation errors, and annotator disagreement. To separate these factors, §4.3 measures class geometry and domain separability, while §4.4 tests sensitivity to boundary quality and annotator disagreement.

Together, these analyses clarify some of the underlying mechanics of ASL representation collapse and show that pretrained models have idiosyncratic strengths and weaknesses in this regard. We release the boundary- and sign-type-annotation data to support future efforts to improving ASL modeling generalizability to OOD data.¹

¹<https://github.com/leekezar/SurveyingCollapseASL>

2 Related Work

To ground our investigation into representation collapse on out-of-distribution sign types, we first review key linguistic properties of the ASL lexicon in §2.1. We then examine dominant sign language modeling paradigms alongside broader structural failures in deep learning in §2.2.

2.1 ASL Lexicography and Phonology

Research in ASL lexicology has shown that the ASL lexicon is typologically rich, comprising a *native core* of established signed words, *non-native* fingerspelled forms, and a highly productive component of classifier constructions (Brentari and Pad-den, 2001). Pertinent to computational modeling, the space of possible productions within each type rely on distinct phonotactic regularities (Yanovich et al., 2016). Because deep learning architectures naturally optimize for the holistic retrieval of high-frequency targets, they may be acutely vulnerable to representation collapse when exposed to certain novel combinations of phonemes. We operationalize these sign types through both the ASL Linguistic Research Project (ASLLRP; Neidle et al. 2022), which documents types in open-ended ASL narratives from Deaf signers, alongside newly-annotated versions of publicly-available STEM content.

For sign tokenization (RQ1), we use the Movement-Hold Model as a potentially robust prior for predicting the boundaries of signed words (Liddell and Johnson, 1989). This framework posits that signed words are generally comprised of *movements*, where some features of the sign are in transition, and *holds*, where all features of the sign are relatively static. The full model is shown to describe a variety of sign types, including compounds and morphological inflections (Liddell, 1984), making it an attractive prior for OOD modeling.

A second set of phonotactic regularities relevant for this work concerns lexical innovation (Sornig, 1981). Signers productively compose morphemes in novel or infrequent ways, such as depicting a referent’s visual characteristics with *classifier constructions*, introducing a reusable term for something not presently established in the lexicon (*neologisms*, *namesigns*), and inflecting the quantity of certain nouns. At present, it is unclear how and to what extent these kinds of productions fit into the geometry of pretrained representation spaces.

Model	Objective	Architecture	Dim.	Data
CiCo (Cheng et al., 2023)	CL + retrieval	Video-text	1024	Isol. signs (multi), text (En.)
SignCLIP (Jiang et al., 2024)	CL + retrieval	Sign-text	768	Isol. signs (multi), text (En.)
SHuBERT (Gueuwou et al., 2025)	SSL + clusters	Video TF	768	Cont. signing (ASL)
Phonologizer (Kezar et al., 2023a)	Phon. + recon.	Pose TF	512	Isol. signs (ASL)
ViFi (Rasheed et al., 2022)	CL	Video-text	512	Gesture/action video + text
VideoMAE (Tong et al., 2022)	SSL + recon.	Video TF	768	Gesture/action video
ISR Baseline (this work)	ISR + phon. + recon.	Pose TF	512	Isol. signs (ASL)

Table 1: Representation models used for clip-level embedding extraction. CL = contrastive learning; SSL = self-supervised learning; ISR = isolated sign recognition; phon. = phonological supervision; recon. = reconstruction; TF = transformer; Isol. = isolated; Cont. = continuous; multi = multilingual; En. = English.

Dataset	Evaluative Purpose	Segments	Signers
ASLLRP	In-domain Pretrain, cont.	1,818	42
ASL Citizen	Lexical Baseline, isol.	62,127	47
Sem-Lex	Lexical Baseline, isol.	84,265	44
ASL STEM Wiki	OOD Target, cont.	2,945	24
Atomic Hands	OOD Target, cont.	3,223	1

Table 2: Source dataset roles, sizes, and number of signers. *cont.* = continuous; *isol.* = isolated sign data.

2.2 ASL Modeling and Representation Collapse

Pre-trained models for ASL recognition and translation tasks address data scarcity in idiosyncratic ways, including self-supervision (Gueuwou et al., 2025), contrastive learning (Jiang et al., 2024), and phonological inductive priors (Kezar et al., 2023a). These methods are trained on ASL datasets which may be augmented with pose estimation (Jiang et al., 2025), English word embeddings (Cheng et al., 2023), or gesture/action video pretraining (Tong et al., 2022) to encourage more generalizable features of language. As suggested in RQ2, these choices necessarily shape the geometry of the learned representation space, but their consequences are not well-understood beyond aggregate statistics.

Representation collapse, or the tendency of a pretrained encoder to distort or lose information in unintended ways, may provide an analytically-rich toolkit for measuring how helpful specific methods are in learning generalizable functions of language. In supervised models, neural collapse describes a late-training regime in which examples concentrate around class means and classifiers organize around those means, yielding high training-set separability while reducing variation within each category (Papayan et al., 2020). In contrastive learning, positive pairs are also pulled towards a class mean, however, those means are simultaneously repulsed from other means via negative pairs (Wang and Isola,

2020; Jing et al., 2022; Jiang et al., 2023). This repulsive force can be agnostic to how negative samples differ, resulting in an even spread of class means that risks marginalizing OOD classes, or biased by the samples’ image in a separate model’s latent space. The latter case is demonstrated in SignCLIP’s use of pretrained English word embeddings to learn visual-semantic representations of ASL signs (Jiang et al., 2024). Lastly, phonological inductive priors, which have both supervised (Kezar et al., 2023a) and self-supervised (Gueuwou et al., 2025) implementations in prior work, are meant to focus models on the reusable, recombinable components of ASL signs, however the incompleteness and inconsistency of phoneme inventories is a representational bottleneck.

3 Data Collection and Preprocessing

We collect ASL signing data from multiple sources and domains to study representation collapse under domain shift. Our preprocessing pipeline has three parts. First, reference and isolated sign resources support tokenizer development, sign type sampling, and ASLLRP-to-STEM transfer (§3.1). Second, continuous videos become pose sequences, candidate clips, visual embeddings, and sentence-context embeddings (§3.2). Third, ASL signers annotate STEM clips for boundary quality and sign type, producing target labels (§3.3).

3.1 Data Sources

Table 2 summarizes the five corpora used in this study. Counts refer to isolated videos for ASL Citizen and Sem-Lex and candidate clips for continuous corpora. The American Sign Language Linguistic Research Project (ASLLRP) is the comparison corpus for ASLLRP-to-STEM transfer (Neidle et al., 2022). ASL Citizen (Desai et al., 2023) and Sem-Lex (Kezar et al., 2023b) provide isolated sign data used to train and evaluate the isolated sign recognition and phonologizer models (Kezar

et al., 2023a). After segmentation, these recognizers identify clips dissimilar to established signed words through recognition uncertainty; details appear in §3.2 and Appendix I.

Our target evaluation uses two STEM corpora. ASL STEM Wiki contains ASL interpretations of Wikipedia articles related to STEM (Yin et al., 2024), while Atomic Hands² is a YouTube channel providing STEM educational content natively produced in ASL and later translated into English, used with limited permission from its creators. These sources provide target STEM clips for annotation, probing, and transfer evaluation.

3.2 Preprocessing and Representation Extraction

For each candidate clip i , the analysis tuple is

$$C_i = \langle x_i^{\text{POSE}}, d_i, y_i, z_i^{\text{SEM}}, \{z_i^m \mid m \in \mathcal{M}\} \rangle,$$

where x_i^{POSE} is the pose sequence, $d_i \in \mathcal{D}$ is the corpus/domain label, $y_i \in \mathcal{Y}$ is the majority sign type label, $z_i^{\text{SEM}} \in Z_{\text{SEM}}$ is the English sentence-context embedding, z_i^m is a clip-level visual embedding from model m , and $\mathcal{M}$ is the set of visual models in Table 1.

Each sentence-level source video s_i is converted into a MediaPipe pose sequence s_i^{POSE} with 75 two-dimensional landmarks per frame. Continuous pose sequences are then segmented into clips x_i^{POSE} using the tokenizer described in §4.1. Candidate clips shorter than 0.35 seconds are excluded before representation extraction because development annotations identified them as fragments of adjacent well-formed segments. Each candidate clip is embedded by the pretrained visual models $\mathcal{M}$ listed in Table 1. We encode its surrounding English sentence or caption with bert-base-uncased, producing a 768-dimensional sentence-context embedding. The space Z_{SEM} represents associated English context, not ASL semantics or gloss identity. All analyses pair each visual embedding with its metadata: dataset, video, signer where available, and sentence-context embedding.

3.3 Annotation Workflow

Four co-authors who are conversant in ASL annotated the STEM clips used in this analysis. The annotators include two Deaf researchers whose primary language is ASL and two hearing researchers

who can converse in ASL. Their average ASL experience is 20.25 years, with a minimum of 10 years and a maximum of 34 years. Three have formal training in ASL linguistics, and all four primarily study STEM topics (biology, cognitive science, human computer interaction, and natural language processing).

Each annotator received a set of 150 video clips to label using a custom interface: 50 clips that are *shared* across annotators for measuring agreement and 100 clips that are *unique* to each annotator. The 50 *shared* clips and 75 of the *unique* clips are those which received the highest mean perplexity score among the evaluated models. This is done to increase coverage over OOD signs, such as scientific neologisms, fingerspelling, and classifier predicates. The remaining 25 clips in the *unique* subset are randomly sampled to estimate sign type frequency by dataset. The candidate pool was balanced across ASL STEM Wiki and Atomic Hands and capped per source video.

For each clip, annotators selected *start*- and *end*-boundary judgments, each *early*, *good*, or *late*. They also marked clips that contained multiple signs, were too short to annotate, or were otherwise invalid or unclear. They additionally annotated nine top-level sign types: *lexical sign*, *fingerspelling*, *loansign*, *name sign*, *number sign*, *compound*, *gesture*, *neologism*, and *classifier*. Classifier annotations could additionally receive one or more of seven subtypes (Appendix A). Multiple top-level labels were permitted when a production crossed categories; analyses requiring one label retained only rows with a single majority label. We compute inter-annotator agreement on the 50 shared clips.

4 Method

We evaluate three levels corresponding to the research questions: tokenizer boundary quality (RQ1, §4.1), linear accessibility of sign type and English sentence-context structure (RQ2, §4.2, §4.3), and geometric, corpus, boundary, and annotation controls (RQ3, §4.4).

4.1 Tokenizer Evaluation

Given a pose sequence $x^{\text{POSE}} = (x_1^{\text{POSE}}, \dots, x_T^{\text{POSE}})$, tokenization predicts an ordered boundary set $B = \{b_1, \dots, b_K\}$ with $1 < b_1 < \dots < b_K < T$. They define candidate sign clips $x_{b_j:b_{j+1}}^{\text{POSE}}$.

²<https://www.youtube.com/@AtomicHands>

The tokenizer follows the Movement-Hold model (§2.1). Holds, the temporal spans where phonological parameters are not in transition, defines a likely separation between morphemes. We operationalize this feature as reductions in hand velocity around time t . Let $m_{\text{ext}}(t)$ denote external movement, i.e. wrist velocity, and $m_{\text{int}}(t)$ denote internal movement, i.e. finger velocity relative to the wrist. It smooths their sum to identify low-motion intervals. Candidate boundaries are placed at local minima of the smoothed velocity profile within low-motion intervals; Appendix B gives the calibrated implementation details.

We also evaluate recognizer uncertainty as a boundary cue, following the prior that true boundaries will yield lower perplexity than boundaries that capture sign fragments. The uncertainty-augmented tokenizer adds normalized ISR lexical uncertainty to the same candidate-selection procedure.

We evaluate tokenizer quality with manually reviewed boundary judgments. We report boundary outcomes by tokenizer, dataset, and sign type in §5.1. We condition downstream sign-type prediction on boundary quality to assess segmentation effects on transfer (§5.2).

4.2 Representation Probes

For each segment i , we analyze its visual embeddings z_i^m using probes trained on one of three targets: domain d_i , sign type y_i , and BERT embedding of the semantic context z_i^{SEM} .

Domain classification $f_{\text{DOM}} : Z_m \rightarrow \mathcal{D}$ measures sensitivity to dataset-specific participants, artifacts, and genre. The probe M_{dom}^m approximates f_{DOM} with a grouped, class-balanced linear classifier that distinguishes ASLLRP from STEM within each supported sign type. We report ROC AUC averaged across test folds; 0.5 (chance) indicates model invariance to dataset source, while 1.0 (perfect recoverability) indicates high sensitivity to domain.

Sign type classification $f_{\text{TYPE}} : Z_m \rightarrow \mathcal{Y}$ asks whether a representation preserves coarse visual properties and high-level phonotactic regularities that distinguish segments with respect to their type. For example, as compared to most established lexical signs, fingerspelling tends to combine low external movement with high internal movement. We train a multinomial linear probe $M_{\text{type}}^m(z_{im}, \theta) \approx p(Y | x_i)$ and report its Macro-F1 and accuracy across test folds.

Semantic prediction, $f_{\text{SEM}} : Z_m \rightarrow Z_{\text{SEM}}$, measures to what extent the preserved phonological contrasts encode the clip’s coarse-grained meaning. Here, we operationalize meaning using the BERT embedding of x_i ’s caption. This JEPA-style task (Assran et al., 2023) is learned with a ridge-linear map:

$$\hat{z}_i^{\text{SEM}} = M_{\text{SEM}}^m(z_i^m, \theta) = A_m z_i^m. \quad (1)$$

The f_{SEM} evaluation uses held-out-caption folds, so captions cannot occur in both train and evaluation portions of a fold. We report this probe overall and by sign type because different sign types convey meaning through different formal cues; within-type contrasts may depend on subtler handshape, movement, or spatial distinctions than those needed for sign type classification.

We evaluate probes with STEM cross-validation, ASLLRP-to-STEM transfer, and held-out-caption, held-out-video, and held-out-signer target folds. Exact fold construction and support counts are reported in Appendix D. We report accuracy and Macro-F1; Macro-F1 is primary because sign type support is unequal.

4.3 Collapse Metrics

While probing experiments point to the possible existence of representation collapse, they do not delineate between different kinds of collapse. Such delineation could be advantageous for explaining and preventing collapse in future models. For example, given a sign type y , the image of X in Z_m , $\{z_i^m | y_i = y\}$ can be flattened along certain dimensions, suggesting loss of articulatory information, or spread across other types, suggesting sensitivity to non-linguistic factors.

We additionally compare each image with its semantic counterpart in Z_{SEM} to contextualize probe performance on the semantic embedding prediction task f_{SEM} . If the visual representations in Z_m occupy a substantially denser region than their semantic counterparts in Z_{SEM} , then m may be incurring a loss of articulatory information that is severe enough to influence downstream performance, while expansion relative to Z_{SEM} may signal brittleness or hallucination.

We operationalize these geometric perspectives below. All three diagnostics are computed on L2-normalized clip embeddings; Appendix E gives their formal definitions.

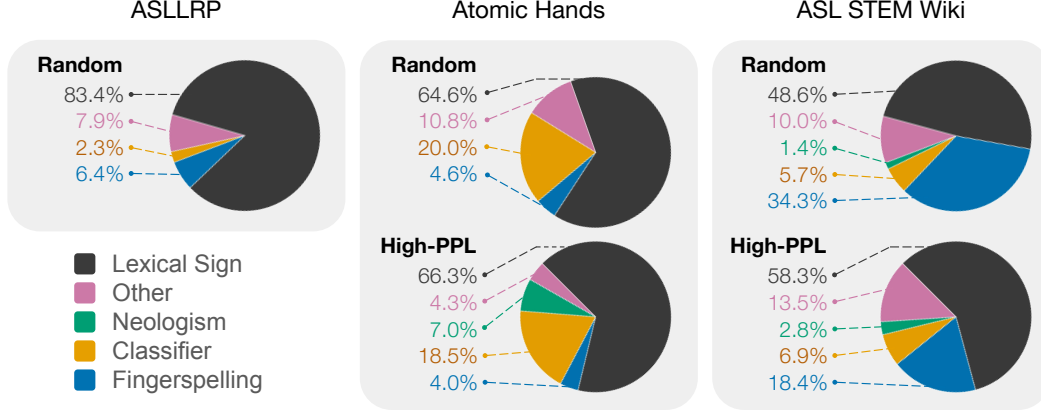


Figure 1: Frequency of major sign type annotations by dataset and sampling strategy (random vs. high perplexity or PPL). As compared to ASLLRP, the STEM data contain more classifiers (+17%) and fingerspelling (+28%).

Effective dimensionality. For model m , sign type y , and domain $D \in \{\text{ASLLRP}, \text{STEM}\}$, let $X_{ym}^D = \{z_i^m : y_i = y, d_i = D\}$ denote the corresponding set of clip embeddings. We summarize its covariance spectrum with participation-ratio effective dimensionality: the number of directions that materially account for within-type variation, rather than the representation’s nominal vector size. The transfer ratio $\rho_{ym}^d = d_{\text{eff}}(X_{ym}^{\text{STEM}})/d_{\text{eff}}(X_{ym}^{\text{ASLLRP}})$ therefore measures whether y occupies fewer, comparable, or more effective directions in STEM than in ASLLRP. Values below 1 indicate *contraction* of within-type variation under transfer; values above 1 *expansion*.

Local sign type structure. For each STEM clip i , $N_k(i)$ contains its $k = 10$ nearest other clips under cosine distance. Target kNN impurity I_k is the mean proportion of neighbors in $N_k(i)$ whose sign type differs from y_i : high impurity means that nearby clips frequently belong to different sign types. Target Fisher separation F_T compares the dispersion of sign-type centroids with the dispersion of clips around their own sign-type centroid. Larger F_T indicates that sign types form more distinct target-domain clusters.

Form-context density. For each clip, we estimate local density in both Z_m and the English sentence-context space Z_{BERT} from the inverse mean cosine distance to its ten nearest neighbors. $D_{m,y}$ is the median, within sign type y , of the log ratio between visual and English-context density. Values near zero indicate similarly concentrated visual and sentence-context neighborhoods; positive values indicate that clips are more locally concentrated in the visual space; negative values indicate greater concentration in the English-context space.

4.4 Data-Fidelity and Statistical Analysis

The collapse metrics described thusfar are influenced not only by the quality of Z_m , but also by factors that are external to the model: annotator agreement, boundary quality, and target-corpus composition. We measure these factors directly to further contextualize the collapse measurements.

Annotation entropy H_i measures how concentrated annotator sign type labels are per clip. The annotation entropy gap Δ_m^H compares probe performance on low- versus high-entropy clips. For **boundary quality**, we report transfer Macro-F1 by segmentation outcome and summarize sensitivity as clean-boundary performance minus non-clean performance. For **target-corpus composition**, sign type frequency gives each type’s share of the annotated corpus sample.

5 Results

We report annotation results (§5.1) followed by probing (§5.2) and collapse metrics (§5.3).

5.1 Annotation Analysis

Table 3 and Appendix B show segmentation errors by feature set, corpus, and sign type. Removing internal or external velocity increases the invalid rate by 12% and 22%, respectively, supporting the Movement-Hold motivation for combining both streams. ISR uncertainty did not improve reviewed-boundary quality: combined velocity gives 9/40 good boundaries (22.5%, Wilson 95% CI 12.3–37.5), compared with 4/20 (20.0%, 8.1–41.6) for velocity plus uncertainty.

For the selected tokenizer, most errors are shifted boundaries rather than unusable clips. Clean-boundary rates are higher for lexical signs (.27) and

Condition	n	Good	1 err.	2 err.	Inv.
Ext. velocity	20	.15	.30	.00	.55
Int. velocity	20	.15	.10	.30	.45
Ext. + int. velocity	40	.23	.38	.08	.33
Velocity + ISR uncertainty	20	.20	.35	.15	.30
<i>Selected tokenizer: corpus breakdown</i>					
ASL STEM Wiki	268	.24	.40	.22	.14
Atomic Hands	182	.17	.41	.23	.19
<i>Selected tokenizer: sign-type breakdown</i>					
Lexical	285	.27	.37	.24	.12
Fingerspelling	53	.02	.55	.36	.08
Classifier	67	.21	.46	.16	.16
Neologism	23	.13	.61	.13	.13

Table 3: Boundary-quality outcomes. “1 err.” and “2 err.” denote shifted boundaries; “Inv.” combines multiple-sign, fragment/too-short, and invalid/unclear outcomes.

Model	I_k	F_T	AUC
ISR	.378	.190	.813
Phonologizer	.366	.175	.823
CiCo	.440	.038	.960
SignCLIP	.408	.067	.975
SHuBERT	.395	.029	.971
ViFi	.408	.037	1.000
VideoMAE	.419	.032	1.000

Table 4: Diagnostics complementary to Figure 2. I_k : target kNN impurity ($\downarrow$); F_T : target Fisher separation ($\uparrow$); AUC: within-type ASLLRP-versus-STEM recoverability ($\downarrow$, chance .5).

classifiers (.21) than neologisms (.13) and fingerspelling (.02), consistent with internally complex productive forms being harder to segment. Annotator agreement is moderate for both sign type ($\alpha = .370$) and boundary category ($\alpha = .381$), with dataset and sign-type decompositions reported in Appendix C.

Figure 1 shows sign-type frequencies from random annotation samples alongside High-PPL samples used for OOD-focused review. The random panels show more productive or less-established STEM material than ASLLRP; High-PPL sampling further enriches these OOD categories. These samples define the target composition for representation probes and collapse diagnostics.

5.2 Representation Collapse and Diagnostics

The representation probes split target-set accessibility from transfer (Figure 2a; Table 7). SignCLIP has the strongest STEM cross-validation Macro-F1 (.202), but the weakest ASLLRP-to-STEM Macro-F1 (.073). ISR shows the opposite profile: lower STEM cross-validation Macro-F1 (.158), but the highest ASLLRP-to-STEM transfer

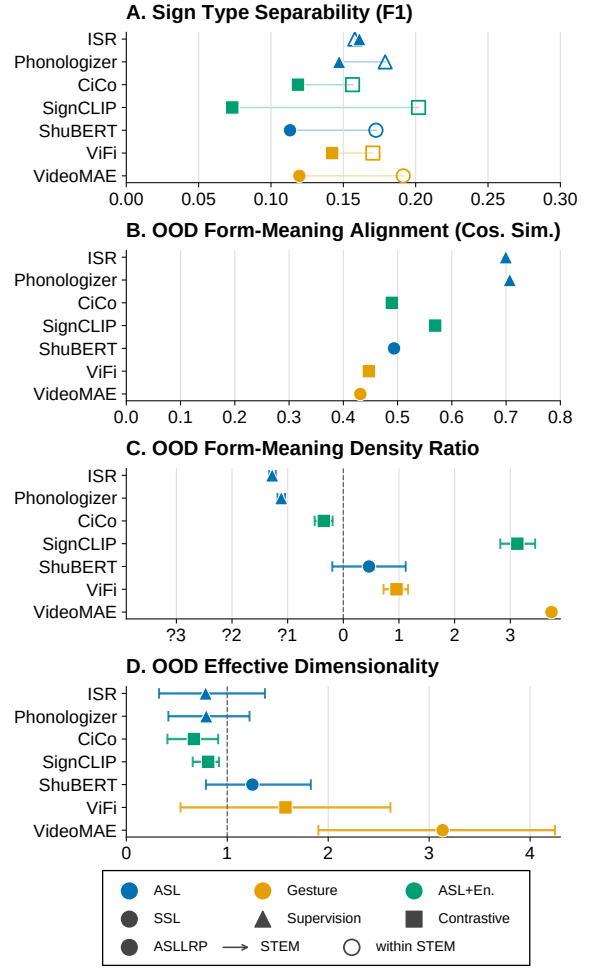


Figure 2: Four OOD diagnostics by representation; error bars are 95% bootstrap intervals where supported.

(.161). Phonologizer is near the top in both settings, with .179 STEM Macro-F1 and .147 ASLLRP-to-STEM Macro-F1. Thus, the strongest model within STEM is not the strongest model under transfer from ASLLRP.

Collapse metrics partially explain this reversal (Figure 2b-d; Table 4). ISR and Phonologizer have the clearest local target geometry, with the lowest kNN impurity (.379 and .367) and highest target Fisher separation (.190 and .175). They also have the lowest mean within-type corpus AUCs (.813 and .823), indicating less recoverable corpus identity. By contrast, CiCo, SignCLIP, ViFi, SHuBERT, and VideoMAE all show stronger corpus recoverability (.960–1.000), suggesting that their representation space is more entangled with dataset source.

Effective-dimensionality ratios show sign-type-specific rather than uniform collapse (Table 9). Model-level ratios range from .670 for CiCo to

3.134 for VideoMAE, but types move oppositely within models. Lexical sign dimensionality expands for ISR, Phonologizer, and SHuBERT, while neologism dimensionality contracts for six of seven models.

Form-to-context results add a second axis of collapse (Figure 2; Table 8). Phonologizer and ISR have the highest held-out-caption cosine (.707 and .699), and their advantage holds across lexical signs, fingerspelling, classifiers, and neologisms. However, visual-to-English-context density is closest to zero for CiCo ($D_m = -.351$), not for ISR or Phonologizer.

5.3 Fidelity Controls

Boundary and annotation controls do not explain representation differences (Appendix 10). Boundary-effect estimates range from -.004 to .096, and annotation-entropy gaps range from -.001 to .013 across models. These effects are smaller than the main transfer reversal, especially SignCLIP’s .129 drop from STEM cross-validation to ASLLRP-to-STEM transfer. The overall pattern is therefore geometric rather than purely annotational, despite the reduced agreement scores: ASL-trained pose models preserve more transferable sign-type structure, while several high-performing target probes also encode stronger corpus identity or stronger target-specific reshaping.

6 Discussion

We return to our central question: *How well do pretrained ASL models generalize to phonologically distinct OOD productions?* The reported metrics provide evidence that models’ representational quality is deeply sensitive to sign type and domain, with noticeably worse performance in STEM contexts as compared to open-ended narratives. Among the tested models, those which encoded the phonological components of ASL emerge as the most generalizable, while those which use English- or gesture-based pretraining tasks seem more reliant on dataset-specific correlations.

RQ1: Tokenizer Generalizability. Boundary quality depends more on articulatory velocity than recognizer uncertainty. Combining internal and external velocity reduced invalid clips, while adding ISR perplexity did not improve the reviewed sample. Segmentation errors affect downstream transfer, but less than the main differences among representation spaces.

RQ2: Collapse across modeling choices. ASL-internal models transfer most reliably. ISR has the highest ASLLRP-to-STEM F1, and Phonologizer remains close while producing the strongest form-to-context cosine. These results support lexical and phonological priors as partial protection against collapse.

English-aligned and gesture/action models show a different profile. SignCLIP performs best under STEM-only validation but worst under ASLLRP-to-STEM transfer, while ViFi and VideoMAE preserve broad target-domain variation but make corpus identity highly recoverable. One possible explanation for the collapse of English-aligned models is the inclusion of signs that do not have a unigram English translation, instead using visual characteristics iconically.

RQ3: Collapse versus evaluation artifacts. The collapse diagnostics show that model differences are not explained by annotation ambiguity or segmentation quality. Boundary effects and annotation entropy gaps are small relative to the spread in transfer F1, domain AUC, effective dimensionality, and Fisher separation. ISR and Phonologizer preserve clearer target sign-type geometry and lower corpus recoverability, while VideoMAE and ViFi show larger geometric shifts and stronger domain sensitivity.

Conclusion. No single representation space captures all of the structure needed for productive ASL under STEM-domain transfer. ASL-specific lexical and phonological models transfer more reliably, English-aligned models perform better in some target-only settings, and gesture/action models preserve broad visual variation that is relatively decoupled from semantic contents. These patterns are consistent with representation collapse as a shift in which articulatory, sign type, and contextual distinctions remain available after pretraining.

Future work should test whether richer phonological labels and other objectives that preserve productive structure can improve transfer without relying too heavily on lexical patterns or corpus-specific cues. Success on these tasks could lead to more externally valid ASL models, and in the case of STEM content, improved educational access for deaf and hard-of-hearing learners.

6.1 Limitations

The main limitation is that the corpora differ in domain, genre, signer population, recording condi-

tions, and annotation structure. Those differences are part of the transfer problem, but they also make it difficult to attribute model behavior solely to linguistic structure. The high domain AUC values show that several models can separate corpora despite the pose estimation preprocessing step.

The STEM target data cover two sources, which don't capture the true breadth and variability of natural STEM ASL signing. ASL STEM Wiki reflects interpreted STEM content, while Atomic Hands reflects ASL-first educational explanations, later translated into English. These sources are valuable but cannot represent all STEM ASL, regional variation, classroom practice, or signer-specific approaches to technical vocabulary.

The sign type inventory also simplifies a more continuous linguistic space. Categories such as lexical sign, fingerspelling, classifier, gesture, compound, and neologism are useful for analysis, but individual productions can fall between or blend these categories. This is especially important for productive signs, lexicalized fingerspelling, and domain-specific neologisms. We preserve overlapping annotations because disagreement can reflect real category ambiguity rather than annotator error or poorly-defined constructs.

Although the annotations include experienced ASL signers, including two Deaf signers with ASL as their primary language, the number of annotators ($n = 4$) limits the stability of agreement estimates. Future work should include more Deaf annotators and a broader range of regional, educational, and linguistic backgrounds.

Additionally, all downstream embeddings depend on the tokenizer, and we did not re-embed the full dataset under every segmentation variant. The current analyses estimate whether boundary quality affects transfer, but they do not fully characterize how each tokenizer would change every representation result.

Finally, the model comparisons are descriptive rather than causal. The models differ in their pre-training data, architectures, objectives, modalities, dimensionality, and language coverage. The observed family-level patterns should therefore be treated as emerging evidence rather than controlled ablations. A causal account would require pre-training experiments that vary one design choice at a time, which is beyond the scope of the current work.

References

- Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabat, Yann LeCun, and Nicolas Ballas. 2023. [Self-supervised learning from images with a joint-embedding predictive architecture](#). *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629.
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. [Sign language recognition, generation, and translation: An interdisciplinary perspective](#). In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '19*, page 16–31, New York, NY, USA. Association for Computing Machinery.
- Diane Brentari and Carol A. Padden. 2001. Native and foreign vocabulary in american sign language: A lexicon with multiple origins. In Diane Brentari, editor, *Foreign Vocabulary in Sign Languages: A Cross-Linguistic Investigation of Word Formation*, pages 87–119. Lawrence Erlbaum Associates, Mahwah, NJ.
- Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. CiCo: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026.
- Aashaka Desai, Lauren Berger, Fyodor O. Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daumé III, Alex X. Lu, Naomi Caselli, and Danielle Bragg. 2023. ASL citizen: A community-sourced dataset for advancing isolated sign language recognition. *arXiv preprint arXiv:2304.05934*.
- Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024. [Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas](#). In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 54–65, Torino, Italia. ELRA and ICCL.
- Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025. [SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria. Association for Computational Linguistics.
- Ruijie Jiang, Thuan Nguyen, Shuchin Aeron, and Prakash Ishwar. 2023. [Hard-negative sampling for](#)

- contrastive learning: Optimal representation geometry and neural- vs dimensional-collapse. *arXiv preprint arXiv:2311.05139*.
- Zifan Jiang, Colin Leong, Amit Moryossef, Oliver Cory, Maksym Ivashechkin, Neha Tarigopula, Biao Zhang, Anne Göhring, Annette Rios, Rico Sennrich, and Sarah Ebling. 2025. [Meaningful pose-based sign language evaluation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 64–80, Suzhou, China. Association for Computational Linguistics.
- Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Sign-CLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193.
- Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2022. Understanding dimensional collapse in contrastive self-supervised learning. *International Conference on Learning Representations (ICLR)*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.
- Lee Kezar, Riley Carlin, Tejas Srinivasan, Zed Sevcikova Sehyr, Naomi Caselli, and Jesse Thomason. 2023a. [Exploring strategies for modeling sign language phonology](#). In *Proceedings of the 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2023)*, pages 577–582.
- Lee Kezar, Elana Pontecorvo, Adele Daniels, Connor Baer, Ruth Ferster, Lauren Christiansen Berger, Jesse Thomason, Zed Sevcikova Sehyr, and Naomi Caselli. 2023b. [The sem-lex benchmark: Modeling ASL signs and their phonemes](#). In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*.
- Scott K. Liddell. 1984. [Think and believe: Sequentiality in american sign language](#). *Language*, 60(2):372–399.
- Scott K. Liddell and Robert E. Johnson. 1989. [American sign language: The phonological base](#). *Sign Language Studies*, pages 195–277.
- Colin P. Lualdi, Barbara Spiecker, Alicia K. Wooten, and Kaitlyn Clark. 2023. [Advancing scientific discourse in american sign language](#). *Nature Reviews Materials*, 8(10):645–650.
- Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022. [ASL video corpora & sign bank: Resources available through the american sign language linguistic research project \(ASLLRP\)](#). *arXiv preprint arXiv:2201.07899*.
- Vardan Papyan, X Y Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117(40):24652–24663.
- Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2022. Finetuned clip models are efficient video learners. *arXiv*.
- Wendy Sandler and Diane Lillo-Martin. 2006. *Sign Language and Linguistic Universals*. Cambridge University Press.
- Karl Sornig. 1981. *Lexical Innovation: A Study of Slang, Colloquialisms, and Casual Speech*. John Benjamins, Amsterdam.
- Rachel Sortino, Christina Kim, Thalia Guettler, Katie McClyman, Lorna Quandt, Alicia Wooten, and Rachel G. Pizzie. 2026. [Producing concept-motivated signs supports learning of stem in american sign language](#). *npj Science of Learning*.
- Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.
- Polina Yanovich, Carol Neidle, and Dimitris Metaxas. 2016. Detection of major ASL sign types in continuous signing for ASL recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3067–3073, Portorož, Slovenia. European Language Resources Association (ELRA).
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.
- Kayo Yin, Chinmay Singh, Fyodor O. Minakov, Vanessa Milan, Hal Daumé III, Cyril Zhang, Alex Xijie Lu, and Danielle Bragg. 2024. [ASL STEM Wiki: Dataset and benchmark for interpreting stem articles](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14474–14490, Miami, Florida, USA. Association for Computational Linguistics.

A Sign Types

Table 5 shows the sign types and the definitions shown to all the annotators. This is largely based on Neidle et al. (2022) with the addition of neologisms, which is pertinent to STEM.

B Tokenization and Boundary Quality

Figure 3 expands Table 3 by showing how the selected combined-velocity tokenizer behaves by sign type. This decomposition shows whether segmentation errors are concentrated in sign types likely to contain short, internally complex, or productive forms. For implementation, the combined velocity signal is $V(t) = m_{\text{ext}}(t) + m_{\text{int}}(t)$ and low-motion interval boundaries are selected by $b = \arg \min_{t \in [t_s, t_e]} \tilde{V}(t)$. The uncertainty-augmented condition scores candidate split frames with $S(t) = \lambda P_b(t) + (1 - \lambda)U(t)$, where $P_b(t)$ is the reference-boundary score and $U(t)$ is ISR uncertainty. The released configuration specifies the locally adaptive velocity thresholds, smoothing windows, minimum-duration constraints, and uncertainty-combination parameters calibrated on reference boundaries.

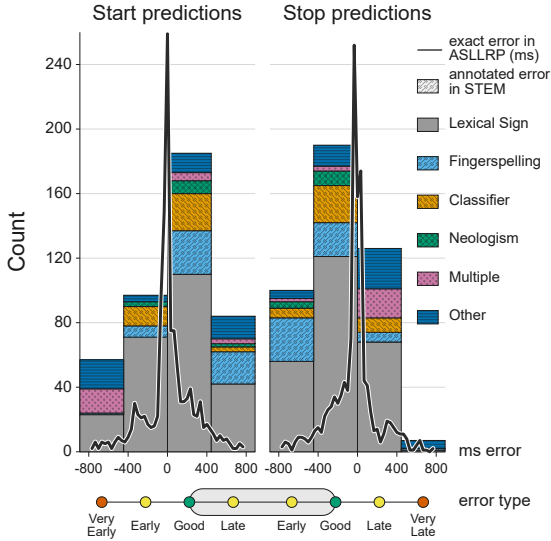


Figure 3: Boundary outcomes for the selected combined-velocity tokenizer by sign type. Bars show annotation categories; the line shows millisecond distance to reference boundaries where available.

C Annotation Agreement

Table 6 reports agreement separately by dataset and by the sign-type modes with enough duplicate annotations to summarize. The overall rows are

averages over the duplicate-annotated clips; the stratified rows show that agreement is not uniform across corpus source or sign type.

D Probe Results

The common probe set contains 454 target clips: 317 lexical signs, 51 fingerspelled forms, 54 classifiers, 12 neologisms, 8 number signs, 3 loansigns, 3 name signs, 3 compounds, and 3 gestures. The five latter cells are low support ($n < 10$). All seven models use the same target rows.

E Metric and Statistical Details

Diagnostic summaries use the same normalized embeddings as the probe experiments. The semantic probe is fit by minimizing $\mathcal{L}_m = \sum_i \|\hat{z}_i^{\text{SEM}} - z_i^{\text{SEM}}\|_2^2$. For model m , sign type c , and domain D , effective dimensionality is $d_{cm}^D = (\sum_j \lambda_j)^2 / \sum_j \lambda_j^2$, where λ_j are eigenvalues of the within-type covariance matrix. We report the STEM/ASLLRP ratio $\rho_{cm}^d = d_{cm}^{\text{STEM}} / d_{cm}^{\text{ASLLRP}}$. For local geometry, I_k is the fraction of k nearest target neighbors whose sign type differs from the query, and $F_T = \text{tr}(S_B) / \text{tr}(S_W)$. For form-context density, $D_{m,c} = \text{median}_{i:y_i=c} \log \frac{(\bar{d}_{im}^{\text{vis}} + 10^{-8})^{-1}}{(\bar{d}_i^{\text{ctx}} + 10^{-8})^{-1}}$. For annotation ambiguity, $H_i = -\sum_c p_i(c) \log p_i(c)$ and $\Delta_m^H = F1_m^{\text{low}} H - F1_m^{\text{high}} H$. Covariance metrics are computed after vector normalization, and model-level summaries average over sign types with target support $n \geq 10$. Nearest-neighbor diagnostics use $k = 10$ cosine neighbors. Density ratios add a small numerical constant (10^{-8}) to avoid division by zero. Full CSV outputs retain support counts for every cell and mark $n < 10$ as low support. Frequency uncertainty uses binomial confidence intervals by corpus and sign type. Model comparisons with clip-level predictions use paired bootstrap resampling over evaluated clips. Sign-type-level geometry metrics use bootstrap resampling over supported sign types.

F Form-to-Context by Sign Type

Table 8 decomposes f_{sem} by sign type. This is the main sign-type-level view of form-to-context prediction: it shows whether the model-level cosine in Figure 2 is driven by lexical signs alone or is also present for fingerspelling, classifiers, and neologisms.

Sign Type	Definition
Lexical Sign	An established, conventional dictionary sign.
Fingerspelling	Manual-alphabet sequences used to represent written or spoken forms.
Name Sign	A sign used to refer to a particular individual, place, or entity.
Number Sign	A sign whose primary function is to express a numerical value or quantify a noun.
Compound	A single lexical item formed from two lexical elements that function together as one word.
Gesture	A non-lexical, imagistic communicative movement not treated as an established lexical sign.
Neologism	A newly created sign for a new concept, often not yet fully standardized.
Classifier	A productive sign where handshape and movement schematically represent a referent class, its shape, handling, spatial relation, or enactment.

Classifier Subtypes	
Body	The signer’s body represents and enacts the referent’s action.
Body Part	A specific body part of the signer represents the corresponding body part of the referent.
Descriptive	Classifiers used to describe visible properties of a referent, especially size, shape, or appearance.
Instrument	Classifiers showing how an object is manipulated or used, usually by the hands.
Locative	Classifiers showing where an object is located, and often also how it moves in space.
Plural	Classifiers marking multiple referents, whether the number is exact or non-exact.
Semantic	Classifiers whose handshape stands for a semantic class of referents, such as person or vehicle.

Table 5: Definitions of major sign types targeted in this work, paraphrased from [Neidle et al. \(2022\)](#).

Task	Group	n	α	Mean κ	Agreement
Sign type	Overall	50	.370	.376	.580
Sign type	ASL STEM Wiki	16	.670	.673	.875
Sign type	Atomic Hands	19	.359	.377	.789
Sign type	Lexical sign	28	-.028	.000	.929
Sign type	Fingerspelling	2	–	–	1.000
Boundary	Overall	50	.381	.391	.480
Boundary	ASL STEM Wiki	16	.393	.421	.500
Boundary	Atomic Hands	19	.575	.582	.684
Boundary	Lexical sign	28	.434	.448	.571
Boundary	Fingerspelling	2	–	–	1.000

Table 6: Inter-annotator agreement on duplicate-annotated clips, decomposed by dataset and available sign-type label modes. Dashes mark low-support cells where α or κ is undefined.

G Geometry and Corpus by Sign Type

Table 9 decomposes two RQ3 diagnostics by sign type: ρ^d , the STEM/ASLLRP effective-dimensionality ratio, and corpus AUC, the recoverability of source identity within a sign type. This table is deliberately sign-type-level because model-level averages can hide different behavior for lexical signs, fingerspelling, classifiers, and neologisms.

H Boundary and Annotation-Entropy Diagnostics

Table 10 reports whether ASLLRP-to-STEM f_{type} changes when target clips are grouped by boundary quality and annotation entropy. The diagnostic symbols used in this table and Figure 2 are summarized in Table 11.

I ISR and Phonologizer Details

The ISR representation is produced by a pose transformer trained on isolated ASL videos. Its backbone is a six-layer, eight-head transformer with hidden size 256, feed-forward size 1024, dropout 0.15, and 48-frame inputs. The backbone is initialized from hand-focused self-supervision and trained to predict 16 phonological feature inventories. A lexical decoder is then trained over 2,477 isolated-sign classes using class-balanced cross-entropy, label smoothing 0.1, auxiliary phonological loss weight 0.1, and reconstruction loss weight 0.05. The recorded validation set contains 7,671 videos; top-1 accuracy is 57.1% and top-5 accuracy is 79.0%. The full model has approximately 8.35 million parameters.

The 512-dimensional ISR vector concatenates two learned 256-dimensional projections: one from the concatenated softmax outputs of the 16 phonological heads, and one from the transformer’s class state concatenated with the mean task state. The Phonologizer uses the same 48-frame, six-layer, hidden-256 task-attention pose backbone and returns the class state concatenated with the mean task state (512 dimensions), without the lexical decoder projection.

J Representation Reproducibility

Pose models. ISR and Phonologizer receive 33 body plus 42 hand landmarks in two dimensions. Inputs are uniformly resampled or padded to 48

Model	$f_{\text{type}}: \text{STEM CV}$ F1	Acc.	$f_{\text{type}}: \text{caption held out}$ F1	Acc.	$f_{\text{type}}: \text{signer held out}$ F1	Acc.	$f_{\text{type}}: \text{ASLLRP} \rightarrow \text{STEM}$ F1	Acc.	f_{sem} Cos.
ISR	.158	.522	.166	.540	.132	.507	.161	.502	.699
Phonologizer	.179	.529	.187	.540	.154	.553	.147	.537	.707
CiCo	.156	.634	.183	.665	.137	.599	.119	.337	.489
SignCLIP	.202	.577	.250	.590	.175	.608	.073	.176	.569
ViFi	.170	.546	.223	.526	.112	.478	.142	.476	.447
SHuBERT	.173	.621	.169	.652	.135	.582	.113	.454	.494
VideoMAE	.191	.597	.228	.632	.141	.608	.120	.436	.431

Table 7: M_{type}^m probe results and M_{sem}^m visual-to-English-context cosine ($n = 454$ target clips). f_{type} columns report Macro-F1 except where accuracy is explicitly labeled; f_{sem} reports cosine similarity. Higher is better for all columns in this table.

Model	Lexical n Cos.		Fingerspelling n Cos.		Classifier n Cos.		Neologism n Cos.	
ISR	317	.713	51	.713	54	.665	12	.722
Phonologizer	317	.712	51	.722	54	.669	12	.708
CiCo	317	.470	51	.551	54	.468	12	.449
SignCLIP	317	.575	51	.618	54	.491	12	.529
ViFi	317	.389	51	.367	54	.309	12	.257
SHuBERT	317	.508	51	.485	54	.454	12	.463
VideoMAE	317	.350	51	.358	54	.344	12	.282

Table 8: Mean cosine for f_{sem} by model and sign type. All four columns have target support $n \geq 10$; other sign types are low support and omitted from this compact table.

frames and normalized by torso position and scale. Their paper embeddings are the full 512-dimensional vectors described above, not older PCA-reduced outputs.

released analysis outputs; software and compute details will be added to the final artifact metadata.

Pose and context extraction. MediaPipe pose sequences contain 33 body landmarks and 21 landmarks from each hand per frame; detailed face landmarks and the depth coordinate are omitted. Missing landmarks are zero-filled before model-specific normalization. English context embeddings mean-pool the final bert-base-uncased layer over non-padding tokens, capped at 64 wordpieces.

Video and contrastive models. CiCo produces 1,024-dimensional clip vectors and SignCLIP 768-dimensional clip vectors from their released visual encoders. SHuBERT and VideoMAE produce frame/token sequences; the main analysis mean-pools the final sequence layer, yielding 768 dimensions. ViFi uniformly samples 16 RGB frames and returns a normalized 512-dimensional clip vector.

Common analysis. The representation run matched 149,317 ASLLRP rows and 525 STEM rows before retaining 454 STEM clips with one majority label. Linear probes use identical row manifests for all models. Neighborhood metrics use cosine distance after L2 normalization, exclude each query itself, and use $k = 10$. Random seeds and the exact row manifests are stored with the

Model	Sign type	n_{ASLLRP}	n_{STEM}	$\rho^d(1)$	Corpus AUC ($\downarrow$, chance=.5)
ISR	Lexical	100	317	1.658	.726
ISR	Fingerspelling	86	51	.800	.849
ISR	Classifier	186	54	.521	.694
ISR	Neologism	2083	12	.166	.985
Phonologizer	Lexical	100	317	1.429	.772
Phonologizer	Fingerspelling	86	51	.880	.857
Phonologizer	Classifier	186	54	.597	.673
Phonologizer	Neologism	2083	12	.262	.993
CiCo	Lexical	100	317	1.016	.976
CiCo	Fingerspelling	86	51	.594	.969
CiCo	Classifier	186	54	.790	.941
CiCo	Neologism	2083	12	.280	.955
SignCLIP	Lexical	100	317	.935	.989
SignCLIP	Fingerspelling	86	51	.903	.986
SignCLIP	Classifier	186	54	.831	.926
SignCLIP	Neologism	2083	12	.576	1.000
ViFi	Lexical	100	317	.922	1.000
ViFi	Fingerspelling	86	51	3.082	1.000
ViFi	Classifier	186	54	2.151	1.000
ViFi	Neologism	2083	12	.152	1.000
SHuBERT	Lexical	100	317	2.100	.980
SHuBERT	Fingerspelling	86	51	1.011	.968
SHuBERT	Classifier	186	54	1.252	.937
SHuBERT	Neologism	2083	12	.635	.999
VideoMAE	Lexical	100	317	3.349	1.000
VideoMAE	Fingerspelling	86	51	4.610	1.000
VideoMAE	Classifier	186	54	3.155	1.000
VideoMAE	Neologism	2083	12	1.421	1.000

Table 9: Sign-type decomposition of transfer geometry and corpus separability. $\rho^d = d_{cm}^{STEM}/d_{cm}^{ASLLRP}$; values near 1 indicate similar effective dimensionality across ASLLRP and STEM. Corpus AUC measures recoverability of ASLLRP versus STEM within each sign type.

Model	Clean		Shifted		Invalid/unclear		Summary	
	n	F1	n	F1	n	F1	$\Delta_b(0)$	$\Delta_H(0)$
ISR	123	.186	298	.142	33	.155	.048	-.001
Phonologizer	123	.145	298	.147	33	.354	-.004	.004
CiCo	123	.173	298	.079	33	.097	.096	.005
SignCLIP	123	.101	298	.099	33	.220	-.001	.004
ViFi	123	.146	298	.134	33	.221	.012	.006
SHuBERT	123	.123	298	.107	33	.182	.014	.013
VideoMAE	123	.139	298	.082	33	.093	.058	.011

Table 10: ASLLRP-to-STEM f_{type} by boundary condition and annotation-entropy summaries. Δ_b is clean Macro-F1 minus nonclean Macro-F1; Δ_H is low-entropy Macro-F1 minus high-entropy Macro-F1. Values near 0 indicate lower sensitivity.

Symbol	Name	Preferred value
I_k	Target kNN impurity	$\downarrow$
F_T	Target Fisher separation	$\uparrow$
ρ^d	STEM/ASLLRP effective dimensionality	1
D_m	Visual-to-context density ratio	0
Δ_b	Boundary-condition F1 gap	0
Δ_H	Annotation-entropy F1 gap	0

Table 11: Symbols used for the descriptive diagnostics.