Surveying Representation Collapse in Pretrained Models of American Sign Language

Anonymous ACL submission

Abstract

The American Sign Language (ASL) lexicon contains heterogeneous sign types, including not only established signed words but also 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, domains where lexical innovation may be more likely to occur. Our analyses provide evidence that (a) pretrained models in general lose a substantial amount of articulatory signal when encoding rarer and more compositional signs, (b) phonology- and gesture-based pretraining can be an effective regularizer for transfer learning, and (c) contrastive alignment with English word embeddings may worsen model generalizability on visually-oriented signs. Future work may seek to improve models' ability to understand and produce these sign types as they appear in STEM or other domains.

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 such computational models can accommodate the linguistic variation found in signed utterances (Bragg et al., 2019; Desai et al., 2024). The American Sign Language (ASL) lexicon is typologically diverse, including not only established signed words, but also *depicting signs*, which reproduce a referent's visual properties; *neologisms* for domain-specific or new concepts, and *fingerspelling*, which generally involves the use of manual letters to express English words (Neidle et al., 2022).

Although these *sign types* collectively appear often in ASL, most individual productions in the latter three sets will be *out-of-distribution* (OOD) for pretrained ASL models, owing to the chronic scarcity of ASL data and high amounts of linguistic variation. 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 is dynamic with respect idiolect, dialect, register, audience, and other factors. These factors of variation drastically increase the sparsity and reduce the frequency of entire sign types, while established signed words are more frequent and less variable in form. *How effectively do pretrained models of ASL induce the generalizable receptive skills necessary to understand OOD productions?*

To examine the potential for systematic *representation collapse* in pretrained ASL models, this work evaluates their 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 significantly more OOD terms as compared to open-ended narratives, likely due to concepts that are naturally suited to depiction[1] and the absence of established terms for specific concepts (Sortino et al., 2026; Lualdi et al., 2023). Separately, progress on modeling STEM content could support educational access for students whose primary language is ASL. Oftentimes, DHH students are often expected to learn complex STEM topics without a consistent lexicon, and may benefit from applications like ASL-based search engines.

[1] For instance, the shape of a curve in a graph, how molecules move in states of matter, or the structure of DNA are easily expressed with depiction.

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 signs for annotation and as a predictor of representation collapse. In §5.1, we compare the relative effectiveness of input signals from external movement, internal movement, and pretrained model perplexity on OOD transfer). We additionally explore the extent to which segmentation error predicts representation collapse 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 §4.2, we probe seven pretrained models on the task of sign type classification and measure performance in terms of transferability to out-of-domain STEM data, form-to-context prediction, and detailed collapse diagnostics (§5.2).

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

The etiology 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. These diagnostics show persistent model differences alongside strong corpus identity in several representation spaces (§5.2).

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.[2]

[2]To preserve anonymity, we will provide the URL to download annotations if accepted.

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 lexicography 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 Padden, 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).

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 (*cf.* "inside" OOD vs "outside" OOD; Lazebnik 2026).

2.2 ASL Modeling and Representation Collapse

Methods for modeling ASL have responded to chronic data scarcity by including self-supervision (Gueuwou et al., 2025), contrastive learning (Jiang et al., 2024), and phonological inductive priors (Kezar et al., 2023a) to increase data efficiency. These methods are trained on ASL datasets which may augmented with pose estimation (Jiang et al., 2025), English word embeddings (Cheng et al., 2023), or gesture (Tong et al., 2022) to encourage more generalizable features of language. As suggested in RQ2, these choices have profound influences on the geometry of the learned representation space, however these influences are poorly 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 (Papyan et al., 2020). In contrastive learning, related geometric pressures are often framed through alignment and uniformity: positive pairs are pulled together, embeddings are spread across the hypersphere, and failures of this balance can concentrate variance into fewer active directions (Wang and Isola, 2020; Jing et al., 2022). Shortcut learning provides a complementary account of why such representations can transfer poorly: models may rely on predictive regularities that are stable in a benchmark distribution and fragile under new test conditions (Geirhos et al., 2020).

For OOD transfer, these accounts point to a shared failure mode in which pretraining preserves high-frequency separations while compressing lower-frequency distinctions into weak directions, sparse neighborhoods, or clusters organized by repeated training-set regularities. In ASL, this can mean that established lexical signs, signer identity, recording conditions, or dataset source dominate the learned geometry (RQ3; Pal et al. 2023). Novelty of the kinds described in §2.1 should expose which pretrained spaces preserve productive linguistic structure and which compress it into training-set regularities.

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
ASL Core	OOD Target, isol.	2,443	26

Table 1: Source dataset roles, sizes, and number of signers. *cont.* = continuous; *isol.* = isolated sign data.

3 Data Collection and Preprocessing

To study representation collapse under domain shift, we collect and preprocess ASL signing data from multiple sources and domains. The data pipeline has three parts. First, we use reference and isolated sign resources to support tokenizer development, sign type sampling, and reference-domain comparison (§3.1). Second, we convert continuous videos into pose sequences, candidate sign clips, visual embeddings, and sentence-context embeddings (§3.2). Third, ASL signers annotate STEM clips for boundary quality and sign type, producing the target labels used in the main analyses (§3.3).

3.1 Data Sources

Table 1 summarizes the six corpora used in this study, including their role, segment counts, and signer coverage. The American Sign Language Linguistic Research Project (ASLLRP) is our primary continuous reference ASL source (Neidle et al., 2022), used for tokenizer evaluation and reference-to-target transfer. ASL Citizen (Desai et al., 2023) and Sem-Lex (Kezar et al., 2023b) provide isolated sign data used to train and evaluate an isolated sign recognition model. The selected model is a 6-layer transformer pretrained on handshape reconstruction and phonological feature prediction (see full details in Appendix B). After continuous videos are segmented, these recognizers also help identify candidate clips that are dissimilar to established signed words via recognition uncertainty.

Our target-domain evaluation uses three STEM corpora. ASL STEM Wiki contains ASL interpretations of Wikipedia articles related to STEM (Yin et al., 2024). Atomic Hands is a YouTube channel containing STEM educational content produced in ASL and later translated into English, used with permission from the creators. These sources provide the target STEM clips used for annotation, probing, and transfer evaluation.

3.2 Preprocessing and Representation Extraction

Each video is converted into a sequence of 2D pose estimates using MediaPipe, with body, face, and hand keypoints ($n = 75$) per frame. Continuous videos are converted into candidate sign clips using the tokenizer described in §4.1. Segments shorter than 0.35 seconds are excluded before representation extraction.

Each candidate clip is embedded by the pretrained visual models listed in Table 2. We also encode the English sentence or caption associated with each segment using BERT, producing a sentence-context embedding used for form-to-meaning probing. All downstream analyses pair each clip-level visual embedding with its segment metadata, including dataset, video, signer when available, and sentence-context embedding.

3.3 Annotation Workflow

Four ASL signers evaluated segmentation quality and labeled sign type for target STEM clips. The annotator pool included two Deaf annotators whose primary language is ASL and two hearing annotators who can converse in ASL. Their average ASL experience is 20.25 years with a minimum of 10 years and a maximum of 34.

Each annotator saw a mix of segments sampled either randomly or weighted by the mean perplexity of models trained to recognize signs. Each annotation pack contained 150 clips: 50 shared random clips, 50 shared high-perplexity clips, and 50 unique high-perplexity clips. The shared clips support agreement analysis, random sampling supports prevalence estimation, and high-perplexity sampling increases coverage of rare and productive sign types. High-perplexity clips are selected using the geometric mean of perplexity scores across pretrained models.

For each clip, annotators selected start-boundary and end-boundary judgments, including early, good, or late start combined with early, good, and late end. They also marked clips that contained multiple signs, were too short to annotate, or were otherwise invalid or unclear. They additionally annotated the sign type as lexical sign, fingerspelling, loansign, name sign, number sign, compound, gesture, neologism, and classifier (+ 7 classifier subtypes; see Appendix A for definitions of these types).

4 Method

Given the in- and out-of-distribution ASL clips, we evaluate whether pretrained visual representations preserve their structure across three levels of analysis. The tokenizer analysis asks whether continuous signing can be divided into usable sign clips from pose, and how that skill generalizes to target data (§4.1). The probe analysis asks whether those clips preserve sign type and sentence-context information in each embedding space (§4.2). The collapse and data-fidelity analyses ask whether changes in class geometry are attributable to transfer, type annotation ambiguity, segmentation quality, or target-corpus composition (§4.3; §4.4).

4.1 Tokenizer Evaluation

Given a pose sequence $X = (x_1, \ldots, x_T)$, tokenization predicts an ordered boundary set $B = \{b_1, \ldots, b_K\}$ with $1 < b_1 < \cdots < b_K < T$. These boundaries define candidate sign clips $X_{b_j:b_{j+1}}$. Tokenizer parameters are tuned on ASLLRP reference boundaries and then applied to STEM target videos.

The proposed tokenizer follows the Movement-Hold model, briefly summarized in §2.1. Because holds correspond to relatively stable articulatory states, local reductions in hand velocity provide a useful cue for candidate sign boundaries. For each frame t, we compute external velocity from wrist keypoints and internal velocity from fingertip motion relative to the wrist. Let $m_{\text{ext}}(t)$ denote external wrist velocity and $m_{\text{int}}(t)$ denote internal finger velocity. The combined velocity signal is

$$V(t) = m_{\text{ext}}(t) + m_{\text{int}}(t). \tag{1}$$

We smooth this signal to obtain $\tilde{V}(t)$.

Candidate boundaries are placed at local minima of the smoothed velocity profile within low-motion intervals. For a candidate low-motion interval $H = [t_s, t_e]$, the selected boundary is

$$b = \arg\min_{t \in [t_s, t_e]} \tilde{V}(t). \tag{2}$$

The implementation uses locally adaptive velocity thresholds and minimum-duration constraints tuned on the reference boundary set.

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. For the uncertainty-

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-CLIP (Rasheed et al., 2022)	CL + ISR	Video–text	512	Gesture, text (En.)
VideoMAE (Tong et al., 2022)	SSL + recon.	Video TF	768	Gesture
ISR Baseline (this work)	ISR + phon. + recon.	Pose TF	512	Isol. signs (ASL)

Table 2: Pretrained representation models used for clip-level embedding extraction. Objective key: CL = contrastive learning; SSL = self-supervised learning; ISR = isolated sign recognition; phon. = phonological supervision; recon. = reconstruction. Architecture key: TF = transformer. Data key: Isol. = isolated; Cont. = continuous; multi = multilingual sign language data; En. = English. Dim. = embedding dimensionality.

augmented tokenizer, each candidate split frame receives a score

$$S(t) = \lambda P_b(t) + (1 - \lambda)U(t), \tag{3}$$

where $P_b(t)$ is the boundary score learned from reference boundary data and $U(t)$ is isolated sign recognizer uncertainty.

Tokenizer quality is evaluated using manually reviewed boundary judgments. We report boundary outcomes by tokenizer, dataset, and sign type in §5.1. We also condition downstream sign-type prediction on boundary quality to assess how strongly segmentation errors affect transfer performance in §5.2.

4.2 Representation Probes

For each segment i and model m, let $z_{im} \in \mathbb{R}^{d_m}$ denote the visual embedding. We first probe whether target STEM embeddings allow linear separation of sign types. A linear classifier g_m is trained and evaluated on target STEM segments:

$$F_m^{\text{target}} = \text{MacroF1}\big(g_m(z_{im}^{\text{target}}), y_i^{\text{target}}\big). \tag{4}$$

We then test whether sign type structure learned from reference ASL transfers to STEM signing. For this transfer setting, the classifier is trained on a reference set balanced across sign types and evaluated on target STEM segments:

$$F_m^{\text{transfer}} = \text{MacroF1}\big(g_m^{\text{ref}}(z_{im}^{\text{target}}), y_i^{\text{target}}\big). \tag{5}$$

To test whether visual form preserves sentence-context information, we evaluate form-to-context prediction. For each segment i, let s_i be the BERT embedding of its English sentence context. For model m, we train a linear map A_m from visual embedding space to sentence-context space by minimizing

$$\mathcal{L}_m = \sum_i \|A_m z_{im} - s_i\|_2^2. \tag{6}$$

We report cosine similarity between $A_m z_{im}$ and s_i, and compare visual and sentence-context neighborhood density as described in §4.3.

4.3 Collapse Metrics

To measure how representation spaces preserve or collapse articulatory information under transfer, we define metrics based on class separability, within-class dimensionality, and source-domain separability. These metrics complement the probe scores by directly measuring the geometry of the embedding space.

From the perspective of between-class separation, collapse corresponds to weaker separation between sign type centroids relative to within-class spread. For model m and domain D, let μ_{cm}^D be the centroid for class c, μ_m^D the global centroid, and z_{im}^D an embedding from domain D. We define

$$\Phi_m^D = \frac{\sum_c n_c \|\mu_{cm}^D - \mu_m^D\|_2^2}{\sum_c \sum_{i:y_i=c} \|z_{im}^D - \mu_{cm}^D\|_2^2}. \tag{7}$$

The target-to-reference Fisher ratio is

$$\Gamma_m = \frac{\Phi_m^{\text{target}}}{\Phi_m^{\text{ref}}}. \tag{8}$$

Values below one indicate weaker class separation in STEM than in reference ASL, and values above one indicate stronger class separation in STEM.

From the perspective of embedding expressivity, collapse corresponds to target sign types occupying a lower-dimensional subspace than their reference counterparts. For class c and model m, we estimate effective dimensionality d_{cm}^D from the eigenvalue spectrum of the class covariance matrix. The target-to-reference dimensionality ratio is

$$\rho_{cm}^d = \frac{d_{cm}^{\text{target}}}{d_{cm}^{\text{ref}}}. \tag{9}$$

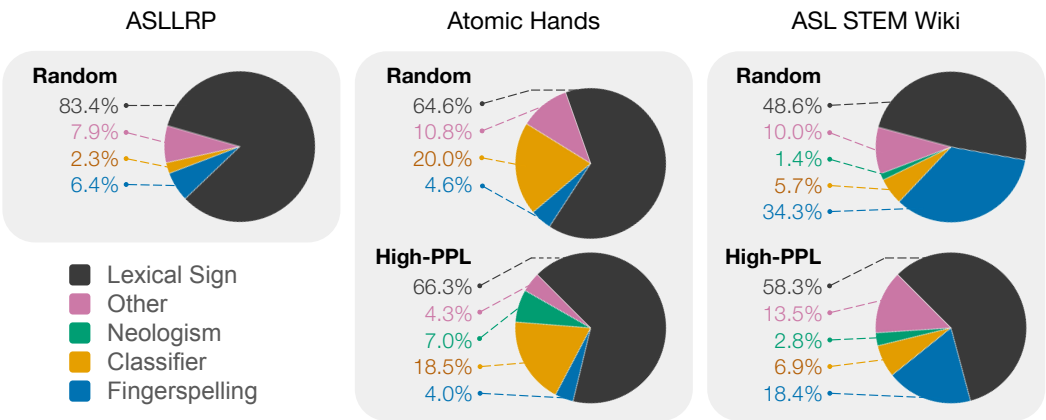

Figure 1: Prevalence of major sign type annotations by dataset and sampling strategy (random vs. high perplexity or PPL). As compared to ASLLRP, Atomic Hands and ASL STEM Wiki contain more classifiers (+17%) and fingerspelling (+28%), respectively.

Values below one indicate compression of the target class manifold, while values above one indicate expansion.

We additionally test whether each embedding space separates corpus identity by training a binary classifier q_m to distinguish reference from target embeddings:

$$AUC_m^{\text{domain}} = \text{AUC}\big(q_m(z_{im}), D_i\big). \tag{10}$$

High domain AUC indicates that the representation strongly separates reference and target data. This metric helps distinguish transferable sign type structure from corpus-specific structure.

We also compute target kNN impurity I_m^{tgt}, the proportion of each target segment's k nearest target-domain neighbors whose sign-type label differs from its own. Higher impurity indicates more locally mixed sign-type neighborhoods.

To compare visual and sentence-context neighborhood structure, we compute a form–meaning density ratio ρ_m^{VS} from local target-domain neighborhoods in visual embedding space and BERT sentence-context space. Values near zero indicate similar local density across the two spaces, while larger absolute values indicate stronger mismatch between visual and sentence-context neighborhoods.

4.4 Data Fidelity Analyses

The main collapse metrics can be affected by label ambiguity, boundary quality, and differences in target-corpus composition. We therefore estimate each factor directly and relate it to probe performance.

For annotation ambiguity, we compute per-segment sign type entropy from overlapping annotations. We then compare model performance on low- and high-entropy segments:

$$\Delta_m^H = F1_m^{\text{low } H} - F1_m^{\text{high } H}. \tag{11}$$

A large gap indicates that performance is concentrated on segments with clearer human agreement.

For boundary quality, we report F1 disaggregated by segmentation error (good, start *or* end error, start *and* end error, invalid). The segmentation-error effect size is the difference between clean-boundary performance and error-condition performance. This measures the extent to which boundary quality influences downstream sign type transfer.

For target-corpus composition, we estimate sign type prevalence separately for each STEM dataset:

$$\pi_{cD} = \frac{n_{cD}}{\sum_{c'} n_{c'D}}. \tag{12}$$

We compare the usage of sign types between interpretations and ASL-first explanations of STEM concepts. These estimates describe the target distributions over which the representation models are evaluated.

5 Results

We report annotation quality and sign-type prevalence first (§5.1), followed by representation-level collapse results and model-specific synthesis (§5.2).

Condition	n	Good	1 err.	2 err.	Inv.
Ext. velocity	20	0.15	0.30	0.00	0.55
Int. velocity	20	0.15	0.10	**0.30**	0.45
Ext. + int. velocity	20	**0.23**	**0.38**	0.08	0.33
Velocity + ISR PPL	20	0.20	0.35	0.15	**0.30**
External + internal velocity breakdown					
ASL STEM Wiki	268	**0.24**	0.40	0.22	0.14
Atomic Hands	182	0.17	**0.41**	**0.23**	**0.19**
lexical	285	**0.27**	0.37	0.24	0.12
fingerspelling	53	0.02	**0.55**	**0.36**	0.08
classifier	67	0.21	0.46	0.16	**0.16**
neologism	23	0.13	0.61	0.13	0.13

Table 3: Boundary-quality outcomes for reviewed tokenizer candidates. "1 err." denotes either the start or end boundary was erroneous; "2 err." denotes both were erroneous; "Inv." (invalid) denotes multiple signs or very brief segments.

5.1 Annotation Analysis

Table 3 reports boundary-quality outcomes for the tokenizer ablation and selected external+internal velocity tokenizer. Segmentation quality is strongly sign-type dependent: lexical signs are most often cleanly bounded, while fingerspelled segments rarely receive clean start and end boundaries. This is consistent with fingerspelling creating a harder segmentation target because one signed word can contain multiple short holds and transitions. ASL STEM Wiki also yields cleaner boundaries than Atomic Hands.

Inter-annotator agreement confirms that boundary quality and sign type are difficult judgments in this setting. For shared clips, boundary-error labels have Krippendorff's $\alpha = 0.381$, mean pairwise $\kappa = 0.391$, and mean pairwise agreement of 0.480. Sign-type labels have Krippendorff's $\alpha = 0.370$, mean pairwise $\kappa = 0.376$, and mean pairwise agreement of 0.580. Sign-type agreement is higher for ASL STEM Wiki than Atomic Hands ($\alpha = 0.670$ vs. 0.359), while boundary-error agreement is higher for Atomic Hands ($\alpha = 0.575$ vs. 0.393). For segments which had a majority winner, their prevalence estimates are shown in Figure 1.

5.2 Representation Collapse and Diagnostics

Figure 2 summarizes OOD sign-type separability, OOD effective dimensionality, form–meaning cosine alignment, and form–meaning density ratio. The main pattern is a tradeoff between transfer stability and target-domain restructuring. ISR and Phonologizer retain the strongest OOD sign-type structure, while VideoMAE expands sharply in target space and shows the largest form–meaning density shift. This suggests that broad visual pretraining can preserve target-domain distinctions in a geometry that transfers weakly from reference ASL.

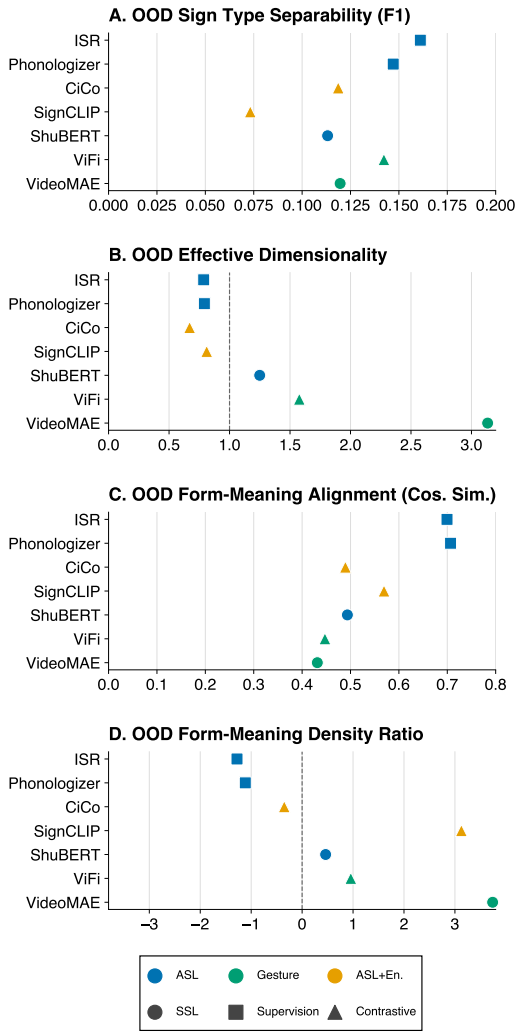

Figure 2: Representation collapse metrics in terms of: (A) linear separability of sign types in STEM data; (B) loss or gain of dimensionality upon transfer; (C) similarity of visual and semantic embeddings; and (D) loss or gain of visual expressivity upon transfer.

Domain identity is easiest to recover from ViFi and VideoMAE embeddings (AUC ≈ 1.000) and hardest from ISR (AUC $= 0.813$) and Phonologizer (AUC $= 0.823$). Target neighborhoods are "most sign-type coherent for Phonologizer ($I_m^{\text{tgt}} = 0.367$) and least coherent for CiCo ($I_m^{\text{tgt}} = 0.440$); target Fisher separation is strongest for ISR ($\Phi_m^{\text{target}} = 0.190$) and Phonologizer ($\Phi_m^{\text{target}} = 0.175$). Segmentation-error effects are largest for CiCo (0.096) and VideoMAE (0.058), and closest to zero for SignCLIP (-0.001) and Phonolo-

gizer (-0.004). Annotation-entropy gaps are small across models, from -0.001 for ISR to 0.013 for ShuBERT, suggesting that label ambiguity is secondary to transfer and geometry differences.

6 Discussion

We return to our central question: *How effectively do pretrained models of ASL induce the generalizable receptive skills necessary to understand OOD productions?* Across metrics, collapse appears as a redistribution of structure: some spaces preserve in-domain sign type separability, others preserve transfer, and others preserve visual variation broadly while strongly encoding domain identity.

RQ1: Tokenizer Generalizability. The tokenizer results show that boundary quality depends more on articulatory velocity than on recognizer uncertainty. The combined-velocity tokenizer produced the highest good-boundary rate, while adding ISR perplexity did not improve the reviewed boundary sample. This supports the premise that hold-based cues are potentially useful for OOD segmentation. Segmentation quality affects downstream probes, but the observed effects are smaller than the main cross-model differences.

RQ2: Collapse across modeling choices. ASL-internal supervision gives the clearest transfer advantage: ISR has the highest OOD F1, and Phonologizer remains close behind while matching ISR on the strongest form-to-context cosine. This supports lexical and phonological priors as partial protection against collapse, while also showing their limits: the phonological label set underspecifies cues such as facial expression and richer external movement structure (Caselli et al., 2017).

English-aligned pretraining shows the opposite profile. SignCLIP has the highest ID F1 but the lowest OOD F1, while CiCo has the highest individual-model ID accuracy but the lowest ID F1. These splits are consistent with text-aligned spaces organizing target clips around frequent or easily lexicalized categories, while sign types such as classifiers, name signs, and neologisms lack stable word-level destinations in English embedding space.

Gesture and video pretraining separate sign types by preserving broad visual variation. ViFi remains competitive on OOD F1, and VideoMAE has strong target-domain F1. However, these spaces also separate much more than sign type: ViFi and VideoMAE reach near-ceiling domain AUC, and VideoMAE shows the largest effective-dimensionality and visual-to-semantic density shifts. That breadth comes at a cost in the form-to-meaning probe, where the figure shows weak alignment between visual structure and sentence-context structure. Self-supervised objectives fit this pattern: they retain visual variation without English labels or isolated sign classes, but they do not determine which variation supports linguistic comprehension.

RQ3: Collapse versus evaluation artifacts. The collapse diagnostics help separate variation in probe scores from variation in embedding structure and corpus identity. VideoMAE changes the most under target-domain transfer, with target classes occupying a much larger space and becoming less aligned with the sentence-context neighborhood structure. ISR shows the opposite profile: its embeddings make corpus identity harder to recover and keep target sign types more separated from one another. Segmentation-error effects are largest for CiCo and VideoMAE, while Phonologizer is closest to zero. Annotation-entropy gaps are small across models, with the largest gap for ShuBERT. Boundary quality and label ambiguity remain important controls, but they do not explain the full spread in ID F1, OOD F1, domain AUC, dimensionality, or Fisher separation.

Conclusion. Taken together, our results suggest that no single representation space captures all of the structure needed for productive ASL under STEM-domain transfer. In these experiments, ASL-specific lexical and phonological models transfer more reliably, English-aligned models perform better on some target-domain probes than on transfer, and gesture-based models preserve broad visual variation while also making domain identity highly recoverable. These patterns are consistent with representation collapse as a shift in which articulatory, sign type, and contextual distinctions remain available after pretraining, rather than as a uniform loss of information.

Future work should test whether richer phonological labels, signer-robust pretraining, and objectives that preserve productive structure can improve transfer without relying too heavily on familiar lexical patterns or corpus-specific cues. Success on these tasks could lead to more externally valid ASL models, offering increased access to information and, in particular, STEM-based educational content.

6.1 Limitations

The main limitation is that the corpora differ in domain, genre, signer population, recording conditions, 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 a Deaf signer with ASL as their primary language, the number of annotators ($n = 3$) 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 pretraining data, architectures, objectives, modalities, dimensionality, and language coverage. The observed family-level patterns should therefore be treated as converging evidence rather than controlled ablations. A causal account would require pretraining experiments that vary one design choice at a time, which is beyond the scope of the current work.

A Sign Types

Table 4 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 Isolated Sign Recognition (ISR) Transformer

The ISR baseline is an ASL-specific isolated sign recognizer built to provide a pose-only reference point for the representation analyses. Unlike the contrastive and video–language models in the main comparison, it is trained entirely within ASL and receives no English supervision. Its role in the paper is therefore to test how much target-domain structure is preserved by a model optimized for isolated lexical recognition with explicit phonological support.

The model is trained through a three-stage ASL-only pipeline. First, a transformer pose encoder is pretrained on unlabeled isolated sign clips using masked/reconstruction-style objectives focused on hand and pose structure. Second, the same encoder is refined through multitask phonological prediction over the isolated sign inventory. Finally, a lexical decoder is trained on top of the phonology backbone to predict isolated sign labels. The resulting checkpoint combines lexical supervision, phonological supervision, and hand/pose reconstruction, rather than relying on a single classification objective.

The architecture uses a transformer-based pose encoder followed by a phonology-conditioned lexical decoder. The input is a 48-frame pose sequence extracted from isolated ASL clips. The configuration uses 6 transformer layers, hidden size 256, 8 attention heads, feed-forward size 1024, and approximately 8.35M parameters.

For the analyses in this paper, we use the model's full 512-dimensional embedding. This embedding fuses phonological posterior information with latent transformer state information, giving the downstream probes access to both predicted linguistic features and learned pose dynamics. For uncertainty-based sampling, we use the entropy of the lexical softmax distribution, so high-entropy clips are those the ISR model treats as less compatible with its isolated-sign vocabulary.

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 4: Definitions of major sign types targeted in this work, paraphrased from Neidle et al. (2022).

References

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.

Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2017. ASL-LEX: A lexical database of american sign language. *Behavior Research Methods*, 49(2):784–801.

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.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

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.

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. SignCLIP: 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)*.

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*.

Teddy Lazebnik. 2026. Introducing "inside" out of distribution. *International Journal of Data Science and Analytics*, 21(43).

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*.

Abhilash Pal, Simon Huber, Cyrine Chaabani, Alessandro Manzotti, and Oscar Koller. 2023. On the importance of signer overlap for sign language detection. *arXiv preprint arXiv:2303.10782*.

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 pretraining. 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 Maugard, and Graham Neubig. 2021. Including sign 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 4038–4060.

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.