ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition

Aashaka Desai[β] **Lauren Berger**[γ] **Fyodor O. Minakov**[α]
Vanessa Milan[α] **Chinmay Singh**[α] **Kriston Pumphrey**[γ] **Richard E. Ladner**[β]
Hal Daumé III[α,δ] **Alex X. Lu**[α] **Naomi Caselli**[γ] **Danielle Bragg**[α]

[α]Microsoft Research [β]University of Washington [γ]Boston University [δ]University of Maryland
{hal3,lualex,dabragg}@microsoft.com, {aashakad,ladner}@cs.washington.edu, nkc@bu.edu

Abstract

Sign languages are used as a primary language by approximately 70 million D/deaf people world-wide. However, most communication technologies operate in spoken and written languages, creating inequities in access. To help tackle this problem, we release ASL Citizen, the largest Isolated Sign Language Recognition (ISLR) dataset to date, collected with consent and containing 83,912 videos for 2,731 distinct signs filmed by 52 signers in a variety of environments. We propose that this dataset be used for sign language dictionary retrieval for American Sign Language (ASL), where a user demonstrates a sign to their own webcam with the aim of retrieving matching signs from a dictionary. We show that training supervised machine learning classifiers with our dataset greatly advances the state-of-the-art on metrics relevant for dictionary retrieval, achieving, for instance, 62% accuracy and a recall-at-10 of 90%, evaluated entirely on videos of users who are not present in the training or validation sets.

1 Introduction

Communication in sign language is an essential part of many people's lives. As 70 million deaf people world-wide primarily use a sign language (WFD, 2022b), the meaningful inclusion of signing deaf people requires widespread access to sign languages both in individual communities and society-wide (§2.1). Towards this, over 100,000 students per year enroll in American Sign Language (ASL) classes (Looney and Lusin, 2019), and the number of UN countries mandating provision of services in sign language has grown from 4 in 1995 to 71 in 2021. (WFD, 2022a).

Despite equal access for signing deaf people becoming an increasingly championed value, most existing information resources (like search engines, news sites, or social media) are written or spoken, and do not offer equitable access. Requiring signing deaf people to navigate information sources in a written language like English necessarily means forcing them to operate in a different, and potentially non-native language. Many tools built around online information sources assume written (or, in some cases, spoken) language input and output, and adapting such tools to sign languages requires a fundamental shift to a visual modality. This has given rise to significant technical challenges, which have motivated the development of computational methods, from sign language recognition to generation and translation (Joksimoski et al., 2022).

In this work, we focus on dictionary retrieval for sign languages. Many existing ASL dictionaries catalog signs with English glosses: an out-of-context translation of a sign into one or more English words (e.g., LOBSTER, CLIMB_LADDER or HAD_ENOUGH). However, English lookup relies on knowing the English translation, and on signs having a 1:1 relationship with English words, which is often not the case. Retrieving dictionary entries through a demonstrated sign may be more natural, but is

Preprint. Under review.

computationally more challenging because of the rich visual format and linguistic complexities. We seek to help address this problem of video-based dictionary retrieval, where a person demonstrates a single sign by video, and the system returns a ranked list of similar signs. Dictionary retrieval fills a practical need for sign language learners, who may see a sign but not know the meaning, and so cannot look it up using a written translation. Moreover, dictionaries can contribute to documenting sign languages, and allow established signers to navigate dictionary resources directly in sign language.

To help advance dictionary retrieval, we collected and release a dataset of isolated ASL signs. Our dataset is intended to support data-driven machine learning methods by overcoming limitations of prior isolated sign language recognition (ISLR) datasets (see Table 1 and §2.2). Machine learning-based development typically requires a large training dataset with appropriate properties (large vocabulary, minimal label noise, and representation of diverse signers and environments). Existing video sign datasets are often filmed in lab settings or scraped from online sources, both of which limit scale and diversity. Alternatively, datasets constructed by scraping web sources do not typically acquire participant consent, which erodes community trust, and also lead to challenges in labeling, as sign languages do not have a standardized annotation system. To overcome such limitations, we build upon recent crowdsourcing proposals (Bragg et al., 2022) to collect and release *ASL Citizen* – the first large-scale crowdsourced sign language dataset. Our dataset is the largest isolated sign dataset to date, newly representative of real-world settings and signer diversity, and collected with permission and transparency.

Using this new dataset, we adapt previous approaches to ISLR (Li et al., 2020; Selvaraj et al., 2021) (§2.3) to the dictionary retrieval task, and release a set of baselines for machine learning researchers to build upon (§4). Our dictionary retrieval problem requires algorithms to return a ranked list of signs, given an input video. In principle, this output can be satisfied by a variety of methods, but we focus on supervised deep learning methods, taking advantage of recent methods for ISLR. We show that even without algorithmic advances, training and testing on our dataset doubles ISLR accuracy compared to previous work, despite spanning a larger vocabulary and using a test set comprised of completely unseen users (§5). We additionally evaluate our dataset against prior datasets by comparing performance on a subset of overlapping glosses, and by comparing performance of learned feature representations from models trained on these datasets, showing further improvements in each case. Finally, through a series of downsampled training set experiments, we show that while dataset size contributes to our improved performance, it is not the only contributing factor.

Throughout this research, we have endeavored to follow a culturally-sensitive and participatory approach to sign language computation. Sign languages are a cornerstone of Deaf culture and identity.[1] In response to growing efforts in sign language computation, some previous works have noted that many of these efforts promote misconceptions or inaccuracies about sign language, exploit sign language as a commodity, and undermine political movements from Deaf communities seeking recognition of sign languages (Bragg et al., 2019; De Meulder, 2021; Yin et al., 2021). Other works question if the technologies being designed actually benefit Deaf communities, and document patterns where technologies are rejected by signing communities for being intrusive, clunky, or insensitive (Harris et al., 2009; Kusters et al., 2017). In this work, we ground our approach in calls issued by disability scholars for better collaboration with Deaf communities, and problem settings that focus on solving real needs (Bragg et al., 2019, 2021; Harris et al., 2009; Yin et al., 2021). In addition to maintaining trust with the communities that sign language technologies are intended to serve, we demonstrate that aligning with these calls improves data collection and problem definition.

In summary, our primary contributions are:

1. We provide a benchmark dataset and metrics for the dictionary retrieval task. Not only does this application have real utility to the signing community, but it grounds ISLR in a real-world problem setting, informing data collection and metrics.

2. We release the largest public dataset of isolated signs to date. Our dataset contains high-quality videos in real-world settings, and was collected and shared with appropriate permissions from contributors.

[1]By established standards, we use "Deaf" to refer to cultural identity, and "deaf" to audiological status.

3. We improve over state-of-the-art ISLR accuracy by more than double, offering improved baselines and code for the community to build upon. We also highlight the impact of data on model performance, and release our code.

For links to the dataset, code, and additional supplementary materials please visit [link TBA].

2 Background and Related Work

2.1 Sign Languages and Deaf Culture

Sign languages are complex languages with large vocabularies, governed by their own phonological rules. Analogous to the sounds of speech, signs are composed of largely discrete elements (e.g., handshape, location, and movement) according to a complex set of rules (Brentari et al., 1998). English translations of isolated signs are called glosses, are written in all-caps, and may be single words or multiple words, typically not corresponding 1-1 to English (just as in any other language translation). Sign execution varies across contexts, signers, and sociolinguistic groups (McCaskill et al., 2011). These factors complicate representative data collection and modeling.

Isolated signs like those generally included in sign language dictionaries, are what some have referred to as "core" parts of the lexicon, and are only a subset of sign languages (Brentari and Padden, 2001). The "non-core" lexicon is generally not well represented in dictionaries or lexical databases, and includes complex constructions like depicting verbs, classifier constructions, and verbs that use time and space in ways that can be difficult to decompose into discrete parts (Fischer and Van der Hulst, 2003; Zwitserlood, 2012). In continuous signing, like in spoken language, coarticulation – the impact of preceding and following signs – affects sign production. Continuous signing also includes grammar that is often expressed with the face, body, and signing space, in addition to the hands (Wilbur, 2013). As such, ISLR only address a fraction of sign language recognition. However, since our goal is dictionary retrieval, this work focuses on isolated signs.

Sign languages also play a critical cultural role in Deaf communities and identity (Hands and Voices, 2022). While our work focuses on American Sign Language (ASL), which is primarily used in North America, over 300 sign languages are used worldwide. Sign languages have been suppressed by political and educational authorities to force deaf individuals to integrate into hearing society by favoring speech and speechreading at the expense of individual welfare (Lane, 1989). These oppressive movements promote many misconceptions that persist today (e.g., that sign languages are lesser languages, or that ASL is signed English), and Deaf activists work to combat these ideas (De Meulder, 2021; Harris et al., 2009; Yin et al., 2021). This cultural context informs our decision to formulate ISLR as a dictionary retrieval problem, which grounds research in a meaningful real-world use case.

2.2 Previous ISLR Datasets

Our work focuses on ASL, which has four main public ISLR datasets: WLASL (Li et al., 2020), Purdue RVL-SLL(Wilbur and Kak, 2006), BOSTON-ASLLVD (Athitsos et al., 2008) and RWTH BOSTON-50 (Zahedi et al., 2005), summarized in Table 1. WLASL offers four different vocabulary sizes, the largest containing 2,000 signs (WLASL-2000 in our tables). While BOSTON-ASLLVD contains a larger vocabulary of 2,742 signs, the number of videos per sign is limited. As discussed above, real-world signs vary greatly by user and across demographics due to dialectal (e.g., geographic region) and sociolectal (e.g., age, gender, identity) variation. Models trained on a small number of dataset contributors, as seen in prior work, may not generalize well to diverse signers (Athitsos et al., 2008; Wilbur and Kak, 2006; Zahedi et al., 2005).

Existing datasets employ a variety of collection and labelling techniques, with varied implications for quality and size. Lab-collected data (Athitsos et al., 2008; Wilbur and Kak, 2006; Zahedi et al., 2005) is typically high-quality with clean labels, but limited in size, participant diversity, and devoid of real-world settings. Datasets scraped from the internet may capture more users or environmental diversity, but varied contributor fluency and difficulty identifying and segmenting signing in videos impacts quality. Labels for scraped data are often unreliable. To minimize variability in labels, prior datasets (Joze and Koller, 2018; Li et al., 2020) primarily scrape from ASL teaching resources and rely on glosses or English text already present in these resources. However, labelling signs has been

Dataset	Vocab Size	Videos	Videos/ sign	Signers	Collection	Consent
RWTH BOSTON-50	50	483	9.7	3 Deaf	Lab	✓
Purdue RVL-SLL	39	546	14.0	14 Deaf	Lab	✓
Boston ASLLVD	2,742	9,794	3.6	6 Deaf	Lab	✓
WLASL-2000	2,000	21,083	10.5	119 Unknown	Scraped	✗
ASL Citizen	**2,731**	**83,912**	**30.7**	**52 Deaf/HH**	**Crowd**	✓

Table 1: Prior ISLR datasets for ASL compared to ASL Citizen. HH stands for hard of hearing.

the subject of significant scholarship, and even linguistically trained annotators struggle due to a lack of conventional notation system (Fenlon et al., 2015; Hochgesang et al., 2018). In addition, videos from teaching resources are filmed in professional studio contexts similar to lab-collected data, and have similar limitations in scale and diversity. Moreover, scraped datasets typically do not have the required permissions from content creators and hosting platforms.

The design of datasets for sign language development has profound implications around issues of fairness, ethics, and responsible AI development (Bragg et al., 2021). Sign language data features identifiable faces and is expensive and labor-intensive to create; consent of all contributors is paramount. Prior work has proposed crowdsourcing sign language videos (Bragg et al., 2022), and addressing ethical concerns during collection as a way to collect larger, representative datasets (Bragg et al., 2020). However, these methods have not been implemented at scale. Our work builds on this prior work, by implementing a crowdsourcing platform with optimized versions of tasks proposed in (Bragg et al., 2022) and partnering closely with Deaf community members throughout. As a result, we present the largest-to-date public sign video dataset, with a large vocabulary and diverse set of Deaf and hard of hearing signers in heterogeneous everyday settings, collected with consent.

2.3 Isolated Sign Language Recognition Methods

The last few years have seen increasing research on isolated sign language recognition (ISLR), as evidenced by the growing number of literature reviews in the space (Adeyanju et al., 2021; Cheok et al., 2019; Cooper et al., 2011; Er-Rady et al., 2017; Joksimoski et al., 2022; Koller, 2020; Rastgoo et al., 2021; Tolba and Elons, 2013; Wadhawan and Kumar, 2021). Earlier approaches rely on handcrafted features and classic machine learning classifiers (Carmona and Climent, 2012; Forster et al., 2013; Monteiro et al., 2016; Ong and Bowden, 2004; Vogler and Metaxas, 1997), typically on datasets with small vocabularies and relatively few videos.

More recent approaches have shifted to deep learning, especially as larger datasets have become available. Appearance-based methods operate directly on video frames as inputs: approaches include spatially pooled convolutional neural networks (Li et al., 2020; Rao et al., 2018) and transformers (Boháček and Hrúz, 2022; De Coster et al., 2020). Alternatively, pose-based methods fit models on keypoints extracted using human pose models (Boháček and Hrúz, 2022; Li et al., 2020; Selvaraj et al., 2021) like OpenPose (Cao et al., 2017) or MediaPipe (Lugaresi et al., 2019) for keypoint detection.

Despite the breadth of research approaches, state-of-the-art ISLR methods still have relatively low recognition performance, achieving around 30% accuracy (Li et al., 2020; Selvaraj et al., 2021) over realistic vocabulary sizes (2,000+ signs), a level unlikely to be useful for any real use case.

2.4 Sign Language Dictionaries

Dictionaries are a meaningful application of ISLR to Deaf community members (Bragg et al., 2019; Huenerfauth and Hanson, 2009): in addition to playing a cultural role in language documentation, they are valuable tools for language users and learners. However, creating effective sign lookup systems is difficult. English-to-ASL dictionaries (e.g. (Signing Savvy, 2022)) accept written queries and so can leverage natural language processing (NLP) techniques for matching, but ASL-to-English dictionaries or ASL-to-ASL dictionaries cannot because there is no standard sign language writing system.

To address these challenges, two main approaches are used: feature-based lookup and example-based lookup. In feature-based lookup, users specify various parameters of the sign they are seeking

(e.g., the handshape used, the location on the body where the sign is made, the motion of the sign, etc.), and a list of top matching signs are returned (Bragg et al., 2015). While this simplifies the lookup problem, unlike English spelling, these features are not conventional and are not widely used or taught, so users may not be familiar with how to make use of them. Novice learners may especially have trouble noticing and remembering these parameters.

In the case of example-based lookup, users demonstrate a sign by video, and a list of top matching signs are returned. While this may be more accessible for users, it is more challenging computationally, now requiring video-based processing to complete the lookup (Xu et al., 2022). Because this is significantly more challenging and largely unsolved, human-computer interaction (HCI) research has primarily focused on understanding potential dictionary use through wizard-of-oz methods and analysing metrics (Alonzo et al., 2019; Hariharan et al., 2018; Hassan et al., 2021, 2022). Our work advances the state of ISLR for dictionary retrieval, potentially enabling functional example-based dictionaries to be created and studied.

3 ASL Citizen Dataset Creation

3.1 Data Collection

We build on prior work (Bragg et al., 2022), which piloted the first crowdsourcing tasks for sign language data collection in a small user study. In this work, we scale data collection with a longer-term deployment with fluent Deaf signers. We deployed a similar collection method to that described in (Bragg et al., 2022), but with enhancements designed to improve recording efficiency and data quality (e.g. displaying a target body contour on the webcam feed to help participants stay in frame). This method secures explicit consent from participants, and curates data appropriate for work on dictionary lookup; participants are asked to contribute videos for a communal dictionary, recording videos in real-world settings, similar to real-world dictionary queries. The task design also eliminates labelling challenges by collecting pre-labelled content. The Appendix provides details on design changes compared to prior work and effects.

On the platform, participants were informed that their videos would contribute to a communal dictionary that displays signing diversity and be released in a public research dataset, and provided explicit consent. Next, each participant was given the full vocabulary of 2,731 signs (from ASL-LEX (Caselli et al., 2017; Sehyr et al., 2021)), sorted such that signs with the fewest videos are shown first, with the goal of encouraging our dataset to be relatively balanced across signs. For each sign, they first viewed a "seed" video of an isolated sign filmed by a highly proficient, trained ASL model. This "seed signer" was a paid research member, and used a high-resolution camera and wireless mouse to ensure recording quality. Participants were then prompted to record their own version of the sign they just saw demonstrated. As in (Bragg et al., 2022), participants could re-play the seed video or their own videos, and re-record or delete videos. We provided an optional English gloss, hidden by default to encourage focus on the ASL. For every 300 videos, participants received a $30 gift card, for up to 3,000 signs. Those who completed the vocabulary re-visited signs with the least recordings. Providing basic demographics was optional.

We took several steps to help ensure that our data collection was culturally sensitive and participatory. Our research team is highly interdisciplinary, including experts in computer vision, NLP, HCI, Deaf studies, and ASL linguistics. This enabled us to identify and address challenges to ISLR in a comprehensive and human-centered way. Our team included several Deaf members and hearing people fluent in ASL and active in the Deaf community. Deaf researchers were involved in every aspect of the research and made direct contact with participants. All recruitment and consent materials were presented in an ASL-first format, featuring short ASL videos. Participants were recruited through relevant email lists and snowball sampling of Deaf researchers' social networks. All procedures were reviewed and approved by IRB.[2]

3.2 Data Verification and Cleaning

To help ensure data quality, we engaged in verification and cleaning procedures under close guidance from our ethics review board. First, we removed empty videos automatically, by removing any

[2]Both Microsoft and Boston University IRBs reviewed the project, with Microsoft serving as the IRB of record (#418).

	Train	Val	Test
Users	35	6	11
Videos	40,605	10,309	32,999
User distribution	60% F	83% F	55% F
Video distribution	54% F	71% F	55% F

Table 2: Statistics for ASL Citizen dataset splits.

Figure 1: Random ASL Citizen video stillframes.

videos that were too small (under 150 KB) or in which YOLOv3 (Redmon and Farhadi, 2018) did not detect a person. Altogether, this filter excluded 50 videos. We also engaged in a manual review of the first and last videos recorded by each participant and random samples throughout, checking for potential sensitive content or anomalous behavior. We removed one user's videos, who recorded many videos without a sign in them. To protect the privacy of people who appear in the background of the video (i.e. besides the participant), we used YOLOv3 to detect if multiple people were present. We also manually identified sets of videos where certificates or other personally identifying objects were visible in the background. For both these sets of videos, we blurred the background using the user segmentation from MediaPipe holistic. This gave us a total of 911 videos with blurred background, which appear in training, validation, and test sets.

3.3 ASL Citizen Dataset Benchmark for Dictionary Retrieval

Our final cleaned dataset, ASL Citizen, contains a total of 83,912 videos corresponding to 2,731 signs recorded by 52 participants, including the seed signer (Table 1).[3] This is the largest ISLR dataset for ASL (or any other sign language) to date. We standardize our glosses to those previously documented by ASL-LEX (Caselli et al., 2017), a database of lexical and phonological properties of signs. This provides standardized sign identifiers independent of English gloss (because glosses are often ambiguous), and allows for researchers to make use of linguistic annotations provided by ASL-LEX (e.g., the handshape of each sign in our dataset). Each sign has multiple recordings ($\mu = 30.7, \sigma = 1.9$).

Of the signers represented in the dataset, 49 identify as Deaf and 3 as hard of hearing; 32 as female and 20 as male; and with ages ranging from 20 to 72 years old ($\mu = 36.16, \sigma = 14.2, n = 49$). These signers come from 16 U.S. states, with between two and 65 years of ASL experience ($\mu = 30, \sigma = 15.12, n = 48$).

We expect our videos to be consistent with our chosen dictionary retrieval task (i.e. participants resemble webcam users demonstrating signs to a dictionary). All users were informed they were contributing to a dictionary and recorded in real-world settings. Since ASL Citizen videos were collected in a variety of settings, they contain varied illumination, background, resolution, and angle. Since our videos are self-recorded, there is also variability in when users start and finish signing in videos (i.e. amount of padding), and the speed, repetition, and execution of signs. We consider this variability to be valuable as it is within the scope of "in-the-wild" dictionary queries, and do not attempt to filter or standardize these variables.

Dataset Splits. We also release standardized splits of our 52 users into training, validation and test sets (Table 2), attempting to balance by female-to-male gender ratio (our dataset does not contain videos from signers who identified themselves as non-binary). Importantly, our splits are established such that users in the validation and test sets are unseen during training. While previous work randomly split videos (Li et al., 2020) (so users in the test set may be already seen in training), we felt it was critical to evaluate on unseen users to align with our dictionary retrieval problem; it is unlikely that a user looking up a sign would be in the training set of a deployed model. Accordingly, the seed signer is placed in the training split. Finally, in our splits, we sought to provide a large test set. While this still leaves sufficient videos for training (more than 2.5× that of the previous largest

[3]Please note that in this paper, we worked with ASL Citizen version 0.9. All statistics about the dataset (number of videos etc.) refer to this dataset version, and all experiments were run with this dataset version.

dataset (Li et al., 2020)), we reasoned that a large test set with multiple users would not only offer more robust estimates of performance, but also leave the potential for a wider range of methods (e.g., unsupervised domain adaptation) in future work.

Sign Ranking and Metrics. Our dictionary retrieval problem requires models to return a ranked list of signs. We consider standard information retrieval metrics: recall-at-K (for K=1, 5, and 10, where recall-at-1 is the same as accuracy), discounted cumulative gain (DCG) and mean reciprocal rank (MRR). Recall-at-K, measured by determining if the correct sign is in the top K rankings, allows us to consider scenarios where users may look at only the first K signs retrieved in response to their query (Hassan et al., 2021). DCG and MRR, on the other hand, evaluate the overall ranking of the correct sign in the entire list. For all of these metrics, a higher score indicates the correct sign is earlier in the ranking, but since DCG uses a log scale, it is more sensitive to order at the top of the ranking. The Appendix provides formulas and additional details.

4 Methods and Training Details

We train two fundamentally different types of machine learning methods on the ASL Citizen dataset: an appearance-based approach, I3D, which is based on a 3D convolutional network applied directly to the frames of the video (Carreira and Zisserman, 2017); and a pose-based approach, ST-GCN, which preprocesses the video to extract pose information, on which we train a temporal graph convolutional network (Yan et al., 2018). Both are classifiers with an output space equal to the size of the vocabulary. For both, to generate a ranked list of retrieved signs, we sort the output probabilities across labels.

4.1 Data Preprocessing

For our I3D model, we preprocess videos by standardizing to 64 frames. Due to variance in sign length and user execution, our dataset videos differ in length. While previous works use random temporal crops to standardize frame lengths during training (Li et al., 2020), we reasoned this practice might alter sign semantics: some signs are compounds of multiple signs, and temporal cropping may reduce the sign to just one root sign. Instead, we standardize training and evaluation videos by skipping frames: for videos longer than 95 frames, we skip every other frame, and for videos longer than 159 frames, we take every third frame. Next, videos shorter than 64 frames are padded with the first or last frame, and longer videos had even numbers of frames removed from the start and end. During training, we augmented with random horizontal flips to simulate left and right handed signers.

For our ST-GCN model, we extracted keypoints using MediaPipe holistic. We use a sparse set of 27 keypoints previously established by OpenHands (Selvaraj et al., 2021). Following previous practice, extracted keypoints are center scaled and normalized using the distance between the shoulder keypoints. Since our videos contain a higher frame-rate than ISLR datasets analyzed by OpenHands, we cap the maximum frames to 128, and downsample frames evenly if the video is longer. As with OpenHands, we apply random shearing and rotation transformations during training as data augmentation.

4.2 Model Structure and Training

We train our I3D model for a maximum of 75 epochs using learning rate 1e-3 and weight decay 1e-8, with an Adam optimizer and ReduceLROnPlateau scheduler with patience 5. As we observed that a cross-entropy loss on the entire video alone led to poor convergence, we also employ a weakly supervised per-frame loss previously used in the Charades Challenge (CVPR 2017 Workshop on Visual Understanding Across Modalities, 2017), where the cross-entropy loss is also applied to each frame in the video (predictions for each frame are produced using linear interpolation of the temporal dimension of the I3D model). Our final loss is an average of the full-video cross-entropy loss and this per-frame cross-entropy loss.

We train our ST-GCN model for a maximum of 75 epochs using a learning rate of 1e-3 using an Adam optimizer and a Cosine Annealing scheduler.

Model	Train Data	Test Data	DCG	MRR	Rec@1	Rec@5	Rec@10
I3D	ASL Citizen	ASL Citizen	**0.7843**	**0.7247**	**0.6210**	**0.8536**	**0.9000**
ST-GCN	ASL Citizen	ASL Citizen	0.7639	0.7001	0.5965	0.8267	0.8780

Table 3: Appearance and pose-based results and baselines. The best results on the ASL Citizen test set are in **bold**. These results represent a phase transition in isolated sign language recognition, up from previous best results in the mid-30% for accuracy (recall-at-1).

Model	Train Data	Test Data	DCG	MRR	Rec@1	Rec@5	Rec@10
I3D	ASL Citizen	ASL Citizen	**0.7843**	**0.7247**	**0.6210**	**0.8536**	**0.9000**
I3D	WLASL-2000	WLASL-2000	--	--	0.3248	0.5731	0.6631
I3D	ASL Citizen	Subset	*0.8573*	*0.8159*	*0.7338*	*0.9163*	*0.9466*
I3D	WLASL-2000	Subset	0.2796	0.1479	0.0850	0.2007	0.2748
ST-GCN	ASL Citizen	ASL Citizen	0.7639	0.7001	0.5965	0.8267	0.8780
ST-GCN	WLASL-2000	WLASL-2000	–	–	0.2140	–	–

Table 4: Results comparing across datasets between ASL Citizen, the prior WLASL-2000 datasets, and a subset of ASL Citizen "matched" to WLASL-2000. The best results on the ASL Citizen test set are in **bold**; on the Subset dataset are in *italics*; and on other datasets are underlined. The grey rows are reproduced from Table 3. Model performance differs substantially between ASL Citizen than WLASL, even on an aligned subset of data.

For each model, we selected the best-performing checkpoint on the validation dataset for analysis on our test dataset. Code and model weights will be released publicly alongside our dataset.

5 Results and Analysis

5.1 Novel Benchmarks

We trained our appearance-based I3D model (Carreira and Zisserman, 2017) on the ASL Citizen dataset to establish an initial baseline, which achieves a top-1 accuracy (recall-at-1) of 62.10%, with a DCG of 0.784 and MRR of 0.725 (see the third row in Table 3).[4] This accuracy is notable, given the difficulty of the problem – our dataset has completely unseen users, and spans one of the largest vocabulary sizes in ISLR to date (2,731 signs). The pose-based ST-GCN model performs similarly, but consistently worse by a few percentage points on all metrics, but still substantially better than any previous reported results on datasets of similar size and complexity. For comparison, due to the number of classes, random guessing would only yield 0.04% expected accuracy.

In previous work, appearance-based and pose-based models have generally shown competitive performance. Pose-based methods reduce information and potentially introduce errors at the keypoint extraction step, at the benefit of making features relevant for signs more accessible and standardized compared to raw pixels. However, this standardization may not outweigh errors and loss of information when the training dataset is diverse enough to train general appearance-based methods. Consistent with this, we note that our pose- based model performance slightly lags behind that of our appearance-based model.

5.2 Comparison to Prior Datasets

We subsequently seek to understand how much our dataset advances the performance of ISLR models in ASL, compared to previous datasets. To do this, we compared our model trained on our dataset, to a public model trained on the previously largest ASL ISLR dataset, WLASL-2000 (Li et al., 2020); see Table 1. We made no changes to model architecture compared to the ASL Citizen I3D model, meaning that any improvements are because of the training data, not model capacity.

[4]Please note that results may vary slightly for future iterations of the dataset and paper. We ask that authors compare to the most up-do-date numbers.

Model	Train Data	Test Data	DCG	MRR	Rec@1	Rec@5	Rec@10
I3D	ASL Citizen	ASL Citizen	**0.7843**	**0.7247**	**0.6210**	**0.8536**	**0.9000**
I3D Features	ASL Citizen	ASL Citizen	0.7427	0.6707	0.5493	0.8215	0.8847
I3D Features	WLASL-2000	ASL Citizen	0.3156	0.1795	0.0986	0.2538	0.3443
I3D Features	Kinetics	ASL Citizen	0.1236	0.0131	0.0043	0.0139	0.0225
I3D Features	ASL Citizen	Subset	0.8295	0.7787	0.6787	0.9032	0.9450
I3D Features	WLASL-2000	Subset	0.4048	0.2716	0.1662	0.3773	0.4854

Table 5: Results comparing feature representations with ranking by cosine-similarity. The grey row is copied from Table 3. Across datasets, the representations learned on the ASL Citizen training data are substantially better than those learned on WLASL-2000 or Kinetics.

A direct comparison of these models is challenging because these models use independent gloss mappings. Not only does this mean that the number of classes differs between models, but the same English gloss in our model may refer to a different sign in WLASL-2000. We overcome this challenge by comparing the models in two distinct ways (benchmarks in Table 4).

First, we directly compared metrics on our test set to previously reported accuracy on the WLASL-2000 test set. While this comparison does not account for potential differences in test set difficulty, we believe our test dataset is more challenging than that of WLASL-2000. First, unlike WLASL-2000, we evaluate on unseen signers. Second, our vocabulary size is larger. Under this comparison, our model achieves a top-1 accuracy of 62.10% on the ASL Citizen test set, while (Li et al., 2020) reports a top-1 accuracy of 32.48% on the WLASL-2000 test set.

Second, we reduced our test set to a subset of glosses that we were confident referred to the same sign in both ASL Citizen and WLASL-2000, and used this test set to compare models trained on ASL Citizen and WLASL-2000. We used a reduced version of ASL Citizen's test set to ensure it would not contain anyone seen during training by *either* model, allowing for a more fair comparison. We excluded any sign with a documented variant in either dataset, and matched glosses only if there was an exact match in English gloss between the two datasets. This procedure resulted in a reduced dataset with 1,075 glosses ("Subset" in our tables). To ensure these glosses referred to the same signs, an author fluent in ASL examined one example per sign from each dataset for 100 random signs, and did not find any discrepancies. To control for both models outputting different numbers of classes, we recalculated the softmax on only the logits corresponding to the 1,075 overlapping glosses (effectively excluding predictions for classes outside of this subset). Under this comparison, top-1 accuracy is 73.38% for our model vs. 8.50% for the WLASL-2000 model. We hypothesize that the drastic drop in accuracy for WLASL-2000 also relative to its original test set (32.48%) is because the model was originally evaluated on seen users, and fails to generalize to unseen users.

Together, these results suggest that our dataset significantly advances the state-of-the-art in ISLR performance for ASL. In the most naive comparison with independent test sets, our model doubles accuracy over the state-of-the-art (62.10% vs. 32.48%). In more standardized comparisons where both models are evaluated on the same test set, our model outperforms prior results by 8.6 times (73.38% vs. 8.50%).

5.3 Comparison of Learned Feature Representation

Finally, we reasoned that while the classifiers of the ASL Citizen and WLASL-2000 models may predict different sets of signs, both models may still be learning features to recognize signs more generally. To assess this, we extracted the globally pooled feature representation before the classification layer of each model, and built a simple nearest-neighbor classifier using the cosine distance from representations of videos in the ASL Citizen training dataset to classify ASL Citizen test data. We report accuracies for this classifier on both the full ASL Citizen test dataset, and the reduced overlapping version, in Table 5. For the full test dataset, our model achieves a top-1 accuracy of 54.93%, while the WLASL-2000 model achieves 9.86% accuracy. For the reduced overlap, our model achieves a top-1 accuracy of 67.87%, while the WLASL-2000 model achieves 16.62% accuracy. This suggests that our dataset enables learning a more robust representation of ASL. We similarly test the performance of I3D feature representations trained on Kinetics (a large-scale human action recognition dataset) (Carreira and Zisserman, 2017): the reported results, which are close to ran-

	DCG	MRR	R@1	R@5	R@10
0%	0.103	0.002	0.000	0.002	0.004
25%	0.387	0.257	0.160	0.353	0.453
32%	0.380	0.250	0.157	0.342	0.439
*47%	0.665	0.580	0.466	0.719	0.791
50%	0.668	0.583	0.467	0.725	0.800
75%	0.724	0.650	0.536	0.792	0.854
100%	0.784	0.725	0.621	0.854	0.900

Table 6: Performance of models with downsampled ASL Citizen training sets. 25% of ASL Citizen gives comparable results to prior approaches, and improvement has not totally asymptoted at 100%. *The 47% split uniformly sampled videos from each sign.

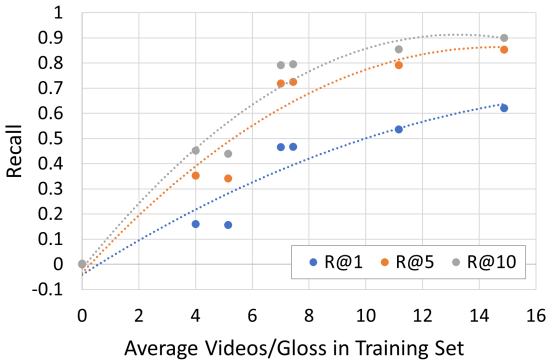

Figure 2: Impact of training data size on recall scores: more videos per sign substantially improves performance. We fit a polynomial function to each sequence.

dom performance, further emphasize out-of-domain models do not generalize to sign language recognition.

5.4 Impact of Dataset Size

Finally, to understand how scaling data collection affects model performance, we systematically downsampled the training dataset. We generated five random splits: 25% of the original training dataset (10,916 videos), 50% (20,302 videos), 75% (30,453 videos), matching the training dataset size of WLASL-2000 (14,055 videos, 32% of our dataset size), and matching the average number of examples per class for WLASL-2000 (19,103 videos, 47% of our dataset size). For all these splits, we ensured that each sign was represented by at least 4 samples, but otherwise sampled at random. We trained an appearance-based I3D model on each of these splits and report results in Table 6, and also experimented with a 0-shot model with no training data by testing performance on an I3D model with randomly initialized weights.

Our results confirm that the scale of our training dataset is critical to performance; as training dataset size decreases, so does accuracy on our test dataset (Figure 2). Interestingly, we observe that when matching the average number of training examples per sign in WLASL-2000 (7 per sign), our model still out-performs previously reported baselines for ISLR (46.59% vs. 32.48%), suggesting the scale of training data may not be the only factor in our improved performance.

6 Discussion and Conclusion

In this work, we introduce a problem formulation for ISLR in the form of dictionary retrieval, provide the largest ISLR dataset to date through a crowdsourcing initiative, and release metrics and baselines showing that our new dataset significantly advances the state-of-the-art for ISLR in ASL.

Our dictionary retrieval formulation is intended to limit misconceptions surrounding sign language. ISLR is a more computationally tractable problem than continuous sign language recognition and significantly more feasible than sign language translation, and has more immediate applications that can benefit Deaf people. Continuous sign recognition requires a system to handle co-articulation, fingerspelling, facial expressions, depictions, and classifiers constructions. Translation requires not only accurate continuous sign recognition, but the ability to move from the syntax and grammar of ASL to English and vice versa. A researcher without this domain expertise may assume that tokenizing a video into a sequence of signs and applying ISLR to these tokens is sufficient for translation. Underestimating the complexity of these aspects of sign language and translation has historically led to objections from Deaf communities (De Meulder, 2021; Erard, 2017; Harris et al.,

2009; Kusters et al., 2017). The dataset presented here will enable technologies like dictionary search, that are focused on reliably classifying signs and do not require considering syntax or even optimal English translation. We discourage researchers from attempting to use this dataset alone (e.g., without also learning from continuous datasets) for more complex applications.

Our data collection process reflects this dictionary lookup problem formulation and yields a large-scale, high-quality ISLR dataset. Traditional lab collection techniques offer high-quality data but limit diversity, whereas scraping data may promise diversity while introducing labelling and fluency errors and compromising consent. For example, scraped videos may degrade quality of labels as videos may only be associated with unstructured English text, or signed by novices. By crowdsourcing data, we balance the trade-off between size, quality and diversity. First, our scalable platform allows users to contribute from their homes and other everyday spaces, capturing real-world diversity and use of signing space representative of dictionary applications. Second, by prompting our contributors with specific signs to demonstrate, we have higher confidence in our labels, which are also automatically generated. Third, recruiting fluent signers from trusted groups ensures our collected data is high-quality and reflects the conventions of ASL. Our study informs future data collection efforts: participatory approaches with meaningful contribution from Deaf researchers can yield not just larger datasets, but higher quality data.

While our benchmark models achieve high levels of ISLR performance on unseen signers, future work is still required to fully solve the dictionary retrieval problem. We have only tackled dictionary retrieval in its cleanest form, and real-life dictionaries present many use cases not fully addressed in this work. First, our supervised models operate on a fixed vocabulary. These models are unable to cope with signs outside of this vocabulary. Since, like spoken languages, sign languages are dynamic, with new signs emerging regularly, future work should consider methods that can adapt to signs that were unseen at training time. Second, we evaluated our models on fluent Deaf signers who can expertly replicate signs. Novice signers would likely have difficulty recalling a sign, and may not execute it perfectly. Therefore, we expect a performance gap for these users, which future models could address.

Future work also includes deepening evaluation. While we considered DCG, MRR, and recall-at-k as metrics in this work, these metrics may not fully align with user preferences. These metrics treat all errors equivalently regardless of severity (with a binary evaluation of relevance). However, metrics that better capture overall list relevance (e.g., by weighting relevance of signs with similar meanings or visual appearance high in the list) may reflect a better user experience (Hassan et al., 2021). Deployed dictionaries also must meet performance measures beyond accuracy (like ease of use or speed). Finally, although we report overall performance, our I3D model achieves accuracies ranging from 43% to 75% across users. Real-world applications are becoming increasingly viable, and future work should explore whether ISLR models are *equitable* – if there are disparities between demographic groups served by models – and how such performance biases might be addressed.

Acknowledgments and Disclosure of Funding

First and foremost, we thank the community members who participated in this research project by recording themselves signing. We also thank Mary Bellard, Miriam Goldberg, Hannah Goldblatt, Paul Oka, Philip Rosenfield, Bill Thies, and the Microsoft Research outreach team for thoughtful discussions, support, and contributions to the crowdsourcing platform.

This work was supported in part by the National Science Foundation Grants: BCS-1625954 and BCS-1918556 to Karen Emmorey and Zed Sehyr, BCS-1918252 and BCS-1625793 to Naomi Caselli, and BCS-1625761 and BCS-1918261 to Ariel Cohen-Goldberg. Additional funding was from the National Institutes of Health National Institute on Deafness and Other Communication Disorders of and Office of Behavioral and Social Science Research under Award Number 1R01DC018279.

References

IA Adeyanju, OO Bello, and MA Adegboye. Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, 12:200056, 2021.

Oliver Alonzo, Abraham Glasser, and Matt Huenerfauth. Effect of automatic sign recognition performance on the usability of video-based search interfaces for sign language dictionaries. In

The 21st International ACM SIGACCESS Conference on Computers and Accessibility, pages 56–67, 2019.

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 182–191, 2022.

Danielle Bragg, Kyle Rector, and Richard E Ladner. A user-powered american sign language dictionary. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1837–1848, 2015.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31, 2019.

Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. Exploring collection of sign language datasets: Privacy, participation, and model performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14, 2020.

Danielle Bragg, Naomi Caselli, Julie A Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E Ladner. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing (TACCESS)*, 14(2):1–45, 2021.

Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. Exploring collection of sign language videos through crowdsourcing. *CSCW 2022*, 2022.

Diane Brentari and Carol A Padden. Native and foreign vocabulary in american sign language: A lexicon with multiple origins. In *Foreign vocabulary in sign languages*, pages 87–119. Psychology Press, 2001.

Diane Brentari et al. *A prosodic model of sign language phonology*. Mit Press, 1998.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

Josep Maria Carmona and Joan Climent. A performance evaluation of hmm and dtw for gesture recognition. In *Iberoamerican Congress on Pattern Recognition*, pages 236–243. Springer, 2012.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

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

Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10 (1):131–153, 2019.

Helen Cooper, Brian Holt, and Richard Bowden. Sign language recognition. In *Visual analysis of humans*, pages 539–562. Springer, 2011.

CVPR 2017 Workshop on Visual Understanding Across Modalities. Charades challenge: Recognize and locate activities taking place in a video, 2017. URL http://vuchallenge.org/charades.html.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Sign language recognition with transformer networks. In *12th international conference on language resources and evaluation*, pages 6018–6024. European Language Resources Association (ELRA), 2020.

Maartje De Meulder. Is "good enough" good enough? ethical and responsible development of sign language technologies. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 12–22, 2021.

Adil Er-Rady, Rdouan Faizi, R Oulad Haj Thami, and H Housni. Automatic sign language recognition: A survey. In *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–7. IEEE, 2017.

Michael Erard. Why sign-language gloves don't help deaf people, 2017. URL https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/.

Jordan Fenlon, Kearsy Cormier, and Adam Schembri. Building bsl signbank: The lemma dilemma revisited. *International Journal of Lexicography*, 28(2):169–206, 2015.

Susan D Fischer and Harry Van der Hulst. Sign language structures. *Deaf studies language and education*, pages 319–331, 2003.

Jens Forster, Christian Oberdörfer, Oscar Koller, and Hermann Ney. Modality combination techniques for continuous sign language recognition. In *Iberian conference on pattern recognition and image analysis*, pages 89–99. Springer, 2013.

Hands and Voices. Deaf culture and community, 2022. https://www.handsandvoices.org/comcon/articles/deafculture.htm.

Dhananjai Hariharan, Sedeeq Al-khazraji, and Matt Huenerfauth. Evaluation of an english word look-up tool for web-browsing with sign language video for deaf readers. In *International Conference on Universal Access in Human-Computer Interaction*, pages 205–215. Springer, 2018.

Raychelle Harris, Heidi M Holmes, and Donna M Mertens. Research ethics in sign language communities. *Sign Language Studies*, 9(2):104–131, 2009.

Saad Hassan, Oliver Alonzo, Abraham Glasser, and Matt Huenerfauth. Effect of sign-recognition performance on the usability of sign-language dictionary search. *ACM Transactions on Accessible Computing (TACCESS)*, 14(4):1–33, 2021.

Saad Hassan, Akhter Al Amin, Alexis Gordon, Sooyeon Lee, and Matt Huenerfauth. Design and evaluation of hybrid search for american sign language to english dictionaries: Making the most of imperfect sign recognition. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3491102.3501986. URL https://doi.org/10.1145/3491102.3501986.

Julie Hochgesang, OA Crasborn, and Diane Lillo-Martin. Building the asl signbank. lemmatization principles for asl. 2018.

Matt Huenerfauth and Vicki Hanson. Sign language in the interface: access for deaf signers. *Universal Access Handbook. NJ: Erlbaum*, 38:14, 2009.

Boban Joksimoski, Eftim Zdravevski, Petre Lameski, Ivan Miguel Pires, Francisco José Melero, Tomás Puebla Martinez, Nuno M Garcia, Martin Mihajlov, Ivan Chorbev, and Vladimir Trajkovik. Technological solutions for sign language recognition: a scoping review of research trends, challenges, and opportunities. *IEEE Access*, 2022.

Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

Annelies Kusters, Maartje De Meulder, and Dai O'Brien. *Innovations in deaf studies: The role of deaf scholars*. Oxford University Press, 2017.

Harlan Lane. *When the Mind Hears: A History of the Deaf*. Knopf Doubleday Publishing Group, London, 1989.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020.

Dennis Looney and Natalia Lusin. Enrollments in languages other than english in united states institutions of higher education, summer 2016 and fall 2016. In *Modern language association*. ERIC, 2019.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Carolyn McCaskill, Ceil Lucas, Robert Bayley, and Joseph Christopher Hill. *The hidden treasure of Black ASL: Its history and structure*. Gallaudet University Press Washington, DC, 2011.

Caio DD Monteiro, Christy Maria Mathew, Ricardo Gutierrez-Osuna, and Frank Shipman. Detecting and identifying sign languages through visual features. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 287–290. IEEE, 2016.

Eng-Jon Ong and Richard Bowden. A boosted classifier tree for hand shape detection. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 889–894. IEEE, 2004.

G Anantha Rao, K Syamala, PVV Kishore, and ASCS Sastry. Deep convolutional neural networks for sign language recognition. In *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, pages 194–197. IEEE, 2018.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. The asl-lex 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277, 2021.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. *arXiv preprint arXiv:2110.05877*, 2021.

Signing Savvy. Signing savvy an ASL sign language video dictionary, 2022. URL https://www.signingsavvy.com/.

MF Tolba and AS Elons. Recent developments in sign language recognition systems. In *2013 8th International Conference on Computer Engineering & Systems (ICCES)*, pages xxxvi–xlii. IEEE, 2013.

Christian Vogler and Dimitris Metaxas. Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 156–161. IEEE, 1997.

Ankita Wadhawan and Parteek Kumar. Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, 28(3):785–813, 2021.

WFD. Legal recognition of national sign languages., 2022a. https://wfdeaf.org/news/the-legal-recognition-of-national-sign-languages/.

WFD. World federation of the deaf, 2022b. URL http://wfdeaf.org/our-work.

Ronnie Wilbur and Avinash C Kak. Purdue rvl-slll american sign language database. 2006.

Ronnie B Wilbur. Phonological and prosodic layering of nonmanuals in american sign language. In *The signs of language revisited*, pages 196–220. Psychology Press, 2013.

Chenchen Xu, Dongxu Li, Hongdong Li, Hanna Suominen, and Ben Swift. Automatic gloss dictionary for sign language learners. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 83–92, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-demo.8. URL https://aclanthology.org/2022.acl-demo.8.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*, 2021.

Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Joint Pattern Recognition Symposium*, pages 401–408. Springer, 2005.

Inge Zwitserlood. Classifiers. In *Sign language: An international handbook*. De Gruyter, 2012.

A Platform modifications

Changes to the platform introduced in prior work (Bragg et al., 2020), with the goal of helping to meet our data collection needs are described below.

Platform change	Intended impact on data collection
Improved recording task flow (automatic countdown to recording start, video upload in the background)	Scale: a more efficient contribution process enables participants to contribute more videos in a set amount of time
Visual design overhaul (updated color scheme, button design, and page layouts; replaced one-off video players with standard video players)	Scale: a better user experience attracts and encourages use, and increases trust in the platform creators, which lowers barriers to contributing
Updated platform structure and navigation (updated landing page with direct links to contribute; updated navigation links in header; updated info page with more project details in ASL and English)	Scale: a reduced learning curve lets participants contribute more data, and increased trust lowers barriers to contributing
Infrastructure scaling (set up infrastructure to handle many parallel contributions, large data storage, backup scripts)	Scale: provides technical capabilities to collect data at scale
Removed participant ability to share content with one another in real-time (instead creating two-phased implementation of collection followed by release)	Scale: removes potential for viewing off-putting content from other users
New set of seed sign videos executed by a well-known fluent signer who is not white-presenting	Participant diversity: representation creates a more welcoming environment and fosters contributions from a wider range of participants
Improved personal data view (enabling users to search through their own videos, sorting videos and demographics into tabs)	Data quality: participants can more easily review and update contributions
Overlaid the outline of a human figure on the webcam feed	Data quality: videos are more likely to capture the entire upper body and are more standardized across participants

Table 7: Summary of platform feature changes, alongside potential impacts on the contributor and resulting dataset.

B Dictionary Retrieval Metrics

For a given query, if *i* is the placement of the desired gloss in the returned list of glosses, we calculate metrics using the following formulae:

- Discounted Cumulative Gain = $\frac{1}{\log_2(i+1)}$, ranges in $[\epsilon, 1]$ with 1 indicating that the correct gloss is always the top ranked item, and $\epsilon = \frac{1}{\log_2(N+1)}$ is the smallest attainable score when the correct gloss is ranked last (in our case, $N = 2729$ so $\epsilon = 0.088$). A completely random ordering with give an average DCG of approximately 0.15.
- Mean Reciprocal Rank = $\frac{1}{i}$, ranges in $[\epsilon, 1]$ with 1 indicating that the correct class is always the top ranked item, and $\epsilon = \frac{1}{N} = 0.00037$. A completely random ordering will give an average MRR of approximately 0.0032.

The overall scores reported are averages of these metrics across respective data splits (e.g., average over all test instances).