Studying and Mitigating Biases in Sign Language Understanding Models

Anonymous ACL submission

Abstract

Crowdsourced sign datasets collected with the involvement of deaf communities, such as the ASL Citizen dataset, represent an important step towards improved accessibility and documentation of signed languages. However, it is important to ensure that these resources benefit people in an equitable manner. Thus, there is a need to understand the potential biases that may result from models trained on sign language datasets. In this work, we utilize the rich information about participant demographics and lexical features present in the ASL Citizen dataset to study and document the biases that may result from models trained on crowdsourced sign datasets. Further, we apply several bias mitigation techniques during model training, and discuss the results and relative success of these techniques. In addition to our analyses and machine learning experiments, with the publication of this work we release the demographic information about the participants in the ASL Citizen dataset to encourage future work in this space.

1 Introduction

The field of natural language processing (NLP) has historically been skewed towards spoken languages. Before 2021, NLP research output in the space of sign language processing (SLP) research was in the minority, with computer vision increasingly dominating this space (Yin et al., 2021). Following works such as Yin et al. (2021), signed languages have received more attention from the NLP community. However, the comparative lack of resources for signed languages compared to spoken languages heightens the difficulty of SLP, and is compounded by the fact that most accessible information (e.g. online resources and social media) is written in a spoken language (Desai et al., 2024).

The ASL Citizen dataset (Desai et al., 2024) was released to help address this resource gap, with the goal of improving *video-based dictionary retrieval* for sign language, where signers demonstrate a particular sign and the system returns a list of similar signs, ranked from most to least similar. Video-based dictionary retrieval systems can help language learners understand the meaning of a sign, and allow signers to access dictionary resources using signed languages (Desai et al., 2024). As a crowd-sourced dataset with videos of individual signs, the ASL Citizen dataset also serves to improve documentation of signed languages. This dataset is the first crowdsourced dataset of videos for isolated signs, and members of deaf communities were involved and compensated for this effort. This dataset is licensed by Microsoft Research and is bound by the Microsoft Research Licensing Terms[1].

Resources such as the ASL Citizen dataset, that improve accessibility and contribute to the documentation of low-resource languages, are critical. However, it is also important to critically analyze these datasets, in order to understand in what conditions (and for what users) these datasets, and models trained on them, are most beneficial. Currently, there is a limited amount of prior work in this space.

To help address this problem, we explore how signer demographics and more latent sources of bias may impact modeling performance. To do this, we analyze demographics in the ASL Citizen dataset, which presents a diversity of signers and vocabulary, and examine how these demographic features, along with lexical and video-level features, may impact model results. Specifically, we present a detailed analysis of the distributions of different demographics, and feature prevalence among demographics. We also present a linguistic analysis of the dataset based on the ASL-Lex annotations for

[1]Terms of use at https://www.microsoft.com/en-us/research/project/asl-citizen/dataset-license/. We are using this dataset in accordance with its intended use.

each sign. Further, we study how these features impact model performance. What characteristics of a sign video may improve dictionary retrieval results, and are there any disparities in performance among different demographics? Finally, we experiment with different debiasing techniques in order to reduce performance gaps without sacrificing overall model accuracy. In addition to publishing our findings, we release the demographic data for the ASL Citizen dataset, so future researchers can continue to work toward the goal of developing equitable sign language processing systems.

In summary, we address the following three research questions:

RQ1 How is the ASL Citizen data distributed, demographically and linguistically?

RQ2 Which demographic and linguistic factors impact dictionary retrieval results in the ASL Citizen dataset?

RQ3 Can we use debiasing strategies to mitigate disparate impacts while maintaining high performance for dictionary retrieval models?

2 Related

Most readily-available information (i.e. online resources and social media) is written, which may limit accessibility for signers. Sign language processing tasks, such as dictionary retrieval, are designed to improve the accessibility of existing systems/resources for Deaf and Hard-of-Hearing (DHH) people. Desai et al. (2024) created the ASL Citizen dataset for the purpose of improving dictionary retrieval.

The ASL Citizen dataset is composed of videos of individual signs for isolated sign language recognition (ISLR). Other ISLR datasets with videos of individual signs have been released, including WL-ASL (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). However, the ASL Citizen dataset is the first large-scale ISLR dataset to be **crowdsourced**. The dataset is made up of crowdsourced videos from ASL signers, where each video corresponds to a particular sign. The corpus is made up of videos for 2731 unique signs, all of which are contained in the ASL-Lex dataset Caselli et al. (2017), a lexical database of signs with annotations including the relative frequency, iconicity, grammatical class, English translations, and phonological properties of the sign. Thus, researchers studying this dataset can also take advantage of the ASL-Lex annotations.

As part of the original data collection effort, demographic information about each participant was collected, but it was not released. With the publication of this work, we release the demographic data in this set, and provide a detailed analysis of this data. Further, using the ASL-Lex features, we analyze the properties of the signs depicted in this dataset, and study how these features, in combination with participant demographics impact model performance. Finally, we qualitatively analyze these videos, and identify some video-level features that may increase or decrease performance.

Motivating our work are previous works indicating that demographics of the signer may impact their signing. For instance, characteristics of particular spoken languages or dialects have been shown to influence gestures, and in turn sign production (Cormier et al., 2010). One example of an ASL dialect is Black ASL, which scholarly evidence has shown to be its own dialect (Toliver-Smith and Gentry, 2017), and for which documentation of dialectical differences dates back to 1965 (Stokoe et al., 1965). Whether an individual speaks Black ASL is likely heavily influenced on their race or ethnicity. An example of geographical differences is Martha's Vineyard, an island off the coast of the United States, where an entire signed language emerged due to the high prevalence of deaf individuals in this community. Hearing and deaf people alike used this language to communicate until the mid-1900s (Kusters, 2010). There is also a distinct Canadian ASL dialect used by signers in English-speaking areas of Canada (Padden, 2010), which is documented in a dictionary (Bailey et al., 2002). Age of language acquisition also impacts ASL production; delayed first-language acquisition affects syntactic knowledge for ASL signers (Boudreault and Mayberry, 2006) and late acquisition (compared to native acquisition) was found to impact sensitivity to verb agreement (Emmorey et al., 1995).

Previous work also indicates the impact of certain features on sign language modeling; for instance, training an ISLR model to predict phonological characteristics of a sign in addition to the sign itself was found to improve model performance by almost 9% (Kezar et al., 2023). (Sarhan et al., 2023) find improved performance when using attention to focus on hand movements in sign videos. However, to our knowledge, there are no existing works that extensively study various sources of model bias on a crowdsourced dataset of sign videos with collected participant demographics. With this work, we aim to address this gap with a systematic analysis of the impact of various participant-level, sign-level, and video-level features, and results from deploying different debiasing techniques.

3 How is the ASL Citizen data distributed, demographically and linguistically?

The ASL Citizen dataset is a crowdsourced dataset containing 83,399 videos of individual signs in ASL from 52 different participants. The dataset contains 2731 unique signs that are included in the ASL-Lex (Caselli et al., 2017) dataset, a dataset with detailed lexical annotations for each sign. The authors of the original work report some demographic statistics, but the demographics of individual (de-identified) participants have not been released. Here, we answer our first research question: how is the ASL Citizen data distributed, demographically and linguistically? We provide a detailed report that includes demographics breakdowns and analyses of various linguistic and video features in the dataset, including the breakdown of these features by gender. We will release the demographics of participants upon publication of this paper.

3.1 Demographic Distributions

In total, the ASL Citizen dataset is comprised of 32 (61.5%) women and 20 (38.5%) men. 21 women are in the training set (60%), 5 are in the validation set (83%), and 6 are in the test set (55%). The vast majority of participants report an ASL level of 6 or 7, and the full distribution of ASL levels can be seen in Figure 4. The participants also list their U.S. states. Using this information, we divide them into four regions as defined by the U.S. Census [2]: Northeast, Midwest, South, and West. We find that more participants in the dataset are from the Northeast than any other region, as shown in Figure 4. We also find that the age range of participants is skewed: participants in their 20s and 30s make up 32 of the 52 participants (see Figure 5).

Participants did not note their ethnicity or race for this dataset. As such, to uncover potential biases related to the participants' perceived skin tone in their videos, we ran the `skin-tone-classifier` Python package from Rejón Pina and Ma on the frame with the first detected face in each video. We found that when we did not specify that the videos were in color, the classifier most often detected them as black and white. When we specified that the videos were in color, the most common skin tone detected (out of the default color palette used in Rejón Pina and Ma) was #81654f. Because the classifier most commonly detected images as black and white, we also tried specifying the video frames as being black and white. When we did this, the most common skin tone detected was #b0b0b0, and the distribution was somewhat different from when the images were specified as being in color. Thus, there may be some errors in the skin tone classification. We plot these results in Figure 6.

3.2 Sign and Video Features

Because the ASL Citizen dataset is composed of signs from ASL-Lex (Caselli et al., 2017), we have access to ASL-Lex's annotated lexical features of each sign for analysis. No works have, to date, studied these features in-depth on the ASL Citizen sign videos. Further, we conduct additional analyses on the video lengths, similarities and differences from the model, and other notable features in the dataset.

Video Length We analyze the distribution of video lengths, in order to study length variation between submitted videos and identify patterns that may explain performance discrepancies between individuals or members of certain demographics. We find that the distribution of video lengths (s) is skewed left, with a longer tail on the right, as shown in Figure 7.

We also study relative video lengths for participants of different ages and genders. To account for differences between which signs were depicted (since participants did not all record the same signs), for each video, we calculate the number of standard deviations the video length is away from the mean for all videos of that sign - in other words, we calculate standard deviations from the mean at the sign level. We find that, while men on average record videos over .3 standard deviations longer than the mean, women on average record videos over 2 standard deviations shorter than the mean. Thus, compared to other videos with the same sign, women record shorter videos than men. We show these results in Figure 8. We also found that, in general, older participants, par-

[2] https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

ticularly those in their 70s, tend to record longer videos on average (again, relative to other videos of the same sign) than younger participants. Upon manual inspection, we found that older participants were more likely to have longer pauses before or after signing than younger participants, which may explain this gap. We also show these results in Figure 8.

Sign Frequency The ASL Citizen dataset is comprised of 2731 signs from the ASL-Lex dataset Caselli et al. (2017), a dataset with expert annotations about properties of each sign including frequency of use, iconicity, and varying phonological properties. To collect sign frequency labels, deaf signers who use ASL were asked to rate signs from 1 to 7 in terms of how often they appear in everyday conversations, where 1 was “very infrequently" and 7 was “very frequently". We plot and compare the distributions for the ASL Citizen dataset and the ASL-Lex dataset in Figure 9, and find that they are very similar.

We also find that there is little variation in sign frequency for participants of different genders. For male participants, the average sign frequency was 4.1592, while the average sign frequency for female participants was 4.1395, indicating that female participants chose slightly less frequently-occurring signs than men.

Sign Iconicity The ASL-Lex dataset also contains crowdsourced annotations for sign iconicity, where non-signing hearing annotators watch videos of a sign and evaluated how much they look like the sign’s meaning from 1 (not iconic) to 7 (very iconic). The ASL-Lex signs have an average iconicity of XX, and the signs in the ASL-Citizen dataset have an average iconicity of 3.379. We plot these distributions in Figure 10, and again find that they are very similar.

We find average iconicity to be 3.378 for women and 3.381 for men. This indicates that, as with frequency, average sign iconicity exhibits only a slight difference between male and female participants.

4 Methods

4.1 Baselines

For our experiments, unless otherwise stated, we use the baseline I3D and ST-GCN models which were trained on the ASL Citizen dataset and released along with the dataset.[3]. Thus, when we refer to the **I3D** model, we mean the 3D convolutional network referred to as such in the ASL Citizen paper, and is trained on preprocessed video frames from the sign videos. When we refer to the **ST-GCN** model, we are again referring to the baseline of the same name in the paper, which is trained on representations of the participants’ poses that are created by extracting key points.

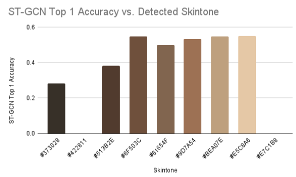

Figure 1: ST-GCN top 1 accuracy scores by detected skin tone. We find that, despite being less represented in the dataset, videos with lighter detected skin tones have higher accuracy scores on average.

5 Which factors impact dictionary retrieval results in the ASL Citizen dataset?

5.1 Participant-level differences

Baseline models perform over 10 percentage points better for male vs. female participants We ran the baseline I3D and ST-GCN models trained on the ASL Citizen dataset (Desai et al., 2024), and, for both models, found an accuracy disparity between participants of different genders. For the I3D model, the overall Top-1 accuracy was 0.6306, while for females it was 0.5914 and for males it was 0.6776; in other words, a gap of over 10 points in favor of male participants was observed. The ST-GCN model saw an even bigger gap; the overall Top-1 accuracy was 0.5944, while the Top-1 accuracy was 0.6838 for males and 0.52 for females.

There is high variation in model accuracy between participants One possible contributor to the above disparities in performance for different genders is the participant-level model accuracy scores. There are 11 participants whose videos are in the test set for the ASL Citizen dataset. Of these 11 participants, 6 are female and 5 are male. When we examine the accuracy scores for each participant, we find high variation between participants

[3]https://github.com/microsoft/ASL-citizen-code

for both models, with over 15-point differences between the highest and lowest accuracy scores for each model (see Table 5. Thus, the large gender gap may partially be explained by this variation, as there are only a few participants of each gender in the test set.

Upon manual inspection, we find several characteristics of user videos that seem to vary between participants. Different participants have different background or lighting quality, and some participants mouth the word being signed while other participants do not. We also found instances of repetition, where the sign is repeated in the video, from P15, who is a female participant. There were also some instances of fingerspelling, where participants fingerspelled the sign before signing it. These and other individual differences may be contributors towards the gender disparity in performance.

The models tended to perform better on lighter skin tones than darker skin tones Despite darker skin tones making up most of the detected skin tones for videos in this dataset (see Figure 6), we found that models averaged better performance when the detected skin tone was lighter. We illustrate this phenomenon for the ST-GCN model in Figure 1. Although we found variations in accuracy between participants in the previous section, the skin tones were categorized at the video level. Thus, these results may not be impacted by the low sample size to the degree that the above results on gender are. However, it is possible that poor lighting in a video may make a participant's detected skin color darker than it actually is. Thus, lighting quality is a potential confounder for these results.

The model performed best on participants in their 20s and 60s The ASL Citizen test set was made up of 11 individuals in their 20s, 30s, 50s, and 60s. We found that, as with gender, model accuracy varied for different age ranges; the highest accuracy scores were achieved for participants in their 20s and 60s. This could be influenced by the proportion of participants in their 20s in the dataset.

5.2 Video-level differences

Performance decreases as the video length diverges from the average For each sign video in the ASL Citizen dataset, we calculated the number of standard deviations (SDs) from the mean for the video length compared to other videos of the same sign. We then placed these values into buckets: less than -2, -2 to -1, -1 to 0, 0 to 1, 1 to 2, and more than 2 SDs from the mean. We find that, on average, the videos farther away from the mean see decreased model performance compared to the videos closest to the mean. The results in full are in Table 1.

Std. devs from mean	I3D Top-1	ST-GCN Top-1
$n < -2$	0.38462	0.3846
$-2 \leq n < -1$	0.5551	0.4862
$-1 \leq n < 0$	0.648	0.5888
$0 \leq n < 1$	**0.6704**	**0.6449**
$1 \leq n < 2$	0.5727	0.5878
$n > 2$	0.3846	0.4668

Table 1: Top-1 accuracy scores for videos within a certain number of SDs away from the mean for videos of the same sign. For both models, videos with lengths closer to the mean yield better model performance.

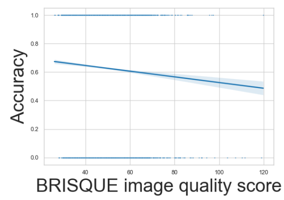

Figure 2: Association between BRISQUE image quality scores and accuracy. Higher BRISQUE scores indicate lower image quality, and vice versa. Thus, higher image quality appears to be associated with better model performance.

Performance decreases when video quality degrades In addition to video length, we studied the impact of video quality on model accuracy. Given that we were studying the quality of individual video frames without a reference image, we used the BRISQUE score (Mittal et al., 2012) to measure image quality of individual frames. Higher BRISQUE scores indicate lower quality, while lower BRISQUE scores indicate higher quality. We found that higher BRISQUE scores correlated negatively with Top-1 model performance for the I3D model, with a Spearman correlation of $\rho = -0.0367$ and a p-value of $p = 1.53x10^{-8}$. We show a scatterplot of these results in Figure 2, along with a linear regression line.

Dissimilarity between participant and seed signer signs negatively impacts model accuracy for the ST-GCN pose model The Frechét distance is often used as an evaluation metric for sign language generation, to study the similarity between generated signs and references (Hwang et al., 2024; Dong et al., 2024) (see § D for more details). In the ASL Citizen dataset, one of the participants is a paid ASL model who records videos for every sign, referred to as the “seed signer".

We studied whether dissimilarity between the participant and seed signer may have a negative impact on model accuracy. To do so, we used the pose models used as input to the ST-GCN model. Every .25 seconds, we measured the distance between the model pose and the participant's pose at that frame, studying the distance between left hands and right hands separately. We found no significant relationship between right hand or left hand distance from the seed signer for the I3D model, and for the ST-GCN model we found a significant negative Spearman correlation between distance from the seed signer and accuracy for the right hand ($\rho = -.0289$, $p = 0.001$). We plot these results, along with lines of best fit, in Figure 11.

When the average signing “speed" is closer to the sign-level average, performance is better In addition to video length, we were interested in studying the average distance between poses over consistent time intervals. We wanted to study how much movement on average occurred within these increments, i.e. the “speed" of sign production. We study this by calculating the pairwise Frechet distance between poses at each 0.25 second interval, with distance calculated between a pose and the pose .25s after, starting from the first frame. We again took this distance for the participants' right hand and left hand. We find that, on average, the farther away a participant's average signing speed is from the mean for that sign, the worse performance is, with especially high performance degradations 2 SDs or more from the mean. We show these results in Table 2.

SD from mean	**I3D** (LH)	**ST-GCN** (LH)	**I3D** (RH)	**ST-GCN** (RH)
$n < -2$	.4627	.2139	.5	.2375
$-2 \leq n < -1$	.6041	.5804	.6121	.5174
$-1 \leq n < 0$	.6503	.6426	.6438	.6351
$0 \leq n < 1$	.6244	.5813	.6423	.6145
$1 \leq n < 2$	.6164	.5261	.616	.5744
$n > 2$	.5711	0.4739	.5619	.5107

Table 2: Number of SDs away from the mean of the sign (in buckets) for the “speed" of signing, i.e. the average Frechet distance between poses every 0.25 seconds, for right hand and left hand. We find that, for both right hand and left hand, the performance degrades as the average “speed" of the sign production in a sign video deviates from the average for that particular sign.

5.3 Sign-level lexical features

The ASL-Lex annotations on this dataset allow us to not only conduct a dataset analysis, but also analyze model performance, and how sign-level features may impact model performance. Below, we present results for four sign-level features annotated in the ASL-Lex dataset: sign frequency, iconicity, phonological complexity, and neighborhood density. We find that several of these features are significantly correlated with model performance, which we discuss below.

Sign frequency, phonological complexity, and neighborhood density are negatively correlated with model accuracy As mentioned in § 3.2, sign frequency annotations were collected from ASL signers, who indicated the frequency of each sign in everyday conversation from 1 (least frequent) to 7 (most frequent). The ASL-Lex 2.0 dataset (Sehyr et al., 2021) also contains a new phonological complexity metric. Using 7 different categories of complexity, scores were calculated by assigning a 0 or 1 to each category (depending on whether that category was present) and adding them together, for a maximum possible scores of 7 (most complex) and a minimum possible score of 0. The highest complexity score in the dataset was a 6. Neighborhood density was calculated based on the number of signs that shared all, or all but one, phonological features with the sign. Intuitively, we expected negative associations with phonological complexity and accuracy as well as neighborhood density and accuracy, and indeed found significant negative correlations ($\rho = -0.0618$, $p = 0.005$ for phonological complexity and $rho = -0.0584$, $p = 0.01$ for neighborhood density). However, we also found a significant negative association between sign frequency and model accuracy, with a correlation of $\rho = -0.057$ and $p = 0.011$. We are unsure of the cause of this negative association, and encourage future researchers to explore this relationship further.

There is no significant correlation between iconicity and model accuracy As mentioned in § 3.2, sign iconicity ratings were also collected for the ASL-Lex dataset, using hearing individuals' judgments regarding how much the sign looks like its English meaning. The hearing individuals assigned ratings from 1 (not iconic at all) to 7 (very iconic). We found a very slight positive

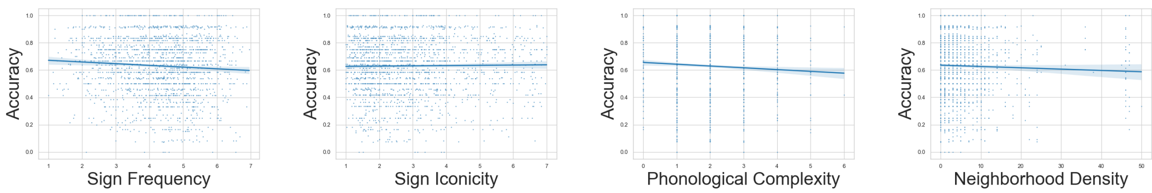

Figure 3: The relationships between sign frequency (left), sign iconicity (center left), phonological complexity (center right), and neighborhood density (right) and top 1 accuracy for the ST-GCN model. We find that sign frequency, phonological complexity, and neighborhood density are all significantly negatively correlated with model accuracy ($p < 0.05$) when calculating Spearman's rank correlation. However, despite a slight positive correlation between iconicity and accuracy, the p-value is not significant.

correlation between sign iconicity and model accuracy ($\rho = 0.044$), which was not significant ($p = 0.8424$). Thus, we conclude that visual similarity to the English word appears not to affect the model's ability to recognize a sign.

5.4 Which features have the greatest impact on model accuracy?

After looking at the impacts of lexical, demographic, and video features on model accuracy, we were interested in studying which features are the most impactful. As such, we study the mutual information between each feature and the Top-1 accuracy for the I3D and ST-GCN models. We study a total of 19 features, where some relate to participant demographics (e.g. age and gender), others relate to the sign lexical features (e.g. sign iconicity), and the rest are characteristics of individual videos (e.g. BRISQUE score and Frechet distances). We find that the five most impactful features are characteristics of the videos themselves (BRISQUE, Frechet from seed signer, and absolute SD of "signing speed"), with BRISQUE video quality scores showing the highest mutual information scores. Out of the lexical features, sign iconicity has the highest mutual information, and out of the demographic features, ASL level has the highest mutual information. The results in full are in Table 6 in Appendix H.

6 Can we mitigate disparate impacts while maintaining higher performance for dictionary retrieval?

6.1 Training on single-gender subsets

We first try to address the gender gap by training on participants of each gender in isolation, and testing performance on male and female participants separately and together. When doing this, we do find a slight difference between the performance gaps for model trained on male-only and female-only subsets. For the model trained on the male-only subset, the Top-1 accuracy for male subjects was .292, and the Top-1 accuracy was .168. For the model trained on the female-only subset, the Top-1 accuracy for male subjects was .291, and the Top-1 accuracy for female subjects was .206. Thus, the model trained only on female subjects had a smaller gap, and higher accuracy parity, between male and female subjects than the model trained on only male subjects. However, both models had low performance overall, so the Top-1 accuracy parity for subjects of different genders (calculated by dividing the female accuracy by the male accuracy) comes out to .7571 for the model trained on all subjects compared to .7079 for the model trained on only female subjects. The model trained on only male subjects has the lowest accuracy parity, at .5746. We show these results in full in Table 7 in Appendix I.

6.2 Training label shift

In addition to training on single-gender subsets, we experiment with a label-shift approach to debiasing. Because ISLR is a multiclass problem, we experiment with the reduction-to-binary approach for debiasing multi-class classification tasks proposed by Alabdulmohsin et al. (2022). We run the label-shift algorithm and train the ST-GCN model on the debiased labels for 25 epochs, and compare the performance of the debiased model to the ST-GCN model without debiasing, which we also train for 25 epochs. We find that the model trained on regular labels actually has a *higher* ratio for female to male accuracy than the debiased model: .7476 for the baseline model, and .7052 for the debiased model. We show these results in full in Table 8 in Appendix J.

Model	Overall			Female participants			Male participants			Parity (Top-1)
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	
ST-GCN	.5238	.7665	.8295	.4406	.6886	.7665	.6236	.8601	.9374	.7065
ST-GCN (VL)	**.5488**	.7923	.8515	**.4666**	.7200	.7941	**.6476**	.8791	.9205	.7205
ST-GCN (VL, fem.)	.5395	**.7926**	**.8538**	.4621	**.7202**	**.7974**	.63	**.8795**	**.9216**	**.7334**

Table 3: Performance of ST-GCN baseline against models that use the resampling strategies discussed in 6.3. We find that both resampling strategies improve accuracy and gender parity over the baseline, and resampling based on video length from only female participants improves gender parity the most.

6.3 Weighted resampling

Although there are large performance discrepancies, on average, between videos from participants of different demographics, particularly gender, based on the results from Table 6, other features are much more heavily tied to model accuracy. Thus, it is likely that these features (in particular, features at the video level) may influence results. But what happens if the impact of videos with potentially-noisy features is reduced during training? We experiment with weighted resampling, where certain features are more likely to be resampled during model training if they have values shown to produce good results. For instance, we show in Table 1 that video lengths closer to the mean for each sign produce higher accuracy scores for both baselines. Thus, we experiment with assigning probabilities for resampling videos in the training set, where the probability of resampling a video is calculated as follows based on the number of SDs from the mean. We explain how we calculate this probability, and present results, for each variable we study in the paragraphs below.

Video length We first experiment with calculating the resampling probability based on video length. Given that videos closer to the mean produced higher accuracy scores, we wanted to resample these videos at a higher rate to reduce training noise. We calculate the probability of resampling as follows, where $l_i(s)$ refers to the length of video i for sign s, mu_s refers to the mean video length of videos depicting sign s, and σ_s refers to the SD for video lengths of videos depicting sign s:

$$P(resample) = \frac{1}{2^{\frac{l_i(s)-\mu_s}{\sigma_s}}} \quad (1)$$

We show the results for this approach in Table 3, represented by the ST-GCN (VL) model. We find that this approach improves upon the baseline ST-GCN model by at least 2 percentage points for all accuracy metrics, and improves gender parity for Top-1 accuracy by 1.4%.

Video length for female participants We then experiment with the exact same resampling process described above, based on number of standard deviations from the mean for video length, but only resample videos from female participants. Because training on an all-female subset yielded a higher test accuracy for female subjects than an all-male subset (Table 7), we wanted to investigate whether restricting our resampled data to female participants improves the gender performance gap. We show these results in Table 3, under the baseline STGCN (VL, fem.). We find that this approach exceeds calculating the resampling probability using video length for participants of all genders for Top-5 and Top-10 accuracy. We also find that this baseline achieves the highest gender parity of all of the baselines, at 2.69% higher than the baseline. Thus, we find evidence that resampling based on video length standard deviations, but only videos from female participants (the group with the lower model accuracy scores), improves gender parity the most over the baseline model.

7 Conclusion

In this work, we address a gap in sign language processing research by studying the biases and performance gaps in sign language resources, and experimenting with strategies to mitigate these biases. We specifically focus on the ASL Citizen dataset, which is the only large-scale crowdsourced ISR dataset. We find performance gaps related to skin tone, participant age, and gender. However, we find that video level features, such as the video quality, signing "speed", and video length, appear to be the most influential features for determining model accuracy. We find that selectively resampling data with video lengths closer to the mean improves overall performance. We also find that doing this resampling strategy for *only* the group with lower model performance (female, when comparing genders) appears to improve the gender parity for model performance.

8 Limitations

While in this work we find and document performance gaps between participants of different demographics such as age and gender, because of the differences between individual participants that we detail above (see Table 5), and the number of participants in the test set (11), it is unclear how much of these differences are due to age or to other underlying factors.

Another limitation is that we focus on a single dataset. This is due in part to the fact that this is the only large-scale crowdsourced dataset for isolated sign language recognition with demographic labels. However, as more crowdsourced sign language resources become available, it is critical that these analyses are repeated on these datasets to assess the generalizability of our results.

9 Ethical Implications

In our analysis of participant demographics, and accompanying features, for the ASL Citizen dataset, we present some characteristics of the dataset that vary between demographics. For instance, we discuss our findings that male participants and older participants typically record longer videos. It is important to emphasize that these findings should not be generalized to all ASL signers, and that they should instead be used to study the characteristics of this dataset in particular.

We also note that participants who chose to denote their demographic information (which was optional) consented for this information to be anonymously released as part of the dataset. No identifiable information about the participants will be released with the publication of this paper; rather, anonymous participant IDs will be accompanied with their demographics.

References

Ibrahim Mansour I Alabdulmohsin, Jessica Schrouff, and Sanmi Koyejo. 2022. A reduction to binary approach for debiasing multiclass datasets. In *NeurIPS 2022*.

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

Carole Sue Bailey, Kathy Dolby, and Hilda Marian Campbell. 2002. *The Canadian dictionary of ASL*. University of Alberta.

Patrick Boudreault and Rachel I Mayberry. 2006. Grammatical processing in american sign language: Age of first-language acquisition effects in relation to syntactic structure. *Language and cognitive processes*, 21(5):608–635.

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*, 49:784–801.

Kearsy Cormier, Adam Schembri, and Bencie Woll. 2010. Diversity across sign languages and spoken languages: Implications for language universals. *Lingua*, 120(12):2664–2667.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2024. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36.

Lu Dong, Lipisha Chaudhary, Fei Xu, Xiao Wang, Mason Lary, and Ifeoma Nwogu. 2024. Signavatar: Sign language 3d motion reconstruction and generation. *Preprint*, arXiv:2405.07974.

Thomas Eiter, Heikki Mannila, and Christian Doppler Labor für Expertensyteme. 1994. Computing discrete fréchet distance.

Karen Emmorey, Ursula Bellugi, Angela Friederici, and Petra Horn. 1995. Effects of age of acquisition on grammatical sensitivity: Evidence from on-line and off-line tasks. *Applied psycholinguistics*, 16(1):1–23.

Eui Jun Hwang, Huije Lee, and Jong C. Park. 2024. Autoregressive sign language production: A gloss-free approach with discrete representations. *Preprint*, arXiv:2309.12179.

Lee Kezar, Jesse Thomason, and Zed Sehyr. 2023. Improving sign recognition with phonology. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737, Dubrovnik, Croatia. Association for Computational Linguistics.

Annelies Kusters. 2010. Deaf utopias? reviewing the sociocultural literature on the world's "martha's vineyard situations". *Journal of deaf studies and deaf education*, 15(1):3–16.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. 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.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708.

Carol Padden. 2010. Sign language geography. *Deaf around the world: The impact of language*, pages 19–37.

René Alejandro Rejón Pina and Chenglong Ma. Classification algorithm for skin color (casco): A new tool to measure skin color in social science research. *Social Science Quarterly*, n/a(n/a).

Noha Sarhan, Christian Wilms, Vanessa Closius, Ulf Brefeld, and Simone Frintrop. 2023. Hands in focus: Sign language recognition via top-down attention.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. 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.

William C Stokoe, Dorothy C Casterline, and Carl G Croneberg. 1965. *A dictionary of American Sign Language on linguistic principles*. Gallaudet College Press, Washington, DC.

Andrea Toliver-Smith and Betholyn Gentry. 2017. Investigating black asl: A systematic review. *American Annals of the Deaf*, 161(5):560–570.

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

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. 2005. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31-September 2, 2005. Proceedings 27*, pages 401–408. Springer.

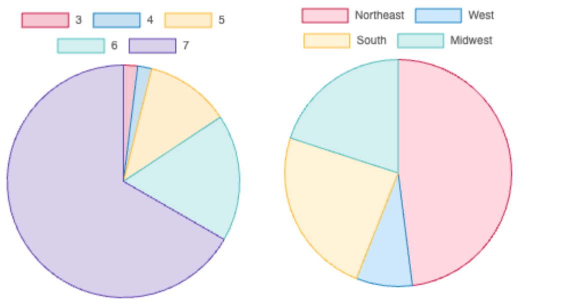

Figure 4: Distribution of ASL levels (left) and regions (right) of participants for the ASL Citizen dataset.

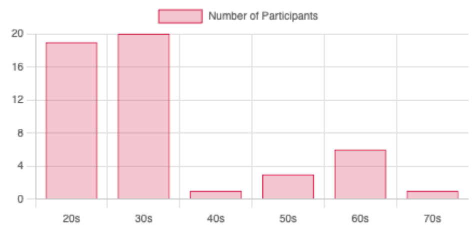

Figure 5: Age ranges of participants in the ASL Citizen dataset. Participants are skewed mostly towards their 20s and 30s, with a lesser skew towards participants in their 60s.

A Participant Demographics

Here, we plot the demographic information discussed in 3.1. Note that providing demographic information was optional, so these numbers will not always add up to the total number of participants (52).

In Figure 4, we plot the distribution of ASL levels and regions associated with the participants in the ASL Citizen dataset. We find that most participants are at an ASL level of 6 of 7, with only one participant each at level 3 or 4. A plurality of participants are from the Northeast, almost half. The West contains the fewest participants.

In Figure 5, we plot the distribution of participants' ages. We find that participants are mostly skewed towards younger adults (20s and 30s) but that there is also a slight skew towards contestants in their 60s. Contestants in their 20s, 30s, 40s, 50s, 60s, and 70s are represented in the dataset, but contestants in their 40s and 70s are not represented in the test set.

In Figure 6, we plot the distribution of skin tones in the dataset when frames are set as color images and black-and-white images. We include black-and-white images because we found that, when an image type was not set, the model detected the images as black-and-white images in the majority of cases. One notable finding is that the skin color model detected lighter skin tones more frequently when the images were set to black-and-white than when they were set to color images. This indicates possible unreliability of the skin color detection; it is possible, for instance, that when the images are set to color, the system classifies the skin colors as darker than they actually are.

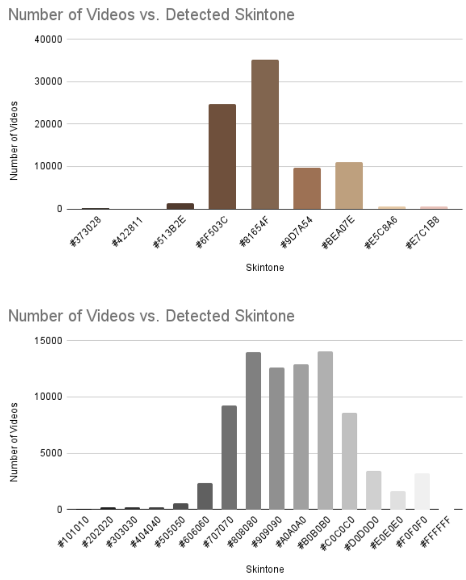

Figure 6: Frequency of detected skin tones of participants in videos when the video frames were set manually to color images (left) and black and white images (right)

B Video Length Distributions

In Figure 7, we find that video lengths have a skewed distribution, where the average video length is higher than the median. In other words, video lengths lower than the mean are more common and vice versa, and there is a long tail to the right. After watching participants' videos, we suspect that this difference in video length is a result of some participants having a tendency to pause for multiple seconds at the beginning of end of their recording. This happens especially often with the first couple of videos that people record.

We also find that female participants have, on average, shorter videos related to their signs than male participants. For each sign video, we calculated the mean and standard deviation for all videos with that sign. We then calculated how many standard deviations those movies were away from the mean.

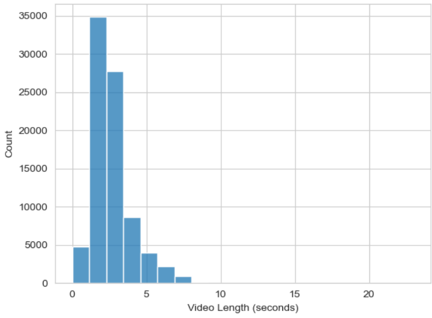

Figure 7: Distribution of video lengths for all sign videos in the ASL Citizen dataset. The distribution is skewed towards the right, with a long tail on the right.

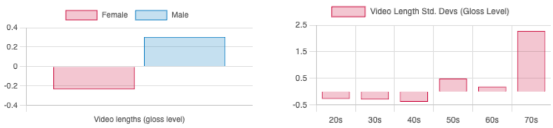

Figure 8: Average number of standard deviations away from the mean at the sign level for male and female participants (left) and participants in their 20s, 30s, 40s, 50s, 60s, and 70s. Relative to other videos of the same sign, women tend to record shorter videos, and older participants tend to record longer videos.

C Lexical Feature Distribution

In addition to getting demographic and video features, we used the ASL-Lex (Caselli et al., 2017) annotations to analyze lexical features in the ASL Citizen dataset. We found that, for sign frequency and iconicity, the distributions are very similar to those in the ASL-Lex dataset. The distributions of both datasets are plotted side-by-side for frequency and iconicity, respectively, in Figures 9 and 10.

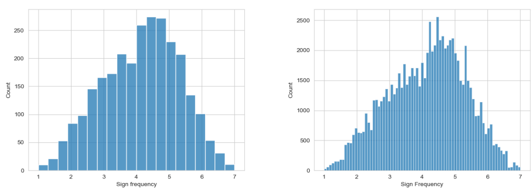

Figure 9: Distributions of labeled sign frequencies for each of the 2731 signs from the ASL-Lex dataset (left) and all of the sign videos in the ASL Citizen dataset (right). The distributions are very similar, indicating that users chosen signs of certain frequencies at a similar rate to how they are distributed in the ASL-Lex dataset.

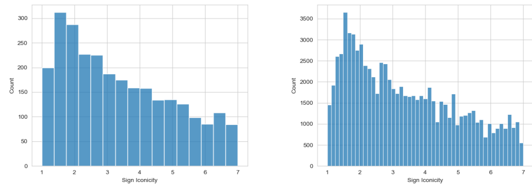

Figure 10: Distribution of sign iconicities in the ASL-Lex dataset (left) and the sign videos recorded in the ASL Citizen dataset (right). Like the sign frequencies, the iconicities in the ASL Citizen videos are distributed similarly to their distribution in the ASL-Lex dataset.

Age range	# in test	I3D Top-1	ST-GCN Top-1
20s	2	.6697	.6076
30s	3	.5689	.5336
40s	0	–	–
50s	2	.549	.5658
60s	3	**.7016**	**.6421**
70s	0	–	–

Table 4: Average accuracy scores for participants of each age range in the test set. There were no participants in their 40s or 70s in the test set, and one participant did not specify their age. We find the highest performance in both models occurs for participants in their 20s and 60s.

D Frechét Distance

The Frechét distance, used as a similarity metric between curves, and is commonly described in the following manner:

> *A* man is walking a dog on a leash: the man can move on one curve, the dog on the other; both may vary their speed, but backtracking is not allowed. What is the length of the shortest leash that is sufficient for traversing both curves?
> - (Eiter et al., 1994)

E Accuracies for different age ranges

In Table 4, we show the Top-1 accuracy scores for the I3D and ST-GCN model for participants of different ages. We find the highest scores occur for participants in their 20s and 30s, with the third highest scores occuring for participants in their 60s. Participants in their 40s and 70s were not represented in the test set.

F Model accuracies for each participant in the test set

In Table 5, we report the accuracy scores for the baseline ST-GCN model on the participants in the

Participant ID	I3D Top-1	ST-GCN Top-1
P6	0.5456	0.4387
P9	0.6586	0.5663
P15	0.4653	0.5757
P17	0.6183	0.4997
P18	0.7065	0.5727
P22	0.5562	0.4671
P35	0.7204	0.7153
P42	0.6041	0.6949
P47	0.7471	0.7886
P48	0.6882	0.6652
P49	0.6327	0.556

Table 5: Model top-1 accuracy scores on the set of videos recorded by each participant in the test set. For both models, there is high variation between participants, with scores ranging from 0.4653 to 0.7204 (I3D) and 0.4387 to 0.7886 (ST-GCN).

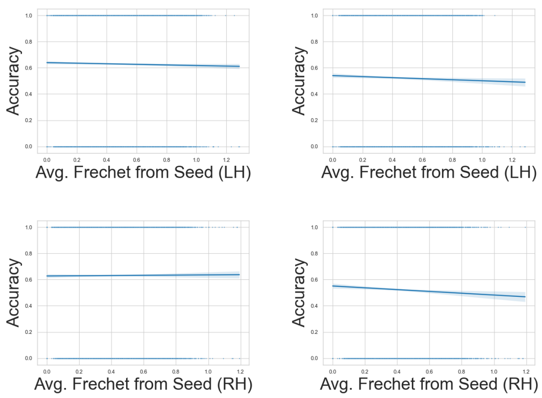

Figure 11: The Frechet distance from the seed (model) signer vs. top-1 accuracy for the I3D model (top) and ST-GCN model (bottom), with the distance between left hands on the left and the distance between right hands on the right.

test set of the ASL Citizen dataset. We find differences of over 20 points between participant averages for both models. P6, P9, P15, P17, P18, and P22 disclosed that they are female, while the other participants disclosed that they are male.

G Frechet distance from seed signer

In Figure 11, we plot the Top-1 accuracies for the I3D and ST-GCN model as a function of the Frechet distance from the seed signer for each sign video (where the seed signer is a recruited ASL model for the ASL Citizen dataset). We find a significant negative correlation between Frechet distance from the seed signer and Top-1 accuracy for the ST-GCN pose model, but no significant correlations for the I3D model.

Feature	**Mut. Info** (ST-GCN)	**Mut. Info** (I3D)
BRISQUE	0.6920	0.6617
Avg. Frechet from seed (RH)	0.6444	0.6217
Abs. Avg. Frechet SD (RH)	0.6390	0.6090
Abs. avg. Frechet SD (LH)	0.6285	0.5641
Avg. Frechet from seed (RH)	0.5889	0.5403
Sign Iconicity	0.0757	0.0508
Sign Frequency	0.0619	0.0440
Abs. avg. Video Length SD	0.0293	0.0399
ASL Level	0.0048	0.0020
Region	0.0034	0.0002
Neighborhood Density	0.0032	0.0026
Number Of Morphemes	0.0026	0.0012
Phonological Complexity	0.0013	0.0006
Lexical Class	0.0007	0.0008
Iconicity Type	0.0002	0.0002
Gender	0	0.0034
Age	0	0.01107
Bounding Box Area (RH)	0	0
Bounding Box Area (LH)	0	0

Table 6: Mutual information for each of the features above and the Top-1 accuracy for the ST-GCN and I3D models, respectively. For both models, the BRISQUE score, average Frechet distance from the model (right hand and left hand) and the absolute value of the number of SDs of the average Frechet distance between frames are the top three features, with the other features far behind. This seemingly indicates that video-level features are the biggest indicator of model accuracy.

H Mutual Information Results

In Table 6, we present the mutual information results in full for each studied variable. We study 19 variables total, spanning demographics, sign lexical features, and video-level features, and calculate the mutual information between each feature and the Top-1 accuracy. We find the highest levels of mutual information to occur for video-level features, suggesting features of individual videos are more impactful for model accuracy than demographic characteristics of the participants. Out of the demographic characteristics, the ASL level of the participant appears to be the most influential with respect to accuracy.

I Results for models trained on single-gender subsets

Here, we report the model results for the ST-GCN model trained on single-gender subsets, comparing models trained on all-male and all-female subsets to the model trained on all of the training data. In Table 7, we report the Top-1, Top-5, and Top-10 accuracy scores for each model.

	Trained on female subjects			Trained on male subjects			Trained on all subjects		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
All	.244	.479	.581	.224	.434	.527	.594	.828	.881
Male	.291	.548	.653	.292	.538	.639	.684	.902	.939
Female	.206	.421	.521	.168	.347	.433	.520	.767	.833

Table 7: Performances for ST-GCN model trained on only male subjects, only female subjects, and all subjects, respectively. We find that the model trained on only female subjects has the lowest performance gap between male and female subjects in the test set, but the ratio of female accuracy to male accuracy is highest for the model trained on all subjects.

J Results for model trained on debiased labels

We report the results for a model trained for 25 epochs on training labels that were debiased using the reduction-to-binary techniques proposed by Alabdulmohsin et al. (2022). We find that the model trained on regular labels actually had a higher accuracy parity score (ratio of female accuracy to male accuracy) than the model trained on debiased labels. We show the Top-1, Top-5, and Top-10 results for each model in Table 8.

	ST-GCN			**ST-GCN (debiased)**		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
All	.5323	.7997	.8622	.4821	.7576	.8265
Male	.6173	.8781	.9254	.5746	.8493	.9014
Female	.4615	.7343	.8096	.4052	.6811	.7641

Table 8: Performances for ST-GCN model trained on regular training labels (left) and debiased training labels (right). We find that the accuracy parity, calculated as the ratio of female to male accuracy, is higher for the model trained on regular training labels than the debiased model.