Investigating Dictionary Expansion for Video-based Sign Language Dictionaries

Aashaka Desai
University of Washington
aashakad@uw.edu

Daniela Massiceti
Microsoft Research
dmassiceti@microsoft.com

Richard Ladner
University of Washington
ladner@cs.washington.edu

Hal Daumé III
University of Maryland
hal3@umd.edu

Danielle Bragg
Microsoft Research
dbragg@microsoft.com

Alex X. Lu
Microsoft Research
lualex@microsoft.com

Abstract

Like most languages, sign languages evolve over time. It is important that sign language dictionaries' vocabularies are updated over time to reflect these changes, such as by adding new signs. However, most dictionary retrieval methods based upon machine learning models only work with fixed vocabularies, and it is unclear how they might support dictionary expansion without retraining. In this work, we explore the feasibility of dictionary expansion for sign language dictionaries using a simple representation-based method. We explore a variety of dictionary expansion scenarios, e.g., varying number of signs added as well as amount of data for these newly added signs. Through our results, we show how performance varies significantly across different scenarios, many of which are reflective of real-world data challenges. Our findings offer implications for the development & maintenance of video-based sign language dictionaries, and highlight directions for future research on dictionary expansion.

1 Introduction

Dictionaries are important resources for sign languages, offering a way to document the many different signs comprising the intricate vocabularies of these languages. In addition to documentation, dictionaries are particularly helpful to novices or language learners, allowing them to easily look up signs they are unfamiliar with. It is crucial that these dictionaries have mechanisms to stay up-to-date with changes in the language, such as the creation and adoption of new signs by signing communities. Here we consider the problem of dictionary expansion, where the vocabulary of an established sign language dictionary is updated to incorporate new signs.

A key factor that makes sign language dictionaries (and thereby their expansion) unique is that sign languages are visual-manual languages. This means that sign language dictionaries typically use video entries to represent each sign in their vocabulary, and need video-based approaches for users to look up signs and query the dictionary. In a video-based dictionary, a user demonstrates a sign to a camera, and the dictionary returns a ranked list of entries from the dictionary that might correspond to that sign. Recent advancements in sign language datasets and modeling show exciting promise in making video-based dictionary retrieval a reality (Hassan et al., 2025). For example, many state-of-the-art isolated sign language recognition models (which are trained to recognize single signs from vocabularies of 2k+ signs) currently achieve recall@10 above 90% (i.e., queried sign is in top-10 results) – this high performance makes it possible to deploy these models for dictionary retrieval.

However, a key limitation of almost all such modeling approaches is that it is unclear how they might support dictionary expansion. The technologies underlying most video-based retrieval methods are often machine learning classifier models trained on a fixed vocabulary, where each sign corresponds to a label in a fixed label set. Incorporating new vocabulary into these models would typically require retraining them from scratch – however, this is can be computationally expensive and relies on the availability of a sufficient amount of high-quality training data, which may be difficult in practice. It also means that third-parties can't adapt existing dictionaries to their needs (e.g., representing local dialects or specialized lexicons). An ideal approach would support vocabulary expansion without retraining the existing model, but this class of approach remains largely unexplored.

In this work, we investigate the feasibility of dictionary expansion without retraining for video-based sign language dictionaries. To overcome fixed vocabulary constraints encountered with current classification approaches, we propose a simple method that instead uses the representations

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22826–22841
November 4-9, 2025

learned by deep learning models for dictionary retrieval, rather than just their final classification layer. Doing so allows us to work with unseen signs through nearest neighbor-type approaches.

We explore how this type of approach performs across a range of circumstances, simulating dictionary expansion in ideal settings as well as more realistic and challenging scenarios. We separately evaluate the performance of expanded models on core signs (those in the original dictionary) as well as newly added signs. We find that our method performs well when adding a small number of new signs ($\sim 1-100$), each with many examples (~ 15 per sign), to a dictionary with an existing large vocabulary – maintaining performance compared to the pre-expansion for both new and old signs. However, performance degrades substantially when we constrain the number of examples for new signs, or attempt to add many new signs to the vocabulary, which we argue are likely scenarios for real-world dictionary expansion. Our work is the first to systematically explore computational challenges dictionary expansion, providing directions for future methods development grounded in real-world needs of sign language dictionaries and their users. These opportunities for future research not only offer a space for technical discovery, but also further the exploration of sign languages as languages (Yin et al., 2021; Desai et al., 2024).

2 Background

Sign Languages and Dictionaries. Sign languages are visual languages, with each sign composed of distinct handshapes, movements, non-manual markers, and other phonological features. There are over 300 sign languages in the world. Our work focuses on American Sign Language (ASL), which is the most common in North America. ASL is culturally significant to Deaf community in the continent, and is also learned by many as a second language (Looney and Lusin, 2019). ASL dictionaries are a valuable resource for this group.

Part of what makes dictionary retrieval in sign language challenging is that two different signs might share nearly all the same phonological features (which can cause them to look visually similar) but have very different meanings. For example in ASL, SORRY and PLEASE only differ in handshape, SUNRISE and SUNSET only differ in movement (examples of *minimal pairs*). Dictionary retrieval methods need to be robust to such dense lexical neighborhoods.

Existing sign language dictionaries support querying in English, through a set of descriptive features, or by video demonstration. Searching by English word or gloss (e.g., SigningSavvy, LifePrint) is a valuable approach for those looking to learn a sign from an English word (i.e., English-to-ASL translation). However, users cannot leverage these dictionaries to look up the meaning of an unfamiliar sign (i.e., ASL-to-English). Allowing users to navigate dictionary search directly in sign languages is complicated as sign languages do not have standardized written forms. One approach is feature-based search (e.g., HandSpeak & (Bragg et al., 2015)) that allows users to search by describing features of a sign (e.g., handshapes, movements). Video-based search allows users to search by demonstrating a sign (returning all dictionary entries that may match the search video). Video-based dictionaries also allow native signers to navigate dictionaries in their primary language– their development is an important avenue of work.

Video-based Dictionaries and Sign Language Recognition. Recent advancements in datasets and modeling have changed the landscape of sign language recognition research. Consider the release of large-scale isolated sign datasets for ASL (e.g., ASL Citizen (Desai et al., 2023), Semlex (Kezar et al., 2023), PopSign (Starner et al., 2023), WLASL (Li et al., 2020), ASLLVD (Athitsos et al., 2008)), each of which contains large vocabularies (over 2000 signs) and multiple samples per sign from different contributors. It is now feasible to use deep learning to train models for single sign recognition, and thus build video-based sign language dictionaries (Hassan et al., 2025). While many have worked to surpass dictionary retrieval performance on these benchmark datasets (e.g., (Gueuwou et al., 2024; Wong et al., 2025)), most of these methods are limited by approaching sign language dictionary lookup as a classification problem – thus making it hard to adapt to changing vocabularies. In this work, we propose a basic method for leveraging these classifiers as dictionaries expand their vocabulary, and demonstrate where this method succeeds and where challenges still remain.

Prior Work on Dictionary Expansion. While some work has explored how sign language recognition models might generalize to unseen data (such as different datasets and different languages, e.g., (Wong et al., 2025)), only few have focused on the

task of dictionary expansion. Huamani-Malca and Bejarano (2023) compares different incremental learning approaches for Peruvian Sign Language dictionaries, and Gupta (2022) explores the same for Indian Sign Language. Both works consider small vocabulary contexts (under 100 signs) under ideal data conditions (multiple examples of the new vocabulary), which might not reflect real-world conditions with larger vocabularies, larger expansion demands and variable amount of examples. While some researchers have explored sign language recognition with data constraints (e.g., limited data for all signs (Bohacek and Hrúz, 2023; Vandendriessche et al., 2025) or unequal data across signs (Kezar et al., 2023) or demographics (Atwell et al., 2024)), they again focus on fixed vocabulary contexts. In this work, we look at the combination of the two contexts i.e., expanding vocabularies *and* data-constrained recognition as this is the most reflective of real-world use of sign language recognition for dictionaries.

3 Experimental Setup

3.1 Task Definition

We consider the task of video-based search in sign language dictionaries. For dictionary retrieval, we aim to map user-submitted video queries $\mathbf{x}$ to glosses[1] y. We assume have access to a pre-trained classifier f that maps videos to a fixed number of glosses N: $f : \mathbf{x} \mapsto \{1, \ldots, N\}$, outputting probability for each gloss in the vocabulary of the dictionary. These probabilities can be ranked and used to retrieve a list of likely matching signs for the user (e.g. the highest probability gloss is the top-ranked retrieval result).

After training this classifier, we discover that M new signs have become commonplace and we wish to add them to our classifier, to yield $f' : \mathbf{x} \mapsto \{1, \ldots, N + M\}$. We wish to expand the dictionary without having to retrain f, as computational resources or time for retraining are limited. To facilitate the expansion of f to f', we assume that we have access to varying number (m) of videos for each of the M new signs (e.g., through crowdsourcing contributions).

Given this new model f', we care about how well it does on *both* the original N signs (i.e., we do not want its performance to degrade dramatically on the old signs in comparison to f), *and* the new M signs (i.e., we want it to correctly recognize the new signs).

3.2 Our Dictionary Expansion Approach

In this work, we propose a method to support dynamic vocabulary expansion of f to f' using learned feature representations. Deep learning models, particularly those trained for classification like f, learn rich intermediate representations that capture semantic and visual structure in the data. These representations (typically extracted from the penultimate layer) can often generalize beyond the specific labels used during training. We adopt a similarity-based retrieval approach using these feature representations. Given a query video of a new sign, we extract its feature representation using f and retrieve the most similar video from a database using a nearest-neighbor search in the feature space. This database consists of feature representations from both the core vocabulary and an expansion set of new signs.

In practice, to expand the dictionary, an administrator would simply need to process new sign videos through the trained (frozen) classifier to extract their feature representations and add them to the retrieval database. No additional training or fine-tuning is required. This provides a lightweight, scalable solution for incorporating new signs into a fixed-vocabulary dictionary by leveraging the generalization capacity of learned embeddings and similarity-based retrieval. Below, we discuss how we simulate a variety of realistic dictionary expansion scenarios, carefully slicing a large ASL dictionary dataset into appropriate subsets.

3.3 Data Setup

We use the ASL Citizen dataset (Desai et al., 2023) as it was collected to support research and development of ASL dictionaries. This dataset contains about 84k videos of 52 d/Deaf and hard-of-hearing contributors fluent in ASL performing single signs. It also contains videos of the seed signer (a highly proficient ASL signer) whose videos were used to prompt data collection. Each video is labeled as 1 of 2731 glosses. The dataset provides standardized train, validation, and test splits by contributor, allowing for testing of model generalization onto unseen users, mimicking real-world use.

To simulate dictionary expansion, we split the ASL Citizen dataset by gloss into two non-overlapping subsets: a "core vocabulary" and an "unseen vocabulary". The core vocabulary is in-

[1]English translations for isolated signs

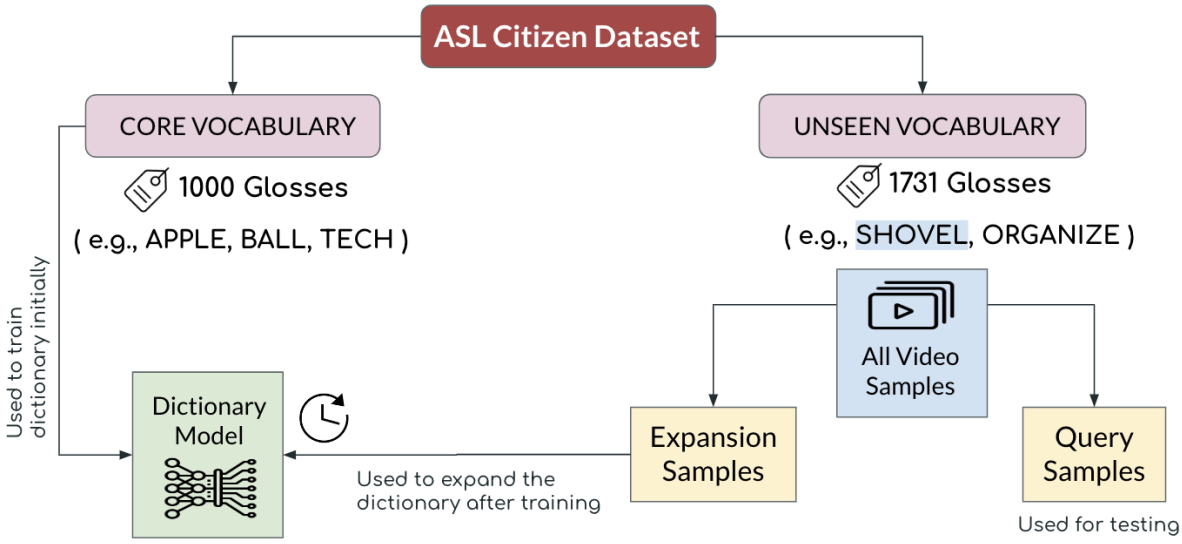

Figure 1: Schematic of Data Setup for Dictionary Expansion

tended to simulate entries that are initially present in a dictionary. We assume that the dictionary administrators have already trained a classifier, using the training dataset associated with the core vocabulary. Held-out test recordings of the core vocabulary can then be used for the purpose of simulating a dictionary query.

The unseen vocabulary are glosses that we reserve to test dictionary expansion under various scenarios. The unseen vocabulary is intended to simulate new signs that will be added to the dictionary at a future date, after machine learning models have already been trained on the core vocabulary. We further split the unseen vocabulary into expansion and query samples. We consider the expansion dataset to be examples the dictionary administrator is in possession of, and can use to expand their machine learning model. The query dataset is then intended to simulate users querying the expanded model, and used to test performance of the expanded models.

Mindful of how the composition of the vocabulary might influence task difficulty, we randomly generated three different core-unseen splits for our dictionary expansion experiments[2]. We repeat our experiments with each split and report average performance and standard deviation across these runs. We note that we did not segment the core and unseen vocabulary by contributor, meaning that signers in the unseen vocabulary may have already been seen in the training dataset of the core vocabulary. We consider this to still be reflective of real-world usage as crowdsourced dictionaries (like ASL Citizen) often sustain repeat contributors.

[2]Splits can be found here: https://github.com/aashakadesai/asl-dict-expansion-splits/

3.4 Dictionary Settings

To understand how the difficulty of dictionary expansion interacts with various data acquisition challenges in the real-world, we begin by declaring a base setting for dictionary expansion. We then vary parameters of this base set-up in controlled ways to simulate these various data challenges.

Base setup In this set-up, we assume we begin with a reasonably large core vocabulary with many examples per sign. We then simulate a scenario where only one new sign is added to the vocabulary, with ample expansion samples for this sign. In our case, our core vocabulary consists of 1,000 signs and total of 14,691 training videos (i.e., $\sim$15 videos per sign, each performed by different contributors). We then expand the vocabulary by 1 sign (going from 1000 to 1,001 signs) using all available expansion samples for that sign (i.e., $\sim$15 videos per sign, each performed by different contributors). We repeat this procedure over all 1,731 signs in our unseen vocabulary and report averaged metrics (see section 3.6), testing on a total of 20,902 videos.

Adding a Sign with a Single Example Gathering multiple samples for a new sign involves considerable effort. For example, for dictionaries that rely upon crowdsourcing, newly added vocabulary may be recently invented signs that have not fully disseminated through the community yet. In these

cases, only a small number of contributors may initially provide videos of the sign. To simulate this scenario, our core vocabulary consists of 1,000 signs (with ∼15 videos per sign, each performed by different contributors). We then expand the vocabulary by 1 sign (going from 1000 to 1,001 signs) using *only one expansion sample for that sign* (in contrast to ∼15 samples used in the base set up). We consider two settings for how the single sample per sign is obtained: In the first setting, we assume that signs are organically contributed by different contributors. We simulate this by randomly sampling a video from the unseen expansion set. In the second setting, we assume a long-term contributor (i.e., someone who has likely recorded examples of most of the core vocabulary) contributes a new sign to the dictionary. We simulate this by using a specific contributor's video to represent core and added signs throughout the dictionary.

Adding Multiple Signs with Multiple Examples In the real world, it is likely that multiple signs will be added to a dictionary. Dictionaries are expected to grow iteratively with contributions over time. This larger expansion context might have a more significant impact on dictionary performance for the core vocabulary. To simulate this scenario, our core vocabulary again consists of 1,000 signs and newly added signs have 15 expansion samples each. Unlike the previous two settings, instead of adding one sign to the dictionary, we sample *a set of signs* (n = 100, 500, 750, or 1000) to add to the dictionary. This allows us to test expanding the vocabulary by 10%, 50%, 75% and 100% respectively. For each stage of expansion, we randomly sample three different sets of signs from the unseen split to be added to the dictionary, as differently composed sets might impact the difficulty of the task uniquely. We report the average performance across these splits as well as corresponding changes in performance of core vocabulary after expansion.

Adding Multiple Signs with Few Examples Next, we were curious about the interaction between adding multiple signs and having limited expansion samples for a new sign on dictionary expansion – a very likely real-world scenario. To simulate this, we used used aspects from each of the above settings: Our core vocabulary consists of 1,000 signs (with ∼15 videos per sign). We then test expanding the the vocabulary by 10%, 50%, 75% and 100% (i.e., adding sets of 100, 500, 750, and 1000 signs respectively). But in this setting, newly added signs have only one expansion sample each. We again consider two settings for how the single sample per sign is obtained: In the first setting, we assume that signs are organically contributed by different contributors. We simulate this by randomly sampling a video from the unseen expansion set. In the second setting, we assume a long-term or hired contributor adds new signs to a dictionary. We simulate this by using the seed signer's video to represent core and added signs throughout the dictionary.

Exploring Smaller Dictionaries Certain sign language dictionaries may be more specialized or in early stages of documenting a lexicon, and thus smaller in vocabulary. To simulate this scenario, we decided to create three new core-unseen splits (100 and 2631 signs respectively) for our experiments. Each of these splits is uniquely composed to reflect various underlying rationales of smaller dictionaries. First, naively, we randomly sampled set of glosses. Second, reflecting a dictionary that might be aiming to be comprehensive but is newer, we sampled a set of semantically distributed glosses (details in Appendix E). Third, to reflect a more specialized dictionary (such as one targeted towards language learners), we sampled a set of glosses matching that use case (from HandSpeak's list of first 100 signs for ASL learners). For each of these splits, our core vocabulary has 100 signs and total of ∼1490 training videos (i.e., ∼15 videos per sign, each performed by different contributors). We first test performance adding a single sign with multiple examples (i.e., newly added signs have ∼15 samples). Then we test adding multiple signs to the dictionary– exploring adding sets of 10, 50, 75, and 100 signs to the dictionary, following the same sampling approach as the large dictionary setting.

3.5 Model Architecture and Training

We use the spatiotemporal graph convolutional network (ST-GCN) (Yan et al., 2018), following previous works on this dataset. We extract keypoints using MediaPipe Holistic, and use the same 27 keypoints outlined in OpenHands (Selvaraj et al., 2022). We preprocess the keypoints using the same procedures outlined in prior work (Selvaraj et al., 2022; Desai et al., 2023). The model is trained on the core vocabulary for 100 epochs using a learning rate 1e-3, an Adam optimizer and a Cosine Annealing scheduler. We select the best performing checkpoint on the core validation set for our

expansion experiments. To extract feature embeddings for our experiments, we use the encoder of the ST-GCN model[3]. We use cosine distance to calculate similarity between embeddings in nearest neighbor search for dictionary retrieval.

3.6 Metrics

We use metrics consistent with prior work on sign language dictionary retrieval to evaluate the returned ranked list of glosses: Discounted Cumulative Gain (DCG, which evaluates overall ranking of the correct sign in the list) and recall-at-k (which considers if the correct is in the top-k rankings).

4 Results

Overall, we find that the performance of our dictionary expansion method varies greatly across different dictionary settings – Figure 2 provides a snapshot of our results. In the following subsections, we walk through the results in each setting and our corresponding analysis. The ranking of settings was consistent across all metrics so we report DCG, but full tables reporting each metric for each experiment can be found in the appendix.

4.1 Core Dictionary Performance

We provide a baseline to contextualize dictionary expansion by first measuring the retrieval performance of seen signs in the core vocabulary only, with no dictionary expansion performed (Table 1 in Appendix A). The classifier trained on the core vocabulary achieves a DCG of 0.7787 and top-1 accuracy of 62.21% when tested on the test split of the core vocabulary (1000 glosses, ∼12038 videos). We then validate our feature representation-based retrieval method: it achieves a DCG of 0.7719 and top-1 accuracy of 61.11% on the same test split of the core vocabulary. Comparing these two approaches, we see that the feature-representation approach performs as well as the classifier on seen (i.e., core) signs. We were further able to improve performance of the feature representation approach by calculating the centroid of all sample videos for each sign and using this instead in our retrieval database of feature representations. Interestingly, this strategy even surpasses the classifier, with a DCG of 0.8050 and top-1 accuracy of 66.04%. Thus, we retrieve the centroid representation of all core and expanded signs for all subsequent experiments.

[3]as the decoder is a single fully connected layer this corresponds to the second-to-last layer of the model

4.2 Base Setup Performance

Having established that using extracted features from a classifier is an effective strategy for dictionary retrieval, we next focused on testing a baseline dictionary expansion scenario that we consider the most ideal setting for dictionary expansion. In this set-up, we add a single sign, with expansion samples equal to the maximum number in the ASL Citizen training dataset for that sign (typically 15).

When we have multiple examples for each sign, we find that the performance of newly added signs is comparable to that of the core vocabulary (Table 2 in Appendix B). Specifically, we see a DCG of 0.8042 and top-1 accuracy of 65.38% on the new vocabulary (compared to a DCG of 0.8050 and top-1 accuracy of 66.04% for the core vocabulary). This suggests that our feature representation method could be effective in this dictionary expansion setting.

4.3 Adding a Sign with One Example

When adding a new sign with just a single example, however, we see a significant degradation in performance (Figure 2). When this sample is randomly selected, the newly added signs achieve a DCG of 0.4492 and top-1 accuracy of 21.74% (Table 2 in Appendix B) – a large drop compared to a DCG of 0.8042 and top-1 accuracy of 65.38% in the base setup with multiple examples per sign.

This demonstrates that data plays a critical role, and we hypothesize that constraining ourselves to a single example video for a sign limits our method's ability to provide a robust representation for retrieval with the new sign. The feature representation of a given video likely contains both information about the sign along with variation specific to a given contributor – we hypothesized that taking the centroid across many samples "averages out" the contributor-specific variation, providing a more robust representation of the sign itself.

Based upon this hypothesis, we reasoned that using a single contributor for all videos in the retrieval database, as opposed to different contributors, may reduce the impact of contributor-specific variation. We experimented with using a specific contributor's videos as expansion samples for the dictionary. We reasoned this would simulate a scenario where a long-term contributor to the dictionary has recorded many, if not all, signs in the vocabulary in addition to the added sign. A benefit of this is that we can then use this signer's videos to con-

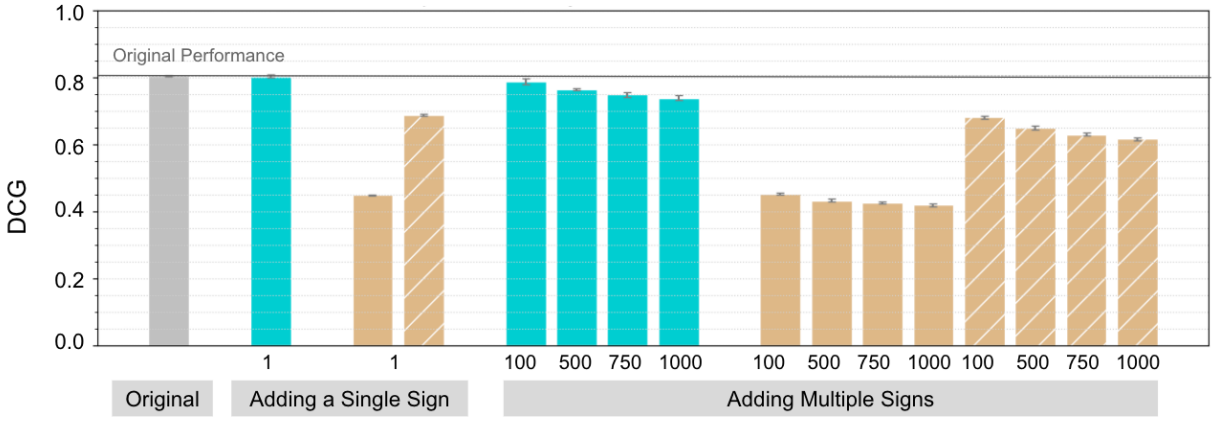

Figure 2: Performance of added vocabulary across different dictionary expansion scenarios. Bars are clustered by scenario and represent DCGs of added signs using centroid representations- scenarios with multiple samples (blue), one random contributor example (solid gold), one long-term contributor example (striped gold).

sistently represent all signs in the dictionary, both core and added – and thus minimize impact of user variation. We tested this approach with videos of five different contributors from the ASL Citizen dataset: the seed signer (P52), P33, P11, P50, P27 – all of whom had recorded videos for (almost all) glosses in the ASL Citizen dataset. We find that dictionary retrieval performance for added signs improves significantly with this single signer set up – in the best case (seed signer), achieving a DCG of 0.6881 and top-1 accuracy of 49.55% after using only one expansion sample per added sign (Table 2). However, we also note that performance varies drastically across contributors: e.g., 24.81% top-1 accuracy using P11's videos vs 42.15% using P27's videos (results for each signer can be found table 3 in Appendix B).

4.4 Adding Many Signs with Many Examples

Having explored expanding the vocabulary of a dictionary by one sign (i.e., going from a vocabulary of 1000 to 1001), we next studied the impact of larger scale expansion that involves adding multiple signs. The left half of Figure 4 and Table 4 summarize performance of core and added signs at different stages of expansion (adding 100, 500, 750, and 1000 signs).

We observed that as more signs are added, dictionary performance degrades. However, even in the largest expansion scenario we tested, dictionary retrieval performance overall remains reasonable for core and added signs at each stage of expansion (Table 4). When adding 100 signs, new signs achieve a DCG of 0.7984 and a top-1 accuracy of 64.33% on average – a slight drop in performance compared to the baseline expansion setup. At this tier, the impact on core vocabulary is also minimal, dropping to a DCG of 0.7968 and top-1 accuracy of 64.84% compared to pre-expansion. We find that performance continues to drop for both core and added vocabulary as we add more signs to the dictionary. The last expansion tier, adding 1000 signs, achieves a DCG of 0.7397 and top-1 accuracy of 56.03% for new signs and a DCG of 0.7436 and top-1 accuracy of 57.27% for core vocabulary. This tier corresponds to effectively doubling the vocabulary of the dictionary and shows a ∼9 point drop in top-1 accuracy and a 0.06 point drop in DCG. While this is not insignificant, we note that top-10 accuracy remains above 85% post-expansion for both core and added signs, meaning that users of a dictionary will continue to find the desired sign among the first 10 entries. This suggests that expanded dictionaries generally maintain similar levels of usability in our simulated setting, even up to 1000 added signs.

4.5 Adding Many Signs with One Example

We next assessed the interaction between adding multiple signs to a dictionary and limited number of expansion samples per sign. Figure 3 summarizes our results multiple stages of expansion and across two different setups– using a random contributor's video for each sign vs. using a specific, long-term contributor's video for all signs.

For both setups, we note a similar overarching trend that performance degrades as we proceed to larger expansion tiers i.e., as we add more signs to the vocabulary. However, we find that performance for core and added signs varies drastically

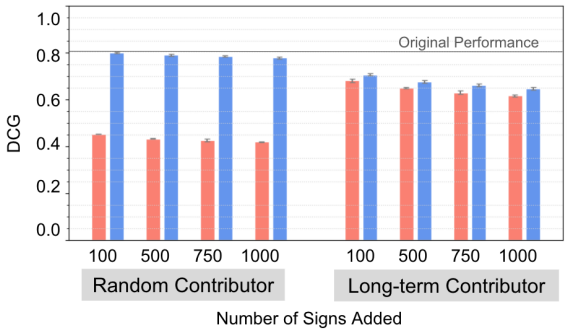

Figure 3: Impact of chosen expansion sample when adding multiple signs with one example. Reporting DCG of added signs (red) and core signs (blue).

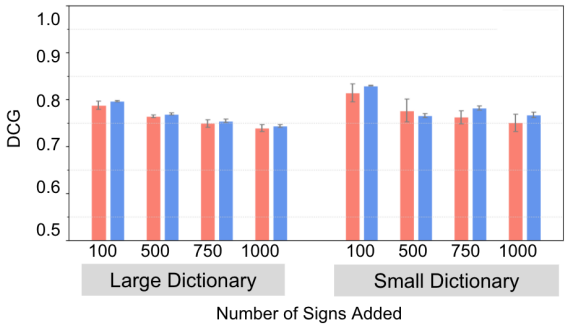

Figure 4: Comparing trends in dictionary expansion across large and small dictionaries.Reporting DCG of added signs (red) and core signs (blue).

in each setup (Figure 3 and Appendix D). When using the random contributor's video for added signs (alongside centroid for core signs), we find large disparities in performance between added signs and core signs – added signs achieve a DCG of ∼0.4 and top-1 accuracy of ∼20% in contrast to core signs with a DCG of ∼0.75 and top-1 accuracy of ∼62% across different stages of expansion (Table 5). When using a long-term contributor's videos, we again note the benefit of reducing user-variation in the retrieval space (Table 6)– added signs achieve a much better performance: a DCG of ∼0.64 and a top-1 accuracy of ∼44%. We note that there are less disparities between core and added signs in this setup; however, we see a bigger drop in core vocabulary performance compared to pre-expansion: achieving a DCG of ∼0.65 and a top-1 accuracy of ∼48% (Table 6). This is likely due to the fact that the seed signer's video does not approximate the latent representation of the sign as well as calculating centroid across multiple users.

4.6 Small Dictionaries

We were curious if the trends we see in dictionary expansion experiments would hold for dictionaries with smaller core vocabularies (such as new or emerging dictionaries). We tested this scenario using models trained on a vocabulary of 100 signs with multiple examples per sign. We then tested expansion using multiple examples per sign.

When adding a single sign to the vocabulary, we find that similar to our results with the larger dictionary, the performance of the added signs is comparable to that of the core vocabulary (first row of tables in Appendix E). When adding multiple signs, this trend continues (Figure 4)– we find small gaps between core vocabulary performance and added vocabulary performance at each stage of expansion. However, for the smaller dictionary, we note much larger variations in performance across different core-unseen splits (as evinced by large deviation bars in Figure 4). For example, with the randomly generated split, adding 10 signs to the vocabulary leads to a 9 point drop in top-1 accuracy for added signs (going from 67.22 to 58.21) (Table 7); whereas the semantically distributed split has almost no drop (Table 8) and the HandSpeak split has 3 point drop (Table 9). We believe this suggests that the lexical composition of the vocabulary (both core and added) is a much more significant factor for small dictionary expansion. Lastly, we note that performance degrades faster with each expansion tier for the small dictionaries. We hypothesize that this is because small dictionaries have not been exposed to sufficient data during training to learn good/robust feature representations for signs.

4.7 Visual and Runtime Analysis

Our representation-based method assumes that signs that are visually similar are placed closer to each other in the embedding space. To evaluate this, we examined the visual similarity of errors made by our method (i.e., confusions) in the base setup. We used phonological labels from ASL-Lex (Sehyr et al., 2021) to identify key parameters (Handshape, Movement, Location) for each sign in our vocabulary and then counted how many of these features differed between confused signs. We find that on average 1.85 features differ between confused signs. In contrast, when sampling 1000 random pairs of signs from the lexicon, 2.38 features differ on average. This indicates that the confusions made by the model are between signs

with fewer differences in phonology (i.e., more visually similar signs). Appendix F provides some examples of common confusions.

Finally, given the motivation around practical deployment, we calculated the time it takes for retrieval as the dictionary expands. As expected with nearest-neighbor search, retrieval time increases linearly with the size of the vocabulary: 10.12ms for 1000 signs vs. 16.08ms for 1500 signs vs. 21.47ms for 2000 signs. At the current vocabulary sizes, we anticipate this increase in time would not significantly impact user experience. However, as the dictionary grows further, implementing techniques for speeding up nearest neighbor search (e.g., using Meta's FAISS package) would be valuable.

5 Discussion

In this work, we explored sign language dictionary expansion across a variety of settings. Our results highlight the feasibility of expanding sign language dictionaries using feature representation-based approach, and demonstrate that performance is contingent on the type and quantity of data available for new vocabulary.

Given enough examples of a sign, we show that incorporating it into an existing dictionary is quite feasible. This is valuable for when small changes need to be made to a dictionary quickly (e.g., recording a new variation or documenting a newly emerging and rapidly adopted sign, like the sign for 'coronavirus').

When adding multiple signs with multiple examples, we show that models are generally robust, albeit at a slightly lower performance than when just a single sign is added. This can inform the frequency at which we retrain models in response to changing vocabularies. A comparison of numbers from the small dictionary experiments (100 core signs) to the large dictionary experiments (1000 core signs) suggest models may need to be retrained at more frequent intervals initially, but less often once they've develop robust representations.

On the other hand, we find that performance degrades when added new signs that have only a limited number of samples. As we would expect, this is the most difficult, but also the most realistic, expansion scenario. This illuminates an important direction for future research: examining new methods for incorporating signs into a dictionary in data-constrained scenarios. We focused on classifiers trained using deep learning in this work; however, alternative training approaches (e.g., meta learning) may result in more robust feature representations for core and added signs. We encourage researchers to take up this line of inquiry.

Looking forward, we emphasize the importance of addressing dictionary expansion for sign languages. Compared to written languages, existing sign language dictionaries represent only a small fraction of the true language lexicon. While many crowdsourcing and documenting initiatives are underway, gathering a snapshot of sign languages that is representative of the many different regional and contextual variations continues to be a challenge. In addition to variations in everyday vernacular, different fields of study (such as STEM disciplines) frequently produce new signs to represent new concepts. Dictionary expansion then offers a crucial pathway to building sustainable and scalable sign language dictionaries.

Our results also offer interesting implications for the maintenance of current crowdsourced dictionaries. Examples available for each sign are expected to grow over time with contributions from users. A representation-based approach like ours (that works with any number of signs and any number of examples per sign) aligns well to the fluctuating vocabulary size and data availability encountered in these dictionaries. Initially, dictionary administrators could work to incorporate a sign with any available sample, and then switch to centroid for these added signs as more examples are collected. While this results in poor added sign performance initially, it allows the sign to be incorporated into the video-based dictionary and solicit further interaction and contributions from dictionary users. To maximize initial performance, dictionary administrators could also work to sustain long-term contributors to record most, if not all, signs being added to the dictionary (as this shows best added sign performance with limited data).

Overall, our work outlines directions for sign language research aligned with community needs and real world settings and furthers the exploration of sign languages as languages in their own right.

6 Limitations

In this work, we reported results on dictionary expansion using only an ST-GCN model. While this architecture is frequently used in sign language recognition and has the benefit of being lightweight,

it is not the only established approach. Appearance-based approaches to sign language recognition (like I3D) might fare differently under expansion scenarios. Exploring dictionary expansion across a variety of architectures as well as robustness of their learned feature representations could be an interesting direction for future research.

To simulate dictionary expansion, we created core-unseen splits from ASL Citizen. This meant that many of the contributors in the unseen split had previously been seen by the model – while this is reflective of a real-world scenario (long-term dictionary contributors), using a completely new dataset for unseen vocabulary would have also allowed us to simulate a dictionary setting with a completely new contributor scenario for added signs.

In our results, we note that the centroid approach outperforms the classifier even in the base dictionary setting – to the best of our knowledge, this has not been discussed in prior literature. While valuable, a core limitation of this approach is that the centroid representation requires multiple training examples to become reliable, which may be difficult leverage with data scarcity.

Our analysis surfaces large variance in dictionary retrieval performance across contributors (section 4.3), we do not investigate these disparities in depth. We hypothesize these disparities could be explained by variations in video quality for these different signers or biases in the underlying model (or both). Disentangling these aspects and more systematically exploring dictionary expansion performance across different contributors is valuable direction for future work.

We also highlight differences in dictionary expansion trends at smaller vs. larger core vocabulary sizes. However, our results are limited to comparing two specific vocabulary sizes (100 vs. 1000) – a more systematic investigation across different vocabulary sizes may have better unearthed overarching trends. Additionally, with our smaller vocabularies we note sensitivity to lexical composition, but our analysis reports aggregates which prevents us from investigating sign-level nuances and conducting a linguistically-informed analysis.

Lastly, we position our work related to evolution of sign languages, but only focus on adding signs to dictionaries and not replacing or removing signs. While our method could be easily adapted to address these, further research is required to investigate impacts on overall dictionary retrieval performance.

References

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.

Katherine Atwell, Danielle Bragg, and Malihe Alikhani. 2024. Studying and mitigating biases in sign language understanding models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 268–283.

Matyáš Bohacek and Marek Hrúz. 2023. Learning from what is already out there: Few-shot sign language recognition with online dictionaries. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6.

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

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, and Danielle Bragg. 2023. Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36:76893–76907.

Aashaka Desai, Maartje De Meulder, Julie A Hochgesang, Annemarie Kocab, and Alex X Lu. 2024. Systemic biases in sign language ai research: A deaf-led call to reevaluate research agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 54–65.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H Liu. 2024. Shubert: Self-supervised sign language representation learning via multi-stream cluster prediction. *arXiv preprint arXiv:2411.16765*.

Rinki Gupta. 2022. Expanding indian sign language recognition system using class incremental learning. In *2022 International Conference on Advances in Computing, Communication and Materials (ICACCM)*, pages 1–5.

Saad Hassan, Matyas Bohacek, Chaelin Kim, and Denise Crochet. 2025. Towards an ai-driven video-based american sign language dictionary: Exploring design and usage experience with learners. *arXiv preprint arXiv:2504.05857*.

Brian Hie, Hyunghoon Cho, Benjamin DeMeo, Bryan Bryson, and Bonnie Berger. 2019. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell systems*, 8(6):483–493.

Joe Huamani-Malca and Gissella Bejarano. 2023. Comparing incremental learning approaches for a growing sign language dictionary. In *Annual International Conference on Information Management and Big Data*, pages 97–106. Springer.

Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. 2023. The sem-lex benchmark: Modeling asl signs and their phonemes. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–10.

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.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh M Khapra. 2022. Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, and 1 others. 2023. Popsign asl v1. 0: An isolated american sign language dataset collected via smartphones. *Advances in Neural Information Processing Systems*, 36:184–196.

Toon Vandendriessche, Mathieu De Coster, Annelies Lejon, and Joni Dambre. 2025. Representing signs as signs: One-shot islr to facilitate functional sign language technologies. *arXiv preprint arXiv:2502.20171*.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2025. Signrep: Enhancing self-supervised sign representations. *arXiv preprint arXiv:2503.08529*.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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.

A Core Dictionary Performance

Table 1 corresponds to section 4.1, where we establish a baseline of the classifier performance and our feature-representation approach using the test split of our core vocabulary.

Setup	DCG	Rec@1	Rec@10
Classifier	0.7787 ± 0.0051	0.6221 ± 0.0101	0.8860 ± 0.0020
All Features	0.7719 ± 0.0016	0.6111 ± 0.0052	0.8815 ± 0.0021
Centroid	0.8050 ± 0.0013	0.6604 ± 0.0040	0.9048 ± 0.0010

Table 1: Establishing baseline performance using classifier and our feature-based approach. The 'all features' setting corresponds to naive use of learned feature representations, and centroid corresponds to calculating centroid for each sign first.

B Adding a Single Sign

Table 2 corresponds to section 4.2 and section 4.3, where we experiment with adding one sign to the dictionary. Table 3 summarized results of our experiments using different contributors when simulating our 'long-term contributor' scenario in section 4.3.

Setting	DCG	Rec@1	Rec@10
Many Examples	0.8042 ± 0.0037	0.6538 ± 0.0047	0.9105 ± 0.0040
One Random Contributor Example	0.4492 ± 0.0016	0.2174 ± 0.0041	0.5396 ± 0.0056
One Long-term Contributor Example	0.6881 ± 0.0025	0.4955 ± 0.0006	0.81205 ± 0.0070

Table 2: Dictionary expansion performance when adding a single sign to the dictionary under a variety of data settings. We can see that amount of data and choice of sample (in low-data settings) are a significant factor in performance.

Setup	DCG	Rec@1	Rec@10
P52 (seed signer)	0.6881 ± 0.0025	0.4955 ± 0.0006	0.8121 ± 0.0070
P33	0.5556 ± 0.0017	0.3275 ± 0.0035	0.6781 ± 0.0014
P11	0.4715 ± 0.0031	0.2482 ± 0.0024	0.5615 ± 0.0043
P50	0.6163 ± 0.0035	0.4004 ± 0.0031	0.7445 ± 0.0056
P27	0.6346 ± 0.0037	0.4215 ± 0.0038	0.7678 ± 0.0038

Table 3: Dictionary expansion performance when adding a single sign to the dictionary with a single example and using a specific contributor's videos to represent all signs. We see performance varies significantly across chosen contributor.

C Adding Many Signs with Many Examples

Table 4 corresponds to section 4.4, where we experiment with adding many signs to the dictionary using all available samples – it summarizes performance of both added signs and core signs with each stage of expansion (adding 100, 500, 750, 1000 signs). We report averages across each the three randomly sampled sets for each stage of expansion.

Setup	DCG	Rec@1	Rec@10
Adding 1	0.8042 ± 0.0037	0.6538 ± 0.0047	0.9105 ± 0.0040
Adding 100	0.7984 ± 0.0090	0.6433 ± 0.0107	0.9078 ± 0.0084
Adding 500	0.7644 ± 0.0031	0.5949 ± 0.0030	0.8815 ± 0.0040
Adding 750	0.7494 ± 0.0080	0.5730 ± 0.0100	0.8689 ± 0.0093
Adding 1000	0.7397 ± 0.0076	0.5603 ± 0.0101	0.8600 ± 0.0072

Setup	DCG	Rec@1	Rec@10
Original	0.8050 ± 0.0013	0.6604 ± 0.0040	0.9048 ± 0.0010
Adding 100	0.7968 ± 0.0015	0.6484 ± 0.0038	0.8989 ± 0.0018
Adding 500	0.7694 ± 0.0025	0.6090 ± 0.0052	0.8774 ± 0.0022
Adding 750	0.7548 ± 0.0034	0.5882 ± 0.0067	0.8657 ± 0.0030
Adding 1000	0.7436 ± 0.0033	0.5727 ± 0.0057	0.8567 ± 0.0037

Table 4: Dictionary Expansion performance when adding multiple signs with multiple examples. Top table is performance of added signs and bottom table is corresponding performance of core signs after dictionary expansion. Top rows in each table are baselines from previous settings.

D Adding a Many Signs with Single Example

Tables 5 and 6 corresponds to section 4.5, where we experiment with adding many signs to the dictionary using a single only one expansion sample. They summarize results with two different setups: using a random contributor's video for each sign vs. using the seed signer's video across the board. We report performance of both added signs and core signs with each stage of expansion (adding 100, 500, 750, 1000 signs). We report averages across each the three randomly sampled sets for each stage of expansion.

Setup	DCG	Rec@1	Rec@10
Adding 100	0.4527 ± 0.0062	0.2203 ± 0.0016	0.5457 ± 0.0163
Adding 500	0.4332 ± 0.0043	0.2016 ± 0.0023	0.5200 ± 0.0094
Adding 750	0.4263 ± 0.0066	0.1971 ± 0.0040	0.5091 ± 0.0114
Adding 1000	0.4192 ± 0.0024	0.1910 ± 0.0013	0.4990 ± 0.0051
Setup	**DCG**	**Rec@1**	**Rec@10**
Adding 100	0.8009 ± 0.0021	0.6543 ± 0.0059	0.9027 ± 0.0009
Adding 500	0.7909 ± 0.0037	0.6402 ± 0.0078	0.8946 ± 0.0006
Adding 750	0.7852 ± 0.0035	0.6327 ± 0.0078	0.8895 ± 0.0010
Adding 1000	0.7792 ± 0.0048	0.6251 ± 0.0085	0.8845 ± 0.0008

Table 5: Dictionary Expansion performance when adding multiple signs with a single example, and using a random contributor's video as a sample. Top table is performance of added signs and bottom table is corresponding performance of core signs after dictionary expansion. We see large disparities between added and core sign performance at each expansion stage.

Setup	DCG	Rec@1	Rec@10
Adding 100	0.6813 ± 0.0062	0.4877 ± 0.0118	0.8038 ± 0.0057
Adding 500	0.6504 ± 0.0035	0.4477 ± 0.0037	0.7743 ± 0.0086
Adding 750	0.6308 ± 0.0074	0.4231 ± 0.0091	0.7561 ± 0.0088
Adding 1000	0.6164 ± 0.0034	0.4047 ± 0.0055	0.7401 ± 0.0075
Setup	**DCG**	**Rec@1**	**Rec@10**
Adding 100	0.7070 ± 0.0049	0.5205 ± 0.0043	0.8284 ± 0.0064
Adding 500	0.6764 ± 0.0056	0.4808 ± 0.0069	0.8005 ± 0.0092
Adding 750	0.6613 ± 0.0052	0.4614 ± 0.0065	0.7855 ± 0.0091
Adding 1000	0.6477 ± 0.0056	0.4443 ± 0.0077	0.7728 ± 0.0087

Table 6: Dictionary Expansion performance when adding multiple signs with a single example and using the seed signer's videos to represent all signs. Top table is performance of added signs and bottom table is corresponding performance of core signs after dictionary expansion. We see improvement in added sign performance compared to the random contributor setup, but core sign performance lags behind.

E Small Dictionaries

Tables 8 and 7 and 9 corresponds to section 4.6, where we experiment with simulating dictionary expansion for small dictionaries. We report results on three different core-unseen splits: randomly generated (Table 7), semantically distributed (Table 8), and specialized (language learning lexicon) (Table 9). To sample the semantically distributed glosses, we extracted GloVE embeddings (Pennington et al., 2014) for all glosses in ASL Citizen dataset (2731) and used geometric sketching (Hie et al., 2019) to sample 100 representative glosses. In the following tables, we report performance of both added signs and core signs with each stage of expansion (adding 100, 500, 750, 1000 signs). We report averages across each the three randomly sampled sets for each stage of expansion.

Setup	DCG	Rec@1	Rec@10
Adding 1	0.8314	0.6722	0.9580
Adding 10	0.7877	0.5821	0.9522
Adding 50	0.7654	0.5735	0.9180
Adding 75	0.7572	0.5714	0.8987
Adding 100	0.7328	0.5285	0.8937

Setup	DCG	Rec@1	Rec@10
Original	0.8302	0.6953	0.9290
Adding 10	0.8311	0.6967	0.9396
Adding 50	0.7925	0.6430	0.9071
Adding 75	0.7783	0.6244	0.8915
Adding 100	0.7625	0.6054	0.8754

Table 7: Performance with small randomly sampled dictionary, adding multiple signs with multiple examples. Top table is performance of added signs and bottom table is corresponding performance of core signs after dictionary expansion. Top rows in each table correspond to establishing a baseline and adding single signs.

Setup	DCG	Rec@1	Rec@10
Adding 1	0.8230	0.6549	0.9551
Adding 10	0.8300	0.6695	0.9491
Adding 50	0.7544	0.5383	0.9236
Adding 75	0.7481	0.5511	0.8959
Adding 100	0.7511	0.5555	0.9015

Setup	DCG	Rec@1	Rec@10
Original	0.8427	0.7073	0.9498
Adding 10	0.8304	0.6965	0.9409
Adding 50	0.7962	0.6530	0.9086
Adding 75	0.7818	0.6305	0.8980
Adding 100	0.7658	0.6056	0.8841

Table 8: Performance with small semantically diverse dictionary, adding multiple signs with multiple examples. Top table is performance of added signs and bottom table is corresponding performance of core signs after dictionary expansion. Top rows in each table correspond to establishing a baseline and adding single signs.

Setup	DCG	Rec@1	Rec@10
Adding 1	0.8528	0.7083	0.9689
Adding 10	0.8261	0.6601	0.9568
Adding 50	0.8112	0.6446	0.9468
Adding 75	0.7835	0.6016	0.9307
Adding 100	0.7692	0.5818	0.9150

Setup	DCG	Rec@1	Rec@10
Original	0.8255	0.6694	0.9507
Adding 10	0.8291	0.6801	0.9481
Adding 50	0.8007	0.6423	0.9233
Adding 75	0.7863	0.6184	0.9139
Adding 100	0.7749	0.6053	0.8964

Table 9: Performance with small specialized lexicon dictionary (for language learners), adding multiple signs with multiple examples. Top table is performance of added signs and bottom table is corresponding performance of core signs after dictionary expansion. Top rows in each table correspond to establishing a baseline and adding single signs.

F Visual Similarity of Common Confusions

We present some common confusions that occurred for an experimental split under the Base Setup scenario (Section 4.2). We present ground truth and predicted glosses alongside phonological labels to describe visual similarity. These phonological labels were derived from ASL-Lex (Sehyr et al., 2021) (a database of phonological properties of signs which all ASL Citizen glosses are cross-referenced against).

Type	Gloss	Handshape	Movement	Location
Truth	SHOVEL1	s	Curved	Neutral
Predicted	DIGUP	a	Curved	Neutral
Truth	ASSEMBLY	a	Straight	Body
Predicted	ATLANTA	a	Curved	Body
Truth	RING	g	Straight	Hand
Predicted	FILTER	5	Straight	Neutral
Truth	VISUALIZE	s	Straight	Head
Predicted	IMAGINE2	s	Straight	Head
Truth	CALM	closed-b	Curved	Neutral
Predicted	QUIET	closed-b	Curved	Head

Table 10: Common Confusions from Base Setup Scenario (Adding One Sign With Many Examples)