A Appendix

This document contains supplementary material for the YouTube-ASL paper.

A.1 Full Qualitative Results

Table 5: The complete set of qualitative examples from our best finetuned and zero-shot models, on sentences sampled from How2Sign by Tarrés et al. [38].

(1)	**Reference**	And that's a great vital point technique for women's self defense.
	Tarrés et al.	It's really a great point for women's self defense.
	Ours (zero-shot)	It's really great, especially for women who are facing barriers.
	Ours (finetuned)	It's really great for women's self defense.
(2)	**Reference**	In this clip I'm going to show you how to tape your cables down.
	Tarrés et al.	In this clip I'm going to show you how to improve push ups.
	Ours (zero-shot)	This video will show how to use the code online.
	Ours (finetuned)	In this clip we're going to show you how to cut a piece of clay.
(3)	**Reference**	In this segment we're going to talk about how to load your still for distillation of lavender essential oil.
	Tarrés et al.	Ok, in this clip, we're going to talk about how to fold the ink for the lid of the oil.
	Ours (zero-shot)	This video will discuss how to submit a digital form for the survey.
	Ours (finetuned)	In this clip we're going to talk about how to feed a set of baiting lizards for a lava field oil.
(4)	**Reference**	You are dancing, and now you are going to need the veil and you are going to just grab the veil as far as possible.
	Tarrés et al.	So, once you're belly dancing, once you've got to have the strap, you're going to need to grab the thumb, and try to avoid it.
	Ours (zero-shot)	he's dancing a lot. Now he needs a hat and a chain
	Ours (finetuned)	Their hopping and dancing is now, they're going to need their squat and squat and they're going to be able to move independently.
(5)	**Reference**	But if you have to setup a new campfire, there's two ways to do it in a very low impact; one is with a mound fire, which we should in the campfire segment earlier and the other way to setup a low impact campfire is to have a fire pan, which is just a steel pan like the top of a trash can.
	Tarrés et al.	And other thing I'm going to talk to you is a little bit more space, a space that's what it's going to do, it's kind of a quick, and then I don't want to take a spray skirt off, and then I don't want it to take it to the top of it.
	Ours (zero-shot)	But if you have to set up a new campfire, you have to set up a campfire. You have to do it in a campfire, or set up a tentfire.
	Ours (finetuned)	But if you have to set up a new campfire, there are two ways to do a low impact fire, one is a cone fire, which we have to do in the tent earlier, and the other one is to set up a campfire in a fire pan.
(6)	**Reference**	So, this is a very important part of the process.
	Tarrés et al.	It's a very important part of the process.
	Ours (zero-shot)	Wash your hands.
	Ours (finetuned)	Alright, let's get started.

B Datasheets for Datasets

We provide documentation of the dataset based on Datasheets for Datasets [6].

[6] https://arxiv.org/pdf/1803.09010.pdf

B.1 Motivation

For what purpose was the dataset created? The dataset was created primarily to serve as training data for ASL to English machine translation; prior datasets are smaller and have fewer unique signers. We used human annotators to identify high-quality ASL videos with well-aligned captions, but it was not feasible to manually correct or align any of the included captions. This is generally sufficient for translation, but slightly less ideal for tasks like ASL to English caption alignment, where consistent alignment standards might be desired. The dataset is probably also less suitable for English to ASL translation, due to the signing variation across videos, though this may be addressed with methods for improved controllability.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This dataset was created by Dave Uthus, Garrett Tanzer and Manfred Georg for Google.

Who funded the creation of the dataset? Google.

B.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Each instance is the video id of a YouTube video.

How many instances are there in total (of each type, if appropriate)? There are 11,093 video ids.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? This contains a subset of videos available on YouTube of people signing in American Sign Language with English captions. This is not strictly representative of the larger set, as we have applied a combination of automatic and manual filtering techniques to find high quality videos with high-quality captions.

What data does each instance consist of? Each instance consists of a single video id, which represents an ASL video with associated English captions.

Is there a label or target associated with each instance? The English captions may be considered the target for each instance, but this depends on the task being attempted.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? No.

Are there recommended data splits (e.g., training, development/validation, testing)? No. YouTube datasets can change over time due to the nature of the platform (videos can be made private or deleted), thus there are no recommended splits.

Are there any errors, sources of noise, or redundancies in the dataset? Yes. During the annotation process, we allowed the annotators to mark whole channels with the same annotations, so there may be some videos which are not of the same quality as the rest of the channel. There may also be minor errors in the videos or captions that the annotators explicitly deemed acceptable.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is not self-contained, as it consists of YouTube video ids only. As such, there is no guarantee that the dataset will remain constant over time. The actual videos themselves fall under YouTube's Terms of Service https://www.youtube.com/static?template=terms.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)? No, the dataset consists of video ids for public videos only, and does not rehost any of the underlying data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? The video ids comprising the dataset itself are random identifiers. The videos referenced by our video ids are hosted by YouTube and therefore subject to YouTube's community guidelines. We or our annotators did not encounter any such content.

Does the dataset identify any subpopulations (e.g., by age, gender)? The dataset does not identify any subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? The video ids comprising the dataset itself are random identifiers. The videos referenced by our video ids include people signing in ASL, which inherently includes the person's appearance. Our dataset does not provide any extra information about these people that was not already publicly available from their uploaded videos, and respects when videos are deleted/made private by virtue of using video ids.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? As with the previous question, the videos referenced by our video ids may contain information about many topics, if the signer chose to discuss that information in their publicly uploaded video. Our dataset does not provide any extra information and respects deletions.

B.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? The videos referenced by each video id instance consist solely of data that is directly observable (uploaded videos and captions). The selection of video ids is implicitly decided by a combination of automatic and manual filtering in order to ensure a relatively high quality level.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? A combination of software programs and manual annotations were used to select preexisting YouTube videos.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Not applicable.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The authors collected the initial collection of video ids, and annotators were hired to help annotate the videos for filtering purposes.

Over what timeframe was the data collected? We collected videos up to January 2022.

Were any ethical review processes conducted (e.g., by an institutional review board)? No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The dataset consists of references to videos that include people, but doesn't collect or release any new information about those people.

Were the individuals in question notified about the data collection? No, as we only provide video ids and no further information about the videos.

Did the individuals in question consent to the collection and use of their data? Not applicable.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? As we only provide ids and not the raw content, users removing the video will make them no longer available for use in our dataset.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? No.

B.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? Yes, we filtered out videos that were not relevant, of poor quality, had poor captions, etc. The result is a list of video ids, which point to unmodified videos.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? The original set of video ids will not be made available.

Is the software that was used to preprocess/clean/label the data available? No.

B.5 Uses

Has the dataset been used for any tasks already? None prior to the baselines provided in this paper.

Is there a repository that links to any or all papers or systems that use the dataset? No.

What (other) tasks could the dataset be used for? In addition to the intended task of ASL to English translation, the dataset could be used for related sign language understanding tasks like caption alignment, and potentially for sign language generation tasks like English to ASL translation, or sign language tasks that do not require captions.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? The dataset was filtered for high-quality videos and captions, which includes a variety of signing styles and skill levels, as long as they can be understood by an ASL user and are captioned correctly. This amount of variety may be unideal for tasks like generation where consistency is preferred. Even though it is varied, it is still not necessarily representative of the signing community as a whole, and should be treated as such.

Are there tasks for which the dataset should not be used? This dataset should not be used as a benchmark for comparing models across time, because the data that YouTube-ASL is derived from will change over time with video deletions or other modifications.

Any other comments?

B.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? The dataset is open sourced.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? GitHub.

When will the dataset be distributed? It is currently available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? We release the YouTube-ASL video ids under CC BY 4.0 International license, while the actual videos/captions on YouTube are preexisting and subject to the YouTube Terms of Service (https://www.youtube.com/static?template=terms).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? See above for license information.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

B.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The authors will be responsible for maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? By contacting any of the authors listed on the publication.

Is there an erratum? No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? It may be updated, and if so, updates will be communicated via the associated GitHub page.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be

retained for a fixed period of time and then deleted)? Our dataset is constructed similarly to past YouTube-related datasets, in that we only provide video ids. Thus, if a user makes their YouTube video private or deletes it, this will then no longer be available for use.

Will older versions of the dataset continue to be supported/hosted/maintained? No, if we need to remove older versions, these will be communicated on the associated GitHub page.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? No, we are currently not planning to allow formal contributions to the dataset at this time, but others may extend the dataset on their own in accordance with the license.

C Additional Information

URL to the data: https://github.com/google-research/google-research/tree/master/youtube_asl

Hosting and maintenance: The data website is on GitHub under Google Research's shared GitHub repository, while the data itself is hosted on Google Research's shared Google Cloud Service.

D Author Statement

The authors bear all responsibility in case of violation of rights, and confirm that this dataset is open-sourced under the CC BY 4.0 International license.