YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus

David Uthus, Garrett Tanzer, Manfred Georg
Google
{duthus,gtanzer,mgeorg}@google.com

Abstract

Machine learning for sign languages is bottlenecked by data. In this paper, we present YouTube-ASL, a large-scale, open-domain corpus of American Sign Language (ASL) videos and accompanying English captions drawn from YouTube. With ~1000 hours of videos and >2500 unique signers, YouTube-ASL is ~3x as large and has ~10x as many unique signers as the largest prior ASL dataset. We train baseline models for ASL to English translation on YouTube-ASL and evaluate them on How2Sign, where we achieve a new finetuned state of the art of 12.39 BLEU and, for the first time, report zero-shot results.

1 Introduction

The primary bottleneck for machine learning research on sign languages is data. As minority languages used by historically marginalized Deaf/Hard of Hearing communities, sign languages lack the plentiful online resources that have facilitated modern machine learning advances [4, 42, 12]. This is compounded by the fact that sign languages have no standardized written form: mining the videos that do exist is more difficult than retrieval for spoken language text. For translation specifically, there is the added problem of finding spoken language captions that are aligned to corresponding sign language content, rather than a voiceover with its own timing. The result is that datasets tend to be constructed by recording new footage in a studio or curating videos from a small number of manually selected content creators, which limits variety.

In order to address these challenges, we present YouTube-ASL, a large-scale, open-domain corpus of American Sign Language (ASL) videos and accompanying English captions, primarily intended for ASL to English machine translation. We mined these videos from YouTube using a two-step process: first, we used automatic content-based annotations to identify potentially relevant captioned videos; and second, we used skilled human annotators to filter out videos with poor quality or misaligned captions. The result is a dataset with 984 hours of high-quality captioned video featuring >2500 unique signers, which is ~3x as large as the largest prior ASL dataset [37] and has ~10x as many unique signers as any sign language dataset to date.

We train simple baseline models for sentence-level ASL to English translation on YouTube-ASL by embedding MediaPipe Holistic landmarks [25, 16] into the T5 language model [32]. Because YouTube videos may be removed over time and therefore cannot form a stable test set—and for comparison to prior work—we evaluate on a standard benchmark, How2Sign [13]. Borrowing from trends in mainstream machine learning [31, 15, 10], we provide not just finetuned but also zero-shot[1] results to test out-of-domain generalization. We achieve a new finetuned state of the art of 12.39 BLEU (vs. the prior SOTA of 8.03 [38]), and for the first time report a zero-shot score, 3.95 BLEU.

[1] Here, "zero-shot" refers to evaluation on an independently constructed benchmark without any kind of domain adaptation. This is different from the use of "zero-shot" in machine translation for transfer to unseen language pairs, or "zero-shot" in prompting for prompts without in-context examples.

Table 1: Summary statistics for different sign language translation datasets. See Section 3.3 for details on how these statistics were derived for YouTube-ASL.

Name	Language	Vocab.	# Hours	# Signers	Source
RWTH-PHOENIX-2014T [5]	DGS	3K	11	9	TV
BOBSL [2]	BSL	77K	1447	39	TV
SWISSTXT [7]	DSGS	-	88	-	TV
VRT-RAW [7]	VGT	-	100	-	TV
CSL-Daily [45]	CSL	2K	23	10	Lab
KETI [21]	KVK	419	28	14	Lab
Public DGS Corpus [18]	DGS	-	50	-	Lab
SP-10 [40]	various	17K	14	79	Web
AfriSign [17]	various	20K	152	-	Web
How2Sign [13]	ASL	16K	79	11	Lab
OpenASL [37]	ASL	33K	288	220	Web
YouTube-ASL (ours)	ASL	60K	984	>2519	Web

We publicly release the YouTube-ASL video IDs.[2] We hope that YouTube-ASL will be useful for tasks such as ASL to English translation and caption alignment—both in the near term to aid in the construction of larger sign language datasets, and eventually to improve accessibility for the Deaf/Hard of Hearing community.

2 Related Work

In this section, we review prior sign language translation datasets and methods for translation from sign languages to spoken languages.

2.1 Sign Language Translation Datasets

Table 1 shows statistics on different sign language translation datasets. There are three main sources for sign language data: ad hoc recorded footage, interpreted TV broadcasts, and online video sharing platforms.

In the first category are datasets that manually recruit signers and record them performing translations of desired phrases, either in a lab setting or with a camera on their personal device. These datasets tend to be small and feature few signers for logistical reasons, and may have exhaustive annotations because the small size of the dataset makes it feasible. This includes datasets such as CSL-Daily [45], with phrases related to daily life in Chinese Sign Language; KETI [21], with phrases related to emergency situations in Korean Sign Language; Public DGS Corpus [18], with elicited dialogues in German Sign Language; and How2Sign [13], with "How To" instructional monologues translated into American Sign Language.

In the second category are datasets that collate interpreted TV programs from a collaborating national broadcaster. These datasets tend to be larger than newly recorded ones, but often use a small number of non-native interpreters and lack fine-grained caption alignment (because the supervision comes from the spoken language audio track). This includes datasets such as RWTH-PHOENIX-2014 [5], with weather forecasts interpreted into German Sign Language; SWISSTXT [7], with news/weather programs interpreted into Swiss German Sign Language; VRT [7], with news programs interpreted into Flemish Sign Language; and BOBSL [2], with BBC programs in many domains interpreted into British Sign Language. At 1447 hours, BOBSL is the largest sign language translation dataset to date (including the present work), but has only 39 signers and speech-aligned subtitles, vs. YouTube-ASL's >2519 signers and sign-aligned captions—though the two datasets are complementary because they are for different languages.

In the third category are datasets that curate content from online video sharing platforms. In prior sign language translation datasets, this content is drawn from a small number of manually selected channels.

[2] https://github.com/google-research/google-research/tree/master/youtube_asl

This includes datasets such as SP-10 [40], with example sentences from an online multilingual sign dictionary; AfriSign [17], with translated Bible passages hosted on the Jehovah's Witnesses website; and OpenASL [37], with videos from three YouTube channels: *DailyMoth*, *Sign1News*, and the National Association of the Deaf. OpenASL is the largest prior ASL dataset and closest work to YouTube-ASL: the key difference is that YouTube-ASL is constructed with open-ended mining from automatic tags, rather than manual channel curation. OpenASL is largely a subset of YouTube-ASL, which—by utilizing the long tail of channels—is ~3x as large and has ~10x as many unique signers.

There are several datasets for easier tasks than translation, like isolated sign recognition and fingerspelling recognition, that mine from the web by ambiguous means. MS-ASL [20], WLASL [23], and ChicagoFSWild [35]/ChicagoFSWild+ [36] are word-level datasets mined from YouTube, sign language-targeted sites like ASLU and ASL-LEX, or other unnamed video sharing platforms. These works do not specify how they retrieved their videos, so it is possible that they used a similar automatic tagging approach to YouTube-ASL, albeit on a more limited scale.

2.2 End-to-End Sign Language Translation

Originally, sign language translation approaches operated on *glosses*, linguistic annotations that represent individual signs, or cascaded translation through glosses as an intermediate step, like speech to text translation often cascades through speech recognition. More recently, due to a variety of deficiencies in glosses and lack of widespread gloss data, the field has shifted to end-to-end modeling with encoder-decoder Transformers, starting with Camgöz et al. [5].

The two main classes of approaches are those that take learned video embeddings as input [6, 38, 27] (via video encoders, primarily I3D [9], pretrained on tasks such as isolated sign recognition), and those that take estimated pose landmarks as input [27] (such as MediaPipe [25] or OpenPose [8]). Some works achieve modest gains given constant data with architectural tweaks like treating different cues in the input video (hands, face) differently [44, 41]. It is unclear to what extent these techniques are necessary or beneficial on larger datasets. Other works seek to benefit from transfer from spoken language or other sign language data [11, 43, 17]. All of these works train and evaluate on splits derived from the same underlying continuous sign language corpus (different datasets across papers), and sometimes multiple such datasets independently in the same paper. In contrast, we train on YouTube-ASL using an uncomplicated approach and evaluate on How2Sign, reporting both finetuned and zero-shot results to get a more robust understanding of our model's state-of-the-art performance.

3 The YouTube-ASL Corpus

YouTube-ASL is a corpus of American Sign Language (ASL) videos with accompanying English captions drawn from YouTube. Video sharing platforms like YouTube are appealing sources of sign language data because they host swaths of diverse content that are more broadly representative of real world conditions than studio footage is. Of course, much of this data is irrelevant or low-quality, so it is imperative to develop cost-effective ways to sift through it.

We used a two-step pipeline to construct the corpus: first, retrieval using automatic content-based annotations, and second, filtering by skilled human annotators at a per-video level. This automatic retrieval step represents a departure from prior continuous sign language corpora and brings us closer to mining approaches from mainstream machine learning.

3.1 Automatically Retrieving Candidate Videos

As described previously in Abu-El-Haija et al. [1], the YouTube video annotation system associates machine-generated tags with each video in the form of Knowledge Graph entities, which are based on the video's metadata, context, and content signals. We retrieved listed public videos tagged as being related to sign language generally or American Sign Language specifically, as of January 2022.[3] This automatic tagging step, while having higher recall than prior works, was flawed in that it was not aware of sign language in the video content itself—to be expected due to the limited nature of current sign language processing. This means that, for example, videos in sign language that do not explicitly mention sign language in the content or context were unlikely to be discovered. This failure

[3] Some video IDs may have been removed from this set over time due to video deletions.

mode was most salient for press conferences with simultaneous interpreters, which tend not to have well-aligned captions anyway.

Given these retrieved videos, we drilled down on those with user-generated captions—i.e., captions that were manually uploaded rather than automatically derived from speech—because speech-derived captions are not tightly aligned with signed content. As a heuristic filtering step, we automatically removed videos with duration <10 seconds or >5 hours, width <480 pixels or height >360 pixels, and frame rate <15fps or >60fps. From inspection, this excluded a negligible amount of desirable videos. The one class of useful videos one might expect this to exclude, short isolated sign videos as used by MS-ASL [20] and WLASL [23], tends to have the label in the video title or description rather than captions, so removing videos under 10 seconds does not have a substantial impact. Finally, we used off-the-shelf person detection tools to exclude videos where none of the captions corresponded to spans with exactly one person present in the video. We limit the scope of our efforts to signing monologues due to the challenges of modeling conversations between multiple signers.

The result was a list of 88,002 candidate videos that might contain ASL with high-quality captions.

3.2 Identifying High-Quality Videos with Skilled Human Annotators

While some smaller datasets like How2Sign [13] use annotators to manually align all captions, this becomes prohibitively expensive for larger datasets. For this reason, OpenASL [37] and BOBSL [2] use annotators to correct only their validation and test sets. We take a coarser-grained approach to annotations but apply it to our entire list of 88,002 candidates: we use humans to identify videos that are roughly suitable and include them in our corpus without modification.

To do so, we hired 3 native ASL users with English proficiency to serve as annotators. The annotators used a bespoke internal tool that would display a given YouTube video and present label options. In order to save time, the annotators were able to mark that their labels held for an entire channel of videos rather than each video individually. Therefore it is possible that certain videos in the corpus are channel outliers and do not meet quality standards, but generally large channels have consistent quality. Each video was labelled by only one annotator unless they brought it up for wider discussion.

Through an iterative process involving written instructions, virtual meetings (through an ASL interpreter or signing project members), and escalations by email for edge cases, we aligned on standards for when to accept a video into the corpus. Some of the reasons for exclusion include: the video's captions do not exclusively correspond to signing; the video is in a sign language other than ASL; the video's captions do not correctly translate the ASL; and the captions are poorly aligned. Notably, in order to increase the size of the corpus, we chose to include videos across all skill levels and signing styles, as long as they were comprehensible to an ASL user and correctly captioned. This variety is beneficial for sign language recognition tasks, where models should be able to understand all signers, but may limit the corpus's usefulness for generation tasks, where consistency and controllability are important.

The result was a list of 11,093 videos whose captions are generally well-aligned English translations of signed ASL content.

3.3 Corpus Statistics

The final, human-filtered YouTube-ASL corpus consists of 11,093 ASL videos with 984 total hours of footage. This is ~3x the size of OpenASL [37], the largest prior ASL dataset, but smaller than BOBSL [2], a British Sign Language dataset. See Table 1 for a comparison between the high-level attributes of YouTube-ASL and prior sign language translation datasets, including total number of hours.

These videos are paired with 610,193 English captions, with a total duration of 813 hours. See Table 2 for statistics on the distribution of captions, as well as Figure 1 for visualizations. The average caption length (8.8 words) and duration (4.8 seconds) are relatively short, which reflects that sentences may be split across multiple captions. We computed vocabulary size by counting the number of distinct strings between whitespace or punctuation across all captions. It is important to keep in mind that in addition to the signing itself, these videos' captions vary in style, literalness of translation (whether the content was originally produced in ASL and translated, or translated into

Table 2: Statistics on the distribution of captions and videos in the YouTube-ASL corpus.

Number of captions	610,193
Caption length (Average / 90^{th} percentile, in characters)	48.9 / 88.0
Caption length (Average / 90^{th} percentile, in words)	8.8 / 16.0
Caption duration (Average / 90^{th} percentile, in seconds)	4.8 / 8.76
Video duration (Average / 90^{th} percentile, in seconds)	318.95 / 675.80

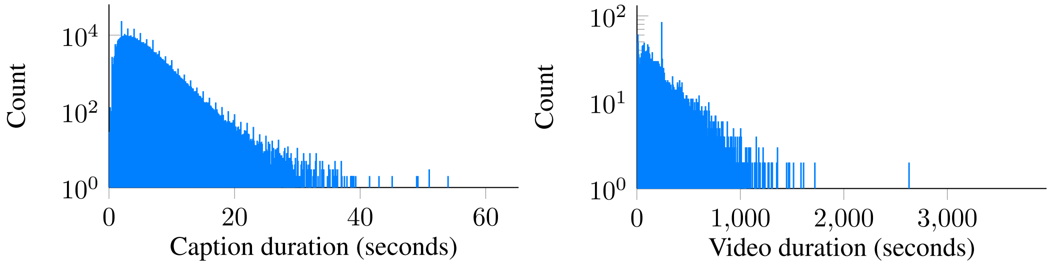

Figure 1: Distribution of caption and video durations. For the video duration graph, we omit 27 videos whose duration exceeds 3600 seconds (between 3610 and 9017 seconds).

ASL from these captions), spelling/grammar correctness, and more. This degree of variability is difficult to quantify in comparisons between datasets.

We use the number of unique channels, 2519, as an approximate lower bound for the number of unique signers in the dataset: some channels may feature many signers, and some signers may appear across multiple channels. Note that with this method, OpenASL [37] would be estimated to have 3 signers, while its authors reached a count of 220 signers using more fine-grained methods. Even this likely underestimate is ~10x the count of any individual sign language dataset to date.

Figure 2 shows the distribution of videos per channel, for channels with at least 20 videos. There are a few channels with many videos—in particular, the two largest channels are the same news channels featured in OpenASL—and then a long tail of channels with fewer videos. This means that the bulk of new footage present in YouTube-ASL but not OpenASL comes from relatively small channels, which helps variety. See Figure 3 for a sense of the distribution of (machine-annotated) topics across videos: they seem more diverse than prior datasets from video sharing platforms but still shaped by typical YouTube use cases, compared to BOBSL's more topic-balanced BBC programming.

4 Baseline Approach

In order to demonstrate the potential of YouTube-ASL, we consider a simple method for sentence-level machine translation from ASL to English built using off-the-shelf components. We use a deliberately barebones approach to avoid introducing inductive bias that helps in more limited settings but becomes harmful with scale.

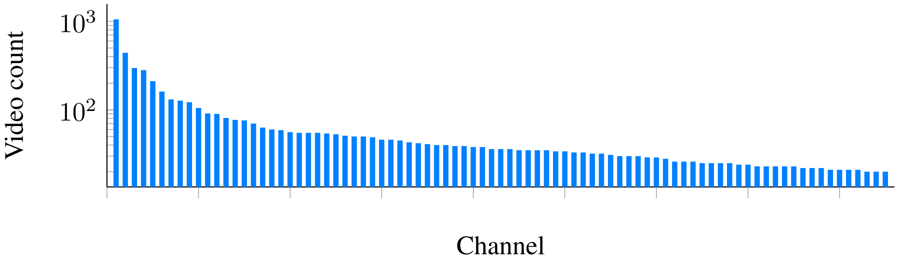

Figure 2: Distribution of videos per channel for channels with at least 20 videos.

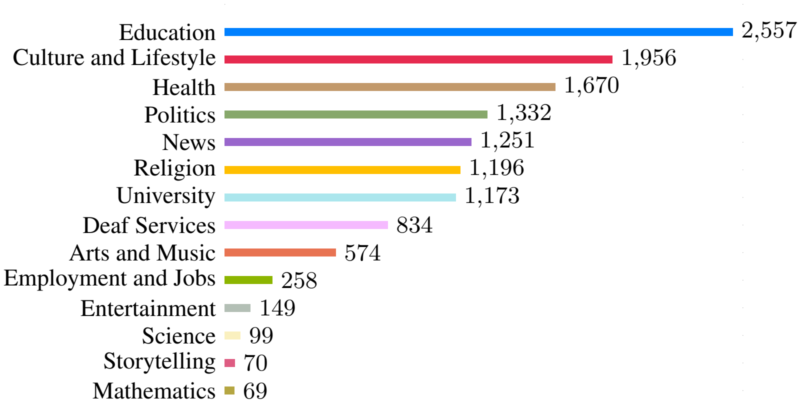

Figure 3: A selection of high-level topics, with the number of YouTube-ASL videos automatically tagged as related to them. Note that a single video can be tagged with more than one topic.

4.1 Preprocessing

For our target English outputs, we use the raw captions from YouTube-ASL. Each training example is clipped to the boundaries of a single caption. We filter out captions with length >300 characters or duration <200ms or >60s, which tend to be malformed, and any captions corresponding to video spans where exactly one person is not present. We do not lowercase the captions or apply any other kind of text normalization.

For our sign language inputs, we use MediaPipe Holistic landmarks [25, 16], rather than raw video. Sign language models that use pose-based inputs have a history of underperforming those that operate on learned video embeddings [20, 26]; it is unclear to what extent this is due to the information bottleneck in the (imperfectly predicted) pose representation, vs. availability of higher quality pretrained video encoders than pretrained pose encoders. Pose inputs offer some benefits like computational efficiency and privacy.

MediaPipe Holistic is a lightweight model that predicts 532 3D landmarks (in x-, y-, and z- image-space coordinates) for the hands, pose, and face of a single human in video footage. For sign language understanding tasks, many of these landmarks are redundant (high-detail face mesh) or unnecessary (lower body), and add undesirable complexity. We discard all but 85 of these points, selected *a priori* using domain knowledge about sign languages:

- For each hand, we use all 21 landmark points.
- For the pose, we use 6 landmark points, for the shoulders, elbows and hips.[4] This discards the lower body and pose landmarks redundant with the hand and face modules.
- For the face, we use 37 landmark points, from the eyes, eyebrows, lips, and face outline.[5]

We normalize the landmarks by scaling them to fit in a unit bounding box across the duration of the clip. We represent landmarks that are not present in a frame with a large negative value. MediaPipe also predicts visibility (self-occlusion) of landmarks within the frame, which we ignore. To reduce sequence length, we discard every second frame. The final preprocessed input is therefore a half-frame rate sequence of 255-dimensional landmark vectors. Note that this half frame rate may vary from 7.5 to 30fps depending on the original video's frame rate, though most end up at 12 to 15 fps.

4.2 Model

Our model is a slightly modified version of T5 [32], which is an encoder-decoder Transformer [39] that has been trained on web-crawled English text. Rather than embed text tokens using a vocabulary

[4] These are indices 11, 12, 13, 14, 23, 24.

[5] These are indices 0, 4, 13, 14, 17, 33, 37, 39, 46, 52, 55, 61, 64, 81, 82, 93, 133, 151, 152, 159, 172, 178, 181, 263, 269, 276, 282, 285, 291, 294, 311, 323, 362, 386, 397, 468, 473.

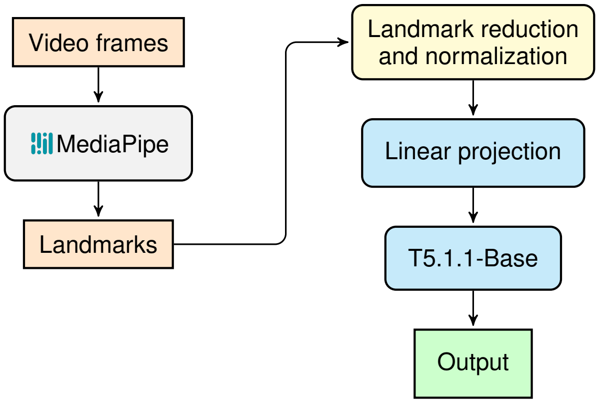

Figure 4: Overview of our model pipeline. Starting from an ASL video clip, we use MediaPipe Holistic to compute 3D landmarks for the face, hands, and body of the subject. We then discard irrelevant landmarks and normalize the remainder. These are concatenated and embedded by a linear projection layer into T5.1.1-Base, which then decodes the English translation. The blue components (Linear projection and T5) are the trainable parameters.

of learned embeddings, we embed each 255-dimensional landmark frame into the encoder using a learned linear projection layer. Otherwise, our architecture is identical to T5.1.1-Base.

We set the encoder context window to 256 tokens (frames) and the decoder context window to 128 tokens, which accommodate the training examples after halving the input frame rate and encoding the target text with T5's SentencePiece vocabulary [22].

5 Experiments

We choose not to provide train, validation, and test splits for YouTube-ASL. Because YouTube videos may be deleted over time, the validation and test splits could not serve as a stable benchmark. We instead evaluate on How2Sign [13], a studio-recorded dataset released under CC BY-NC 4.0 consisting of "How To" instructional narratives translated from English into ASL. This also allows us to integrate trends towards more robust evaluation from speech and text modeling [31, 15, 10], where models trained on large web corpora are evaluated both zero-shot and finetuned on independently constructed benchmarks.

Practices for constructing test sets in prior sign language dataset works are mixed. For example, OpenASL [37] and AfriSign [17] construct their test sets by randomly splitting at the sentence level; SP-10 [40] does the same but with multiway translations identified as a single sentence. How2Sign [13] samples document-level narratives rather than individual sentences, but most signers are shared between the train and test sets, and some narratives are present in both the train and test sets, translated by different signers. BOBSL [2] invests substantial effort into creating signer-independent, topic-balanced splits; this is perhaps why its translation baseline scores only 1.00 BLEU despite the dataset's size. Zero-shot evaluation lets us sidestep these issues and get a better sense of the model's quality for real use.

5.1 Setup

We ablate across four different training schedules:

- **H2S**. We train only on How2Sign, not YouTube-ASL, for a like-for-like comparison with prior methods.
- **YT-ASL**: We train only on YouTube-ASL, and evaluate on How2Sign zero-shot.
- **YT-ASL + H2S**: We train on a mixture of How2Sign and YouTube-ASL, mixed in proportion to the size of the datasets.
- **YT-ASL → H2S**: We train on YouTube-ASL, then finetune on How2Sign.

Table 3: Metrics for ASL to English translation on How2Sign. Our models either train from scratch or finetune a pretrained T5 checkpoint, and are trained on How2Sign (H2S) only, YouTube-ASL (YT-ASL) only, a mixture of H2S and YT-ASL, or YT-ASL and then finetuned on H2S.

Approach	Training Schedule	BLEU-1	BLEU-2	BLEU-3	BLEU	BLEURT
Álvarez et al. [3]	H2S	17.40	7.69	3.97	2.21	-
GloFE-VN [24]	H2S	14.94	7.27	3.93	2.24	31.65
Tarrés et al.[38]	H2S	34.01	19.30	12.18	8.03	-
Ours (no pretraining)	H2S	13.92	4.69	1.82	0.86	30.65
	YT-ASL	14.53	5.47	2.61	1.41	29.55
	YT-ASL + H2S	28.60	14.56	8.68	5.60	37.72
	YT-ASL → H2S	28.38	15.41	9.55	6.26	39.40
Ours (pretrained)	H2S	14.96	5.11	2.26	1.22	29.98
	YT-ASL	20.93	10.35	6.14	3.95	34.98
	YT-ASL + H2S	36.35	23.00	16.13	11.89	44.78
	YT-ASL → H2S	**37.82**	**24.13**	**16.92**	**12.39**	**46.63**

We also ablate the effect of pretraining on English text by comparing models trained from scratch using the T5.1.1-Base architecture, vs. finetuned from the T5.1.1-Base pretrained checkpoint.

We train with a batch size of 128 and learning rate of 0.001 with Adafactor [34]; other hyperparameters are the T5X defaults. For models trained solely on How2Sign data, we train for 20,000 steps. For models trained on YouTube-ASL (including with How2Sign mixed in), we train for 200,000 steps. When finetuning on How2Sign after training on YouTube-ASL, we finetune for an additional 5,000 steps. Each 1,000 steps takes approximately 0.25 TPUv4-hours.

Following prior work, we present BLEU [28] and BLEURT [33] scores. BLEU scores are computed using SacreBLEU [29] version 2, with all default options. BLEURT scores are computed using checkpoint BLEURT-20 [30, 14]. We decode using beam search with a beam width of 5.

5.2 Results

See Table 3 for metrics comparing our models to prior works on How2Sign [3, 24, 38]. The best results come from training on YouTube-ASL from a pretrained checkpoint, then finetuning on How2Sign, which achieves 12.39 BLEU vs. the state of the art of 8.03 BLEU [38]. The base model achieves 3.95 BLEU zero-shot, which is nontrivial but substantially worse than the finetuned score. Factors that could contribute to this gap include train/test leakage of signers and narratives, How2Sign's narrow domain, and the extra ~10% training data it represents.

Results are substantially worse when training from scratch, which suggests that T5's English pretraining gives the model a better initialization, as De Coster et al. [11] found for frozen pretrained language models. Results are absymal when trained without YouTube-ASL. The most direct comparison of our approach to prior work is T5 trained from scratch on How2Sign only, which reaches just 0.86 BLEU, despite training on the same data as Tarrés et al. [38]'s 8.03 BLEU. This might be explained by their use of a pretrained video encoder and various decisions they made to optimize for small amounts of data (smaller network, more text normalization, careful hyperparameter sweep), whereas we used a less tuned configuration that was intended for larger datasets.

See Table 4 for qualitative examples of the translations produced by our best finetuned and zero-shot models, on sentences sampled from How2Sign by Tarrés et al. [38]. The translations capture elements of the reference translation but are clearly not yet of usable quality. The zero-shot predictions hew less closely to the references, but the errors usually make sense in light of the sign language input. For example, in (1), the sign used to mean "defense" also means "barrier".

6 Conclusion

In this paper, we presented YouTube-ASL, a new, publicly available parallel corpus for American Sign Language and English that is ~3x the size and has ~10x as many unique signers as the largest

Table 4: Qualitative examples from our best finetuned and zero-shot models, on sentences sampled from How2Sign by Tarrés et al. [38]. See Table 5 in the Appendix for the complete set of examples.

(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.

prior ASL dataset. Our key improvement over prior work is that we used automatic tagging followed by human filtering to increase mining recall without harming precision. We demonstrated the value of this data with a simple baseline built from off-the-shelf components (MediaPipe Holistic and T5) that achieves a new finetuned state of the art in ASL to English translation on How2Sign, 12.39 BLEU. We also reported a zero-shot score of 3.95 BLEU, a first for sign language translation. We hope that YouTube-ASL will be immediately useful for research on methods for sign language translation and caption alignment, as well as tools for automatic annotation/filtering of new sign language datasets. Because YouTube-ASL has so much signer variety, including across dialect and skill level, it may be less useful for generation than recognition tasks.

While our baseline improves upon prior work, even the finetuned translations are subjectively low-quality and are not yet useful in the real world. We hope that more refined modeling approaches will provide better results with the same data, but despite our and prior efforts, ASL is still a low-resource language by modern standards [19]—let alone the many other sign languages of the world, most of which are even less resourced. Future work may look to address this by mining broader datasets with other kinds of supervision, and as model quality improves, it should perform more comprehensive evaluations to understand differences across domains, dialects, levels of fluency, signer appearance, and other such factors.

7 Ethical Considerations

Sign language datasets pose privacy challenges, because the signer's appearance (body, facial expressions), which is personally identifying, is a vehicle for the language itself. Video anonymization techniques are not yet mature enough to be useful in this regard. Our corpus is composed of sign language content that uploaders made publicly visible on YouTube, and we release only video IDs so that changes to the underlying videos are automatically reflected in the corpus. While the corpus covers a broader variety of channels than prior works, this does not mean it is necessarily representative of the signing population—or even if it were representative, that models trained on it would work equally well for everyone.

We train our models on reduced poses as a form of anonymization, but this is not suitable for all modeling approaches and may harm model quality. Until sign language translation models are closer to usable quality, there is little risk of societal harm, except that individuals or organizations mistakenly rely on models that are inadequate. As we approach that point, sign language processing will adopt the risks of natural language processing in general, but with a great potential to improve accessibility for Deaf/Hard of Hearing people.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark, 2016. URL https://arxiv.org/abs/1609.08675.

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BBC-Oxford British Sign Language dataset, 2021. URL https://arxiv.org/abs/2111.03635.

[3] Patricia Cabot Álvarez, Xavier Giró Nieto, and Laia Tarrés Benet. Sign language translation based on transformers for the How2Sign dataset. 2022. URL https://imatge.upc.edu/web/publications/sign-language-translation-based-transformers-how2sign-dataset.

[4] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366762. doi: 10.1145/3308561.3353774. URL https://doi.org/10.1145/3308561.3353774.

[5] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. URL https://doi.org/10.1109/CVPR.2018.00812.

[6] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL https://doi.org/10.1109/CVPR42600.2020.01004.

[7] Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. Content4all open research sign language translation datasets, 2021. URL https://arxiv.org/abs/2105.02351.

[8] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(01):172–186, jan 2021. ISSN 1939-3539. doi: 10.1109/TPAMI.2019.2929257. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2929257.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. URL https://doi.org/10.1109/CVPR.2017.502.

[10] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech, 2022.

[11] Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. Frozen pretrained transformers for neural sign language translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 88–97, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-at4ssl.10.

[12] Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27, 2023. URL https://doi.org/10.1007/s10209-023-00992-1.

[13] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A large-scale multimodal dataset for continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2735–2744, June 2021. URL https://doi.org/10.1109/CVPR46437.2021.00276.

[14] Google-Research. Google-research/bleurt: Bleurt is a metric for natural language generation based on transfer learning. URL https://github.com/google-research/bleurt.

[15] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation, 2021.

[16] Ivan Grishchenko and Valentin Bazarevsky. Mediapipe holistic - simultaneous face, hand and pose prediction, on device, Dec 2020. URL https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html.

[17] Shester Gueuwou, Kate Takyi, Mathias Müller, Marco Stanley Nyarko, Richard Adade, and Rose-Mary Owusuaa Mensah Gyening. Afrisign: Machine translation for african sign languages. In *4th Workshop on African Natural Language Processing*, 2023. URL https://openreview.net/forum?id=EHldk3J2xk.

[18] Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-54-2. URL https://aclanthology.org/2020.signlang-1.12.

[19] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world, 2021. URL https://arxiv.org/abs/2004.09095.

[20] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding american sign language. *CoRR*, abs/1812.01053, 2018. URL http://arxiv.org/abs/1812.01053.

[21] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13), 2019. ISSN 2076-3417. doi: 10.3390/app9132683. URL https://www.mdpi.com/2076-3417/9/13/2683.

[22] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

[23] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020. URL https://doi.org/10.1109/WACV45572.2020.9093512.

[24] Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. Gloss-free end-to-end sign language translation, 2023. URL https://arxiv.org/abs/2305.12876.

[25] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines, 2019. URL https://arxiv.org/abs/1906.08172.

[26] Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgöz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 3434–3440, June 2021. URL https://doi.org/10.1109/CVPRW53098.2021.00382.

[27] Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.71.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[29] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

[30] Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for mt. In *Proceedings of EMNLP*, 2021.

[31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435. URL https://dl.acm.org/doi/abs/10.5555/3455716.3455856.

[33] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

[34] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/shazeer18a.html.

[35] Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Jonathan Michaux, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. American Sign Language fingerspelling recognition in the wild. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 145–152, 2018. URL https://doi.org/10.1109/SLT.2018.8639639.

[36] Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Fingerspelling recognition in the wild with iterative visual attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. URL https://doi.org/10.1109/ICCV.2019.00550.

[37] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video, 2022. URL https://arxiv.org/abs/2205.12870.

[38] Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. Sign language translation from instructional videos, 2023. URL https://arxiv.org/abs/2304.06371.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[40] Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. MLSLT: Towards multilingual sign language translation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5109, 2022. URL https://doi.org/10.1109/CVPR52688.2022.00505.

[41] Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.525. URL https://aclanthology.org/2020.coling-main.525.

[42] Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing, 2021. URL https://arxiv.org/abs/2105.05222.

[43] Biao Zhang, Mathias Müller, and Rico Sennrich. SLTUNET: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=EBS4C77p_5S.

[44] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition, 2020.

[45] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, June 2021. URL https://doi.org/10.1109/CVPR46437.2021.00137.

A Appendix

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.