

Towards Privacy-Aware Sign Language Translation at Scale

Phillip Rust^{♥,†} Bowen Shi[◇] Skyler Wang^{◇,♣}
Necati Cihan Camgöz[◇] Jean Maillard^{◇,‡}

♥ University of Copenhagen ◇ FAIR at Meta ◇ Meta ♣ UC Berkeley
† p.rust@di.ku.dk ‡ jeanm@meta.com

Abstract

A major impediment to the advancement of sign language translation (SLT) is data scarcity. Much of the sign language data currently available on the web cannot be used for training supervised models due to the lack of aligned captions. Furthermore, scaling SLT using large-scale web-scraped datasets bears privacy risks due to the presence of biometric information, which the responsible development of SLT technologies should account for. In this work, we propose a two-stage framework for privacy-aware SLT at scale that addresses both of these issues. We introduce SSVP-SLT, which leverages self-supervised video pretraining on anonymized and unannotated videos, followed by supervised SLT finetuning on a curated parallel dataset. SSVP-SLT achieves state-of-the-art finetuned and zero-shot gloss-free SLT performance on the How2Sign dataset, outperforming the strongest respective baselines by over 3 BLEU-4. Based on controlled experiments, we further discuss the advantages and limitations of self-supervised pretraining and anonymization via facial obfuscation for SLT.¹

1 Introduction

Used by millions worldwide, sign languages play a crucial role in facilitating communication for many d/Deaf and hard-of-hearing individuals. Visual in nature, these languages make use of the co-articulated features of hands (i.e., finger positioning, shape, movement, palm orientation, etc.), body postures, gaze, mouth gestures, mouthings and facial expressions to convey meaning (Stokoe, 1980). Globally, there are more than 300 sign languages, each with their own grammar and vocabulary.² American Sign Language (ASL) alone

¹Code and data are available at <https://github.com/facebookresearch/ssvp-slt>.

²<https://www.un.org/en/observances/sign-languages-day>

| STAGE | I. | II. |
|------------|------------------|-----------------------------|
| DATA | | |
| SCALE | Large | Smaller |
| SOURCE | Web-mined | Hand-curated |
| ANNOTATED | ✗ | ✓ |
| ANONYMIZED | ✓ | ✓/✗ (with explicit consent) |
| TRAINING | Self-supervised | Supervised |
| OUTPUT | Pretrained model | Translations |

Table 1: Our proposed generic, scalable and privacy-aware SLT framework. We make no assumptions about model architecture and anonymization method.

is estimated to have more than half a million native users, ranking it among the most commonly used languages in the United States (Mitchell and Young, 2022).

Despite the prevalence of sign languages, they are still under-served by translation technology. Besides under-investment (Yin et al., 2021) and the inherent difficulty of SLT,³ another key explanation for this imbalance is the lack of sufficiently large, clean, and labeled parallel corpora. Current state-of-the-art SLT systems require detailed and time aligned annotations (Zhou et al., 2023; Uthus et al., 2023), which is not scalable, as annotating sign language data is a labour intensive task and can only be done by proficient signers.

We argue that a promising solution to SLT’s data scarcity is to utilize publicly available *unannotated* sign language data.⁴ In other domains of computer vision and NLP, a common practice is to pretrain on large-scale unannotated web datasets and later finetune on curated, task-specific datasets (Devlin et al., 2019; Radford et al., 2018, 2019; Raffel et al., 2020; He et al., 2022). This practice is largely unexplored in the SLT domain and comes with additional challenges. In particular, moving

³Results of the WMT 2023 SLT task evince this difficulty; the best system only achieved ~1 BLEU (Müller et al., 2023).

⁴For example, Uthus et al. (2023) filtered their Youtube-ASL dataset from ~88K to 11K videos based largely on the availability and quality of English captions.

to large-scale sign language processing makes it increasingly difficult to control the composition of the training data. Because sign language videos typically feature faces and upper bodies and thus are biometrically identifying, such research may exacerbate privacy risks. Hence, developing sign language technologies responsibly requires us to account for these risks and explore techniques to protect privacy.

In this work, we study the effectiveness of self-supervised video pretraining for SLT, under consideration of the aforementioned privacy risks. We first propose a *generic, scalable and privacy-aware* two-stage framework for SLT, summarized in Table 1. We introduce *SSVP-SLT* (Self-Supervised Video Pretraining for Sign Language Translation), an implementation of this framework consisting of two or optionally three stages: pretraining a continuous sign language encoder via masked autoencoding (MAE; He et al., 2022) on anonymized video, then optionally bridging the modality gap via CLIP-style video-text pretraining (Radford et al., 2021), and finally training an SLT system via supervised finetuning using extracted features from the pretrained model. Our best performing models achieve 15.5 BLEU finetuned and 7.1 BLEU zero-shot on the How2Sign dataset (Duarte et al., 2021), surpassing SOTA in both settings by over 3 BLEU while using data anonymized via facial obfuscation. We also introduce a new ASL-to-English SLT benchmark dataset, *DailyMoth-70h*, consisting of ~70h of continuous signing in native ASL. We then evaluate the downstream performance impact and discuss the benefits and limitations of facial blurring to achieve anonymization. Through controlled ablation studies of SSVP-SLT, we identify what factors contribute to a strong pretraining and finetuning recipe. We conclude by discussing opportunities and challenges of self-supervised pretraining for sign language processing.

2 Background and Related Work

Gloss-free SLT Glosses are a way of representing individual signs into a written form. Being monotonically aligned to signs, they can be a useful medium between sign and spoken languages. Most SLT approaches to date rely on them (Chen et al., 2022a,b; Zhang et al., 2023). The task of predicting glosses from continuous signing is typically performed via gloss supervision either

jointly or in a cascaded approach with supervised SLT finetuning (Camgöz et al., 2018, 2020).

However, glosses are also considered an incomplete and inaccurate representation of sign language (Camgöz et al., 2020; Müller et al., 2023). Furthermore gloss annotation is a labour intensive task. Due to these constraints, there is a growing body of research on gloss-free SLT. Most such approaches incorporate techniques aimed at reducing the modality gap, such as training the visual encoder via sign spotting (Tarrés et al., 2023; Shi et al., 2022),⁵ adding inductive bias in the attention mechanism (Yin et al., 2023), using conceptual anchor words (Lin et al., 2023), or performing visual-language pretraining (Zhou et al., 2023). Uthus et al. (2023) also benefit from a pretrained text model (T5; Raffel et al., 2020). Similar to Zhou et al. (2023), we also leverage language supervision to reduce the modality gap, albeit in conjunction with self-supervised video pretraining.

Sign Language Video Anonymization Sign language videos typically feature signers’ faces, which convey essential linguistic information. However, in virtual domains, particularly in spaces involving the discussion of sensitive topics, exposing one’s face (and identity) may lead to various forms of personal risks. Such exposures could even lead to harm associated with professional or insurance discrimination. For these reasons, the d/Deaf and hard-of-hearing community has long expressed interest in anonymization and privacy protection techniques for sign language content (Lee et al., 2021), and sign language video anonymization has, in recent years, become an active area of research (Isard, 2020; Xia et al., 2024).

For general-domain images and videos, simple anonymization techniques such as facial obfuscation via overlaying or blurring are widely used and accepted (Frome et al., 2009; Yang et al., 2022). In the sign language domain, such techniques may be inadequate due to the loss of information in the facial region (Lee et al., 2021). More specifically, in signed language, non-manual features such as mouthing, eyebrow and head movements are used extensively to enrich grammar. Certain signs with similar manual features are only disambiguated through mouth morphemes. Moreover,

⁵While Tarrés et al. (2023) do not explicitly mention the use of sign spotting, they rely on features extracted from an I3D model (Carreira and Zisserman, 2017) by Duarte et al. (2022), who used an iterative sign spotting technique.

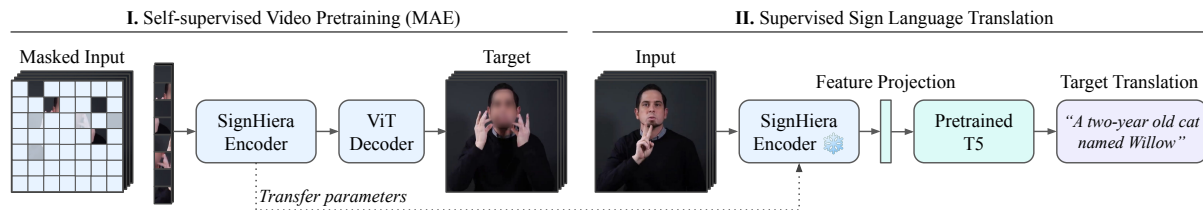


Figure 1: Overview of our two-stage SSVP-SLT method. The first stage consists of training a SignHiera encoder via masked autoencoding (MAE) on *blurred* video frames. In the second stage, a pretrained T5 model is finetuned for SLT while the pretrained SignHiera is kept frozen (❄️). The input video in the second stage *can be unblurred*.

facial expressions are often used to indicate emphasis, negation, a question, etc. (Baker-Shenk, 1985; Valli and Lucas, 2000; Neidle et al., 2000).

Recent work has focused on anonymizing sign language content via avatars (Tze et al., 2022; Lee et al., 2021) and transferred or synthetic human appearances (Saunders et al., 2021; Xia et al., 2022, 2024). While promising, these approaches are nascent and we are unaware of studies that determine to what extent models can learn to recover or disambiguate obfuscated linguistic information from context. That being said, human studies suggest that signers struggle to comprehend content anonymized in such a way (Lee et al., 2021).

Lacking an obvious alternative, in this work we return to the relatively straightforward technique of facial blurring. Despite its limitations, we demonstrate that blurring can raise privacy protection with little performance degradation.⁶ This, we argue, can facilitate large-scale video anonymization when applied to publicly available sign language data.

Masked Autoencoding for Video and Beyond

Following its success as a self-supervised learning paradigm in the image domain (He et al., 2022), MAE has been widely applied in other areas, including audio (Huang et al., 2022), language (Rust et al., 2023), and video (Feichtenhofer et al., 2022; Tong et al., 2022; Wang et al., 2023). Considering that MAEs have been shown to be capable of acquiring both language and basic video understanding from pixels alone, it is conceivable that high-quality sign language representations can be learned directly from RGB video data via MAE, given enough data. Recently, Sandoval-Castaneda et al. (2023) explored MAE among other self-supervised learning techniques in the context of isolated sign language recogni-

⁶In Appendix A we discuss issues with pose landmarks, often promoted as a privacy-preserving alternative to video.

tion (ISLR) and found it to be highly useful. MAE has, however, not yet been successfully applied to SLT. In SLT, videos are much longer, and learning high-quality representations requires models to capture long-range spatiotemporal dependencies. Our usage of MAE, or self-supervised pretraining in general, therefore stands in contrast to recent SLT methods, gloss-based and gloss-free methods alike, which instead fully rely on supervised training that requires annotated captions (Zhou et al., 2023; Tarrés et al., 2023; Lin et al., 2023; Uthus et al., 2023).

3 Generic Framework

We first outline a generic, scalable and privacy-aware two-stage transfer learning framework for SLT (See Table 1).

In **Stage I**, we train a model, with the goal to learn high-quality continuous sign language representations, via self-supervised learning. The data used at this stage is *always anonymized*. We make no assumptions on how the data may be anonymized, i.e. face blurring as discussed in §2, or more sophisticated methods such as using synthetic appearances.

In **Stage II**, we finetune the model from the first stage in a supervised manner using a smaller and hand-curated parallel dataset. Ideally, the finetuning dataset, being manageable in size, can be unanonymized after gaining explicit consent from signers in the data to minimize information loss.

4 Method

The base implementation of our framework is designed as a two-step approach, termed **SSVP-SLT**. We provide a high-level overview in Fig. 1.

Self-Supervised Video Pretraining (MAE) We first aimed to pretrain a capable sign language encoder on video data alone—no gloss, pseudo-gloss, or spoken-language text annotations—

allowing us to leverage large amounts of unannotated sign language video in the future, alleviating the data scarcity issue in SLT training.

Our sign video encoder, *SignHiera*, builds on Hiera (Ryali et al., 2023), a vision transformer that combines a hierarchical architecture, shown to be crucial for learning phonetically meaningful sign representations (Sandoval-Castaneda et al., 2023) with masked autoencoding (MAE), a widely used self-supervised learning paradigm (He et al., 2022). Its hierarchical architecture also makes Hiera more efficient to train than other MAE-based video transformers such as VideoMAE (Tong et al., 2022; Wang et al., 2023) or MAE-ST (Feichtenhofer et al., 2022).

Hiera embeds a video into a sequence of spatio-temporal tokens. A large percentage of tokens is randomly masked, while the rest is passed through a hierarchical transformer stack with several pooling operations. The decoder receives fused multi-scale features extracted before each pooling operation and processes them via a lightweight transformer stack. A final linear projection yields pixel logits. The loss is computed as the normalized mean squared error between the original and predicted pixel values of the masked tokens.

SignHiera is initialized from the original Hiera-Base-16 \times 224 checkpoint pretrained on Kinetics-400 (Kay et al., 2017). In order to capture longer-range spatio-temporal dependencies in signed utterances, we increased the clip length from 16 to 128 video frames, leading to an 8 \times sequence length, and accordingly resized and reinitialized the position embeddings. We further added attention masking to accommodate shorter videos with temporal padding and added a third Q-pooling operation after the last encoder stage to save GPU memory. We trained with a masking ratio of 0.9.

Supervised SLT Finetuning The translation network is an encoder-decoder transformer model (Vaswani et al., 2017). Our default configuration uses a pretrained T5-v1.1 (Raffel et al., 2020; Shazeer, 2020), following Uthus et al. (2023). We also experimented with BART (Lewis et al., 2020) and Tarrés et al. (2023)’s setup, training a \sim 10M parameter transformer from scratch.

The only difference from a text transformer is that our translation network takes in video features extracted from the pretrained SignHiera. We used SignHiera’s final intermediate representations, which are of size $\mathbb{R}^{B \times \frac{T}{2} \times \frac{H}{32} \times \frac{W}{32} \times D}$, where

B is the batch size, $T=128$, $H=W=224$ is the input video size, and $D=768$ is the feature size. We mean-pooled over the spatial dimensions to obtain feature vectors of size $\mathbb{R}^{B \times \frac{T}{2} \times D}$. Videos shorter than 128 frames were padded for encoding, and the padding was then removed from the extracted features. For longer videos, we used a sliding window with stride $\frac{T}{2}$ and concatenated the resulting features. A linear projection $W_{\text{proj}} \in \mathbb{R}^{D \times D'}$ mapped the extracted features to a sequence of size $\mathbb{R}^{B \times S \times D'}$, with S being the sequence length of the extracted features and D' the transformer’s hidden size. This sequence was processed by the transformer as usual (Vaswani et al., 2017).

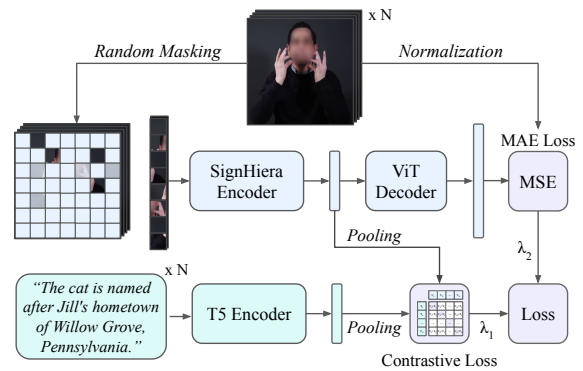


Figure 2: Overview of our LSP extension.

Adding Language-supervised Pretraining We also experimented with extending SSVP-SLT with a language-supervised pretraining (LSP) step to bridge the modality gap between the input videos and text translations. Bridging this gap may improve gloss-free SLT performance, as discussed in §2. In what follows, we refer to SSVP-SLT with an additional LSP stage as **SSVP-SLT-LSP**. The LSP stage fits *in between* the self-supervised MAE pretraining and the supervised SLT finetuning stage. Since language supervision requires annotations, the LSP stage should be considered a part of stage II in our generic framework.

Our LSP approach is illustrated in Figure 2. We first initialized a CLIP model (Radford et al., 2021) with our MAE-pretrained SignHiera encoder as the vision tower and a pretrained T5-v1.1 encoder as the text tower. We then jointly trained the CLIP model via contrastive video-text pretraining and the SignHiera model via MAE. The goal is to help the SignHiera encoder, which is involved in both tasks, learn strong continuous sign representations grounded in the target

modality (text). The videos were masked with a 90% ratio even for computing the contrastive loss, which is similar to FLIP (Li et al., 2023), and enabled end-to-end training by drastically reducing the memory footprint. The two losses (MAE and contrastive) were balanced via GradNorm (Chen et al., 2018), which helped stabilize training, compared to using fixed loss weights.⁷

5 Experimental Setup

5.1 Datasets

Youtube-ASL (YT) We used Youtube-ASL (Uthus et al., 2023), the largest available ASL training dataset with ~1000h of in-the-wild video and over 2500 signers, both during pretraining and supervised finetuning.⁸ We used a version in which all signers’ faces are blurred for privacy.

How2Sign (H2S) We also used How2Sign (Duarte et al., 2021), an ASL dataset with ~80h of video and nine different signers in a green screen studio setting, for pretraining, finetuning, and downstream evaluation of our SLT models. Again, we used a version with blurred faces only.

DailyMoth-70h To isolate the impact of face blurring during pretraining and finetuning on SLT performance, we relied on a new dataset, which we name *DailyMoth-70h*. This dataset contains over 70h of video of a single signer from the ASL news page TheDailyMoth and was obtained through a license agreement with TheDailyMoth’s host.⁹ We used both unblurred and blurred dataset versions and report dataset statistics in Appendix B.

5.2 Training and Evaluation Protocols

We briefly describe our training and evaluation protocols. The full configurations for pretraining and SLT training are listed in Appendix C.1.

Face blurring We used an internal face blurring software and ensured its reliability on the YT and H2S datasets via a combination of automatic and manual verification techniques. Example frames sampled from two blurred videos from

⁷In contrast to Zhou et al. (2023), we did not jointly train the text decoder as doing so did not improve performance in preliminary experiments and led to training instabilities.

⁸For the lack of a readily available larger, unannotated dataset, Youtube-ASL fits both dimensions of our framework: the large, publicly-available, unannotated dataset and the curated, parallel dataset.

⁹<https://www.dailymoth.com/>

the DailyMoth-70h data are shown in Appendix Figure 5.

MAE pretraining We largely follow the Hiera pretraining recipe from Ryali et al. (2023). In our default configuration, we trained for 800 effective epochs (accounting for repeated sampling as in Feichtenhofer et al. (2022)) with the AdamW optimizer (Kingma and Ba, 2015; Loshchilov and Hutter, 2019). The learning rate was set to $8e-4$ with linear warmup over 120 epochs and cosine decay to $1e-5$. The batch size was 2×128 , with 2 being the repeated sampling factor. Similar to Zhou et al. (2023), we employed video data augmentation via random cropping, horizontal flipping, and RandAug (Cubuk et al., 2020). We used a 128×2 temporal sampling strategy, i.e., sampling 128 frames with a stride of 2, which fully accommodates ~85–95% of videos in the data.

SLT finetuning When finetuning only on How2Sign or DailyMoth-70h, we closely followed the setup of Tarrés et al. (2023), training a ~10M parameters transformer from scratch; see Appendix C.1 for more details. For How2Sign, we reused their lowercase unigram tokenizer (vocab size 7K). For DailyMoth-70h, we trained a cased tokenizer (unigram, 7K vocab size), which we found to work better due to the large proportion of named entities in the data.

When finetuning on Youtube-ASL, as we needed a model with more capacity we relied on a pretrained T5-v1.1 with default tokenizer, following Uthus et al. (2023). We trained for up to 100 epochs with early stopping, batch size 512 and AdamW with peak rate $5e-4$, linear warmup over two epochs and cosine decay to $1e-7$. We used dropout of 0.3 and 0.2 label smoothing in the cross-entropy loss. We did not use video data augmentation unless specified otherwise.

Language-supervised pretraining We performed 200 epochs of LSP with a batch size of 512 on top of 600 epochs of MAE-only pretraining. We did not use repeated sampling, which is incompatible with the contrastive loss. We again used AdamW, warming up to a learning rate of $1e-4$ over ten epochs, followed by cosine decay to $1e-6$. The GradNorm optimizer has a one epoch warmup, a peak learning rate of $1e-2$, and decays to $1e-4$. Data augmentation and temporal sampling are the same as for MAE pretraining.

| METHOD | BLUR | FT DATA | BLEU-1 | BLEU-2 | BLEU-3 | BLEU | ROUGE-L | BLEURT |
|---------------------------------------------------|------|----------|-------------|-------------|-------------|-------------|-------------|-------------|
| Baselines | | | | | | | | |
| Lin et al. (2023) | ✗ | H2S | 14.9 | 7.3 | 3.9 | 2.2 | 12.6 | 31.7 |
| Tarrés et al. (2023) | ✗ | H2S | 34.0 | 19.3 | 12.2 | 8.0 | — | — |
| ----- | | | | | | | | |
| Uthus et al. (2023) | ✗ | H2S | 15.0 | 5.1 | 2.3 | 1.2 | — | 30.0 |
| | | YT | 20.9 | 10.4 | 6.1 | 4.0 | — | 35.0 |
| | | YT + H2S | 36.3 | 23.0 | 16.1 | 11.9 | — | 44.8 |
| | | YT → H2S | 37.8 | 24.1 | 16.9 | 12.4 | — | 46.6 |
| Ours | | | | | | | | |
| SSVP-SLT ₈₀₀ ^{H2S} | ✓ | H2S | 30.2 | 16.7 | 10.5 | 7.0 | 25.7 | 39.3 |
| ----- | | | | | | | | |
| SSVP-SLT ₈₀₀ ^{YT} | ✓ | H2S | 38.1 | 23.7 | 16.3 | 11.7 | 33.8 | 44.2 |
| | | YT | 29.2 | 16.6 | 10.7 | 7.1 | 28.3 | 41.8 |
| | | YT + H2S | 41.6 | 27.2 | 19.3 | 14.3 | 36.8 | 48.6 |
| | | YT → H2S | 41.9 | 27.7 | 19.8 | 14.7 | 37.8 | 49.3 |
| ----- | | | | | | | | |
| SSVP-SLT ₈₀₀ ^{YT+H2S} | ✓ | H2S | 38.9 | 24.1 | 16.4 | 11.6 | 34.0 | 44.5 |
| | | YT | 29.1 | 17.0 | 11.1 | 7.5 | 28.6 | 41.6 |
| | | YT + H2S | 41.8 | 27.4 | 19.5 | 14.3 | 36.9 | 48.9 |
| | | YT → H2S | 41.8 | 27.4 | 19.6 | 14.6 | 37.7 | 49.0 |
| ----- | | | | | | | | |
| SSVP-SLT-LSP _{600→200} ^{YT+H2S} | ✓ | YT + H2S | 43.2 | 28.8 | 20.8 | 15.5 | 38.4 | 49.6 |

Table 2: How2Sign test performance of SSVP-SLT in different pretraining configurations compared to baselines. The BLUR column denotes whether faces in the train and eval data are blurred. FT DATA indicates the finetuning configuration; respectively, YT+H2S and YT→H2S refer to training on the two datasets jointly or consecutively.

Evaluation We used beam search with 5 beams and no length penalty. We evaluated every epoch when finetuning on How2Sign or Dailymoth-70h and every 500 steps for Youtube-ASL. We kept the checkpoint with the highest validation BLEU-4 and evaluated it on the respective test set.

Notation Below, we use superscript and subscript to indicate the pretraining dataset and number of epochs, respectively. For instance, SSVP-SLT₈₀₀^{YT+H2S} refers to 800 epochs of MAE pretraining on Youtube-ASL and How2Sign. For SSVP-SLT-LSP, 600→200 denotes 600 epochs of MAE pretraining followed by 200 epochs of LSP.

5.3 Baselines

Lin et al. (2023) propose to bridge the visual and text modalities via contrastive anchoring of encoded visual features to embeddings of conceptual words in the target sequence. Tarrés et al. (2023) is the SOTA on How2Sign without additional SLT data, training a 6+3 layer transformer from scratch on features from an I3D model (Carreira and Zisserman, 2017). The I3D model was first pretrained on Kinetics (Carreira and Zisserman, 2017) and BOBSL (Albanie et al., 2021), and finetuned on How2Sign for sign language recognition using annotations generated via iterative sign spotting (Duarte et al., 2022). Uthus et al. (2023)

is the current SOTA on How2Sign, and finetunes a pretrained T5-v1.1-Base model for SLT directly on pose landmarks extracted from YouTube-ASL and How2Sign videos.

5.4 Evaluation Measures

We report BLEU via SacreBLEU (Papineni et al., 2002; Post, 2018).¹⁰ We also report ROUGE-L (Lin, 2004) and BLEURT (Sellam et al., 2020) from the BLEURT-20 checkpoint, which has been shown to correlate well with human judgments.

6 Results and Discussion

Comparison against the state-of-the-art We present our main results in Table 2. Our best models significantly improve over the previous 12.4 BLEU state-of-the-art by Uthus et al. (2023). SSVP-SLT achieves 14.7 and 14.6 BLEU when pretraining on YT and YT+H2S respectively. Our best overall model, utilizing SSVP-SLT-LSP, achieves 15.5 BLEU, a 3.1 point improvement over the baseline. When pretraining and finetuning on YT only, we also observe a 3.1 BLEU improvement (4.0 vs. 7.1) over the previous SOTA in the zero-shot setting. These results demonstrate

¹⁰nrefs:1|case:mixed|
eff:no|tok:13a|smooth:exp|version:2.3.1

the overall effectiveness of SSVP-SLT and, more broadly, self-supervised pretraining for SLT.

Pretraining on YT+H2S performs almost the same as training on YT only, with the YT-only models even sometimes performing best. Given the distributional gap between the datasets (in-the-wild YT vs. studio H2S video) and the fact that the YT+H2S models consumed more data, this finding is somewhat surprising. While this may be due in part to randomness in the training dynamics, it could also mean that sufficient finetuning can compensate for not accessing the H2S data at pretraining, presumably because the pretraining set is sufficiently large and diverse. This encouraging result suggests that we can pretrain on large data independently of knowing what our finetuning and inference dataset will be—a crucial requirement for practical SLT applications.

We find that YT data is beneficial both for pretraining and finetuning, which emphasizes the importance of training on large and diverse data and suggests that we can expect further gains from scaling to large public unannotated video.

Finally, we find that bridging the modality gap via language-supervised pretraining yields a 1.2 BLEU improvement over its MAE-only counterpart. Given enough annotated data, the technique can be employed independently of self-supervised pretraining at little extra cost.

| BLUR | | BLEU | ROUGE-L | BLEURT |
|----------|-----|------|---------|--------|
| PRETRAIN | SLT | | | |
| ✗ | ✗ | 28.8 | 50.9 | 51.7 |
| ✗ | ✓ | 28.1 | 50.6 | 51.4 |
| ✓ | ✗ | 28.2 | 50.3 | 51.0 |
| ✓ | ✓ | 27.5 | 49.6 | 50.4 |

Table 3: Performance on *unblurred* test data for SSVP-SLT trained and evaluated on DailyMoth-70h with or without facial blurring during pretraining and SLT.

What’s the effect of blurring? We isolate the impact of facial obfuscation via blurring on SLT performance by training SSVP-SLT models on DailyMoth-70h with and without blurring during pretraining and SLT training. We report the results in Table 3. As expected, performance is best when not blurring (28.8 BLEU) and worst when blurring at finetuning time (28.1 and 27.5 BLEU). Crucially, some performance can be recovered after pretraining on blurred data when performing SLT on unblurred data (28.2 BLEU), validating

our proposed framework (see Table 1).¹¹ This means we can pretrain in a privacy-aware manner without sacrificing too much performance.

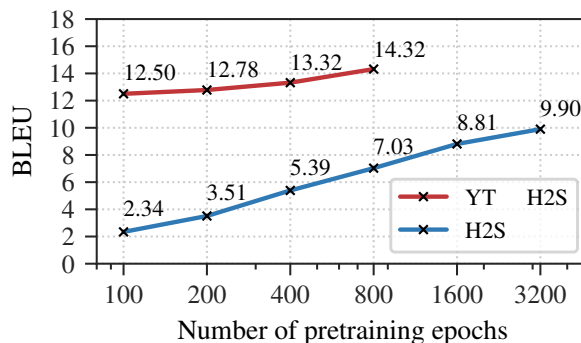


Figure 3: How2Sign test BLEU of SSVP-SLT after pretraining on YouTube-ASL and How2Sign or How2Sign only and finetuning on the same data.

How long should you pretrain? The sign language MAE task is intricate. The model first needs to abstract away surface-level information such as background and signer appearance. It then needs to implicitly acquire an understanding of ASL, capturing long-range dependencies in the video. It is therefore worth investigating basic scaling properties. In Figure 3, we show downstream SLT performance over the course of pretraining. Similar to Ryali et al. (2023) and He et al. (2022), we observe consistent downstream improvements as pretraining progresses, suggesting that the models are not overfitting to the training data even after extensive pretraining. These results underscore the task’s effectiveness and indicate that further scaling would likely yield additional gains.

| CLIP SIZE | PRETRAINING EPOCHS | BLEU | ROUGE-L | BLEURT |
|-----------------|--------------------|-------------|-------------|-------------|
| H2S | | | | |
| 128 | 800 | 7.0 | 25.7 | 39.3 |
| 128 | 100 | 2.3 | 14.3 | 33.6 |
| 16 | 800 | 5.0 | 20.2 | 36.4 |
| YT + H2S | | | | |
| 128 | 800 | 14.3 | 36.9 | 48.9 |
| 128 | 100 | 12.5 | 34.2 | 46.1 |
| 16 | 800 | 10.4 | 31.7 | 44.3 |

Table 4: How2Sign test performance of SSVP-SLT when pretraining on (YouTube-ASL and) How2Sign with a clip size of 16 versus 128 video frames.

¹¹We hypothesize that even more performance could be recovered if the SignHiera video encoder was unfrozen during SLT training, allowing adaptation to the facial information.

Is encoding longer clips necessary? Increasing clip length is costly due to the poor scaling of global attention, raising the question of whether encoding longer clips is needed. Our results indicate that the answer is *yes*. Table 4 compares the performance of the original 16-frame Hieras with our 128-frame SignHiera. While 16-frame Hieras achieve non-trivial performance after 800 epochs, it is substantially outperformed by 800 epoch SignHiera (7.0 vs 5.0 BLEU for H2S and 14.3 vs 10.4 BLEU for YT+H2S). This may be partially explained by the fact that SignHiera sees up to 8× as much data every step. However, we also see that 800-epoch 16-frame Hieras lag far behind even a 100-epoch SignHiera (which has seen roughly the same number of tokens) when training on a large dataset (12.5 vs 10.4 BLEU in the YT+H2S setting). When training on less data (H2S), 16-frame Hieras are still worse than 400-epoch SignHiera (5.39 vs 5.0 BLEU, see Figure 3). Overall, this suggests that certain information cannot easily be acquired from shorter clips.

| MODEL | PARAM | PT | BLEU | ROUGE-L | BLEURT |
|---------|-------|----|-------------|-------------|-------------|
| BART | 140M | ✗ | <u>14.0</u> | <u>36.8</u> | <u>48.3</u> |
| | | ✓ | 13.5 | 36.2 | 48.1 |
| T5-v1.1 | 248M | ✗ | 11.6 | 35.0 | 46.1 |
| | | ✓ | <u>14.3</u> | <u>36.9</u> | <u>48.9</u> |

Table 5: How2Sign test performance of SSV-SLT₈₀₀^{YT+H2S} when finetuning BART and T5, initialized randomly (PT = ✗) or from the pretrained model (✓).

How to choose the text model? We investigate how the architecture and initialization of the text transformer affects performance. Table 5 compares pretrained and randomly initialized BART (Lewis et al., 2020), the English monolingual counterpart to mBART (Liu et al., 2020), which has previously been successfully adapted for German and Chinese sign languages (De Coster and Dambre, 2022; Chen et al., 2022a; Zhou et al., 2023), and T5-v1.1 as used by Uthus et al. (2023).

Overall, T5 outperforms BART, possibly due to larger capacity, but the gap is small. While it is worse to finetune a randomly initialized T5 model compared to the pretrained one (corroborating findings by Uthus et al. (2023)), for BART we find the opposite result. We conclude that whether text pretraining is helpful needs to be evaluated on a case-by-case basis. It may be worth investigating in the future whether an additional pretraining

or finetuning step may lead to better adaptation of the text model to sign language.

| AUG | BLEU | ROUGE-L | BLEURT |
|-----|-------------|-------------|-------------|
| ✗ | 14.3 | 36.9 | 48.9 |
| ✓ | 14.7 | 37.2 | 49.0 |

Table 6: How2Sign test performance of SSV-SLT₈₀₀^{YT+H2S} with and without finetuning augmentation.

Should we augment data at finetuning? Augmentation such as flipping, cropping, and RandAug may improve generalization, but it comes at a high storage cost at finetuning time, as the video features are extracted offline. Is it worth the cost? We compared using 60 epochs of augmented videos (a 60-fold increase in storage) with not using any augmentation. The results in Table 6 show that using augmentation yields a reasonable 0.4 BLEU gain, suggesting that augmentation can be useful when storage is not a major concern.

| INITIALIZATION | MAE | CLIP | BLEU | ROUGE-L | BLEURT |
|-------------------------------------------|-----|------|-------------|-------------|-------------|
| Hiera ₈₀₀ ^{K400} | ✗ | ✓ | 2.1 | 14.9 | 35.0 |
| | ✓ | ✓ | 11.0 | 32.1 | 44.7 |
| SSVP-SLT ₆₀₀ ^{YT+H2S} | ✓ | ✗ | 14.3 | 36.9 | 48.9 |
| | ✓ | ✓ | 15.5 | 38.4 | 49.6 |

Table 7: How2Sign test performance when including (✓) or removing (✗) the MAE and CLIP objectives and pretraining from the original Hieras₈₀₀^{K400} or SSV-SLT₆₀₀^{YT+H2S} checkpoint for 200 epochs on YT+H2S, followed by finetuning on the same data.

Are both pretraining objectives necessary? In §6, we saw that language-supervised video-text pretraining is highly effective when combined with MAE. Are both necessary? We compared pretraining for 200 epochs with either and both objectives, initializing from a 600-epoch SSV-SLT checkpoint, and also performed 200 epochs of CLIP-only pretraining from the original pretrained Hieras. The results in Table 7 show that removing either objective results in a performance drop. The drop is larger when removing MAE, indicating its continued importance after 600 epochs of MAE-only training. Initializing from the original Hieras results in very poor performance (2.1 BLEU), suggesting that language-supervised pretraining alone is not useful in our setting. Considering that language supervision has previously been shown to be effective in isolation (Zhou et al., 2023), this may be due to the FLIP-style masking and the fact

that we do not jointly pretrain T5. We also emphasize that language-supervised pretraining falls under stage II of our framework as it requires annotations; it can, therefore, only serve as an addition to self-supervised pretraining, but not a replacement.

7 Conclusion

Through controlled experiments, we studied the effectiveness of self-supervised pretraining for SLT while considering privacy risks. We introduce SSVP-SLT, a novel, scalable, and privacy-aware SLT method that leverages masked autoencoding on anonymized video. It achieves state-of-the-art ASL-to-English translation performance on the How2Sign benchmark, outperforming the best previous models in the finetuned and zero-shot setting by over 3 BLEU.

Our results demonstrate the promise of self-supervised learning to alleviate the data scarcity issue and further scale up sign language processing in the future. We found that video anonymization, even via simple techniques such as face blurring, has relatively little negative impact on downstream performance, further proving that we can build more proficient systems without neglecting important privacy concerns. We hope that this work, alongside the code and data we release, will spur future developments that benefit the d/Deaf and hard of hearing communities.

Limitations

Compute Requirements Currently, SSVP-SLT requires access to substantial compute to train at the scale of Youtube-ASL (600K videos). This is primarily due to the high dimensionality of video data, exacerbated by the long clip length and information density in sign language content, which creates a data-loading bottleneck and increases the memory footprint, especially in combination with a transformer architecture. Our longest pretraining run in full precision (fp32) took approximately two weeks on 64 A100 GPUs. We believe that it will be important to drive down this cost in the future and make large-scale video pretraining more accessible. While many simple interventions such as mixed precision, gradient accumulation, and gradient checkpointing could drastically reduce the memory footprint, they usually come at the cost of training time or stability. In general, we note that this limitation is not unique to our approach but often not apparent due to training being

conducted on nearly 100× smaller datasets such as RWTH-Phoenix-Weather 2014T (7K videos; Camgöz et al., 2018).

Anonymization We rely on face blurring for video anonymization, which is known to incur a loss of linguistic information (see §2). In the future, it will be worth investigating more sophisticated methods, such as using synthetic appearances. Also, largely due to a lack of linguistic tools for continuous signing, we did not investigate what effects anonymization may have on the translations *qualitatively*. For instance, it would be instructive to know whether the model successfully disambiguates certain phonemes with similar manual features through context in the absence of facial information.

Languages Due to the availability of sufficiently large datasets for our pretraining experiments, we only experiment with American Sign Language and English, the two highest-resource signed and spoken languages. We aim to diversify this language selection in the future.

Ethics Statement

Regarding performance, our models may contain demographic biases and underperform for certain race, gender, and age groups. For instance, even though the YouTube-ASL dataset (a dataset we used for pretraining and supervised finetuning) features over 2500 signers, the authors did not provide demographic details of these signers. Similarly, our DailyMoth-70h dataset includes only one signer (white, male, and early middle-aged). As such, it is unclear how our models perform for underrepresented users, who, aside from having diverse identities, may introduce different accents or regional variations of ASL that our models do not adequately capture. We call for future research on SLT to be more explicit about documenting demographic biases in their datasets and models.

Lastly, we emphasize that anonymization inherently does not offer any formal privacy guarantees in contrast to frameworks such as differential privacy (Dwork, 2006), which fundamentally comes at a (often substantial) cost in utility (Geng et al., 2020). As such, while our work (and the use of facial obfuscation in general) represents an important first step towards comprehensively protecting the privacy of signers, it should not be relied on in sensitive or high-stakes applications.

Acknowledgements

We thank members of the Seamless Communication team at FAIR for helpful feedback and discussions throughout this project. We also thank Florian Metze and Amanda Duarte for their prompt assistance with our questions about the How2Sign dataset.

References

- Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. [Bbc-oxford british sign language dataset](#). *arXiv preprint*.
- Charlotte Baker-Shenk. 1985. [The facial behavior of deaf signers: Evidence of a complex language](#). *American Annals of the Deaf*, 130(4).
- Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. [Neural sign language translation](#). In *CVPR*.
- Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. [Sign language transformers: Joint end-to-end sign language recognition and translation](#). In *CVPR*.
- João Carreira and Andrew Zisserman. 2017. [Quo vadis, action recognition? A new model and the kinetics dataset](#). In *CVPR*.
- Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. [A simple multi-modality transfer learning baseline for sign language translation](#). In *CVPR*.
- Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022b. [Two-stream network for sign language recognition and translation](#). In *NeurIPS*.
- Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. [GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks](#). In *ICML*.
- Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc Le. 2020. [Randaugment: Practical automated data augmentation with a reduced search space](#). In *NeurIPS*.
- Mathieu De Coster and Joni Dambre. 2022. [Leveraging frozen pretrained written language models for neural sign language translation](#). *Information*, 13(5).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *NAACL*.
- Amanda Cardoso Duarte, Samuel Albanie, Xavier Giró-i-Nieto, and Gül Varol. 2022. [Sign language video retrieval with free-form textual queries](#). In *CVPR*.
- Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-i-Nieto. 2021. [How2sign: A large-scale multimodal dataset for continuous american sign language](#). In *CVPR*.
- Cynthia Dwork. 2006. [Differential privacy](#). In *ICALP*.
- Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. 2022. [Masked autoencoders as spatiotemporal learners](#). In *NeurIPS*.
- Andrea Frome, German Cheung, Ahmad Abdulkaader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. 2009. [Large-scale privacy protection in google street view](#). In *ICCV*.
- Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. 2020. [Tight analysis of privacy and utility tradeoff in approximate differential privacy](#). In *AISTATS*.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2022. [Masked autoencoders are scalable vision learners](#). In *CVPR*.
- Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. [Masked autoencoders that listen](#). In *NeurIPS*.
- Amy Isard. 2020. [Approaches to the anonymisation of sign language corpora](#). In *LREC Workshop on the Representation and Processing of Sign Languages*.
- Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. [The kinetics human action video dataset](#). *arXiv preprint*.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *ICLR*.
- Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2021. [American sign language video anonymization to support online participation of deaf and hard of hearing users](#). In *ASSETS*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *ACL*.
- Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023. [Scaling language-image pre-training via masking](#). In *CVPR*.

- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*.
- Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. [Gloss-free end-to-end sign language translation](#). In *ACL*.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *TACL*, 8.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *ICLR*.
- Ross E Mitchell and Travas A Young. 2022. [How Many People Use Sign Language? A National Health Survey-Based Estimate](#). *The Journal of Deaf Studies and Deaf Education*, 28(1).
- Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. 2021. [Evaluating the immediate applicability of pose estimation for sign language recognition](#). In *CVPR Workshops*.
- Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023. [Findings of the second WMT shared task on sign language translation \(WMT-SLT23\)](#). In *WMT*.
- Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. [Considerations for meaningful sign language machine translation based on glosses](#). In *ACL*.
- Carol J. Neidle, Judy Kegl, Benjamin Bahan, Dawn MacLaughlin, and Robert G. Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. MIT Press.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *ACL*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *NeurIPS*.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *WMT*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *ICML*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#). *OpenAI Technical Report*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI Technical Report*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *JMLR (21)*.
- Phillip Rust, Jonas F. Lotz, Emanuele Bugliarelli, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. [Language modelling with pixels](#). In *ICLR*.
- Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. 2023. [Hiera: A hierarchical vision transformer without the bells-and-whistles](#). In *ICML*.
- Marcelo Sandoval-Castaneda, Yanhong Li, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. [Self-supervised video transformers for isolated sign language recognition](#). *arXiv preprint*.
- Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2021. [Anonymsign: Novel human appearance synthesis for sign language video anonymisation](#). In *FG*.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *ACL*.
- Noam Shazeer. 2020. [Glu variants improve transformer](#). *arXiv preprint*.
- Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. [Open-domain sign language translation learned from online video](#). In *EMNLP*.
- William C Stokoe. 1980. Sign language structure. *Annual review of anthropology*, 9(1).
- Laia Tarrés, Gerard I. Gállego, Amanda Cardoso Duarte, Jordi Torres, and Xavier Giró-i-Nieto. 2023. [Sign language translation from instructional videos](#). In *CVPR Workshops*.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. *Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training*. In *NeurIPS*.

Christina O. Tze, Panagiotis P. Filntisis, Anastasios Roussos, and Petros Maragos. 2022. *Cartoonized anonymization of sign language videos*. In *IVSMP*.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. *Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus*. In *NeurIPS*.

Clayton Valli and Ceil Lucas. 2000. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. In *NeurIPS*.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. *Videomae V2: scaling video masked autoencoders with dual masking*. In *CVPR*.

Zhaoyang Xia, Yuxiao Chen, Qilong Zhangli, Matt Huenerfauth, Carol Neidle, and Dimitri Metaxas. 2022. *Sign language video anonymization*. In *LREC Workshop on the Representation and Processing of Sign Languages*.

Zhaoyang Xia, Carol Neidle, and Dimitris N. Metaxas. 2024. *Diffslva: Harnessing diffusion models for sign language video anonymization*. In *LREC Workshop on the Representation and Processing of Sign Languages*.

Kaiyu Yang, Jacqueline H. Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2022. *A study of face obfuscation in imagenet*. In *ICML*.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. *Gloss attention for gloss-free sign language translation*. In *CVPR*.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. *Including signed languages in natural language processing*. In *ACL*.

Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. *SLTUNET: A simple unified model for sign language translation*. In *ICLR*.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. *Gloss-free sign language translation: Improving from visual-language pretraining*. In *ICCV*.

A Pose landmarks vs RGB Video

Pose landmarks (e.g., from MediaPipe Holistic) are cited as a privacy-preserving alternative to RGB video for SLT (Moryossef et al., 2021; Uthus et al., 2023). While they may indeed offer benefits in terms of efficiency and generalization, we argue that they do not offer meaningful privacy protection either. For instance, using a sufficiently large number of facial landmarks that are estimated accurately results in what is essentially a scan of the facial geometry, a biometric identifier according to legislation like the Biometric Information Privacy Act (BIPA).¹² Despite abstracting away shallow information about a person’s appearance, pose information could, therefore, be (mis)used in similar ways as de-anonymized video. Analogous to facial obfuscation in video, one could reduce the number of facial landmarks or add noise to them to hinder re-identification, but doing so also results in (arguably even more) loss of information.

B DailyMoth-70h Dataset

We introduce *DailyMoth-70h*, a dataset containing over 70h of video with aligned English captions of a single native ASL signer (white, male, and early middle-aged) from the ASL news page TheDailyMoth.¹³ We obtained the data via a license agreement with the host of TheDailyMoth.

Download and License The fully self-contained dataset is available under a CC BY-NC license at <https://github.com/facebookresearch/ssvp.slt>.

Statistics We provide detailed dataset statistics in Table 8 and Figure 4.

Purpose The dataset is fully self-contained and can therefore serve as a new benchmark for studying the task of single-signer translation (e.g., for building personalized systems). Furthermore, sign language translation involves overcoming several challenges such as generalizing over signers, their appearances and their signing styles as well as mapping spatio-temporal signed utterances to spoken words. DailyMoth-70h can be used to disentangle some of these challenges by eliminating the signer and style variances and allow researchers to

¹²<https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57>

¹³<https://www.dailymoth.com/>

| | TRAIN | VALIDATION | TEST | FULL |
|-------------------------------------------------------------------------|--------------------|--------------------|--------------------|--------------------|
| Raw data | | | | |
| Number of signers | — | — | — | 1 |
| Number of videos | — | — | — | 496 |
| Video duration (hours) | — | — | — | 76.9 |
| Time frame | 01/21–04/23 | 02/22–04/23 | 02/22–04/23 | 01/21–04/23 |
| Segmented data | | | | |
| Number of clips | 41 412 | 2789 | 4185 | 48 386 |
| Clip duration (hours) | 65.8 | 4.0 | 6.0 | 75.8 |
| Vocabulary (words) | 18 495 | 4803 | 6040 | 19 694 |
| Duration in seconds per clip (mean / std / 90 th percentile) | 5.7 / 2.4 / 8.9 | 5.2 / 2.1 / 7.9 | 5.2 / 2.1 / 7.9 | 5.6 / 2.3 / 8.7 |
| Characters per caption (mean / std / 90 th percentile) | 43.9 / 12.7 / 58.0 | 44.1 / 12.8 / 59.0 | 43.3 / 12.9 / 59.0 | 43.9 / 12.7 / 59.0 |
| Words per caption (mean / std / 90 th percentile) | 8.6 / 2.4 / 12.0 | 8.7 / 2.4 / 12.0 | 8.5 / 2.4 / 12.0 | 8.6 / 2.4 / 12.0 |

Table 8: DailyMoth-70h dataset statistics

ablate their models more focused on the sign-to-spoken language mapping.

Preprocessing We first segmented the raw videos into clips of ~5.6 seconds on average based on their aligned English captions. Each entry in the SubRip subtitle (SRT) file, which maps video timestamps to captions, was chosen to be a distinct datapoint. Accordingly, example clips are often sentence fragments rather than full sentences.

Afterwards, the segmented video clips were automatically cropped such that the signer is approximately in the center of the frame and resized to 224×224 pixels. The preprocessed clips were saved in their native framerates, which are either 24 or 30 fps.

Next, many videos had burned-in captions which, if not removed, would reduce the translation task to a simple OCR task. We, therefore, used an off-the-shelf text detection model to identify burned-in captions in the videos, and blurred the captions conservatively. Although the blurring may be imperfect due to errors made by the text detector, this intervention should nevertheless solve the concern of models shortcutting the SLT task for the most part.

Finally, we split the data into training, validation, and test splits. The proportions (85% / 6% / 9%) were chosen to approximately match How2Sign. The validation and test examples were randomly sampled from the subset of videos posted after January 2022, which avoids data leakage from Youtube-ASL or OpenASL (Shi et al., 2022), both of which have cut-off dates before/during January 2022, into the DailyMoth-70h evaluation splits. The training examples were randomly sampled from the full range of dates (January 2021 to April 2023).

C Reproducibility

C.1 Model and Training Configurations

We report our pretraining configurations for SSVP-SLT in Table 9 and SSVP-SLT-LSP in Table 10. Our finetuning configurations are listed in Table 11 for Youtube-ASL (+ How2Sign) and Table 12 for How2Sign-only and DailyMoth-70h.

C.2 Code

Our implementation uses Python 3.10 and PyTorch 2.0.1 (Paszke et al., 2019) compiled with CUDA 11.7. The code is available under a CC BY-NC license at https://github.com/facebookresearch/ssvp_slt.

C.3 Hardware & Runtime

We ran our experiments on NVIDIA A100 80GB and V100 32GB GPUs. On Youtube-ASL (+ How2Sign), pretraining took ~20 minutes (SSVP-SLT) / 30 minutes (SSVP-SLT-LSP) per effective epoch on 64 A100 or 128 V100 GPUs. On How2sign or DailyMoth-70h, an effective epoch of SSVP-SLT pretraining took ~3 minutes. Finetuning and evaluating T5 and BART on Youtube-ASL (+ How2Sign) took ~8 and 4 minutes, respectively, per epoch on 32 V100 GPUs. Training T5 was slower due to training in full precision as opposed to fp16 and using a smaller batch size with gradient accumulation. Finetuning and evaluating with Tarrés et al. (2023)’s setup on How2Sign or DailyMoth-70h took ~1 minute per epoch on a single V100 GPU.

D Qualitative Examples

In Table 13, we provide qualitative examples of our best-performing model (15.5 BLEU

on How2Sign), compared against the best-performing models from Tarrés et al. (2023), Uthus et al. (2023), as well as the reference translations. The examples were picked from the How2Sign test split by Tarrés et al. (2023). Examples (3)–(5) are, anecdotally, more challenging than the average test example. We find that our model is mostly on-topic and matches the syntactic structure, although it can still struggle with repetitions and the mixing-up of signs. Our model’s failure patterns are more similar to Uthus et al. (2023)’s models—possibly a result of fine-tuning the same base model (T5-v1.1) on the same datasets (Youtube-ASL and How2Sign). For instance, in example (3), both models mistakenly predict the verb “feed” (and mispredict everything that comes after) but flawlessly match the syntactic structure of the reference translation. Overall, both baselines appear to exhibit a higher occurrence of (complete) mis-translation, which aligns with our quantitative results.

| PARAMETER | VALUE |
|-----------------------------|---------------------------------|
| Decoder blocks | 8 |
| Decoder heads | 8 |
| Mask ratio | 0.9 |
| Drop path rate | 0.2 |
| Video size (T, C, H, W) | (128, 3, 224, 224) |
| Sampling Rate | 2 |
| Face Blur | ✓ |
| Random Crop | ✓ |
| Horizontal Flip | ✓ |
| RandAug | ✓(4, 7) |
| Repeated sampling | 2 |
| Optimizer | AdamW |
| Optimizer momentum | $\beta_1 = 0.9, \beta_2 = 0.95$ |
| Weight decay | 0.05 |
| Peak learning rate | $8e-4$ |
| Learning rate schedule | Cosine Decay |
| Minimum learning rate | $1e-5$ |
| Effective warmup epochs | 120 |
| Effective epochs | 800 |
| Effective batch size | 256 |
| Gradient clipping | — |
| Precision | fp32 |

Table 9: SSVP-SLT pretraining settings

| PARAMETER | VALUE |
|-----------------------------|---------------------------------|
| CLIP Text tower | T5-v1.1-base |
| CLIP Feature pooling | mean |
| CLIP Feature projection | 2-layer MLP |
| Decoder blocks | 8 |
| Decoder heads | 8 |
| Mask ratio | 0.9 |
| Drop path rate | 0.2 |
| Video size (T, C, H, W) | (128, 3, 224, 224) |
| Sampling Rate | 2 |
| Face Blur | ✓ |
| Random Crop | ✓ |
| Horizontal Flip | ✓ |
| RandAug | ✓(4, 7) |
| Repeated sampling | 1 |
| Optimizer | AdamW |
| Optimizer momentum | $\beta_1 = 0.9, \beta_2 = 0.95$ |
| Weight decay | 0.05 |
| GradNorm α | 1.0 |
| Peak learning rate | $M = 1e-4, GN = 1e-2$ |
| Learning rate schedule | Cosine Decay |
| Minimum learning rate | $M = 1e-6, GN = 1e-4$ |
| Effective warmup epochs | $M = 10, GN = 1$ |
| Effective epochs | 200 |
| Effective batch size | 512 |
| Gradient clipping | 1.0 |
| Precision | fp32 |

Table 10: SSVP-SLT-LSP pretraining settings. “M” refers to the main optimizer while “GN” refers to the GradNorm optimizer.

| PARAMETER | VALUE |
|-----------------------------|---------------------------------|
| Model & Tokenizer | T5-v1.1 |
| Dropout probability | 0.3 |
| Label smoothing | 0.2 |
| Number of beams | 5 |
| Video size (T, C, H, W) | ($T, 3, 224, 224$) |
| Sampling Rate | 2 |
| Face Blur | ✓ |
| Random Crop | ✗ |
| Horizontal Flip | ✗ |
| RandAug | ✗ |
| Min sequence length | 0 |
| Max sequence length | 1024 |
| Max target length | 128 |
| Optimizer | AdamW |
| Optimizer momentum | $\beta_1 = 0.9, \beta_2 = 0.95$ |
| Weight decay | $1e-1$ |
| Peak learning rate | $5e-4$ |
| Learning rate schedule | Cosine Decay |
| Minimum learning rate | $1e-7$ |
| Warmup epochs | 2 |
| Epochs | 100 |
| Batch size | 256 |
| Early stopping | ✓ |
| Gradient clipping | 1.0 |
| Precision | fp32 |

Table 11: Finetuning settings for Youtube-ASL.

| PARAMETER | VALUE |
|-----------------------------|---------------------------------|
| Encoder layers | 6 |
| Decoder layers | 3 |
| Attention heads | 4 |
| Embedding dim | 256 |
| FFN embedding dim | 1024 |
| Output dim | 256 |
| Layerdrop | 0.0 |
| Activation function | ReLU |
| LayerNorm Before | ✓ |
| LayerNorm Embedding | ✓ |
| Scale embeddings | ✓ |
| Decoder share embeddings | ✓ |
| Vocab size | 7000 |
| Lowercase tokenizer | H2S = ✓, DM = ✗ |
| Truecase outputs | H2S = ✓, DM = ✗ |
| Dropout probability | 0.3 |
| Label smoothing | 0.2 |
| Number of beams | 5 |
| Video size (T, C, H, W) | ($T, 3, 224, 224$) |
| Sampling Rate | 2 |
| Face Blur | H2S = ✓, DM = ✓/✗ |
| Random Crop | ✗ |
| Horizontal Flip | ✗ |
| RandAug | ✗ |
| Min sequence length | 0 |
| Max sequence length | 1024 |
| Max target length | 128 |
| Optimizer | AdamW |
| Optimizer momentum | $\beta_1 = 0.9, \beta_2 = 0.95$ |
| Weight decay | $1e-1$ |
| Peak learning rate | $1e-2$ |
| Learning rate schedule | Cosine Decay |
| Minimum learning rate | $1e-4$ |
| Warmup epochs | 10 |
| Epochs | 200 |
| Batch size | 32 |
| Early stopping | ✓ |
| Gradient clipping | 1.0 |
| Precision | fp16 |

Table 12: Finetuning settings for How2Sign (H2S) & DailyMoth-70h (DM).

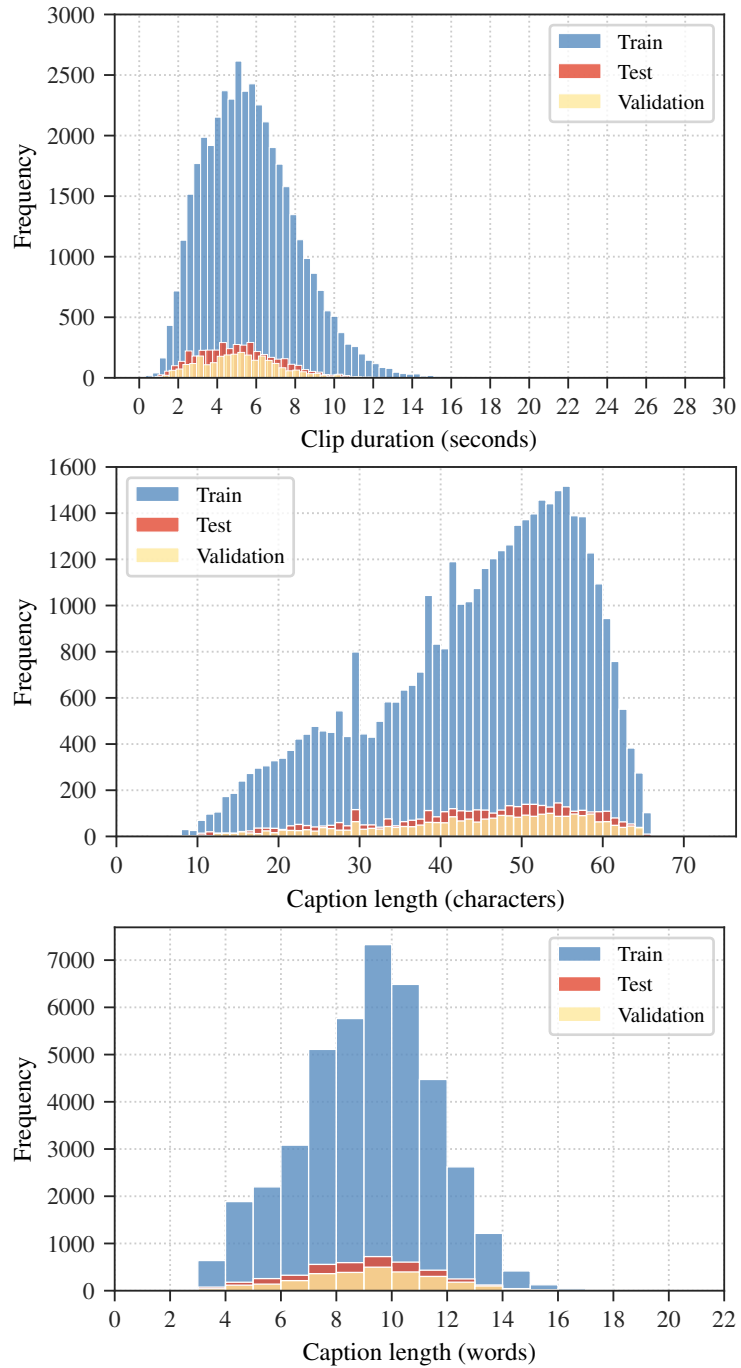
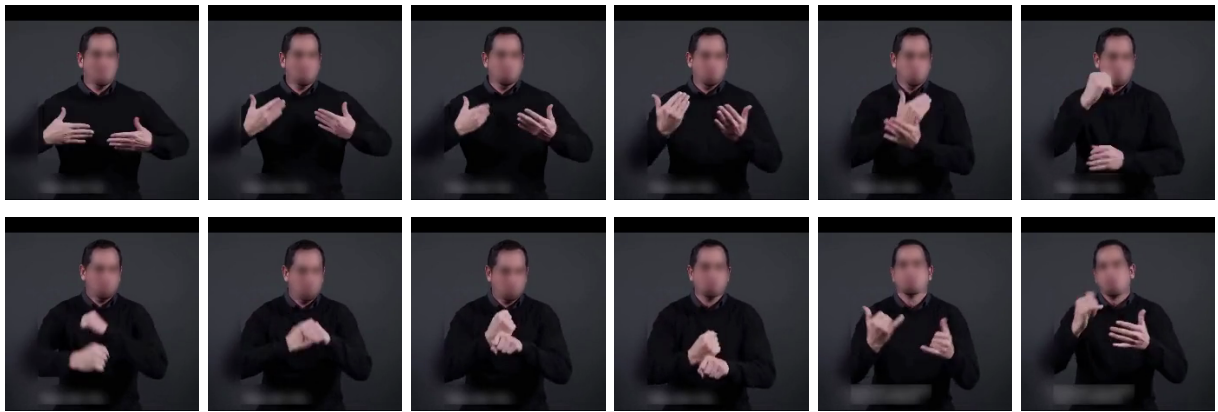


Figure 4: DailyMoth-70h dataset split statistics



(a) "Hello, welcome to *The Daily Moth*."



(b) "Happy New Year"

Figure 5: Example frames sampled from two videos in the blurred version of the DailyMoth-70h training split.

| | | |
|-----|------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (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. |
| | Uthus et al. | It's really great for women's self defense. |
| | Ours | This is a really great point 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. |
| | Uthus et al. | In this clip we're going to show you how to cut a piece of clay. |
| | Ours | In this clip I'm going to show you how to clip the cable, the cable. |
| (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. |
| | Uthus et al. | In this clip we're going to talk about how to feed a set of baiting lizards for a lava field oil. |
| | Ours | In this clip we're going to talk about how to feed the trail for draining clean for laborer 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. |
| | Uthus et al. | 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. |
| | Ours | So that she's going to get her hips up as far as she can, and now she's going to lift her head up as far as possible. |
| (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. |
| | Uthus et al. | 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. |
| | Ours | But if you have to set up a new campfire, this is one way to do it in a low impact. One is a monk fire. One is a campfire. The other one is to set a campfire in a campfire. That's just a post like the top of the post. |
| (6) | Reference | So, this is a very important part of the process. |
| | Tarrés et al. | It's a very important part of the process. |
| | Uthus et al. | Alright, let's get started. |
| | Ours | It's an important part of the process. |

Table 13: Qualitative translation examples from our best-performing model compared to Tarrés et al. (2023), Uthus et al. (2023), and the reference translations. The examples were picked from the How2Sign test set by Tarrés et al. (2023) and do not necessarily accurately reflect progress on the task. We see that our model is mostly on-topic, but can still struggle with repetitions and the mixing-up of signs.