Segment, Embed, and Align: A Universal Recipe for Aligning Subtitles to Signing

Anonymous ACL submission

Abstract

The goal of this work is to develop a universal approach for aligning subtitles (i.e., spoken language text with corresponding timestamps) to continuous sign language videos. Prior approaches typically rely on end-to-end training tied to a specific language or dataset, which limits their generality. In contrast, our method Segment, Embed, and Align (SEA) provides a single framework that works across multiple languages and domains. SEA leverages two pretrained models: the first to segment a video frame sequence into individual signs and the second to embed the video clip of each sign into a shared latent space with text. Alignment is subsequently performed with a lightweight dynamic programming procedure that runs efficiently on CPUs within a minute, even for hour-long episodes. SEA is flexible and can adapt to a wide range of scenarios, utilizing resources from small lexicons to large continuous corpora. Experiments on four sign language datasets demonstrate state-of-the-art alignment performance, highlighting the potential of SEA to generate high-quality parallel data for advancing sign language processing.

1 Introduction

Sign language alignment is the task of aligning spoken language text to continuous sign language videos. Formally, given a video and a sequence of weakly temporally aligned subtitle units, the alignment task adjusts the start/end timestamps of the subtitles to better match the signing (Figure 1). This matters in practice for at least two reasons.

First, the development of sign language translation systems (Müller et al., 2022, 2023; De Coster et al., 2023) relies on *parallel data* between sign language videos and spoken language text. While small-scale studio or lab datasets, such as How2Sign (Duarte et al., 2021) and CSL-Daily (Zhou et al., 2021), offer high-quality manual alignment, large-scale broadcast datasets, like BOBSL (Albanie et al., 2021), often exhibit noisy alignment. Subtitles for broadcast content are typically generated based on the original audio track and therefore show a non-deterministic temporal offset relative to the signed interpretation (Müller et al., 2022). Improving alignment quality benefits the training and evaluation of translation models.

Second, user-generated sign language content on online platforms like YouTube (Uthus et al., 2023; Tanzer and Zhang, 2025) gains from *automatic subtitling*. Good alignment cuts the manual effort of post-hoc captioning and improves accessibility at scale (Gernsbacher, 2015). Accurate alignment also helps sign language linguists when creating and annotating corpora—e.g., the DGS Corpus (Hanke et al., 2020)—by pre-segmenting material and synchronizing text and signing tiers.

The most reliable way to align is through human annotators (ideally, native signers) with tools such as ELAN (Wittenburg et al., 2006), VIA (Dutta and Zisserman, 2019), or iLex (Hanke, 2002). However, manual alignment is prohibitively time-consuming and costly. As reported by Bull et al. (2021) in the context of BOBSL, an expert fluent in sign language takes approximately 10–15 hours to align subtitles to one hour of continuous sign language video. On the other hand, the per-hour rate is ∼40 USD according to the WMT-SLT 22 campaign (Müller et al., 2022). Previous methods for sign language alignment have either relied directly on manually aligned data or indirectly on an end-to-end network trained on the former. For example, Bull et al. (2021) finetunes a subtitle aligner network on 17.7 hours of manual annotation (§2).

In this work, we propose *Segment, Embed, and Align* (SEA) as a general solution to the text-signing alignment problem, which is effective across various languages and data sources and requires almost zero direct in-domain supervision for the alignment task. SEA consists of three modular steps: (1) *Segment* continuous sign language

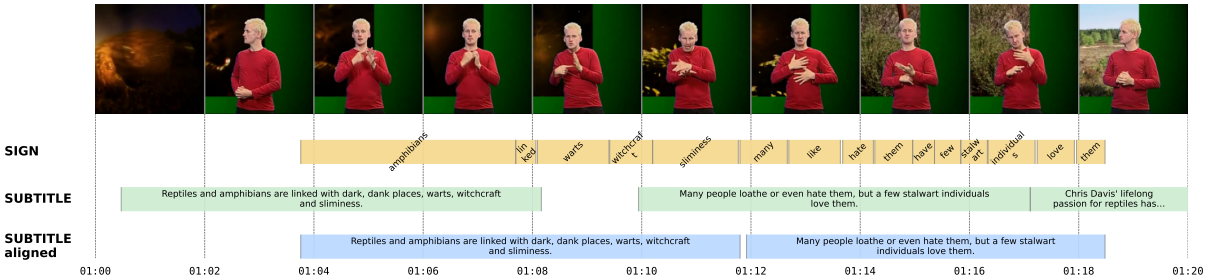


Figure 1: **Aligning subtitles to continuous signing:** In broadcast interpreting and other sign language corpora, original subtitles (green) frequently lag or lead the actual signing (yellow) by non-deterministic amounts. Our alignment method, SEA, produces time-corrected subtitles (blue) that better correspond to the signed content. Keyframes are sampled at the midpoint of each span and may not include all annotated signs present in that interval.

video frames into individual meaningful units (e.g., signs); (2) *Embed* the sequences of signing and subtitle units of a video into a shared latent space; (3) *Align* the subtitle units to the signs based on a similarity matrix formed by text and signing embeddings and the prior subtitle locations. We highlight that the first two modules are facilitated by pretrained sign language *segmentation* (Moryossef et al., 2023) and *embedding* (Jiang et al., 2024b) models that transfer to alignment as a downstream task. The alignment module is a lightweight *global* optimization based on dynamic programming (DP). It runs swiftly on CPUs at the episode level without requiring gradient-based operations. In contrast, prior work performs training/inference in *local* windows of about 20 seconds, followed by a slow dynamic time warping (DTW) process to resolve local conflicts (Bull et al., 2021).

SEA's modular design offers greater generalisability than end-to-end, in-domain supervised methods: individual components can be seamlessly swapped in a plug-and-play manner for different languages while keeping the framework unchanged. Our experiments (§4) reveal that the segmentation component, in particular, exhibits relatively strong cross-linguistic universality, while the embedding component benefits from language-specific expertise. SEA achieves state-of-the-art (SOTA) performance across four different benchmarks (BOBSL, How2Sign, WMT-SLT, and SwissSLi (Jiang et al., 2024a)) and three sign languages: British Sign Language (BSL), American Sign Language (ASL), and Swiss German Sign Language (DSGS). All SEA components are open source to enable the curation of better-aligned sign language datasets and future improvements[1].

[1]Repository links will be released after anonymous review.

2 Related Work

Video-text alignment. General video–text alignment has been widely studied on HowTo100M (Miech et al., 2019), a large corpus of instructional videos paired with narrations derived from automatic speech recognition, which provides weak supervision. Building on this, Miech et al. (2020) proposes MIL-NCE to learn joint video–text representations through multi-instance contrastive learning, addressing narration–visual misalignment. Han et al. (2022) further uses weak labels to estimate sentence alignability and predict temporal spans, enabling sentence-level localization in long videos.

Signing-subtitle alignment. Specifically developed for sign language videos, Bull et al. (2021) introduces a model named Subtitle Aligner Transformer (SAT) that is first pretrained on large-scale data for the task of sign localisation with weak supervision and then finetuned on 17.7 hours of manually aligned subtitle-signing data (Albanie et al., 2020). SAT utilizes BERT (Devlin et al., 2019) to embed the subtitles and I3D (Carreira and Zisserman, 2017) to embed the visual signal, then cross-attends them in an encoder-decoder architecture. The model outputs frame-level predictions on whether each frame belongs to the queried subtitle or not within a 20-second search window.

A contemporary study, Jang et al. (2025a), extends SAT to a new SOTA on BOBSL by three additional contributions: (1) Preprocessing the subtitles with specific grammatical rules; (2) A selective alignment loss with negative sampling of irrelevant video clips; (3) Iterative self-training with refined pseudo-labels. These modifications optimize SAT, but in contrast to SEA, a fair amount of manually aligned data remains key to achieving strong performance. Furthermore, although SAT is end-to-end

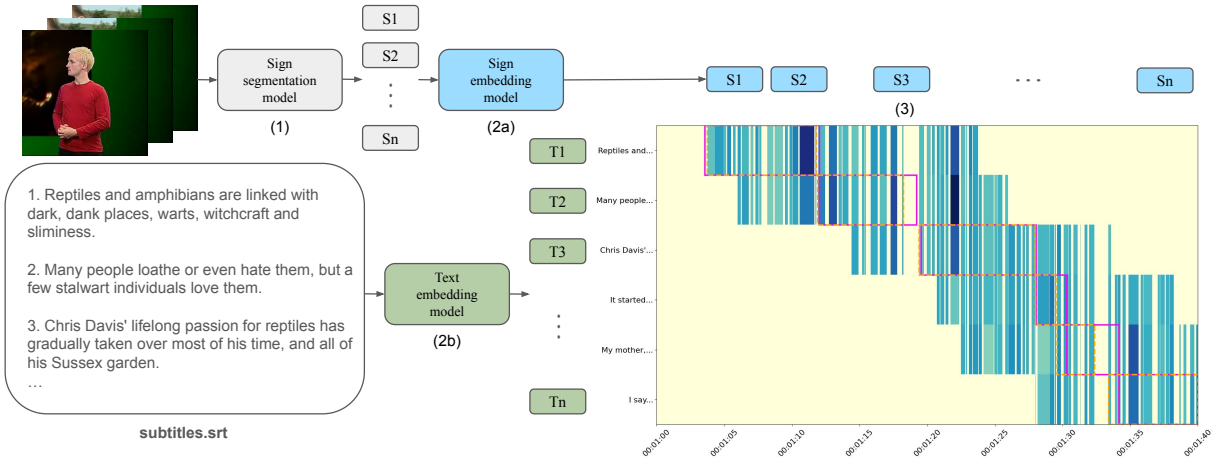


Figure 2: **Our method SEA consists of three modular steps:** (1) *Segment* video frames of continuous signing into individual signs; (2) *Embedd* each sign (s_1 to s_n; 2a) and subtitle unit (t_1 to t_n; 2b) into a shared latent space, with their dot product similarities encoded by a similarity matrix; (3) *Align* subtitles to signing based on the text-sign similarities and the original temporal location of the subtitle units. The similarity matrix is illustrated as a heatmap over time, with darker bars indicating a higher similarity between a sign and a subtitle; similarities of the originally remote signs are zeroed out to ensure locality. The dashed/solid boxes in the heatmap indicate the predicted/manually aligned subtitle locations, respectively.

at the sentence level, Bull et al. (2021); Jang et al. (2025a) must follow local model predictions using a DTW process to resolve the temporal conflicts of individual predictions. On the other hand, the alignment in the SEA algorithm runs globally at the episode level, so no local window is needed.

Tanzer (2025) proposes a multitasking network with two branches, caption alignment and translation, inspired by Whisper (Radford et al., 2023). However, the reported alignment performance, measured by frame accuracy, lags far behind a model-free baseline where the caption duration is scaled in proportion to the caption length. We present SEA as a training-free alternative for the alignment task, as opposed to prior work.

3 Method

In this section, we describe the SEA method (Figure 2) step by step. We first introduce how we leverage pretrained sign language *segmentation* (§3.1) and *embedding* (§3.2) models, and then present the key *alignment* step (§3.3).

3.1 Segmentation

Given a continuous stream of video frames, also known as an episode, we first segment it into *meaningful units*, such as signs[2]. This process is analogous to text tokenization. While the idea sounds natural, in practice, there is no standardized tool like those in NLP (Sennrich et al., 2016; Kudo and Richardson, 2018), and the sign boundaries are not clearly defined linguistically (Hanke et al., 2012).

In this work, we adopt the automatic sign segmentation model of Moryossef et al. (2023), trained on ∼73 hours of annotated data from the Public DGS Corpus. It passes MediaPipe Holistic (Grishchenko and Bazarevsky, 2020) poses into a lightweight LSTM (Hochreiter and Schmidhuber, 1997) model and then produces frame-level predictions. Despite being trained only on German Sign Language (DGS), the model shows successful zero-shot transfer to other sign languages[3], which is attributed to (1) the iconicity and finite set of hand/body configurations in sign languages, and (2) the use of MediaPipe pose representations that capture and focus on these motion primitives.

3.2 Embedding

After sign segmentation, for each episode, we now have a sequence of *signs* and a sequence of *subtitle units*. Each *sign* is a sequence of continuous video frames of signing, and each *subtitle* is a sequence of text tokens[4]. The next goal is to convert them into sequences of *alignable vectors*.

We employ the SignCLIP (Jiang et al., 2024b)

[2]Practically, the signing units can be on a sub-sign level due to an imperfect segmentation model.

[3]Zero-shot transfer (0.76 ROC-AUC on *O*-tag) was reported for French Sign Language in Moryossef et al. (2023), and we find the segmentation also useful for ASL/BSL/DSGS.

[4]Subtitles may or may not form well-formed sentences.

model, which re-purposes CLIP (Radford et al.) to project text and signing into the same latent space. Subtitle units are embedded with a BERT-like text encoder, while signing video clips are converted to a sequence of MediaPipe poses and then encoded by a pose encoder initialized from the same BERT weights and finetuned for signing. We use a multilingual checkpoint pretrained on *SpreadtheSign* by Jiang et al. (2024b), which covers 41 sign languages and English on the text side only. We also show the benefits of language-specific finetuning for SignCLIP on BSL/English, ASL/English, and DSGS/German data in §4.

3.3 Alignment

The next goal is to assign each subtitle to a contiguous span of predicted sign segments, and then rewrite the subtitle's timestamps to reflect that span. We optimize a single objective that balances two kinds of evidence. (1) Temporal/prosodic costs: how far the subtitle must move, how much the duration changes, and how big the inter-sign gaps are within a subtitle—signing tends to be continuous, and big gaps indicate prosodic pauses that lead to the next phrase. (2) Semantic similarity costs: a reward for assigning the subtitle to signs whose embeddings are similar to the text. We consider two variants within the same framework: *Segment and Align* uses only the timing/prosodic costs (no semantics), while *Segment, Embed, and Align* augments the same objective with the semantic term. Both variants are solved by a single global DP procedure over the full episode. We mathematically define the alignment algorithm next, present implementation details in §3.4, and provide a pseudocode description in Appendix §A for additional clarity.

Task. Given subtitle units $t_1, \ldots, t_n$ with times $(\mathrm{start}(t_i), \mathrm{end}(t_i))$ and signs $s_1, \ldots, s_m$ with $(\mathrm{start}(s_j), \mathrm{end}(s_j))$, we assign to each t_i a contiguous span $s_l, \ldots, s_r$ and shift t_i to that span's boundaries $(\mathrm{start}(s_l), \mathrm{end}(s_r))$.

Cost function. For a pair $(t_i, s_l, \ldots, s_r)$, a *cost function* $\phi(t_i, s_{l:r})$ is defined by a weighted sum of: onset distance $|\mathrm{start}(t_i) - \mathrm{start}(s_l)|$, offset distance $|\mathrm{end}(t_i) - \mathrm{end}(s_r)|$, duration difference $|(\mathrm{end}(t_i) - \mathrm{start}(t_i)) - (\mathrm{end}(s_r) - \mathrm{start}(s_l))|$ with weight w_{dur}, and internal inter-sign gaps $G(l, r) = \sum_{j=l}^{r-1} \max(0, \mathrm{start}(s_{j+1}) - \mathrm{end}(s_j))$ with weight w_{gap}.

Segment and Align. We select non-overlapping spans (l_i, r_i) for t_i that minimize the total cost $\sum_i \phi(t_i, s_{l_i:r_i})$ using the individual costs above. A standard episode-level DP algorithm is employed to determine the global optimum, which iterates over the subtitles and signs once their matching costs are set up. The DP is similar to, but distinct from, DTW: DTW treats each unit independently, except for path continuity; whereas SEA considers inter-sign gaps as part of the cost. We append the formal definitions of the DP in Appendix §A.

Segment, Embed, and Align. We enrich alignment with a text–sign similarity matrix. Let u_i and v_j be embeddings for t_i and s_j; define raw similarities by dot product $M^0_{ij} = \langle u_i, v_j \rangle$. For each row i, we keep only $window_size = 50$ temporally nearest signs to a subtitle to ensure locality (others masked to 0) and apply a row-wise *softmax* normalization to obtain the final M. The total span similarity is $\Sigma(i; l, r) = \sum_{j=l}^{r} M_{ij}$, and we add a negative term $-w_{\mathrm{sim}} \Sigma(i; l, r)$ to the cost function. It pushes the algorithm to align subtitles to a span of similar signs based on the weight w_{sim}. The DP optimization process is identical to the above.

3.4 Implementation Details

Segmentation. We use the `E4s-1` checkpoint from Moryossef et al. (2023) as the sign segmentation model, with two controllable decoding parameters `b-threshold` and `o-threshold` ranging from 0 to 100 to decide when to start/end a sign.

For BSL, we explore finetuning the segmentation model on 3.3 hours (149 videos) of the BSL Corpus (Schembri et al., 2013), a video corpus of conversational BSL with gloss-level annotations. Given the small size, we use *negative sampling*: for each annotated sentence, a non-signing clip from BOBSL (10+ seconds away from any detected sign, lasting 1–30s) is randomly prepended, appended, or substituted. It helps the model to recognize negatives, as the data is densely packed with signing; the best model is named *Segmentation-BSLCP*.

We also evaluate the BSL Corpus–trained models of Renz et al. (2021a,b) in §4.5. With less supervision and a focus on boundary detection rather than individual signs, they often produce false positives on non-signing segments. In contrast, Moryossef et al. (2023) addresses it through a BIO tagging scheme (Ramshaw and Marcus, 1999), which detects active signing. We omit experimenting with the recent follow-up of He et al. (2025) as

Dataset	Languages	Data source	#episodes (train/dev/test)	#subtitles	#hours	#signers	Offsets	FPS
BOBSL	BSL/English	Interpretation of mixed programs	55 (16/4/35)	31,479	∼46	22	2.70/2.70	25
How2Sign	ASL/English	Studio-recorded re-signing	2,528 (2,212/132/184)	35,191	∼80	9	1.57/2.13	24/30/50/60
WMT-SLT SRF	DSGS/German	Live interpretation of daily news	29 (21/4/4)	7,071	∼16	3	1.06/1.75	25
SwissSLi mitenand	DSGS/German	Offline, deaf translation of one program	14 (0/7/7)	554	∼1	2	0.49/0.56	25

Table 1: Summary of datasets and their statistics. *Offsets* corresponds to the temporal shifts added to the start/end of the subtitles in the *Original*$^+$ baseline, where the offsets correspond to the median (or mean) differences between the original and manually aligned subtitle timings, calculated from the training (or validation for SwissSLi) set.

neither the model nor the code has been available.

Embedding. We use the *SignCLIP-multilingual* (`E7.2`; `baseline_temporal`) checkpoint for sign and subtitle embeddings by default. ISO 639-3 language codes are prepended in the text prompts for aligning different language pairs: `<en> <bfi>` for BSL/English; `<en> <ase>` for ASL/English; and `<de> <sgg>` for DSGS/German.

We further finetune *SignCLIP-multilingual* with language-specific data. For BSL/English, we use sign spottings from BOBSL (Momeni et al., 2022), following Raude et al. (2024); Jang et al. (2025b) but with a contrastive text objective instead of classification over 8,697 signs, totaling ∼3.5M examples. For ASL/English, we aggregate ∼200K examples from PopSign (Starner et al., 2023), ASL Citizen (Desai et al., 2023), and Sem-Lex (Kezar et al., 2023), three isolated sign language recognition (ISLR) datasets. We normalize text by using lowercase glosses without variations. For DSGS/German, we finetune on Signsuisse (Müller et al., 2023), which provides 16,213 lexical items each paired with an example sentence. The resulting models are named *SignCLIP-BSL*, *SignCLIP-ASL*, and *SignCLIP-Suisse*, respectively. Finetuning takes less than 3 days on an NVIDIA RTX A6000 GPU with 48 GB of GPU memory.

Alignment. We start by experimenting with *Segment and Align* without *Embed*. For datasets where a training set is available, we conduct a random search to identify the optimal parameter combination, including the DP weights and parameters, the segmentation model decoding parameters, and fixed offset values applied before and after the DP process. This search process learns dataset-specific bias, such as the common temporal delay in interpretation/translation, as summarized in Table 1. We then incorporate the similarity matrix provided by the embeddings with the weight w_{sim}, which reflects the strength of the embeddings. Additional details and the best parameter combinations are listed in Appendix A.

4 Experiments

This section introduces the datasets (§4.1), our evaluation protocol, and the baselines (§4.2). We then present how our SEA method performs both quantitatively (§4.3), compared to previous SOTA approaches, and qualitatively (§4.4). Finally, we ablate different components of our approach (§4.5).

4.1 Dataset Overview

We select continuous sign language datasets that include *parallel* subtitles before and after manual alignment, such as BOBSL, How2Sign, and WMT-SLT SRF. We present an overview of the statistics of the datasets in Table 1. For BOBSL and How2Sign, we use the default split. For WMT-SLT SRF, we re-split the original training set of the WMT-SLT 22 campaign to obtain 4 episodes for dev and test, respectively, as the subtitles before manual alignment are not available for these sets.

We additionally hire DSGS experts to annotate (annotation details in Appendix C) 14 episodes of the *mitenand* program from the SwissSLi dataset, which acts as a test-only set in a highly low-resource setting. This dataset is intended to be a DSGS/German dataset of superior quality, as it was translated by deaf translators offline rather than live interpreted, unlike WMT-SLT SRF, where the hearing interpreters tend to revert to spoken-language grammar and drop words under time pressure.

4.2 Evaluation and Baselines

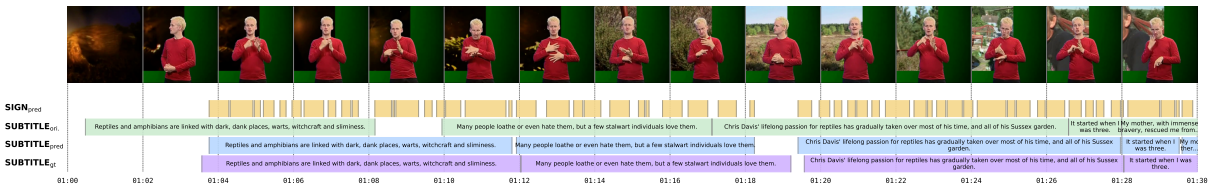
We follow the evaluation protocol proposed by Bull et al. (2021), but prioritize the *F1-score* over frame accuracy, since SEA does not predict at frame level. Specifically, *F1@0.50* counts a subtitle as correct when the predicted segment overlaps the ground truth with intersection-over-union (IoU) ≥ 0.50.

We construct two baselines per dataset: *Original* refers to the original subtitles from the datasets, typically intended to align with the speech signal; *Original*$^+$ refers to adding a fixed offset to the start and end position of each subtitle[5]. The fixed offset

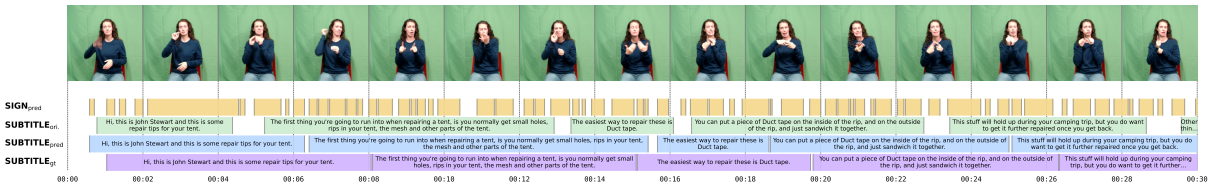
[5]Equivalent to S_{audio} and S^+_{audio} in Bull et al. (2021).

	BOBSL		How2Sign		WMT-SLT		SwissSLi	
	Val	**Test**	**Val**	**Test**	**Val**	**Test**	**Val**	**Test**
STATIC BASELINES								
Original alignment (usually audio-based)	29.09	14.11	30.63	33.06	47.83	46.85	69.58	60.48
Original$^+$ (fixed offsets from training data)	49.63	44.61	31.91	36.21	74.17	74.83	–	–
END-TO-END TRAINED BASELINES								
SAT (Bull et al., 2021)	60.98	54.57	–	–	75.32	75.32	–	–
SAT$^+$ (Jang et al., 2025a)	74.30	63.81	–	–	–	–	–	–
SEGMENT AND ALIGN								
Segmentation (Moryossef et al., 2023)	66.24	49.58	33.38	36.17	**75.66**	76.83	71.48	84.19
Segmentation-BSLCP	63.61	47.75	–	–	–	–	–	–
SEGMENT, EMBED, AND ALIGN								
SignCLIP-multilingual	66.70	50.68	35.51	37.51	75.16	76.43	**71.86**	**85.57**
SignCLIP-BSL/ASL/DSGS-finetuned	72.78	54.50	**38.32**	**39.57**	75.34	**77.69**	**71.86**	85.22
SignCLIP-BSL + SAT$^+$ subtitles	**74.81**	**65.68**	–	–	–	–	–	–

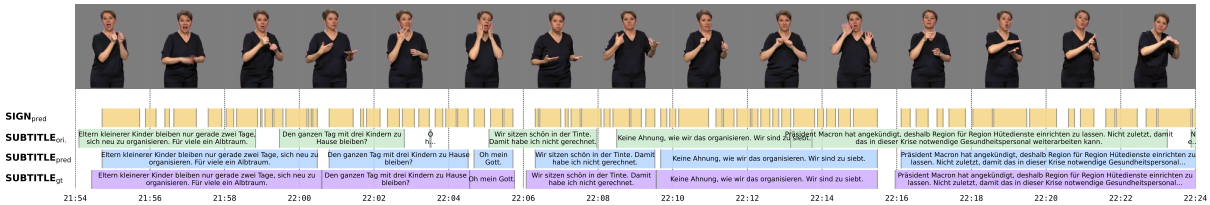
Table 2: SEA alignment performance (F1@0.50) on the validation and test sets across four sign language datasets. SEA is SOTA on all test sets and we report two variants: (1) *Segment and Align* (DP with timing/prosodic costs only); (2) *Segment, Embed, and Align* (same DP + SignCLIP text–sign similarity; multilingual or finetuned). Baselines include static offsets and SAT/SAT$^+$. For BOBSL, we also test SEA initialized from SAT$^+$ predicted subtitles.



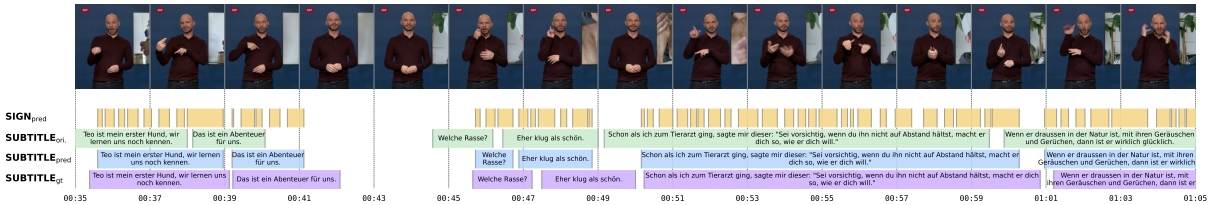
(a) BOBSL (English→BSL; video id: `5224144816887051284`)



(b) How2Sign (English→ASL; video id: `ETOZLBScxWY-3-rgb_front`)



(c) WMT-SLT SRF (German→DSGS; video id: `srf.2020-03-13`)



(d) SwissSLi mitenand (German→DSGS; video id: `2023-02-16_ein_hund_fürs`)

Figure 3: Qualitative results: For each dataset, we sample a 30-second validation clip and show keyframes every 2 seconds. Rows: predicted signs from the segmentation model (yellow), original subtitles (green), SEA-aligned subtitles (blue), and expert-aligned ground truth (purple). In general, segmentation identifies the signing frames of interest; SEA then shifts subtitles toward higher text–sign similarity—for example, in BSL the first subtitle begins at the fingerspelling AMPNIBIS ("amphibians") and ends at the sign for "sliminess".

numbers are set to the median differences between the default and the manually aligned subtitles in the training set[6], as collected in Table 1. For BOBSL, we keep +2.7 seconds — the *average* shift between audio and manually-aligned sentences in the manual training set, as specified by prior work (Albanie et al., 2021). However, we find that using median values is more robust to outliers for other datasets.

For BOBSL, we have two additional baselines: *SAT* (Bull et al., 2021) and *SAT*$^+$ (Jang et al., 2025a). We also train the SAT model on the WMT-SLT data to test the applicability of SAT on a different dataset, using the same I3D visual backbone but domain-specific alignment supervision.

4.3 Experimental Results and Discussion

Table 2 reports SEA's performance: it attains SOTA on all four test sets—surpassing SAT/SAT$^+$ on BOBSL and establishing new SOTA on three additional, cross-language/domain datasets not previously evaluated. Key points follow.

The original alignment quality varies across datasets. Fixed offsets significantly improve interpreted datasets (e.g., BOBSL, WMT-SLT SRF), where a typical interpretation delay is expected; however, the effect is weaker for datasets with different production workflows. Note that alignment on How2Sign is inherently hard because some random appearances of signing do not correspond to any subtitle, especially near episode starts and ends.

End-to-end training by SAT/SAT$^+$ remains a good choice when supervision is available. The effectiveness also relies on a strong in-domain visual backbone network (I3D for BOBSL), which does not necessarily excel on another dataset or sign language, as seen in the case of WMT-SLT. We do not further explore SAT$^+$ in this work and refer to additional optimizations by Jang et al. (2025a).

Adjusting original alignment based on the prosodic cues from segmentation already helps. In *Segment and Align*, the segmentation plays two main roles: (1) detecting in which temporal spans signing happens (against the signer making a pause/rest or being absent in the picture); (2) segmenting meaningful units to draw decision boundaries on, such as a prosodic pause. However, the lack of semantic understanding limits performance, especially when such cues are rare, e.g., How2Sign, where signers tend to sign continuously without many breaks in a studio recording setting.

The embedding module further enhances the ability to discern boundaries and achieves SOTA. Consistent improvements across datasets and languages are observed after the *embed* step is added to the pipeline, even with the *SignCLIP-multilingual* model, which did not see any DSGS/German data during the pretraining phase. Language-specific finetuning further enhances the alignment quality: For BOBSL, finetuning on in-domain sign spottings, although noisy, brings a boost of ∼4 to 6 *F1@0.50* on the validation and test sets. For How2Sign, finetuning on three out-of-domain but large-scale ISLR datasets also clearly helps. The effect of finetuning on Signsuisse, a small lexicon resource of a different domain, is relatively weak as seen for WMT-SLT and SwissSLi.

Prior alignment and dataset biases matter. For BOBSL, we additionally initialize the prior location of the subtitles by the model output from Jang et al. (2025a), as a controlled experiment to explore SEA's sensitivity and robustness to prior locations of different quality. When initialized with the subtitles automatically aligned by *SAT*$^+$, SEA inherits the gains and further refines them. When initialized with subtitles in a challenging original position, such as in How2Sign, SEA still improves alignment but less markedly. It is also worth mentioning the importance of dataset-specific biases. In the random search process, we have learned similar offset values as in the *Original*$^+$ baseline and found that a fixed 1-second offset applied after the DP process is always useful in the datasets evaluated[7].

4.4 Qualitative Results

We present qualitative results in Figure 3, which include a 30-second sample for each of the evaluated datasets, taken from SEA's global alignment. In general, the sign segmentation accurately identifies temporal regions of interest and excludes non-signing parts, such as pauses during signing. The segmentation and embedding modules, together, help the final alignment step move the original subtitles to locations closer to the human-aligned ground truth. However, because SEA still tends to stay tied to the original subtitle timings, when the latter are poor, it affects the final alignment quality, which can sometimes cause errors to propagate. In

[6] Unavailable for SwissSLi, validation set is used instead.

[7] "Useful" in the sense of better evaluation results since annotators tend to give extra duration after the signing stops.

our ASL example, the first subtitle misses several signs—including TENT—which is tricky since the word "tent" also appears in the next sentence.

4.5 Ablation Study

We ablate the effects of different segmentation and embedding methods in Table 3 and Table 4.

Segmentation. We explore different segmentation models introduced in §3.4 and measure their impact on the *Segment and Align* approach. We evaluate segmentation performance on the BSL Corpus (BSLCP), equipped with human-timed gloss annotations, and alignment on the BOBSL validation set using the same segmentation model. We cannot evaluate segmentation on BOBSL since it lacks equivalent gloss-level human annotation with accurate timings. Renz et al. (2021a,b) performs better than Moryossef et al. (2023) on segmentation evaluation, but the final alignment is poorer. We speculate three reasons: (1) BSL Corpus might not be representative of BOBSL due to the domain gap. (2) Renz et al. (2021a,b), trained to excel at discerning signing boundaries, produce many false positives of signing when used for alignment. (3) Segmentation of linguistically well-defined signs might not be a hard requirement for downstream tasks such as alignment. Upon inspection, Moryossef et al. (2023) produces slightly over-segmented output, but SEA seems to operate well on the *sub-word* (Sennrich et al., 2016) level. Finetuning Moryossef et al. (2023) on the BSLCP training data achieves SOTA segmentation performance (even outperforming follow-up work; He et al. (2025)), but falls short of the original model's alignment capabilities (*Segmentation-BSLCP* in Table 2). We thus adhere to the original segmentation model from Moryossef et al. (2023) for SEA.

Embedding. We also examine the dynamics between the strength of the embedding models, measured by retrieval-based ISLR, and the final alignment results. For BSL, we test ISLR on the BOBSL sign spottings, and for ASL, we test ISLR on the PopSign ASL dataset. A consistent positive correlation is observed between ISLR and alignment across languages. For BOBSL, we further test a specific gloss-based alternative of the embedding module, i.e., using the pseudo/human glosses produced in the previous CSLR2 (Raude et al., 2024) work. We fill in the similarity matrix with a value of 1 if a sign segment's gloss is in a subtitle, and 0 otherwise. This alternative can be viewed as a rigid assignment of signs (analogous to how a human expert would approach it, as shown in Figure 1) rather than a soft similarity; we find that the latter approach works better.

Model	BSLCP Seg.		BOBSL Align.
	F1	mF1S	F1@50
Renz et al. (2021a)	47.71	–	57.98
Renz et al. (2021b)	–	–	57.53
He et al. (2025)	50.18	–	–
Moryossef et al. (2023)	31.13	33.32	**66.24**
finetuned on BSLCP	51.46	55.46	–
finetuned + negative sampling	**52.09**	**56.04**	63.61

Table 3: Sign segmentation results on BSL Corpus test set and alignment results on BOBSL validation set using the same segmentation model in *Segment and Align*.

Lang.	Model	ISLR	BOBSL Align.
BSL (BOBSL)	SignCLIP-multilingual	0.5	66.70
	SignCLIP-BSL	**43.0**	**72.78**
	CSLR2 pseudo glosses	–	68.80
	CSLR2 human glosses*	–	78.75
ASL (PopSign)	SignCLIP-multilingual	3.0	35.51
	SignCLIP-ASL	**84.3**	**38.32**

Table 4: ISLR results on BOBSL sign spottings and PopSign ASL test sets with respect to alignment scores on validation sets in *Segment, Embed, and Align*. Results using CSLR2 glosses are also included as an ablation, and the human glosses* are marked as an oracle.

5 Conclusion

We introduce *Segment, Embed, and Align* (SEA), a universal method for aligning subtitles to sign language videos. Splitting the problem into sign *segmentation*, cross-modal *embedding*, and global *alignment* via DP, SEA attains SOTA alignment results across diverse languages and data sources. The modular design of SEA enables plug-and-play improvements—language-specific finetuning strengthens embeddings without requiring end-to-end alignment training on in-domain supervision.

Beyond establishing strong empirical results, SEA has practical value: it can harvest higher-quality parallel text–sign data at scale and serve as a flexible component in broader sign language processing pipelines. We release all components of SEA to encourage reproducible research and future advances.

Limitations

Our work has the following limitations: (1) The embedding and alignment models can be possibly improved iteratively with regard to each other, and we leave this to future work due to time and space limits. (2) Although already evaluated extensively on four diverse datasets, the robustness to dataset-specific biases and data quality can be further improved. For example, a mechanism for the algorithm to drop irrelevant signing or subtitles would alleviate the bad cases observed in How2Sign, and merging segments when the boundary is difficult to discern would improve accuracy and be useful for paragraph-level downstream applications. (3) Human-in-the-loop workflows, such as post-editing and evaluation, are not studied in this work, but we believe that it will be valuable to establish a semi-automatic process with human annotators' intervention to further enhance alignment quality and practical usability.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Proc. ECCV*.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv*.

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. 2021. Aligning subtitles in sign language videos. In *Proc. ICCV*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. CVPR*.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daumé III, Alex X. Lu, Naomi Caselli, and Danielle Bragg. 2023. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. In *NeurIPS 2023 Datasets and Benchmarks Track*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Proc. CVPR*.

Abhishek Dutta and Andrew Zisserman. 2019. The VIA annotation software for images, audio and video. In *Proc. ACM MM*.

Morton Ann Gernsbacher. 2015. Video captions benefit everyone. *Policy insights from the behavioral and brain sciences*, 2(1):195–202.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic.

Tengda Han, Weidi Xie, and Andrew Zisserman. 2022. Temporal alignment networks for long-term video. In *Proc. CVPR*.

Thomas Hanke. 2002. iLex - a tool for sign language lexicography and corpus analysis. In *Proc. LREC*.

Thomas Hanke, Silke Matthes, Anja Regen, and Satu Worseck. 2012. Where does a sign start and end? segmentation of continuous signing. In *Proc. LREC*.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. 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*.

Low Jian He, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Hands-on: Segmenting individual signs from continuous sequences. *arXiv*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Youngjoon Jang, Jeongsoo Choi, Junseok Ahn, and Joon Son Chung. 2025a. Deep understanding of sign language for sign to subtitle alignment. *arXiv*.

Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, and Andrew Zisserman. 2025b. Lost in translation, found in context: Sign language translation with contextual cues. In *CVPR*.

Zifan Jiang, Anne Göhring, Amit Moryossef, Rico Sennrich, and Sarah Ebling. 2024a. SwissSLi: The multi-parallel sign language corpus for Switzerland. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024b. SignCLIP: Connecting text and sign language by contrastive learning. In *Proc. EMNLP*.

Lee Kezar, Elana Pontecorvo, Adele Daniels, Connor Baer, Ruth Ferster, Lauren Berger, Jesse Thomason, Zed Sevcikova Sehyr, and Naomi Caselli. 2023. The sem-lex benchmark: Modeling asl signs and their phonemes. In *Proc. ACM SIGACCESSS*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. EMNLP*.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Liliane Momeni, Hannah Bull, K R Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *ECCV*.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Proc. EMNLP (Findings)*.

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 *Proceedings of the Eighth Conference on Machine Translation (WMT)*.

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. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. Learning transferable visual models from natural language supervision. In *Proc. ICML*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*.

Charles Raude, KR Prajwal, Liliane Momeni, Hannah Bull, Samuel Albanie, Andrew Zisserman, and Gül Varol. 2024. A tale of two languages: Large-vocabulary continuous sign language recognition from spoken language supervision. *arXiv*.

Katrin Renz, Nicolaj C. Stache, Samuel Albanie, and Gül Varol. 2021a. Sign language segmentation with temporal convolutional networks. In *Proc. ICASSP*.

Katrin Renz, Nicolaj C. Stache, Neil Fox, Gül Varol, and Samuel Albanie. 2021b. Sign segmentation with changepoint-modulated pseudo-labelling. In *Proc. CVPRW*.

Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. 2013. Building the british sign language corpus.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. ACL*.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam S. Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Anandi Vempala, Alec Tan, Jocelyn Heath, Unnathi Kumar, Priyanka Mosur, Tavenner Hall, and 5 others. 2023. Popsign asl v1.0: An isolated american sign language dataset collected via smartphones. In *Proc. NeurIPS*.

Garrett Tanzer. 2025. FLEURS-ASL: Including American Sign Language in massively multilingual multitask evaluation. In *Proc. NAACL*.

Garrett Tanzer and Biao Zhang. 2025. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. In *Proc. ICLR*.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus. In *Proc. NeurIPS*.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proc. LREC*.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proc. CVPR*.

A Extended Implementation Details

In this section, we provide an additional explanation of the alignment algorithm in SEA with Python-like pseudocode (Figure 4). Following §3, we first obtain segmented signs (line 2), and then embed these signs together with the subtitles (line 5-6).

Similarity matrix. The subtitle and signing embeddings form a similarity matrix $\mathbb{R}^{N_{\text{subtitles}} \times N_{\text{signs}}}$, where each entry is the dot product between a subtitle and a sign embedding, measuring their similarity (line 7). We explore two variants for assigning groups of consecutive signs to subtitle units, without/with the use of a similarity matrix. The subtitles are then shifted according to the matched signs. Following Bull et al. (2021), alignment adjusts subtitle timings to match the signing. Unlike frame-level classification in local windows, SEA globally assigns groups of consecutive signs to subtitle units and shifts subtitles accordingly.

Segment and Align. A *cost function* is defined for assigning a `span` of signs to a subtitle `cue` (line 12). The cost is a weighted sum of four terms: (a) distance between subtitle start and first sign; (b) distance between subtitle end and last sign; (c) duration difference between subtitle and span; (d) internal gaps within the span. By minimizing this cost globally, subtitles are aligned to signing using only prosodic cues and prior subtitle timing, referred to as *Segment and Align*.

Segment, Embed, and Align. To incorporate semantic associations between text and signing, we add a negative cost term, (e) `total_similarity`, which sums the subtitle–sign similarities from the matrix. With this addition, the process becomes *Segment, Embed, and Align*.

Since episodes can be almost 1 hour long and only nearby signs are assumed *alignable* with each subtitle, we restrict locality with a `win_size` (window size) parameter, zeroing out scores outside this window of closest neighbors (line 21). Remaining scores are row-normalized with `Softmax`, yielding a matrix of values concentrated around the temporal diagonal (line 22). We employ a dynamic programming algorithm to determine the global optimum (line 25). A `max_gap` constraint further picks the most similar sign subgroup for each subtitle with gaps no larger than this threshold, regulating the maximum gap of signs allowed within a subtitle, and discarding the rest (line 30).

```
1  # §3.1 Segment
2  signs = segmenter(video)
3
4  # §3.2 Embed
5  v_emb = embed_sign(signs)
6  t_emb = embed_text(subtitles)
7  sim_matrix = dot_product(t_emb, v_emb)
8
9  # §3.3 Align
10 # Define alignment cost function
11 # between a subtitle cue and a span of signs
12 def cost(cue, span):
13   a = abs(cue.start - span.start)
14   b = abs(cue.end - span.end)
15   c = abs(cue.duration - span.duration)
16   d = compute_internal_gap(span)
17   e = compute_total_similarity(cue, span)
18   return a + b + w_dur*c + w_gap*d - w_sim*e
19
20 # Filter and normalize similarity scores
21 sim_matrix = zero_out(sim_matrix, win_size)
22 sim_matrix = softmax_normalize(sim_matrix)
23
24 # Run dynamic programming
25 dp = aligner(signs,subtitles,cost,sim_matrix)
26 bounds = backtrack(dp)
27
28 # Postprocessing:
29 # refine span with best subgroup
30 for i in range(len(subtitles)):
31   group = signs[bounds[i]:bounds[i+1]]
32   subgroups = split_on_gap(group, max_gap)
33   alignments[i] = \\
34     pick_best(subgroups, subtitles[i], cost)
35
36 # Finally, update subtitle timestamps
37 for i in range(len(subtitles)):
38   subtitles[i].start = alignments[i][0].start
39   subtitles[i].end   = alignments[i][-1].end
```

Figure 4: Python-like pseudocode for SEA: segment, embed, and align using DP with a weighted cost function, followed by refining spans and updating subtitles.

Finally, the subtitles' start/end times are updated (line 37).

DP implementation. We maintain a table $dp[i, j]$, of the best total cost after aligning the first i subtitles to the first j signs:

$$dp[i, j] = \min_{i-1 \leq k < j} \{ dp[i-1, k] + \phi(t_i, s_{k:j-1}) \}$$

$dp[0, 0] = 0$, $dp[0, j > 0] = \infty$. After filling the table iteratively via the recurrence above, we do *not* require including all signs since false positives are often found in the episode end; instead we pick the best ending $j^\star = \arg\min_{0 \leq j \leq m} dp[n, j]$ and backtrack from $(n, j^\star)$ to $(0, 0)$ by the stored predecessors to recover the subtitle–sign assignment.

Parameter random search. For each of the datasets evaluated, we first conduct a random pa-

Parameter	BOBSL	How2Sign	WMT-SLT	SwissSLi
Start offset before DP	2.60	1.30	1.40	0.49
End offset before DP	2.10	1.50	1.60	0.56
Start offset after DP	0.00	0.00	0.00	0.00
End offset after DP	1.00	1.00	1.00	1.00
`b-threshold` for segmentation	30.00	40.00	20.00	20.00
`o-threshold` for segmentation	50.00	50.00	30.00	30.00
Window size (`win_size`)	50.00	50.00	50.00	50.00
Max inter-sign gap (`max_gap`)	10.00	8.00	6.00	6.00
Duration distance weight (`w_dur`)	1.00	5.00	0.50	0.50
Inter-sign gap weight (`w_gap`)	5.00	0.80	5.00	5.00
Similarity weight (`w_sim`)	10.00	10.00	5.00	1.00

Table 5: Best parameter settings per dataset (offsets and gaps in seconds).

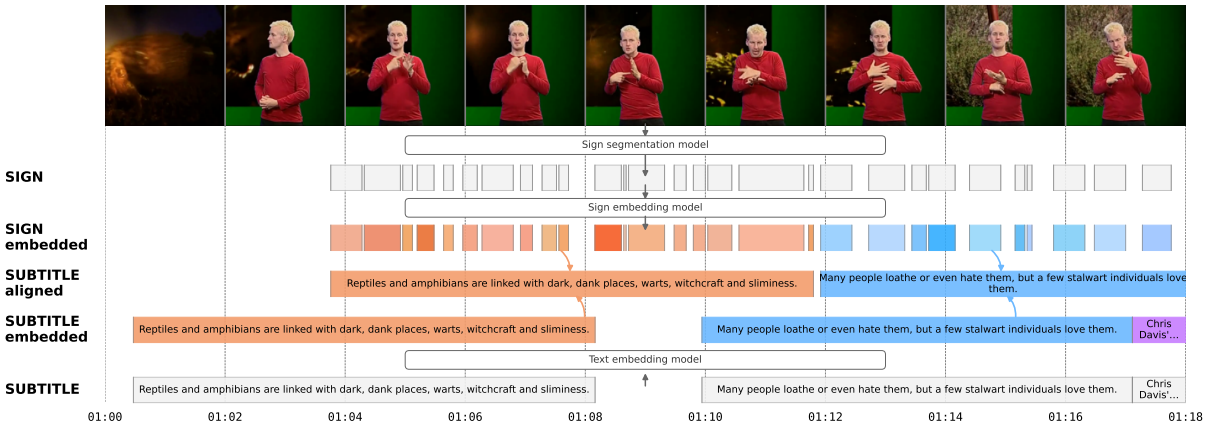


Figure 5: **Aligning subtitles to continuous signing using SEA:** From top down, signs are first segmented, then embedded (colored). From bottom up, subtitles are embedded into the same latent space as signs and then aligned accordingly to the direction of similarities.

rameter search of up to 10 000 iterations (approximately 1 day, running on 128 CPU workers) to find the optimal parameter combination under the condition of *Segment and Align*. We then inherit the best parameter combination when introducing the *Embed* step, only slightly adjusting the weight for similarity—`w_sim`—with the rest unchanged, as searching again with the embedding step is significantly more costly.

The values of these parameters are presented in Table 5. We note that these parameters might reflect dataset-specific biases learned from the training set. For SwissSLi, which lacks a training set, we transfer most parameters from WMT-SLT, a dataset of the same languages. The similarity weights for both are set relatively small, as the corresponding embedding models are relatively weakly finetuned on a small amount of data in DSGS/German.

B SEA Alignment Illustrated on BOBSL

In addition to the description in §A, Figure 5 provides a more visual, step-by-step illustration of how SEA alignment works using the example video with id `5224144816887051284` from BOBSL. Signs (from top down) and subtitles (from bottom up) are aligned based on their similarities in the embedding space.

C Human Annotation Details on SwissSLi

We hire DSGS experts to annotate 14 episodes of the *mitenand* program from the SwissSLi dataset (Jiang et al., 2024a). The main task is to manually align the subtitles of *mitenand* in German to better match the signing in DSGS.

The annotators are either DSGS interpreter students or deaf interpreters. For the three students, annotations were performed as part of a course project in exchange for 3 ECTS course credits. Participation was voluntary. For deaf interpreters, the annotation was compensated at the rate of ∼40 USD per hour with informed consent.

We append the annotation instructions below:

Annotation instructions. The videos we are using are from the mitenand in Gebärdensprache program. They are translated by deaf translators. You are asked to correct the annotation of the videos by ELAN.

You can find the videos and ELAN files in SWITCHdrive. After you are finished, upload the new ELAN files to the same location.

When you open an ELAN file with the corresponding video under the same folder, you will see four existing annotation tiers:

SUBTITLE: the German sentences from the subtitle/caption file, whose timings might not be perfectly aligned to the signing. SENTENCE: empty sentences detected by our AI model, which might also be inaccurate, but could serve as an indication for correcting the timings of the subtitles. SIGN: empty signs detected by our AI model, which might also be inaccurate, but could serve as an initial proposal for individual sign annotations. GLOSS: signs in the SIGN tier filled automatically with the glosses from the gloss files the deaf translators use. Note that the number of actual signs might vary from the detected ones.

You have two tasks:

On the sentence level, copy the SUBTITLE tier and create a new SUBTITLE_CORRECTED tier, where the start and end times of each sentence are corrected to match the signing. On the sign level, copy the GLOSS tier and create a new GLOSS_CORRECTED tier, where the start and end times of each gloss are corrected to match the signing, and the gloss name is corrected if the current one is wrong (just use intuitive glosses, they will be changed to match our DSGS lexical database later). In addition, remove superfluous gloss annotation as well as segments such as "Mit Untertiteln von SWISS TXT" and add missing gloss annotation if needed. If necessary, split segments (e.g., split "SAGEN NEIN" into "SAGEN" and "NEIN"). If lexical meaning is expressed nonmanually, please use a gloss in square brackets; e.g., if the meaning of "no" is expressed with a headshake, please create a segment "[NO]". Please apply narrow segmentation following the description from this article: "Transitional movements between signs do not count as part of either sign". Therefore, usually, there are gaps between two signs during which the articulators move from the end of one sign to the beginning of the next."

It would be great if you could also note a few things during the process:

How much time does it take you to finish each task? What is the most time-consuming part? Does the output of the AI model help you with the process, for example, the SENTENCE tier on sentence timing correction and the automatically generated glosses on gloss annotation? Any issue with the model's automatic annotation or the glosses from the deaf translators?