TTIC's Submission to WMT-SLT 23

Marcelo Sandoval-Castañeda
TTI-Chicago
marcelo@ttic.edu

Yanhong Li
TTI-Chicago
yanhongli@ttic.edu

Bowen Shi
Meta AI
bshi@meta.com

Diane Brentari
The University of Chicago
dbrentari@uchicago.edu

Karen Livescu
TTI-Chicago
klivescu@ttic.edu

Gregory Shakhnarovich
TTI-Chicago
gregory@ttic.edu

Abstract

We describe TTIC's submission to the WMT 2023 Sign Language Translation shared task on the Swiss-German Sign Language (DSGS) to German track. Our approach explores the advantages of using large-scale self-supervised pre-training in the task of sign language translation, over more traditional approaches that rely heavily on supervision, along with costly labels such as gloss annotations. The proposed model consists of a VideoSwin transformer for image encoding, and a T5 model adapted to receive VideoSwin features as input instead of text. On WMT-SLT 22's development set, this system achieves 2.03 BLEU score, a 59% increase over the previous best reported performance. On the official test set, our primary submission achieves 1.1 BLEU score and 17.0 chrF score. It also achieves the highest human evaluation score among all the participants.

1 Introduction

Sign language translation (SLT) is the task of translating a signed language to a written language, typically the lingua franca of the region the signed language is utilized. In recent years, SLT has received increased attention from the natural language processing (NLP) and computer vision (CV) communities.

The best-performing SLT models primarily rely on glosses (Zhou et al., 2021; Chen et al., 2022), a combination of morpheme translations into the target language along with differentiating phonological features like handshape and location. However, annotating glosses is expensive (Müller et al., 2023b), and recent research has begun to move away from gloss-based translation (Shi et al., 2022a; Uthus et al., 2023; Lin et al., 2023), particularly in regimes where larger datasets are available.

In this paper, we study large-scale self-supervision and noisy supervision for Swiss-German Sign Language (DSGS from the German *Deutschschweizer Gebärdensprache*) to German SLT, as part of the WMT-SLT 23 shared task (Müller et al., 2023a). Given recent findings on self-supervised transformers' perfomance on isolated sign recognition and feature extraction (Sandoval-Castañeda et al., 2023), we utilize a VideoSwin (Liu et al., 2022) visual feature extractor with BEVT pre-training (Wang et al., 2022). Additionally, we use T5 (Raffel et al., 2020) as a sequence-to-sequence translation model into German because of its state-of-the-art performance on American Sign Language (ASL) to English SLT with pose input (Uthus et al., 2023). Depending on the generation algorithm, our model achieves either the highest BLEU score (Papineni et al., 2002) or the highest chrF (Popović, 2015) in the task's leaderboard. With top-k beam sampling, it achieves 0.8 BLEU and 17.3 chrF, and with diverse beam search (Vijayakumar et al., 2016), it achieves 1.1 BLEU and 17.0 chrF.

2 Method

Our model follows the most common gloss-free translation architecture, composed of a visual encoding backbone and a transformer-based model for sequence modeling. Our visual backbone is a Video Swin Transformer (VideoSwin) and our sequence-to-sequence model is a Text-to-Text Transfer Transformer (T5).

2.1 VideoSwin

VideoSwin is an architecture proposed as an extension of the shifted-window transformer (Liu et al., 2021), a hierarchical vision transformer that relies on windowed self-attention for computational efficiency. We pre-train a VideoSwin using video-only BEVT pre-training (Wang et al., 2022) on OpenASL (Shi et al., 2022a), using the codebook from a discrete variational autoencoder (dVAE) (Ramesh et al., 2021) to produce the labels in the self-supervision objective. Though OpenASL is originally a sign language translation dataset, we ig-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 344–350
December 6–7, 2023.

nore the English translations and train exclusively on the dataset's videos. Then, we fine-tune on the gloss-based version (Dafnis et al., 2022; Neidle and Ballard, 2022) of WLASL2000 (Li et al., 2020) for supervised isolated sign language recognition.

Given a video with dimensions $16 \times 224 \times 224$, that is, 16 frames of height 224 pixels and width 224 pixels, VideoSwin first divides the input into patches of shape $2 \times 4 \times 4$ and produces a 128-dimensional vector representation for each patch, producing a tensor of shape $8 \times 56 \times 56 \times 128$. After the first two windowed self-attention blocks, patch representations are divided into non-overlapping groups of four spatially contiguous patches, which are then projected into a single 256-dimensional vector each. This is done again after two windowed self-attention blocks, and once more after eighteen windowed self-attention blocks. The resulting tensor after these patch merging steps has dimensions $8 \times 7 \times 7 \times 1024$.

For translation, we pad the video at the end such that the number of frames is a multiple of 16, divide it into non-overlapping segments of 16 contiguous frames, and run each segment independently through the model. The visual features extracted from the model are the output of the last windowed self-attention block from VideoSwin for each video segment. Then, we concatenate them across the time dimension, and remove the model's outputs that correspond to the padding frames. This is done both during training and during inference. More formally:

$$f_{1:\lceil T/2 \rceil} = M^v(I_{1:T}) \quad (1)$$

where $I_{1:T}$ is a sequence of T image frames, M^v is our VideoSwin model, and $f_{1:\lceil T/2 \rceil}$ is the resulting sequence of visual features, with dimensions $\lceil T/2 \rceil \times 7 \times 7 \times 1024$.

2.2 T5

T5 is a standard encoder–decoder text transformer (Raffel et al., 2020). Recent research has found that T5 pre-trained on English and fine-tuned for ASL to English translation produces state-of-the-art results using pose input (Uthus et al., 2023). We use a T5 model pre-trained on the German Colossal Cleaned Common Crawl (GC4) corpus, which is a cleaned and pre-processed German-only corpus based on Common Crawl. We take pre-trained checkpoints[1] from HuggingFace (Wolf et al., 2020).

[1] https://huggingface.co/GermanT5

Since our sequence of visual features $f_{1:\lceil T/2 \rceil}$ has dimensions $\lceil T/2 \rceil \times 7 \times 7 \times 1024$, we project these into a single vector per timestep, $\lceil T/2 \rceil \times 1024$. To this end, we use a simple convolutional layer with kernel size $1 \times 7 \times 7$. We replace the word embeddings layer from the T5 model with this convolutional layer. This is the only component trained from scratch in our DSGS to German translation model.

2.3 Training Loss

We use cross-entropy loss for BEVT pre-training, isolated sign language recognition (ISLR) fine-tuning, text-to-text pre-training, and features-to-text translation.

2.4 Inference

We expand on the effect of generation algorithms in Section 4.5. For our primary submission, our generation algorithm of choice is diverse beam search (Vijayakumar et al., 2016), with 5 beams, 5 beam groups, and a diversity penalty of 1.

3 Experimental Setup

3.1 Data

We use both last year's and this year's WMT-SLT datasets. Last year's training dataset is composed of data from FocusNews and SRF, both news TV programs, consisting of 17,207 manually aligned DSGS–German pairs, for a total of 35 hours. German text is obtained from the subtitles that correspond to the original spoken German content, and DSGS video is obtained from live translators. Manual alignment is necessary to ensure that each translated sentence in the video is assigned the correct German sentence. In contrast, this year's dataset consists of 231,834 DSGS–German pairs without any manual alignment, for a total of 437 hours, of only SRF data. Last year's SRF data is a subset of this year's dataset, with the key difference that the superset does not contain manually aligned and verified German translations.

Additionally, we use OpenASL (Shi et al., 2022a), a dataset consisting of 288 hours of ASL-English pairs, for the self-supervised pre-training of our visual encoder. In this pre-training we also employ the labels produced by the codebook of a dVAE, which was separately trained on Conceptual Captions (Sharma et al., 2018). For the second stage of pre-training of our visual encoder,

we fine-tune the pre-trained model on the gloss-based version of WLASL2000 (Li et al., 2020), a 14-hour dataset consisting of 19,673 isolated sign ASL videos and 1535 gloss labels (Neidle and Ballard, 2022).

Lastly, the checkpoint we use for T5 is pre-trained on the GC4 corpus. GC4 is a German-only corpus that contains 40.8 billion tokens in total. This is a subset of Common Crawl where the primary language is German extracted between 2015 and 2021.

3.2 Training

Our visual backbone is VideoSwin's base version. It consists of 88.1 million parameters, and is composed of 2 windowed self-attention blocks with 128 hidden dimensions at stage 1, 2 with 256 hidden dimensions at stage 2, 18 with 512 hidden dimensions at stage 3, and 2 with 1024 dimensions at stage 4. We pre-train it in two stages. First, we train it for 150 epochs on OpenASL via video-only BEVT where the labels are produced by the codebook of a dVAE, with a learning rate of 0.0005 on a cosine schedule with 10 warmup epochs and batch size of 128 across 8 GPUs. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and 0.05 weight decay. In the second stage, we train it on gloss-based WLASL2000[2] for classification for 120 epochs, this time with a learning rate of 0.0003 on a cosine schedule with 2.5 warmup epochs and a batch size of 256 across 8 GPUs. Again, we use AdamW as our optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and 0.001 weight decay. Our VideoSwin backbone is then frozen for the rest of our model's training.

For translation, we adapt T5's efficient-large (Tay et al., 2022) version using a convolutional layer to project our representations. This model is composed of 1.09 billion parameters, with 36 self-attention blocks in the encoder and 36 self-attention blocks in the decoder. To tokenize the target translations, we use a SentencePiece tokenizer trained on the same data as the German-only T5, with a vocabulary size of 32,128. We train it in two stages, using both WMT-SLT 22 and WMT-SLT 23 data. WMT-SLT 23 translations are weakly supervised labels, since there is no guarantee of alignment between the video and the corresponding text translations. Therefore, our pipeline uses it as a large, noisy dataset to train the model which will be eventually further fine-tuned with WMT-SLT 22, which has manually verified labels. First, we train it for 8500 steps on WMT-SLT 23's dataset, with a learning rate of 0.001 on a linearly decreasing schedule and a batch size of 64 across 8 GPUs. We use Adafactor (Shazeer and Stern, 2018) as the optimizer. For the second stage, we train the model for 1500 steps on WMT-SLT 22's dataset, with a learning rate of 0.0002 on a linearly decreasing schedule with a batch size of 64 across 8 GPUs. We also use Adafactor at this stage.

3.3 Evaluation

We evaluate our systems and compare them with last year's submissions, since we use the same validation set, using BLEU-1, BLEU-2, BLEU-3 and BLEU-4.

4 Experimental Results

Table 1 shows the performance of our model on WMT-SLT 22's development set, compared to the highest reported BLEU-4 scores reported on the test set by human evaluation (Müller et al., 2022). We also include MSMUNICH's model based on AV-HuBERT (Shi et al., 2022c), since it achieved the highest BLEU-4 score on the development set. Our model performs at least 81% better than the others in all metrics, and 99% better in BLEU-4, which is the metric used in the challenge's leaderboard.

We additionally perform several ablations to quantify the impact of our model's several moving parts. Our ablations are performed using T5's efficient-base configuration with 619 million parameters for time efficiency, unless otherwise specified.

4.1 Visual Backbone

We first evaluate the effect of our choice of visual backbone and pre-training tasks. We compare our VideoSwin backbone with two other models. First, we take a standard I3D model (Carreira and Zisserman, 2017) trained on the ISLR component of the BBC-Oxford British Sign Language dataset (Albanie et al., 2020), called BSL5K (Varol et al., 2021), since I3D is the most commonly used backbone for SL translation. Previous literature suggests that diversity of isolated signs leads to

[2]The original data can be downloaded here: https://dxli94.github.io/WLASL/ And the gloss-based labels can be downloaded here: https://dai.cs.rutgers.edu/dai/s/aboutwlasl

Model	Backbone	Translation Data	B1	B2	B3	B4
MSMUNICH (Dey et al., 2022)	AV-HuBERT	WMT-SLT 22	–	–	–	1.28
MSMUNICH (Dey et al., 2022)	I3D	WMT-SLT 22	–	–	–	0.77
UZH (Müller et al., 2022)	OpenPose	WMT-SLT 22	–	–	–	0.59
TTIC (Shi et al., 2022b)	I3D	WMT-SLT 22	8.36	2.92	1.55	1.02
Ours	**VideoSwin**	**WMT-SLT 22 + 23**	**15.19**	**5.62**	**3.06**	**2.03**

Table 1: Performance of our model on WMT-SLT 22's development set compared to WMT-SLT 22's highest reported scores. B1, B2, B3, and B4 stand for BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively.

better representations for downstream tasks like translation, and BSL5K is the largest and most diverse ISLR dataset to our knowledge. We also include an I3D model trained on WLASL2000 for comparison. Second, we also include a version of our pipeline where we replace OpenASL with the WMT-SLT 23 training data without translations for self-supervised pre-training. However, we do not include WLASL2000 fine-tuning for this model, given the language differences between DSGS and ASL.

As Table 2 shows, There is significant deterioration from shifting our self-supervised BEVT VideoSwin backbone to any of the fully supervised I3Ds. Similarly, despite being pre-trained in a different language, OpenASL pre-training performs much better than WMT-SLT 23 pre-training, despite being the smaller training set (288 vs. 437 hours). This is likely a product of OpenASL's far superior diversity in backgrounds, which are masked in WMT-SLT 23, topics (social media content vs. news), and signers (220 vs. 4).

Backbone	Data	B1	B2	B3	B4
I3D	ASL	12.15	2.96	1.31	0.79
I3D	BSL	12.79	2.80	1.12	0.59
BEVT	DSGS	12.43	3.34	1.72	1.16
BEVT	**ASL**	**15.16**	**5.20**	**2.75**	**1.82**

Table 2: Impact of visual backbone and training data on our model's performance. I3D refers to Inception3D models and BEVT refers to BEVT VideoSwin models. We group our pre-training data by language: BSL refers to BSL5K, DSGS refers to WMT-SLT 23, and ASL refers to OpenASL (if BEVT) and WLASL2000.

4.2 Translation Pre-Training

We also consider different combinations of our two DSGS to German translation datasets. In our training set-up, the model is first trained on WMT-SLT 23's weakly supervised labels, and then fine-tuned on WMT-SLT 22's manually aligned labels. We compare this to settings where we use either only WMT-SLT 23 data or only WMT-SLT 22 data. Using only WMT-SLT 22 data is equivalent to WMT-SLT 22's challenge.

From Table 3, we can see that despite the possible misalignments in WMT-SLT 23, training on a larger set of translation pairs is superior to using only WMT-SLT 22 data. However, the best performance we obtain comes from first training on the potentially noisy but large WMT-SLT 23, and then fine-tuning on WMT-SLT 22 for fewer steps.

W22	W23	B1	B2	B3	B4
✗	✓	14.28	4.33	2.27	1.58
✓	✗	13.47	4.30	2.19	1.42
✓	✓	**15.16**	**5.20**	**2.75**	**1.82**

Table 3: Impact of weak supervision translation labels on our model's performance. W22 refers to training on WMT-SLT 22 data and W23 refers to training on WMT-SLT 23 data. Where both are used, the model is trained on WMT-SLT 23 first and then on WMT-SLT 22.

4.3 Sequence-to-Sequence Model

In addition to T5, we also adapt Whisper (Radford et al., 2023) for DSGS to German translation and test it. The intuition behind it is that audio and video both have a time dimension that corresponds to seconds, whereas text does not. We adapt it in a similar fashion to T5, with the addition of a $4\times$ bicubic interpolation step right before the convolutional layer. We do so because Whisper receives input with 50 tokens per second, whereas our VideoSwin features produce one representation every two frames, for 12.5 every second, since the video is at 25 frames per second.

Results in Table 4 suggest that using a text-to-text model performs significantly better than a speech-to-text one.

Model	B1	B2	B3	B4
Whisper	15.08	4.26	2.04	1.29
T5	**15.16**	**5.20**	**2.75**	**1.82**

Table 4: Impact of sequence-to-sequence component of our model on translation performance.

4.4 Model Size

Next, we consider model size in Table 5. Due to computational and time constraints, we only evaluate T5-efficient-small, T5-efficient-base, and T5-efficient-large, with 142 million, 619 million, and 1.09 billion parameters respectively. As expected, larger models correspond to better performance.

Size	Params	B1	B2	B3	B4
Small	142m	15.43	5.13	2.47	1.52
Base	619m	15.16	5.20	2.75	1.82
Large	**1.09b**	**15.19**	**5.62**	**3.06**	**2.03**

Table 5: Impact of model size on our model's performance.

4.5 Decoding Algorithm

Last, we evaluate the effect of different choices of decoding algorithm on test set performance, using our best performing model, T5-efficient-large. We compare the results generated from the following algorithms: greedy decoding, top-k sampling (Fan et al., 2018), beam search, top-k beam sampling, and diverse beam search (Vijayakumar et al., 2016), with $k = 50$ and beam width set to 5. Table 6 shows our results from this experiment, revealing that diverse beam search and top-k beam sampling represent the most significant improvements from the greedy decoding baseline. We choose diverse beam search for our primary submission to the challenge, as it is the only one that improves both BLEU and chrF scores from our baseline.

Generation Algorithm	B4	chrF
Greedy Decoding	0.9	16.0
Top-k Sampling	0.8	16.3
Beam Search	0.9	17.2
Top-k Beam Sampling	0.8	17.3
Diverse Beam Search	**1.1**	**17.0**

Table 6: Impact of generation algorithm for our best model in WMT-SLT 23's test set.

5 Conclusion

Our experiments evaluate a hierarchical vision transformer on the task of sign language translation for the first time, and demonstrate superior performance over I3D-based translation models. We also show the benefits of using large datasets and self-supervised models for sign language translation, outperforming all previous fully supervised approaches to this task. Our final model achieves highest BLEU-4 score, highest chrF score, and highest human evaluation score among all participants of the task. However, translation quality remains extremely low.

Acknowledgements

This work was supported in part by the TRI University 2.0 program. We thank Shester Gueuwou for helpful discussions about sign languages and translation.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Konstantinos M. Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri Metaxas. 2022. Bidirectional skeleton-based isolated sign recognition using graph convolution networks and transfer learning. In *13th International Conference on Language Resources and Evaluation (LREC)*.

Subhadeep Dey, Abhilash Pal, Cyrine Chaabani, and Oscar Koller. 2022. Clean text and full-body transformer: Microsoft's submission to the WMT22 shared task on sign language translation. *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-bonet, Anne Goering, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023a. 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, et al. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023b. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Carol Neidle and Carey Ballard. 2022. Why alternative gloss labels will increase the value of the WLASL dataset. *ASL-LRP Project Report No. 21*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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. *Journal of Machine Learning Research (JMLR)*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Marcelo Sandoval-Castañeda, Yanhong Li, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. Self-supervised video transformers for isolated sign language recognition. *arXiv preprint arXiv:2309.02450*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022a. Open-domain sign language translation learned from online video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022b. TTIC's WMT-SLT 22 sign language translation system. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022c. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations (ICLR)*.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pretraining and finetuning transformers. In *International Conference on Learning Representations (ICLR)*.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A large-scale, open-domain american sign language-english parallel corpus. *arXiv preprint arXiv:2306.15162*.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y. Jiang, L. Zhou, and L. Yuan. 2022. BEVT: BERT Pretraining of Video Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2021. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*.