From Sign Spottings to Spoken Language: Fine-Tuning Large Language Models for Improved Translation

Markus Hoehn* **Lane Burgett** **Semyon Zharkov** **Andrei Mazin**

Manhattan High School
Manhattan, KS 66502
{markush, lanbur, semyonzharkov, amazin}@ksu.edu

Abstract

The task of sign language translation is crucial for enhancing communication for the deaf and hard-of-hearing community. In this paper, we build upon previous work by enhancing the large language model (LLM) component of a hybrid approach for sign language translation from continuous video streams. The prior hybrid approach employed a GPT-3.5 Turbo prompt to generate translations from identified sign spottings. We improve this by incorporating fine-tuning techniques and revising the prompt. We employ a sign spotter from existing literature to identify individual gestures within the video stream. These identified gestures are then processed by an LLM, which constructs grammatically correct and coherent sentences. Our evaluation of two models demonstrates significant improvements, with the fine-tuned Gemini-1.0 Pro model showing the most successful enhancement in translation accuracy and coherence.

1 Introduction

The main form of communication for millions of deaf and hard of hearing people around the world is sign language. Sign languages are complex communication systems that utilize hand gestures, facial expressions, and body language to convey meaning [11]. Converting sign language to spoken language is a task known as Sign Language Translation (SLT) and can help people communicate to deaf and hard of hearing people.

SLT, when treated as a Neural Machine Translation (NMT) task, struggles due to the challenges of aligning and tokenizing continuous sign language videos and the grammatical differences and word order between spoken and sign languages [3]. To address this issue some researchers approach sign language as a two-step process: first, Sign Language Recognition (SLR) identifies individual signs (i.e., sign spotting) within the video, and then, these signs are translated into meaningful sentences to achieve SLT [4, 5, 16] .

Large Language Models (LLMs), trained on extensive web-scale text corpora, have demonstrated significant capabilities in natural language processing tasks, such as multilingual translation. Their ability to understand and generate text across multiple languages with diverse syntactic and lexical properties highlights their versatility and effectiveness [10]. Given the distinct grammatical structures of spoken and sign languages and the different sequences of sign glosses (written representations of signs) and spoken word order, LLMs hold promise for converting sign spottings (identifications of individual signs) into spoken sentences due to their rich semantic understanding.

*Corresponding Author

Preprint. Under review.

A novel approach in the field, introduced recently in [13], involved using a sign spotter to identify hand gestures (glosses) for each video frame, subsequently processed by a large language model (LLM) to convert to spoken language. However, the LLM component was limited to prompting the 'GPT-3.5 Turbo' version of ChatGPT with a brief task description, lacking the integration of fine-tuning [15].

Our contributions to the approach introduced in [13] include: (1) Modifying the system input prompts for better sentence generation. (2) Fine-tuning LLMs specifically for converting prompted sign spottings into coherent sentences [15]. (3) Evaluating the performance of these fine-tuned models to identify the top performers.

2 Related Work

Early tasks in sign language automation primarily focused on SLR. Initially, due to technical limitations, research centered on hand-crafted features that analyzed hand shape and motion. Subsequently, pose, face, and mouth features were extensively incorporated into recognition pipelines [6, 2, 1, 7].

Modern approaches to SLT incorporate deep learning architectures, particularly Convolutional Neural Networks (CNNs), to learn robust features directly from video data. For instance, some papers proposed a CNN-based architecture for sign language translation, achieving significant improvements over traditional methods [3]. These deep learning approaches can capture complex spatial and temporal relationships within sign language gestures, leading to more accurate recognition [5, 16].

More recently, researchers have proposed leveraging LLMs and their large linguistic capability to advance SLT [14]. Approaches utilizing gloss-free translations performed considerably worse in terms of BLEU score [9] compared to their gloss-based counterparts [14]. Most notably, Sincan et al. [13] have proposed a method of translating sign spottings to spoken language through an LLM. In this paper, the authors sent a basic prompt to a pre-trained 'GPT-3.5 Turbo' model. The model did not include fine-tuning. We leverage the approach by Sincan et al. [13] and specifically explore the benefits of fine-tuning on two models.

Fine-tuning refers to a transfer learning approach where a pre-trained model's parameters are adjusted using new data. This adjustment can involve training the entire neural network or only specific layers [15]. For example, an LLM can be provided with a prompted input of sign spottings and fine-tuned to generate coherent spoken language.

3 Method

We use the Spotter from [13] and then feed Spottings over 0.7 confidence into our LLM component.

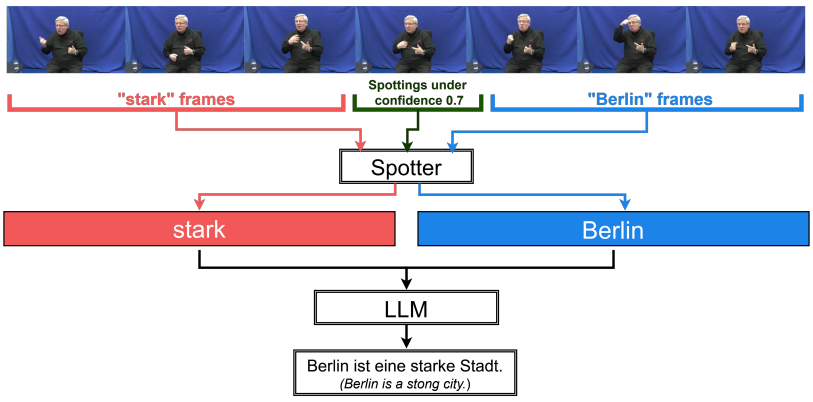

Figure 1: An overview of the proposed sign language translation architecture.

We use a modified system prompt for all our LLM components. For Gemini-1.0 Pro, it is added as another user prompt. Adapted from [13], this prompt is altered to always generate a sentence, eliminating the "No Translation" option. This ensures better comparison with fine-tuned models, which do not produce "No Translation" due to their training data.

```
Prompt: "You are a helpful assistant designed to generate a sentence based
on the list of words entered by the user. You need to strictly follow
these rules:

    1. The user will only give the list of German words separated by a
       space. You must generate a meaningful German sentence from them,
       even if you have to guess.
    2. Only provide a response containing the generated sentence. If you
       cannot create a coherent German sentence from the words, still make
       an attempt to form a German sentence using the given words."
```

We fine-tune two models, 'GPT-3.5 Turbo' and 'Gemini-1.0 Pro,' using glosses (except the INDEX gloss which stands for pointing) directly from the dataset and separately the spotter's sign spottings on the dataset's videos.

4 Experiments

We use the MeineDGS dataset which is a comprehensive linguistic resource for German Sign Language (DGS) [8]. The videos feature natural conversations between two deaf participants. We adhere to a sign language translation protocol for our splits [11] to compare to [13], which includes 40,230 training sentences, 4,996 development sentences, and 4,997 test sentences.

We fine-tune models (with hyperparameters: epochs = 1, batch size = 26, LR multiplier = 2) using the protocol's train split for training and for GPT-3.5 Turbo we use the protocol's dev split for training validation [11]. For user input, we use the glosses in the split (excluding the INDEX gloss) or the predicted sign spottings. For model output, we use the provided spoken language directly from [8].

We use BLEU [9] and BLEURT [12] metrics to evaluate our approach.[2] We use the sacreBLEU implementation for BLEU scores and BLEURT-20 checkpoints for BLEURT scores, evaluating various LLM implementations [9, 12]. The evaluation are conducted on the protocol's test split [11], comparing to the original spoken language sentences from [8].

Table 1: Benchmarks of different models

Method	BLEU1	BLEU2	BLEU3	BLEU4	BLEURT
On Spottings					
GPT-3.5 Turbo (scores from [13])	14.82	4.19	**1.45**	**0.64**	21.62
GPT-3.5 Turbo (modified prompt)	19.23	1.32	0.21	0.04	27.26
GPT-3.5 Turbo (fine-tuned on glosses)	14.97	0.99	0.11	0.01	27.54
GPT-3.5 Turbo (fine-tuned on spottings)	14.38	0.80	0.12	0.02	25.59
Gemini-1.0 Pro (modified prompt)	19.28	1.40	0.24	0.05	27.72
Gemini-1.0 Pro (fine-tuned on glosses)	24.09	3.13	0.80	0.29	35.03
Gemini-1.0 Pro (fine-tuned on spottings)	**26.73**	**3.60**	0.92	0.37	**35.41**
On Glosses					
GPT-3.5 Turbo (modified prompt)	28.38	4.50	1.30	0.48	40.73
GPT-3.5 Turbo (fine-tuned on glosses)	20.87	2.80	0.69	0.20	39.85
GPT-3.5 Turbo (fine-tuned on spottings)	18.15	1.95	0.46	0.14	33.85
Gemini-1.0 Pro (modified prompt)	28.70	4.74	1.36	0.48	42.54
Gemini-1.0 Pro (fine-tuned on glosses)	36.55	9.86	**4.09**	**1.98**	**51.65**
Gemini-1.0 Pro (fine-tuned on spottings)	**37.47**	**10.05**	4.08	1.95	49.30

[2]The BLEU and BLEURT metrics were computed on a system with a Ryzen 9 3950X processor and 128GB of memory. Sign spottings were computed on an NVIDIA A40 GPU.

Table 2: Selected Example from MeineDGS test split

Glosses	TAUB-GEHORLOS KENNENLERNEN BEIDE GUT
Spotter	TAUB-GEHORLOS KENNENLERNEN BEIDE GUT
Spoken Language	Er war gehörlos, wir haben uns kennengelernt und gut verstanden. *(He was deaf, we met and got along well.)*
GPT-3.5 Turbo ([13] prompt)	No Translation
GPT-3.5 Turbo (modified prompt)	Beide können sich gut kennenlernen, obwohl einer taub und der andere gehörlos ist. *(They both get to know each other well, even though one is deaf and the other is deaf.)*
GPT-3.5 Turbo (fine-tuned on glosses)	Und ich habe viele taube Kinder kennengelernt, es war alles gut. *(And I met a lot of deaf children, everything was good.)*
GPT-3.5 Turbo (fine-tuned on spottings)	Meine Freundin und ich kennen uns. *(My girlfriend and I know each other.)*
Gemini-1.0 Pro ([13] prompt)	Taub-Gehörlose lernen sich beide gut kennen. *(Both deaf people get to know each other well.)*
Gemini-1.0 Pro (modified prompt)	Taub-Gehörlose kennenlernen beide gut. *(Deaf people get to know both people well.)*
Gemini-1.0 Pro (fine-tuned on glosses)	Ich habe einen Gehörlosen kennengelernt und wir verstehen uns gut. *(I met a deaf person and we get along well.)*
Gemini-1.0 Pro (fine-tuned on spottings)	Ich habe einen Gehörlosen kennengelernt, der war sehr nett. *(I met a deaf person who was very nice.)*

Table 2 includes results from using the prompt in [13] via our calling code.

5 Discussion

Fine-tuning improvements among BLEURT and BLEU scores, particularly with Gemini-1.0 Pro, show potential for application to other datasets with different spotters. This facilitates a complete gloss-based sign language translation system, enhancing the LLM's ability to generate coherent and accurate translations, proving effective in SLT. Fine-tuning also provides qualitative improvements and prevents outputs like "The generated sentence is..." and other unwanted text, such as English text, which can occur even with an explicit system prompt on a pre-trained model. However, fine-tuning has a drawback: without training data designating some spottings as untranslatable, the model won't acknowledge its inability to translate and may generate long sentences, even with a system prompt addressing this limitation. We observe that fine-tuning increases the use of some words and phrases disconnected from the input spottings/glosses. Further exploration of prompt engineering to discourage these behaviors is warranted. These behaviors may also explain the differences between BLEU scores among different n-gram precisions. Our benchmarks on the dataset's glosses show capabilities for a theoretically perfect spotter. We recognize that Gemini-1.0 Pro outperforms GPT-3.5 Turbo for this task when fine-tuned. Finally, SLT has significant potential for improving communication between deaf and hearing people. One promising application is sign language teaching, where models like ours can act as virtual tutors, providing instant feedback and translation during practice sessions.

References

[1] Epameinondas Antonakos, Anastasios (Tassos) Roussos, and Stefanos Zafeiriou. A survey on mouth modeling and analysis for sign language recognition. 05 2015.

[2] Patrick Buehler, Andrew Zisserman, and Mark Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968, 2009.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. *CoRR*, abs/2003.13830, 2020.

[5] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033, 2020.

[6] Ali Farhadi, David Forsyth, and Ryan White. Transfer learning in sign language. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[7] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 12 2015.

[8] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release, 2020.

[9] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[11] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696, 2020.

[13] Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. Using an llm to turn sign spottings into spoken language sentences. *arXiv preprint arXiv:2403.10434*, 2024.

[14] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. `https://D2L.ai`.

[16] Biao Zhang, Mathias Müller, and Rico Sennrich. Sltunet: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract claims that we have improved the LLM component of the hybrid approach to SLT. Our metrics support this claim and we provide insights into potential further innovations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of our approach are discussed in the paper's discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explain all the parameters, dataset splits, and models we use in our experiments section. The paper provides ample description for reproducing the main claims and conclusions in the method and experiments sections.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper does not provide open access to the code; however, it mentions the tools and models used and includes descriptions on how they can be faithfully reproduced in our method and experiments section.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All necessary details about fine-tuning and testing were included in the experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: No, each entry in our table is from running the score calculation on the whole test split once. However the large test split should minimize deviations between runs. Error bars are not reported because it would be too expensive to run our tests multiple times using the model APIs. There were also rate limitations to how fast we could test the models.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Specifications for the system used to train and test our solution are listed in the footnotes of our experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper fully adheres to the NeurIPS code of ethics and does not violate any of its guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss how our proposed solution can enhance engagement with the deaf community at the end of our discussion section. There are no immediate negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks no data or models are released.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All conditions for the code and data used are met and cited. We cite the Spotter we used and cite the DGS corpus which may freely be used for scientific research.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.