Prompting with Sign Parameters for Low-resource Sign Language Instruction Generation

Md. Tariquzzaman, Md Farhan Ishmam, Saiyma Sittul Muna, Md. Kamrul Hasan, Hasan Mahmud
Department of Computer Science and Engineering
Islamic University of Technology, Bangladesh
{tariquzzaman,farhanishmam,saiymasittul,hasank,hasan}@iut-dhaka.edu

Abstract

Sign Language (SL) enables two-way communication for the deaf and hard-of-hearing community, yet many sign languages remain under-resourced in the AI space. Sign Language Instruction Generation (SLIG) produces step-by-step textual instructions that enable non-SL users to imitate and learn SL gestures, promoting two-way interaction. We introduce BdSLIG, the first Bengali SLIG dataset, used to evaluate Vision Language Models (VLMs) (i) on under-resourced SLIG tasks, and (ii) on long-tail visual concepts, as Bengali SL is unlikely to appear in the VLM pre-training data. To enhance zero-shot performance, we introduce Sign Parameter-Infused (SPI) prompting, which integrates standard SL parameters, like hand shape, motion, and orientation, directly into the textual prompts. Subsuming standard sign parameters into the prompt makes the instructions more structured and reproducible than free-form natural text from vanilla prompting. We envision that our work would promote inclusivity and advancement in SL learning systems for the under-resourced communities.

1. Introduction

Sign Language (SL) differs fundamentally from spoken or written languages as it inherently requires communication in the visual modality. At a lexical level, these SL gestures are typically encapsulated in a short video of a few seconds [1, 2]. With the current advancements in computer vision models [3–5], there has been significant work on automated processing of SL video clips, primarily for American Sign Language (ASL) recognition tasks [6, 7].

Despite the proliferation of ASL recognition works, we highlight two major research gaps in this domain. Firstly, the overwhelming focus on ASL works has marginalized non-ASL languages [8], limiting the *inclusivity* and *generalizability* of the models [9]. Secondly, SL recognition addresses only one half of the SL communication process, *i.e.*, understanding the SL user. The other half, *i.e.*, communicating back to the SL user, requires *learning* and *producing* the SL gestures - a task which is relatively less popular.

Current literature has provided methods for non-ASL and under-resourced SL recognition tasks [8, 10]. For learning SL, LLM-based automated SL Instruction Generation (SLIG) [11] creates textual instructions that guide the users to replicate the SL gesture. Automated gesture synthesis [12] can also be a plausible option for communicating back to the SL user, albeit with limited real-life applicability. However, there has been no prior work on the intersection of the aforementioned research problems, *i.e.*, SLIG for low-resource languages. Our work serves as a pilot project to address this problem space.

SLIG in non-ASL languages is critical, as each SL possesses unique gestural cues, shaped by linguistic, cultural, and regional factors [13, 14]. These variations mirror linguistic differences in spoken or written languages. We chose Bengali for our non-ASL evaluation language as it remains an under-resourced language in the automated SL space [15], while being native to 300 million active speakers, making it a high-impact choice for real-world applicability. Bengali SL also predominantly focuses on recognition tasks [2, 16, 17]; hence, our proposed BdSLIG is the first low-resource and Bengali SLIG dataset.

We hypothesize that a Vision Language Model (VLM) [18, 19], when given a frame or short video of a sign, can generate a textual description (*e.g.*, "both hands raised near face, wrists bent inwards") that captures the sign's meaning and form. Such descriptions could serve as interactive instructions for learners (analogous to language learning feedback [20]) and also provide interpretability by explaining what the model "sees." This idea builds on prior work in explainable sign technology, *e.g.*, the Learn2Sign system [21], which provided feedback on location, movement, and handshape to ASL learners. Since Bengali Sign Language is a low-resource language and sign recognition models are still underdeveloped, it is essential to analyze how large models behave in this context. Prior research [22] has shown

that general LLMs may produce overly generic or semantically off-base sign descriptions without domain-specific guidance. In light of these limitations, our contributions can be summarized as follows:

- We propose the first dataset to evaluate the performance of VLMs on the under-resourced task of Bengali sign language instruction generation.
- We introduce Sign Parameter-Infused (SPI) prompting, which integrates standard sign language parameters, *e.g.*, hand shape, motion, and orientation, directly into the textual prompts of the VLMs. Experiments reveal SPI prompting generally leads to better performance.
- Our experiments also serve as an evaluation of the performance of VLMs on long-tail distributed and semantically fine-grained visual concepts.

2. Methodology

In this section, we first define the sign language parameters (§2.1) that will be relevant to our prompting technique (§2.3). We also construct the BdSLIG dataset (§2.2) and define our evaluation framework (§2.4).

2.1. Sign Language Parameters

Sign language parameters are essential visual cues used to identify or describe a sign language gesture [23]. Our work uses the following parameter categories, defined below with a few associated keyword tags:

1. **Handshape**: The morphological configuration, *e.g.*, raised, extended, pointed, bent, spread, twisted, of the fingers and palms.
2. **Movement Type**: The trajectory or action of the hands, *e.g.*, static, circular, squeezing, grabbing, rotating, pulling.
3. **Location**: The placement of the hands while producing the sign, *e.g.*, near head, forehead, eye, ear, nose, mouth, chest, shoulder.
4. **Palm Orientation**: The direction the palm is facing relative to the signer's body, *e.g.*, inward, outward, upward, downward, left, right.
5. **Spatial Interaction**: The role of one or both hands in the sign's spatial scheme, *e.g.*, right-hand active, left-hand active, both-hands symmetric, both-hands asymmetric.
6. **Temporal Dynamics**: The temporal pattern of the movement onset, duration, and repetition, *e.g.*, sustained hold, repeated oscillation, single tap, repeated tapping.
7. **Facial Cues**: Facial expressions, *e.g.*, neutral, brows-raised, brows-furrowed, mouth-open, mouth-closed, lip-pursed, that can disambiguate the sign language.

These sign-language parameters serve two roles in our work: (i) to establish the annotation guidelines for human annotation of the BdSLIG dataset, and (ii) to define the Sign Parameter-Infused (SPI) prompting for evaluating zero-shot capabilities of vision-language models (VLMs). Fig. 1 provides a visual overview of these parameters, where each category is illustrated with example video frames sampled from the BdSLIG dataset.

2.2. BdSLIG Dataset

The images of BdSLIG are sourced from the BdSLW60 dataset [15], a 60-word Bengali sign language recognition dataset. For each word, we selected one representative video and manually annotated the textual sign instructions. We hired three trained annotators following the annotation guidelines with our defined sign parameters (§2.1). The annotations were validated in two stages: first for grammatical and semantic consistency of the text, then by a Bengali sign language domain expert. Both the annotators and validators received monetary compensation for their work.

2.3. Prompting Framework

Our prompting framework consists of three stages: (i) Frame Sampling, (ii) Sign Parameter Injection, and (iii) Instruction Generation. The second stage uses Sign Parameter-Infused (SPI) prompting, whereas the corresponding vanilla prompting strategy disregards this step.

2.3.1. Frame Sampling

A video input is taken from the framerate-standardized version, where all videos are adjusted to 30 frames per second. As inferring a high number of frames can be challenging to the VLM and increase the computational overhead, we uniformly sample every 20^{th} frame from the video. This reduces the input length and thereby decreases the context fed to the VLM while retaining most of the temporal information.

The limitations of uniform sampling of video frames in the context of sign languages are two-fold. Firstly, the motion of the sign is not uniform, *i.e.*, there can be periods of high and low activities. Hence, extracting an equal number of frames from high and low activities should not be optimal. Secondly, every k^{th} frame sampled uniformly may not be a *key frame*. We define the key frames as the minimum set of frames that capture the motion of the sign language. While there can be subjectivity in picking the key frames, the extraction procedure will surely not be uniform. Despite these limitations, we decided to adopt uniform sampling for its simplicity, and as our work serves as a pilot study.

2.3.2. Sign Parameter Injection

Given a set of sign language parameters $\mathcal{S}$ and a base textual prompt $\mathcal{P}$, we define our Sign Parameter-Infused (SPI) prompt $f(\mathcal{P}, \mathcal{S})$. Specifically, our SPI prompting describes the visual motor cues, as seen in Fig. 1, to obtain a stepwise, narrative-style instruction set.

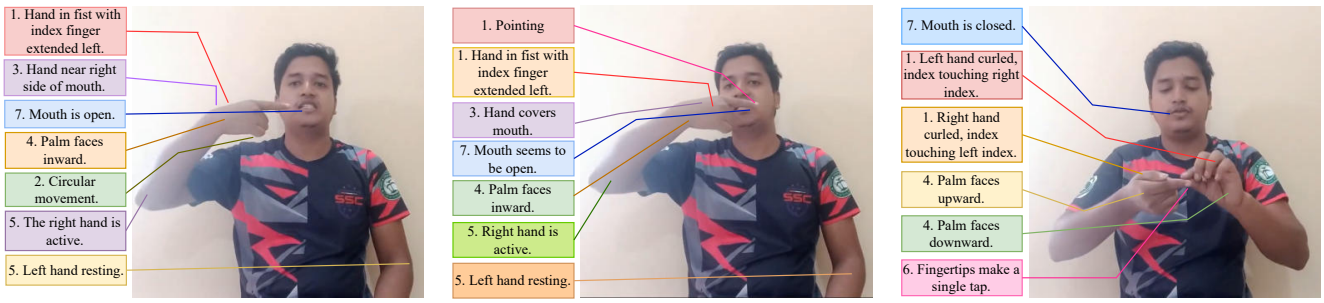

Figure 1. Parameter-based description of the `Toothpaste` sign (video taken from the BdSLIG dataset). The numbers denote the parameter numbers mentioned in §2.1. The parameter descriptions are given to elucidate and exemplify the given SL parameters and *are not* part of the BdSLIG dataset.

2.3.3. Instruction Generation

We define the generated instruction as a sequence of answer tokens, $\mathbf{a} = \langle a_1, a_2, \ldots, a_n \rangle$. For a generative Vision Language Model (VLM), g, taking a video feed or a set of video frames $\mathcal{V}$ and a prompt $\mathcal{P}$, we define instruction generation for vanilla prompting:

$$a_i = \arg\max \mathbb{P}(a|s(\mathcal{V}), \mathcal{P}, a_{<i}), \tag{1}$$

where $s(\cdot)$ is the video sampling function. Similarly, for SPI prompting, the generated instructions are:

$$a_i = \arg\max \mathbb{P}(a|s(\mathcal{V}), f(\mathcal{P}, \mathcal{S}), a_{<i}), \tag{2}$$

where $\mathcal{S}$ is the set of sign language parameters.

2.4. Evaluation Framework

We primarily have two evaluation goals: (i) assessing the quality of the generated instructions on the BdSLIG dataset and (ii) evaluating the efficacy of SPI prompting over vanilla prompting. A quantitative assessment can be performed by benchmarking state-of-the-art VLMs (§2.4.1) and comparing vanilla and SPI prompting strategies using standard text evaluation metrics (§2.4.2).

2.4.1. Models

We evaluate four state-of-the-art VLMs: (i) GPT 4.1-mini, (ii) GPT 4.1, (iii) Gemini 2.0 Flash, and (iv) Gemini 2.5 Pro. All these are proprietary models with reasonably high scores in LLM benchmarks, *e.g.*, LMArena leaderboard[1].

2.4.2. Evaluation Metrics

The evaluation of sign language instructions using functions, heuristics, and metrics can be tricky. We first emphasize that ideal evaluation can only be achieved by a human, preferably a sign language educator or expert. Text matching or similarity metrics *do not* evaluate the correctness of the generated instruction. The metrics capture text-level and semantic similarity. The suite of metrics used in our benchmarks is: ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR, and BERTScore. Each metric has its own limitations, and we believe using all would provide a holistic view of the generated textual instructions.

[1] https://lmarena.ai/leaderboard

3. Results and Analysis

This section is centered around our benchmark results from Tab. 1. We also present several analyses and discussions on the validity of the results and scopes for improvement.

3.1. SPI *vs*. Vanilla Prompting

Tab. 1 reports how well VLM-generated instructions align with our expert annotation from the BdSLIG dataset. SPI prompting consistently outperforms Vanilla Prompting across most metrics, especially METEOR and BERTScore. These metrics emphasize semantic similarity and are less sensitive to surface-level lexical variations, suggesting that SPI prompting helps the VLMs generate more structured and semantically faithful instructions, crucial in low-resource sign language contexts.

3.2. Larger *vs*. Smaller Models

Larger models like GPT-4.1 and Gemini-2.5-Pro benefit most, showing strong gains in both lexical and semantic alignment. While smaller variants like GPT-4.1-mini and Gemini 2.0 Flash also show improvement under the SPI setting, their limited capacity to interpret nuanced motion cues leads to weaker alignment with ground-truth annotations. Overall, GPT-4.1 takes the crown by achieving the highest score in all metrics except BERTScore.

3.3. Can VLMs generate acceptable instructions?

Answering this question can be quite nuanced. The general text scores, both lexical and semantic matching, can be deemed relatively low. If we look at other text generation tasks, *e.g.*, machine translation, the generative models tend

Prompting	Model	Metric					
		ROUGE-1	**ROUGE-2**	**ROUGE-L**	**BLEU**	**METEOR**	**BERTScore**
Vanilla	GPT 4.1-mini	0.526	0.232	0.342	0.174	0.396	0.396
	GPT 4.1	0.528	0.220	0.335	0.149	0.416	0.394
	Gemini 2.0 Flash	0.464	0.209	0.328	0.141	0.329	0.416
	Gemini 2.5 Pro	0.505	0.224	0.338	0.166	0.354	0.396
SPI (Ours)	GPT 4.1-mini	0.522 ↓	0.224 ↓	0.326 ↓	0.162 ↓	0.447 ↑	0.396 =
	GPT 4.1	**0.553** ↑	**0.271** ↑	**0.384** ↑	**0.196** ↑	**0.492** ↑	0.441 ↑
	Gemini 2.0 Flash	0.505 ↑	0.250 ↑	0.365 ↑	0.168 ↑	0.359 ↑	**0.456** ↑
	Gemini 2.5 Pro	0.507 ↑	0.212 ↓	0.319 ↓	0.141 ↓	0.365 ↑	0.387 ↓

Table 1. Evaluation of Vanilla and SPI prompting on our BdSLIG dataset.

to perform much better. However, most such tasks have *objective* ground truths, *i.e.*, there is less flexibility in annotating the answer. Instructions required to perform or imitate a sign are very *subjective*, *i.e.*, there can be multiple ways to write the instructions to perform the same sign.

3.4. How accurate are the metrics?

Following §2.4.2, we emphasize that lexical and semantic similarity metrics are not suitable for evaluating sign language instructions. To elucidate, we present two first-step instructions for the sign word `toothpaste` (Fig. 1).

Sample-1: Instruction Step-1 for *toothpaste*

The right hand, with fingers gently curved and palm facing toward the signer, is brought up to the mouth area. The fingertips lightly touch or hover near the lips, with the hand at chin or mouth level. The left hand remains at rest.

Sample-2: Instruction Step-1 for *toothpaste*

The right hand moves toward the mouth with the index finger fully extended and pointing straight ahead, while the other fingers are curled into a loose fist. At the same time, the mouth opens slightly, revealing the teeth clearly.

These two sample instructions essentially imitate the same hand motion. To a human learner, the differences in execution are small. Yet, due to the variations in wording, order of description, and emphasis on certain body positions, the text evaluation metrics would produce a considerably lower score. This is evident as most of these metrics quantify surface-level lexical overlap. Metrics like BertScore convey some form of semantic similarity, but not in the context of sign language. A possible approach is to use a larger and more contextual metric to evaluate the text pairs, *e.g.*, using LLM-based evaluation methods [24].

3.5. Long-tail Evaluation

The current generation of VLMs is trained on large-scale web corpora [4]. For most evaluation tasks, *e.g.*, question answering on natural images, it is highly likely that the evaluation data is present in the pre-training corpus of the VLM [25]. However, as Bengali is a low-resource language, it is highly unlikely that the VLMs are pre-trained on Bengali sign language and the associated instruction data, *i.e.*, it is in the tail end of the pre-training corpus distribution. Hence, BdSLIG can be a good benchmark for visual long-tail distribution. Following this setting, GPT-4.1 takes the crown as the best-performing model in both prompting strategies as per Tab. 1. Hence, GPT-4.1 should have relatively better visual understanding and grounding.

4. Conclusion

We present (i) BdSLIG, a novel Bengali sign language instruction generation dataset, and (ii) Sign Parameter-Infused (SPI) prompting, a prompting strategy that injects sign language parameters into vision-language prompts. Instruction generation following standard sign parameters makes the description structured and reproducible, unlike free-form text. Canonicalizing the descriptors can be suitable for classification, retrieval, or alignment tasks. Our experiments investigate the efficacy of SPI prompting, long-tail evaluation of state-of-the-art models, and metrics for evaluating sign language instruction. Our work serves as a pilot project to promote research in multimodal generative systems for individuals with accessibility needs.

Reproducibility

Our dataset is available on HuggingFace: `https://huggingface.co/datasets/aplycaebous/BdSLIG` and the code is available on GitHub: `https://github.com/tariquzzamanf/SPIP`. The reproducibility of results on certain proprietary models can vary based on the time and date of model inference.

References

[1] Kshitij Bantupalli and Ying Xie. American sign language recognition using deep learning and computer vision. In *2018 IEEE international conference on big data (big data)*, pages 4896–4899. IEEE, 2018. 1

[2] SM Rayeed, Sidratul Tamzida Tuba, Hasan Mahmud, Mumtahin Habib Ullah Mazumder Md, Saddam Hossain Mukta Md, and Kamrul Hasan Md. Bdsl47: A complete depth-based bangla sign alphabet and digit dataset. *Data in Brief*, 51:109799, 2023. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[5] Javier Selva, Anders Skaarup Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B. Moeslund, and Albert Clapés. Video transformers: A survey. *CoRR*, abs/2201.05991, 2022. 1

[6] Shikhar Sharma and Krishan Kumar. Asl-3dcnn: American sign language recognition technique using 3-d convolutional neural networks. *Multimedia Tools and Applications*, 80(17):26319–26331, 2021. 1

[7] Carman KM Lee, Kam KH Ng, Chun-Hsien Chen, Henry CW Lau, Sui Ying Chung, and Tiffany Tsoi. American sign language recognition and training method with recurrent neural network. *Expert Systems with Applications*, 167:114403, 2021. 1

[8] Gokul NC, Manideep Ladi, Sumit Negi, Prem Selvaraj, Pratyush Kumar, and Mitesh Khapra. Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. *Advances in Neural Information Processing Systems*, 35:36202–36215, 2022. 1

[9] Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Johanna Mesch, and Marc Schulder, editors, *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 54–65, Torino, Italia, May 2024. ELRA and ICCL. 1

[10] Kaveesh Charuka, Sandareka Wickramanayake, Thanuja D Ambegoda, Pasan Madhushan, and Dineth Wijesooriya. Sign language recognition for low resource languages using few shot learning. In *International Conference on Neural Information Processing*, pages 203–214. Springer, 2023. 1

[11] Mert Inan, Katherine Atwell, Anthony Sicilia, Lorna Quandt, and Malihe Alikhani. Generating signed language instructions in large-scale dialogue systems. *arXiv preprint arXiv:2410.14026*, 2024. 1

[12] Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. Signllm: Sign language production large language models. *arXiv preprint arXiv:2405.10718*, 2024. 1

[13] Ceil Lucas and Robert Bayley. The importance of variation research for deaf communities. Technical report, University of Pennsylvania Working Papers in Linguistics, 2001. 1

[14] Hannah Lutzenberger, Katie Mudd, Rose Stamp, and Adam Schembri. The social structure of signing communities and lexical variation: A cross-linguistic comparison of three unrelated sign languages. *Glossa*, 8(1):1–40, 2023. 1

[15] Husne Ara Rubaiyeat, Hasan Mahmud, Ahsan Habib, and Md Kamrul Hasan. Bdslw60: A word-level bangla sign language dataset. *arXiv preprint arXiv:2402.08635*, 2024. 1, 2

[16] Md Johir Raihan, Mainul Islam Labib, Abdullah Al Jaid Jim, Jun Jiat Tiang, Uzzal Biswas, and Abdullah-Al Nahid. Bengali-sign: A machine learning-based bengali sign language interpretation for deaf and non-verbal people. *Sensors*, 24(16):5351, 2024. 1

[17] Abu Saleh Musa Miah, Jungpil Shin, Md Al Mehedi Hasan, and Md Abdur Rahim. Bensignnet: Bengali sign language alphabet recognition using concatenated segmentation and convolutional neural network. *Applied Sciences*, 12(8):3933, 2022. 1

[18] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. 1

[19] Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, 106:102270, 2024. 1

[20] Yoshiyuki Nakata, WL Quint Oga-Baldwin, and Atsuko Tsuda. Student perceptions of feedback and self-regulated language learning: A mixed-methods investigation. *System*, 131:103654, 2025. 1

[21] Prajwal Paudyal, Junghyo Lee, Azamat Kamzin, Mohamad Soudki, Ayan Banerjee, and Kornepati Sandeep. Learn2sign: Explainable ai for sign language learning. 02 2019. 1

[22] Mert Inan, Anthony Sicilia, and Malihe Alikhani. SignAlignLM: Integrating multimodal sign language processing into large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3691–3706, Vienna, Austria, July 2025. Association for Computational Linguistics. 1

[23] Wendy Sandler and Diane Lillo-Martin. *Sign Language and Linguistic Universals*. Cambridge University Press, Cambridge, 2006. 2

[24] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024. 4

[25] Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18471–18480, 2024. 4