Breaking the Silence: A Dataset and Benchmark for Bangla Text-to-Gloss Translation

Anonymous ACL submission

Abstract

Gloss is a written approximation that bridges Sign Language (SL) and its corresponding spoken language. Despite a deaf and hard-of-hearing population of at least 3 million in Bangladesh (Alauddin and Joarder, 2004), Bangla Sign Language (BdSL) remains largely understudied, with no prior work on Bangla text-to-gloss translation and no publicly available datasets. To address this gap, we construct the first Bangla text-to-gloss dataset, consisting of 1,000 manually annotated and 4,000 synthetically generated Bangla sentence–gloss pairs, along with 159 expert human-annotated pairs used as a test set. Our experimental framework performs a comparative analysis between several fine-tuned open-source models and a leading closed-source LLM to evaluate their performance in low-resource BdSL translation. GPT-5.4 achieves the best overall performance, while a fine-tuned mBART model performs competitively despite being approximately 100× smaller. Qwen-3 outperforms all other models in human evaluation. This work introduces the first dataset and trained model for Bangla text-to-gloss translation. It also demonstrates the effectiveness of systematically generated synthetic data for addressing challenges in low-resource sign language translation.

1 Introduction

Sign language has been the primary method of communication for deaf people for centuries (Cantin and Encrevé, 2022). But deaf people often use lip reading, text-based communication, or other media (Barnett, 2002) to interact with others. Thus Sign Language Translation (SLT) improves the quality of life of the deaf people by helping them communicate more easily with non-signers while allowing each party to use their preferred language. In addition, the deaf community has expressed their preference in using sign over other media (Middleton et al., 2010).

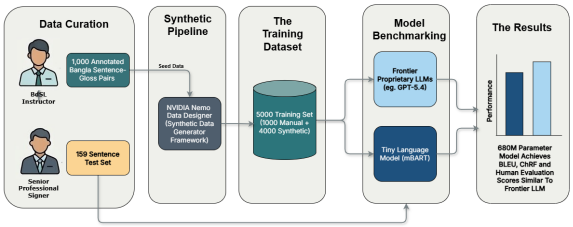

Figure 1: Overview of the Bangla Text-to-Gloss (T2G) Framework

In the context of Sign Language (SL), gloss is considered as the written approximation of SL which uses words as "labels" for each sign along with various grammatical notes. A sign can have multiple meanings depending on the context of the sentence. For example, the English sentence "I am going to the store." is translated into gloss as "I GO STORE". Gloss thus acts as an intermediary between the SL and it's corresponding language.

Bangla, a language spoken by 272.7 million individuals (Sultana et al., 2025), also has a substantial deaf community that uses the Bangla Sign Language daily. According to the government defined disability categories, among the people of Bangladesh, hearing disability is at 0.29% (Bangladesh Bureau of Statistics), although (Alauddin and Joarder, 2004) suggests figures as high as 3 million. However, BdSL is a low-resource language. Consequently, translation efforts suffer from quality of data and under-articulated signing (Babbitt and Mansueto, 2012; Walsh et al., 2024). In our work, we introduce a novel dataset representing Bangla sentences and their gloss form. Research has shown that the use of glosses as intermediaries improves the translation process (Gómez et al., 2021).

To address this gap, we present our contributions in this paper.

- We curated a dataset of 1,000 hand-annotated Bangla sentences and their corresponding

glosses, annotated by a certified BdSL instructor. To ensure a rigorous evaluation, a separate test set of 159 sentences was independently developed by a senior professional BdSL signer.

- We generate 4000 synthetic gloss with The NeMo Data Designer (2025) framework using the manually labeled data as the foundational seed for the generation process

- We benchmark a diverse set of proprietary and open-source models to evaluate their performance on our dataset, and develop a tiny language model with 680M parameters that achieves performance comparable to GPT 5.4.

Our synthetic data generation framework can greatly reduce the cost of annotating huge amounts of gloss data with experts. We show that a small amount of high-quality/annotated data can be augmented systematically by using the data as seed to LLM based synthetic data generation framework.

2 Literature Review

2.1 Text-to-Gloss

Text-to-gloss is a necessary task for sign language translation (SLT). Research has shown that considering gloss form can improve the performance of SLT compared to end-to-end neural networks. Additionally, for tasks like text-to-sign, text-to-gloss translation is a necessary step. Although the Hamburg Notation System (HamNoSys) can provide phonetic level notation (Hanke, 2004), the literature has shown that such an unsupervised approach can improve performance with supervised gloss notations. Despite such a necessity, text-to-gloss is a low-resource domain.

For the text-to-gloss task, Stoll et al. (2018) used an RNN-based encoder-decoder with GRUs. They evaluated their model on the PHOENIX-Weather-2014T dataset and achieved performance comparable to that of Camgoz et al. for the opposite translation direction. This RNN-based text-to-gloss model provides a useful baseline for our text-to-gloss transformer model. Babbit and Mansueto et al. use the POS tag and multilingual dataset to improve the performance of the text-to-gloss model (Babbitt and Mansueto, 2012).

2.2 Bangla text-to-gloss

Moryossef and Goldberg (2021) describes 20 distinct tasks for sign language translation. Among them, BdSL only covers the video (sign) to text task (Zeeon et al., 2024; Sams et al., 2023; Sonare et al., 2021; Alam et al., 2021; Hoque et al., 2016; Harini et al., 2020). Few works have focused on the text-to-sign and text-to-pose tasks (Shahriar et al., 2017; Sarkar et al., 2009; Hoque et al., 2016). Bangla, a morphologically rich language, lacks extensive research in gloss annotation. Although corpora such as the Bangla POS-tagged corpus or Bangla dependency treebanks exist, they focus primarily on syntactic and lexical tagging rather than morpheme-level glossing. There is no prior work on the Bangla text-to-gloss task to the best of our knowledge. In fact, there is a general scarcity of text-to-gloss dataset across languages. Moryossef et al. (2021) summarizes publicly available text-to-gloss datasets in an attempt to demonstrate the small size of corpus (Table-1).

Language Pair	Gloss Text Pairs	Gloss / Spoken
Signum DGS–German (von Agris and Kraiss, 2007)	780	565 / 1,051
NCSLGR ASL–English (SignStream, 2007)	1,875	2,484 / 3,104
RWTH-PHOENIX-Weather-2014T (Camgoz et al., 2018)	8,257	1,870 / 4,839
French SL–French (Limsi, 2019)	2,904	2,266 / 5,028

Table 1: Publicly available SLT corpora with gloss annotations and spoken language translations.

2.3 LLM for synthetic data generation

Incorporating synthetic data with real-world data has been shown to improve model adaptability and contextual understanding, particularly in domain-specific applications. This hybrid approach often outperforms models trained solely on real or synthetic data, as it provides a diverse and enriched training set (Zhezherau and Yanockin, 2024). Recent advances in leveraging large language models (LLMs) such as GPT-4o have demonstrated their efficacy in generating high-quality synthetic data for specialized tasks, including conversational semantic frame analysis. Matta et al. (2024) highlights the cost-efficiency of using LLM-generated data when combined with human-labeled data. The study shows that as budget constraints become more stringent, incorporating synthetic data significantly improves model performance compared to relying solely on human-labeled datasets. This indicates that LLMs can play a pivotal role in resource-limited settings, offering scalable data generation while maintaining relevance to the application domain. Smaller models tend to exhibit biases towards their generated data, while larger models show more reliability (Maheshwari et al., 2024).

3 Methodology

3.1 Field Study

We started our research work by interviewing BdSL teachers and professional signers. We discovered the following findings from the interviews: (i) There is no set standard for BdSL. As such, it varies between different schools and even between individual students as they adopt the teachings for ease of use. (ii) BdSL does not have a strict grammatical structure, rather rule of thumbs. However some common rules exist such as use of infinitive verb in place of detailed verb forms. Despite this, governmental and academic institutions in Bangladesh have made standardization efforts, including the publication of a four-volume guide titled "ইশারায় বাংলা ভাষার সহায়িকা" (Isaray Bangla Bhasar Sahayika) (Guide to Bangla Sign Language). We curated part of our test dataset from this book. We also took the help of a professional Sign Language Interpreter to compile the rest of the test dataset. Every sentence in the test data was manually re-verified by the professional to ensure linguistic accuracy to prioritize quality for this benchmark. To reflect the variations we observed in our field study, we used another annotator to prepare our training dataset. The details about the annotators can be found at C.

3.2 Constructing Bangla T2G dataset

We construct a Bangla text-to-gloss (T2G) dataset which, to the best of our knowledge, is the first for Bangla language. We consider two data sources for Bangla T2G - human annotated data and synthetic data generated using LLM.

Curating hand annotated data

As noted in our field study, we found that BdSL users differ in their use of signs. As such, our annotators worked independently to curate the training and the test dataset. All contributors provided informed consent and were briefed on the study's objectives. To create the training data, we provided the annotator with a 1000 sentences that comprises of different aspects of everyday life, such as Family life, Transportation, News discussions, School, Weather, Life lessons, etc in Bangla. We asked the annotator to translate the sentences into their gloss form. We used this manually created data as ground truth to generate the synthetic data.

Dataset	Train	Test
Human Annotated	1,000	159
Synthetic	4,000	–

Table 2: Combined datasets across gloss experiments.

Synthetic data generation using LLM

LLM has been widely used for generating synthetic data in low-resource domain and has shown sufficient reliability (Zehady et al., 2024; del Barrio et al., 2024). We take a sample of 100 sentences from the manually annotated dataset to generate gloss form for 100 Bangla sentences with The NeMo Data Designer (2025) framework and `GPT-4o` as the underlying LLM. In our prompt (provided in Appendix B), we mention the rule of thumbs of BdSL, instruct the LLM to follow the patterns seen in the seed dataset and apply it all together to generate a new sentence-gloss pair for a new topic. We then evaluate these 100 text-to-gloss translations with our human annotator who created the training dataset who marked **74.75%** of sentences as correct. Subsequently, we generate gloss representations for an additional 4000 Bangla sentences using the same framework.

We combine the above two data sources to construct our final dataset (full details in Table-2), creating *bangla-gloss*.

3.3 Closed-source Baselines

`GPT-5.4` represents the state-of-the-art in generative LLM. It was incorporated into this study to serve as a state-of-the-art baseline for few-shot translation capabilities.

3.4 Trained Models for T2G Translation

We fine-tune four open-source models on our proposed training set: (i) **MarianMT** (Tiedemann and Nygaard, 2004), a dedicated NMT encoder-decoder pre-trained on OPUS; (ii) **mBART-50** (Liu et al., 2020), a multilingual seq2seq model whose bidirectional encoder and autoregressive decoder are well-suited for low-resource translation; (iii) **Qwen-3** (Yang et al., 2025), selected for its strong multilingual performance; and (iv) **Phi-4** (Abdin et al., 2024), included to assess whether a compact general-purpose LM can compete with dedicated translation models. Additional implementation details are provided in Appendix A.

3.5 Evaluation Scheme

Following Babbitt and Mansueto (2012); Moryossef et al. (2021), we use three metrics -

Model	Type	Parameters	BLEU	ChrF	Human Eval.
GPT-5.4	Causal LM	–	**39.26**	**73.75**	67.76%
Qwen-3	Causal LM	8B	30.82	70.16	**72.27**%
Phi-4†	Causal LM	4B	2.26	26.95	0%
mBart-large	Seq2Seq	0.68B	34.81	70.81	63.76%
MarianMT†	Seq2Seq	0.14B	7.11	44.60	2.1%

† Frequently produces misspelled tokens and incoherent gloss sequences, resulting in near-zero human acceptability.

Table 3: Automatic (BLEU-4, ChrF) and human evaluation scores on the *bangla-gloss* test set (159 sentences). Proprietary models are prompted few-shot; open-source models are fine-tuned on our training data. Human evaluation reflects acceptability as judged by a certified BdSL signer (Appendix C).

BLEU-4 (Papineni et al., 2002), chrF (Popović, 2015) and Human Evaluation to evaluate the text-to-gloss translation models. These metrics provide insight into the surface-level similarity of generated outputs with reference glosses, with higher-order BLEU scores emphasizing longer, contiguous n-gram matches. We do not use METEOR as their authors state that Bangla is not a supported language[1]. Although COMET supports Bangla, we do not use it because it hasn't seen BdSL in it's training data, making it unsuitable for BdSL evaluation[2].

4 Result & Analysis

The performance of all models on the *bangla-gloss* benchmark is summarized in Table 3. We organize our analysis around three key observations.

Automatic Metrics. GPT-5.4 achieves the highest BLEU (39.26) and ChrF (73.75), establishing an upper bound among the models evaluated. Among open-source models, fine-tuned mBART attains the second-best BLEU (34.81) and ChrF (70.81) despite being roughly $\sim$100$\times$ smaller, demonstrating the effectiveness of domain-specific fine-tuning of pretrained multilingual seq2seq models for low-resource translation. Qwen-3 follows closely with a BLEU of 30.82 and ChrF of 70.16. MarianMT and Phi-4, by contrast, yield substantially lower scores, suggesting that neither a small dedicated NMT model nor a general-purpose small LM can adequately capture BdSL gloss structure under our training regime.

Human Evaluation. Human judgments reveal a notably different ranking. Qwen-3 receives the highest human evaluation score (72.27%), surpassing GPT-5.4 (67.76%) and mBART (63.76%). This suggests that Qwen-3 produces glosses that are semantically adequate to native signers even when their surface forms diverge from the reference. MarianMT and Phi-4 receive near-zero human scores (2.1% and 0%, respectively), as both models frequently generate misspelled tokens and incoherent output.

Metric Discordance. The gap between automatic and human scores—most notably Qwen-3 ranking first in human evaluation but third in BLEU—underscores a known limitation of reference-based metrics in gloss translation: because BdSL lacks a single standardized gloss convention, valid translations may differ lexically from the reference while remaining acceptable to native signers. This finding aligns with prior observations on metric reliability in low-resource SLT (Moryossef et al., 2021) and motivates future work on reference-free evaluation for BdSL.

5 Conclusion & Future Work

In this work, we fine-tuned mBart, Qwen, Phi and MarianMT models for the BdSL text-to-gloss task. The mBart yielded a BLEU score of 34.81 and a ChRF score of 70.81 while Qwen-3 acieved a human evaluation score of 72.27%. Notably, their performances are comparable to GPT-5.4, validating the effectiveness of domain-specific adaptation for low-resource sign language translation. Additionally, we also observe that synthetic data generated using Bangla gloss generation rules can indeed effectively improve text-to-gloss performance. This finding supports the conclusion proposed in Moryossef et al. (2021). In future works, we aim to extend the dataset by incorporating regional variations. Additionally, we want to incorporate the insights from the linguists into the Bangla gloss generation rule of thumbs to generate better glosses.

[1] https://www.cs.cmu.edu/~alavie/METEOR/
[2] github.com/Unbabel/COMET

Limitations

Although our dataset was annotated by people who knows Bangla, it is possible that deaf people outside of the Bangladesh might follow different patterns of BdSL which are not represented in our work. Our dataset primarily consists of hand-annotated examples; however, the volume of such annotations remains limited. This constraint affects the coverage and diversity of linguistic constructions, particularly for complex or less frequently used sentence structures. Bangla exhibits unique syntactic and morphological characteristics that are reflected in Bangla Sign Language (BdSL). As such, a set of glossing rules tailored specifically to BdSL would likely produce more accurate and linguistically faithful representations.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Md Shahinur Alam, Mahib Tanvir, Dip Kumar Saha, and Sajal K Das. 2021. Two dimensional convolutional neural network approach for real-time bangla sign language characters recognition and translation. *SN Computer Science*, 2:1–13.

Mohammad Alauddin and Abul Hasnat Joarder. 2004. Deafness in bangladesh. In *Hearing Impairment: An Invisible Disability How You Can Live With a Hearing Impairment*, pages 64–69. Springer.

Luke Babbitt and Jenna Mansueto. 2012. Text2gloss: Translation into sign language gloss with transformers. In *Proceedings of the Stanford CS224N Custom Project*.

Bangladesh Bureau of Statistics. National survey on persons with disabilities (nspd) 2021. Accessed: 25-6-2025.

Steven Barnett. 2002. Communication with deaf and hard-of-hearing people: a guide for medical education. *Academic Medicine*, 77(7):694–700.

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

Yann Cantin and Florence Encrevé. 2022. Perspectives: On the historicalness of sign languages. *Frontiers in Communication*, 7:801862.

David Alonso del Barrio, Max Tiel, and Daniel Gatica-Perez. 2024. Human interest or conflict? leveraging llms for automated framing analysis in tv shows.

Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th workshop on building and using comparable corpora (BUCC 2021)*, pages 18–27.

Thomas Hanke. 2004. Hamnosys–representing sign language data in language resources and language processing contexts. In *sign-lang@ LREC 2004*, pages 1–6. European Language Resources Association (ELRA).

R. Harini, R. Janani, S. Keerthana, S. Madhubala, and S. Venkatasubramanian. 2020. Sign language translation. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 883–886.

Md. Tazimul Hoque, Md. Rifat-Ut-Tauwab, Md. Fasihul Kabir, Farhana Sarker, Mohammad Nurul Huda, and Khandaker Abdullah-Al-Mamun. 2016. Automated bangla sign language translation system: Prospects, limitations and applications. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 856–862.

Limsi. 2019. Dicta-sign-lsf-v2. ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.

Yuxian Liu, Shijie Yang, Junyang Li, Xu Du, Haitao Yang, and Xipeng Liu. 2020. Multilingual denoising pre-training for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 3008–3018. Association for Computational Linguistics.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *arXiv preprint arXiv:2409.11968*.

Shiho Matta, Yin Jou Huang, Fei Cheng, Hirokazu Kiyomaru, and Yugo Murawaki. 2024. Investigating cost-efficiency of llm-generated training data for conversational semantic frame analysis. *arXiv preprint arXiv:2410.06550*.

Anna Middleton, Graham H Turner, Maria Bitner-Glindzicz, Peter Lewis, Martin Richards, Angus Clarke, and Dafydd Stephens. 2010. Preferences for communication in clinic from deaf people: A cross-sectional study. *Journal of Evaluation in Clinical Practice*, 16(4):811–817.

Amit Moryossef and Yoav Goldberg. 2021. Sign Language Processing. https://sign-language-processing.github.io/.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.

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*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Ataher Sams, Ahsan Habib Akash, and S. M. Mahbubur Rahman. 2023. Signbd-word: Video-based bangla word-level sign language and pose translation. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Biswajit Sarkar, Kaushik Datta, CD Datta, Debranjan Sarkar, Shashanka J Dutta, Indranil Das Roy, Amalesh Paul, Joshim Uddin Molla, and Anirban Paul. 2009. A translator for bangla text to sign language. In *2009 Annual IEEE India Conference*, pages 1–4. IEEE.

Rhythm Shahriar, AGM Zaman, Tanvir Ahmed, Saqib Mahtab Khan, and HM Maruf. 2017. A communication platform between bangla and sign language. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 1–4. IEEE.

NCSLGR SignStream. 2007. *Volumes 2–7*.

Babita Sonare, Aditya Padgal, Yash Gaikwad, and Aniket Patil. 2021. Video-based sign language translation system using machine learning. In *2021 2nd International Conference for Emerging Technology (INCET)*, pages 1–4.

Stefan Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2018. Text2gloss: Sequence learning for text representation in sign language translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Nusrat Sultana, Rumana Yasmin, Bijon Mallik, and Mohammad Shorif Uddin. 2025. Onubad: A comprehensive dataset for automated conversion of bangla regional dialects into standard bengali dialect. *Data in Brief*.

NVIDIA The NeMo Data Designer. 2025. Nemo data designer: A framework for generating synthetic data from scratch or based on your own seed data. `https://github.com/NVIDIA-NeMo/DataDesigner`. GitHub Repository.

Jörg Tiedemann and Lars Nygaard. 2004. The opus corpus-parallel and free: http://logos. uio. no/opus. In *LREC*.

Ulrich von Agris and K. Kraiss. 2007. Towards a video corpus for signer-independent continuous sign language recognition. In *Gesture in Human-Computer Interaction and Simulation*.

Harry Walsh, Abolfazl Ravanshad, Mariam Rahmani, and Richard Bowden. 2024. A data-driven representation for sign language production. In *Proceedings of the 18th International Conference on Automatic Face and Gesture Recognition (FG 2024)*. Institute of Electrical and Electronics Engineers (IEEE).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Iftekhar E Mahbub Zeeon, Mir Mahathir Mohammad, and Muhammad Abdullah Adnan. 2024. Btvsl: A novel sentence-level annotated dataset for bangla sign language translation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE.

Abdullah Khan Zehady, Safi Al Mamun, Naymul Islam, and Santu Karmaker. 2024. Bongllama: Llama for bangla language. *arXiv preprint arXiv:2410.21200*.

Alexey Zhezherau and Alexei Yanockin. 2024. Hybrid training approaches for llms: Leveraging real and synthetic data to enhance model performance in domain-specific applications. *arXiv preprint arXiv:2410.09168*.

A Training Details

All models were fine-tuned on the *bangla-gloss* training set with a learning rate of 2×10^{-5}. Model-specific details are provided below.

- **MarianMT:** A transformer encoder-decoder architecture trained on OPUS, a multilingual corpus of openly available documents (Tiedemann and Nygaard, 2004). It serves as a dedicated NMT baseline, in contrast to the general-purpose LMs listed below.
- **mBART-50:** We select mBART-50 for its bidirectional Transformer architecture. Its 12-layer encoder and 12-layer decoder (with 1024-dimensional hidden states) are explicitly optimized for sequence-to-sequence reconstruction, making it superior to models that rely on separately initialized decoders such as BERT (Liu et al., 2020).
- **Qwen-3:** Selected for its strong multilingual performance, particularly on Indo-European languages (Yang et al., 2025). The model was fine-tuned with a maximum sequence length of 2048.
- **Phi-4:** A compact general-purpose language model included to test whether small LMs can match dedicated translation models or larger LLMs on the T2G task.

B Prompt for BdSL Synthetic Gloss Generation

Prompt: Sentence-to-Gloss Mapping

You are a linguistically careful assistant trained to generate Bangla Sign Language (BdSL) gloss approximations from Bangla sentences. You are given a seed dataset with two columns:

{{Text}} translates to {{Gloss}}

Your task is to learn the mapping pattern between Sentence -> Gloss from the examples and then generate ONE new, high-quality gloss for ONE Bangla sentence for the topic {{Topics}}.

Strict instructions:
1. Use Bangla words, written in Bangla script.
2. Use the correct infinitive form of verbs wherever applicable. Example: “কাজ করে” -> “কাজ করা”, “গাইতে” -> “গাওয়া”. Do not invent verb forms not seen or implied in the seed data.
3. Follow BdSL-preferred ordering as observed in the seed Gloss column.
4. Do not paraphrase or add information.
5. Do not remove core meaning.
6. If a sentence is ambiguous, choose the most neutral, literal gloss.
7. No numbering, no quotes, no extra whitespace.
8. If the input sentence contains: multiple clauses, temporal expressions, conditionals, negation, modality, or questions, the gloss MUST preserve that complexity.
9. Final output MUST strictly follow this format:

BanglaSentence -> BanglaGloss

Constraints:
- Exactly ONE line.
- Exactly ONE ”->”.
- Left side MUST be the input sentence verbatim.
- Right side MUST be the generated BdSL gloss.
- A simpler gloss than the input sentence is INVALID.
- Any deviation is an error.

C Annotator Details

The training dataset was annotated by a graduate student from a university’s education research institute with professional experience at a government school for the deaf. The test set was developed in collaboration with a professional sign language interpreter currently active in national broadcast media. While the annotators are hearing professionals, we mitigated ‘hearing-centric’ bias by validating our linguistic framework and findings through structured interviews with veteran BdSL educators. The primary annotator was recruited via institutional networks and was compensated at a rate exceeding local professional standards for specialized linguistic tasks, reflecting the high workload of training data preparation. The test set validator contributed their time on a voluntary basis as a professional consultant. All participants provided written informed consent. They were briefed on the project’s objectives, the data’s intended use for academic research, and the anonymized nature of the eventual public release. Annotators were instructed to use the natural flow of BdSL. No personally identifiable information is included in the dataset. This study was conducted under the supervision of the authors’ institutional department. In the absence of a formal Institutional Review Board (IRB) for non-medical linguistic research, the protocol followed the university’s internal research code of conduct to ensure the protection of all participants.