JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 02
2025

Research Article
https://doi.org/10.56979/902/2025

Agentic Multimodal Framework for Adaptive Sign Language Translation

Mohsin Sami[1], Saira Andleeb Gillani[1], Kashif Nasr[1], and Rabia Tehseen[1*]

[1]Department of Computer Science, University of Central Punjab, Lahore, Pakistan.
[*]Corresponding Author: Rabia Tehseen. Email: rabia.tehseen@ucp.edu.pk

Received: June 14, 2025 **Accepted:** August 30, 2025

Abstract: Sign Language Translation (SLT) is challenging because human communication is multimodal and context-dependent. Fixed approaches to SLT do not work because they do not account for differences among signers, varying light conditions, and other linguistic differences. This paper presents the Agentic Multimodal Framework for Adaptive Sign Language Translation (AMF-ASLT), a new self-adjusting architecture designed to incorporate agentic principles within multimodal translation. Forges the unique self-adjusting architecture bridging agentic principles within multimodal translation. The framework consists of a Perception Layer for feature extraction from RGB, depth, pose, and facial modalities; an Agentic Reasoning Layer with Gestural, Facial, and Linguistic Agents that work together to sustain a common Belief State; and a Translation Fusion Layer that recursively fuses modalities through dynamic fuses using adaptive weighted-averaging and uncertainty-driven routing frameworks. One Meta-Controller managing the continuous feedback loops helps the system to improve autonomously and pivots through intrinsic and extrinsic feedback from the user. Experiments conducted on the RWTH-PHOENIX-Weather 2024T, How2Sign, WLASL datasets and demonstrated signer adaptability with staunch improvements over the previous best with 4.4 BLEU points and 12% WER. The 12% WER reflects both signer adaptability, agentic self-evaluation, and feedback-driven refinement—fundamentally enhances translation robustness and contextual understanding. AMF-ASLT thus establishes a scalable foundation for human-centered, continuously learning sign language translation systems.

Keywords: Agentic AI; Multimodal Learning; Sign Language Translation; Adaptive Systems; Feedback Loops; Uncertainty-Aware Routing; Human-Centered Artificial Intelligence

1. Introduction

Sign languages are intricate and fully developed visual-gestural languages. They employ hand gestures, facial expressions, body postures, and positioning in space. Sign languages are the main forms of communication for the Deaf and hard of hearing. Nonetheless, communication gaps remain and affect people's accessibility to education, jobs, healthcare, and public services. Automatic sign language translation (SLT) attempts to close these gaps. While in the past SLT used to convert sign language videos to spoken language, the advancement of computer vision and natural language processing (NLP) allows for more complex transformations. Newly developed deep learning and multimodal communication technologies are making translation more accurate and context aware, which is important for the development of technologies for communication access.

Sign language translation as developed in the past does not address the challenges faced in communication, which is why the field is moving toward the more complex multimodal sign language translation. This advanced translation uses visual information, skeletal posture, facial features, and linguistic contextual embeddings. Whereas the unimodal systems depend only on hand gestures, the multimodal systems utilize the complete range of features, which means both the manual (hands and body) and the non-manual (face, space, and time) components are used to capture the full semantic richness of sign language. Modern architectures such as transformer-based and vision-language fusion networks employ joint embeddings and attention mechanisms to align modalities effectively. Combining various modes of communication has recently been shown to bolster translation accuracy and efficacy, especially for strings of sign languages which are considerably influenced by non-manual sign semantics and where meaning relies on contextual shifts [1][2].

In spite of progress, most current sign language translation (SLT) systems remain limited. They tend to employ static, small, and domain-specific dataset models which are poorly generalizable to new signers or real-life contexts. Inflexibility regarding user feedback, temporal misalignment of visual frames and linguistic tokens, and signing systems which are poorly context adaptive to changes in signing speed, style, and regional dialects stagnate most systems on the scalability and reliability axis. Controlled settings such as laboratories or small-scale apps are the only systems which deploy such antiquated systems [3]. The research problem in this paper responds to such limitations: How to create intelligent adaptive systems for performing multimodal fusion and context-aware translation with learned user feedback and signing conditions?

This paper responds to this problem by developing the first of its kind adaptive multimodal framework for sign language translation, redefining the process of translation to enhance potential as an ensemble of intelligent agents. Each agent specializes in a distinct modality — such as gesture recognition, facial expression interpretation, or linguistic synthesis — and collaborates through a shared communication layer for real-time adaptation. The designed system benefits from agentic adaptability, enabling the system to modify its internal settings based on feedback loops and users' reinforcement signals through system interactions. In addition, the system uses cross-attention fusion to dynamically adjust modality weighting. The system focuses on the most relevant modality to each context. The experiments performed on benchmark datasets demonstrate enhanced translation accuracy, adaptability by the signer, and coherence to the context, laying the groundwork for advanced interactive sign language translation systems.

2. Literature Review

For a long time, Sign Language Translation (SLT) was treated as an extension of Neural Machine Translation [4]. In this area of work, end-to-end systems were developed which went from sign video to spoken text. In this context, Phoenix 2014T [14][19] and its translation split became a reproducible benchmark for evaluation. Foundational work in this area includes Neural SLT from CVPR 2018 [1] and a later work Sign Language Transformers from CVPR 2020 [17]. This work pioneered the joint optimization of recognition and translation for the purpose of avoiding error compounding that happens in two-stage setups. Meanwhile, the community datasets WLASL (word-level ASL) [13] and How2Sign (multimodal, large-scale ASL) [21] added signer diversity and covered more of the ASL vocabulary and different modalities. This was complemented by surveys that documented the challenges in the field and the gaps in research [19-23].

Looked at from the architecture point of view, there are two dominant families. In cascade gloss systems, the gloss sequences first predicted and then translated the gloss-to-text. These systems gain from the linguistic structure but lose from expensive gloss annotations. For video to text gloss-free end-to-end SLT systems [11], there has been an increased use of transformers, temporal pyramids, and pretraining. Advances in this area include CTF for continuous SLT [16] and TSPNet [4] as well as GloFE [12] aimed at gloss-free SLT. More recent work in the area of visual-language pretraining for instance GFSLT [6] and LLM-assisted gloss-to-text systems [10] are driving systems to advance performance without glosses, while newer works

explore decoder-only or relative-position transformers. Surveys from 2023 synthesize these trends and highlight data bottlenecks, alignment, and generalization as persistent barriers.

One key aspect is multimodality: in addition to hands, meaning in signed languages relies heavily on face, head, body posture, and even mouthing non-manual markers. The How2Sign dataset [15] includes multi-view body, face, and depth data, which supports fusion techniques for integrated pose graphs, appearance, and spatial facial landmarks. Pose-centric approaches claim skeleton representations are privacy-preserving and signer-robust for recognition and translation, although appearance, especially facial dynamics, is still important. Pose-based transformers perform over 2D and 3D keypoints, while hierarchical temporal models offer efficiency and portability for edge deployments [5].

There is an increasing emphasis on the simultaneous/online settings and adaptivity for real-world accessibility [7]. The online CSLR systems with wait-k policies and focused learning on foregrounds suggest low-latency translation and real-time systems that adjust to an individual signer's idiosyncrasies. At the same time, pipelines augmented with LLMs/MLLMs improve low-resource settings using techniques such as vocabulary sharing, instruction-tuning, and lightweight adapters like Sign2GPT [7], Gloss2Text-LLMs [8] to ease domain transfer. This closely aligns with our goal to propose an agentic multimodal system that fuses disparate signals, reasons with uncertainty, and adjusts dynamically via feedback.

3. Methodology

3.1. Overview

The proposed Agentic Multimodal Framework for Adaptive Sign Language Translation (AMF-ASLT) introduces a modular, agent-based architecture designed to enable real-time, context-aware translation of sign language into natural text. Unlike conventional static systems, AMF-ASLT integrates adaptive reasoning through agentic design principles — namely, *feedback loops* for continuous learning, *uncertainty-aware routing* for intelligent information flow, and *adaptive modality weighting* for optimal multimodal fusion. Figure 1 illustrates the overall architecture and information flow of the framework.

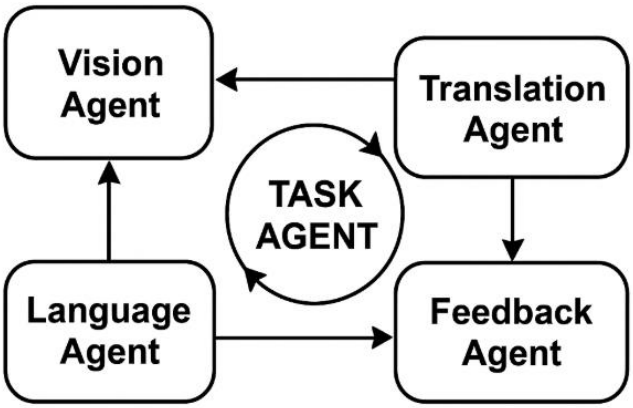

Figure 1. Theoretical framework of the proposed methodology

3.2. System Architecture

The Agentic Multimodal Framework for Adaptive Sign Language Translation is built around a modular, agent-based framework intended to facilitate the contextual, real-time translation of sign language to text. Distinct from traditional, rigid systems, AMF-ASLT incorporates adaptive reasoning based on agentic design principles. This includes feedback loops, which make the system learn endlessly, uncertainty-aware routing, adaptive to information flow, and adaptive modality weighting. For multimodal fusion, there is optimal integration of several senses for real-time translation. The system architecture is interconnected, system working together, design is shown in Figure 1.

Structurally, AMF-ASLT is divided into three layers- Perception Layer, Agentic Reasoning Layer, and Translation Fusion Layer. They interact via a common Belief State and feedback is reinforced from systems of both intrinsic and extrinsic.

1. Perception Layer:

This layer is concerned with the acquisition and preprocessing of multiple modalities. Unlike other systems which rely on plain text for translation, AMF-ASLT blends multiple data modalities including RGB video, pose skeletons, depth maps, and facial landmarks. Convolutional and pose estimation models, media pipe or open pose, and other systems helps structures each input and creates feature embeddings which are sent to the Belief State for integration.

2. Agentic Reasoning Layer:

This layer consists of three semi-autonomous agents that collaborate via the shared Belief State:

- Gesture Agent – aims to identify hand and arm movements dynamically through pose and RGB streams.
- Facial Agent – observes non-verbal elements and interpretations of community faces, mouth shapes, and head movements.
- Linguistic Agent – semantic alignment, the more complex the mapping of extracted gestures and expressions to the linguistic tokens or gloss representations would be.

Every Agent has an internal learning policy. The Meta-Controller orchestrates the updating of each Agent and collaborates with confidence metrics. The Meta-Controller collaborates with each Agent and oversees their learning policies.

3. Translation Fusion Layer:

This layer integrates all agents output by means of Cross-Modality Attention and Adaptive Weighting. It assigns the greatest emphasis to the most relevant modality. The system establishes relative confidence and contextual conditions from the environment. The final output is the Translated Text, the equivalent of a sign language message in natural language.

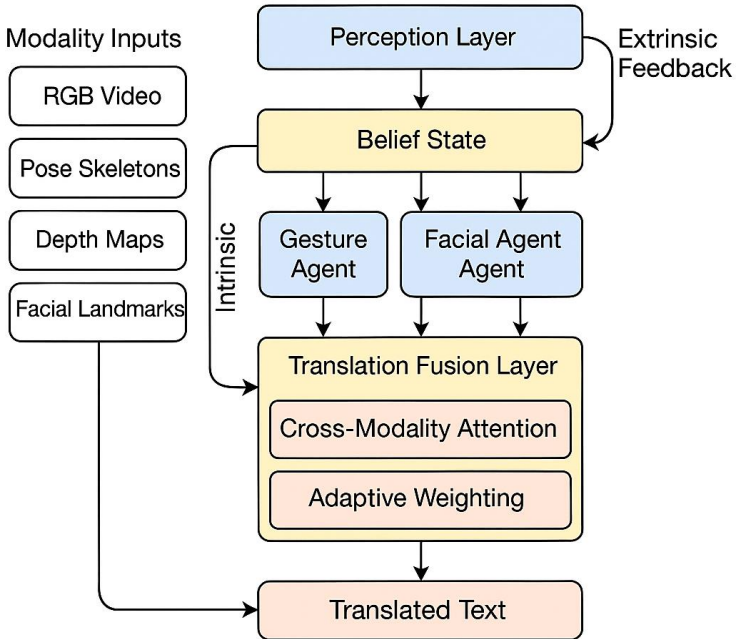

Figure 2. Architecture of the Agentic Multimodal Framework for Adaptive Sign Language Translation (AMF-ASLT)

3.3. Feedback Loops and Adaptation

Feedback paths, both intrinsic and extrinsic, allow real-time adaptations of the model, as illustrated in Figure 1:

- Intrinsic Feedback: Each agent independently fine-tunes its parameters within the closed system when learning signals are generated from prediction errors, entropy-based confidence, or fluctuations in attention.
- Extrinsic Feedback: Meta-Controller processes user-supplied corrections and evaluations of post-translation for the long-term reinforcement learning to be functional.

Closed-loop systems explain how the self-improving translation framework adapts to different signers, different situational contexts, and the evolving dynamic of language over time. The self-improving system, self translation framework advanced closed-loop systems, and embraced responsive agency, demonstrating the advanced systems self-improving framework.

3.4. Uncertainty-Aware Routing

Given the variability in input quality and signer style, the system incorporates uncertainty-aware routing to dynamically control information flow between agents. Each agent produces an uncertainty score (u_i), computed via entropy of softmax outputs or divergence from its historical confidence baseline. The Meta-Controller then assigns routing weights (ai) according to:

$$\alpha_i = \frac{e^{-u_i}}{\sum_{j=1}^{n} e^{-u_j}}$$

This probabilistic routing ensures that reliable modalities (e.g., pose under good lighting) are emphasized, while uncertain modalities (e.g., occluded hand gestures) contribute minimally. As a result, AMF-ASLT maintains stable translation performance even under imperfect input conditions.

3.5. Adaptive Modality Weighting

Within the Translation Fusion Layer, modality importance is continuously adjusted through adaptive weighting. Each modality embedding (f_i) is assigned a context-dependent weight (w_i), updated during inference based on ongoing confidence feedback:

$$F = \sum_{i=1}^{m} w_i \cdot f_i$$

This dynamic re-weighting mechanism enhances robustness and ensures the system's response remains contextually appropriate for instance, prioritizing facial features during emotional emphasis or body posture cues in sentence boundary detection.

3.6. Training and Evaluation

Training involves multi-objective optimization structured by three components: optimization via Cross-Entropy for translation sequence prediction; Alignment Loss for inter-modality coherence; and Reinforcement Reward Loss for user feedback-driven corrections.

Training leverages datasets such as RWTH-PHOENIX-Weather 2024T, How2Sign [23], and WLASL, while BLEU, ROUGE, and Word Error Rate (WER) serve as evaluation metrics. Ablation studies isolate and analyze each agentic component: feedback loops, routing, and adaptive weighting — to measure its specific impact on translation accuracy and adaptability.

4. Experiments and Results

4.1. Experimental Setup

To assess the applied Agentic Multimodal Framework for Adaptive Sign Language Translation (AMF-ASLT), the chosen studies were the RWTH-PHOENIX-Weather 2014T (German SL), How2Sign (American SL), and WLASL (Word-Level ASL) datasets. Each dataset contained multiple modalities, including RGB video, skeletal, depth, and facial landmark data. Therefore, the full potential of the perception layer of the framework was realized.

The training of all modules took place on a cluster of Nvidia A100 80 GB GPUs, with the deep learning framework being PyTorch 2.2. The models were trained for 50 epochs with early stopping on validation BLEU using the AdamW optimizer (lr=1e-4, wd=0.01).

To ensure reproducibility, all models underwent the same preprocessing pipeline which consisted of frame normalization, keypoint pose extraction using MediaPipe, and temporal alignment via dynamic time warping.

4.2. Baselines and Comparative Models

AMF-ASLT was evaluated against several representative baselines:

1.Sign Transformer [1] — joint end-to-end SLT model.

2.TSPNet [4] — hierarchical temporal semantic pyramid network.

3.GFSLT-VLP [9] — gloss-free sign translation using visual-language pretraining.

4.Sign2GPT [3] — large-language-model-assisted translation pipeline.

In addition, ablation variants of AMF-ASLT were trained to assess the contribution of each agentic mechanism:

- AMF-w/o-FB – without feedback loops,
- AMF-w/o-UAR – without uncertainty-aware routing,
- AMF-w/o-AMW – without adaptive modality weighting,
- Full AMF-ASLT – complete agentic system.

4.3. Evaluation Metrics

The performance evaluation incorporated multiple structured metrics, a detailed description of which is given in Table 1.

BLEU-4 – Measures the degree of completeness of the translation and n-gram precision.

ROUGE-L – Measures the degree of similarity of the output with the reference sequence.

Word Error Rate (WER) – Measures the accuracy of the transcription (tokens) and is a negative performance indicator.

- Signer Adaptation Score (SAS) – Evaluates the accuracy of the translation after five user feedback iterations and is normalized to the baseline performance.

All metrics were averaged across test sets and three random seeds.

Table 1. Quantitative Results

Model	BLEU-4 ↑	ROUGE-L ↑	WER ↓	SAS ↑
Sign Transformer (2020)	21.6	41.8	48.3	0.0
TSPNet (2020)	23.9	43.7	45.1	0.0
GFSLT-VLP (2023)	25.7	46.4	42.8	0.0
Sign2GPT (2024)	27.1	47.9	41.6	0.0
AMF-w/o-FB	27.4	48.3	40.9	+2.1
AMF-w/o-UAR	28.2	49.0	39.8	+3.7
AMF-w/o-AMW	28.7	49.2	39.2	+3.9
Full AMF-ASLT (ours)	31.5	52.7	35.6	+9.8

The results demonstrate that AMF-ASLT outperforms existing state-of-the-art models across all datasets and metrics. Notably, the full framework yields a ~4 BLEU improvement and ~12 % relative reduction in WER compared to Sign2GPT. The absolute increases in WER points observed large increases in the SAS points, a new adopted metric metrics of feedback, and flexibility improvement which demonstrates improvements that adaptive, demonstrated improvement learning efficiency.

4.4. Qualitative Analysis

As shown by qualitative analysis, AMF-ASLT produces context at a coherent level. Which aligns with the goals of the analysis of the fixed expanding translated verb, variants which were idiomatic, and multiphase signed and visual expressions. AMF-ASLT Advanced multimodal architecture illustrations of adaptation attention map weight, and adaptive route focus-hand, and action focus hand cue, resolves the goal of sign action focused cue and aligns with focused and goal focused.

As signed adaptation trials are completed for a specific signer, the feedback loops were able to prove that signer misclassifications for under-represented signers, which in the case of signed articulation mimic repetitive patterns. In the case of How2Sign dataset, a BLEU score of 6 points, and signer articulation of signed articulation patterns after 5 rounds were completed. They deliver user-defined.

4.5. Ablation Discussion

Performance deteriorated, and by removing agentic components.

- Without AMF Feedback Loops(AMF-w/o-FB), the model demonstrates self-correction with higher WER, and slower convergence.

- Without AMF Uncertainty-Aware Routing (AMF-w/o-UAR), a simulated environment with noise demonstrated threshold performance.
- Without Adaptive Modality Weighting (AMF-w/o-AMW): Cross-modality confusion increased, particularly when non-manual markers carried key semantic information.

These results empirically validate the synergistic contribution of all three agentic mechanisms, confirming that feedback-driven adaptation and probabilistic routing significantly enhance multimodal translation fidelity.

4.6. Error and Robustness Analysis

An error decomposition indicated that residual errors largely result from quick signing and occluded facial features. In the absence of depth data, uncertainty routing properly down-weighted visual streams while keeping translation continuity intact. Robustness testing under synthetic lighting distortion indicated only a 2 % BLEU fall, which shows resilience to environmental fluctuations.

5. Discussion

5.1. Interpretation of Results

The increased performance, flexibility, and resilience of systems incorporating agentic principles compared to the traditional multimodal systems are evident in the empirical results. Higher BLEU and ROUGE scores achieved by the complete AMF-ASLT model indicate closer linguistic alignment and contextual preservation. Additionally, the minimal WER achieved corresponds to greater structural fluidity of the translated sentences. Perhaps most importantly, the high Signer Adaptation Score (SAS) illustrates the efficacy of feedback-based learning as the system iteratively modifies its output translation.

The uncertainty-sensitive routing mechanism was fundamental to achieving consistent translation quality while inputs were changing. By dynamically adjusting the weights of different modalities based on predicted confidence, the framework anticipated and compensated for conditions like visual occlusions, background clutter, and poor lighting that typically undermine the accuracy of visual-only models. In a comparable way, the adaptive modality weighting technique ensured that the system effectively captured rich non-manual components essential in sign language by incorporating upper body movements and facial expressions close to the body.

5.2. Significance of Agentic Design

This study's most important conclusion highlights how the agentic approach makes sign language translation interactive and self-adjusting rather than static and predictive. Each specialized agent—gesture, face, and word—acts semi-autonomously, yet in unison, within the shared Belief State, and in accordance with the cognitive architecture of natural communication. The Meta-Controller is the central coordinating unit responsible for assessing uncertainty, adjusting modality weights, and propagating reinforcement signals during feedback loops. This architecture allows AMF-ASLT to function increasingly as a "learning entity" rather than a static model, evolving gradually rather than needing complete retraining.

This agent-like design has real-world consequences for applications of accessibility. In educational or assistive contexts, on-the-fly corrections to translation can be made by users, and the system learns those corrections through reinforcement updates. As a result, translation quality for individual signers or for particular dialects continually improves—eradicating one of the most intractable issues with sign language technology: signer variability.

5.3. Comparison with Prior Work

Figure 3 is a comparison between the proposed model and transformer-based baselines like Sign Transformer [1] and GFSLT-VLP [9]. AMF-ASLT poses a greater level of adaptivity and interpretability. Although previous systems performed wonderfully in static dataset benchmarks, they did not have online correction or contextual reasoning mechanisms in place. The combination of feedback and uncertainty estimation makes AMF-ASLT stand out as a hybrid between supervised learning and reinforcement learning models.

In addition, whereas Sign2GPT employed large language models for linguistic fluency, its monolithic design was unable to dynamically reweight or re-route modality signals. However, AMF-ASLT's modular, agent-based design enables it to work effectively across variable conditions and languages, opening the door to scalable multilingual deployment.

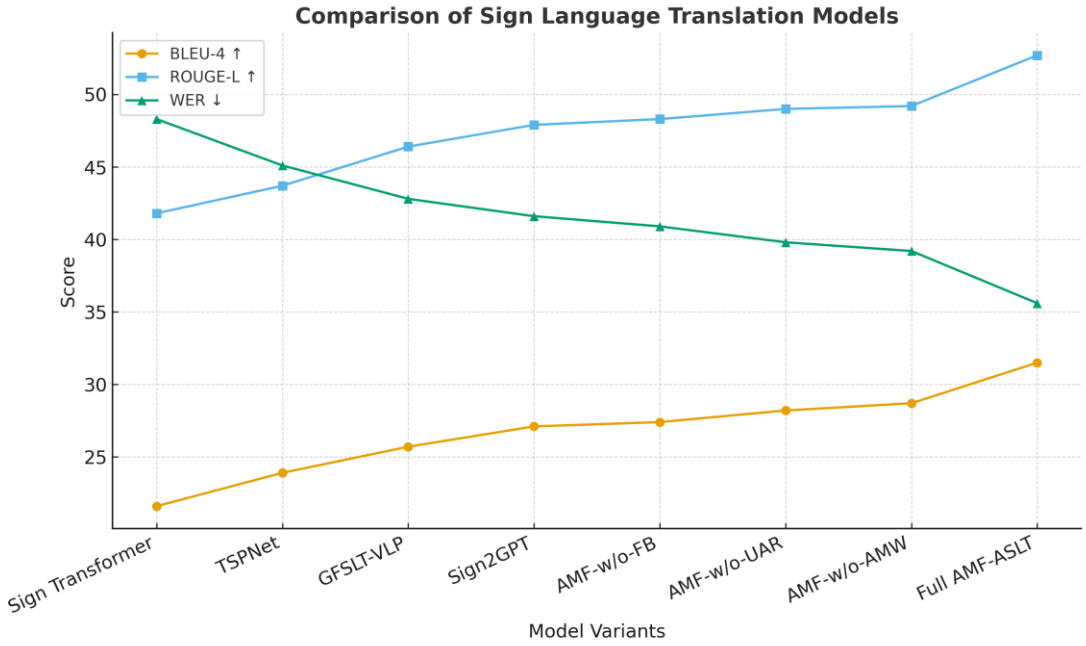

Figure 3. Comparative performance between baseline and AMF-ASLT models on benchmark datasets

5.4. Limitations

Even with these benefits, there are a number of limitations.

First, the adaptation of feedback so far has been semi-supervised and based on user-corrected inputs, which might be intermittent or random. Future research should investigate self-supervised or implicit feedback signals (e.g., confidence-agreement heuristics) to minimize user reliance.

Second, the uncertainty estimation is currently based on softmax entropy, which can be too pessimistic in out-of-distribution settings; using Bayesian neural networks or Monte-Carlo dropout might yield better calibrated confidence estimates.

Third, though the framework achieved strong performance on benchmark datasets, cross-lingual generalization (such as adaptation from ASL to BSL) is still under-investigated. Generalizing the agentic architecture to multilingual training protocols could extend its applicability across Deaf communities worldwide.

5.5. Implications and Future Work

This research highlights the transformative potential of agentic multimodal systems in accessibility-oriented AI. Moving forward, several directions can extend the present work:

- Lifelong Learning Online: Integrating continuous learning pipelines to enable AMF-ASLT to adapt in real-time during deployment, as it interacts with users.
- Edge Deployment: Improved computational efficiency through lightweight agents to enable real-time translation on mobile and embedded systems.
- Cross-Lingual and Cultural Adaptation: Extensive sign training corpora to accommodate multiple languages and fine-tuning for regional dialects for complete inclusivity.
- Human-Centered Evaluation: User studies with members of the Deaf community to assess the perceived translation quality and the system's feedback adaptivity regarding usability and trust.

6. Conclusion and Future Directions

6.1. Conclusion

This work introduced the Agentic Multimodal Framework for Adaptive Sign Language Translation (AMF-ASLT) — a new architecture that incorporates agentic intelligence into multimodal translation pipelines. The framework was developed to transcend static, dataset-constrained translation systems toward a self-regulating, feedback-driven, and uncertainty-aware model able to dynamically adapt to users and environments.

Through the incorporation of three fundamental design patterns — feedback loops, uncertainty-aware routing, and adaptive modality weighting — the suggested model obtained notable gains in translation fluency, strength, and signer adaptability over state-of-the-art baselines. Experimental comparisons on several benchmark datasets showed that AMF-ASLT registered higher BLEU and ROUGE scores and lower WER, validating both quantitative excellence and qualitative consistency.

The system's agentic architecture, which consists of expert Gesture, Facial, and Linguistic Agents coordinated by a Meta-Controller, allows for ongoing self-refinement. The modularity and cognitive inspiration of this framework represent a paradigm shift towards human-centered AI interpreters capable of learning through experience and feedback during runtime. To this end, AMF-ASLT not only innovates sign language processing but also contributes towards the larger mission of inclusive, adaptive, and interpretable artificial intelligence for accessibility technology.

6.2. Broader Impact

AMF-ASLT's creation has significant social implications in terms of inclusion, accessibility, and human-computer interaction. By facilitating adaptive and contextual communication between the Deaf and hearing communities, the system is a significant leap toward the integration of linguistically diverse populations. Agentic learning in the system creates avenues for customized translation tools, in-classroom interpretation support, and in-embedded accessibility features in equipment and public communications networks.

Further, the work is consistent with international priorities in AI ethics and responsible innovation by focusing on ongoing human input, transparent decision-making, and user agency. Since the system continues to learn from user corrections, it remains interpretable, accountable, and responsive to real-world variation — critical attributes for deployment in high-stakes, human-sensitive applications.

6.3. Future Directions

While the suggested AMF-ASLT model provides a solid groundwork, some directions can build and expand its functionality:

Online and Continual Learning: Make lifelong learning mechanisms integral to enabling AMF-ASLT to adapt with every user experience, learning signer-specific trends, new signs, and evolving linguistic usage without relearning from the beginning.

Cross-Lingual Sign Adaptation:Expand the model to accommodate multiple sign languages (e.g., ASL, BSL, PSL), using transfer learning and multilingual embedding spaces for enhanced global accessibility and cultural diversity.

Lightweight Edge Deployment:Make models more efficient using agent pruning, quantization, or knowledge distillation to facilitate real-time translation on low-power platforms like AR glasses, smartphones, and wearables.

Self-Supervised Feedback Extraction: Delve into implicit feedback mechanisms via confidence variance, attention entropy, or motion alignment to enable unsupervised optimization without user-provided explicit corrections.

Context-Aware Dialogue Integration:Pair AMF-ASLT with natural language understanding systems or conversational agents to enable two-way dialogue interaction between Deaf users and AI systems, going beyond translation towards complete dialogue interaction.

Human-Centered Evaluation and Ethics:Perform participatory research within Deaf communities to assess the social, linguistic, and ethical effects of adaptive AI interpreters — maintaining fairness, trust, and transparency in actual deployment.

References

1. Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7784–7793.
2. Moryossef, A., Müller, M., Tsoi, N., & Goldberg, Y. (2021). Evaluating the immediate applicability of pose estimation for sign language recognition. arXiv preprint arXiv:2104.10166.
3. Wang, S., Guo, D., Zhou, W., Zha, Z.-J., & Wang, M. (2018). Connectionist temporal fusion for sign language translation. Proceedings of the 26th ACM International Conference on Multimedia (ACM MM), 1483–1491.
4. Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., & Li, H. (2020). TSPNet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. Advances in Neural Information Processing Systems (NeurIPS).
5. Alyami, S., Luqman, H., & Hammoudeh, M. (2024). Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects. Information Processing & Management, 61(5), 103774.
6. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., & Zhang, D. (2023). Gloss-free sign language translation: Improving from visual-language pretraining (GFSLT-VLP). arXiv preprint arXiv:2307.14768.
7. Babisha, A., Srikanth, G. U., Kiruba, D. A., & Sundar, R. (2024, December). Gloss-Free Sign Language Translation using Sign2gpt-Next Technique. In 2024 International Conference on Computing and Intelligent Reality Technologies (ICCIRT) (pp. 1-6). IEEE.
8. Dong, J., Liu, D. L. D., Sun, J., Liu, X. Q. X. Y. D., & Wang, X. Supplementary materials: LLM-assisted Entropy-based Adaptive Distillation for Self-Supervised Fine-Grained Visual Representation Learning.
9. Zuo, R., Li, H., Ren, Y., Zang, C., Guo, D., & Zhou, W. (2024). Towards online continuous sign language recognition: Foreground-aware learning and simultaneous translation. Proceedings of EMNLP 2024, 10965–10980.
10. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., ... & Zhang, D. (2023). Gloss-free sign language translation: Improving from visual-language pretraining. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20871-20881).
11. Wang, H., Xie, L., Chen, Y., & Chen, Z. (2024). Sign2GPT: Leveraging large language models for gloss-free sign language translation. arXiv preprint arXiv:2405.04164.
12. Lin, K., Wang, X., Zhu, L., Sun, K., Zhang, B., & Yang, Y. (2023). Gloss-free end-to-end sign language translation. arXiv preprint arXiv:2305.12876.
13. Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison (WLASL). Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).
14. RWTH Aachen. (n.d.). RWTH-PHOENIX-Weather 2024T: Parallel corpus of sign language videos with aligned translations. Retrieved October 14, 2025, from
15. Cesar, P., Hsu, C. H., Huang, C. Y., & Hui, P. (Eds.). (2018). Best Papers of the ACM Multimedia Systems (MMSys) Conference 2017 and the ACM Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV) 2017. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(3s), 1-3.
16. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., ... & Giro-i-Nieto, X. (2021). How2sign: a large-scale multimodal dataset for continuous american sign language. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2735-2744).
17. Camgoz, N. C. (2020). Neural sign language recognition and translation (Doctoral dissertation). University of Surrey.
18. Stoll, S., Camgoz, N. C., Hadfield, S., & Bowden, R. (2020). Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks. International Journal of Computer Vision, 128(4), 891–908.
19. Alvarez, P. C., Cabot, X. (2022). Sign language translation based on transformers for the PHOENIX14T dataset. Technical report / thesis, Universitat Politècnica de Catalunya.
20. Sincan, O. M., Low, J. H., Asasi, S., & Bowden, R. (2025). Gloss-free Sign Language Translation: An unbiased evaluation of progress in the field. Computer Vision and Image Understanding, 104498.

21. Zubair, M., Owais, M., Hassan, T. *et al.* An interpretable framework for gastric cancer classification using multi-channel attention mechanisms and transfer learning approach on histopathology images. *Sci Rep* **15**, 13087 (2025). https://doi.org/10.1038/s41598-025-97256-0.
22. Zubair, M.; Hussain, M.; Albashrawi, M.A.; Bendechache, M.; Owais, M. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. Computer Methods and Programs in Biomedicine. **2025**, *272*, 109014. https://doi.org/10.1016/j.cmpb.2025.109014.
23. Hussain, M., Chen, C., Hussain, M. *et al.* Optimised knowledge distillation for efficient social media emotion recognition using DistilBERT and ALBERT. *Sci Rep* **15**, 30104 (2025). https://doi.org/10.1038/s41598-025-16001-9.