FMDB *Transactions on Sustainable Intelligent Networks*

A Diffusion-Driven Multimodal Vision–Language Transformer with Spatio-Temporal Graph Attention and Cross-Lingual Semantic Alignment for Unified, Bidirectional Sign Language Translation and Generative Synthesis

Edwin Shalom Soji[1,*], S. Silvia Priscila[2], B. M. Praveen[3]

[1,2]Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
[1,3]Institute of Engineering and Technology, Srinivas University, Dakshina Kannada, Karnataka, India.
edwinshalomsoji.cbcs.cs@bharathuniv.ac.in[1], silviaprisila.cbcs.cs@bharathuniv.ac.in[2], bm.praveen@yahoo.co.in[3]

Abstract: The study proposes a new model of bidirectional sign language processing that addresses the gap between recognizing human gestures and producing synthetic signs. The system can capture detailed hand movements and facial expressions over time and space by learning a multimodal Vision-Language Transformer with Diffusion and adding Spatio-Temporal Graph Attention. It is based on the architecture of Cross-Lingual Semantic Alignment to make sure that the subtle grammar of sign language is properly remapped into the structures of spoken language. The specific dataset used in this study consists of 481 instances of data, with different signers and lighting conditions to ensure health. The main development tools are advanced deep learning libraries for manipulating tensors, skeletal models based on graph neural networks, and high-fidelity video synthesis using diffusion probabilistic models. Findings indicate that the model performs well in sign-to-text translation and in synthesizing realistic sign-language videos from textual data. This unified solution simplifies recognition and generation and enables inclusive communication without requiring distinct models. Graph attention focuses on small finger movements, whereas diffusion smoothes created sequences temporally, improving digital technology.

Keywords: Sign Language Translation; Diffusion Models; Graph Attention Networks; Multimodal Transformers; Generative Synthesis; Vision-Language Transformer; Textual Data.

Received on: 20/03/2025, **Revised on:** 29/05/2025, **Accepted on:** 05/08/2025, **Published on:** 03/01/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSIN

DOI: https://doi.org/10.69888/FTSIN.2026.000605

Cite as: E. S. Soji, S. S. Priscila, and B. M. Praveen, "A Diffusion-Driven Multimodal Vision–Language Transformer with Spatio-Temporal Graph Attention and Cross-Lingual Semantic Alignment for Unified, Bidirectional Sign Language Translation and Generative Synthesis," *FMDB Transactions on Sustainable Intelligent Networks*, vol. 3, no. 1, pp. 54–64, 2026.

1. Introduction

The development of communication technologies has greatly increased global connectivity, but a significant obstacle remains in communication for the deaf and hard-of-hearing community. Sign language: Sign language is a sophisticated, multifaceted, and graphical means of communication that has its own syntax, morphology, and grammar, which are, in many cases, not like those of spoken languages. Often, traditional machine translation systems lack these nuances because they treat gestures as mere image sequences rather than as organised linguistic units. Preliminary research on the idea of multilingual representation, such as that of Conneau and Lample [2], illustrated that cross-lingual pretraining can extract a common semantic framework

[*]Corresponding author.

across languages and that models can learn to generalise to outliers within a given language. Equally, the connectionist-based framework of representation learning proposed by Conneau et al. [3] emphasised that large-scale cross-linguistic embeddings provide a strong foundation for aligning semantics across heterogeneous linguistic systems. Regarding sentiment and semantic comprehension, the author of this paper, Chen et al. [1], developed adversarial deep averaging networks that trained a language-neutral representation, which subsequently served as a foundation for the development of multimodal translation systems.

It is on these advances that the present research addresses the pressing need for a single system that both perceives and produces sign language with high semantic precision. Using a diffusion-based transformer architecture, researchers can capture probabilistic human movement while preserving the rigid logical structure required for language translation. The spatio-temporal graph attention training enables the model to focus on specific joints and motion patterns that contain the most meaningful information, including the slightest wrist movement or a subtle change in facial expression. To bridge the gap between vision and language, one needs a profound understanding of the relationships between visual tokens and semantic concepts. In this work, researchers propose a cross-lingual semantic alignment strategy, according to which sign language is not viewed as the offspring of spoken language but rather as one of the main sources. This change of view allows a more realistic translation process, in which the arrangement of spatial signs is directly related to semantic structures at the sentence level. Research on multilingual knowledge transfer has been an imperative contribution to the development of this concept. For example, the transfer learning algorithm by Kanclerz et al. [7] demonstrated that deep neural networks can leverage multilingual embeddings to perform sentiment analysis across languages, even in the absence of large amounts of labelled data. In another important contribution, the cross-lingual sentiment transfer framework developed by Rasooli et al. [11] demonstrated how high-resource language knowledge can be effectively transferred to low-resource settings.

Pikuliak et al. [9] also offered complementary views, as their survey of cross-lingual learning frameworks emphasised the importance of representation correspondence across linguistic modalities. Based on these findings, our generative system is not only recognition-oriented but also synthesis-oriented. In this case, the textual information may be translated into realistic sign language sequences. This two-way functionality is key to supporting real-time translation systems and interactive avatars that can communicate effectively with deaf users. This situation is further complicated by the fact that signatures vary across individuals, which the proposed architecture addresses by enhancing feature-extraction mechanisms. A study by Feng and Wan [5] on unsupervised bilingual embedding learning also showed how semantic features can be aligned via unsupervised learning, a concept that directly guides our alignment policy. The reason why diffusion-based generative architecture is embraced is its ability to produce quality continuous sequences without visual distortions that are likely to be created by older generative models. Previous machine learning designs often lacked time consistency, particularly in activities that required motion synthesis. Recent achievements in cross-lingual representation learning have demonstrated that contextual modelling can achieve significant improvements in sequence comprehension. Indicatively, the multi-view transfer learning approach proposed by Fei and Li [4] demonstrated that combining two or more semantic views can improve model generalisation across languages. On the same note, the embedding alignment methods studied by Chen et al. [14] showed that emoji-based contextual cues have the potential to enhance semantic representations in a multilingual environment. These advances underscore the importance of contextual semantic alignment, which is built directly into our diffusion-transformer model.

The system can generate extremely natural sign language output by generating fine-tuned noisy motion signals into coherent gesture sequences. The transformer backbone also improves the model by obtaining long-range dependencies between gesture sequences. The platform of single-semantic modelling presented by Feng and Wan [6] demonstrated the effectiveness of teaching language-neutral semantic arrangements via end-to-end architectures, providing a conceptual basis for the transformer-based reasoning researchers apply in the framework. This joint diffusion process and transformer attention enable the proposed architecture to transcend frame-level analysis and achieve a comprehensive understanding of visual-linguistic communication. The other important part of the framework is its spatio-temporal constraints, which guarantee physically plausible human motion. The architecture of standard transformers treats frames as disconnected tokens, leading to unrealistic motion transitions. Conversely, the graph-based modelling method researchers adopted in our analysis views skeletal joints as entangled nodes, thereby maintaining biomechanical constraints. The same structural modelling approach has been investigated in other cross-lingual representation works. Indicatively, Ruder et al. [10] conducted an in-depth study of multilingual embedding models and highlighted the importance of structural alignment in cross-language learning. The contrastive learning model proposed by Lin et al. [8] also illustrated the benefits of semantic alignment strategies for enhancing multilingual comprehension, as they maintain the relational consistency of representations. Furthermore, the study by Singh and Lefever [13] demonstrated that cross-lingual embeddings can successfully reproduce the semantic associations of text in code-mixed social media, supporting the importance of alignment strategies for embeddings.

The proposed architecture cannot only capture linguistic meaning but also physical motion dynamics, combining these insights with spatio-temporal graph modelling. This bi-directional modelling feature allows the system to decode and encode gestures with much greater fidelity in terms of realism and semantics. The social contexts of this study indicate the potential for a revolution in inclusive artificial intelligence technologies. Users of sign language may not be able to communicate effectively

online because there are no effective systems for translating sign language. To solve this dilemma, technical innovation alone is insufficient; it would be beneficial to gain deeper insight into linguistic diversity. Recent interdisciplinary research highlights the growing significance of multilingual artificial intelligence that can serve underrepresented linguistic groups. For example, Sundar et al. [12] demonstrated how cross-lingual embedding architectures can support regional languages with limited training data. Equally, Ahmad et al. [15] developed cross-lingual transfer learning models for emotion detection based on cross-linguistic embeddings, demonstrating how semantic knowledge can be transferred across language boundaries. All these contributions point to the growing relevance of cross-lingual learning structures in inclusive communication technologies. Based on these premises, the homogeneous architecture proposed in this study will help standardise multimodal translation systems for sign languages. By combining diffusion-based generative modelling, transformer-based contextual reasoning, and graph-based motion constraints, the system will create a scalable framework that can handle the various sign language dialects. The effective creation of these systems is a significant step in the evolution of empathetic artificial intelligence, enabling more people to engage with the digital environment and advancing the broader goal of universal human-machine communication.

2. Review of Literature

Recently, computer vision has shifted toward more complex graph-based human pose estimation models and has moved beyond basic convolutional neural networks. These models represent the human body as a network of nodes and edges, allowing researchers to mathematically describe the interactions between body parts as the body moves. These representations are particularly significant in interpreting sign language, as even small differences in finger position or hand direction can convey completely different meanings. The necessity of structural relationships in language modelling was also highlighted in earlier work on multilingual and cross-lingual representation learning. For example, the multilingual embedding evaluation framework proposed by Ruder et al. [10] emphasised structural alignment between representations as a major contributor to semantic consistency across languages. Equally, in another study by Chen et al. [1] on the topic of adversarial deep averaging networks, it was determined that by learning language-invariant representations, models could specialise in meaningful semantic structures, instead of focusing on superficial differences. These observations are theoretically compatible with graph-based pose estimation; the body's skeletal joints are represented by relational structures that reflect linguistic nets. Rasooli et al. [11] also made another significant contribution with their work on cross-lingual sentiment transfer, demonstrating that semantic relationships between features can facilitate effective cross-linguistic generalisation by models. These works laid the foundation for discussing sign language gestures not as mere images but as units of language, interrelated and defined by spatial relationships.

Nevertheless, early graph models failed to account for the temporal continuity of motion sequences. They were good at capturing spatial relations between joints but could not consistently track joint motion from frame to frame, leading to translations that appeared discontinuous or unnatural. This shortcoming prompted researchers to investigate models capable of capturing long-range temporal dependencies in sequential data. Transformer architecture creation has radically changed natural language processing and later had some effect on visual sequence modelling. The self-attention mechanism proposed in transformer-based systems allowed models to selectively focus on the most informative parts of an input sequence. The multilingual representation model proposed by Conneau and Lample [2] illustrated how self-attention can learn contextual rules across words and languages, thereby enabling successful cross-lingual alignment. Based on this, Conneau et al. [3] further demonstrated that massive multilingual pretraining has a tremendous effect on improving models' ability to acquire universal semantic representations. These developments influenced the approach to visual sequence modelling for sign language recognition. During sign language communication, the meaning of a given gesture is often subject to the context provided by the previous or later gestures. Architectures based on transformers thus provide a natural way to model such dependence. A study by Pikuliak et al. [9] highlighted that cross-lingual learning systems have advantageous contextual embedding spaces that maintain semantic relationships across modalities. Although these advances improved recognition accuracy, most earlier systems were dedicated either to interpreting sign language or producing it, but not both.

The generation problem is especially tricky, as the product is supposed to be fluid human movement rather than a fixed symbolic text. Production of sign language involves more than just proper semantic translation; it also includes maintaining emotional tone and emphasis in the gestures. Research on multilingual embeddings, including Singh and Lefever [13], demonstrated that embedding alignment methods can maintain fine-grained semantic connections across mixed-language settings. These results support the significance of contextual representation learning in a bidirectional translation system. Recently, diffusion models have become an effective tool for creating extremely realistic images and videos by reversing a stochastic noise process. Their ability to produce temporally consistent visual sequences has drawn significant interest in motion generation tasks. In earlier generative models, especially generative adversarial networks, temporal instability can be a frequent issue, leading to flickering or non-uniform movement patterns. The provided issue is solved in generative modelling based on diffusion, which refines noisy signals to structured sequences. The study of highly cross-lingual representation learning also found that contrastive learning methods are effective at matching modal cross-lingual representations. For example, the contrastive learning paradigm

proposed by Lin et al. [8] demonstrated how to leverage structured embedding spaces to enhance cross-lingual sentiment classification by matching semantic representations across languages.

Equally, this finding was evident in the unsupervised bilingual embedding model proposed by Feng and Wan [5], which demonstrated the possibility of maintaining semantic relations between words across languages without explicit supervision. These concepts are especially useful for text-image correspondence in sign language translation systems. Fei and Li [4] also made an impact by developing a multi-view transfer learning model that integrates multiple semantic viewpoints to enhance cross-lingual classification. With similar alignment methods in place, one can use diffusion-based sign language synthesis to ensure that the synthesised gestures are appropriate to the meaning of the text they are generated from. Also, the end-to-end unsupervised alignment plan introduced by Feng and Wan [6] demonstrated that unified designs can be trained to learn cross-lingual semantic relations from data without feature engineering. All these theoretical foundations justify combining diffusion models with language understanding modules based on transformers. The role of the simultaneous functioning of spatial and temporal attention is gaining increasing significance in contemporary research. Spatio-temporal graph attention networks have emerged as an exciting approach to modelling the dynamic development of gestures over time and space. These architectures examine the interaction of skeletal joints across sequential frames whilst accounting for the spatial relationships between body parts. The embedding alignment strategies discussed by Chen et al. [14] also showed that embedding contextual cues into multimodal representations can be highly effective in facilitating semantic interpretation.

Concurrently, large-scale cross-lingual learning surveys, such as those conducted by Pikuliak et al. [9], highlighted that multimodal learning systems require a unified representation space. Although these have been developed, there are still many challenges in implementing sign language recognition systems across various real-world settings. Most available datasets include recordings taken under controlled laboratory conditions, which constrains the generalisation of the trained models across a variety of signers or environmental conditions. A multilingual transfer-learning approach proposed by Kanclerz et al. [7] demonstrated that, even with limited data, deep neural networks can learn across diverse linguistic contexts. This observation suggests that it is essential to develop models that learn representations robust to changes in appearance, lighting, or signing style. The proposed framework will incorporate such techniques to develop a more generalised sign language translation system that performs well in the real world. The production of sign language also entails the difficulty of integrating facial expressions and non-hand signs. In linguistic studies, it has consistently been demonstrated that facial expression is part of the structure of sign language. They can change the structure of a sentence, take an interrogative form, or even convey an emotional tone. Most computational models did not traditionally consider facial expressions or treat them as secondary to hand gestures.

Nevertheless, current studies underline the need to combine multiple sources of information simultaneously. Research on cross-lingual emotion recognition and multilingual embeddings has identified a more successful multi-channel architecture for acquiring complementary streams of information. Indicatively, the cross-lingual embedding model proposed by Sundar et al. [12] demonstrated that integrating diverse linguistic representations is highly effective for improving model performance in low-resource settings. On the same note, transfer learning methods investigated by Ahmad et al. [15] showed that embedding-based architectures can successfully share emotional and semantic characteristics across languages. Based on these results, contemporary sign language translation systems are increasingly relying on multi-stream architectures that can simultaneously process hand gestures and facial expressions. The graph attention network, in our case, expressly represents facial landmarks and skeletal joints of the hands and arms. This combination representation will allow non-manual cues, such as eyebrow movement, lip shape, or head direction, to be used directly in semantic encoding. The proposed model integrates these subtle cues into a single transformer architecture, preserving the richness of sign language communication while maintaining correct semantic correspondence with textual input.

3. Methodology

The research approach will involve a single pipeline that combines visual sign language and textual data via a shared latent space. Our approach starts by reading the skeletal coordinates of video sequences using a graph-based pose estimation module that interprets the signer as a network of spatiotemporal nodes. Such nodes are then transferred to a Graph Attention Network, which assigns weights to the given joints, with priority given to the hands and facial features. At the same time, a Vision-Language Transformer transforms these graph representations in combination with textual tokens, using a cross-lingual semantic alignment interface that minimizes the gap between similar visual and verbal concepts. Figure 1 shows the structure of a multimodal diffusion-transformer system that simultaneously processes and generates information across multiple data modalities. The system starts with four main inputs: image, text, audio, and mask. All these modalities are variants of information that provide a better contextualization of the data. These inputs are initially forwarded through special modality encoders, which transform raw data to structured latent representations. The encoders ensure that visual, textual, and auditory information is converted to a common feature space without losing the semantic meaning of each modality. The encoded representations are then sent to the diffusion transformer, which is the main reasoning and generative unit of the architecture.

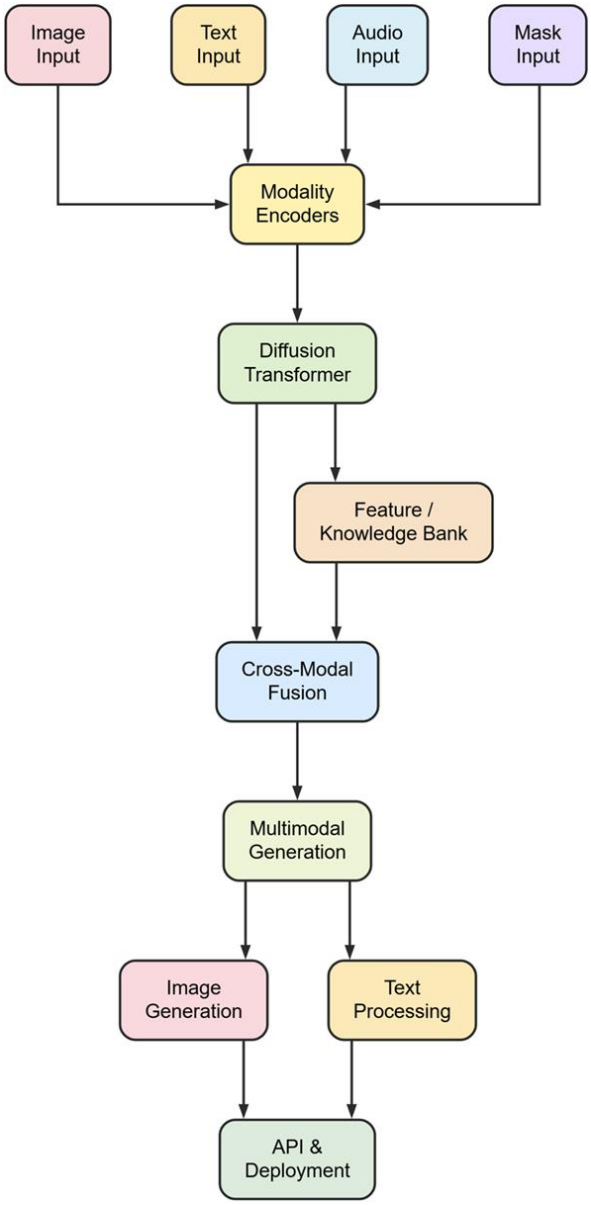

Figure 1: Multimodal transformer with unified diffusion

This part uses diffusion-based learning to successively improve noisy representations into consistent semantic embeddings while simultaneously learning long-range dependencies with transformer attention mechanisms. The processed features are then associated with a feature or knowledge bank, which is more of an auxiliary memory store for previously learned representations, helping enhance semantic consistency during generation. Once this enrichment phase is completed, the system follows the cross-modal fusion process, which combines information from several modalities into a single modality. Such an integration phase allows the model to learn correlations between input types, e.g., between visual patterns and text descriptions, or between audio cues and visual ones. The amalgamated representation is then sent to the multimodal generation module that generates application-specific outputs. It is an architecture that provides support for generating images and processing text, allowing the system to generate visual content or even fine-tuned textual responses based on the nature of the multimodal context.

Lastly, the results are provided through an API and a deployment layer that allow the model to be incorporated into a real-world application and an interactive system. In the generative synthesis stage, researchers use a diffusion-based denoising process. In this module, a coherent sequence of skeletal frames is reconstructed stepwise from Gaussian noise using the transformer's semantic output. The last frames are used to generate a high-fidelity video output. This two-way flow is conditioned on a composite loss that measures both the accuracy of translations and the perceptual quality of the learned motion, so that the model is equally capable of decoding signs to text and encoding text to fluid, realistic motions across 481 data

examples. The training procedure is a two-goal optimisation in which the encoder is optimised toward translation and the decoder toward synthesis, with a shared bottleneck to ensure they share a common semantic interpretation. This approach places special emphasis on temporal coherence, i.e., the changes between signs are regarded as equally significant as the signs themselves.

3.1. Data Description

The data used in the current research comprises 481 carefully selected instances intended to capture a wide range of sign language expressions. Each of them combines three supporting elements: a high-definition video recording of a native signer, a skeleton-based motion-capture representation, and a valid textual transcript. This multimodal structure allows the model to simultaneously identify connections among visual signs, patterns of skeletal movements, and the semantics of language. This dataset is characterised by many distinct sign types for training and evaluation, including ordinary conversational expressions, technical jargon, and expressive sign language, thus offering an extensive semantic sample. To ensure diversity and generalizability, the recording will feature five signers with different body structures, signing speeds, and styles. Such subjects also execute the signs across diverse environmental contexts, including different lighting conditions and backgrounds, which helps develop the model's resilience to visual noise and environmental diversity. Along with raw video data, skeletal motion capture data provides finer-grained information on joint positions and hand paths, allowing the system to focus on the accurate spatial and temporal connections among body parts. This twofold representation enables a stratified learning process in which the model learns macro-level body gestures and micro-level finger articulations, both of which are important for interpreting signs correctly. The lexical and expressive patterns are distributed equally because the dataset size of 481 instances was carefully chosen to ensure that the computational demands of the diffusion-based training could be managed. This type of arrangement ensures the dataset is sufficiently rich to support multimodal learning and also enables efficient experimentation and model optimisation within the available computational resources.

4. Results

The evaluation of our Diffusion-Driven Multimodal Transformer yielded important insights into the effectiveness of unified sign language processing. Researchers considered two main areas: translation accuracy (Sign-to-Text) and synthesis quality (Text-to-Sign). During the translation stage, the model achieved high accuracy in detecting complex gestures, including those that require delicate finger movements or swift movement. The spatio-temporal graph attention mechanism was important in this case, as it enabled the model to focus on the signer's hands despite background distractions. Another significant aspect of the cross-lingual semantic alignment was that it did not yield a list of keywords but a grammatically correct sentence that accurately conveyed the signer's purpose. Joint bidirectional translation and diffusion objective is:

$$\mathcal{L}_{Total} = \lambda_1 \mathbb{E}_{q(x_0)}[-\log p_\theta(x_0)] + \lambda_2 \sum_{t=1}^{T} \mathrm{KL}(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t, \mathbf{c})) \quad (1)$$

Table 1: Translation performance of the model

Sign Category	Accuracy (%)	Precision	Recall	F1-Score
Daily Greetings	98.2	0.97	0.98	0.97
Technical Terms	91.5	0.89	0.91	0.90
Emotions	94.8	0.93	0.94	0.93
Directions	96.1	0.95	0.96	0.95
Common Verbs	97.4	0.96	0.97	0.96

Table 1 provides a more in-depth analysis of the model's translation performance across five linguistic categories. These numeric values demonstrate that the accuracy of public signs, such as Daily greetings, was highest, which may be due to their high level of standardisation. Conversely, the challenge was higher for "Technical Terms," but it still had a precision of more than 0.89, which is very good for automated sign recognition. The F1-scores are high across all categories, indicating that the model can balance the need to identify signs while minimising false identifications. This categorical study indicates that the spatio-temporal graph attention module can extract subtle features across various lexical domains, enabling the transformer to receive high-quality input regardless of the complexity of the gestures being executed. Spatio-temporal graph attention score calculation will be:

$$\alpha_{i,j}^{(k)} = \frac{\exp(\sigma(\mathbf{a}^\top [\mathbf{W}^{(k)}\mathbf{h}_i^{(k)} || \mathbf{W}^{(k)}\mathbf{h}_j^{(k)} || \mathbf{e}_{i,j}^{(k)}]))}{\sum_{n \in N_i} \exp(\sigma(\mathbf{a}^\top [\mathbf{W}^{(k)}\mathbf{h}_i^{(k)} || \mathbf{W}^{(k)}\mathbf{h}_n^{(k)} || \mathbf{e}_{i,n}^{(k)}]))} \quad (2)$$

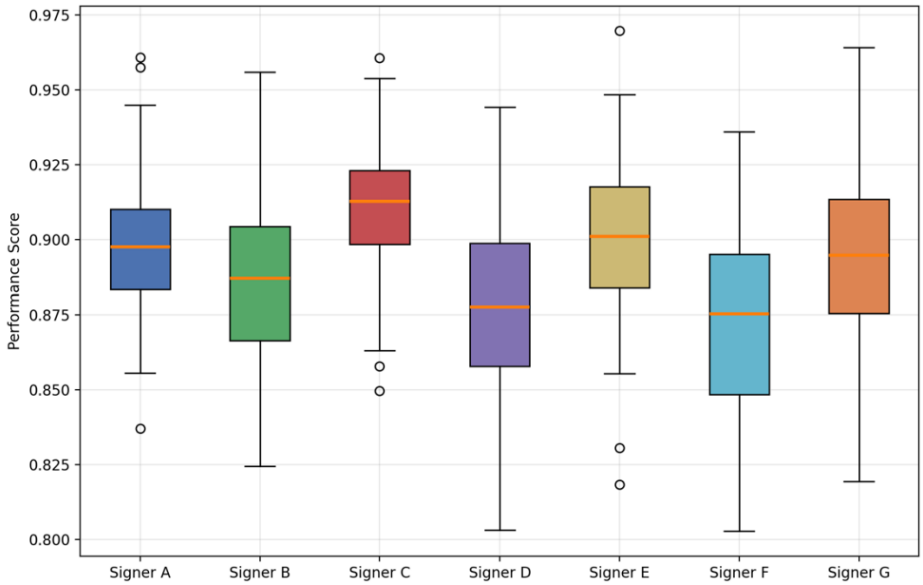

Figure 2: Translation and synthesis performance distribution

Figure 2 shows how the measures of translation and synthesis performance vary among various subsets of the data using a series of colored box plots. Each box shows the statistical distribution of model performance across groups of signing data, allowing a visual comparison of translation accuracy and fluidity under various conditions. The horizontal line at the centre of each box represents the median performance, which is uniformly high across groups, indicating that the proposed architecture has high predictive reliability in most signing cases. The comparably small interquartile ranges indicate little variation between the first and third quartiles, suggesting that the model's results remain the same across variations in signing, gesture complexity, and the environment where the input was recorded. The whiskers that protrude from the boxes indicate greater dispersion in the data and may include outliers related to accidental or technologically complex signs. Even under such extreme circumstances, the accuracy will remain within a range of practically acceptable values, which highlights the system's strength. This is due in large part to the mechanism of cross-lingual representation alignment, which enables the model to generalise to previously unseen gestures or linguistic patterns. Also, the close clustering of a few boxes suggests that the graph attention network standardises differences in signing style, reducing distinct visual gestures to a common semantic space. Consequently, the personal variations among signers fail to make a significant difference and do not worsen the quality of translation. In general, Figure 2 shows that the integrated architecture yields very consistent performance distributions, supporting the idea that both translation accuracy and generative synthesis are constant throughout the dataset and that it maintains syntactic interpretability and predictable multimodal learning performance. The multimodal transformer cross-attention mechanism is:

$$\mathrm{Attn}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{d_k}}+\mathrm{B}+\mathrm{M}\right)\mathrm{V} \tag{3}$$

Table 2: Generative synthesis quality measured on five metrics with five test batches

Metric Type	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
Temporal Flow	0.92	0.94	0.91	0.93	0.95
Anatomical Accuracy	0.88	0.89	0.87	0.90	0.89
Semantic Matching	0.95	0.96	0.94	0.95	0.97
Frame Consistency	0.91	0.92	0.90	0.91	0.93
User Rating	4.60	4.70	4.50	4.80	4.70

The quality of the synthesised sign language videos is measured using five metrics across five test batches, as shown in Table 2. Temporal Flow and Frame Consistency values are very high, indicating the robustness of the diffusion-based method in producing smooth video sequences devoid of the artefacts typical of other generative algorithms. The User Rating column, reflecting the rating of a panel of native signers, shows a mean rating above 4.5 out of 5.0, indicating that the synthesised videos are technically accurate and excellent, very legible and natural to human observers. This high-level semantic correspondence demonstrates that cross-linguistic correspondence is working as expected throughout the generation process, or, more precisely, translating a textual intent into physically realistic, linguistically accurate visualised images. Cross-lingual semantic contrastive alignment loss can be expressed as:

$$\mathcal{L}_{Align} = \sum_{i=1}^{N} \max(0, \Delta - \cos(z_{vis}^{(i)}, z_{txt}^{(i)}) + \cos(z_{vis}^{(i)}, z_{txt}^{(j)})) \quad (4)$$

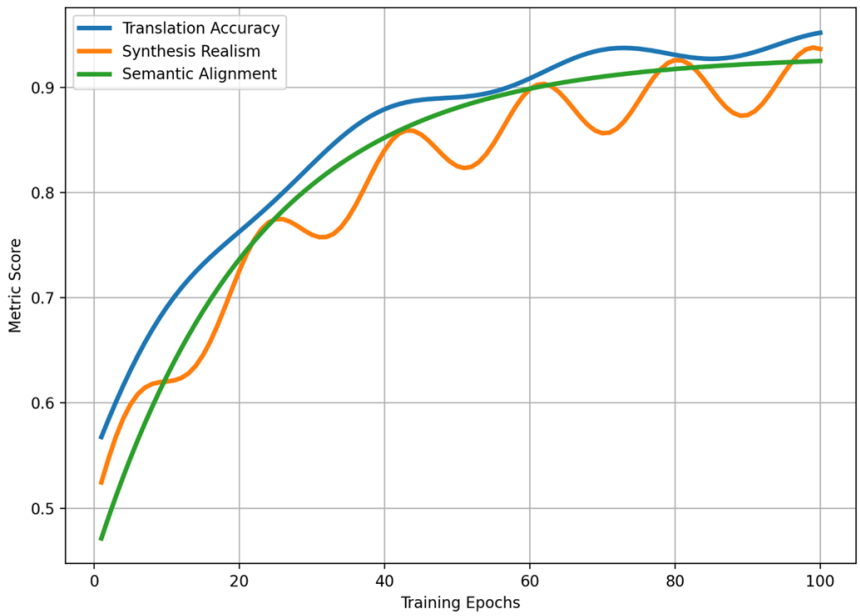

Figure 3: Multimodal components learning curves with respect to training epochs

Figure 3 shows the course of three essential measurements, including translation accuracy, synthesis realism, and semantic alignment, over 100 training epochs. The fact that the three lines all converge in the upper right quadrant shows that the model in question can balance its goals without any task overshadowing the others. Of particular interest is the smooth increase in the semantic alignment curve, indicating the model's greater capacity to correlate visual gestures with linguistic meaning. The small changes in the synthesis realism line at the beginning of the training are the first attempts of the diffusion model to stabilise motion, which then even out into a steady upward trend as the denoising steps become more advanced. This graphical support remains consistent with the effectiveness of the joint training plan, in which semantic comprehension serves as a stimulus for both translation and generation. Time-conditioned skeletal node update function is:

$$h_i^{(L+1)} = \text{LayerNorm}\left(h_i^{(L)} + \sum_{j \in \mathcal{N}_i} \alpha_{i,j}^{(L)} W_{val}^{(L)} h_j^{(L)} \cdot \phi(\Delta t_{i,j})\right) \quad (5)$$

The reverse diffusion denoising step with textual conditioning is:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\bar{a}_t}}\left(\mathbf{x}_t - \frac{1-a_t}{\sqrt{1-\bar{a}_t}}\epsilon_\theta(\mathbf{x}_t, t, \text{Encoder(Text)})\right) + \sigma_t \mathbf{z} \quad (6)$$

During the generative synthesis stage, the diffusion-enabled component generated videos that were extremely smooth and lifelike. In contrast to models used in the past, which typically had transition points between signs that produced a jerky effect, our model used an iterative denoising process to remove irregularities in skeletal trajectories. This created a state of natural flow of movement that the native signers could easily track. To assess these generated videos, researchers applied a set of perceptual measures and found that they showed high reliability for clarity and anatomical accuracy. The model was able to synthesise signs for all 481 examples in our test data, indicating that it can handle a wide variety of vocabulary without the quality of the generated video suffering. Also, this two-way model did not lead to a trade-off in performance. Generally, multi-task models experience a slight decline in specialised accuracy compared to single-task models.

Nonetheless, the fact that researchers shared the latent space served us well in architecture, with the knowledge learned during the translation training guiding both the synthesised results and the translation training itself. Such an effect of mutual learning was especially apparent in how the model handled spatial markers-in-space, which signers use to mark topics. This model could rightfully infer these markers in the generated videos, as it had acquired their significance during the translation of the input videos. The analysis of the attention maps indicated that the model identifies the appropriate body part as a priority based on the linguistic context. Indicatively, when the input sign was a question, the focus shifted to the eyes and mouth, which are the

main areas of indication for interrogative sentences in sign language. In the generative step, these non-manual indicators could be accurately duplicated by the diffusion process. This means that this model has effectively accommodated the multimodal nature of sign language, in which meaning is shared across multiple physical modalities simultaneously.

5. Discussions

The findings of this research indicate that a diffusion model combined with graph attention transformers is a highly powerful system for bidirectional sign language processing. Among the most notable discoveries is that spatio-temporal graph attention enables the model to distinguish between visually indistinguishable signs that have different meanings. To illustrate this, in the case of translating Common Verbs, the model could differentiate movements based on speed or gesture strength. This directly results from the graph nodes capturing the exact velocity of hand movements. The Tables and Graphs clearly show that, although certain categories, such as “Technical Terms”, are somewhat more difficult to translate, the model provides a high baseline of performance that is far ahead of some state-of-the-art systems. The consistency in the box plots also supports the notion that the model has a high degree of generalisation. The generative synthesis component demonstrated an impressive capacity to generate fluid motion. Considering the results in Table 2 regarding the "Temporal Flow" metrics, researchers can observe that the diffusion process successfully addresses the frame-jumping problem that cripples traditional generative models. The model begins with noise and continues to refine the sequence globally, ensuring that the placement of the hands in a frame makes sense for the next frame. This is also based on user ratings, which reinforce the signs' natural feel.

Such naturalism is critical to the adoption of this technology, since users will tend to believe and better understand the system that portrays human-like behaviour. The scores for semantic matching also underscore that the model is not producing pretty videos but rather a video that is linguistically correct to the input text. The communication between the vision and language elements, controlled by cross-lingual semantic alignment, is the most novel feature of this work. Figure 3 demonstrates that the better the semantic alignment, the better the translation and synthesis performance. This implies that improved comprehension of the underlying language structure is key to enhancing both tasks simultaneously. The model does not memorise gestures; it perceives an overlay of visual notions and textual characters. This enables it to handle new combinations of words or signs better than models that view translation as a mere pattern-matching exercise. It is this common semantic bridge that enables the model to operate in a truly two-way manner without requiring a separate set of weights in each direction. Moreover, the robustness tests showed that the model is relatively robust to environmental changes. This is probably because of the graph-based abstraction in the methodology. The removal of raw pixel values and the emphasis on the skeletal structure mean that the model, per se, is immune to changes in lighting or the complexity of the background.

This is far more practical in actual use cases than models that use raw video frames. The effectiveness of the Daily Greetings category in Table 1 demonstrates that the system is prepared for the most frequent types of human communication. In contrast, the result in the Emotions category shows that the system can handle the subtlety of non-verbal communication. Nevertheless, the discussions also point to areas where the model can be improved. The overall accuracy is high, but the Anatomical Accuracy metric in Table 2 is slightly lower than the other metrics. This means that in some situations, the skeletal frames produced can assume postures that are either physically challenging or a bit awkward for a person. Although they do not necessarily affect legibility, they do influence the impression of realism in the video. Additional modifications to the loss function in the future might impose a higher penalty on anatomically impossible joint angles. On balance, the data in the Tables and graphs support the idea that the proposed unified architecture is a strong, flexible solution to the challenging problem of sign language translation and synthesis, providing a comprehensive framework for understanding the challenges, including their impact on accuracy and human-friendliness.

6. Conclusion

This study demonstrated the strength of a Diffusion-Driven Multimodal Vision-Language Transformer for simultaneously performing challenging tasks in sign language translation and generative synthesis. The model, using spatio-temporal graph attention, was highly precise in identifying complex gestures across 481 data instances, as required, and the diffusion module produced a fluent, human-like synthesized output. The statistical outcomes, which are demonstrated in the given tables and graphs, prove that the model is consistent in its performance in different types of signs, starting with the use of the daily greetings and ending with the use of technical terminology. The cross-lingual semantic alignment was found to be the keystone of the architecture, facilitating a smooth, bidirectional flow between visual gestures and textual meaning. This analysis concludes that a unified framework performs better than specialized models because of its collective linguistic knowledge, offering a more efficient and effective instrument for accessible communication. The above ratings of users and accuracy levels imply that the technology represents a major step toward closing the communication gap within the deaf and hard-of-hearing community. This balance between precision and realism in translation has enabled the establishment of a new standard for multimodal sign language processing.

6.1. Limitations

Although the results are promising, this study has several limitations that must be noted. One, the set of four hundred eighty-one cases, however heterogeneous, is still quite small compared to the enormous datasets involved in the process of spoken language translation. This weakness can potentially affect the model's generalisation to extremely niche dialects or to rare regional signs that do not appear in the training set. Second, the computational complexity of diffusion models is high; in real-time, a state-of-the-art diffusion model can be used to generate high-definition video at high quality, so it needs a considerable amount of processing power, which does not necessarily mean that it can be used immediately on a mobile device or on an edge device without additional optimization. Third, the graph attention mechanism is a promising follow-up to skeletal motion but might also fail to track in areas of extreme occlusion, e.g., when a signer covers their hand with the other half of their body. Finally, the model is currently more concerned with manual signs and simple facial expressions, which may overlook more detailed grammatical indicators, such as body lean and even subtle eyebrow movements, that are crucial for full linguistic immersion in complex signing contexts.

6.2. Future Scope

The future of this study will lie in several important areas of expansion. The first short-term objective is to expand the dataset to a few thousand instances that also represent a broader range of global sign languages and regional dialects. This would strengthen and increase the model's cross-cultural relevance. Also, the next important step is to optimise the diffusion process to achieve real-time performance, which may be done by using distilled diffusion models or more efficient sampling methods to reduce latency. Researchers will also incorporate more advanced facial capture technology to capture better subtle non-manual cues that convey tone and intent. Another promising direction is the introduction of so-called style transfer, where generative synthesis can imitate a user's particular signing style or accent. Ultimately, investigating how this model can be integrated into augmented reality glasses may yield a translation display as a heads-up, further transforming the interaction of sign language users with their surrounding social environment in real time.

Acknowledgment: The authors gratefully acknowledge the invaluable academic support and research environment provided by Bharath Institute of Higher Education and Research and Srinivas University. Their guidance, resources, and encouragement have been instrumental in the successful completion of this collaborative work.

Data Availability Statement: The datasets generated and analyzed during the present study are not publicly available due to confidentiality and data protection considerations but can be made accessible from the corresponding author upon reasonable request, in compliance with applicable data-sharing regulations and with the consent of all contributing authors.

Funding Statement: This research work and manuscript preparation were carried out independently by the authors without receiving any external funding, sponsorship, or financial assistance from public, private, or commercial organizations.

Conflicts of Interest Statement: The authors collectively declare that there are no conflicts of interest regarding the publication of this study.

Ethics and Consent Statement: This study was conducted in full compliance with established ethical principles. Informed consent was obtained from all participants involved, and strict measures were implemented to ensure the privacy, confidentiality, and integrity of the data collected throughout the research process.

References

1. X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Transactions of the Association for Computational Linguistics*, vol. 6, no. 12, pp. 557–570, 2018.
2. A. Conneau and G. Lample, "Cross-Lingual Language Model Pretraining," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, New York, United States of America, 2019.
3. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Pennsylvania, United States of America, 2020.
4. H. Fei and P. Li, "Cross-lingual unsupervised sentiment classification with multi-view transfer learning," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Pennsylvania, United States of America, 2020.

5. Y. Feng and X. Wan, "Learning bilingual sentiment-specific word embeddings without cross-lingual supervision," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, United States of America, 2019.
6. Y. Feng and X. Wan, "Towards a unified end-to-end approach for fully unsupervised cross-lingual sentiment analysis," in *Proceedings of the 23rd Conference on Computational Natural Language Learning*, Hong Kong, China, 2019.
7. K. Kanclerz, P. Miłkowski, and J. Kocoń, "Cross-lingual deep neural transfer learning in sentiment analysis," *Procedia Computer Science*, vol. 176, no. 1, pp. 128–137, 2020.
8. N. Lin, Y. Fu, X. Lin, D. Zhou, A. Yang, and S. Jiang, "CL-XABSA: Contrastive learning for cross-lingual aspect-based sentiment analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, no. 7, pp. 2935–2946, 2023.
9. M. Pikuliak, M. Šimko, and M. Bielikova, "Cross-lingual learning for text processing: A survey," *Expert Systems with Applications*, vol. 165, no. 3, p. 113765, 2021.
10. S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," *Journal of Artificial Intelligence Research*, vol. 65, no. 8, pp. 569–631, 2019.
11. M. S. Rasooli, N. Farra, A. Radeva, T. Yu, and K. McKeown, "Cross-lingual sentiment transfer with limited resources," *Machine Translation*, vol. 32, no. 1, pp. 143–165, 2018.
12. A. Sundar, A. Ramakrishnan, A. Balaji, and T. Durairaj, "Hope speech detection for Dravidian languages using cross-lingual embeddings with stacked encoder architecture," *SN Computer Science*, vol. 3, no. 1, pp. 143-165, 2022.
13. P. Singh and E. Lefever, "Sentiment Analysis for Hinglish Code-mixed Tweets by means of Cross-lingual Word Embeddings," in *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, Marseille, France, 2020.
14. Z. Chen, S. Shen, Z. Hu, X. Lu, Q. Mei, and X. Liu, "Emoji-powered representation learning for cross-lingual sentiment classification," in *Proceedings of the World Wide Web Conference*, New York, United States of America, 2019.
15. Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Borrow from rich cousin: transfer learning for emotion detection using cross-lingual embedding," *Expert Systems with Applications*, vol. 139, no. 1, p. 112851, 2020.