A Grammar-Aware Multimodal Transformer for Structured ASL-to-English Translation

Enshirah Altarawneh
Department of Computer Engineering, Faculty of Engineering, The Hashemite University, Zarqa, Jordan
enshirah@hu.edu.jo (corresponding author)

Jawdat S. Alkasassbeh
Department of Electrical Engineering, Faculty of Engineering Technology, Al-Balqa Applied University, Amman, Jordan
jawdat1983@bau.edu.jo

Esraa Alshdaifat
Department of Information Technology, Faculty of Prince Al Hussein Bin Abdallah II for Information Technology, The Hashemite University, Zarqa, Jordan
esraa@hu.edu.jo

Aws Al-Qaisi
College of Engineering and Technology, American University of the Middle East, Egaila, Kuwait
aws.al-qaisi@aum.edu.kw

Maen Takruri
College of Engineering and Technology, American University of the Middle East, Egaila, Kuwait
maen.takruri@aum.edu.kw

Received: 16 February 2026 | Revised: 4 March 2026 and 17 March 2026 | Accepted: 18 March 2026

 DOI: https://doi.org/10.48084/etasr.18203

ABSTRACT

While the process of automatically translating American Sign Language (ASL) into English remains challenging due to the inherent complexities of creating signs in space-time and due to the existence of its own grammatical structure, one of the primary objectives of this study was to create an ASL-to-English Translation Framework incorporating grammatical representations. Utilizing a formalized grammatical model of ASL rather than simply viewing ASL as a series of unconnected, unrelated signs or motions, our method views ASL as a structurally based means of communicating that is comparable to spoken languages. In addition, our system captures spatial and temporal interrelations in ASL by processing multimodal input data consisting of Red Green Blue (RGB) color video frames, 2D/3D body pose keypoints, and hand landmark information while employing a transformer-based architectural design. Moreover, we created ASL grammar tokens which represent intermediate expressions of characteristics, including whether a given sign is negative, whether a subject has been explicitly referenced, etc. The utilization of these tokens facilitates a transition from the ASL representation to the corresponding English representation. The proposed methodology was tested via experiments conducted on two publicly accessible benchmark datasets: Word-Level American Sign Language (WLASL) and Microsoft American Sign Language (MS-ASL). Results indicated that the proposed methodology outperformed the current state-of-the-art methodologies. Significant improvements in Bilingual Evaluation Understudy (BLEU-4) scores (+5.6 and +5.0 relative to baselines) were realized for WLASL and MS-ASL, respectively. Additional evaluation metrics utilized to assess increased lexical accuracy and semantic coherence included Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L), Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Consensus-based Image Description Evaluation (CIDEr). Lastly, our grammar token prediction module achieved an exceptionally high accuracy rate of 95.1%, thereby providing further justification for the employment of structural linguistic modeling concurrent with multimodal feature fusion within the translation pipeline. The results suggest that combining multimodal feature fusion with grammar-aware representations provides substantial improvement over previously

employed methods for translating ASL into English and provides a foundation for future generations of ASL-to-English translation systems.

Keywords-American Sign Language (ASL); RGB video; grammar-aware translation; multimodal learning; transformer; CNN-LSTM; pose estimation; Bilingual Evaluation Understudy (BLEU)

I. INTRODUCTION

Sign languages, such as the American Sign Language (ASL), are full-fledged natural languages with unique grammatical structures that can be distinctly different from spoken languages in terms of syntax, morphology, and discourse structure, while incorporating non-linear word ordering and non-manual signals (e.g., facial expression and body position) that provide the context for meaning [1, 2]. The linguistic characteristics of ASL present a significant technical barrier for the development of practical automatic ASL recognition and translation systems.

In the early days of ASL recognition, researchers focused on the identification of individual signs and finger-spelling through visual characteristics using traditional Machine Learning (ML) methods [3]. In recent years, the rapid advancement in Deep Learning (DL) has enabled the utilization of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for the analysis of both the spatial and temporal components of signed sequences [4, 5]. Although CNNs and RNNs were successful in recognizing isolated or gloss-level signs, they are insufficient for performing an analysis of continuous sign language translation, including the examination of temporal dependencies and higher-order linguistic structures. Additionally, advancements in Sequence-to-Sequence (S2S) models and transformer-based architecture have enabled end-to-end translation of sign language video sequences into written spoken-language text [6-8]. However, most existing translation methods still treat the problem as a direct visual-to-text mapping task and implicitly assume structural correspondence between signed and spoken language grammar. In contrast, ASL conveys grammatical information through topic-comment structures, flexible tense marking, and non-manual cues rather than explicit lexical markers [9]. Consequently, failure to account for these structural properties often produces translations that are grammatically inconsistent or semantically incomplete. Previous studies have further shown that gloss-based or purely visual feature extraction methods are insufficient for modeling higher-level linguistic phenomena such as modality, negation, and discourse structure [10, 11]. These limitations highlight the need for translation models that explicitly incorporate linguistic modeling to bridge the structural differences between signed and spoken languages.

To combat these limitations, this paper presents a grammar-aware multimodal approach for translating ASL to English in real time. Unlike previous approaches that directly map visual features or gloss sequences to text, the proposed method introduces structured ASL grammar tokens as an intermediate representation encoding high-level linguistic properties, including sentence modality, temporal reference, negation, and topicalization. The grammar-aware representation is integrated within a transformer-based S2S architecture that combines Red Green Blue (RGB) video frames, body pose keypoints, and hand landmarks via multimodal fusion. By integrating explicit structural guidance into the translation pipeline without requiring word-level gloss annotations, the proposed framework generates more coherent, semantically faithful sentence-level translations while maintaining real-time performance.

The experimental results demonstrate that adding grammar-aware representations significantly improved the translation quality over state-of-the-art methods based upon either visual or gloss-based modeling, when evaluated on the publicly available Word-Level American Sign Language (WLASL) and Microsoft American Sign Language (MS-ASL) benchmarks.

II. RELATED WORK

Early ASL recognition systems primarily focused on recognizing hand gestures, including alphabets, numbers, and isolated signs, by mostly employing CNN-based architectures for static ASL handshape recognition from RGB images. Later, these methods were extended to dynamic gesture recognition and small-vocabulary tasks using CNN-Long Short-Term Memory (CNN-LSTM) models combined with handcrafted temporal features. Although these approaches can achieve high levels of accuracy in identifying specific words and symbols, they are limited to the recognition of single-word sequences and are not designed to translate complete sentences. Research has focused on developing ASL recognition systems that utilize ML algorithms. For example, a real-time Automatic Speech Recognition (ASR) system was developed in [6] that utilizes CNNs to recognize ASL hand gestures in real time. Additionally, a real-time ASL interpretation system was introduced in [12] that utilized keypoint tracking combined with DL models. These models are very efficient at recognizing gestures in real-time and are well-suited for many applications; however, like all previously discussed models, they remained limited to gesture classification and did not address higher-level linguistic structures, such as sentence-level semantics and grammatical organization.

Gloss-based approaches have also been investigated, where an intermediate gloss representation is generated before translation into spoken-language text [7, 8]. While gloss-based approaches can produce well-aligned translations, they require expensive manual annotation and often fail to capture important grammatical aspects of ASL, including spatial agreement, topicalization, and sentence modality [9]. End-to-end translation models instead attempt to directly map visual features to spoken-language sentences without gloss supervision [5]. Although these approaches reduce annotation requirements and improve scalability, they generally lack interpretability and cannot directly learn the linguistic structure of ASL from visual data. In contrast, the proposed methodology introduces a grammar-aware ASL-to-English translation framework in which ASL grammatical structure is explicitly modeled through an intermediate representation.

Unlike prior real-time translation systems based solely on isolated gesture recognition [6, 12, 13] and unlike conventional gloss-based approaches [7, 8], the proposed framework introduces explicit linguistic modeling within the translation pipeline.

A summary of some of the more relevant recent research on ASL recognition and translation is provided in Table I, including the datasets, model architectures, input modalities, and target tasks employed.

TABLE I. SUMMARY OF REPRESENTATIVE DATASETS, MODELS, AND TASKS IN ASL AND SIGN LANGUAGE RECOGNITION LITERATURE

Ref	Dataset (size/type)	Main ML/DL models	Accuracy	Input modality	Target / task
[14]	Kaggle ASL alphabets and digits (2,520 images, 36 signs)	Basic CNN, VGG-16, ResNet-50	ResNet-50: 94.05%	RGB images	Isolated letters and digits
[15]	Custom ASL alphabets, digits, symbols (64k+ images); Sign Language MNIST	SNDA (DenseNet + attention), InceptionV3, ResNet	SNDA: 99.76%; ResNet: 99.98%	RGB images	Letters, digits, static symbols
[16]	WLASL word-level dataset (>2,000 words, >100 signers)	CNNs, Pose-TGCN (temporal GCN)	Top-10 accuracy ≈ 62.63%	RGB video, 2D pose keypoints	Isolated word recognition (large vocabulary)
[17]	Custom ASL alphabet image dataset under varying conditions	Custom CNN	Up to 99.38%	RGB images	ASL alphabet letters
[18]	Multiple multilingual sign corpora, benchmark datasets	CNN + Bi-LSTM; NMT + MediaPipe + GAN	>95% classification; BLEU ≈ 38.06	Video, pose keypoints	Continuous recognition, translation, video generation
[19]	Boston ASL Lexicon Video Dataset (100 words, >3,300 samples, 6 signers)	3D-CNN (ASL-3DCNN)	Real-time speed: 0.19 s/frame	RGB video	Dynamic isolated words
[20]	Continuous GSL, ASL, CSL corpora	CNN encoders + S2S, Transformer, attention, RL-NMT	BLEU-2 to BLEU-4 scores	RGB video, pose, face, hand landmarks	Sentence-level translation
[21]	Survey of SLR work (2014-2021), multiple datasets	CNN, RNN/LSTM, 3D-CNN, hybrid vision-sensor models	Reported ranges (no single metric)	RGB, depth, skeleton, wearables	Letters, words, sentences, continuous SLR
[22]	ASL alphabet dataset (87,000+ RGB images)	AlexNet, ConvNeXt, EfficientNet, ResNet-50, ViT	ResNet-50: 99.98%; EfficientNet: 99.95%	RGB images	ASL alphabet letters
[23]	Wearable inertial motion capture dataset (300 sentences, 7,400 recordings)	CNN + Bi-LSTM + CTC; LSTM encoder–decoder with attention	99.07% (word); 97.34% (sentence); WER 16.63%	IMU sensors	Continuous sentences and gestures
[24]	Multiple RGB/RGB-D datasets; Greek SLR corpus	CNN + LSTM/BLSTM, CTC-based models	≈90–96% (task-dependent)	RGB, depth, skeleton	Gloss-level and sentence-level recognition
[25]	Custom ASL A-Z static image dataset	Two custom CNNs + Softmax	98.44%	RGB images	ASL alphabet letters
[26]	ASL dataset with letters, words, digits; mixed images and video	KNN, Naive Bayes, SVM, YOLOv5, CNN, LSTM hybrids	Up to 98.01% (SVM + LSTM)	RGB images and video	Letters, short words, digits
[27]	ASL video corpus	Inception CNN + RNN	Accuracy not explicitly reported	RGB video	Sign sequences (words/phrases)
[28]	Multiple SL datasets (alphabets, numbers, words, sentences)	CNN, RNN, hybrid DL, CTC	≈90%+ alphabets; 93.67% mixed tasks	RGB, RGB-D, sensors	Letters, numbers, words, some sentences
[29]	Continuous SLR benchmark datasets (survey)	2D/3D-CNNs, CNN+RNN, Transformers, CTC	N/A (review paper)	RGB, depth, pose, skeleton	Continuous sentence- and gloss-level SLR
Proposed Work	Public ASL datasets (WLASL, MS-ASL)	Multimodal CNN + LSTM + Transformer with ASL grammar token embedding	BLEU score improvement over gloss-based models	RGB video + 2D/3D pose, hand landmarks	Sentence-level ASL-to-English translation, grammar-aware, real-time

Visual Geometry Group 16 (VGG-16), Residual Network with 50 layers (ResNet-50), Modified National Institute of Standards and Technology (MNIST), Self-Normalizing DenseNet with Attention (SNDA), Temporal Graph Convolutional Network (TGCN), Bidirectional Long Short-Term Memory (Bi-LSTM), You Only Look Once (YOLO), American Sign Language 3D Convolutional Neural Network (ASL-3DCNN), Greek Sign Language (GSL), Chinese Sign Language (CSL), Sign Language Recognition (SLR), Connectionist Temporal Classification (CTC), Neural Machine Translation + MediaPipe + Generative Adversarial Network (NMT + MediaPipe + GAN), Reinforcement Learning-based Neural Machine Translation (RL-NMT), Inertial Measurement Unit (IMU), Alex Krizhevsky Network (AlexNet), Convolutional Neural Network Next (ConvNeXt), Efficient Neural Network (EfficientNet), Vision Transformer (ViT), Word Error Rate (WER), Red Green Blue + Depth (RGB-D), K-Nearest Neighbors (KNN), Not Applicable (N/A)

III. METHODOLOGY

The overall proposed system, shown in Figure 1, comprises five main stages: i) multimodal visual feature extraction using YOLO-based hand, face, and body detection, ii) temporal sign representation learning, iii) ASL grammar token inference, iv) grammar-aware sequence translation, and v) sentence-level English text generation.

To develop a set of independent sign video sequences to support sentence-level evaluation and comparison of isolated sign video-to-text translations using the same type of independent sign video sequencing (as WLASL and MS-ASL), we developed a time-sequence based structure representing the sentence in a sequential way via a concatenated sequence of individually annotated word-level sign videos. These sign videos were derived from a predetermined order of words which follow a valid ASL syntactic structure. This is expected to help preserve the inherent temporal ordering of the individual signs within a given sentence's structural sequence of signs. From these three types of video streams (RGB frames, 2D/3D pose keypoints, and hand landmark data) we generated multimodal features at each segmentation point across the total temporal interval of the sign sequence. For each of these segmentations, we aggregated the generated multimodal features into a singular embedding representation for the entire sign sequence. After generating the sign sequence embeddings, they were input into a grammar-aware transformer module that generates intermediate grammar tokens to determine the sentence type, negation, temporal references, and topicalization inherent in the original sentence. Lastly, paired up each of the generated sign sequences with their corresponding reference English sentences to calculate the sentence-level translation quality metrics (BLEU-4, ROUGE-L, METEOR and CIDEr) to validate performance.

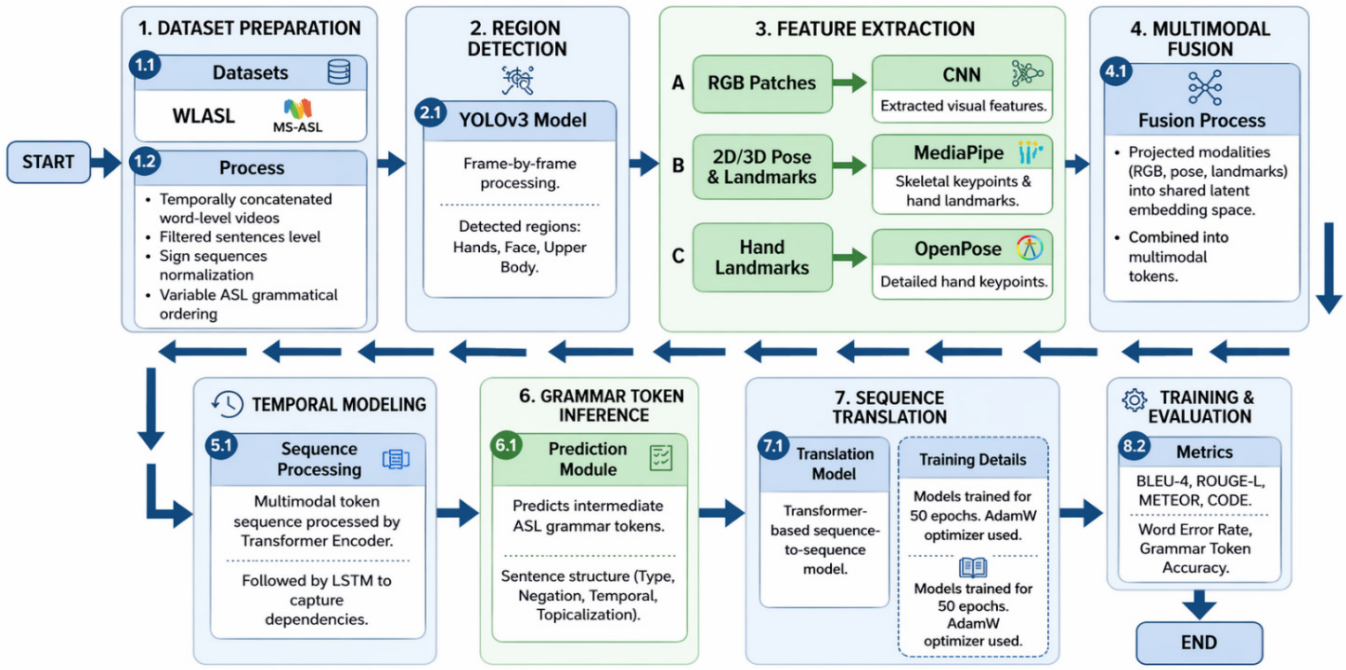

Fig. 1. Proposed grammar-aware multimodal ASL-to-English translation framework.

A. *Multimodal Visual Feature Extraction with YOLO*

The inputs in the model consist of RGB video sequences from two publicly available ASL datasets: WLASL [16] and MS-ASL [30]. WLASL contains large-vocabulary word-level signs (approximately 2,000 words from 119 signers), whereas MS-ASL provides isolated signs with significant signer variation (approximately 1,000 glosses from 222 signers). The use of both datasets ensures diversity in lexical content, signer appearance, and contextual variations. Table II summarizes the dataset specifications.

Each video frame I_t is first processed using YOLOv3 [31] to detect and crop key regions corresponding to the hands, face, and upper body. Let R_t^h, R_t^f, and R_t^b denote the cropped regions of the hands, face, and upper body, respectively:

$$R_t^h, R_t^f, R_t^b = \text{YOLO}(I_t) \quad (1)$$

This region-based cropping reduces background noise and enables subsequent pose and landmark estimation to focus on the most informative visual components for ASL gesture modeling. From each cropped region, 2D/3D body pose keypoints P_t and detailed hand landmarks H_t are extracted using state-of-the-art pose estimation frameworks, such as MediaPipe [32], OpenPose [33], and High-Resolution Network (HRNet) [34], as well as raw network packet datasets [35].

TABLE II. SPECIFICATION OF ASL DATASETS USED FOR TRAINING AND EVALUATION

Dataset	Total videos	Vocabulary size	Signers	Train/validation /test samples	Annotation type
[16]	21,083	2,000 words	119	14,758 / 3,162 / 3,163	Word-level labels
[30]	25,513	1,000 glosses	222	17,860 / 3,826 / 3,827	Isolated sign labels

Raw RGB patches extracted from the YOLO regions R_t are further processed by a CNN to obtain appearance-based features. The modality-specific embeddings are defined as:

$$F_t^{\text{RGB}} = \phi_{\text{CNN}}(R_t)$$
$$F_t^{\text{pose}} = \phi_{\text{pose}}(P_t)$$
$$F_t^{\text{hand}} = \phi_{\text{hand}}(H_t) \quad (2)$$

where F_t^{RGB} represents the RGB frame at time step t, ϕ_{CNN} denotes the CNN-based RGB feature extraction function, ϕ_{pose} denotes the pose feature extraction network, and ϕ_{hand} denotes the hand landmark feature extraction network. Each embedding is projected into a shared latent space:

$$Z_t^m = W^m F_t^m + b^m, m \in \{\text{RGB}, \text{pose}, \text{hand}\} \quad (3)$$

where Z_t^m denotes the projected modality-specific embedding for the modality m, W represents the projection matrix, F_t denotes the feature vector at time step t, and b represents the learnable bias vector associated with modality m.

The concatenated multimodal feature token Z_t, formed by combining RGB, pose, and hand embeddings at time step t, is defined as:

$$Z_t = \left[Z_t^{\text{RGB}}; Z_t^{\text{pose}}; Z_t^{\text{hand}}\right] \quad (4)$$

Positional encoding $\tilde{Z}_t$ is then applied to preserve temporal order:

$$\tilde{Z}_t = Z_t + \text{PE}(t) \quad (5)$$

The temporal sequence of multimodal tokens $\tilde{Z}_{1:T}$ is subsequently fed into a transformer encoder for cross-modal fusion, producing the output sequence H:

$$H = \text{TransformerEncoder}(\tilde{Z}_{1:T}) \quad (6)$$

The self-attention mechanism, implemented as scaled dot-product attention for learnable query Q, key K, and value V, is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V \quad (7)$$

where d_k denotes the dimensionality of the key vectors used in the softmax normalization. The query, key, and value matrices are defined as:

$$Q = HW_Q$$
$$K = HW_K$$
$$V = HW_V \quad (8)$$

The resulting contextualized multimodal representation is then obtained as:

$$\hat{F}_t = H_t \quad (9)$$

The fused multimodal features $\hat{F}_t$ are subsequently passed through an LSTM to model sequential dependencies h_t [12, 20]:

$$h_t = \text{LSTM}(\hat{F}_t, h_{t-1}) \quad (10)$$

The sequence of hidden states $H_{1:T} = [h_1, \ldots, h_T]$ captures the temporal dynamics of ASL gestures. An attention mechanism is then applied to emphasize the most informative time steps:

$$\hat{H} = \sum_{t=1}^{T} \alpha_t h_t, \alpha_t = \text{softmax}(\text{score}(h_t)) \quad (11)$$

where α_t denotes the attention weight at time step t, and $\text{score}(h_t)$ represents the attention scoring function applied to the hidden state h_t.

B. Grammar-Aware Token Inference

ASL grammar tokens are inferred from the temporal representations $H_{1:T}$, capturing sentence type, negation, topicalization, and temporal references. Let $G(\cdot)$ denote the grammar token inference function for the predicted grammar token g_i, which belongs to the grammar token vocabulary V_{grammar}:

$$G = G(H_{1:T}) = [g_1, g_2, \ldots, g_L], g_i \in V_{\text{grammar}} \quad (12)$$

The conditional probability of the grammar token sequence is formulated as:

$$p(G \mid H_{1:T}) = \prod_{i=1}^{L} p(g_i \mid g_1, \ldots, g_{i-1}, H_{1:T}) \quad (13)$$

This structured intermediary representation bridges ASL grammar and English syntax, facilitating grammar-aware translation [16, 19].

C. Grammar-Aware Sequence Translation

Once the ASL grammar token sequence $G = [g_1, g_2, \ldots, g_L]$ is inferred, it is passed to a transformer-based S2S model to generate the output English sentence. The transformer leverages multi-head attention to align ASL grammar tokens with English words while capturing long-range dependencies and grammatical structures. Formally, the translation process is defined through the conditional probability of the output sentence given the grammar tokens ($p(Y \mid G)$):

$$p(Y \mid G) = \prod_{t=1}^{T} p(y_t \mid y_1, \ldots, y_{t-1}, G) \quad (14)$$

where $Y = [y_1, y_2, \ldots, y_T]$ denotes the output English sentence. Each decoder step attends to the entire token sequence G, allowing the model to consider the global grammatical context. Positional encodings are applied to G to preserve token order, and cross-attention layers in the transformer decoder ensure that English word generation respects ASL token structure [18, 20].

Furthermore, a coverage-aware attention mechanism is introduced that allows monitoring and keeping track of which

grammar tokens have received input from the model, thus eliminating redundancy and omission of essential content. This is particularly important in ASL translation, where temporal ordering and coarticulation of gestures significantly affect meaning. Finally, the decoder outputs a probability distribution over the English vocabulary at each time step. The final sentence is generated using either greedy decoding or beam search to ensure grammatical coherence and semantic consistency.

D. Evaluation Metrics

To evaluate the proposed framework, multiple complementary metrics are employed for the overall assessment of the performance of the proposed system's gesture recognition, grammar model, and its ability to translate at the sentence level into English, including:

- BLEU-4 [36, 37], which measures n-gram overlap between predicted and reference sentences, and is a standard metric for S2S translation.
- ROUGE-L [38], which evaluates recall-oriented n-gram overlap and the completeness of the generated translation.
- METEOR [18, 39], which uses both synonymy and stemming to reduce many of the limitations of the BLEU metric, to evaluate semantic equivalence.
- WER [16, 20, 23], which measures word-level insertions, deletions, and substitutions; often used when evaluating continuous ASL recognition systems.
- Sign recognition accuracy [16, 17, 19], which evaluates the correct identification of individual ASL signs.
- Grammar token accuracy [18, 20], which evaluates the accuracy of predicted ASL grammar tokens in order to ensure that the output sentences are properly structured and semantically coherent.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

All experiments were implemented in PyTorch and executed on an NVIDIA RTX 4090 Graphics Processing Unit (GPU). The model was optimized using AdamW with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. A cosine annealing scheduler was applied, and early stopping was triggered based on validation BLEU-4 performance.

Training was conducted for 50 epochs with a batch size of 16. Official train/validation/test splits from WLASL and MS-ASL were used to ensure fair comparisons. Input frames were resized to 224×224, and 2D/3D pose and hand landmarks were extracted using MediaPipe. Baseline models were re-implemented under identical training conditions and hyperparameter settings to maintain consistency. All of the most important training and architecture parameters for the model are depicted in Table III, including learning rate, optimizer options, batch size, number of transformer layers and attention heads, embedding dimensions, dropout probability, and total training epochs.

B. Quantitative Performance Analysis

The overall translation performance of the proposed framework was evaluated against several state-of-the-art baselines. Table IV summarizes the results (BLEU-4, ROUGE-L, METEOR, and CIDEr) on the WLASL and MS-ASL datasets, while Figure 2 illustrates the results on the WLASL dataset.

TABLE III. HYPERPARAMETER CONFIGURATION

Parameter	Value
Optimizer	AdamW
Initial learning rate	1e-4
Weight decay	1e-4
Batch size	16
Epochs	50
Transformer layers	6
Attention heads	8
Hidden dimension	512
Dropout	0.1
Beam size	5
Scheduler	Cosine annealing
Early stopping	Based on validation BLEU-4

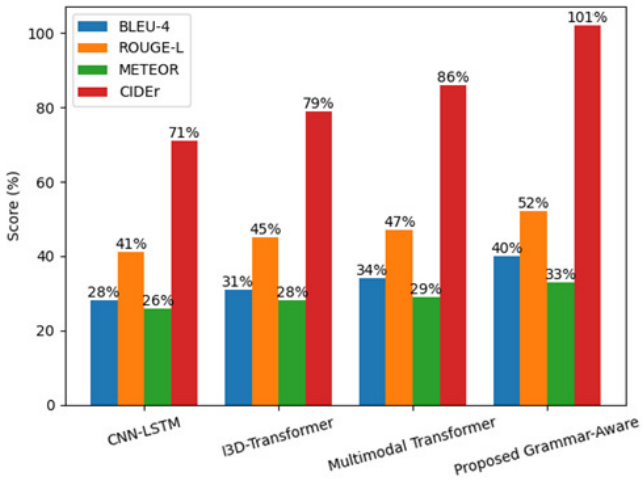

Fig. 2. Sentence-level translation metrics for all evaluated models on the WLASL dataset.

TABLE IV. PERFORMANCE COMPARISON ON WLASL AND MS-ASL DATASETS

Model	WLASL				MS-ASL			
	BLEU-4	ROUGE-L	METEOR	CIDEr	BLEU-4	ROUGE-L	METEOR	CIDEr
CNN-LSTM	28.4	41.2	25.7	71.3	26.9	39.8	24.1	68.5
Inflated 3D Convolutional Network (I3D)-Transformer	31.6	44.5	27.8	79.4	29.8	42.1	26.3	75.2
Multimodal Transformer	34.2	47.3	29.1	85.7	32.5	45.4	27.9	82.3
Proposed Model	39.8	52.6	33.4	101.2	37.5	50.2	31.8	96.4

The proposed grammar-aware model consistently outperforms all baselines across both datasets. Specifically, on WLASL, the model achieved a BLEU-4 improvement of +5.6 over the strongest baseline, whereas on MS-ASL, the improvement reached +5.0 BLEU-4. Consistent improvements were also observed for ROUGE-L, METEOR, and CIDEr, and their statistical significance was verified via paired bootstrap resampling ($p < 0.01$), confirming enhancements in both lexical overlap and semantic adequacy. Furthermore, to interpret the model's focus across different gestures and modalities, attention weights from the transformer-based fusion module are visualized as a heatmap in Figure 3. The heatmap demonstrates that the model dynamically adjusts to critical hand movements and pose transitions during translation, confirming that multimodal fusion effectively captures the temporal dependencies required for grammar-aware ASL-to-English translation.

C. Recognition Performance

In conjunction with the metrics used to assess how well the models translate at the word level, we evaluated them for gesture recognition and grammar token prediction. In addition to those evaluations, we used the same three metrics to evaluate performance as we did when assessing at the word level. Our evaluation results are shown in Figure 4. The result indicated that the proposed grammar-aware model had lower WER than the other three models, better sign recognition, and greater grammar token accuracy, suggesting that a grammar-aware approach can enhance an ASL interpreter's understanding of the ASL words' lexical content and its grammatical structure.

D. Performance Across Sentence Lengths

Figure 5 presents a side-by-side comparison of BLEU-4 scores across the WLASL and MS-ASL datasets, which showcases that the second-best performing model, besides the proposed one, is the multimodal transformer model, serving as the baseline model. To further evaluate the performance of the proposed model over the baseline, their performance was tested across varying sentence lengths, while calculating the BLEU-4 scores in each instance (Table V). The results reveal that improvements for short ASL sequences (1-3 signs) were relatively modest, while for sequences containing more than seven signs, the proposed model achieved a +10.3 BLEU-4 improvement, demonstrating enhanced capability in modeling long-range temporal dependencies and hierarchical sentence structure. This behavior reflects the intended advantage of grammar-aware modeling.

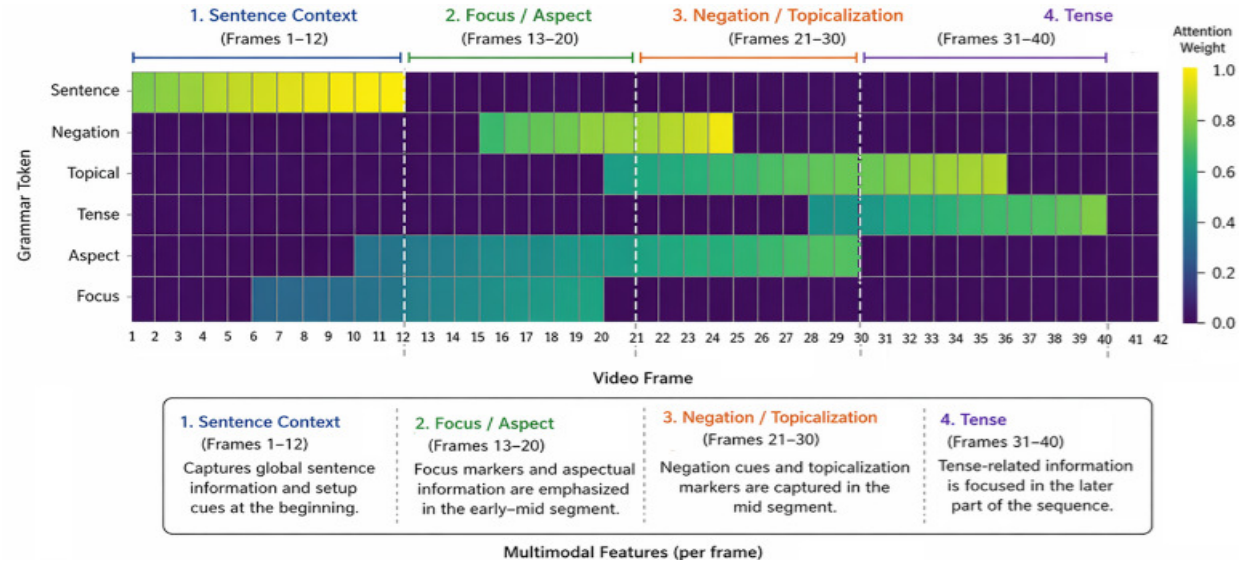

Fig. 3. Heatmap of attention weights from the transformer-based fusion module.

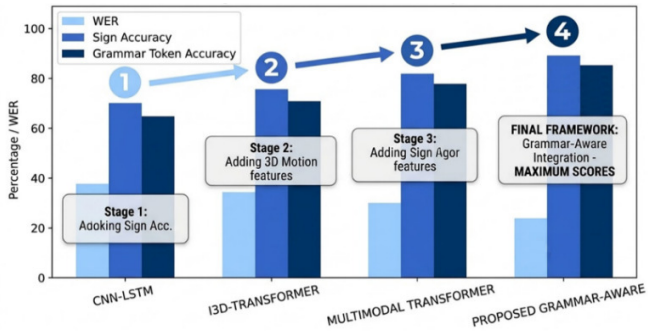

Fig. 4. Gesture recognition and grammar-token prediction metrics across models.

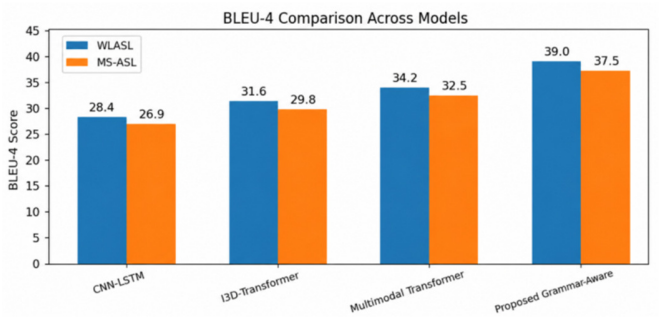

Fig. 5. BLEU-4 score for all evaluated models on the WLASL and MS-ASL dataset.

TABLE V. BLEU-4 PERFORMANCE ACROSS SENTENCE LENGTHS ON WLASL

Sentence length	Multimodal transformer	Proposed	Improvement
1-3 signs	47.2	49.1	+1.9
4-6 signs	36.8	42.5	+5.7
7+ signs	23.4	33.7	+10.3

Additionally, Figure 6 visualizes the attention weights generated by the transformer-based fusion module for a different representative ASL sample consisting of 20 processed frames after temporal preprocessing. Darker shades indicate higher attention weights, emphasizing critical signs required for sentence-level translation. The shorter sequence length reflects natural variability in ASL sentence duration and demonstrates the model's adaptive attention behavior across varying temporal contexts.

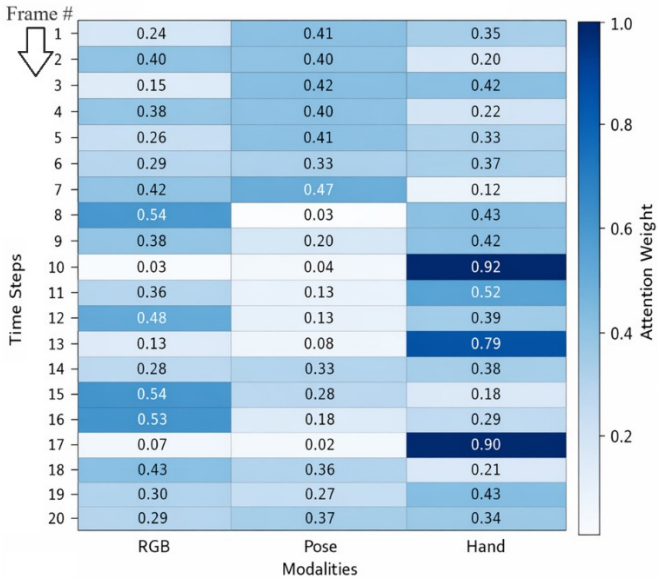

Fig. 6. Transformer attention heatmap over time and modalities.

E. Impact of Grammar Token Modeling

To directly evaluate grammatical understanding, grammar-token prediction accuracy was measured, as summarized in Table VI. Moreover, Figure 7 demonstrates a layered bar representation of correctly and incorrectly classified grammar tokens. Sentence type and negation achieved accuracies above 93%, whereas temporal markers and topicalization remained within the 86-89% range, indicating the relatively higher complexity associated with these grammatical components.

The high classification accuracy across grammatical categories indicates that the model effectively captures structured linguistic representations. However, performance for temporal markers and topicalization was slightly lower, highlighting the difficulty of modeling non-manual cues such as facial expressions and head movements.

TABLE VI. GRAMMAR TOKEN CLASSIFICATION ACCURACY

Grammar category	Accuracy (%)
Sentence type	95.1
Negation	93.6
Temporal markers	89.4
Topicalization	86.8

F. Ablation Study

An ablation study was also conducted on the proposed model to quantify the contribution of each component, as displayed in Table VII. The analysis showed that removing grammar tokens resulted in the highest performance reduction (-5.1 BLEU-4), confirming that explicit linguistic modeling is the primary factor driving translation quality.

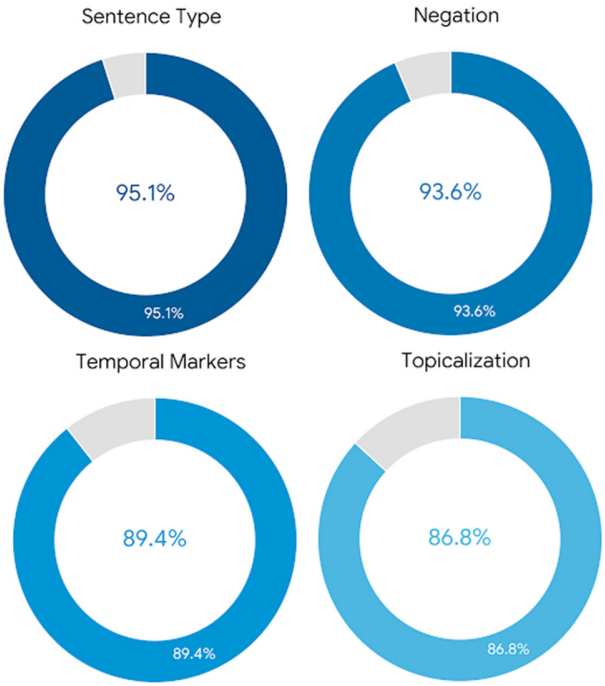

Fig. 7. Grammar token prediction accuracy across categories.

TABLE VII. ABLATION STUDY ON WLASL DATASET

Model variant	BLEU-4
Full model	39.8
Without grammar tokens	34.7
Without hand landmarks	36.2
Without pose stream	37.4

G. Error Analysis and Future Suggestions

Despite high overall performance, the proposed model showcased persisting errors in rapid co-articulation sequences and visually similar handshapes with subtle motion differences. Moreover, grammar-token misclassifications occurred mainly during temporal transitions and topicalization, which depend heavily on non-manual cues. To address these limitations, incorporating explicit facial-expression modeling represents a significant next step for improving grammatical inference.

Beyond algorithmic refinements, the practical deployment of real-time ASL translation systems necessitates robust and secure data transmission, especially when working in assistive environments and on cloud computing platforms, where the integrity of the communications process and its security are extremely important [40]. Consequently, future research must focus on optimizing computational and transmission resources through efficient, adaptive system designs, leveraging adaptive techniques already established in wireless communication [41].

V. CONCLUSION

The proposed grammar-aware multimodal American Sign Language (ASL)-to-English translation framework is designed to translate ASL into English by modeling structured ASL grammar rather than isolated gestures. By combining Red Green Blue (RGB) frame data, 2D and 3D body pose keypoints, and hand landmark information within a transformer-based architecture, the system captures the dynamic spatial and temporal characteristics of ASL gestures. Additionally, the intermediate grammar-token generation stage models key ASL grammatical components, including sentence type, negation, topicalization, and temporal reference, enabling more effective mapping between ASL grammar and English syntactic structure.

The proposed framework demonstrated consistent improvements over all evaluated state-of-the-art baselines on both the Word-Level American Sign Language (WLASL) and Microsoft American Sign Language (MS-ASL) datasets. Specifically, the model achieved Bilingual Evaluation Understudy (BLEU-4) metric improvements of +5.6 on WLASL and +5.0 on MS-ASL. Consistent metric gains were also observed for Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) (+5 to +7), Metric for Evaluation of Translation with Explicit Ordering (METEOR) (+4 to +6), and Consensus-based Image Description Evaluation (CIDEr) (+15 to +20), indicating improved lexical and semantic alignment. Furthermore, grammar-token classification accuracy across sentence type (95.1%), negation (93.6%), temporal markers (89.4%), and topicalization (86.8%) confirms the model's effectiveness in learning structured linguistic patterns. Moreover, ablation studies further demonstrated that grammar-aware modeling contributed the largest performance improvement, while multimodal fusion provided complementary benefits.

Overall, the results demonstrate that explicit grammar modeling combined with multimodal temporal representations can substantially improve the accuracy and coherence of ASL-to-English translation. Future work may incorporate additional non-manual cues, including facial expressions and head movements, and extend the framework toward real-time translation systems to further improve accessibility and communication for the Deaf community.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no competing interests that could have influenced the work reported in this manuscript.

ACKNOWLEDGEMENT

Not applicable to this work.

DATA AVAILABILITY

The datasets generated and/or analyzed during the current study are available in [16, 31].

REFERENCES

[1] W. C. Stokoe, "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf," *Journal of Deaf Studies and Deaf Education*, vol. 10, no. 1, pp. 3–37, Jan. 2005, https://doi.org/10.1093/deafed/eni001.

[2] R. Pfau, M. Steinbach, and B. Woll, Eds., *Sign Language: An International Handbook*. DE GRUYTER, 2012.

[3] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998, https://doi.org/10.1109/34.735811.

[4] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, Dec. 2015, https://doi.org/10.1016/j.cviu.2015.09.013.

[5] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video," *International Journal of Computer Vision*, vol. 126, no. 2–4, pp. 430–439, Apr. 2018, https://doi.org/10.1007/s11263-016-0957-7.

[6] T. A. Patil et al., "Real-Time American Sign Language Recognition System Using Deep Learning and Computer Vision," *International Journal of Scientific Research in Engineering and Management*, vol. 09, no. 06, pp. 1–9, June 2025, https://doi.org/10.55041/IJSREM50739.

[7] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural Sign Language Translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 7784–7793, https://doi.org/10.1109/CVPR.2018.00812.

[8] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 10020–10030, https://doi.org/10.1109/CVPR42600.2020.01004.

[9] S. K. Liddell, *Grammar, Gesture, and Meaning in American Sign Language*, 1st ed. Cambridge University Press, 2003.

[10] R. Zuo and B. Mak, "Improving Continuous Sign Language Recognition with Consistency Constraints and Signer Removal." arXiv, 2022, https://doi.org/10.48550/ARXIV.2212.13023.

[11] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, Louisiana, USA, 2018.

[12] B. Alsharif, E. Alalwany, A. Ibrahim, I. Mahgoub, and M. Ilyas, "Real-Time American Sign Language Interpretation Using Deep Learning and Keypoint Tracking," *Sensors*, vol. 25, no. 7, Mar. 2025, Art. no. 2138, https://doi.org/10.3390/s25072138.

[13] S. Shekhar, "Real-time Sign Language to Text Conversion using Deep Learning Models," in *2024 3rd International Conference for Advancement in Technology (ICONAT)*, Sept. 2024, pp. 1–7, https://doi.org/10.1109/ICONAT61936.2024.10774717.

[14] V. S. Ganesh Reddy, B. AnkammaRao, C. Manvitha, K. Priyanka, V. Phani Kumar Sistla, and V. K. Kishore Kolli, "Performance Evaluation of Various Deep Learning Models for Sign Language Recognition," in *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, Feb. 2025, pp. 189–194, https://doi.org/10.1109/ESIC64052.2025.10962660.

[15] E. Hassan, M. Y. Shams, T. Abd El-Hafeez, and M. Elseddik, "A novel model for expanding horizons in sign Language recognition," *Scientific Reports*, vol. 15, no. 1, July 2025, Art. no. 24358, https://doi.org/10.1038/s41598-025-09643-2.

[16] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 1448–1458, https://doi.org/10.1109/WACV45572.2020.9093512.

[17] A. Kasapbaşi, A. E. A. Elbushra, O. Al-Hardanee, and A. Yilmaz, "DeepASLR: A CNN based human computer interface for American Sign Language recognition for hearing-impaired individuals," *Computer Methods and Programs in Biomedicine Update*, vol. 2, 2022, Art. no. 100048, https://doi.org/10.1016/j.cmpbup.2021.100048.

[18] B. Natarajan *et al.*, "Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation," *IEEE Access*, vol. 10, pp. 104358–104374, 2022, https://doi.org/10.1109/ACCESS.2022.3210543.

[19] S. Sharma and K. Kumar, "ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 26319–26331, July 2021, https://doi.org/10.1007/s11042-021-10768-5.

[20] T. Ananthanarayana *et al.*, "Deep Learning Methods for Sign Language Translation," *ACM Transactions on Accessible Computing*, vol. 14, no. 4, pp. 1–30, Dec. 2021, https://doi.org/10.1145/3477498.

[21] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," *IEEE Access*, vol. 9, pp. 126917–126951, 2021, https://doi.org/10.1109/ACCESS.2021.3110912.

[22] B. Alsharif, A. S. Altaher, A. Altaher, M. Ilyas, and E. Alalwany, "Deep Learning Technology to Recognize American Sign Language Alphabet," *Sensors*, vol. 23, no. 18, Sept. 2023, Art. no. 7970, https://doi.org/10.3390/s23187970.

[23] Y. Gu, H. Oku, and M. Todoh, "American Sign Language Recognition and Translation Using Perception Neuron Wearable Inertial Motion Capture System," *Sensors*, vol. 24, no. 2, Jan. 2024, Art. no. 453, https://doi.org/10.3390/s24020453.

[24] N. Adaloglou *et al.*, "A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1750–1762, 2022, https://doi.org/10.1109/TMM.2021.3070438.

[25] P. Rakshit, S. Paul, and S. Dey, "Sign language detection using convolutional neural network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, no. 4, pp. 2399–2424, Apr. 2024, https://doi.org/10.1007/s12652-024-04761-7.

[26] N. Shanthi, C. Sharmila, M. Muthuraja, S. Janupritha, P. Kavin, and J. Keerthi, "Unveiling the Power of Machine Learning and Deep Learning in Advancing American Sign Language Recognition," in *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, Apr. 2024, pp. 360–369, https://doi.org/10.1109/ICC-ROBINS60238.2024.10534028.

[27] K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 4896–4899, https://doi.org/10.1109/BigData.2018.8622141.

[28] A. Sultan, W. Makram, M. Kayed, and A. A. Ali, "Sign language identification and recognition: A comparative study," *Open Computer Science*, vol. 12, no. 1, pp. 191–210, May 2022, https://doi.org/10.1515/comp-2022-0240.

[29] A. Khan *et al.*, "Deep Learning Approaches for Continuous Sign Language Recognition: A Comprehensive Review," *IEEE Access*, vol. 13, pp. 55524–55544, 2025, https://doi.org/10.1109/ACCESS.2025.3554046.

[30] H. Vaezi Joze and O. Koller, "MS-ASL: a large-scale data set and benchmark for understanding american sign language," in *The British Machine Vision Conference (BMVC)*, Sept. 2019.

[31] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement." arXiv, Apr. 2018, https://doi.org/10.48550/arXiv.1804.02767.

[32] C. Lugaresi *et al.*, "MediaPipe: a framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[33] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, Jan. 2021, https://doi.org/10.1109/TPAMI.2019.2929257.

[34] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021, https://doi.org/10.1109/TPAMI.2020.2983686.

[35] H. A. Al-Ofeishat *et al.*, "Analysis and Comparison of Raw Network Packet Datasets Using Machine Learning Classification and Grey Wolf Optimization," *International Journal of Advances in Soft Computing and its Applications*, vol. 17, no. 1, Mar. 2025, https://doi.org/10.15849/IJASCA.250330.13.

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001, Art. no. 311, https://doi.org/10.3115/1073083.1073135.

[37] M. Zouidine and M. Khalil, "Large Language Models for Arabic Sentiment Analysis and Machine Translation," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20737–20742, Apr. 2025, https://doi.org/10.48084/etasr.9584.

[38] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Text Summarization Branches Out*, July 2004, pp. 74–81.

[39] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, June 2005, pp. 65–72.

[40] A. Manasreh, A. A. M. Sharadqh, J. S. Alkasassbeh, and A. Al-Qaisi, "Ensuring telecommunication network security through cryptology: a case of 4G and 5G LTE cellular network providers," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, Dec. 2019, Art. no. 4860, https://doi.org/10.11591/ijece.v9i6.pp4860-4865.

[41] J. S. Alkasassbeh, A. K. Al-Qaisi, M. Al-Hunaity, and J. Alkasassbeh, "Maximize Saving Transmitted Power in Wireless Communication System Using Adaptive Modulation Technique," *International Journal on Communications Antenna and Propagation (IRECAP)*, vol. 6, no. 2, Apr. 2016, Art. no. 61, https://doi.org/10.15866/irecap.v6i2.8486.