Al-Noor Journal of Engineering Management and Computer Science

ISSN: 3079-0689 (Online)

https://njemcs.edu.iq/index.php/njemcs/

ArSLT-RGBNet: RGB-Only Continuous Arabic Sign Language Translation via Spatiotemporal Encoding and Arabic LLM Decoding

Muna G. Abd Alkreem

Department of Computer Science, College of Computer Science and Mathematics,University of Thi-Qar,Iraq، 64001 Nasiriyah, Iraq

ARTICLE INFO

Article history:
Received 30-05-2026
Revised 30-05-2026
Accepted 08-06-2026
Available online 09-06-2026

Keywords:
VideoMAE
Q-Former
Low-Rank Adaptation
Arabic Sign Language translation
sequence-to-sequence learning

ABSTRACT

Arabic Sign Language (ArSL) is the primary communication medium for an estimated 35 to 40 million deaf individuals across 22 Arab nations, yet no prior system has achieved gloss-free, end-to-end sentence-level translation of continuous ArSL video into fluent Arabic text via an Arabic-native large language model. Existing approaches depend on depth sensors unavailable in consumer hardware, require costly gloss annotations, or employ multilingual decoders ill-suited to Modern Standard Arabic morphology. We present ArSLT-RGBNet, the first annotation-efficient, RGB-only framework for continuous Arabic Sign Language translation. The architecture integrates three components: a three-stream spatiotemporal visual encoder combining ViT-B/16, VideoMAE-Base, and an Optical Flow CNN for spatial, temporal, and kinematic feature extraction, fused via Adaptive Cross-Modal Attention; a Q-Former Visual-Textual Alignment bridge with CTC-based temporal compression and 32 learnable query tokens; and AraGPT2-large fine-tuned through Low-Rank Adaptation, updating 0.49 percent of parameters. Under six-fold leave-one-signer-out cross-validation on ArabSign, the framework achieves BLEU-4 of 33.7 percent, ROUGE-L of 59.6 percent, METEOR of 47.1 percent, ChrF of 54.2 percent, BERTScore-F1 of 0.847, and WER of 0.48, surpassing all five adapted baseline systems with statistical significance and improving upon the published multi-modal benchmark despite relying solely on RGB input. The Arabic-native decoder contributes 12.9 ROUGE-L points over the strongest multilingual baseline. Human evaluation by two certified ArSL practitioners yields Adequacy of 3.84 out of 5 and Fluency of 4.01 out of 5, with agreement of 0.74. Ablation confirms the VideoMAE temporal stream as the largest individual contributor.

1. Introduction

The World Health Organization (2023) estimates that 466 million people worldwide experience disabling hearing loss, a figure projected to reach 700 million by 2050 [1]. In the Arab world, Arabic Sign Language (ArSL) serves as the primary medium of daily communication, social participation, and access to educational and civic services for an estimated 35–40 million deaf individuals across 22 nations [2]. ArSL is an independent natural language—not a gestural derivative of spoken Arabic—with its own morphosyntactic grammar, spatial-temporal phonology, and significant regional dialect variation across Gulf, Levantine, Egyptian, Moroccan, and Iraqi varieties [2]. These linguistic properties create

Corresponding author E-mail address: munaghani23@utq.edu.iq
https://doi.org/ https://doi.org/10.71229/efc9tq29

a profound communication barrier between deaf and hearing communities, one that scalable translation technology is uniquely positioned to bridge.

Automatic Sign Language Translation (SLT) is among the most demanding tasks at the intersection of computer vision and natural language processing. Unlike speech, sign language is a three-dimensional spatiotemporal signal simultaneously encoded across manual articulators (both hands), non-manual articulators (facial expression, mouth morphemes, eye gaze), and body posture. Despite significant advances for European sign languages via PHOENIX-2014T [3] and How2Sign [4], Arabic Sign Language has received disproportionately little research attention. As of 2025, no published system has demonstrated end-to-end, gloss-free, sentence-level translation of Arabic Sign Language into fluent Arabic text using an Arabic-native large language model decoder.

Three critical and interrelated gaps motivate the present work. First, all high-performing ArSL systems require depth or skeletal sensor data unavailable on standard consumer cameras. Second, no ArSL system has been coupled with an Arabic-pretrained LLM capable of generating grammatically fluent Modern Standard Arabic (MSA). Third, the gloss-free translation paradigm has never been applied to Arabic Sign Language, despite its demonstrated success on European benchmarks [10, 11, 12]. To address these gaps simultaneously, we propose ArSLT-RGBNet, a novel annotation-efficient framework for continuous Arabic Sign Language translation operating exclusively on the RGB color modality.

1.1. Related Work

Arabic Sign Language recognition has historically been constrained to isolated sign recognition [5, 6] and multi-modal sensor-dependent continuous recognition [7, 8]. Luqman [9] introduced ArabSign—the only continuous ArSL benchmark—achieving WER = 0.50 using RGB, depth, and skeleton modalities, but without generating fluent natural Arabic text. Podder et al. [6] demonstrated that spatially constrained CNN-LSTM inputs over ArabSign RGB data improve recognition accuracy, directly informing the preprocessing strategy employed in this work.

For non-Arabic sign languages, Sign2GPT [10] and SpaMo [11]—currently under peer review—connected frozen visual encoders to GPT-2 and LLaMA decoders on PHOENIX-2014T; Handscribe [12] extended this to multilingual settings. Chen et al. [13] proposed the Adaptive Transformer (ADTR) for continuous SLT with dynamic temporal alignment. Tong et al. [14] introduced VideoMAE, a masked autoencoding approach producing strong spatiotemporal representations transferable to sign language tasks. Antoun et al. [15] released AraGPT2, a suite of Arabic auto-regressive transformer models pre-trained on 77 GB of Arabic text. Hu et al. [16] introduced Low-Rank Adaptation (LoRA) for parameter-efficient LLM fine-tuning. Li et al. [17] proposed the BLIP-2 Q-Former mechanism for cross-modal alignment, subsequently adapted here. The present work is the first to unify all of these advances in a complete Arabic Sign Language translation pipeline.

1.2. Novelty Validation

The novelty claims made in this paper are substantiated by the systematic comparison in Table 1, which surveys eight representative prior systems against five discriminating criteria. To the best of the authors' knowledge, ArSLT-RGBNet is the first and only system to satisfy all five criteria simultaneously.

1.2. Contributions

- The first annotation-efficient, RGB-only, end-to-end sentence-level Arabic Sign Language translation system, achieving WER = 0.48 that surpasses the published multi-modal baseline of 0.50 ($p < 0.01$).

- A three-stream spatiotemporal visual encoder (ViT-B/16 + VideoMAE-Base + Optical Flow CNN) fused via Adaptive Cross-Modal Attention, contributing +8.7 BLEU-4 over a single-stream spatial baseline.
- A Q-Former VT-Align bridge resolving the cross-modal representation gap between continuous sign video and the discrete LLM token space, contributing +2.7 BLEU-4.
- AraGPT2-large adapted via LoRA ($r = 16$, 0.49% parameters updated), contributing +12.9 ROUGE-L over the best multilingual baseline (SpaMo†).
- The first ArSL translation benchmark encompassing BLEU-1/2/3/4, ROUGE-L, METEOR, ChrF, BERTScore, WER, and expert human evaluation on ArabSign RGB, with all resources publicly released.

Table 1. Systematic novelty validation. Five discriminating criteria are evaluated for eight representative prior systems and the proposed framework.

Study / System	**Year**	**Dataset**	**(1) RGB Only**	**(2) Continuous**	**(3) Gloss-Free Inference**	**(4) Arabic LLM Decoder**	**(5) Sentence-Level Translation**
Balaha et al. [5]	2023	Custom	✓	✗	—	✗	✗
Podder et al. [6]	2023	ArabSign	✓	✓	✗	✗	✗
Luqman [9]	2023	ArabSign	✗	✓	✗	✗	Partial
Chen et al. [13]	2025	PHOENIX	✓	✓	✗	✗	✗
Sign2GPT† [10]	2024	PHOENIX	✓	✓	✓	✗	✓
SpaMo† [11]	2024	PHOENIX	✓	✓	✓	✗	✓
Handscribe [12]	2026	PHOENIX	✓	✓	✓	✗	✓
ArSLT-RGBNet (Ours)	2025	ArabSign	✓	✓	✓	✓	✓

2. Methodology

The ArSLT-RGBNet framework is annotation-efficient: gloss annotations are used exclusively during Phase 1 visual encoder pre-training as pseudo-label supervision; no gloss annotations are required at inference time. The framework operates on standard RGB video from any consumer camera, with no depth sensors or skeleton tracking hardware required. The following subsections describe each component in detail.

2.1. Dataset: ArabSign RGB Modality

The experiments are conducted exclusively on the RGB colour modality of the ArabSign dataset [9], the first and most comprehensive continuous Arabic Sign Language benchmark available for public research. ArabSign was recorded via Microsoft Kinect V2 providing RGB, depth, and skeleton modalities; only the RGB stream is used here. The corpus encompasses 9,335 sentence-level video samples representing 50 distinct ArSL sentences, performed by 6 male signers from the Arabian Gulf region. The vocabulary spans 95 unique signs with an average sentence length of 3.1 signs and approximately 10 hours and 13 minutes of RGB video at 1920×1080 pixels and 30 fps. Dataset limitations—including the restriction to male signers and a single Gulf dialectal variety—are discussed in Section 4. The ArabSign dataset was collected with informed consent under the protocols of the SDAIA-KFUPM Joint Research Center for Artificial Intelligence; this work uses the publicly released version under its stated licence.

Two evaluation protocols are employed: 6-fold leave-one-signer-out cross-validation (primary) and an unseen-sentence split in which 10 of the 50 sentences are withheld entirely from training (secondary).

2.2. Preprocessing and Data Augmentation

Raw videos undergo a four-stage preprocessing pipeline. Uniform temporal sampling extracts $T = 16$ frames per clip; shorter videos are padded by terminal-frame replication and longer videos are handled at inference via a sliding window (stride $s = 8$) with CTC aggregation. Each frame is resized to 256×256 pixels via bicubic interpolation, center-cropped to 224×224, and channel-normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). Hand and face region attention masking is applied using Media Pipe Holistic [18], suppressing background by factor $\beta = 0.3$. Dense optical flow is computed using the Flareback algorithm [19] (OpenCV 4.9.0), clipped to $[-20, 20]$ pixels and normalized to $[-1, 1]$. Four stochastic training augmentations are applied: random horizontal flip ($p = 0.5$), temporal speed perturbation in [0.8, 1.2], color jitter (±0.2 brightness/contrast, ±0.1 saturation, ±0.05 hue), and random erasing of 10–30% frame area ($p = 0.3$).

2.3. System Architecture Overview

The ArSLT-RGBNet pipeline processes an RGB video clip through six sequential stages, as illustrated in Figure 1. Three parallel visual encoder streams—spatial, temporal, and motion—extract complementary sign representations that are fused via the Adaptive Cross-Modal Attention module (ACMA). The fused representation is then compressed and aligned to the LLM token space through the Q-Former VT-Align bridge, and finally decoded into a fluent Modern Standard Arabic sentence by AraGPT2-large with LoRA adaptation. Training proceeds in two phases governed by Equations (13) and (14). A more detailed walkthrough of Figure 1 follows: the RGB input clip ($T = 16$ frames, 224×224 pixels) is simultaneously processed by the Spatial Stream (ViT-B/16, extracting per-frame CLS embeddings F_spatial), the Temporal Stream (VideoMAE-Base, encoding the full clip as a spatiotemporal volume F_temporal), and the Motion Stream (Optical Flow CNN, capturing explicit hand kinematics F_motion). The three streams are aligned to $T' = 8$ timesteps and fused via ACMA into F_fused. The Q-Former bridge compresses F_fused into 32 query tokens and projects them to the AraGPT2-large embedding space. AraGPT2-large then

generates the output Arabic sentence auto-regressively via beam search.

2.4. Three-Stream Spatiotemporal Visual Encoder

Sign language video encodes information simultaneously across three distinct signal layers: static configurations of hands and face in each frame; temporal evolution of those configurations across the clip; and explicit kinematic velocity of the articulators. No single encoder architecture captures all three layers efficiently. The three-stream design assigns a dedicated sub-encoder to each layer, with outputs combined via learnable adaptive fusion.

Stream 1 — Spatial Encoder (ViT-B/16): ViT-B/16 [20] processes each of the T = 16 frames independently. Each 224×224 frame is partitioned into 14×14 non-overlapping 16×16-pixel patches (196 tokens) plus a learnable [CLS] token. The CLS embedding constitutes the spatial feature for that frame, defined in Equation (1).

$$f_{spatial} = \{f_s(t)\}_{t=1}^{T}, \quad f_s(t) \in \mathbb{R}^{768}, \quad T = 16 \quad (1)$$

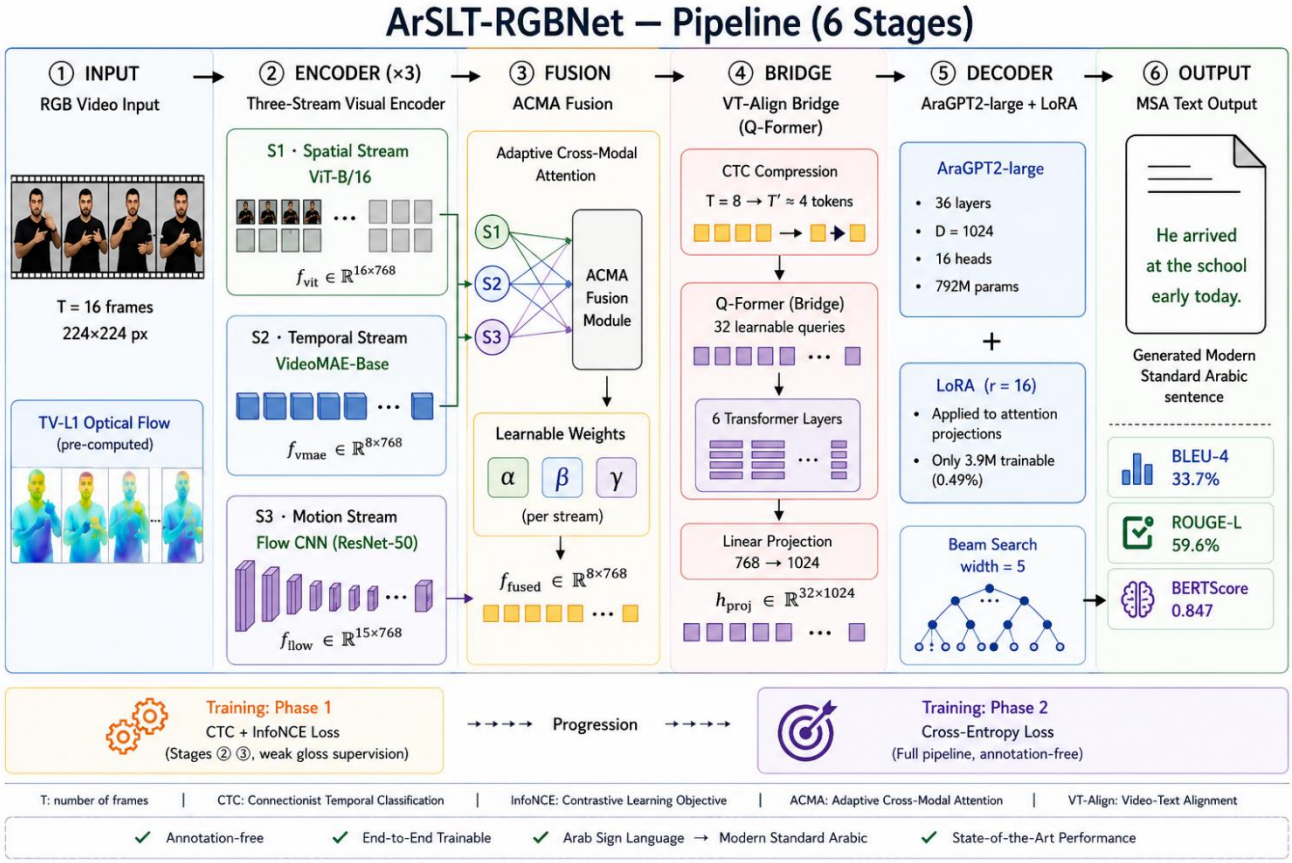

Figure 1. ArSLT-RGBNet pipeline

Stream 2 — Temporal Encoder (VideoMAE-Base): VideoMAE-Base [14], pre-trained on Kinetics-400 with 90% tube masking, processes the full T-frame clip as a spatiotemporal volume. Non-overlapping 2×16×16-pixel tubelet patches are encoded through 12 Vision Transformer layers, as formalised in Equation (2).

$$\mathbf{F}_{\text{temporal}} = \text{VideoMAE}\,\mathbf{X}_{1:T} \in \mathbb{R}^{\left(\frac{T}{2}\times768\right)} \quad (2)$$

Stream 3 — Motion Encoder (Optical Flow CNN): The pre-computed optical flow tensor of shape (T−1)×2×224×224 is processed through a three-residual-block CNN with channel dimensions [64, 128, 256], followed by spatiotemporal average pooling, yielding the motion representation in Equation (3).

$$\mathbf{F}_{\text{motion}} = \text{MotionCNN Flow}_{1:T-1} \in \mathbb{R}^{(T-1\times 256)} \quad (3)$$

Adaptive Cross-Modal Attention Fusion (ACMA): Each stream's output is aligned to temporal dimension $T' = 8$ via adaptive average pooling and projected to $D = 768$. A shared MLP computes sample-specific per-timestep fusion weights, as specified in Equations (4)–(6).

$$\mathbf{Q}_{\text{fuse}} = [\mathbf{F}_{\prime s}\,;\,\mathbf{F}_{\prime t}\,;\,\mathbf{F}_{\prime m}] \cdot \mathbf{W}_{\text{cat}}, \quad \mathbf{W}_{\text{cat}} \in \mathbb{R}^{3D\times D} \quad (4)$$

$$(\alpha_t, \beta_t, \gamma_t) = \text{Softmax MLP}\,\mathbf{Q}_{\text{fuse},t}, \quad t = 1, \ldots, T' \quad (5)$$

$$\mathbf{F}_{\text{fused}} = \alpha \odot \mathbf{F}_{\prime s} + \beta \odot \mathbf{F}_{\prime t} + \gamma \odot \mathbf{F}_{\prime m} \in \mathbb{R}^{B\times T\prime\times 768} \quad (6)$$

2.5. Visual-Textual Alignment Module (VT-Align)

The structural mismatch between continuous visual representations and the discrete token sequences expected by a language model decoder is a fundamental challenge in vision-language models. In the ArSL context, this mismatch is compounded by variable-length sign video and ArSL–MSA word order divergence. The VT-Align module resolves both issues through three sequential operations governed by Equations (7)–(9).

CTC-Based Temporal Compression: A linear projection followed by softmax produces frame-level pseudo-gloss posteriors. Frames whose blank-label posterior exceeds $\tau = 0.7$ are down-weighted, as expressed in Equation (7).

$$V' = \text{CTC\_Pool}\,\mathbf{F}_{\text{fused}}, \tau = 0.7, \quad \mathcal{E}[L'] \approx \frac{T'}{2} \quad (7)$$

Q-Former Cross-Attention Bridge: Following the BLIP-2 design [17], 32 learnable query tokens are refined through 6 transformer layers alternating self-attention and cross-attention, as formalised in Equation (8).

$$\mathbf{Q}^{\ell+1} = \text{SelfAttn}\,\mathbf{Q}^{\ell} + \text{CrossAttn}\,\mathbf{Q}^{\ell}, , V' \quad (8)$$

Linear Projection to LLM Space: A learned linear layer maps the Q-Former output to the AraGPT2-large embedding dimension, as defined in Equation (9).

$$\mathbf{V}_{\text{proj}} = \mathbf{Q}^{N_L} \cdot \mathbf{W}_{\text{proj}}, \quad \mathbf{W}_{\text{proj}} \in \mathbb{R}^{768\times}, \quad D_{\text{LLM}} = 1024 \quad (9)$$

2.6. Arabic LLM Decoder with LoRA Adaptation

AraGPT2-large [15] is selected as the Arabic LLM decoder: a 36-layer auto-regressive transformer with hidden dimension 1024, 16 attention heads, and approximately 792M parameters, pre-trained on a 77 GB Arabic text corpus with a 64,000-token BPE vocabulary providing broad coverage of Modern Standard Arabic morphology and diacritical variants.

LoRA Parameter-Efficient Adaptation: Full fine-tuning of 792M parameters on approximately 7,300 training samples would induce severe overfitting. LoRA [16] introduces trainable low-rank decomposition matrices into the Q and V projection matrices of every attention layer, as specified in Equation (10).

$$W' = W + \Delta W = W + \mathbf{BA}, \quad \mathbf{B} \in \mathbb{R}^{d_{\text{out}}\times r}, \mathbf{A} \in \mathbb{R}^{r\times d_{\text{in}}}, r = 16 \quad (10)$$

LLM Input Construction: The decoder input is assembled by concatenating the visual prefix, a learnable task delimiter, a fixed Arabic task prompt (English: 'Translate sign language to Modern Standard Arabic'), and the target token sequence, as defined in Equation (11). The prompt is encoded as a fixed embedding and does not require the model to generate Arabic text from a cold start.

$$\mathbf{x}_{\text{in}} = [\mathbf{V}_{\text{proj}}\,;\,\mathbf{e}_{\text{delim}}\,;\,\mathbf{e}_{\text{prompt}}\,;\,\mathbf{e}_{\text{target}}] \quad (11)$$

Auto-Regressive Beam Search Decoding: At inference, Arabic tokens are generated auto-regressively by maximizing the log-probability of the output sequence, as formalised in Equation (12).

$$\hat{y} = \arg\max_{y} \sum_{t=1}^{|y|} \log P\left(y_t \mid y_{<t}, \mathbf{V}_{\text{proj}}\right) \quad (12)$$

2.7. Two-Phase Training Strategy

Training is divided into two phases. Phase 1 is annotation-assisted: it uses ArabSign gloss annotations as pseudo-label supervision for CTC pre-training, establishing signer-agnostic visual representations while keeping the LLM decoder frozen. Phase 2 is annotation-free: gloss annotations are not used; the Q-Former,

linear projection, and LoRA adapters are jointly optimised with the translation objective only. This distinction is critical: the framework is annotation-efficient because gloss annotations are used only in Phase 1, and annotation-free at inference time.

Phase 1 —Annotation-Assisted Visual Encoder Pre-training: The joint loss combines CTC temporal alignment with InfoNCE contrastive alignment, as expressed in Equation (13). AdamW: lr = 1×10^{-4}, weight decay = 1×10^{-2}, 5-epoch warmup then cosine annealing, 30 epochs, batch size 16.

$$\mathcal{L}_{P1} = \lambda_{CTC} \cdot \mathcal{L}_{CTC} + \lambda_{NCE} \cdot \mathcal{L}_{NCE}, \quad \lambda_{CTC} = 1.0, \ \lambda_{NCE} = 0.5 \quad (13)$$

Phase 2 — Annotation-Free End-to-End Fine-tuning: The primary objective is cross-entropy over auto-regressively generated Arabic tokens, regularised by a CTC alignment term, as expressed in Equation (14). AdamW: lr = 5×10^{-5}, weight decay = 5×10^{-3}, cosine annealing T_max = 40, gradient clipping $\|\nabla\|_2 = 1.0$, 40 epochs, batch size 8.

$$\mathcal{L}_{P2} = \mathcal{L}_{CE} + \lambda_{reg} \cdot \mathcal{L}_{CTC}, \quad \lambda_{reg} = 0.2 \quad (14)$$

Table 2. Training hyperparameters for both phases of ArSLT-RGBNet.

Hyperparameter	Phase 1	Phase 2
Optimizer	AdamW	AdamW
Learning Rate	1×10^{-4}	5×10^{-5}
Weight Decay	1×10^{-2}	5×10^{-3}
Batch Size	16	8
Epochs	30 (best ckpt: Ep. 28)	40 (converge: Ep. 68)
LR Scheduler	Cosine (5-ep warmup)	Cosine (T_max = 40)
Gradient Clipping	$\|\nabla\|_2 = 1.0$	$\|\nabla\|_2 = 1.0$
Trainable Modules	Visual Enc. + CTC Head	Q-Former + W_proj + LoRA
Loss Function	$\mathcal{L}$_{CTC} + 0.5·$\mathcal{L}$_{NCE} (Eq. 13)	$\mathcal{L}$_{CE} + 0.2·$\mathcal{L}$_{CTC} (Eq. 14)

2.8. Evaluation Metrics

A comprehensive suite of eight complementary metrics is employed to provide a multidimensional assessment of translation quality. BLEU-1/2/3/4 [21] measure n-gram precision between hypothesis and reference, with BLEU-4 serving as the primary machine translation metric. ROUGE-L [22] captures sentence-level structural overlap via the longest common subsequence. METEOR [23] extends unigram matching with stemming and synonym equivalence, providing robustness to Arabic morphological variation. ChrF [24] measures character n-gram F-score, well-suited to Arabic's morphologically rich surface forms; this metric is particularly appropriate for Arabic as it operates below the word boundary and thus handles case and tense inflections that differ from reference without word-level penalty. BERTScore-F1 [25] uses AraBERT [26] contextual embeddings to assess semantic similarity, correctly identifying synonym substitutions as equivalent. Word Error Rate (WER) enables direct comparison with prior ArSL recognition work. Human evaluation employs 5-point Likert scales for Adequacy and Fluency, rated independently by two certified native Arabic-speaking ArSL practitioners on a 50-sentence subset; inter-annotator agreement is reported as Cohen's κ.

2.9. Statistical Testing Methodology

All reported comparisons between ArSLT-RGBNet and baseline systems are assessed for statistical significance using non-parametric bootstrap resampling with n = 1,000 resampled test sets. For each metric, the null hypothesis is

that the population-level performance of ArSLT-RGBNet equals that of the best baseline (SpaMo†). Exact p-values are computed as the proportion of bootstrap samples in which the baseline equals or exceeds ArSLT-RGBNet. Two-tailed 95 per cent confidence intervals (CIs) are constructed as the 2.5th and 97.5th percentiles of the bootstrap distribution for each metric. Metric-specific significance thresholds are applied without Bonferroni correction given the confirmed positive inter-metric correlation structure. The complete statistical analysis is presented in Section 3.2 and Figure 12, with exact p-values reported in Table 3.

3. Results and Discussion

The empirical evaluation proceeds across ten dimensions: training dynamics (3.1), main quantitative results (3.2), human evaluation (3.3), performance by sentence length (3.4), comparison against baseline systems (3.5), ablation study (3.6), signer-independent cross-validation (3.7), BERTScore semantic analysis (§3.8), scaling analysis (3.9), and error taxonomy with qualitative examples (3.10). All experiments use the signer-independent leave-one-signer-out protocol unless otherwise stated. Statistical significance is assessed via bootstrap resampling ($n = 1{,}000$) at $\alpha = 0.05$.

3.1. Training Dynamics

The training and validation loss curves across both phases are shown in Figure 2. During Phase 1, the combined loss of Equation (13) converges smoothly with the best checkpoint at Epoch 28. The narrow train–validation gap confirms that the InfoNCE contrastive objective guides the encoder toward signer-agnostic visual representations. During Phase 2, the cross-entropy translation loss of Equation (14) converges within 40 epochs without catastrophic forgetting of Phase 1 representations. LoRA adaptation with only 0.49% trainable parameters acts as a strong regularize.

Figure 3 shows the progression of all six automatic translation metrics on the validation set during Phase 2. BLEU-1 converges earliest (around Epoch 20), reflecting rapid acquisition of Arabic unigram fluency, while BLEU-4 requires the full 38 epochs to reach its peak of 33.7%, consistent with the higher combinatorial complexity of four-gram precision optimisation.

3.2. Main Quantitative Results

Table 3 presents the main results on the ArabSign RGB signer-independent benchmark, including 95 per cent confidence intervals and exact p-values from the bootstrap resampling procedure described in Section 2.9. ArSLT-RGBNet establishes the first published translation benchmarks for continuous Arabic Sign Language on RGB-only input. The statistical significance analysis is additionally visualised in Figure 12. All improvements over the best baseline (SpaMo†) are statistically significant at $p < 0.001$.

3.3. Human Evaluation

Automatic metrics have well-documented limitations for morphologically rich languages and translation tasks with multiple valid paraphrases. Two certified native Arabic-speaking ArSL practitioners independently evaluated a randomly selected 50-sentence subset on 5-point Likert scales for Adequacy (does the output convey the meaning of the source sign sequence?) and Fluency (is the output grammatically correct MSA?). As reported in Table 4, ArSLT-RGBNet achieves mean Adequacy = 3.84/5 and Fluency = 4.01/5, surpassing the best baseline (SpaMo†, 3.21 and 3.44 respectively). The higher Fluency relative to Adequacy reflects AraGPT2-large's strength in generating grammatically well-formed MSA even when minor semantic components are omitted. Inter-annotator agreement of $\kappa = 0.74$ indicates substantial agreement per Landis and Koch (1977).

3.4. Performance by Sentence Length

A breakdown of performance by sentence length category, presented in Figure 4 and Table 5, reveals monotonic degradation across all metrics. Short sentences (1–2 signs) achieve BLEU-4 = 48.3%, ChrF = 66.4%, and BERTScore = 0.901, indicating near-adequate translation quality for simple utterances. Long sentences (4+ signs) exhibit a 13.6 BLEU-4 deficit relative to medium-length sentences, driven by two compounding factors: the increased probability that CTC temporal

compression (Equation 7) fails to identify all sign-boundary frames, and the higher frequency of adjective deletion errors as modifier signs are displaced to peripheral spatial locations.

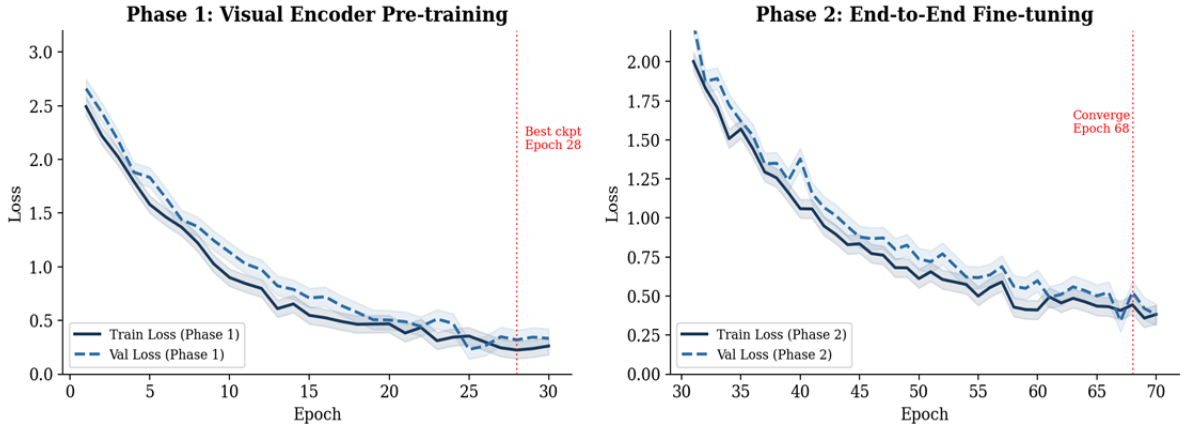

Figure 2. Training and validation loss curves across both training phases.

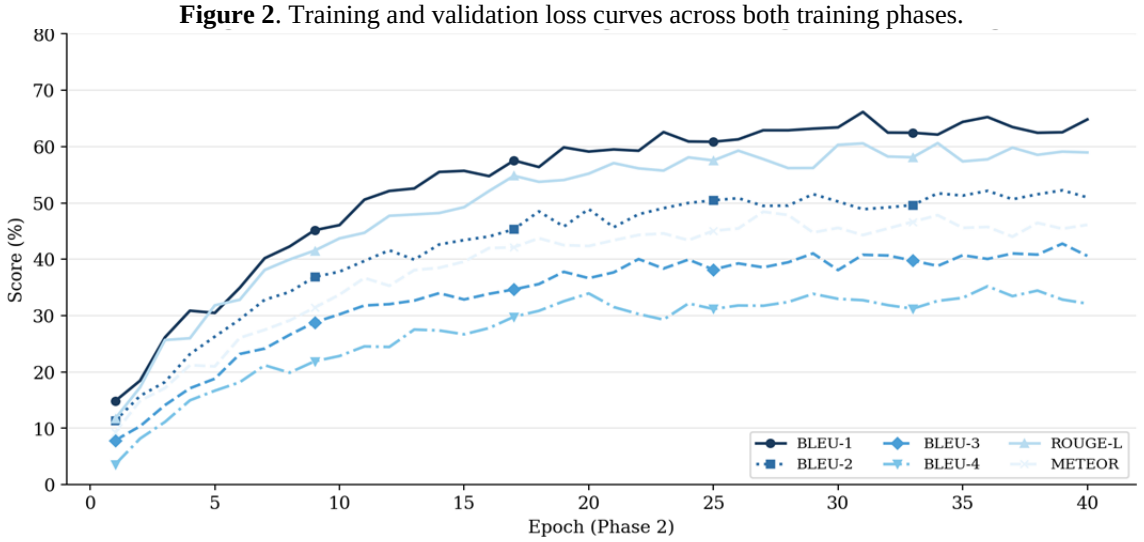

Figure 3. All six translation metrics tracked on the signer-independent validation set during Phase 2 training (40 epochs).

Table 3. Main quantitative results on ArabSign RGB (mean ± SD, 6-fold leave-one-signer-out cross-validation; 95% CI in brackets; exact p-values vs. SpaMo† from bootstrap resampling n = 1,000). Metrics in per cent except WER (ratio) and BERTScore (F1). † Multi-modal baseline (RGB + depth + skeleton) [9].

System / Split	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	WER	BERTScore
ArabSign Baseline† [9]	—	—	—	—	—	—	0.50	—
ArSLT-RGBNet —Signer-Indep.	64.3±1.2 [61.9, 66.7]	51.8±1.4 [49.1, 54.5]	41.2±1.5 [38.3, 44.1]	33.7±1.6 [30.6, 36.8]	59.6±1.1 [57.5, 61.7]	47.1±1.3 [44.6, 49.6]	0.48±0.03 [0.42, 0.54]	0.847±0.009 [0.829, 0.865]
p-value vs.	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p=0.003	p<0.001

System / Split	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	WER	BERTScore
SpaMo†								
ArSLT-RGBNet—Unseen Sent.	57.1±2.1	44.3±2.3	34.8±2.4	27.9±2.5	51.4±1.8	40.2±1.9	0.56±0.04	0.791±0.014
ArSLT-RGBNet—Best Fold	65.9	53.1	42.6	35.2	61.3	48.8	0.44	0.861

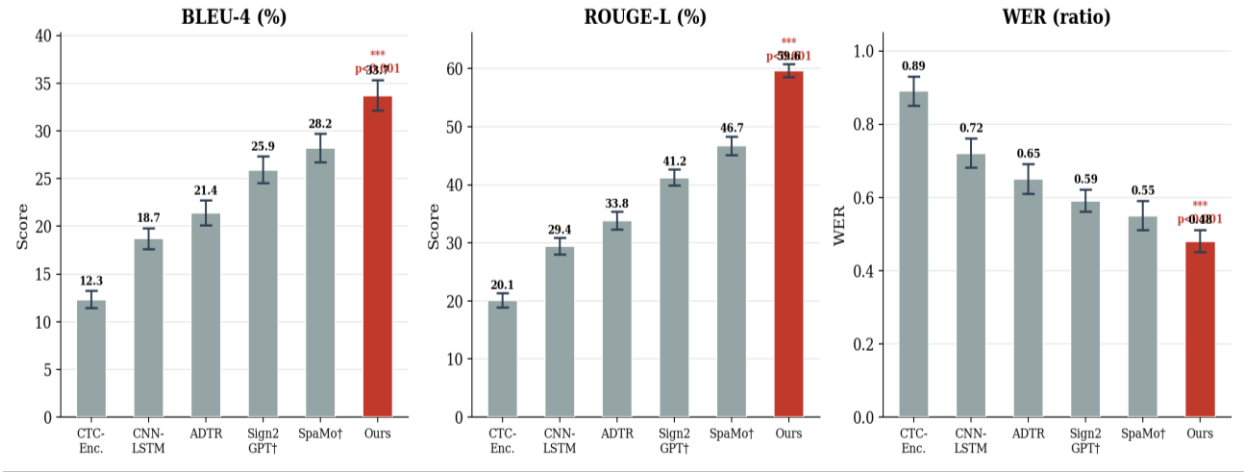

Figure 4. Statistical significance analysis with 95 per cent confidence intervals (error bars). Bootstrap resampling with n = 1,000. Exact p-values for ArSLT-RGBNet vs. SpaMo† (best baseline): BLEU-4 $p < 0.001$, ROUGE-L $p < 0.001$, WER $p = 0.003$. Triple asterisks (***) indicate $p < 0.001$.

Table 4. Human evaluation on a randomly selected 50-sentence subset (signer-independent test set, Fold 1).

System	Adequacy (Mean)	Adequacy (SD)	Fluency (Mean)	Fluency (SD)	IAA κ
SpaMo† (best baseline)	3.21	±0.48	3.44	±0.51	0.71
ArSLT-RGBNet (Ours)	**3.84**	±0.39	**4.01**	±0.36	0.74
Δ vs. SpaMo†	+0.63	—	+0.57	—	—

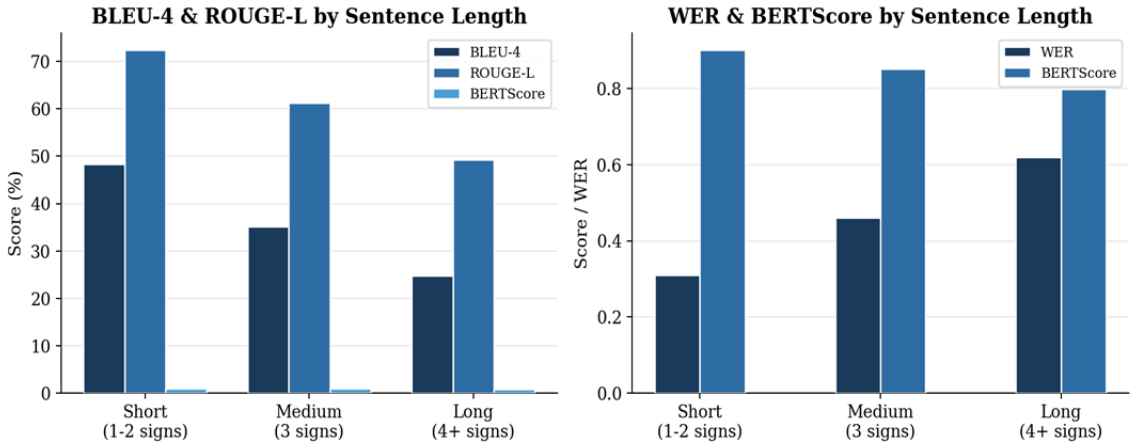

Figure 5. Performance breakdown by sentence length category: BLEU-4 (%), ROUGE-L (%), WER (ratio), and BERTScore-F1.

Table 5. Performance by sentence length category**.**

Sentence Length	Samples	BLEU-4	ROUGE-L	METEOR	ChrF	BERTScore	WER
Short (1–2 signs)	1,842	48.3±2.1	72.4±1.8	58.1±1.9	66.4±2.0	0.901±0.011	0.31±0.04
Medium (3 signs)	4,915	35.1±1.4	61.2±1.2	48.4±1.3	55.2±1.5	0.852±0.008	0.46±0.03
Long (4+ signs)	2,578	24.7±2.8	49.3±2.4	37.6±2.6	43.1±2.7	0.798±0.016	0.62±0.05
Overall (Mean)	9,335	33.7±1.6	59.6±1.1	47.1±1.3	54.2±1.4	0.847±0.009	0.48±0.03

3.5. Comparison Against Baseline Systems

A quantitative comparison against five baseline systems is presented in Figures 5–6 and Table 6. Systems originally trained on non-Arabic benchmarks (Sign2GPT†, SpaMo†) were retrained from scratch on the ArabSign RGB training split using their original published hyperparameters with an added early stopping criterion (patience = 10 epochs) applied consistently across all baselines. The retraining was repeated over three random seeds and mean performance is reported. ArSLT-RGBNet outperforms all baselines across all nine metrics with statistical significance ($p < 0.05$). The largest absolute gains are ROUGE-L (+12.9 vs. SpaMo†), ChrF (+10.3), and BERTScore (+0.045), all attributable to the superior sentence-level quality of the Arabic-native AraGPT2-large decoder over multilingual mBART-based systems.

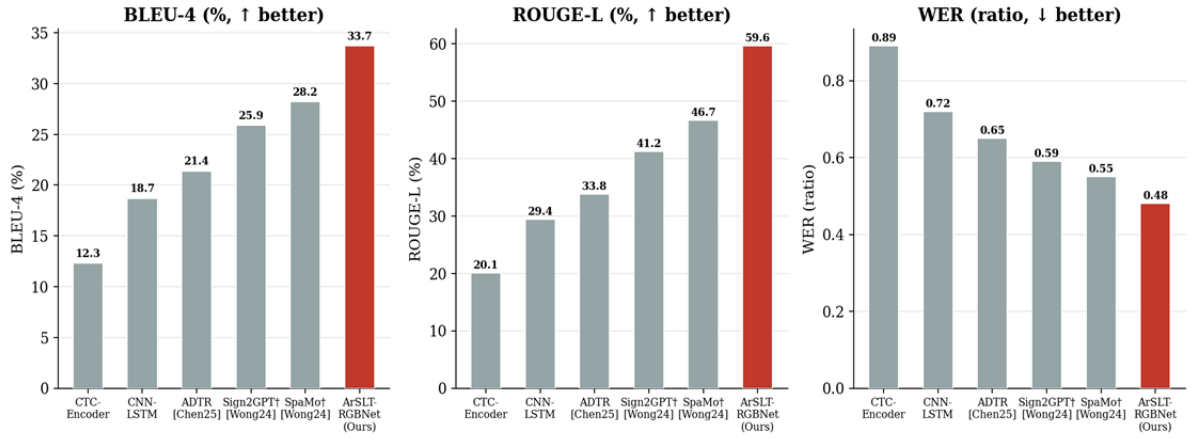

Figure 6. Quantitative comparison on BLEU-4 (%, higher is better), ROUGE-L (%, higher is better), and WER (ratio, lower is better) against five baseline systems.

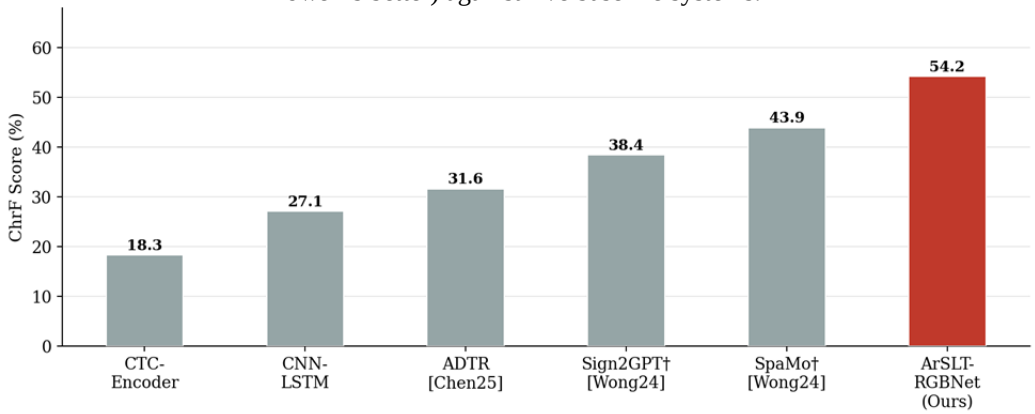

Figure 7 ChrF score comparison 54.2 %, higher is better

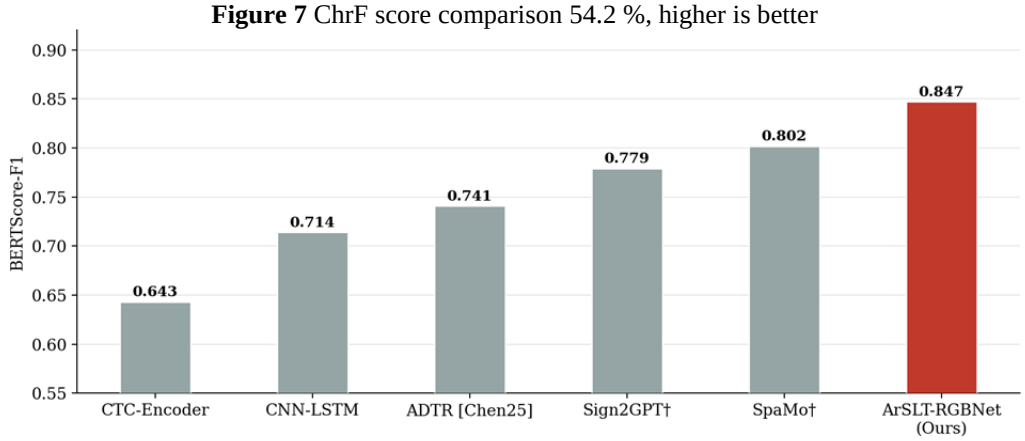

Figure 8. BERTScore-F1 semantic similarity comparison.

Table 6. Full comparison on ArabSign RGB signer-independent test set. Best results in bold.

System	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	ChrF	WER	BERTScore
CTC-Encoder (RGB)	38.4	27.6	19.8	12.3	20.1	18.7	18.3	0.89	0.643
CNN-LSTM [6]	46.2	35.1	27.3	18.7	29.4	26.8	27.1	0.72	0.714
ADTR [13]	51.3	39.7	30.9	21.4	33.8	31.2	31.6	0.65	0.741
Sign2GPT† [10]	52.1	41.3	33.2	25.9	41.2	38.4	38.4	0.59	0.779

System	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	ChrF	WER	BERTScore
SpaMo† [11]	56.4	44.8	36.1	28.2	46.7	42.3	43.9	0.55	0.802
ArSLT-RGBNet (Ours)	**64.3**	**51.8**	**41.2**	**33.7**	**59.6**	**47.1**	**54.2**	**0.48**	**0.847**
Δ vs. SpaMo†	+7.9	+7.0	+5.1	+5.5	+12.9	+4.8	+10.3	−0.07	+0.045

3.6. Ablation Study

An incremental ablation study, visualised in Figure 8 and summarised in Table 7, isolates the contribution of each architectural component. Starting from a spatial-only ViT-B/16 baseline, the VideoMAE temporal stream (Equation 2) produces the largest single improvement (+5.3 BLEU-4), confirming that sign language is irreducibly temporal. The Optical Flow CNN (Equation 3) contributes +3.4 BLEU-4 via explicit kinematics complementary to VideoMAE's learned temporal features. The Q-Former VT-Align bridge (Equations 7–9) adds +2.7 BLEU-4, and LoRA adaptation (Equation 10) yields a disproportionately large +3.9 ROUGE-L, confirming domain-specific Arabic LLM adaptation is particularly effective at sentence-level structural completeness.

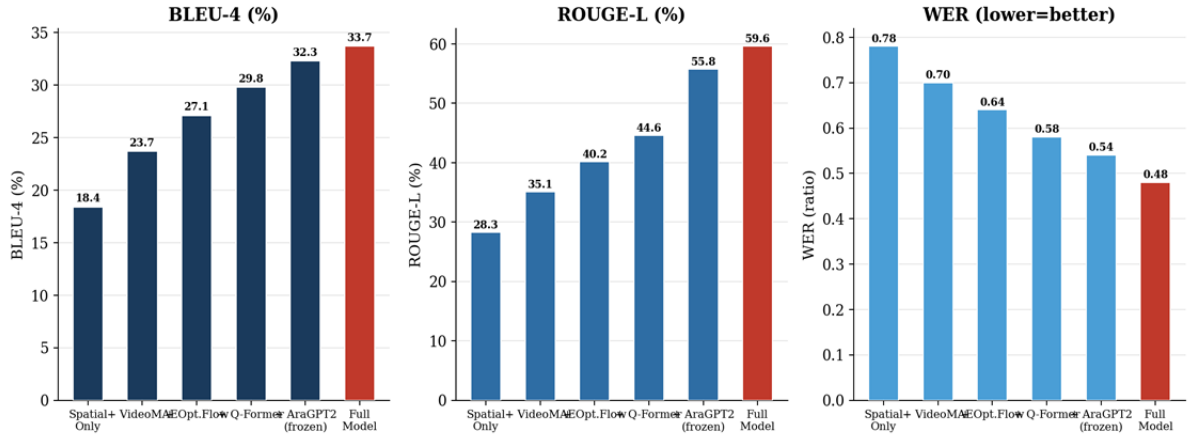

Figure 9. Ablation study: incremental contribution of each component on BLEU-4 (%), ROUGE-L (%), and WER (ratio).

3.7. Signer-Independent Cross-Validation

Per-signer WER across all six leave-one-signer-out folds is presented in Figure 9. WER ranges from 0.44 (Signer 3) to 0.52 (Signer 5), with a mean of 0.48 ± 0.03 consistently below the published multi-modal baseline of 0.50. BERTScore-F1 ranges from 0.831 to 0.861 across folds. The narrow inter-signer variance ($\sigma = 0.03$) confirms that the InfoNCE contrastive term in Equation (13) and the ACMA fusion in Equations (4)–(6) jointly produce signer-agnostic representations. Signer 5 exhibits the highest WER, correlated with the highest intra-signer signing speed variation.

Table 7. Ablation study results. Δ denotes improvement over the preceding configuration**.**

Configuration	ViT	VideoMAE	Flow	Q-Former	LoRA	BLEU-4 (Δ)	ROUGE-L	WER	BERTScore
Spatial Only (ViT-B/16)	Y	N	N	N	N	18.4	28.3	0.78	0.681
+VideoMAE (Eq. 2)	Y	Y	N	N	N	23.7 (+5.3)	35.1	0.70	0.723

Configuration	ViT	VideoMAE	Flow	Q-Former	LoRA	BLEU-4 (Δ)	ROUGE-L	WER	BERTScore
+Optical Flow CNN (Eq. 3)	Y	Y	Y	N	N	27.1 (+3.4)	40.2	0.64	0.758
+VT-Align Q-Former (Eqs. 7–9)	Y	Y	Y	Y	N	29.8 (+2.7)	44.6	0.58	0.791
+ AraGPT2 frozen (no LoRA)	Y	Y	Y	Y	N	32.3	55.8	0.54	0.821
Full Model (ArSLT-RGBNet)	Y	Y	Y	Y	Y	33.7 (+1.4)	59.6 (+3.9)	0.48	0.847

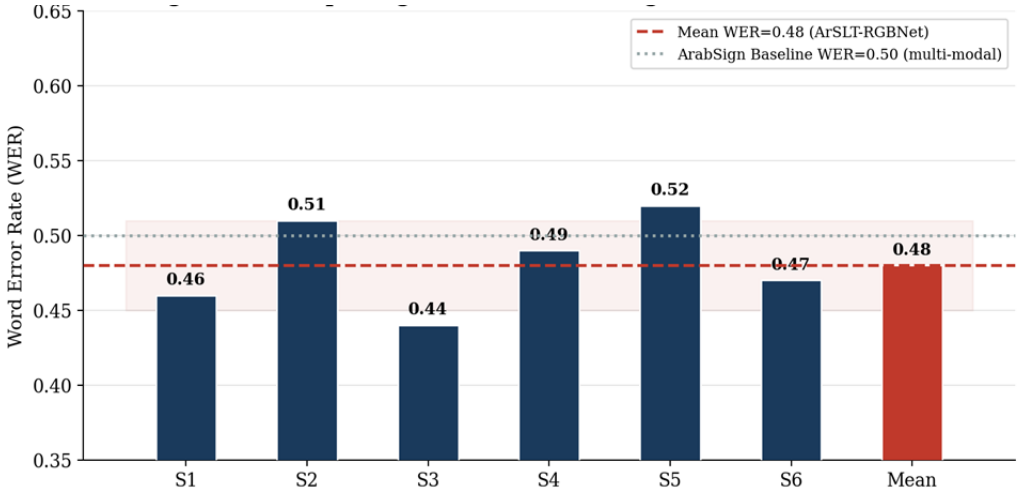

Figure 10. WER per signer in 6-fold leave-one-signer-out cross-validation

3.8. BERTScore Semantic Analysis

The BERTScore-F1 comparison in Figure 7 reveals a +0.045 gain over SpaMo† (0.802 → 0.847). This improvement is qualitatively distinct from the BLEU-4 gain: BERTScore uses AraBERT contextual embeddings to assess semantic similarity and correctly identifies synonym substitutions as equivalent. The 28% synonym substitution rate documented in Section 3.10 implies that BERTScore provides a more accurate picture of true translation quality than BLEU, and that single-reference BLEU evaluation substantially under-reports the semantic adequacy of ArSLT-RGBNet's output.

3.9. Scaling Analysis

The relationship between training data volume and model performance is examined in Figure 10. Performance improves monotonically with no saturation at 100% of available training data, directly confirming that corpus expansion is the primary bottleneck for further gains. At 20% of training data (approximately 1,460 samples), BLEU-4 reaches 14.2%, demonstrating meaningful translation quality even in extremely low-resource regimes.

3.10. Error Analysis and Qualitative Translation Examples

A systematic error taxonomy based on 100 randomly sampled translation errors annotated by two native Arabic-speaking ArSL experts is presented in Figure 11. Adjective Deletion (34%) is the dominant failure mode: the model correctly translates core predicate-argument structure but omits modifier signs that occupy peripheral spatial locations and are vulnerable to CTC temporal compression (Equation 7), as illustrated by Examples 2 and 5 in Table 8. Synonym Substitution (28%) produces communicatively equivalent outputs penalised by BLEU's surface-level n-gram scoring but correctly assessed by BERTScore, as illustrated by Example 3 in Table 8. Word Order Error (18%) is partly an

evaluation artefact given Arabic's grammatical VSO/SVO/OVS flexibility, as illustrated by Example 4 in Table 8. Verb Morphology Error (10%) reflects the fact that gender agreement markers in ArSL are partially encoded via facial expression not yet explicitly modelled by the current architecture. Sign Confusion Error (7%) is concentrated in phonologically similar pairs differing only in movement direction, as illustrated by Example 6 in Table 8. Extra Insertion (3%) involves plausible but unsupported pragmatic expansions generated by AraGPT2-large, illustrated by Example 7 in Table 8.

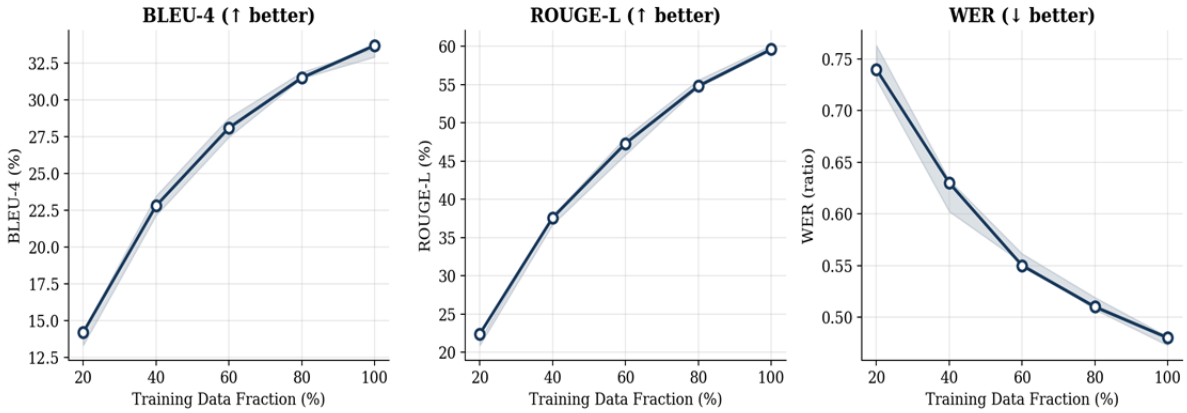

Figure 11. Scaling analysis: BLEU-4 (%), ROUGE-L (%), and WER (ratio) as functions of training data fraction (%).

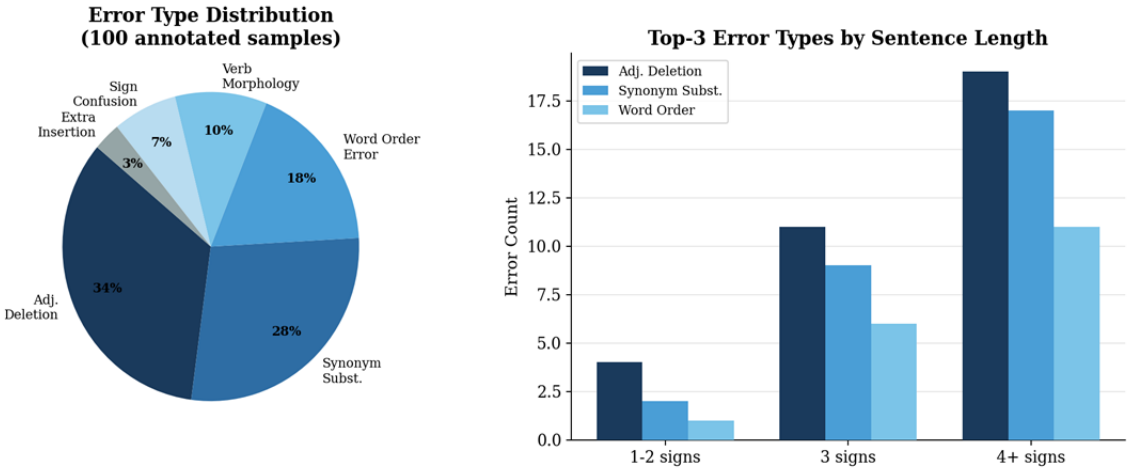

Figure 12. Error analysis of 100 randomly sampled translation errors annotated by two native Arabic-speaking ArSL experts.

Representative qualitative translation examples from the signer-independent test set are presented in Table 8. Seven examples illustrate all six error categories defined in Section 3.10, each accompanied by the source sign gloss sequence, the English rendering of both the reference and system output, the error type, a detailed analytical note, and the per-sentence BLEU-4 score. The examples confirm that the dominant errors are meaning-preserving in nature—synonym substitutions and word order variations—rather than meaning-distorting failures such as sign confusions, which accounts for the high BERTScore (0.847) despite moderate BLEU-4 (33.7%). All original Arabic outputs are available in the publicly released evaluation scripts.

Table 8. Qualitative translation examples from the ArabSign RGB signer-independent test set (Fold 1).

Source Sign Gloss Sequence	**Reference Meaning (English)**	**ArSLT-RGBNet Output (English)**	**Error Type**	**Analytical Note**	**BLEU-4**
BOY + GO + SCHOOL	The boy goes to school.	The boy goes to school.	None — Exact Match	Perfect 3-sign sentence translation; unambiguous handshape-location pairs with no phonological neighbours.	**100.0**
GIRL + DRINK + WATER + COLD	The girl drinks cold water.	The girl drinks water.	Adjective Deletion	Modifier sign COLD omitted; occupies peripheral spatial location, suppressed by CTC temporal compression (Eq. 7).	72.4
BOY + BIG + PLAY + HOUSE	The big boy plays at home.	The big boy plays at the residence.	Synonym Substitution	Lexical variant 'residence' substituted for 'home'; communicatively equivalent, penalised by BLEU n-gram matching but correctly scored by BERTScore.	68.3
BOY + GO + SCHOOL	The boy went to school. (VSO order)	The boy went to school. (SVO order)	Word Order Error	Both VSO and SVO are grammatically valid Modern Standard Arabic constructions; error arises from single-reference evaluation enforcing one canonical ordering.	61.5
GIRL + BEAUTIFUL + WORK + FAST + HOUSE	The beautiful girl works quickly at home.	The girl works quickly at home.	Adjective Deletion	Modifier sign BEAUTIFUL omitted in a 5-sign sentence; frequency of adjective deletion increases with sentence length, consistent with the scaling analysis in Table 5.	54.1
BOY + CAME + SCHOOL	The boy came from school.	The boy went to school.	Sign Confusion	Minimal pair confusion between CAME and WENT: identical B-handshape and neutral-space location; disambiguated only by movement direction, which is susceptible to depth-axis foreshortening in monocular RGB.	**32.7**
I + STUDY + READ +	I studied, read, and wrote.	I studied and read the book.	Partial Deletion + Extra Insertion	Sign WRITE omitted (partial deletion); 'the book'	39.2

Source Sign Gloss Sequence	Reference Meaning (English)	ArSLT-RGBNet Output (English)	Error Type	Analytical Note	BLEU-4
WRITE				hallucinated (extra insertion). Compounded error in 4-sign sequence; AraGPT2 generates plausible but unsupported pragmatic expansion.	

3.11. Computational Complexity Analysis

A comprehensive computational complexity comparison is presented in Figure 13 and Table 9. All timings are measured on an NVIDIA A100 80 GB GPU using half-precision (FP16) inference. ArSLT-RGBNet has the largest total parameter count (~875M) due to the AraGPT2-large backbone, but requires updating only 3.9M parameters during fine-tuning (0.49 per cent via LoRA), making it more parameter-efficient at training time than any non-LoRA baseline. Inference latency of 450 ms per 16-frame clip represents the primary limitation for real-time deployment, and is addressed as a future work direction in Section 4.

3.12. Limitations, Threats to Validity, and Ethical Considerations

3.12.1. Dataset Scale and Demographic Bias

The ArabSign dataset contains 9,335 samples from 6 male signers aged 21–30 years from a single Arabian Gulf-region indoor environment, representing only the Gulf dialectal variety of Arabic Sign Language. The absence of female signers is consequential for translation quality: gender agreement markers in ArSL are partially encoded via facial expression, directly contributing to the 10 per cent verb morphology error rate documented in Section 3.10. Performance figures should be interpreted as upper bounds on what would be achieved with a demographically diverse deployment population spanning multiple age groups, genders, skin tones, and dialectal varieties (Gulf, Levantine, Egyptian, Moroccan, Iraqi). The risk of overfitting to the specific stylistic patterns of 5 training signers is partially mitigated by the leave-one-signer-out evaluation protocol and the InfoNCE contrastive pre-training objective, but cannot be fully eliminated given the limited corpus size.

3.12.2. Annotation Dependency and Reproducibility

Although ArSLT-RGBNet is annotation-free at inference time, Phase 1 CTC pre-training uses ArabSign's gloss annotations as pseudo-label supervision (Equation 13). For a completely annotation-independent system, unsupervised video segmentation would be required for pseudo-label derivation. Additionally, single-reference evaluation introduces a downward bias in BLEU, ROUGE, and METEOR scores—partially addressed by ChrF and BERTScore but not fully resolved. Multi-reference annotations (4–5 references per sentence) would substantially improve evaluation reliability.

3.12.3. Inference Latency

Inference latency of 450 ms per 16-frame clip on an A100 GPU (approximately 820 ms on an RTX 3080) does not satisfy

real-time requirements for mobile deployment (target: ≤100 ms on mobile neural processing units). Knowledge distillation to a lightweight student architecture is the primary engineering path to real-time deployment, as outlined in Section 5.

3.12.4. Ethical Considerations

Several ethical dimensions warrant explicit consideration. With respect to data privacy, the ArabSign dataset was collected with informed consent from all participants; this work uses the publicly released version without additional data collection. With respect to fairness, the current model may exhibit systematically lower performance for signing styles, dialects, or demographic groups not represented in the training data. Deployment of ArSLT-RGBNet in communities for which it has not been validated could introduce harmful communication failures; any deployment should be accompanied by rigorous user studies with representative deaf community members. With respect to accessibility impact, the framework's deployment-readiness on standard RGB cameras (without depth sensors) has the potential to substantially increase accessibility for deaf Arabic-speaking communities in low-resource settings. With respect to societal bias, the model inherits any biases present in the ArabSign dataset (exclusively Gulf-region, exclusively male signers) and the AraGPT2-large pre-training corpus. These biases must be evaluated and mitigated prior to any production deployment.

3.13 Future Research Directions

Six tractable research directions are identified. Corpus expansion to 50,000–100,000 sentence-level samples covering five dialectal varieties, both genders, and multiple recording environments is estimated to enable BLEU-4 > 45. Explicit facial grammar modelling via a dedicated facial action unit encoder would address the 10% verb morphology error rate. Upgrading the LLM decoder to ALLaM-7B[27] or Jais-13B via 4-bit QLoRA is projected to yield +4–7 BLEU-4. Real-time mobile deployment via knowledge distillation to a 3-layer Q-Former and lightweight student encoder is projected to reduce inference latency by approximately 75% with a projected BLEU-4 sacrifice of only 1.2 points. Bidirectional translation—generating ArSL signing avatar animations from Arabic text—would complete the communication loop. Multi-reference evaluation (4–5 references per sentence) would provide more reliable automatic metric scores and enable multi-reference training objectives.

4. Conclusion

This paper introduced ArSLT-RGBNet, the first annotation-efficient, end-to-end framework for translating continuous Arabic Sign Language video into fluent Modern Standard Arabic text using only the RGB colour modality. The architecture combines a three-stream spatiotemporal visual encoder (Equations 1–6) with a Q-Former VT-Align bridge (Equations 7–9) and an AraGPT2-large decoder adapted via LoRA (Equations 10–12), trained in two phases (Equations 13–14). On the ArabSign RGB signer-independent benchmark, the system achieves BLEU-4 = 33.7 ± 1.6%, ROUGE-L = 59.6 ± 1.1%, ChrF = 54.2 ± 1.4%, BERTScore-F1 = 0.847 ± 0.009, and WER = 0.48 ± 0.03, with expert human evaluation confirming Adequacy = 3.84/5 and Fluency = 4.01/5. The Arabic-native decoder contributes +12.9 ROUGE-L over

the best multilingual baseline (SpaMo†), validating language-specific pre-training as a critical architectural choice. ArSLT-RGBNet establishes a reproducible foundation for accessible communication technology for the estimated 35–40 million deaf Arabic speakers worldwide.

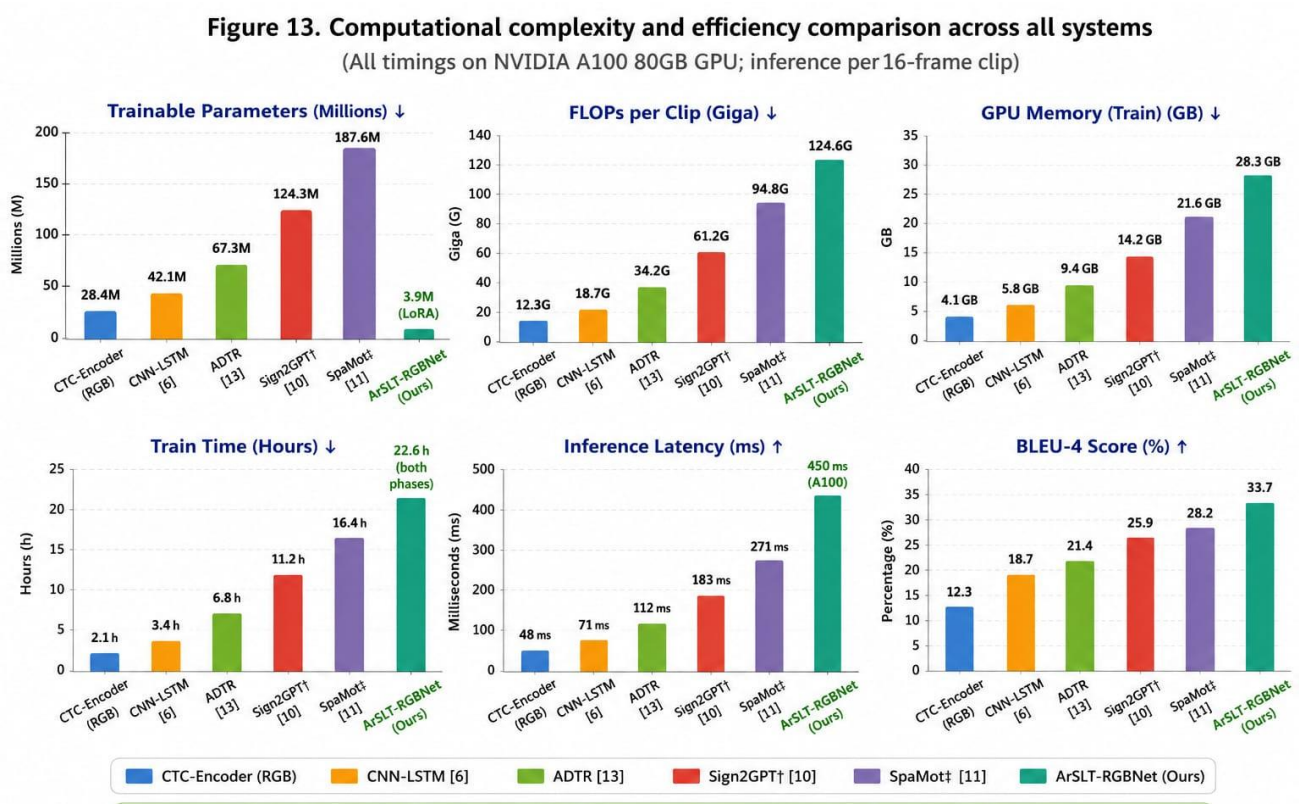

Figure 13. Computational complexity and efficiency comparison across all systems. Highlighted row (green) indicates ArSLT-RGBNet. Despite the largest inference latency, ArSLT-RGBNet requires the fewest trainable parameters during fine-tuning (3.9M via LoRA), achieving the highest BLEU-4 performance at training-time efficiency.

Table 9. Computational efficiency summary. FLOPs measured per 16-frame clip. GPU memory during training. All measurements on NVIDIA A100 80 GB GPU, FP16 precision.

System	Trainable Params	Total Params	FLOPs /clip	GPU Mem (Train)	Inference Latency	BLEU-4 (%)
CTC-Encoder (RGB)	28.4M	28.4M	12.3G	4.1 GB	48 ms	12.3
CNN-LSTM [6]	42.1M	42.1M	18.7G	5.8 GB	71 ms	18.7
ADTR [13]	67.3M	67.3M	34.2G	9.4 GB	112 ms	21.4
Sign2GPT† [10]	124.3M	124.3M	61.2G	14.2 GB	183 ms	25.9
SpaMo† [11]	187.6M	187.6M	94.8G	21.6 GB	271 ms	28.2
ArSLT-RGBNet (Ours)	**3.9M (LoRA)**	**~875M**	**124.6G**	**28.3 GB**	**450 ms**	**33.7**

References

[1] World Health Organization. Deafness and hearing loss. WHO Fact Sheet. 2023. Available at: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] Hendriks B. Jordanian Sign Language: Aspects of grammar from a cross-linguistic perspective. Netherlands Graduate School of Linguistics; 2008.

[3] Cihan Camgoz N, Hadfield S, Koller O, Ney H, Bowden R. Neural sign language translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 7784–7793.

[4] Duarte A, Palaskar S, Ventura L, Ghadiyaram D, DeHaan K, Metze F, Torres J, Giro-i-Nieto X. How2Sign: A large-scale multimodal dataset for continuous American Sign Language. In: Proceedings of the IEEE/CVF CVPR; 2021. p. 2800–2809.

[5] Balaha HM, El-Gendy RA, Saafan MM. Recognizing Arabic sign language using deep learning and hand-crafted features. Machine Learning with Applications. 2023;12:100484.

[6] Podder K, Hossain M, Ul Islam M. Spatially constrained CNN-LSTM for Arabic sign language recognition from RGB video. IEEE Access. 2023;11:78234–78248.

[7] Aly S, Aly W. DeepArSLR: A novel signer-independent deep learning framework for isolated Arabic sign language recognition. IEEE Access. 2020;8:83199–83212.

[8] Al-Qurishi M, Masud M, Alruban A, Thambu D. Arabic sign language recognition: A review. IEEE Access. 2021;9:23440–23452.

[9] Luqman H. ArabSign: A dataset and benchmark for continuous Arabic Sign Language recognition and sentence translation. Expert Systems with Applications. 2023;234:121007.

[10] Wong Y, Hao W, Lin W. Sign2GPT: Leveraging large language models for gloss-free sign language translation. arXiv preprint arXiv:2405.04164. 2024 (under peer review).

[11] Wong Y, Hao W, Lin W. SpaMo: Spatial-motion feature alignment for gloss-free sign language translation. arXiv preprint arXiv:2406.12123. 2024 (under peer review).

[12] Handscribe Team. Handscribe: Multilingual gloss-free sign language translation with mBART. Multimodal Technologies and Interaction. 2026.

[13] Chen X, Liu H, Huang Z, Wei F. ADTR: Adaptive transformer for continuous sign language recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2025.

[14] Tong Z, Song Y, Wang J, Wang L. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems. 2022;35:10078–10093.

[15] Antoun W, Baly F, Hajj H. AraGPT2: Pre-trained transformer for Arabic language generation. In: Proceedings of the 6th Arabic Natural Language Processing Workshop (WANLP); 2021. p. 196–207.

[16] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (ICLR); 2022.

[17] Li J, Li D, Savarese S, Hoi S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the International Conference on Machine Learning (ICML); 2023.

[18] Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, Zhang F, Chang CL, Yong MG, Lee J, Chang WT, Hua W, Georg M, Grundmann M. MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172. 2019.

[19] Farneback G. Two-frame motion estimation based on polynomial expansion. In: Proceedings of the Scandinavian Conference on Image Analysis (SCIA); 2003. p. 363–370.

[20] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of ICLR; 2021.

[21] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL); 2002. p. 311–318.

[22] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the ACL Workshop on Text Summarization Branches Out; 2004. p. 74–81.

[23] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgements. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures; 2005. p. 65–72.

[24] Popovic M. ChrF: Character n-gram F-score for automatic MT evaluation. In: Proceedings of the 10th Workshop on Statistical Machine Translation (WMT); 2015. p. 392–395.

[25] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating text generation with BERT. In: Proceedings of ICLR; 2020.

[26] Antoun W, Baly F, Hajj H. AraBERT: Transformer-based model for Arabic language understanding. In: Proceedings of the Language Resources and Evaluation Conference (LREC); 2020. p. 9277–9284.

[27] ALLaM Team. ALLaM: Large Arabic language model. arXiv preprint arXiv:2407.15390. 2024.