King Fahd University of Petroleum and Minerals
Information and Computer Science Department

LEVERAGING LARGE VISION MODELS FOR CONTINUOUS SIGN LANGUAGE RECOGNITION

A Thesis Proposal Presented to the

DEANSHIP OF GRADUATE STUDIES

Submitted by:

Ahmed Abul Hasanaath
ID: G202302610

Advisor:

Dr. Hamzah Luqman

November, 2024

Table of Contents

List of Figures

List of Tables

Abbreviations and acronyms

SL	Sign Language
ASL	American Sign Language
BSL	British Sign Language
FSL	French Sign Language
DGS	Deutsche Gebärdensprache
ArSL	Arabic Sign Language
SLR	Sign Language Recognition
ISLR	Isolated Sign Language Recognition
CSLR	Continuous Sign Language Recognition
HMM	Hidden Markov Model
CTC	Connectionist Temporal Classification
LLMs	Large Language Models
LVMs	Large Vision Models
SLT	Sign Language Translation
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long-Short Term Memory
3DCNN	Three-dimensional Convolutional Neural Network
GCN	Graph Convolutional Network
ViT	Vision Transformer
GANs	Generative Adversarial Networks
SLP	Sign Language Production
MTL	Multi-task Learning

Chapter 1

Introduction

Effective communication is essential for human interaction. According to the World Health Organization, the prevalence of hearing loss is on the rise with estimates suggesting that by 2050, one in every 10 people will have hearing loss [1]. Sign Language (SL) serves as an essential communication tool for millions of hard-hearing people worldwide. SLs are fully developed and structured forms of communication that rely on visual-manual modalities rather than auditory-vocal means. It incorporates hand gestures, facial expressions, and body movements to convey meaning. Unlike spoken languages, which rely on sounds and phonetics, SLs use a combination of spatial positioning, movement, and visual cues, including eye gaze, facial expression, and even the speed of the signs, to express complex thoughts, emotions, and nuances. SL consists of several components including hand shape, location, movement and orientation. Facial expressions are used usually to form the grammar and syntax of SL, allowing for the expression of not only words but full sentences, questions, and abstract concepts. SL also uses unique syntax rules that follow a "topic-comment" structure that differs significantly from spoken languages' subject-verb-object ordering.

It is important to recognize that SL is not a universal language; instead, there are many distinct sign languages used around the world. Despite a shared reliance on visual communication, SLs are not universal [2]. Similar to spoken languages, SLs differ widely across regions and cultures. There are hundreds of sign languages around the world [3], each with its own grammar, vocabulary, and linguistic nuances [3]. For example, American SL (ASL) differs significantly from British SL (BSL), not only in their vocabulary but also in their grammatical structures and cultural nuances. Similarly, French SL (FSL) and German SL (DGS) have their own unique characteristics, reflecting the cultural and linguistic differences of the countries in which they are used. This diversity highlights the importance of recognizing the sign languages used in each country to integrate the deaf community. It also underscores the need for appropriate language education and support for individuals who rely on these languages for communication.

Despite the widespread use of SL, the development of technologies that can accurately recognize and translate it into spoken language remains a challenging and under-explored domain. Sign Language Recognition (SLR) aims to bridge the gap between technology and the deaf community by enabling real-time translation of SL into written or spoken forms.

SLR can be classified into Isolated SLR (ISLR) and Continuous SLR (CSLR). In ISLR, the system is tasked with recognizing individual sign wrds where each sign is clearly separated in one video. This approach is typically applied in controlled environments, where signers pause between each sign, making it easier for models to learn and classify individual gestures. In contrast, CSLR tackles the more challenging problem of recognizing a sequence of signs performed in a continuous motion, reflecting how SL is used in communication. Unlike ISLR, there are no explicit boundaries between signs in CSLR, which means the system must account for the complex temporal dynamics of coarticulation, where the execution of one sign influences the movement and appearance of the next sign. Additionally, CSLR models must handle varying signing speeds, signer-specific variations, and changes in context. These challenges add to the complexity of CSLR and makes CSLR significantly more difficult to address than ISLR. CSLR can be further divided into two main approaches: gloss-based and gloss-free CSLR. A gloss refers to the label of the sign gesture representing the meaning of a sign. It typically corresponds to a word or phrase in the spoken language (e.g., English, Arabic). In gloss-based CSLR, the system first maps SL gestures to a sequence of intermediate symbolic representations, called glosses, which are essentially transcriptions of the signs. This intermediate representation simplifies the translation process by breaking down SL into smaller, manageable units. In contrast, gloss-free CSLR attempts to directly map signs to spoken or written language without the use of intermediate glosses. This task is more complex than gloss-based task due to the lack of this structured representation. This thesis specifically focuses on gloss-based SLR, where the goal is to develop systems that can accurately recognize and map SL gestures to their corresponding glosses.

This thesis will focus on CSLR, aiming to advance methods for recognizing sequences of signs in natural communication. In a typical CSLR framework, as illustrated in Figure 1.1, the process starts with a video input, which consists of frames capturing continuous gestures. This video is first passed through a spatial module that extracts spatial features, representing key visual information such as the hand shapes, movements, facial expressions, and body posture in each frame. These features are crucial as SL relies on both manual and non-manual components. The extracted spatial features are then forwarded to a temporal module, which processes the sequence over time to generate spatio-temporal features. These features capture the appearance and the dynamic motion of signs, which are essential in recognizing how signs transition fluidly from one sign to the next in the sentence. The spatio-temporal features are then fed into an alignment module to generate gloss alignment proposals. In CSLR, the alignment task involves matching the spatio-temporal features with a sequence of glosses from a predefined vocabulary. However, this is particularly challenging because continuous signing lacks clear temporal alignment between signs and their corresponding glosses. To address this issue, several techniques have been utilized in the literature, such as Hidden Markov Models (HMMs) [2] and Connectionist Temporal Classification (CTC) [4]. Prior to CTC's, HMMs were used, but these required explicit segmentation, making them less adaptable to continuous input. CTC is a specialized loss function designed for sequence-to-sequence problems where the input and output sequences may not have direct, one-to-one alignment. CTC predicts the glosses without requiring predefined temporal boundaries, allowing for more flexibility in handling continuous sign sentences. To evaluate the CSLR systems, alignment proposals from the CTC are compared against the ground truth gloss sequence, which is the correct ordered sequence of glosses corresponding to the signs in the video.

The goal of the CSLR system is to learn to correctly predict the gloss sequence while handling the challenges of continuous input, including dealing with temporal boundaries, variability in signers, and the inherent complexity of SL gestures. The performance of CSLR systems is measured using various evaluation metrics, depending on the task. Accuracy measures the percentage of correctly predicted signs and it is used mainly for evaluating ISLR systems. Word Error Rate (WER) is used to evaluate the system's performance in recognizing sequences of words or signs. It calculates the number of insertions, deletions, and substitutions required to transform the predicted sequence into the ground truth sequence. Bilingual Evaluation Understudy (BLEU) is a metric commonly used in machine translation tasks. In CSLR, it is used to evaluate the quality of translation from SL glosses to natural language text. It is used mainly with gloss-free CSLR systems.

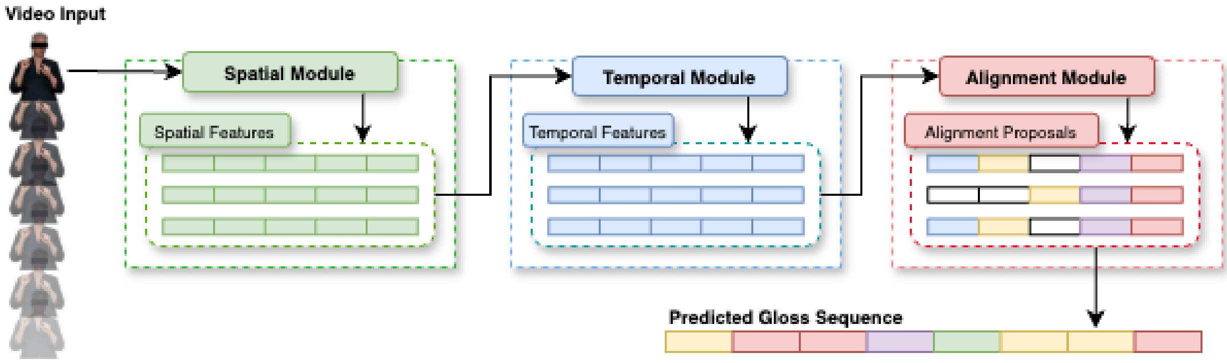


Figure 1.1: General CSLR Framework

The motivation for this thesis arises from the growing need for more accurate and robust real time systems capable of recognizing continuous sign language sentences in real-world, continuous communication settings. Current approaches to CSLR often struggle with the inherent challenges posed by SL, including variability in signer styles, coarticulation between signs, and the lack of clear temporal boundaries between gestures. Additionally, the limited availability of large-scale and diverse datasets further complicates the development of generalized CSLR models. As a result, the recognition accuracy of existing models tends to decline when applied to real-time scenarios with diverse signers, signing speeds, and environments. To address these challenges, this thesis proposes a novel approach that leverages the strengths of diffusion models to pre-train encoders specifically designed for CSLR. Two diffusion models will be pretrained: one focused on capturing sparse temporal sequences, where key sign gestures occur with significant gaps in time, and the other on dense temporal sequences. Where signs occur in rapid succession with minimal separation. The two diffusion models will help the system learn different temporal characteristics of SL, ensuring that the model is better equipped to handle a wide range of signing speeds and styles. This thesis will also explore the utilization of Large Language Models (LLMs) along with Large Vision Models (LVMs) in a contrastive combination for CSLR. The LVMs in the proposed methodology will be built using encoders from the pre-trained diffusion models. The pretrained encoders from the diffusion models will be integrated into a SlowFast network architecture. In this design, the slow pathway will capture long-term, low-frequency motion patterns from the sparse temporal sequences, while the fast pathway will focus on short-term, high-frequency motions from the dense temporal sequences. By combining the strengths of both sparse and dense temporal modeling, the system aims to improve the accuracy and robustness of CSLR, particularly in recognizing continuous signing in natural, unsegmented communication. This integration of diffusion models into a SlowFast network represents a novel approach to address the critical challenges in CSLR and serves as the primary focus of this thesis.

The thesis is organized as follows. Chapter 2 presents a literature review, where we explore available datasets, data acquisition methods, and various CSLR approaches. In Chapter 3, we identify the research gaps and outline the proposed work. This chapter will include a gap analysis, the presentation of research objectives, and a detailed discussion of the proposed methodology. Additionally, we address the expected limitations of the study and conclude with the expected thesis timeline.

Chapter 2

Literature Review

The increasing prevalence of diverse SLs has highlighted the urgent need for effective communication tools tailored for the deaf and hard hearing people. CSLR plays a crucial role in facilitating this communication by translating continuous SL gestures into written or spoken language. This literature review aims to explore the foundational aspects of CSLR, providing a comprehensive overview of its datasets and approaches. The first section lists the various sign acquisition methods and modalities utilized for collecting sign gestures. Understanding how these datasets re constructed is critical for evaluating the quality and applicability of CSLR systems. The section section will list the available CSLR benchmark datasets, providing both a brief overview and a detailed description of their characteristics. Finally, in the third section, we investigate the methodologies employed in CSLR, including advancements in machine learning and computer vision techniques that enhance recognition accuracy and efficiency. This exploration will focus on the evolution of various approaches, from traditional methods to more recent innovations utilizing deep learning and hybrid models. Figure 2.1 summarizes the taxonomy of CSLR.

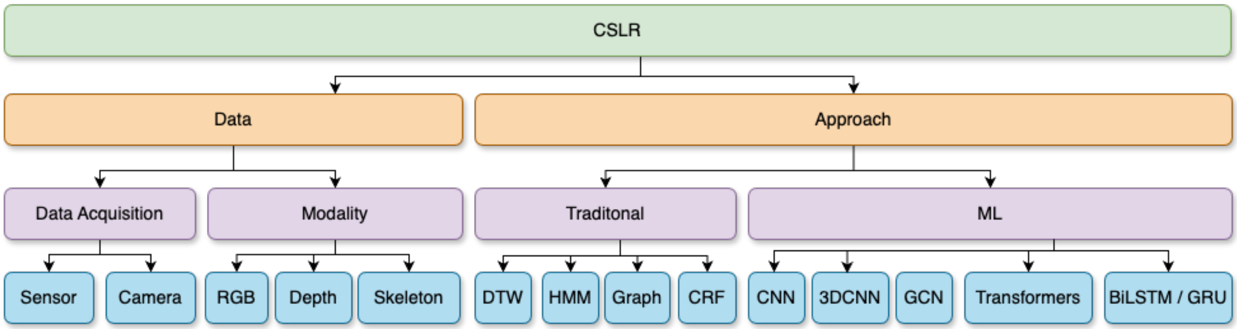
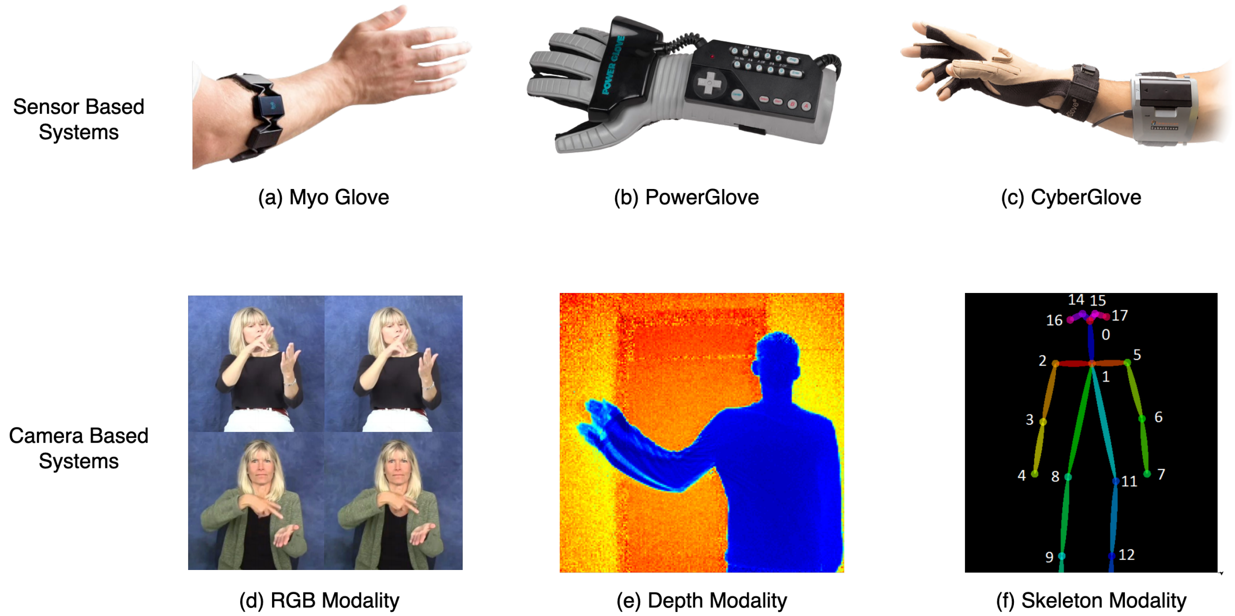


Figure 2.1: CSLR Taxonomy

2.1 Data Acquisition

Data acquisition techniques can be broadly categorized into vision-based and sensor-based methods. Vision-based approaches utilize video cameras to capture signs, enabling a rich visual representation of the signing process. Conversely, sensor-based methods utilize specialized devices, such as data gloves or armbands, to track and collect sign data through direct sensor inputs [5, 6, 7]. Figure 2.2 illustrates both sensor-based and vision-based data acquisition methods.

Various modalities can captured using vision-based systems, each presenting unique advantages that cater to specific needs. The modalities include RGB, depth and skeleton. RGB-based systems utilize standard video cameras to capture the color and texture of the signer's appearance, hands, and background, making them widely accessible. However, they are often sensitive to lighting conditions and background noise [8]. In contrast, depth sensors provide 3D information by measuring the distance between the camera and the object, which is particularly beneficial for managing occlusions and comprehending the geometry of signing actions [9, 10, 11]. Additionally, skeleton-based systems provide an effective means of representation by extracting key points that correspond to the signer's joints, including hands, arms, and body, thereby simplifying the analysis of signing movements [12, 13, 11].



(a) Myo Glove

(b) PowerGlove

(c) CyberGlove

(d) RGB Modality

(e) Depth Modality

(f) Skeleton Modality

Figure 2.2: Illustration of Data Acqusition Methods in SLR. The first row illustrates (a-c) three different sensor-based wearables used for data acquisition and the second row (d-f) illustrates the vision-based methods.

2.2 Available CSLR Datasets

Publicly available datasets for SLR vary across several aspects including language sign level, annoations and size. SLR datasets can be broadly classified based on the sign level into isolated sign language and continuous sign language. Isolated sign language datasets consist of discrete signs that exist outside of a conversational context where each sign gesture is presented in video one or image. In contrast, continuous sign language datasets capture recordings of complete signed sentences, reflecting the fluid nature of real-life communication of deaf people. The majority of publicly available sign language datasets focus on the word level, where each sign corresponds to a specific word, thereby limiting the contextual understanding that can be derived from them [14, 15, 16, 17]. However, datasets that include sentences or conversational contexts are crucial for developing systems that can handle the complexities of natural sign language understanding, including variations in signing speed, context, and simultaneous use of non-manual features. Language is another factor among SLR datasets, with six major sign languages predominantly covered in CSLR research: German Sign Language (GSL) [18, 15, 16], Chinese Sign Language (CSL) [10, 19], American Sign Language (ASL) [14, 20, 21], and Arabic Sign Language [22, 17]. Additional variations among SLR datasets include vocabulary size, number of sentences, number of signers, and modalities used. This section will delve into the available datasets categorized by sign languages—specifically American, German, Chinese, and Arabic—highlighting their unique characteristics, components, and applicability for Continuous Sign Language Recognition. Table 2.1 summarizes these datasets.

2.2.1 American Sign Language Datasets

There are three prominent CSLR datasets for ASL, namely Purdue RVL-SLL [14], RWTH-Boston-104 [20], and ASL-Homework [21]. First, the Purdue RVL-SLL dataset was developed by Purdue University's Robot Vision Lab (RVL). The dataset consists of a large vocabulary of 104 unique ASL signs, specifically designed for continuous sign language recognition. Signed by 14 signers, it consists of approximately 600 unique sentences that cover a wide variety of topics. The RWTH-BOSTON-104 dataset was developed by the RWTH Aachen University in collaboration with Boston University. This dataset consists of 104 distinct signs and contains 201 sentences, offering a diverse set of signed phrases for training machine learning models. Purdue RVL-SLL and RWTH-BOSTON-104 datasets exist in the RGB modality whereas ASL-Homework has been gathered in both RGB and depth modalities. The ASL-Homework dataset, created as part of a project at Boston University, is specifically designed for ASL recognition research in educational settings. It consists of a vocabulary of around 2048 signs and contains 6841 sentences, making it one of the more extensive resources available for continuous ASL sentence recognition.

2.2.2 German Continuous Sign Language Datasets

German CSLR datasets have been extensively used in literature as benchmark datasets to evaluate CSLR systems. RWTH-PHOENIX-Weather-2012 [18] is among the first datasets developed for CSLR. This dataset was groundbreaking for German Sign Language (Deutsche Gebärdensprache, DGS) research, providing a large-scale resource of sign language data. It consists of 1081 unique signs and 2640 sentences, making it the largest publicly available dataset for DGS at the time. This dataset played a pivotal role in advancing CSLR, offering a rich vocabulary and diverse sentences derived from televised weather reports. As its impact on the research community grew, RWTH-PHOENIX-Weather was further developed into

RWTH-PHOENIX-Weather-2014 [15], a significantly expanded version with an increased vocabulary. The RWTH-PHOENIX-Weather-2014 dataset remains one of the most prominent benchmarks for German SLR. With an enlarged vocabulary of approximately 2,048 signs and 6841 sentences, this dataset doubled the scope of its predecessor. It consists of real-world weather forecast recordings, which add a layer of complexity due to the natural conversational speed and diversity in signer expressions. However, one of the challenges posed by RWTH-PHOENIX-Weather-2014 is the high frequency of rare signs—approximately 30% of the vocabulary appears only once in the training data—making it difficult for machine learning models to generalize. The RWTH-PHOENIX-Weather-2014-T dataset [16] is a refined version of the widely used Phoenix2014 dataset, specifically designed to enhance both CSLR and Sign Language Translation (SLT) tasks. It contains 1066 signs and 8257 unique sentences. The Pheonix datasets have become crucial benchmarking datasets in CSLR and SLT research, offering a well-rounded resource that simulates real-world signing scenarios with complex sentence structures and signer variability.

2.2.3 Chinese Continuous Sign Language Datasets

The CSL dataset [10], introduced in 2016, is a CSL resource designed for CSLR research. It consists of approximately 100 sentences signed by 50 different individuals. The dataset was was developed in a controlled lab environment and it is structured to facilitate two types of evaluation: CSL Split I for signer-independent evaluation and CSL Split II for testing on unseen sentences. Despite the inclusion of many signers, the dataset has a relatively small number of unique sentences (100) that limits its variability and the generalizability of models trained on it. However, the subsequent release of CSL-Daily [19] sought to address the limitations. CSL-Daily expands the vocabulary to 2,000 signs spread across 6,598 sentences, thus providing greater diversity and more realistic scenarios for training SLR systems. The expansion of the dataset with CSL-Daily significantly broadened the scope of research possibilities within CSL recognition. The CSL dataset is available in three modalities: RGB, depth and skeleton modalities, whereas the CSL-Daily dataset only exists in the RGB modality.

2.2.4 Arabic Sign Language Datasets

The ArSL for Deaf Drivers dataset [22] is a dataset designed for the recognition of ArSL in the context of driving. This dataset aims to support the development of systems that assist deaf drivers by recognizing commonly used signs that may be relevant to driving scenarios. It contains a vocabulary of around 215 unique sentences, covering essential driving commands, safety instructions, and vehicle control signals. The ArSL for Deaf Drivers dataset is particularly valuable for researchers focused on building assistive technologies for deaf drivers, helping bridge the communication gap in driving contexts. On the other hand, the ArabSign dataset [17] is an important resource for Arabic SLR covering the general domain area. It consists of a relatively small vocabulary of 95 signs and includes 50 unique sentences. The dataset offers multi-modal recordings, including high-quality video of manual signs and non-manual markers like facial expressions and body posture, making it suitable for capturing the complexity of sign language communication. However, these datasets are limited in the vocabulary and number of sentences. There is another dataset called JUMLA-QSL-22 dataset

[23]. This dataset focuses on phrases and sentences commonly used in healthcare settings, containing 6,300 records of 900 sentences. The data collection process includes a diverse set of participants, both hearing-impaired individuals and sign interpreters, to capture variations in signing styles and speeds, ensuring linguistic diversity. The use of true depth cameras allows for comprehensive recordings from multiple angles, capturing intricate signing movements and non-manual markers like facial expressions and body posture. In the context of the arabic language, the need for the development of more comprehensive datasets remains.

Table 2.1: Summary of surveyed CSLR datasets

Dataset	Year	Sign Language	Unique Sentences	Modality	Domain
Purdue RVL-SLL [14]	2002	American	600	RGB	Disasters
RWTH-BOSTON-104 [20]	2007	American	201	RGB	General
RWTH-PHOENIX-Weather [18]	2012	German	2640	RGB	Weather
RWTH-PHOENIX-Weather-2014 [15]	2014	German	6841	RGB	Weather
RWTH-PHOENIX-Weather-2014-T [16]	2018	German	8257	RGB	Weather
CSL [10]	2019	Chinese	100	RGB, Depth, Skeleton	General
ArSL for Deaf Drivers [22]	2021	Arabic	215	RGB	Deaf Drivers
CSL-Daily [19]	2021	Chinese	6598	RGB	General
ArabSign [17]	2022	Arabic	50	RGB, Depth, Skeleton	General
ASL-Homework [21]	2022	American	NA	RGB, Depth	General
JUMLA-QSL-22 [23]	2023	Arabic	900	RGB, Depth	Healthcare

2.3 CSLR Approaches

Several techniques have been proposed for CSLR. These approaches can be categorized based on the types of models and algorithms used to handle the temporal dynamics and spatial features of sign language sequences. Each method comes with its strengths in terms of capturing different aspects of the signing process, from frame-level features to long-term temporal dependencies. In this section, we will explore various techniques, starting with traditional methods like CNNs combined with HMMs to more recent advances using Graph Convolutional Networks (GCNs) and Transformer-based architectures. The subsections will delve into how these approaches address the challenges of CSLR and their contributions to the field.

2.3.1 CNN with HMM

Convolutional Neural Networks (CNNs) and HMMs ha ve been effectively employed to address the challenges of modeling both spatial and temporal information in sign language videos. The combination of CNNs, known for their ability to extract discriminative spatial features, and HMMs, capable of modeling temporal sequences, was a popular approach in the literature prior to the advent of RNNs and transformers [2]. In 2015, Koller et al (2015). [24] tackled the issue of mouth shape recognition, which plays a critical role in SL, by employing deep CNNs for weakly supervised learning. They integrated CNN outputs with HMMs to improve the classification of mouth shapes, highlighting the importance of facial features in CSLR and achieving superior performance on a dataset consisting of only 201 sentences. In 2016, Koller et al (2016) [25] introduced a hybrid CNN-HMM model for CSLR, using a Bayesian framework to interpret CNN outputs within the sequential structure of HMMs. This end-to-end approach demonstrated significant performance improvements, with gains ranging from 15% to 38% on various benchmarks. Moving forward to 2019, Koller et al (2019) [26] proposed a weakly supervised multi-stream CNN-LSTM-HMM framework to uncover sequential parallelism in sign language videos. By synchronizing multiple streams at critical junctures using HMMs. This method addressed the complexity of identifying sign language attributes that lack strong individual discriminative power. The approach led to enhanced performance in lip-reading and hand shape recognition.

2.3.2 RNNs based approaches

Although HMMs were utilized extensively in CSLR research [2]. HMMs lack the ability to capture global context. This issue motivated researchers to shift towards RNNs [27, 28, 29, 30, 31, 32, 33, 10, 34, 35, 36]. Long Short-Term Memory (LSTM), a type of RNN, have an innate ability to "remember" over longer contextual lengths. LSTMs were used for the first time for CSLR in [27] and [28]. Cihan et al. (2017) [27] introduced an architecture consisting of SubUNets, a network comprising CNNs, BiLSTMs, and CTC. Unlike previous methods, which often depend on predefined frame labels to train classifiers, the methodology of [28] addressed the issue of noisy labels in video data that are frequently overlooked in existing datasets. This approach treats training labels as weak labels, iteratively refining the label-to-frame alignment in a weakly supervised manner. Similarly, Cui et al. (2017) of [29] proposed a weakly supervised framework for CSLR and they focused on cases where ordered gloss labels are available but exact temporal locations are missing. In 2023, Hu et al. (2024) [30] introduced AdaSize, a novel solution aimed at enhancing the efficiency of CSLR by addressing spatial redundancy in video frames. AdaSize employs a dynamic end-to-end learnable model for determining frame resolution. This model enables significant reductions in computational load and memory usage while maintaining accuracy comparable to state-of-the-art methods. Through their experiments, the authors demonstrated that AdaSize provided insightful visual analyses of spatial redundancy in CSLR datasets, further boosting the recognition process. Despite these advancements, existing approaches have struggled with efficiency, particularly concerning computational load and resource utilization.

Non-intrusive sensing is integral to CSLR systems in order to keep a natural and comfortable experience for signers. Non-intrusive sensing refers to the ability to gather data or monitor a subject without directly interfering with or disrupting their natural behavior, environment, or body. In the context of SLR, this means capturing SL gestures without requiring the user to wear specialized equipment, use complex devices, or be in a controlled environment. Sharma et al. (2023) [31] presented a deep transfer learning framework for Indian Sign Language, leveraging data from isolated signs to improve continuous sentence recognition. A limitation of their system is the necessity for Inertial Measurement Units (IMUs) to be placed on the hands and fingers of signers, making it an intrusive system. Cross-modality learning is an effective approach to solve the problem of non-intrusive sensing. In 2017, Fand et al. (2017) [32] introduced DeepASL, a groundbreaking deep learning-based framework designed for ubiquitous and non-intrusive word and sentence-level ASL translation. They utilized infrared light for sensing and trained a bidirectional RNN. In 2019, Hu et al. (2023) [33] proposed a prior-aware cross-modality augmentation learning method for CSLR. Their approach generated pseudo video-text pairs through cross-modality editing techniques guided by textual grammar and visual pose priors, creating authentic hard examples to enhance the model's learning.

Multi-stream RNN architectures have emerged as a pivotal approach in the CSLR landscape to effectively address the complexities inherent in processing sign language data. Huang et al. (2018) [10] introduced the Hierarchical Attention Network with Latent Space (LS-HAN) framework for CSLR. The proposed model was designed to tackle challenges in continuous SLR by eliminating the need for temporal segmentation, which can propagate errors and complicate the recognition process. The LS-HAN architecture integrates a two-stream CNN for generating video feature representations, a Latent Space (LS) for bridging semantic gaps, and a Hierarchical Attention Network (HAN) for recognition. Building on this foundation, the Chen et al. (2022) [34] presented a two-stream SLR model with a dual visual encoder to enhance SLR and SLT by addressing the visual redundancy in raw RGB video data. By incorporating two distinct streams—one for raw video input and another for keypoint sequences from a keypoint estimator—this model employs bidirectional lateral connections and a sign pyramid network to facilitate effective communication between the streams. The ability of TwoStream-SLR to seamlessly extend into a translation model, TwoStream-SLT, underscores its versatility. In a more recent contribution, Ahn et al. (2024) [35] explored the SlowFast network, a two-pathway architecture that captures spatial and dynamic features by operating at distinct temporal resolutions. This approach allows for the separate capture of critical aspects of SL such as hand shapes and movements. It introduced two feature fusion methods—Bi-directional Feature Fusion (BFF) and Pathway Feature Enhancement (PFE)—to enhance the transfer and representation of both spatial and dynamic semantics. Continuing this trend, Zheng et al. (2023) [36] also proposed the Contrastive Visual-Textual Transformation for SLR (CVT-SLR), addressing the weakly supervised nature of SLR, which often relies on textual gloss annotations. This approach utilizes a variational autoencoder (VAE) to align visual and textual modalities while leveraging pretrained contextual knowledge, alongside a contrastive cross-modal alignment algorithm that enhances consistency constraints. These multi-stream RNN architectures illustrate a robust evolution in CSLR, emphasizing the importance of effectively leveraging diverse data modalities to enhance recognition accuracy and efficiency.

2.3.3 Temporal Convolutions and 3DCNNs

Temporal convolutions have become an essential technique in CSLR to effectively address the intricacies of video data and the inherent challenges of recognizing SL glosses and their temporal boundaries [2]. Several techniques have been proposed for CSLR using temporal convolutions [37, 38, 39, 40, 41]. In 2020, Papastratis et al. (2020) [37] proposed a cross-modal learning approach to enhance vision-based CSLR by integrating text information, which allows for improved modeling of intra-gloss dependencies. The proposed framework integrated temporal convolutions with 2DCNNs and BiLSTMs. Two robust encoding networks have been employed to generate video and text embeddings, which are then aligned into a joint latent representation to produce more descriptive video-based features. These features are then jointly classified with a decoder. Similarly, Min et al. (2021) [38] tackled the issue of overfitting in vision-based CSLR by introducing a Visual Alignment Constraint (VAC). This method enhances the feature extractor through alignment supervision with auxiliary losses that ensure the alignment of feature predictions. Building on these advancements, Min et al. (2022) of [39] explored the limitations of traditional CTC and introduced RadialCTC, a novel objective function that preserves the iterative alignment mechanism while constraining sequence features on a hypersphere. Hu et al. (2023) [40] investigated the role of human body trajectories in CSLR through their proposed correlation network (CorrNet). The proposed network captures cross-frame trajectories essential for sign identification by dynamically computing correlation maps between adjacent frames. This framework significantly enhances the ability to recognize signs based on local temporal movements. Concurrently, another work introduced a temporal super-resolution network (TSRNet) [41] by incorporating frame-level and temporal feature extraction to minimize resource requirements while maintaining performance.

Three-dimensional convolutional neural networks (3DCNNs) have emerged as a powerful tool in CSLR that effectively capture spatial and temporal features from video data. Several works have been proposed that utilize 3DCNNs for CSLR [42, 43, 44, 45]. In 2018, Pu et al. (2018) [42] introduced a deep neural architecture that integrates a 3D residual convolutional network (3D-ResNet) for visual feature extraction, paired with a stacked dilated convolutional network and CTC to learn the mapping from sequential features to sentence-level labels. This approach addressed the challenges of training deep networks, as the authors proposed an iterative optimization strategy that generates pseudo-labels for video clips that enhance the feature representation of the 3D-ResNet. The subsequent year, the authors of [43] focused on enhancing the effectiveness of pseudo labels in the CNN-RNN-CTC framework by proposing a dynamic pseudo label decoding method. This technique utilized dynamic programming to identify a reasonable alignment path, ensuring that the generated pseudo labels aligned with the natural word order of sign language. A temporal ensemble module integrated features across different time scales, further boosting recognition performance. Another significant contribution came from Yang et al. (2019) [44] with the Structured Feature Network (SF-Net). This network learns multiple levels of semantic information from the data to effectively encode frame, gloss, and sentence-level information into a unified feature representation. Building on these advancements, in 2021, Huang et al. (2021) [45] proposed a boundary-adaptive encoder that effectively captures the hierarchical nature of SL signals. Their method incorporated a location-based window attention model during decoding to enhance long sequence modeling. It also leverages sign language subword units that address both isolated and continuous recognition within a unified framework.

2.3.4 GCN Based Approaches

GCNs have emerged as a powerful tool for addressing the complexities of CSLR by effectively capturing spatial-temporal relationships in SL data. In 2021, a study introduced a Self-Mutual Knowledge Distillation (SMKD) method that utilizes GCNs to enhance the recognition of spatial and temporal features by employing both visual and contextual modules that focus on short-term and long-term information, respectively [46]. This approach highlights the importance of optimizing the visual module to improve feature extraction while sharing weights between classifiers to strengthen discriminative capabilities across modalities. Building on this foundation, Li et al. (2022) [47] proposed a Multi-View Spatial-Temporal Network that leverages GCNs to process RGB and skeleton data. This approach captures the intricate spatial-temporal dynamics inherent in sign language. The network integrates a multi-view spatial-temporal feature e xtractor Network to learn from multiple perspectives, which is complemented by a Transformer-based encoder that excels in modeling long-term dependencies, culminating in a CTC decoder for comprehensive meaning prediction.

2.3.5 Transformer Based Networks

Vision Transformers (ViTs) have emerged as a powerful tool in CSLR by effectively addressing the challenges posed by the complex nature of SL and the limitations of traditional convolutional approaches. The spatial-temporal transformer network (STTN) [48] encoded sign language videos into predicted sequences aligned with text. This model incorporates a chunking technique to manage computational complexity while extracting global and local features efficiently. The cross-modal contextualized sequence transduction (C2ST) model [49] addressed the limitations of traditional CSLR frameworks by incorporating contextual knowledge from gloss sequences into video representation learning. This approach integrated linguistic features and introduced a contextualized sequence transduction loss. This improved alignment learning by overcoming the independence assumptions of conventional methods.

Language models (LMs) have emerged as a transformative architecture in CSLR. LMs have significantly enhanced the modeling of temporal dynamics and contextual dependencies in sign language data. The introduction of SignBERT [50] in 2021 marked a pivotal advancement in the field, as it employed a self-supervised pre-training approach that incorporated a model-aware hand prior. By treating hand poses as visual tokens and utilizing masking strategies for token reconstruction, SignBERT effectively integrated hand prior knowledge to enhance hierarchical context modeling in sign sequences. Later, SignBERT+ [51] was proposed to by address the overfitting tendencies through self-supervised learning and introducing multilevel masked modeling strategies. This allowed the framework to leverage existing data more effectively by enhancing the representation of sign language context while still utilizing a transformer-based architecture to model relationships across frames. Recently, a multiscale temporal network was designed to capture varying temporal features of sign language [52]. This network not only incorporated transformer modules for improved feature encoding and innovatively utilized a multiscale temporal block (MST-block) to learn temporal features at different scales that significantly improved the accuracy of CSLR. In a complementary effort, the introduction of SignCLIP [53] in 2024 further exemplifies the versatility of transformer-based approaches by repurposing the Contrastive Language-Image Pretraining (CLIP) framework to connect spoken language text and sign language videos within a unified representation space. By leveraging large-scale multilingual video-text pairs, SignCLIP efficiently learns useful visual representations.

2.3.6 Generative Models

Generative modeling is a class of machine learning techniques that focuses on learning the underlying distribution of data in order to generate new, synthetic instances that resemble the training data [54]. These models aim to capture complex patterns and structures within the data. By doing so, they enable the creation of realistic samples that can be used in various applications. These applications include art generation, text synthesis, and, notably, sign language production. Among the prominent types of generative models are Generative Adversarial Networks (GANs) and diffusion models, both of which have gained traction in recent years. Figure 2.3 illustrates a high level overview of GANs and diffusion models.

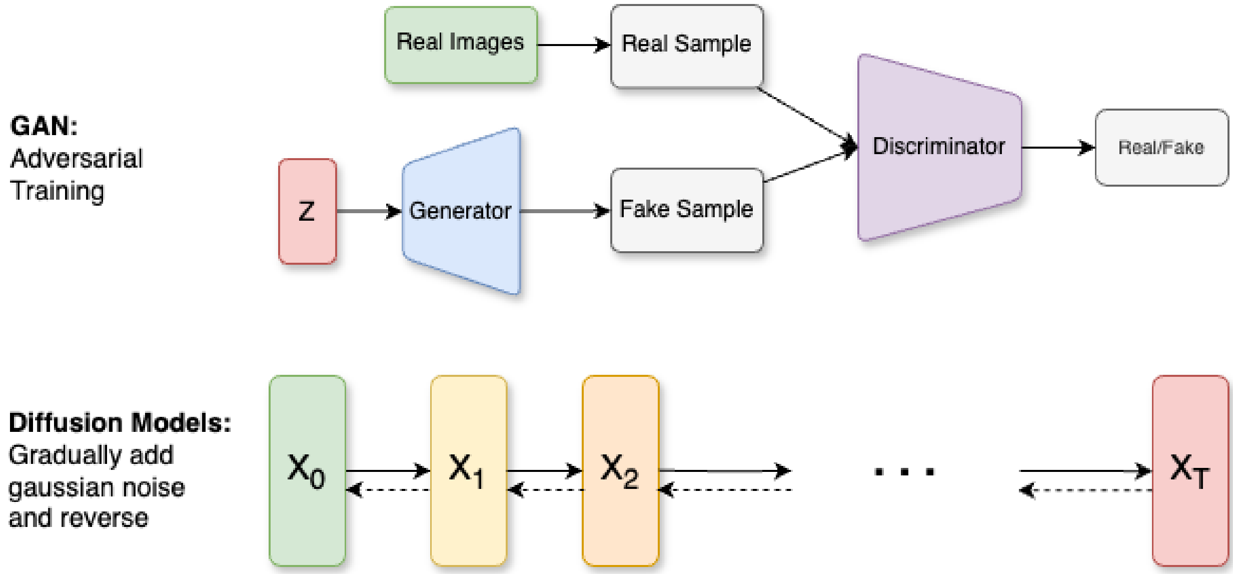


Figure 2.3: Illustration of GANs and Diffusion Models

GANs were first introduced in [55] and have since been used in many applications. GANs consist of two neural networks—the generator and the discriminator. The generator creates synthetic samples, while the discriminator evaluates their authenticity against real data. Through iterative training, the generator learns to produce increasingly realistic outputs, effectively bridging the gap between synthetic and real data. This adversarial training process has shown its effectiveness in applications requiring high-quality data generation, such as image and video synthesis. On the other hand, diffusion models, first introduced in [56], are a newer class of generative models that operate by gradually transforming a simple distribution into a complex one through a series of denoising steps. By starting with random noise and iteratively refining it based on learned patterns from the training data, diffusion models can generate high-quality outputs with impressive fidelity. Recently, diffusion models have emerged as powerful tools in the domain of video generation, showcasing their capability to produce high-fidelity videos from various inputs [57]. For instance, frameworks like Imagen Video [58] leverage a cascade of diffusion models to generate detailed videos conditioned on text prompts, enabling impressive levels of control and artistic diversity. Additionally, unified discrete diffusion approaches have been introduced to facilitate robot policy learning by generating future video predictions from actionless human videos, highlighting the versatility and potential of diffusion models in both creative and practical applications [59]. Both GANs and diffusion models have emerged as powerful tools for generating complex data structures, making them suitable for tasks like sign language production, where realism and accuracy are paramount.

While generative models have shown promise in various applications, they have yet to be fully leveraged for CSLR. Instead, their utilization has primarily focused on Sign Language Production (SLP), where the exploration of generative models remains somewhat limited. Rastgoo et al. (2021) [60] surveyed a few approaches within SLP, specifically mentioning the use of avatars, neural machine translation (NMT), motion graphs, and GANs. Despite these advancements, there is still considerable potential for further investigation into generative models for both SLP and CSLR, particularly in developing more sophisticated and realistic sign language generation techniques that can bridge the gap between spoken and signed communication. This section explores some of the works in the domain of sign language that have utilized GANs and diffusion models.

Generative Adversarial Networks

The use of GANs in SL applications, particularly for CSLR, remains relatively underexplored, with existing literature primarily focusing on SLP. Natarajan et al. (2022) [61] tackled the significant challenges of high-quality sign language video generation by introducing a framework called Dynamic GAN. This model addresses the prevalent issues of blurred effects and subpar video quality seen in previous methods by utilizing skeletal pose information and person images as inputs to generate photo-realistic SL videos. By employing a U-Net-like architecture in the generator phase, the Dynamic GAN effectively creates target frames from skeletal poses, while the VGG-19 framework classifies generated samples according to their corresponding word classes. This approach distinguishes itself from existing by relying on animation or avatars, demonstrating superior performance across multiple benchmark datasets, including RWTH-PHOENIX-Weather 2014T and a self-created dataset for Indian Sign Language. Similarly, Elakkiya et al. (2021) [62] proposed hyperparameter-optimized GAN (H-GAN) to explore the classification of manual and non-manual gestures. This method addresses the complexities arising from the combination of hand, face, and body postures, which can lead to various occlusions. Despite these advancements, the overall application of GANs in CSLR and sign language recognition is still limited, indicating a significant opportunity for future research.

Diffusion Models

Advancements in diffusion models have begun to pave the way for more effective approaches to SLP, yet there remains a noticeable lack of work utilizing these models in CSLR and SL applications in general. In the context of SL, diffusion models have only been used for SLP and SLT. Baltatzis et al. (2024) [63] aimed to address the lack of realism due to the heavy reliance on 2D data. They proposed a novel diffusion-based SLP model trained on a large-scale dataset of 4D signing avatars paired with their text transcripts, showcasing substantial improvements in generating dynamic 3D avatar sequences. This is achieved through a diffusion process that incorporates an anatomically informed graph neural network based on the SMPL-X body skeleton, resulting in superior quantitative and qualitative performance compared to prior methods. Additionally, Tang et al. (2024) [64] introduced a Gloss-driven Conditional Diffusion Model (GCDM) that tackles the complexities of converting text or audio sentences into sign language videos. Furthermore, Fang et al. (2023) [65] addressed a critical gap in SLP by proposing SignDiff, a dual-condition diffusion pre-training model designed for continuous ASL production. This model leverages the How2Sign dataset to generate sign language representations from skeletal poses, employing a novel Frame Reinforcement Network (FR-Net) that enhances alignment between text lexical symbols and dense sign language pose frames. Similarly, the authors of [66] propose SinDiff, a transformer-based diffusion framework that utilizes dynamic attention and global context for spoken-driven sign language generation, achieving improved accuracy over conventional methods. Finally, the authors of [67] present the G2P-DDM model, which transforms sign gloss sequences into corresponding sign pose sequences using a discrete denoising diffusion architecture. Despite these promising developments, the application of diffusion models in CSLR and SLP remains limited, highlighting an opportunity for further research.

Chapter 3

Research Problem and Proposed Work

This chapter outlines the core research problem, identifies gaps in the current literature, and presents the proposed work aimed at addressing these issues. The chapter begins with an analysis of existing research gaps in CSLR in effort to highlight the shortcomings of existing approaches in the field and the unmet needs in CSLR. Building on this analysis, the research objectives section clearly defines the goals and scope of this thesis. The proposed methodology section details the framework that will be proposed to achieve these objectives. Additionally, the expected limitations subsection discusses potential challenges and constraints that could impact the research outcomes. Finally, the timeline section provides a structured schedule, mapping out the key phases of the thesis to ensure timely completion of each stage.

3.1 Gap Analysis

This section critically examines the key challenges within the current research landscape of CSLR and identifies significant underexplored areas. A prominent issue is the insufficient integration of LVMs and generative models, which could significantly enhance CSLR systems by effectively addressing the sequential and contextual nuances of SL. Moreover, the potential of emerging generative techniques, particularly diffusion models, remains largely untapped, presenting opportunities for advancing CSLR performance. The limited availability of annotated CSLR datasets further complicates the development and validation of robust systems. Additionally, the complexity inherent in fusing diverse data modalities such as RGB video, depth, and skeleton data poses challenges in fully capturing the richness of sign language. These research gaps highlight critical areas that will inform the proposed methodologies in this work.

3.1.1 Under-Exploitation of Large Vision Models

Current CSLR research predominantly frames the task as a video understanding problem, overlooking the potential of LVMs to enhance recognition through language modeling. Few studies have ventured to incorporate LVM techniques into CSLR to improve recognition accuracy [36, 50, 49, 68]. Integrating LVMs in CSLR systems while addressing could gain a deeper semantic understanding of sign language, ultimately leading to improved accuracy and effectiveness.

Furthermore, the limitations in existing techniques need to be addressed. These limitations include high computational costs and inadequate temporal modeling. LVMs require substantial resources for training and inference. Guo et al. [68] used 1D TCNs as an alternative to 3D CNNs to reduce computational costs. Additionally, many LVMs, originally designed for static images, need to be adapted to video based tasks for CSLR. Existing approaches such as [53] perform simple averaging across the temporal dimension to produce video embeddings. This process can be enhanced with temporal modeling techniques such as LSTMs and transformers. The aforementioned limitations can be addressed to improve the effectiveness of LVMs in CSLR.

3.1.2 Generative Models for CSLR

Generative models, particularly diffusion models, offer innovative pathways to address various gaps in CSLR. One significant gap is the absence of effective generative feature learning methods. Generative models excel at modeling the underlying distribution of data, which can facilitate enhanced feature extraction and representation from high-dimensional and multimodal data, including RGB, depth, and pose information. While the application of diffusion models in SLR remains scarce, the majority of existing research has concentrated on SLP [63, 64, 65, 66, 67]. For CSLR, leveraging diffusion models could enhance recognition performance through pre-training encoders on unconditional or conditional sign language data, employing disjoint or multi-task learning approaches. This could involve reconstructing images or pose data while integrating masked modeling techniques to bolster feature learning.

3.1.3 Insufficient Annotated CSLR Datasets

The progression of CSLR research heavily relies on the availability of well-annotated and diverse datasets. Currently, the scope of publicly available datasets is limited, primarily focusing on a narrow selection of sign languages. For instance, the RWTH-PHOENIX-Weather-2014 dataset covers DGS with a relatively extensive vocabulary, while datasets like FluentSigners-50 target CSL. However, many sign languages, such as Arabic, Brazilian, and Indian Sign Languages, are underrepresented in CSLR research, thereby restricting the generalizability and inclusivity of existing models [2]. To address this gap, developing new datasets for less-explored sign languages is crucial, facilitating the understanding of their unique characteristics and enabling the creation of tailored CSLR models.

To address the limitations in CSLR dataset diversity, one objective of this thesis is to develop a novel dataset specifically designed for underrepresented sign languages using smartphone selfie cameras. The Selfi dataset will involve individuals signing while holding the camera with one hand and using the other hand for signing. This approach to collecting SL datasets has not been explored in the literature.

3.2 Research Objectives

The primary goal of this thesis is to advance research in CSLR through the application of LVMs to address some of the challenges identified in the gap analysis. The scope of this thesis is as follows:

1. **Comprehensive Literature Review:** A thorough review of the state-of-the-art in CSLR will be conducted focusing on the application of large vision models, MTL, and multimodal fusion techniques. This review will also cover existing datasets, annotation methods, and performance metrics used in CSLR.

2. **Exploring Large Vision Models for CSLR:** Investigate the use of LVMs to enhance the performance of CSLR systems. This exploration will include the following methodologies:

   - **VQVAE Pretraining Methodology:** Implement and evaluate a VQVAE-based pretraining methodology for improving the robustness of CSLR. This will involve training the model on both sparse and dense sign language datasets to learn a shared codebook that effectively captures the visual and temporal characteristics of sign language.

   - **Diffusion-Based Text Generation:** Develop and assess a diffusion model for generating text embeddings conditioned on video features. This will include training the model to denoise embeddings and evaluating its performance in generating accurate gloss sequences in conjunction with the CSLR system.

   - **Contrastive Learning Framework:** Design and implement a contrastive learning framework for CSLR that aligns visual and text embeddings. This will involve experimenting with different contrastive objectives and negative sampling strategies to enhance the model's ability to generalize across diverse sign language datasets.

3. **Performance Evaluation** To evaluate the proposed CSLR approaches comprehensively, experimentation will involve multiple sign languages and diverse datasets. This will help determine the effectiveness of each approach across varying linguistic and visual features, ensuring the models generalize well to different languages and dataset structures. Comparative studies will examine the performance of each methodology across tasks and datasets to identify the most robust and accurate CSLR solutions.

4. **Publication and Contribution to the Research Community:** The results of this research will be published in relevant academic journals and conferences, with the goal of making the findings and newly developed methods available to the CSLR research community. The dataset, models, and code will be shared as benchmarks for future work in this domain.

3.3 Proposed Methodology

In this section, we propose three innovative approaches that utilize LVMs for CSLR. The first approach involves the utilization of a Variational Quantized Variational Autoencoder

(VQVAE) model for generative feature learning. The VQVAE will be trained on a diverse set of prominent CSLR datasets, with a focus on conditioning it solely on language inputs. Once pretrained, the VQVAE encoder will serve as a foundation for training a SlowFast network for CSLR. The second approach focuses on image-conditioned text generation using DMs. Here, we will construct a DM that generates textual representations of sign gloss sequences by conditioning the gloss text embeddings on their corresponding video features. Lastly, the third approach investigates the integration of a CLIP-based architecture for CSLR. This approach aims to establish a framework for aligning visual and linguistic modalities in a contrastive setup. Collectively, the proposed approaches aim to study the effectiveness of LVMs for CSLR.

3.3.1 VQVAE Enhanced Encoding for CSLR

Figure 3.1 illustrates a high-level overview of our VQVAE-based architecture. This methodology unfolds in two distinct phases, beginning with the pretraining of the VQVAE on a combination of prominent CSLR datasets, with the model conditioned on the corresponding language inputs. During the first phase, we will train two VQVAE models simultaneously, utilizing a shared codebook to promote consistency in feature representation. One model will focus on sparse temporal data, while the other will concentrate on dense temporal data. This dual training approach allows for a comprehensive understanding of the temporal dynamics present in sign language.

The VQVAE models will first undergo training aimed at reconstructing individual frames from the input video. The loss function for VQVAE training can be defined as follows:

$$\begin{aligned}
\mathcal{L}_{\text{VQVAE}} &= \mathcal{L}_{\text{reconstruction}} + \beta\mathcal{L}_{\text{commitment}} + \lambda\mathcal{L}_{\text{codebook}}, && (3.1)\\
\mathcal{L}_{\text{reconstruction}} &= \|\mathbf{x} - \hat{\mathbf{x}}\|^2, && (3.2)\\
\mathcal{L}_{\text{commitment}} &= \|\mathbf{z} - \mathbf{e}_k\|^2, && (3.3)\\
\mathcal{L}_{\text{codebook}} &= \|\mathbf{z} - \text{sg}(\mathbf{e}_k)\|^2. && (3.4)
\end{aligned}$$

where, $\mathbf{x}$ represents the original input, $\hat{\mathbf{x}}$ denotes the reconstructed output, $\mathbf{z}$ is the latent variable, while $\mathbf{e}_k$ is the codebook vector corresponding to the assigned code for each input. The term **sg** indicates the stop-gradient operation to prevent gradients from flowing through the codebook during backpropagation. This foundational step ensures that the models effectively learn to capture essential features from static images before adapting to the complexities of video data. Next, we will modify the VQVAE architecture by incorporating a 1D convolutional layer after each 2D convolutional layer. This adaptation is crucial for transitioning from a frame-level representation to a video-level representation, enabling the model to process and understand the sequential nature of sign language effectively.

In the second phase of this methodology, we will integrate the video-level encoders derived from the pretrained VQVAE models into a SlowFast network architecture. This dual-architecture approach allows the network to effectively balance the need for both temporal and spatial feature extraction, crucial for accurately interpreting the dynamic gestures in sign language. The integration process will involve fine-tuning the SlowFast network to optimize its performance for CSLR. In addition to the VQVAE-derived encoders, we will implement a sequence

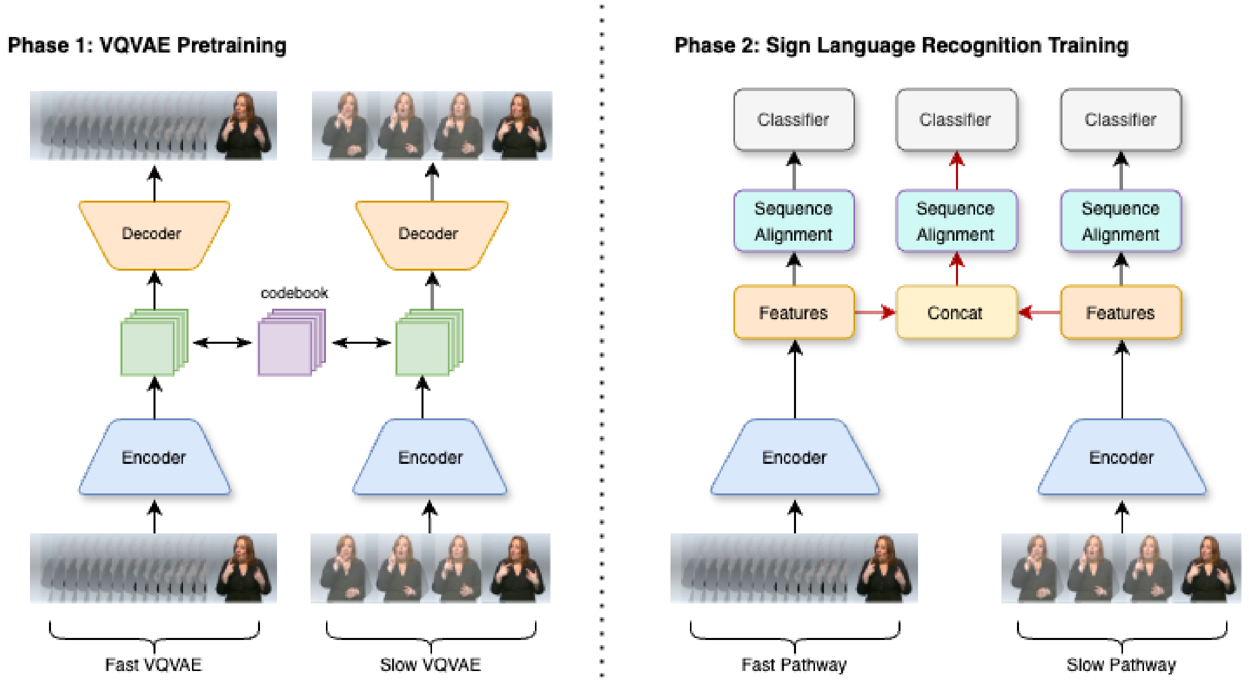


Figure 3.1: Overview of the VQVAE based architecture

alignment model on top of the encoders to ensure that the input video sequences are aligned correctly with their corresponding linguistic outputs. Following this alignment step, a classifier will be introduced to predict the gloss sequences from the aligned representations. The entire network will be trained using CTC loss.

3.3.2 Image-Conditioned Text Generation Using Discrete Diffusion Models

Figure 3.2 illustrates a high-level overview of our diffusion based architecture. In the second methodology, we employ a diffusion model to generate text embeddings conditioned on sign language video features. The process is structured in two main stages: a forward noising process and a reverse denoising process. Additionally, a separate decoder network, trained independently, is employed to generate the final gloss sequence.

To begin, the diffusion model takes an averaged text embedding as input. This embedding, denoted by $\mathbf{e}_{\text{avg}}$, is calculated as the average of individual gloss embeddings in the target gloss sequence, effectively summarizing the sequence. In the forward process, $\mathbf{e}_{\text{avg}}$ is progressively noised across several timesteps $t = 1, 2, ..., T$, adding random Gaussian noise at each step. This forward process is defined by a Markov chain, which iteratively corrupts the embedding.

In the reverse process, the diffusion model learns to reconstruct the original embedding by progressively denoising the input. This reverse process is trained to approximate the original, clean $\mathbf{e}_{\text{avg}}$ from its noisy versions, thus capturing the meaningful structure of the gloss sequence embedding. The training objective of the diffusion model, referred to as the Variational Lower

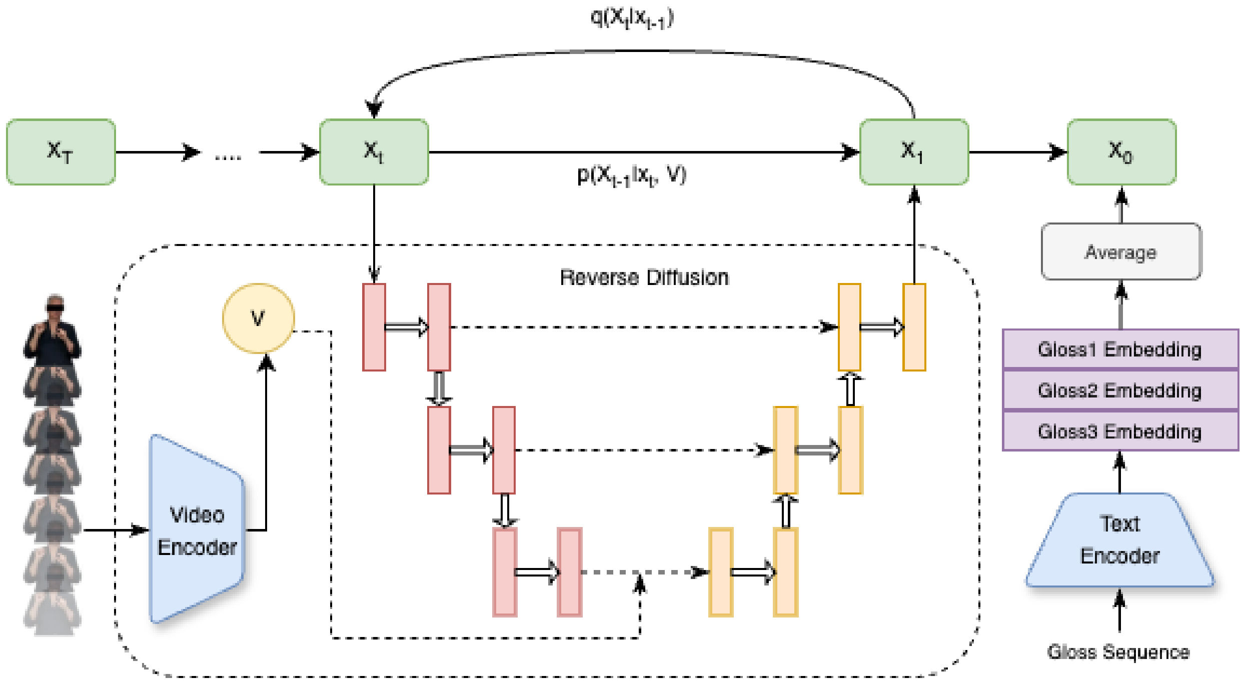


Figure 3.2: Overview of the diffusion based architecture

Bound (VLB) loss, can be formulated as:

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_q \left[\text{DKL} \left(q(\mathbf{e}_{t-1}|\mathbf{e}_t) \parallel p_\theta(\mathbf{e}_{t-1}|\mathbf{e}_t) \right) \right], \tag{3.5}$$

where $q(\mathbf{e}_{t-1}|\mathbf{e}_t)$ represents the true forward noising distribution, and $p_\theta(\mathbf{e}_{t-1}|\mathbf{e}_t)$ represents the learned reverse process. The Kullback-Leibler Divergence (DKL) measures the discrepancy between the true forward process and the model's reverse process, encouraging the model to generate embeddings that closely match the original gloss embedding structure.

Once the text embedding generation via the diffusion model is complete, a separate decoder network, trained independently, is used to generate the full gloss sequence. This decoder is essentially a CSLR model which takes as input the video features from the sign language video. The decoder is conditioned on the averaged text embedding $\mathbf{e}_{\text{avg}}$ generated by the diffusion model, which allows it to better align its generated gloss sequence with the text representation produced by the diffusion process. The decoder model is trained using the CTC loss.

3.3.3 Contrastive Learning for CSLR

Figure 3.3 illustrates a high-level overview of our CLIP based architecture. In the third methodology, we explore the use of contrastive learning to enhance CSLR. Contrastive learning aims to bring the representations of similar inputs closer together in the embedding space, while pushing dissimilar representations farther apart. For CSLR, this technique is especially useful in learning discriminative features that distinguish different signs, glosses, and sequences, thus improving recognition accuracy. The primary goal of this approach is to learn an embedding space in which the features of video segments that represent the same gloss are closer together than those representing different glosses. We achieve this by training a model with a contrastive loss, where positive pairs (video segments with the same gloss) are encouraged to have high similarity, while negative pairs (video segments with different glosses) are encouraged to have low similarity.

To formalize this, let $\mathbf{v}_i$ and $\mathbf{v}_j$ represent the embeddings of two video segments. If $\mathbf{v}_i$ and $\mathbf{v}_j$ correspond to the same gloss, they form a positive pair; otherwise, they form a negative pair. The contrastive loss function $\mathcal{L}_{\text{contrastive}}$ is defined as follows:

$$\mathcal{L}_{\text{contrastive}} = \sum_{(i,j)\in\mathcal{P}} (1 - \cos(\mathbf{v}_i, \mathbf{v}_j)) + \sum_{(i,j)\in\mathcal{N}} \max\left(0, \cos(\mathbf{v}_i, \mathbf{v}_j) - \delta\right), \tag{3.6}$$

where $\mathcal{P}$ denotes the set of positive pairs, and $\mathcal{N}$ denotes the set of negative pairs. The function $\cos(\mathbf{v}_i, \mathbf{v}_j)$ computes the cosine similarity between embeddings $\mathbf{v}_i$ and $\mathbf{v}_j$, encouraging high similarity for positive pairs and low similarity for negative pairs. The margin δ defines a threshold, ensuring that the similarity between negative pairs does not exceed a certain value.

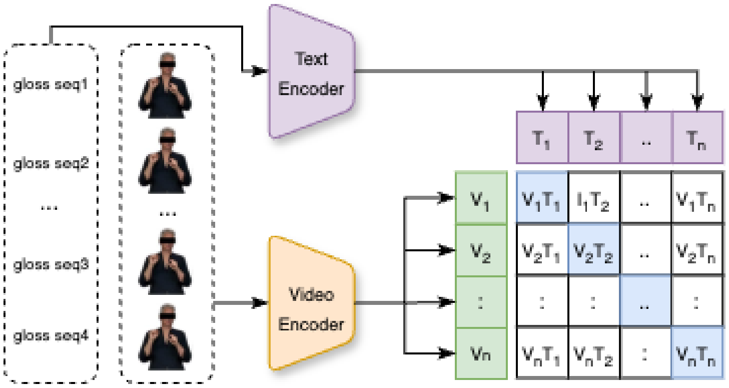


Figure 3.3: Overview of the CLIP-based architecture

To generate positive and negative pairs, we will employ a strategy based on temporal alignment and gloss annotations. Positive pairs are constructed by sampling segments within the same video or across videos that correspond to the same gloss label, while negative pairs are formed by sampling segments from different gloss labels. This allows the model to learn robust, gloss-specific representations that are invariant to variations within each gloss, such as signer differences and slight variations in execution.

Additionally, we incorporate an augmentation mechanism to further enrich the learning process. For each video segment, we apply spatial and temporal augmentations, such as random cropping, scaling, and temporal jittering, to create multiple augmented views of the same gloss. These augmented views are used as additional positive pairs, reinforcing the model's ability to recognize the gloss under different conditions.

The overall loss function combines the contrastive loss with a classification loss $\mathcal{L}_{\text{class}}$ to guide the model towards accurate gloss classification. The total loss $\mathcal{L}_{\text{total}}$ is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrastive}} + \alpha \mathcal{L}_{\text{class}}, \tag{3.7}$$

where α is a weighting parameter that balances the influence of the contrastive and classification losses. The classification loss $\mathcal{L}_{\text{class}}$ is defined as the cross-entropy loss over the gloss labels, ensuring that the model not only learns to distinguish between similar and dissimilar segments but also accurately predicts the gloss labels for each segment.

3.4 Expected Limitations

Despite the potential advancements proposed in this research, several limitations may arise during the course of the study. These limitations include:

- **Dataset Limitations:** The performance of our models is highly dependent on the quality and diversity of the available CSLR datasets. Most datasets are limited in terms of the variety of signers, glosses, and regional variations in sign language. Additionally, the datasets may lack annotations for CSLR, especially for more nuanced gestures and complex sentences. These limitations may affect the generalizability of our models to new sign language inputs.

- **Model Complexity and Computational Requirements:** The proposed methodologies, especially those involving diffusion models and contrastive learning, are computationally intensive. Training these models may require significant computational resources, including large-scale GPUs and extended training time.

- **Real-Time Performance Constraints:** One of the practical challenges in sign language recognition is achieving real-time performance. Although the proposed methodologies may excel in accuracy, they may not be suitable for real-time applications due to latency introduced by complex model architectures and inference time. Ensuring real-time usability may require further model optimization or architectural simplification, potentially at the expense of accuracy.

- **Signer-Independence Challenge:** One of the key challenges in CSLR is achieving signer-independence, where the model can accurately recognize signs from unseen signers. Even with data augmentation or pose-based techniques, the model may struggle to generalize well across different signers with varying appearances, signing styles, and speeds. This could lead to lower performance on real-world, signer-independent tasks.

- **Diffusion Model Limitations:** While diffusion models show great promise, their application to CSLR is relatively novel and unexplored. Issues related to model convergence and learning from temporal and spatial features may emerge. Additionally, the long training times and complexity associated with diffusion models may limit experimentation and fine-tuning.

- **Time Constraints:** Given the scope of the research—developing new models and experimenting with multiple techniques—time may be a limiting factor. Some research avenues may require further exploration beyond the timeframe of this thesis.

These limitations highlight the challenges associated with advancing CSLR through the proposed methodologies. Addressing these issues in future work will be essential for developing more robust, generalizable, and interpretable models that can perform effectively across diverse contexts and applications.

3.5 Project Timeline

The timeline shown in Table 3.1 outlines the systematic approach that will be taken in order to ensure timely completion of this thesis. The project is divided into several phases, each focusing on specific tasks that build upon the previous work. Starting with foundational model development in the setup phase, the project will progress through the implementation of diffusion models, followed by the training processes, and concluding with evaluation and benchmarking. This structured timeline ensures that the research objectives are met efficiently and effectively.

Table 3.1: Thesis Timeline

Phase	**Tasks**	**Deadline**
Setup	Build a simple model for CSLR using 2D CNN Replace with 3D CNN Add 2-stream RNN	Mid December 2024
Methodology 1	Pretrain VQVAE for CSLR on sparse and dense data streams with shared codebook Extend VQVAE with temporal layers for video-level representations Train the model for frame reconstruction	End of January 2024
Methodology 2	Build a single-stream multi-task diffusion model Pretrain diffusion model for generating text embeddings from gloss sequence Integrate diffusion model for denoising embedding during generation	End of February 2025
Methodology 2	Train separate decoder for gloss sequence generation Condition decoder on video features and averaged text embeddings Train decoder model with CTC loss for accurate gloss alignment	End of March 2025
Methodology 3	Implement contrastive learning framework for CSLR Design contrastive objective to align visual and text embeddings Experiment with different negative sampling strategies	End of April 2025
Evaluation	Conduct evaluations, ablation studies, and benchmarking on CSLR tasks Measure impact of VQVAE pretraining, diffusion model, and contrastive learning Analyze model performance on diverse CSLR datasets	End of May 2025

References

[1] *World report on hearing — who.int.* https://www.who.int/publications/i/item/9789240020481. [Accessed 02-10-2024].

[2] Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. "Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects". In: *Information Processing & Management* 61.5 (2024), p. 103774.

[3] Danielle Bragg et al. "Exploring collection of sign language datasets: Privacy, participation, and model performance". In: *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility.* 2020, pp. 1–14.

[4] Alex Graves and Alex Graves. "Connectionist temporal classification". In: *Supervised sequence labelling with recurrent neural networks* (2012), pp. 61–93.

[5] Karush Suri and Rinki Gupta. "Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory". In: *Computers & Electrical Engineering* 78 (2019), pp. 493–503.

[6] Mohamed Hassan, Khaled Assaleh, and Tamer Shanableh. "Multiple proposals for continuous arabic sign language recognition". In: *Sensing and Imaging* 20.1 (2019), p. 4.

[7] Deniz Ekiz et al. "Sign sentence recognition with smart watches". In: *2017 25th Signal Processing and Communications Applications Conference (SIU)*. IEEE. 2017, pp. 1–4.

[8] Nikolas Adaloglou et al. "A comprehensive study on deep learning-based methods for sign language recognition". In: *IEEE transactions on multimedia* 24 (2021), pp. 1750–1762.

[9] Ildar Kagirov et al. "TheRuSLan: Database of Russian sign language". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 6079–6085.

[10] Jie Huang et al. "Video-based sign language recognition without temporal segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[11] Maher Jebali, Abdesselem Dakhli, and Mohammed Jemni. "Vision-based continuous sign language recognition using multimodal sensor fusion". In: *Evolving Systems* 12.4 (2021), pp. 1031–1044.

[12] Wisnu Aditya et al. "Novel spatio-temporal continuous sign language recognition using an attentive multi-feature network". In: *Sensors* 22.17 (2022), p. 6452.

[13] Heike Brock, Iva Farag, and Kazuhiro Nakadai. "Recognition of non-manual content in continuous japanese sign language". In: *Sensors* 20.19 (2020), p. 5621.

[14] Aleix M Martínez et al. "Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language". In: *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE. 2002, pp. 167–172.

[15] Oscar Koller, Jens Forster, and Hermann Ney. "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers". In: *Computer Vision and Image Understanding* 141 (2015), pp. 108–125.

[16] Necati Cihan Camgöz et al. "Rwth-phoenix-weather 2014 t: Parallel corpus of sign language video, gloss and translation". In: *CVPR, Salt Lake City, UT* 3 (2018), p. 6.

[17] Hamzah Luqman. "ArabSign: a multi-modality dataset and benchmark for continuous Arabic Sign Language recognition". In: *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2023, pp. 1–8.

[18] Jens Forster et al. "RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus." In: *LREC*. Vol. 9. 2012, pp. 3785–3789.

[19] Hao Zhou et al. "Improving sign language translation with monolingual data by sign back-translation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1316–1325.

[20] Philippe Dreuw and Hermann Ney. "SignSpeak-Bridging the gap between signers and speakers". In: *Beitrag in dieser Sitzung* (2009).

[21] Saad Hassan et al. "ASL-Homework-RGBD Dataset: An annotated dataset of 45 fluent and non-fluent signers performing American Sign Language homeworks". In: *arXiv preprint arXiv:2207.04021* (2022).

[22] Samah Abbas, Hassanin Al-Barhamtoshy, and Fahad Alotaibi. "Towards an Arabic Sign Language (ArSL) corpus for deaf drivers". In: *PeerJ Computer Science* 7 (2021), e741.

[23] Oussama El Ghoul, Maryam Aziz, and Achraf Othman. "JUMLA-QSL-22: A Novel Qatari Sign Language Continuous Dataset". In: *IEEE Access* (2023).

[24] Oscar Koller, Hermann Ney, and Richard Bowden. "Deep learning of mouth shapes for sign language". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 85–91.

[25] Oscar Koller et al. "Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition." In: *BMVC*. 2016, pp. 136–1.

[26] Oscar Koller et al. "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos". In: *IEEE transactions on pattern analysis and machine intelligence* 42.9 (2019), pp. 2306–2320.

[27] Necati Cihan Camgoz et al. "Subunets: End-to-end hand shape and continuous sign language recognition". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3056–3065.

[28] Oscar Koller, Sepehr Zargaran, and Hermann Ney. "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4297–4305.

[29] Runpeng Cui, Hu Liu, and Changshui Zhang. "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7361–7369.

[30] Lianyu Hu et al. "Scalable frame resolution for efficient continuous sign language recognition". In: *Pattern Recognition* 145 (2024), p. 109903.

[31] Sneha Sharma, Rinki Gupta, and A Kumar. "Continuous sign language recognition using isolated signs data and deep transfer learning". In: *Journal of Ambient Intelligence and Humanized Computing* (2023), pp. 1–12.

[32] Biyi Fang, Jillian Co, and Mi Zhang. “Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation”. In: *Proceedings of the 15th ACM conference on embedded network sensor systems*. 2017, pp. 1–13.

[33] Hezhen Hu et al. “Prior-aware cross modality augmentation learning for continuous sign language recognition”. In: *IEEE Transactions on Multimedia* 26 (2023), pp. 593–606.

[34] Yutong Chen et al. “Two-stream network for sign language recognition and translation”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17043–17056.

[35] Junseok Ahn, Youngjoon Jang, and Joon Son Chung. “Slowfast Network for Continuous Sign Language Recognition”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 3920–3924.

[36] Jiangbin Zheng et al. “Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 23141–23150.

[37] Ilias Papastratis et al. “Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space”. In: *IEEE Access* 8 (2020), pp. 91170–91180.

[38] Yuecong Min et al. “Visual alignment constraint for continuous sign language recognition”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11542–11551.

[39] Yuecong Min et al. “Deep radial embedding for visual sequence learning”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 240–256.

[40] Lianyu Hu et al. “Continuous sign language recognition with correlation network”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2529–2539.

[41] Qidan Zhu et al. “Continuous sign language recognition via temporal super-resolution network”. In: *Arabian Journal for Science and Engineering* 48.8 (2023), pp. 10697–10711.

[42] Junfu Pu, Wengang Zhou, and Houqiang Li. “Dilated convolutional network with iterative optimization for continuous sign language recognition.” In: *IJCAI*. Vol. 3. 2018, p. 7.

[43] Hao Zhou, Wengang Zhou, and Houqiang Li. “Dynamic pseudo label decoding for continuous sign language recognition”. In: *2019 IEEE International conference on multimedia and expo (ICME)*. IEEE. 2019, pp. 1282–1287.

[44] Zhaoyang Yang et al. “Sf-net: Structured feature network for continuous sign language recognition”. In: *arXiv preprint arXiv:1908.01341* (2019).

[45] Shiliang Huang and Zhongfu Ye. “Boundary-adaptive encoder with attention method for Chinese sign language recognition”. In: *IEEE Access* 9 (2021), pp. 70948–70960.

[46] Aiming Hao, Yuecong Min, and Xilin Chen. “Self-mutual distillation learning for continuous sign language recognition”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11303–11312.

[47] Ronghui Li and Lu Meng. “Multi-view spatial-temporal network for continuous sign language recognition”. In: *arXiv preprint arXiv:2204.08747* (2022).

[48] Zhenchao Cui et al. “Spatial–temporal transformer for end-to-end sign language recognition”. In: *Complex & Intelligent Systems* 9.4 (2023), pp. 4645–4656.

[49] Huaiwen Zhang et al. "C2st: Cross-modal contextualized sequence transduction for continuous sign language recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 21053–21062.

[50] Hezhen Hu et al. "SignBERT: Pre-training of hand-model-aware representation for sign language recognition". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11087–11096.

[51] Hezhen Hu et al. "Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (2023), pp. 11221–11239.

[52] Qidan Zhu et al. "Multiscale temporal network for continuous sign language recognition". In: *Journal of Electronic Imaging* 33.2 (2024), pp. 023059–023059.

[53] Zifan Jiang et al. "SignCLIP: Connecting Text and Sign Language by Contrastive Learning". In: *arXiv preprint arXiv:2407.01264* (2024).

[54] Priyanka Gupta et al. "Generative AI: A systematic review using topic modelling techniques". In: *Data and Information Management* (2024), p. 100066.

[55] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[56] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265.

[57] Ling Yang et al. "Diffusion models: A comprehensive survey of methods and applications". In: *ACM Computing Surveys* 56.4 (2023), pp. 1–39.

[58] Jonathan Ho et al. "Imagen video: High definition video generation with diffusion models". In: *arXiv preprint arXiv:2210.02303* (2022).

[59] Haoran He et al. "Large-scale actionless video pre-training via discrete diffusion for efficient policy learning". In: *arXiv preprint arXiv:2402.14407* (2024).

[60] Razieh Rastgoo et al. "Sign language production: A review". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 3451–3461.

[61] B Natarajan and R Elakkiya. "Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks". In: *Soft Computing* 26.23 (2022), pp. 13153–13175.

[62] R Elakkiya, Pandi Vijayakumar, and Neeraj Kumar. "An optimized generative adversarial network based continuous sign language classification". In: *Expert Systems with Applications* 182 (2021), p. 115276.

[63] Vasileios Baltatzis et al. "Neural Sign Actors: A diffusion model for 3D sign language production from text". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 1985–1995.

[64] Shengeng Tang et al. "Gloss-driven Conditional Diffusion Models for Sign Language Production". In: *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

[65] Sen Fang et al. "SignDiff: Learning Diffusion Models for American Sign Language Production". In: *arXiv preprint arXiv:2308.16082* (2023).

[66] Wuyan Liang and Xiaolong Xu. "Sindiff: Spoken-to-Sign Language Generation Based Transformer Diffusion Model". In: *Available at SSRN 4611530* ().

[67] Pan Xie et al. “G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 6. 2024, pp. 6234–6242.

[68] Leming Guo et al. “Distilling cross-temporal contexts for continuous sign language recognition”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 10771–10780.