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

LEVERAGING DIFFUSION MODELS AND MULTIMODAL FUSION FOR CONTINUOUS SIGN LANGUAGE RECOGNITION THROUGH MULTI-TASK LEARNING AND LANGUAGE MODELING

A Thesis Proposal Presented to the

DEANSHIP OF GRADUATE STUDIES

Submitted by:

Ahmed Abul Hasanaath
ID: G202302610

October, 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, but for individuals who are deaf or hard of hearing, traditional spoken language often poses significant challenges. 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 be affected [1]. SL (SL) serves as an essential tool for millions of hard-hearing people worldwide. Sign languages 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, making it an essential tool for the deaf and hard of hearing community. Unlike spoken languages, which rely on sounds and phonetics, sign languages 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. These components combine 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, often following 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. Much like spoken languages, sign languages differ widely across regions and cultures. There are hundreds of sign languages around the world, each with its own grammar, vocabulary, and linguistic nuances. 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 specific SL used within a particular community and 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. Within this field, there are two main approaches: gloss-based and gloss-free SLR. 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 SLR, 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 SLR attempts to directly map signs to spoken or written language without the use of intermediate glosses, a much more complex 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. Within the domain of gloss-based SLR, there are two primary subcategories: Isolated Sign Language Recognition (ISLR) and Continuous Sign Language Recognition (CSLR). In ISLR, the system is tasked with recognizing individual signs presented in isolation, meaning each sign is clearly separated, and there are no transitions or coarticulation effects between consecutive signs. 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 sequences of signs performed in continuous, fluid motion, reflecting how SL is used in natural 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. Additionally, CSLR models must handle varying signing speeds, signer-specific variations, and changes in context, all of which add to the complexity of recognition. These real-world factors make CSLR significantly more difficult to address than isolated recognition.

This thesis specifically focuses on CSLR, aiming to advance methods for recognizing sequences of signs in natural, unsegmented 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 not only the appearance but also the dynamic motion of signs, which is essential in recognizing how signs transition fluidly from one to the next in continuous signing. Next, the spatio-temporal features are 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 boundaries between individual signs. To address this issue, several techniques have been utilized in the literature, such as Hidden Markov Models (HMMs) and Connectionist Temporal Classification (CTC), both of which are designed to handle the temporal dynamics and sequential nature of continuous SL. Prior to CTC's widespread adoption, HMMs were used [2], but these required explicit segmentation, making them less adaptable to continuous input. CTC [3]

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 sequences. The 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 in a sequence. 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 or video to natural language text.

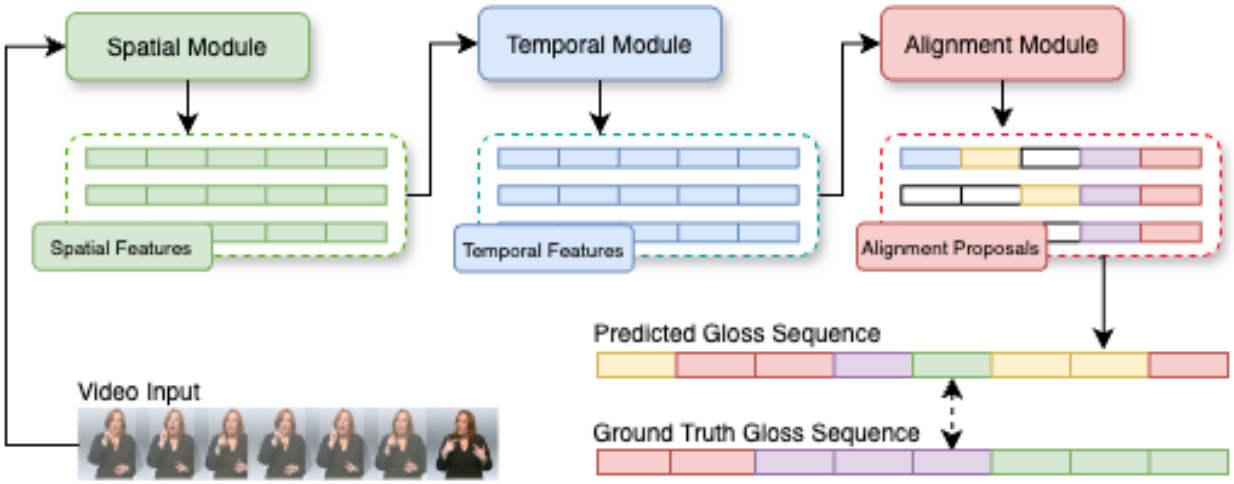

Figure 1.1: General CSLR Framework

The motivation for this thesis arises from the growing need for more accurate and robust systems capable of recognizing SL 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, 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-world 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 explores the utilization of Large Language Models (LLMs) along with Large Vision Models (LVMs) in a contrastive combination as a CSLR system. The LVMs in the proposed methodology will be built using the encoders from the pre-trained diffusion models. The pretrained encoders from the diffusion models also be 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 addressing the critical challenges in CSLR and serves as the primary focus of this thesis.

The thesis is organized into two key sections. Section 2 presents the literature review, where we explore available datasets, data acquisition methods, and various approaches to CSLR. In Section 3, we identify the research gaps and outline the proposed work. This section includes 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 project 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-of-hearing communities. CSLR plays a crucial role in facilitating this communication by translating fluid, unsegmented 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 subsection examines the various data acquisition methods and modalities utilized in the collection of sign language data. Understanding how these datasets are constructed is critical for evaluating the quality and applicability of CSLR systems. Furthermore, this subsection will delve into the distinct features of SL, including both manual and non-manual components, which are essential for accurate recognition. A detailed review of available datasets will also be provided, highlighting their characteristics and relevance to current research. In the second subsection, 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.

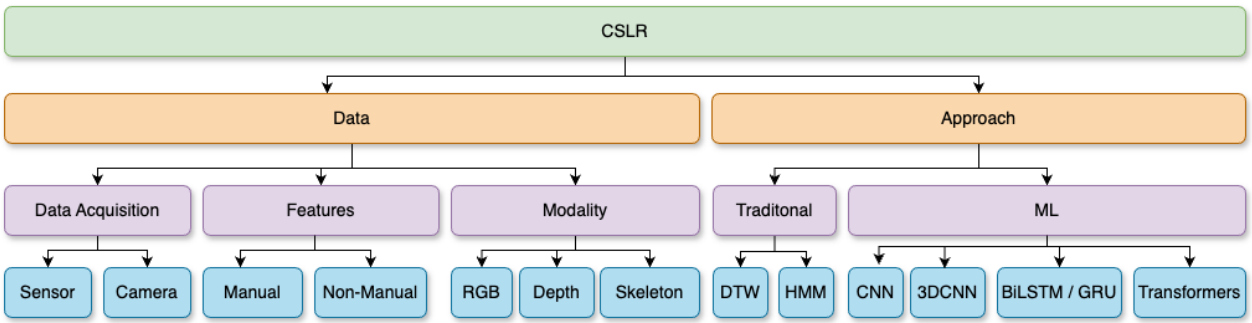

Figure 2.1: CSLR Taxonomy

2.1 Available Datasets

Effective data acquisition is fundamental to the success of SLR systems, which can be broadly categorized into vision-based and sensor-based methods. Vision-based approaches utilize cameras to capture signs as videos or images, 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, as seen in [4, 5, 6]. Figure 2.2 illustrates both sensor-based and camera-based data acquisition methods. Within camera-based systems, various modalities can be employed for data capture, each presenting unique advantages that cater to specific needs. The modalities include RGB, depth and skeleton. RGB-based systems utilize standard video data 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, as noted in [7]. 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 [8, 9, 10]. 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 [11, 12, 10]. To achieve accurate SLR, CSLR systems rely heavily on extracting meaningful features from the captured data, which can be categorized into manual and non-manual features. Manual features encompass essential hand-related movements crucial for identifying the core components of signs, such as hand shape, hand orientation, hand location, and hand motion [2]. In contrast, non-manual features capture movements of other body parts, including facial expressions and body posture, enriching the context and nuances of the sign language. These non-manual features include mouth motion, head nodding, body pose, facial expressions, and eye gaze, all of which can significantly influence the meaning or tone of the signs [2].

Publicly available datasets for SLR vary across several aspects, one of which is the category of the datasets. SLR datasets can be broadly classified into isolated sign language and continuous sign language. Isolated sign language datasets consist of discrete signs that exist outside of a conversational context, making them suitable for training models to recognize individual signs. In contrast, continuous sign language datasets capture recordings of complete signed sentences, reflecting the fluid nature of real-life communication. 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 [13, 14, 15, 16]. However, datasets that include sentences or conversational contexts are crucial for developing systems that can handle the complexities of natural sign language, including variations in signing speed, context, and simultaneous use of non-manual features. Language is another distinguishing factor among SLR datasets, with six major sign languages predominantly covered in CSLR research: German Sign Language (GSL) [17, 14, 15], Chinese Sign Language (CSL) [9, 18], American Sign Language (ASL) [13, 19, 20], and Arabic Sign Language [21, 16]. 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,

Figure 2.2: Illustration of Data Acqusition Methods in SLR. The first row illustrates 3 three different sensor-based wearables used for data acquisition. The second row illustrates the camera-based methods; (a) illustrates the RGB modality, (b) illustrates the depth modality and (c) illustrates the skeleton modality

and Arabic—highlighting their unique characteristics, components, and applicability for Continuous Sign Language Recognition. Table 2.1 summarizes these datasets.

2.1.1 American Sign Language Datasets

There are three prominent CSLR datasets for ASL. First, the Purdue RVL-SLL [13] dataset, developed by Purdue University's Robot Vision Lab (RVL), is a valuable resource for ASL recognition research. The dataset includes a large vocabulary of 104 unique ASL signs, specifically designed for continuous sign language recognition. Signed by 14 signers, it consists of approximately 600 video recordings of unique sentences that cover a wide variety of topics, making it suitable for both word-level and sentence-level sign language recognition tasks. Second, The RWTH-BOSTON-104 dataset [19], developed by the RWTH Aachen University in collaboration with Boston University, is a significant resource for ASL recognition, specifically designed to aid research in CSLR. It features a vocabulary of 104 distinct signs and contains 201 sentences, offering a diverse array of signed phrases for training machine learning models. Finally, The ASL-Homework dataset [20], created as part of a project at Boston University, is specifically designed for ASL recognition research in educational settings. It features a vocabulary of around 2048 signs and contains 6841 sentences, making it one of the more extensive resources available for continuous ASL sentence recognition. 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 Purdue RVL-SLL dataset covers the disaster domain whereas the other two do not have a particular domain.

2.1.2 German Sign Language Datasets

German CSLR datasets have been extensively used in literature as benchmark datasets to evaluate the proficiency of CSLR systems. RWTH-PHOENIX-Weather-2012 is among the first. Initially developed by [17], the RWTH-PHOENIX-Weather-2012 dataset was groundbreaking for German Sign Language (Deutsche Gebärdensprache, DGS) research, providing a large-scale resource of sign language data. It featured 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 [14], a significantly expanded version with an increased vocabulary, marking an evolution in SL dataset resources. 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 features real-world weather forecast recordings, which add a layer of complexity due to the natural conversational speed and diversity in signer expressions. 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 [15] 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.

2.1.3 Chinese Sign Language Datasets

The CSL dataset, introduced by [9] in 2016, is a Chinese Sign Language resource designed for CSLR research, consisting of approximately 100 sentences signed by 50 different individuals. Developed in a controlled lab environment, the dataset 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's relatively small number of unique sentences (100) limits its variability, which could potentially constrain the generalizability of models trained on it. However, the subsequent release of CSL-Daily [18] sought to address this limitation. CSL-Daily expands the vocabulary to 2,000 signs spread across 6,598 sentences, thus providing greater diversity and more realistic scenarios for training sign language recognition systems. The expansion of the dataset with CSL-Daily significantly broadened the scope of research possibilities within Chinese Sign Language recognition. The CSL dataset exists with RGB, depth and skeleton modalities whereas the CSL-Daily dataset only exists in the RGB modality.

2.1.4 Arabic Sign Language Datasets

The ArSL for Deaf Drivers dataset [21] is a specialized resource 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 signs, 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 [16] is an important resource for Arabic SLR covering the general domain area. It features 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. In the context of the arabic language, there is a need for the development of more comprehensive datasets.

Table 2.1: Summary of surveyed CSLR datasets

Dataset	Year	Sign Language	Signs	Modality	Domain
Purdue RVL-SLL [13]	2002	American	600	RGB	Disasters
RWTH-BOSTON-104 [19]	2007	American	201	RGB	General
RWTH-PHOENIX-Weather [17]	2012	German	2640	RGB	Weather
RWTH-PHOENIX-Weather-2014 [14]	2014	German	6841	RGB	Weather
RWTH-PHOENIX-Weather-2014-T [15]	2018	German	8257	RGB	Weather
CSL [9]	2019	Chinese	100	RGB, Depth, Skeleton	General
ArSL for Deaf Drivers [21]	2021	Arabic	215	RGB	Deaf Drivers
CSL-Daily [18]	2021	Chinese	6598	RGB	General
ArabSign [16]	2022	American	50	RGB, Depth, Skeleton	General
ASL-Homework [20]	2022	American	NA	(RGB, Depth)	General

2.2 CSLR Approaches

2.2.1 CNN and HMM

Convolutional Neural Networks (CNNs) and HMMs have 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, the authors of [22] 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 despite limited annotated data. In 2016, [23] introduced a hybrid CNN-HMM model for continuous sign language recognition, 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, [24] 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.

2.2.2 Capturing Global Context with RNNs

HMMs were utilized extensively in CSLR research [2]. However, HMMs lack the ability to capture global context, prompting researchers to shift towards RNNs. LSTMs, a type of RNN, have an innate ability to "remember" over longer contextual lengths, and the authors of [25] and [26] were the first to use LSTMs for CSLR. The proposed methodology in [25] introduced an architecture that consisted 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 [26] addresses the issue of noisy labels in video data, frequently overlooked in existing datasets. This approach treats training labels as weak labels, iteratively refining the label-to-image alignment in a weakly supervised manner. Similarly, the authors of [27] propose a weakly supervised framework for CSLR, focusing on cases where ordered gloss labels are available but exact temporal locations are not. Despite these advancements, existing approaches have struggled with efficiency, particularly concerning computational load and resource utilization. In 2023, [28] 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 task for determining frame resolution, enabling 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 not only improved throughput but also provided insightful visual analyses of spatial redundancy in CSLR datasets, further optimizing the recognition process.

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. The authors of [29] 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 was 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, the authors of [30] 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, [31] 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.

2.2.3 Multi-stream RNNs

Multi-stream RNN architectures have emerged as a pivotal approach in the CSLR landscape, effectively addressing the complexities inherent in processing sign language data. The authors of [9] introduce the Hierarchical Attention Network with Latent Space (LS-HAN), a framework 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 authors of [32] present TwoStream-SLR, a dual visual encoder that enhances 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 interaction techniques such as 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, the authors of [33] explore 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 like hand shapes and movements, introducing 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, the authors of [34] also propose the Contrastive Visual-Textual Transformation for SLR (CVT-SLR), addressing the weakly supervised nature of sign language recognition, 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. Together, 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.

2.2.4 Temporal Convolutions and 3DCNNs

Temporal convolutions have become an essential technique in CSLR, effectively addressing the intricacies of video data and the inherent challenges of recognizing SL glosses and their temporal boundaries [2]. Various works have been proposed that utilize temporal convolutions. In 2020, the authors of [35] proposed a cross-modal learning approach to enhance vision-based CSLR by integrating text information, which allows for improved modeling of intra-gloss dependencies. Their framework integrated temporal convolutions with 2DCNNs and BiLSTMs. They employed two robust encoding networks to generate video and text embeddings, which are then aligned into a joint latent representation, effectively producing more descriptive video-based features that are jointly classified with a decoder. Following this, the authors of [36] in 2021 tackled the issue of overfitting in vision-based CSLR by introducing a Visual Alignment Constraint (VAC). This innovative method enhances the feature extractor through alignment supervision with auxiliary losses that ensure the alignment of feature predictions. Building on these advancements, the authors of [37] in 2022 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. In 2023, the authors of [38] investigated the role of human body trajectories in CSLR through their proposed correlation network (CorrNet), which 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 study from 2023 introduced a temporal super-resolution network (TSRNet) [39]. This model incorporates frame-level and temporal feature extraction to minimize resource requirements while maintaining performance, positioning the TSRNet as a generator in an adversarial framework to enhance semantic information recovery.

Three-dimensional convolutional neural networks (3DCNNs) have emerged as a powerful tool in CSLR, effectively capturing spatial and temporal features from video data. Various works have been proposed that utilize 3DCNNs. In 2018, the authors of [40] introduced a novel 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 implemented an iterative optimization strategy that generated pseudo-labels for video clips, enhancing the feature representation of the 3D-ResNet. The subsequent year, the authors of [41] 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, while a temporal ensemble module integrated features across different time scales, further boosting recognition performance. In 2019, another significant contribution came from the authors of [42] with the Structured Feature Network (SF-Net), which learned multiple levels of semantic information from the data, effectively encoding frame, gloss, and sentence-level information into a unified feature representation. Building on these advancements, in 2021, the authors of [43] proposed a boundary-adaptive encoder that effectively captured the hierarchical nature of SL signals. Their method incorporated a location-based window attention model during decoding to enhance long sequence modeling and leveraged sign language subword units, thus addressing both isolated and continuous recognition within a unified framework.

2.2.5 Graph Convolutional Networks

Graph Convolutional Networks (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 [44]. 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, a 2022 paper proposed a Multi-View Spatial-Temporal Network that leverages GCNs to process RGB and skeleton data,

effectively capturing the intricate spatial-temporal dynamics inherent in sign language [45]. The network integrates a Multi-View Spatial-Temporal Feature Extractor 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.

2.2.6 Transformer Based Networks

Transformers have emerged as a transformative architecture in CSLR, significantly enhancing the modeling of temporal dynamics and contextual dependencies in sign language data. In 2021, the introduction of SignBERT [46] 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. Building upon this foundation, SignBERT+ [47] further refined the approach in 2023 by addressing overfitting tendencies through self-supervised learning and introducing multilevel masked modeling strategies. This allowed the framework to leverage existing data more effectively, enhancing the representation of sign language context while still utilizing a transformer-based architecture to model relationships across time. The trend continued in 2024 with the development of a multiscale temporal network designed to capture varying temporal features of sign language [48]. This network not only incorporated transformer modules for improved feature encoding but also innovatively utilized a multiscale temporal block (MST-block) to learn temporal features at different scales, significantly improving accuracy in CSLR. In a complementary effort, the introduction of SignCLIP [49] 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.

Vision Transformers (ViTs) have emerged as a powerful tool in CSLR, effectively addressing the challenges posed by the complex nature of SL and the limitations of traditional convolutional approaches. In 2022, the introduction of a multi-view spatial-temporal continuous sign language recognition network marked a significant advance in the field, leveraging a ViT-based encoder to learn long-term dependencies in sign language data [45]. This architecture integrates a Multi-View Spatial-Temporal Feature Extractor Network (MSTN) that captures spatial-temporal features directly from RGB and skeleton data, enhancing the model's ability to process the intricate dynamics of sign language. Building on this momentum, the following year saw the proposal of the Spatial-Temporal Transformer Network (STTN) [50], which innovatively encoded sign language videos into predicted sequences aligned with text. This model incorporated a chunking technique to manage computational complexity while extracting global and local features efficiently. Furthermore, in 2023, the Cross-modal Contextualized Sequence Transduction (C2ST) model [51] addressed the limitations of traditional CSLR frameworks by incorporating contextual knowledge from gloss sequences into video representation learning. This approach not only integrated linguistic features

but also introduced a contextualized sequence transduction loss, which improved alignment learning by overcoming the independence assumptions of conventional methods.

2.2.7 Generative Models for Sign Language

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. These models aim to capture complex patterns and structures within the data, enabling the creation of realistic samples that can be utilized in various applications, including 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.

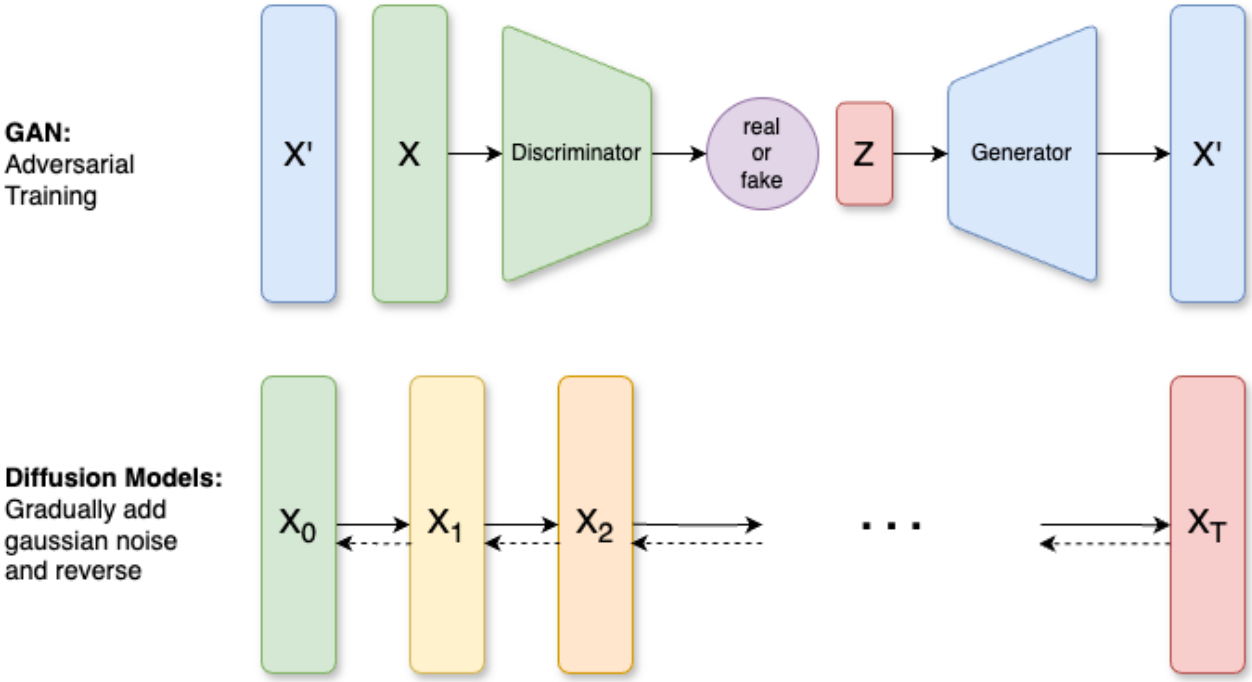

Figure 2.3: Illustration of GANs and Diffusion Models

GANs were first introduced in [52] 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 proven particularly effective in applications requiring high-quality data generation, such as image and video synthesis. On the other hand, diffusion models, first introduced in [53], 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. For instance, frameworks like Imagen Video [54] 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 [55]. 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. The review paper [56] from 2021 highlights only 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. The authors of [57] tackle the significant challenges of high-quality sign language video generation by introducing a novel framework called Dynamic GAN. This model addresses the prevalent issues of blurred effects and subpar video quality seen in previous methods, 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 methods that rely 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, the authors of [58] proposed hyperparameter-optimized GAN (H-GAN) to explore the classification of manual and non-manual gestures, addressing the complexities arising from the combination of hand, face, and body postures,60 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. Many existing SLP methods rely heavily on 2D data, which

limits the realism of the generated motions. In response, the authors of [59] propose 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, the authors of [60] introduce a Gloss-driven Conditional Diffusion Model (GCDM) that tackles the complexities of converting text or audio sentences into sign language videos. Furthermore, the authors of [61] address a critical gap in SLP by proposing SignDiff, a dual-condition diffusion pre-training model designed for continuous American Sign Language (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 [62] 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 [63] 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 study, establishing the foundation for the work. The Proposed Methodology follows, detailing the techniques and frameworks that will be employed to achieve these objectives. Additionally, the Expected Limitations subsection discusses potential challenges and constraints that could impact the research outcomes. Finally, the Project Timeline provides a structured schedule, mapping out the key phases of the project to ensure timely completion of each stage. Together, these subsections provide a comprehensive overview of the research approach, laying the groundwork for the successful execution of the study.

3.1 Gap Analysis

This section examines key challenges in the current research landscape of CSLR, identifying areas that have been underexplored and offer potential for innovation. One of the central issues is the limited utilization of language models, which have yet to fully capitalize on their ability to enhance CSLR systems by effectively handling the sequential and contextual aspects of SL. Additionally, the field has only begun to explore the potential of diffusion models, an emerging class of generative models with significant promise for improving CSLR performance. Another critical challenge is the scarcity of annotated CSLR datasets, which hampers the development and testing of robust systems. The complexity of fusing multiple data modalities—such as RGB, depth, and skeleton data—adds further difficulty in capturing the full spectrum of sign language. Moreover, multi-task learning (MTL) presents an opportunity to tackle these challenges by enabling models to learn various aspects of sign language, from hand gestures to facial expressions, simultaneously. Together, these research gaps provide important insights that will inform the direction of the proposed work.

3.1.1 Under-Exploitation of Language Modeling

Many studies treat CSLR primarily as a video understanding task, neglecting its aspect as a language modeling task. Few studies have explored this angle to improve gloss accuracy, with notable exceptions including [34], [46], [51], and [64]. Integrating language modeling techniques could provide a more nuanced understanding of sign language semantics, thereby enhancing CSLR systems' accuracy and effectiveness.

3.1.2 Opportunities with Diffusion Models

Diffusion models offer novel approaches to addressing several gaps in CSLR. One key gap is the lack of effective generative feature learning techniques. Generative models have the ability to model the underlying distribution of data. This ability can allow for enhanced feature extraction and representation and can also address the difficulty of learning from both high-dimensional and multimodal data such as RGB, depth, or pose information. Few studies have used diffusion models for SLR. The majority of the studies in the literature have used diffusion models for SLP [59, 60, 61, 62, 63]. For the purpose of CSLR, diffusion models can be utilized by pre-training the model. Leveraging diffusion models for pre-training the encoder on unconditional or conditional SL data could enhance recognition performance. This approach can be executed in a disjoint or multi-task setup, experimenting with reconstructing images or pose data, and incorporating masked modeling to enhance feature learning.

3.1.3 Lack of Annotated CSLR Datasets

The availability of annotated and diverse CSLR datasets is crucial for advancing research in CSLR. Currently, publicly available datasets are limited and primarily cover only a few SLs. For instance, the RWTH-PHOENIX-Weather-2014 dataset encompasses GSL with a relatively large vocabulary, while datasets like FluentSigners-50 focus on CSL and involve a considerable number of signers. However, many sign languages remain underrepresented in CSLR research, such as Arabic, Brazilian, and Indian Sign Languages, limiting the generalizability and inclusivity of current models [2]. Addressing this gap requires the development of new datasets for less-researched sign languages, which will help in understanding their specific features and creating CSLR models tailored to them.

3.1.4 Multimodal Fusion Complexity

Integrating multiple modalities such as RGB video, depth data, skeletal information and textual data, can enhance CSLR accuracy. Research indicates that combining RGB with pose data can achieve lower Word Error Rates (WERs), as seen in studies by [32] and [65]. However, multi-modal CSLR systems are computationally intensive and require further research to effectively fuse these diverse data sources without introducing excessive computational burdens. Developing methods to streamline this integration is crucial for practical deployment.

3.1.5 Potential of Multi-Task Learning

MTL involves training models on related tasks simultaneously to exploit shared information and improve generalization. In CSLR, joint training with SLT has shown promise, as both tasks benefit from shared visual features [66]. Collaborative approaches, such as Multilingual CSLR, have demonstrated success by leveraging shared low-level visual patterns across different SLs, enabling more generalizable and robust models [67]. By incorporating datasets from multiple SLs, models can learn common patterns in hand shapes, movements, and facial expressions that transcend individual languages, enhancing their ability to generalize across different linguistic and cultural contexts. This approach can be particularly valuable in improving the performance of CSLR systems for underrepresented or low-resource sign languages by transferring knowledge from larger, more annotated datasets of other sign languages. Moreover, utilizing these multilingual datasets allows models to capture a wider variety of signing styles and variations, thereby boosting recognition accuracy and making the systems more adaptable to new or unseen languages. Furthermore, the recent advancements in LVMs open up promising opportunities for MTL in CSLR. For instance, leveraging human video pre-training for robotic policy learning, as demonstrated in the unified discrete diffusion approach, could be adapted for CSLR to address the scarcity of annotated data by learning from vast, unlabeled SL videos [55]. Similarly, the capabilities of text-conditioned video generation frameworks like Imagen Video offer insights into how fine-grained sign language features, such as hand movements and facial expressions, could be modeled with high fidelity, improving recognition accuracy [54]. The DiffSLVA framework introduces novel techniques for anonymizing sign language videos while preserving linguistic content, utilizing pre-trained large-scale diffusion models for zero-shot text-guided tasks [68], which demonstrates the versatility of these models in handling complex multimodal tasks without large annotated datasets. Additionally, in this context, the development of models such as InternVL [69], which scales up to 6 billion parameters and aligns with large language models using extensive image-text data, can further enhance CSLR systems by achieving state-of-the-art performance on various visual-linguistic tasks. Finally, models like SignCLIP, which aligns spoken language text with sign language videos in a shared space through contrastive learning, points to the potential of MTL where various SL processing tasks, such as recognition and translation, could benefit from unified learning architectures [49].

3.2 Research Objectives

The primary goal of this thesis is to advance research in CSLR through the application of diffusion models and the exploration of key 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 diffusion models, language modeling, MTL, and multimodal fusion techniques. This review will also cover existing datasets, annotation methods, and performance metrics used in CSLR, with a particular emphasis on addressing the research gaps in the field.
2. **Development of Diffusion Model Techniques for CSLR:** This research will

explore how diffusion models can be effectively applied to CSLR tasks. This includes pre-training encoders with diffusion models, experimenting with multi-modal fusion using these models, and investigating contrastive learning approaches for CSLR. The outcomes will provide novel methodologies for integrating generative models into CSLR.

3. **Exploring Multimodal Fusion Complexity:** This objective will focus on optimizing multimodal fusion techniques by integrating RGB video, pose, and textual information to improve CSLR performance without excessive computational overhead. The research will evaluate different fusion strategies and diffusion models to balance computational efficiency and accuracy.
4. **Exploring MTL:** This thesis aims to implement MTL approaches that combine CSLR with related tasks such as gesture recognition and multilingual training. The goal is to create models capable of sharing information across tasks to improve generalization and performance in CSLR.
5. **Incorporating Language Modeling into CSLR:** This research will investigate the integration of advanced language modeling techniques such as LVMs into CSLR systems to improve the semantic and syntactic accuracy of gloss prediction. By exploring models that treat CSLR as both a video understanding and a language modeling task, the study will develop more holistic and accurate recognition systems.
6. **Performance Evaluation and Optimization:** The thesis will involve extensive experimentation to determine which approaches yield the best results in CSLR. This includes optimizing diffusion model techniques, multimodal fusion, and MTL approaches for real-world applications. Comparative studies will be conducted to establish the most effective methods for CSLR across different datasets and tasks.
7. **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

Figure 3.1 illustrates a high-level overview of the proposed methodology, which integrates diffusion models, multimodal fusion, MTL for CSLR. The methodology consists of two distinct training phases.

Phase 1: This phase focuses on pretraining two distinct diffusion models. The goal of this phase is to pretrain an encoder on a generative task and explore how the integration of MTL can improve the representation of sign language features in the diffusion model encoders. Specifically, two diffusion models will be trained. The first model will be trained on data with a *dense temporal stride* (referred to as the fast pathway) and another on data with a *sparse temporal stride* (referred to as the slow pathway). These two models are designed to focus in different aspects of the data. The *fast pathway* will focus on capturing *temporal features*

by processing frequent, smaller time intervals, while the *slow pathway* will prioritize learning *spatial features* while still considering temporal dependencies. In diffusion models, the data x is progressively corrupted with noise, modeled by the forward process $q(x_{t-1}|x_t)$, where t is the time step, and x0 represents the clean data. The goal of the model is to reverse this process using a learned model $p_\theta(x_{t-1}|x_t)$, parameterized by θ. In this case, the diffusion model learns to denoise sign language features by minimizing the variational lower bound (VLB) of the negative log-likelihood:

$$\mathcal{L}_{VLB} = \mathbb{E}[\mathrm{DKL}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t))] \tag{3.1}$$

Phase 2: This phase leverages the pretrained encoders from the diffusion models for feature extraction to train a CSLR classifier. The extracted spatio-temporal features are fed into two distinct architectures for comparison. The first architecture integrates the diffusion model encoders with a language model, which is trained contrastively with the encoders. The contrastive loss encourages the model to produce similar representations for semantically related gloss sequences and sign sequences while differentiating unrelated ones. The contrastive loss $L_{contrastive}$ can be formulated as:

$$\mathcal{L}_{contrastive} = -\log\left(\frac{\exp(\mathrm{sim}(z_i, z_j))}{\sum_k \exp(\mathrm{sim}(z_i, z_k))}\right) \tag{3.2}$$

where z_i and z_j are the embeddings of related sign and gloss sequences, and sim denotes a similarity function such as cosine similarity. The second architecture employs the diffusion model encoders combined with a CTC loss for alignment. CTC aligns the predicted glosses with the ground truth sequence by maximizing the probability of the correct gloss sequence y given the input features x. CTC loss calculates the negative log likelihood of the correct gloss sequence y given the predicted sequence x:

$$L_{CTC} = -\log P(x|y) \tag{3.3}$$

During evaluation, the contrastive learning model (diffusion encoders + language model), and the CTC-based model (diffusion encoders + classifier with CTC alignment). will be compared. The evaluation will assess how well these models generalize to unseen sign language sequences, taking into account the performance on temporal and spatial aspects of the data. This detailed exploration of the two training phases will help to identify the most effective method for tackling the unique challenges of CSLR. This section will elaborate on the key components of the proposed framework.

3.3.1 Diffusion Model Pre-training

Diffusion models will be used in this work to improve feature learning. This will be achieved by pre-training diffusion models in the following ways:

- **Unconditional or Conditional Pre-training:** In phase 1, the diffusion models will be pre-trained using both conditional and unconditional SL data by employing the

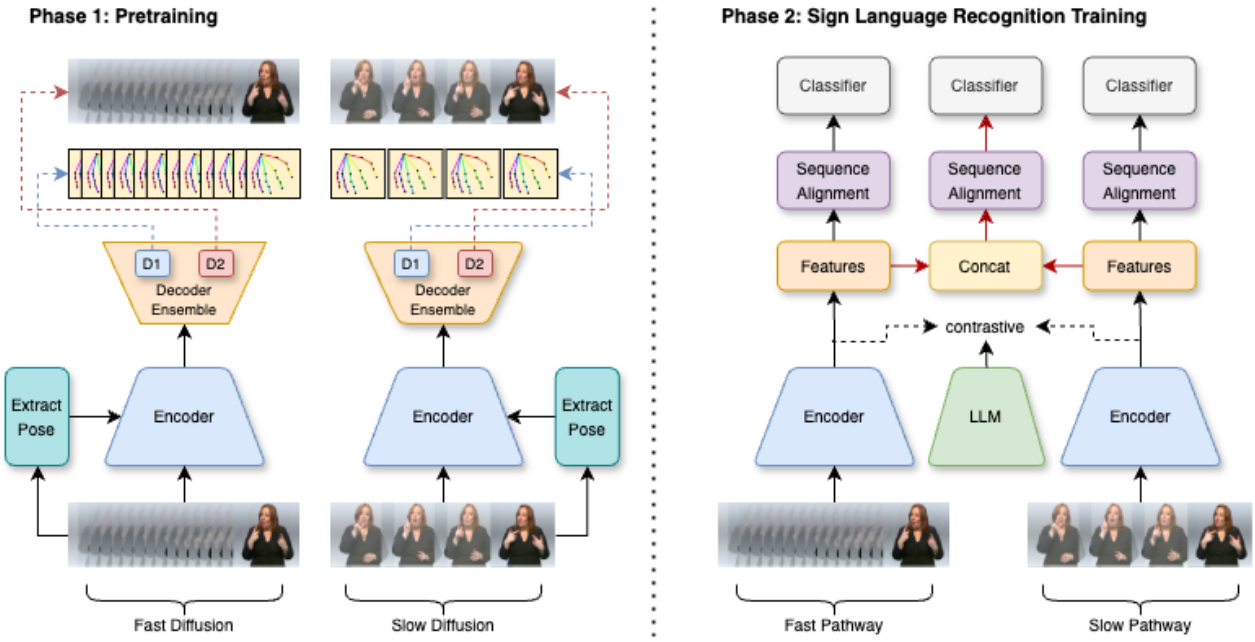

Figure 3.1: High level overview of the proposed methodology

Classifier Free Guidance (CFG) concept [70]. The unconditional setup will involve modeling the general distribution of sign language data, while the conditional setup will include the use of labels or glosses to guide the learning process. The goal is to use the diffusion models as a pre-training mechanism to enhance the encoder's ability to extract meaningful features from both pose and RGB data.

- **Feature Learning:** In phase 2, the pre-trained encoder from the diffusion model will be utilized to extract temporal and spatial features, focusing on improving the ability to capture subtle variations in sign language gestures across different signers.

3.3.2 Dual-Stream Network (SlowFast Pathways)

The proposed architecture will incorporate a dual-stream network, similar to the SlowFast model [33]. The fast pathway will focus on capturing fine-grained temporal changes in the video, while the slow pathway will capture more global, spatial features. This design will further help the model deal with both short-term and long-term temporal dependencies, ensuring a comprehensive representation of SL features.

3.3.3 Contrastive Learning with Language Modeling

Incorporating language modeling is critical for understanding the linguistic structure of SL. The proposed framework will adopt a contrastive training approach where the diffusion model encoder will be trained contrastively alongside a text encoder. This process will allow the model to learn a shared representation between sign language sequences and their corresponding glosses or textual descriptions. By aligning SL video features with textual

features, the model will capture the language structure of the signs, improving its ability to generate glosses and bridge the gap between video understanding and language modeling.

3.3.4 Multimodal Fusion

One of the major challenges in CSLR is the integration of multiple data modalities, such as RGB video and pose information. To address this, the research will focus on pre-training with multimodal data. The diffusion model encoder will be pre-trained using both pose and RGB data to create a unified representation of sign language gestures. This multimodal fusion is expected to enhance the model's ability to capture both the appearance and motion aspects of SL. The fused features will provide richer input for downstream tasks, such as gloss generation and recognition, improving the model's robustness across different signers and environments.

3.3.5 Multi-task Learning

To further enhance generalization and performance, the methodology will incorporate a MTL setup where the model will be trained to perform multiple tasks simultaneously:

- **Task 1: RGB Sign Generation:** The diffusion model decoder will be trained to generate realistic RGB video sequences of signs, focusing on reconstructing the original SL data.
- **Task 2: Pose Data Generation:** Simultaneously, the diffusion model decoder will be trained to generate corresponding pose data (i.e., skeletal representations of signers) based on the input video sequences. This multi-modality learning improve the overall performance on both video generation and recognition tasks.

3.3.6 Model Evaluation

The trained models will be evaluated on multiple CSLR benchmarks, focusing on performance metrics such as WER, accuracy, and generalization across unseen signers. Experiments will be conducted to assess the contribution of diffusion models in both single-task and multi-task settings, as well as the effectiveness of multimodal fusion and contrastive learning with language models.

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:

1. **Computational Complexity:** The integration of diffusion models, multi-modal fusion, and MTL increases computational demands. Training models that incorporate temporal and spatial features, especially using diffusion models, requires significant computational power and memory. This could limit the size of experiments or slow down the research process, particularly when experimenting with large-scale datasets or complex multi-task architectures.

2. **Multimodal Fusion Complexity:** While multimodal fusion of RGB, pose, and textual data is expected to improve recognition accuracy, it also adds complexity to the model. Efficiently integrating these modalities without causing computational overhead or diminishing performance remains a challenge. Additionally, managing missing or noisy data from one or more modalities may affect the system's robustness.
3. **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.
4. **Diffusion Model Limitations:** While diffusion models show great promise, their application to CSLR is relatively novel and unexplored. Issues related to model convergence, learning from temporal and spatial features, and effectively combining generative and discriminative tasks in a multi-task setting may emerge. Additionally, the long training times and complexity associated with diffusion models may limit experimentation and fine-tuning.
5. **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, such as refining multi-task learning models or fully exploring diffusion models, may require further exploration beyond the timeframe of this thesis.

3.5 Project Timeline

The timeline outlines the systematic approach that will be taken in order to ensure timely completion of this project. The project is divided into distinct 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 summarizes the timeline.

Table 3.1: Project Timeline

Phase	Tasks	Deadline
Setup	Build a simple model for CSLR using 2D CNN Replace with 3D CNN Add 2 stream RNN	End of October 2024
Phase 1	Build a single stream multi-task diffusion model for RGB, pose, or both Pretrain Fast and Slow Diffusion	End of November 2024
Phase 2	Train a CTC decoder with/without LLM	End of December 2024
Misc	Try experimenting with GNNs for processing pose data	End of January 2025
Evaluation, Ablation, and Benchmarking	Conduct evaluations, ablation studies, and benchmarking	End of February 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] Alex Graves and Alex Graves. "Connectionist temporal classification". In: *Supervised sequence labelling with recurrent neural networks* (2012), pp. 61–93.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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).

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

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

[44] 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.

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

[46] 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.

[47] 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.

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

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

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

[51] 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.

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

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

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

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

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

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

[63] 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.

[64] 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.

[65] Ronglai Zuo and Brian Mak. “Improving continuous sign language recognition with consistency constraints and signer removal”. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 20.6 (2024), pp. 1–25.

[66] Necati Cihan Camgoz et al. “Sign language transformers: Joint end-to-end sign language recognition and translation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10023–10033.

[67] Hezhen Hu et al. “Collaborative multilingual continuous sign language recognition: A unified framework”. In: *IEEE Transactions on Multimedia* 25 (2022), pp. 7559–7570.

[68] Zhaoyang Xia, Carol Neidle, and Dimitris N Metaxas. “DiffSLVA: Harnessing Diffusion Models for Sign Language Video Anonymization”. In: *arXiv preprint arXiv:2311.16060* (2023).

[69] Zhe Chen et al. “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 24185–24198.

[70] Jonathan Ho and Tim Salimans. “Classifier-free diffusion guidance”. In: *arXiv preprint arXiv:2207.12598* (2022).