DIFFSIGN: A DIFFERENT APPROACH TO CONTINUOUS SIGN LANGUAGE RECOGNITION WITH DIFFUSION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal alignment is a general way for continuous sign language recognition (CSLR) tasks. However, Due to the weakly supervised nature of CSLR, manual alignment often fails to map sign frames to glosses accurately. In this paper, we propose a diffusion-based framework, achieving CSLR in a new view based on cross-modal generation, leveraging the inherent semantic consistency between sign videos and glosses. To address the issue of ambiguous boundaries in sign videos, we have also developed a contrastive learning-based feature enhancement strategy, which serves as a more sophisticated alternative to the simple attention mechanisms commonly used in text-to-image generation tasks. Extensive experiments on three public sign language recognition datasets demonstrate the effectiveness of generation way in CSLR and it can achieve better performance than state-of-the-art methods. The code of our method will be available upon acceptance.

1 INTRODUCTION

Sign language is an essential communication bridge between deaf and hearing individuals. Efforts to map sign language videos to textual glosses, known as sign language recognition, have garnered significant interest recently. This field can be divided into isolated sign language recognition (ISLR) and continuous sign language recognition (CSLR), depending on the number of signs in a video. Given its closer alignment with real-life situations, CSLR has attracted more attention from researchers.

From a machine learning perspective, CSLR can be considered a weakly supervised task, as it lacks specific gloss-level annotations for each sign. To tackle this challenge, techniques such as cross-modal alignment (Chen et al., 2024a; Pu et al., 2019) and explicit consistency constraints (Min et al., 2021; Zuo & Mak, 2022) have been utilized. However, in a weakly supervised context, achieving frame-level alignment is an ill-posed problem and can result in insertion or deletion errors (Park et al., 2008), adversely affecting the final recognition performance. Meanwhile, similar to verbal language, the combination of sign words is dynamic, leading to varied transitions between signs. In such scenarios, alignments trained on limited data may not generalize well to broader settings.

As incorrect alignment may introduce some significant challenges to CSLR, an alternative can be minimizing manual intervention and fully leveraging the inherent semantic relationship between sign videos and gloss sequences. In general, the CSLR process is to transfer a video containing a series of consecutive signs to a sequence of natural language words, namely glosses, making it a form of cross-modal generation task. While numerous generative models like GANs (Radford et al., 2015) and VAEs (Kingma & Welling, 2013) exist, they may not be ideally suited for the complex demands of video-to-text generation. Unlike most image-to-image style transfer tasks, CSLR must adhere to restricted ground truth for glosses, and the differences between the input and output in CSLR are substantial. For example, CSLR must transform sign video inputs into textual outputs, presenting challenges that could lead to mode collapse or gradient vanishing issues, especially when using complex training strategies like those employed in GANs. Instead of focusing solely on low-level pixel-wise matching, the semantic information should be used more sufficiently in such a cross-modal generation. In other words, we should learn more about the semantic consistency of

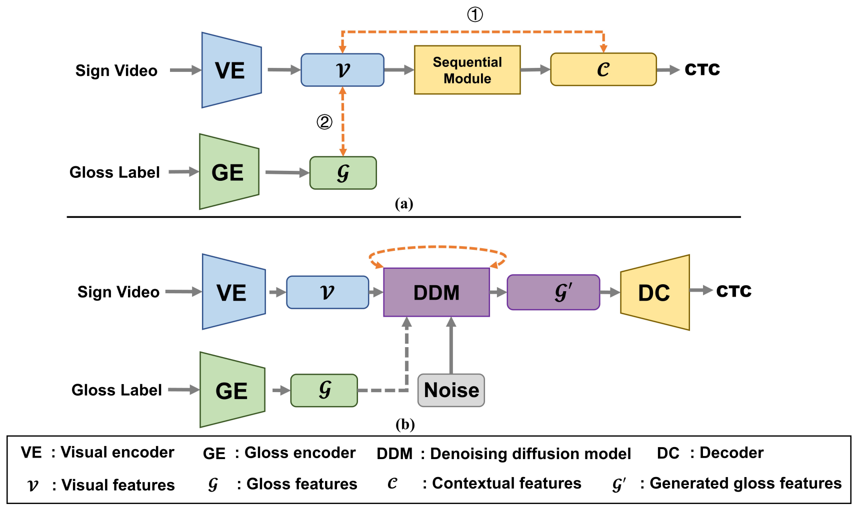

Figure 1: (a) A mainstream CSLR framework based on cross-modal alignment of fixed patterns. ① denotes the approaches such as self-distillation and iterative training, while ② denotes the means like sentence-level contrastive learning and knowledge distillation. (b) Our proposed CSLR framework based on cross-modal generation. Gloss features are provided only at training stage. We transform CSLR into a cross-modal generation, fully exploring the potential of diffusion models in cross-modal feature correlation.

sign videos and glosses in their data distribution. Denoising diffusion models (DDMs) have emerged as a promising solution, often producing higher quality samples and garnering increased attention from researchers. In the process of noising and denoising, DDM focuses not on individual samples but on the distribution itself. This adaptive exploration of cross-modal feature relationships is what we need to obtain accurate glosses. As illustrated in Fig.1, by viewing CSLR as a cross-modal generation task, the dependence on gloss-level labels becomes unnecessary.

Another challenge is handling the uncertainties and ambiguities inherent in certain sign video segments. As proposed in (Rombach et al., 2022), attention injection is utilized to capture crucial information for enhanced generation performance. Unlike typical text-to-image generation tasks, CSLR lacks an explicit mapping framework. Given the nature of sign videos, some frames serve merely as transitions between two sign actions and lack semantic content. These transitional segments might inadvertently produce features similar to non-existent sign actions in the video, a phenomenon exacerbated by coarticulation. Such ambiguous features present challenges for DDMs in accurately interpreting visual-textual associations and generating the correct gloss sequences. To address this, it is essential to refine the feature representation to better capitalize on semantic correlations. Considering that video clips and gloss text represent two facets of the same sign word, we can leverage this semantic link by enhancing the distinction between transitional features and all glosses, while narrowing the gap between clearly semantic features and their corresponding glosses.

The main contributions are summarized as follows:

- We provide a novel view that transferring CSLR into a cross-modal generation task, and propose a DDM-based generation framework DiffSign, which avoids the wrong predictions caused by inaccurate alignments in traditional CSLR methods.
- We propose a gloss-level feature enhancement method based on contrastive learning to alleviate the semantic ambiguity present in visual features, ensuring a clear distinction among the visual features that represent different glosses.
- The proposed DiffSign achieves state-of-the-art results on three widely used CSLR datasets (PHOENIX-2014, PHOENIX-2014T, CSL-Daily). Sufficient ablation experiments are demonstrated, providing interpretability and reproducibility for the proposed architecture.

2 Related Work

2.1 Continuous Sign Language Recognition

Recent CSLR methods (Min et al., 2021; Zuo & Mak, 2022; Guo et al., 2023) have primarily focused on achieving more accurate correspondence between video segments and their corresponding glosses. (Zuo & Mak, 2022; Cheng et al., 2023) applies cross-modal contrastive learning at the sentence-level, improving the alignment globally. To achieve more accurate alignment, some methods use iterative training (Cui et al., 2019; Pu et al., 2019), knowledge distillation (Min et al., 2021), or gloss-level cross-modal contrastive learning (Chen et al., 2024a) to accomplish finer-grained alignment. (Chen et al., 2024a) uses Dynamic Time Warping (DTW) to establish the correspondence between visual features and gloss features. It forms clip-gloss pairs through a fixed similarity measurement. Due to the absence of gloss-level labels, alignment becomes an ill-posed problem, making it challenging to achieve precise correspondences through fixed patterns. Therefore, in our approach, we no longer perform cross-modal alignment. Instead, we transform CSLR into a cross-modal generation task. We explore the relationships between cross-modal features during the denoising process of DDM, obtaining more precise gloss sequences through generation.

2.2 Denoising Diffusion Model

The DDM includes a forward Gaussian diffusion noising process and a reverse denoising generation process. It can iteratively denoise the input Gaussian noise under the guidance of guiding information, ultimately generating a target aligned with the guiding information. DDMs have demonstrated impressive performance in cross-modal generation, such as text-to-visual generation with models like Stable Diffusion (Rombach et al., 2022), LGD (Song et al., 2023), and UniDiffuser (Bao et al., 2023). Additionally, the visual-to-text task of image captioning, generative approaches are also utilized (Luo et al., 2023). Whether it is ”text-to-image” or ”image-to-text,” the core lies in the correspondence of cross-modal features. High-quality correspondence of cross-modal features is essential for achieving text-based image editing. Therefore, DDMs have a strong ability to explore and learn the relationships between cross-modal features which is suitable for forming clip-gloss correspondence in CSLR. Additionally, (Chen et al., 2024b) points out that applying DDMs as denoising autoencoders to recognition tasks can extract linearly separable representations of images. Based on this idea, (Zheng et al., 2023; Guo et al., 2023) apply generative models such as VAEs and DDMs as denoising autoencoders to the CSLR task in order to optimize the visual representation. Since they utilize generative models as denoising autoencoders, there is no longer a need for guiding information from other modalities. This results in a loss of cross-modal characteristics and fails to fully explore cross-modal feature associations. In our approach, we pioneeringly complete the CSLR task through a cross-modal generation process and fully leverage the ability of DDMs in cross-modal feature association.

3 Proposed Method

3.1 Preliminaries

Suppose we have a sign language video, which is encoded to $\mathcal{V} \in \mathbb{R}^{N \times D}$. And the gloss label is encoded to a fixed-length sequence $\mathcal{G} \in \mathbb{R}^{N' \times D}$ of length N', where D denotes the feature dimension. Then our target is to generate $\mathcal{G}'$ from noise with the guidance of $\mathcal{V}$.

3.2 Diffusion Model based Gloss Generation

3.2.1 Diffusion Denoising Detail

As the sign video contains tens to hundreds of frames, directly applying DDMs at the pixel level incurs significant costs in terms of computational resources and time. Then we perform the noising and denoising at the feature level with the latent diffusion model (LDM). The architecture of our DiffSign is illustrated in Fig. 2. The diffusion process of LDM is the same as that of pixel-level DDM (Ho et al., 2020), except that we are adding noise to the gloss features $\mathcal{G}$. We progressively add

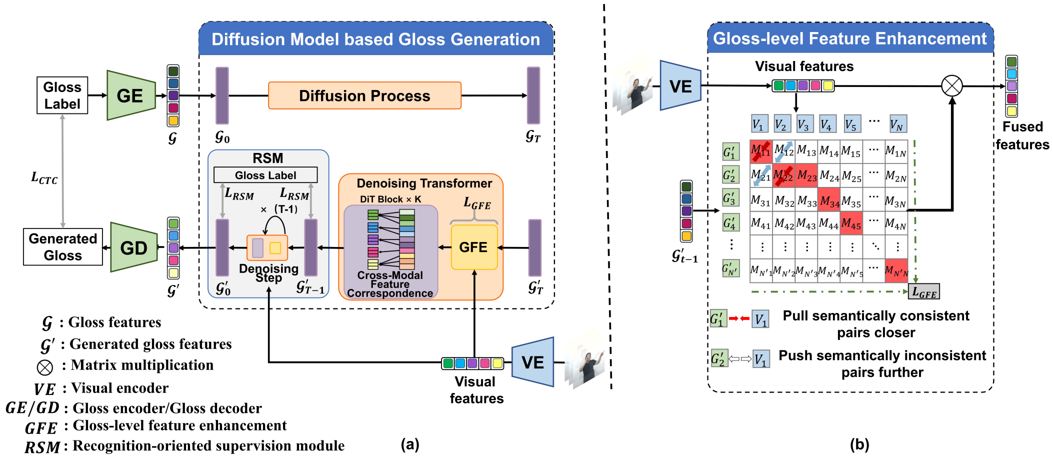

Figure 2: An overview of our proposed framework. (a) The core idea of transforming the CSLR into cross-modal generation is demonstrated, leveraging the powerful ability of diffusion models in understanding cross-modal feature correlations, obtaining more accurate gloss sequences. (b) It demonstrates the implementation details of gloss-level feature enhancement and how it alleviates the ambiguity and uncertainty issues in visual features through gloss-level contrastive learning.

multi-level Gaussian noise to the gloss features in a Markov chain manner to obtain $\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_T$. Each step can be represented as:

$$q(\mathcal{G}_t|\mathcal{G}_{t-1}) = \mathcal{N}(\mathcal{G}_t; \sqrt{1-\beta_t} \cdot \mathcal{G}_{t-1}, \beta_t \mathbb{I}), \tag{1}$$

where $\mathcal{N}$ denotes the normal distribution and β_t is the weight controlling the noise for step t. Therefore, given $\mathcal{G}_0(\mathcal{G})$, $\mathcal{G}_t$ can be expressed as:

$$\mathcal{G}_t = \sqrt{\overline{\alpha_t}} \cdot \mathcal{G}_0 + \sqrt{1-\overline{\alpha_t}} \cdot \varepsilon, \tag{2}$$

where ε represents Gaussian noise. We set the parameters for all diffusion steps as follows: $\beta_t \in (0, 1)$ for $t = 1, \ldots, T$, $\alpha_t = 1 - \beta_t$, and $\overline{\alpha_t} = \prod_{k=1}^{t} \alpha_k$.

During the denoising phase, the Gaussian noise $\mathcal{G}_T^{'}$ is combined with the visual guidance $\mathcal{V}$ through cross-attention. Utilizing the predicted noise from the DDM, the process iteratively reduces noise to generate the gloss features $\mathcal{G}^{'}$. The entire denoising process is carried out in the manner of DDPM (Ho et al., 2020) and can be represented as:

$$p_\theta(\mathcal{G}_{0:T}^{'}) = p(\mathcal{G}_T^{'}) \cdot \prod_{t=1}^{T} p_\theta(\mathcal{G}_{t-1}^{'}|\mathcal{G}_t^{'}), \tag{3}$$

where θ represents the parameters of the model. Each step of denoising follows a normal distribution, which can be represented as:

$$p_\theta(\mathcal{G}_{t-1}^{'}|\mathcal{G}_t^{'}) = \mathcal{N}(\mathcal{G}_{t-1}^{'}; \mu_\theta(\mathcal{G}_t^{'}, t), \Sigma_\theta(\mathcal{G}_t^{'}, t)). \tag{4}$$

The mean $\mu_\theta(\mathcal{G}_t^{'}, t)$ and variance $\Sigma_\theta(\mathcal{G}_t^{'}, t)$ denote the model's predictions for the reverse process, which adhere to a normal distribution. For the reason of simplifying the model and providing more accurate predictions, we only use the model to predict the mean, while the variance is set to a fixed value $\Sigma_\theta = \beta_t \mathbb{I}$.

Since we try to achieve cross-modal generation, the traditional backbones like U-Net for image-oriented noise prediction are not suitable. Inspired by (Peebles & Xie, 2023) who serializes image features and performs diffusion-denoising in the latent space, we employ diffusion transformer (DiT) for denoising. Based on the idea of directly transforming CSLR into cross-modal generation to obtain gloss features with clear boundary information, we need to generate the gloss features $\mathcal{G}^{'}$

using the diffusion model guided by the visual features $\mathcal{V}$. The entire "video-to-text" process can be modeled as:

$$p_\theta(\mathcal{G}^{'}|\mathcal{V}) = \prod_{t=1}^{T} p_\theta(\mathcal{G}^{'}_{t-1}|\mathcal{G}^{'}_t, \mathcal{V}), \tag{5}$$

where $p_\theta(\mathcal{G}^{'}_{t-1}|\mathcal{G}^{'}_t, \mathcal{V})$ not only represents the recovery of the gloss features during the iterative sampling process but also signifies the model's continuous learning of the cross-modal contextual information between the input gloss features $\mathcal{G}^{'}_t$ and the visual guidance $\mathcal{V}$. The entire process integrates the embeddings in $\mathcal{V}$ that represent the same sign action and uses them to update the corresponding token in $\mathcal{G}^{'}_t$. After generating $\mathcal{G}^{'}$, we use a transformer-based decoder to decode the predictions. Through this "video-to-text" approach, we have innovatively explored the potential of diffusion models in cross-modal feature correspondence, resulting in gloss features that contain richer cross-modal contextual information, and ultimately decoding more accurate gloss sequences.

3.2.2 Recognition-oriented Supervision Module

Although the powerful capability of diffusion model has been proven in generating images and videos (Ho et al., 2022; Rombach et al., 2022), it is still challenging to adapt it to high-level vision tasks like CSLR. As the measurement for generation quality like FID mainly focuses on pixel-wise information, during the later stages of the iterative sampling process only low-level supervision is given. It makes the generation quality of features negatively correlated with the recognition accuracy (Chen et al., 2024b). These low-level features are not useful for generating gloss labels, and we need to make the diffusion model pay more attention to semantic and other high-level features during the sampling process. Therefore, we modify the denoising module of the diffusion model, and add a step-wise constraint, which can be formulated as:

$$\mathcal{L}_{RSM} = \sum_{t=0}^{T} \delta_t \cdot [-\log p(\mathbb{G}|\mathcal{G}^{'}_t; \theta)], \tag{6}$$

where $\mathbb{G}$ denotes the ground truth gloss label. It can apply connectionist temporal classification constraints between the gloss features $\mathcal{G}^{'}_t$ and the gloss label. In the initial few iterations, since the generated gloss features $\mathcal{G}^{'}_t$ is mostly noise, we use δ_t to balance the loss values at different stages to prevent excessive gradients and ensure stable training.

3.3 Gloss-level Feature Representation Enhancement with Contrastive Learning

Continuous sign language videos, as raw signals, are influenced by coarticulation, leading to the presence of semantically ambiguous features in the visual sequences. These features, occurring in the transitional parts between adjacent sign actions, may resemble certain gloss features due to changes in sign actions, even though they do not inherently contain semantic information. To ensure differences between features representing different sign actions in visual features are more distinct and to help the diffusion model understand the visual-textual relationship, we propose gloss-level feature enhancement based on contrastive learning.

We choose to perform gloss-level feature enhancement during the fusion of $\mathcal{V}$ and $\mathcal{G}^{'}_t$ at the sampling stage. This is because during the training stage of the diffusion model, the intensity of adding noise is random, and we cannot effectively balance the losses generated by each batch under the gloss-level feature enhancement constraint. The recognition-oriented supervision module is added during sampling stage for the same reason. Due to our use of cross-attention to combine $\mathcal{V}$ and $\mathcal{G}^{'}_t$, a dot product-based attention matrix $M \in \mathbb{R}^{N^{'} \cdot N}$ is generated. Here, M_{ij} represents the similarity between the i-th token and the j-th embedding. In the CSLR task, this indicates the probability that the i-th token and the j-th embedding represent the same sign action. As shown in fig.3, we select the token with the highest similarity for each embedding to form positive pairs and pair it with the remaining tokens to form negative pairs. For those transitional segments in continuous sign language videos that do not contain semantic actions, there is no need to select corresponding tokens for the embeddings obtained from them. We have set a similarity threshold τ, and when the similarity between a certain embedding and all tokens is less than τ, we consider that this embedding is derived from transitions encoding, and no longer match tokens for it. Therefore, during gloss-level contrastive learning, this embedding (anchor) has no positive sample pairs and forms negative sample pairs with all tokens to increase the distance between them. By using this method to create a gap between sign language actions, it helps the model understand the relationship between visual embeddings and gloss tokens, leading to the generation of a more accurate gloss representation $\mathcal{G}'$. Due to the high noise and low reliability of the gloss tokens generated in the initial sampling stage, we still use δ_t as weights to balance the losses in different sampling stages. The constraint of gloss-level feature enhancement can be represented as:

$$\mathcal{L}_{GFE} = -\log \sum_{t=0}^{T} \sum_{i=1}^{N} \frac{\delta_t \cdot exp(\mathcal{V}_i \cdot \mathcal{G}'_{t+})}{\sum_{j=1}^{N'} exp(\mathcal{V}_i \cdot \mathcal{G}'_{tj})}; \tag{7}$$

where $\mathcal{G}'_t$ represents the gloss token sequences generated at the t-th sampling step, $\mathcal{G}'_{t+}$ denotes the token in $\mathcal{G}'_t$ that serves as a positive sample, and $\mathcal{G}'_{tj}$ represents the j-th token in $\mathcal{G}'_t$. And $\mathcal{L}_{GFE}$ represents the total gloss-level feature enhancement loss.

With these designs, the final loss can be expressed as:

$$\mathcal{L} = \mathcal{L}_{CTC} + \mathcal{L}_{RSM} + \mathcal{L}_{GFE} + \gamma \cdot \mathcal{L}_{DDM} \tag{8}$$

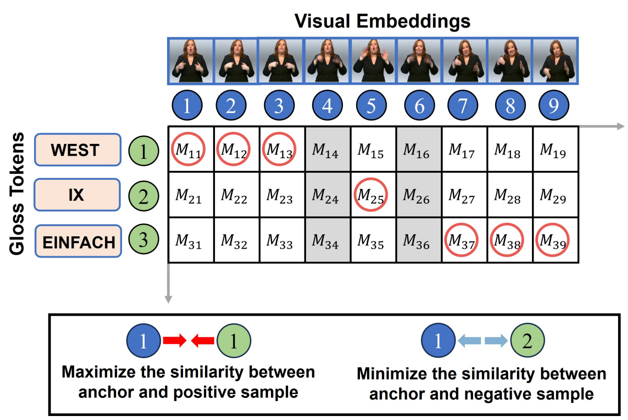

Figure 3: The attention matrix produced during gloss-level feature enhancement. The token-embedding pairs circled in red form the positive sample pairs and the rest of pairs form the negative sample pairs. In particular, the gray blocks consist of embeddings that represent the transitional parts between sign actions.

4 Experiment

4.1 Experiment Setup

The proposed method is implemented with Pytorch on two NVIDIA RTX 4090GPUs. It takes ResNet-34 (He et al., 2016) as the visual encoder and a module derived from mBART (Liu et al., 2020) as the gloss encoder. A decoder consisting of 4 transformer encoder layers is equipped to decode the final prediction. We adopt a 12-layer DiT to predict the noise. The training process consists of two stages. We first train our network without diffusion model based gloss generation module for 40 epochs to obtain meaningful visual features and gloss features. In the second stage, we freeze the parameters of visual encoder, gloss encoder and decoder to train the rest of model for 60 epochs. The Adam optimizer is adopted, and the initial learning rate is set to 10^{-4} for both stages. The learning rate decays (0.2) at epochs 20 and 35 for stage one and decays (0.2) at epochs 30 and 50 for stage two. The weight decay of 10^{-4} and batch size of 2. The diffusion time step is set to 1000, and we set β_t increasing linearly from $\beta_1 = 0.0001$ to $\beta_T = 0.99$. The hyperparameter γ is set to 10.0. The weights in Eqs.6 and Eqs.7 are set from $\delta_1 = 0.0001$ to $\delta_T = 0.01$, and linear weight schedule is adopted. The threshold $\tau = 0.15$. Word Error Rate(WER) (Park et al., 2008) is utilized as the evaluation metric. Lower WER refers to higher recognition accuracy.

4.2 Comparison with State-of-the-art Methods

The performance of our DiffSign is evaluated on three widely used SLR datasets, which are PHOENIX-2014 (Koller et al., 2015), PHOENIX-2014T (Camgoz et al., 2018) and CSL-Daily (Zhou et al., 2021a).

PHOENIX-2014. Table 1 presents a comparison between several state-of-the-art methods on PHOENIX-2014 dataset. Compared to (Hu et al., 2023), which adopts self-enhanced correlation calculation to capture hand trajectories but solely depends on a LSTM-based sequential module to complete alignment, our method can achieve 2.2% and 2.1% improvement on dev and test subset, respectively. We also find that (Chen et al., 2022) which incorporates the keypoint sequence outperforms (Hu et al., 2023) which relies on RGB stream only. This suggests that fusing more vision-based information such as optical flow, skeleton keypoints, etc. can improve recognition to a limited extent, 0.4% and 0.6% on dev and test subset. As the explicit alignment of the fixed pattern is adopted, the performance difference between some of the recent state-of-the-art methods is not significant. The difference in WER between top-tier methods like (Zhang et al., 2023), (Chen et al., 2022) and (Ahn et al., 2024) is only about 0.5%. By transforming cross-modal alignment into cross-modal generation and fully exploring the cross-modal feature correspondence ability of the diffusion denoising model, our method achieves an improvement of nearly 1% in the dev subset compared to SOTA work (Zhang et al., 2023).

Table 1: Comparison with state-of-the-art methods on PHOENIX-2014 dataset.

Methods	Dev(%)	Test(%)
SubUNets (Cihan Camgoz et al., 2017)	40.8	40.7
IAN (Pu et al., 2019)	37.1	36.7
CNN-LSTM-HMMs* (Koller et al., 2019)	26.0	26.0
SFL (Niu & Mak, 2020)	24.9	25.3
DNF(RGB) (Cui et al., 2019)	23.8	24.4
FCN (Cheng et al., 2020)	23.7	23.9
CMA (Pu et al., 2020)	21.3	21.9
VAC (Min et al., 2021)	21.2	21.9
STMC* (Zhou et al., 2021b)	21.1	20.7
C^2SLR* (Zuo & Mak, 2022)	20.5	20.4
CVT-SLR (Zheng et al., 2023)	19.8	20.1
CorrNet (Hu et al., 2023)	18.8	19.4
TwoStream-SLR* (Chen et al., 2022)	18.4	18.8
SlowFast (Ahn et al., 2024)	18.0	18.3
C^{2}ST (Zhang et al., 2023)	17.5	17.7
Ours	**16.6**	**17.1**

"*" indicates the utilization of more cues such as extra face or hand features acquired by heavy pose-estimation networks or pre-extracted heatmaps.

PHOENIX-2014T. We demonstrate the performance of several methods on both dev and test sets of PHOENIX-2014T in Table 2. Our method still outperforms rest approaches on both of these subsets. (Zheng et al., 2023) is similar to our approach in integrating the generative model into the CSLR task; however, it introduces the generative model as a denoising autoencoder, overlooking the generative model's ability to explore cross-modal feature associations. (Zhou et al., 2021b; Zuo & Mak, 2022) introduce extra facial and hand features acquired by heavy pose-estimation networks or pre-extracted heatmaps. Although adding more modal features as inputs can improve the representation ability of the fused features, it also inevitably introduces redundant information, which affects the final recognition results. Our method accomplishes CSLR through a more elegant single-cue framework and achieves more accurate recognition.

CSL-Daily is a recently released large-scale Chinese sign language dataset for both continuous sign language recognition and translation. The content of the dataset is centered around daily life. It has the largest vocabulary size (20K) among commonly used CSLR datasets. Table 3 shows that our method still achieves the best result on this challenging dataset. The excellent performance of our method on CSL-Daily dataset demonstrates the feasibility of converting CSLR to cross-modal generation and helps us outperform the SOTA work (Zhang et al., 2023), which recurrently fuses gloss representations from all previous time steps with the current time visual representation.

Table 2: Comparison with state-of-the-art methods on PHOENIX-2014T dataset.

Methods	Dev(%)	Test(%)
SFL (Niu & Mak, 2020)	25.1	26.1
DNF(RGB) (Cui et al., 2019)	23.3	25.1
FCN (Cheng et al., 2020)	23.3	25.1
SignBT (Zhou et al., 2021a)	22.7	23.9
CNN-LSTM-HMMs* (Koller et al., 2019)	22.1	24.1
C^2SLR* (Zuo & Mak, 2022)	20.2	20.4
STMC* (Zhou et al., 2021b)	19.6	21.0
CVT-SLR (Zheng et al., 2023)	19.4	20.3
CorrNet (Hu et al., 2023)	18.9	20.5
TwoStream-SLR* (Chen et al., 2022)	17.7	19.3
SlowFast (Ahn et al., 2024)	17.7	18.7
C^2ST (Zhang et al., 2023)	17.3	18.9
Ours	**16.5**	**17.8**

Table 3: Comparison with state-of-the-art methods on CSL-Daily dataset.

Methods	Dev(%)	Test(%)
SubUNets (Cihan Camgoz et al., 2017)	41.4	41.0
LS-HAN (Huang et al., 2018)	39.0	39.4
SignBT (Zhou et al., 2021a)	33.2	33.2
FCN (Cheng et al., 2020)	33.2	32.5
DNF(RGB) (Cui et al., 2019)	32.8	32.4
CorrNet (Hu et al., 2023)	30.6	30.1
TwoStream-SLR* (Chen et al., 2022)	25.4	25.3
SlowFast (Ahn et al., 2024)	25.5	24.9
C^2ST (Zhang et al., 2023)	25.9	25.8
Ours	**24.3**	**23.9**

4.3 Ablation Study

In this section, we begin by performing comprehensive experiments on each component of our framework to thoroughly assess the effectiveness of our designs. Next, we analyze the impact of hyperparameters utilized in our network.To clearly illustrate the functionality of our designs, we also compare our method with prior approaches based on traditional recognition frameworks. We use CSL-Daily as our benchmark for evaluation.

Study on each component. Table 4 presents an analysis of the effectiveness of each proposed component of our network. The first row of the table indicates that no proposed modules are applied; instead, the extracted visual features are directly fed into the decoder for recognition results. Our model reaches a Word Error Rate (WER) of 26.1% on the CSL-Daily Dev set by adopting diffusion model based gloss generation (DGG) module to accomplish CSLR through cross-modal generation. When we incorporate gloss-level feature enhancement (GFE) to mitigate the ambiguity and uncertainty of sign language videos caused by coarticulation and to ensure sharp differences between visual features representing different glosses, the WER improves to 24.9%. Furthermore, by implementing recognition-oriented supervision module (RSM) for enhanced supervision and better adaptation to downstream recognition tasks, we achieve a final WER of 24.3%.

Table 4: Study the effects of each component of the proposed network on the CSL-Daily dataset.

DGG	GFE	RSM	Dev(%)	Test(%)
✘	✘	✘	31.2	30.8
✔	✘	✘	26.1	25.4
✔	✔	✘	24.9	24.5
✔	✔	✔	**24.3**	**23.9**

Study on weight δ_t of gloss-level feature enhancement and recognition-oriented supervision module. As the target gloss features are gradually refined during the sampling process, we need to impose different levels of constraints on the gloss features generated in each iteration to achieve a balance in the loss value. We use a series of weighting factors $\delta_t \in (0,1)_{t=1}^{T}$ to accomplish the task. We compare our default configuration where δ_t linearly increases from $\delta_1 = 10^{-4}$ to $\delta_T = 0.01$ with other different settings. The comparison is shown in Table 5. In the initial few iterations, the quality of the generated gloss features is low and contains a significant amount of noise, which does not effectively represent the individual glosses in the sentence. If we set the initial weights too high, we will encounter a significantly large loss, which can lead to vibrations during training and result in non-convergence. Conversely, if we set the weights at the end stage too low, it will reduce the overall influence of the constraints, making it impossible to alleviate the uncertainty and ambiguity of sign videos or to better adapt to downstream recognition tasks.

Table 5: Study the effects of weight δ_t of gloss-level feature enhancement and recognition-oriented supervision module.

δ_1	δ_T	Dev(%)	Test(%)
10^{-4}	0.01	**24.3**	**23.9**
10^{-3}	0.01	26.7	26.2
10^{-2}	0.01	29.3	28.5
10^{-4}	0.1	28.4	28.0
10^{-4}	0.02	25.5	25.0
10^{-4}	0.01	**24.3**	**23.9**
10^{-4}	0.001	25.9	25.1

Comparison with cross-modal alignment methods. Unlike other methods focus on cross-modal alignment, we transform the CSLR task into a cross-modal generation task, fully exploring the potential of diffusion models in corresponding cross-modal features, which has resulted in more accurate gloss sequences. As shown in Table 6, we compare our cross-model generation method with other gloss-level or sentence-level cross-modal alignment methods to further validate its superiority. For fairness considerations, we will use ResNet34 as the visual encoder for all methods. Compared to previous methods based on cross-modal alignment, our method leverages cross-modal generation to generate more accurate gloss sequences.

Table 6: Comparison with other cross-model alignment methods.

Methods	Dev(%)	Test(%)
SEC (Zuo & Mak, 2022)	28.1	27.7
Visual Enhancement (VE) (Min et al., 2021)	28.0	27.5
Visual Alignment (VA) (Min et al., 2021)	27.6	27.0
VE+VA (Min et al., 2021)	26.9	26.3
IAN (Pu et al., 2019)	29.5	28.8
DNF (Cui et al., 2019)	31.2	30.6
CVT-SLR (Zheng et al., 2023)	27.3	26.9
C^2ST (Zhang et al., 2023)	26.1	25.9
Ours	**24.3**	**23.9**

4.4 VISUALIZATION

Visualization of the gloss predictions obtained through cross-modal generation. Fig.4 shows how our network present a more accurate gloss sequence when compared with several top-tier methods (Zhang et al., 2023; Chen et al., 2022; Hu et al., 2023). As we mentioned earlier, gloss-level alignment in the CSLR task is an ill-posed problem, and continuous sign language videos are affected by coarticulation, which increases the uncertainty and ambiguity during the transitions of sign actions, further complicating the alignment process. Relying solely on the idea of explicit cross-modal alignment presents performance bottlenecks. Our method transforms the CSLR into cross-modal generation, allowing powerful diffusion models to fully understand the relationships between visual embeddings and gloss tokens. This enables the models to obtain more accurate results. As shown in Fig.4, when there are consecutive identical glosses, previous methods was not able to effectively identify gaps between these same glosses, leading to only partial recognition of glosses “press” or direct misidentification. In contrast, our proposed network distinguishes transitions from other sign actions through gloss-level feature enhancement. By leveraging cross-modal generation for gloss prediction, we can accurately identify each instance of ”press” along with the gaps between them.

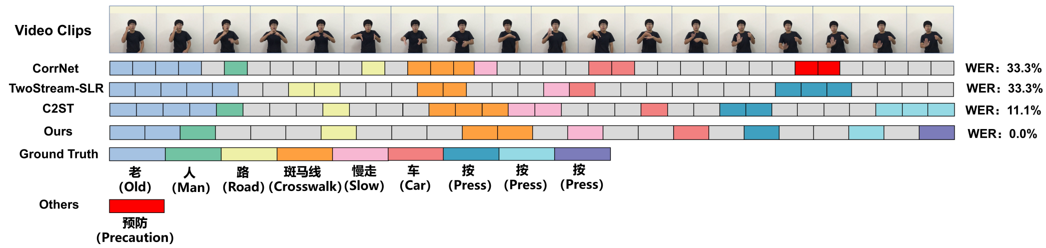

Figure 4: Beam search decode results for several top-tier methods and our proposed method on CSL-Daily dataset. The grey blocks represent the transitions that contain no semantic actions. Our proposed method accurately recognizes multiple consecutive and identical glosses as well as the transitions between them.

Visualization for gloss-level feature enhancement and recognition-oriented supervision module. Fig.5 illustrates the impact of GFE and RSM on the word error rate (WER) during the training process. Incorporating either GFE or RSM results in a reduction of the final WER, with reductions of 1.2% and 1.0%, respectively. Due to the poor quality of the generated gloss features in the early stages of training, the large gradient values resulting from the added constraints caused some oscillations in the network, leading to an increase in WER during the first 15 epochs. However, after completing 60 training epochs, the application of either constraint significantly reduced the network's WER. When we simultaneously apply both constraints, the WER is further reduced, the WER decreases by 1.8% compared to the scenario without any constraints, showing the effectiveness of the two constraints.

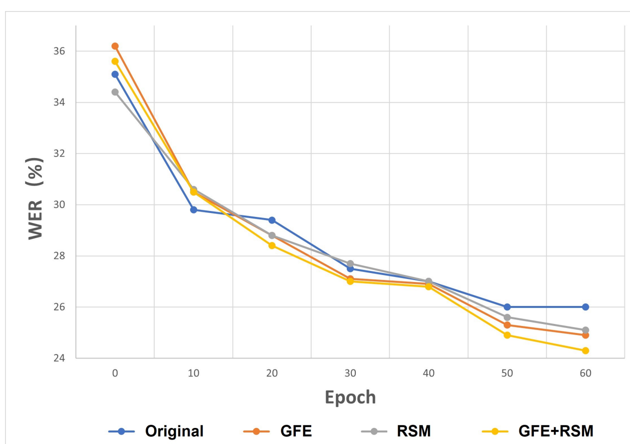

Figure 5: Visualization of the impact of different constraints on WER in the CSL-Daily dataset.

5 Conclusion

This paper propose a diffusion model based network equipped with recognition-oriented supervision module to complete CSLR through cross-modal generation and fully explore the potential of diffusion models in cross-modal feature correspondence. A contrastive learning based gloss-level feature representation enhancement strategy is proposed to optimize visual features and mitigate the ambiguity and uncertainty inherent in sign language videos. Our network achieves state-of-the-art results on several datasets, including Phoenix-2014, Phoenix-2014T, and CSL-Daily.

REFERENCES

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)*, pp. 3920–3924. IEEE, 2024.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7784–7793, 2018.

Hao Chen, Jiaze Wang, Ziyu Guo, Jinpeng Li, Donghao Zhou, Bian Wu, Chenyong Guan, Guangyong Chen, and Pheng-Ann Heng. Signvtcl: Multi-modal continuous sign language recognition enhanced by visual-textual contrastive learning. *arXiv preprint arXiv:2401.11847*, 2024a.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024b.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056, 2022.

Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 697–714. Springer, 2020.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19016–19026, 2023.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 3056–3065, 2017.

Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891, 2019.

Leming Guo, Wanli Xue, Ze Kang, Yuxi Zhou, Tiantian Yuan, Zan Gao, and Shengyong Chen. Denoising-diffusion alignment for continuous sign language recognition. *arXiv preprint arXiv:2305.03614*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2529–2539, 2023.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320, 2019.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23359–23368, 2023.

Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11542–11551, 2021.

Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pp. 172–186. Springer, 2020.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. An empirical analysis of word error rate and keyword error rate. In *Interspeech*, volume 2008, pp. 2070–2073, 2008.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4165–4174, 2019.

Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting continuous sign language recognition via cross modality augmentation. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1497–1505, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pp. 32483–32498. PMLR, 2023.

Huaiwen Zhang, Zihang Guo, Yang Yang, Xin Liu, and De Hu. C2st: Cross-modal contextualized sequence transduction for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21053–21062, 2023.

Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z Li. 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*, pp. 23141–23150, 2023.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316–1325, 2021a.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779, 2021b.

Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5131–5140, 2022.