Denoising-Diffusion Alignment for Continuous Sign Language Recognition

Leming Guo[1], Wanli Xue[1*], Yuxi Zhou[1], Ze Kang[1], Tiantian Yuan[2], Zan Gao[1], and Shengyong Chen[1]

[1] School of Computer Science and Engineering, Tianjin University of Technology
China
[2] Technical College for the Deaf, Tianjin University of Technology
China
{glm,kz}@stud.tjut.edu.cn,xuewanli@email.tjut.edu.cn
joy_yuxi@pku.edu.cn,yuantt@tjut.edu.cn,zangaonsh4522@gmail.com,sy@ieee.org

ABSTRACT

Continuous sign language recognition (CSLR) aims to promote active and accessible communication for the hearing impaired, by recognizing signs in untrimmed sign language videos to textual glosses sequentially. The key challenge of CSLR is how to achieve the cross-modality alignment between videos and gloss sequences. However, the current cross-modality paradigms of CSLR overlook using the glosses context to guide the video clips for global temporal context alignment, which further affects the visual to gloss mapping and is detrimental to recognition performance. To tackle this problem, we propose a novel **D**enoising-**D**iffusion global **A**lignment (DDA), which consists of a denoising-diffusion autoencoder and DDA loss function. DDA leverages diffusion-based global alignment techniques to align video with gloss sequence, facilitating global temporal context alignment. Specifically, DDA first proposes the auxiliary condition diffusion to conduct the gloss-part noised bimodal representations for video and gloss sequence in a common low-dimensional latent space. To address the problem of the recognition-oriented alignment knowledge represented in the diffusion denoising process cannot be feedback. The DDA further proposes the Denoising-Diffusion Autoencoder, which adds a decoder in the auxiliary condition diffusion to denoise the partial noisy bimodal representations via the designed DDA loss in self-supervised. Notably, our DDA loss not only performs the denoising but also achieves the alignment knowledge transfer. In the denoising process, each video clip representation of video can be reliably guided to re-establish the global temporal context between them via denoising the gloss sequence representation. Experiments on three public benchmarks demonstrate that our DDA achieves state-of-the-art performances and confirm the feasibility of DDA for video representation enhancement. Furthermore, DDA also can be a plug-and-play optimization to generalize other CSLR methods.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**.

KEYWORDS

Continuous Sign Language Recognition, Global Temporal Context Alignment, Diffusion denoising

1 INTRODUCTION

Sign language is a convenient communication way between hearing-impaired people. One simple way to promote the active integration of hearing-impaired people into society is for hearing people to be capable of reading and understanding sign language. However, it is difficult for hearing people to learn sign language. Continuous Sign Language Recognition (CSLR) involves recognizing signs in an untrimmed video to textual glosses[1] sequentially, which facilitates barrier-free communication for hearing-impaired people, and it is an important manifestation of social good.

CSLR, as a typical cross-modality recognition task, requires an effective cross-modal alignment paradigm. However, current alignment paradigms of CSLR ignore employing the gloss context to guide the visual representations for global temporal context alignment. These paradigms mainly fall into the following categories: (1) Employing Connectionist Temporal Classification (CTC) to map visual representations to gloss space [13, 18, 25, 28] (as shown in Figure 1(a)). (2) Based on CTC, employ the language model to learn gloss context and map both visual and gloss representations to a common high-dimensional space to close their distributions [30, 45] (illustrated in Figure 1(b)). (3) Mapping visual representations at each time step with gloss representations (extracted from pre-trained language model) of all previous time steps into multiple high-dimensional hybrid spaces, which conducts local temporal context alignment [31, 44] (depicted in Figure 1(c)).

---

[1]Gloss: each word from the text sentence of the annotated sign language video.
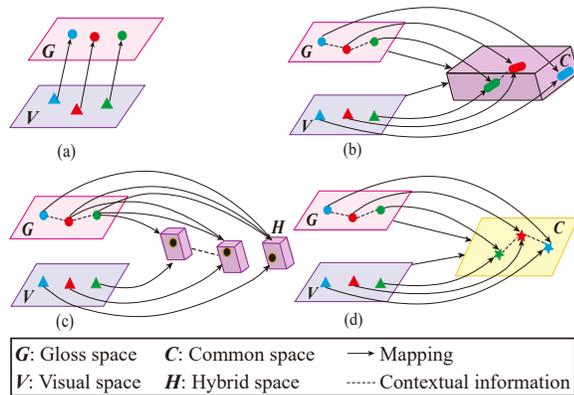
author name and author name, et al.

Figure 1: Cross-modality alignment paradigms investigated in CSLR. (a) Video clip→individual gloss mapping. (b) High-dimensional common latent space learning contains two modalities sequence. (c) Multi-hybrid spaces learning for each video clip and previous time steps' glosses. (d) Low-dimensional common space learning contains two modalities sequence with global temporal context, a novel paradigm first proposed by our DDA.

Although these paradigms achieve decent performance, the alignment of Figure 1(a), (b), and (c) are unable to build the global temporal context alignment among each video clip guided by the gloss sequence, thus further affecting the visual-gloss mapping and the model's recognition performance. *As such, properly discovering the global temporal context alignment of CSLR is of great importance.*

The current methods that enable advanced cross-modal alignment are mainly based on contrastive learning [24] or diffusion model [2, 3, 41], but the former is data-hungry [32]. Therefore, we propose the auxiliary condition diffusion to perform the global temporal context alignment in a diffusion manner. Additionally, [7] has concluded that rather than the diffusion process, the denoising process is the key to driving the recognition-oriented representation capability of the diffusion model. However, the denoising process works at the inference, which cannot feedback the powerful alignment knowledge in representations to the backbone.

To solve these challenges, we propose a novel **D**enoising-**D**iffusion **A**lignment (DDA), containing the Denoising-Diffusion Autoencoder and DDA loss function, to explore the diffusion-based global alignment between video and gloss sequence to facilitate the global temporal context alignment. Firstly, the proposed auxiliary condition diffusion merges the representations of both the gloss sequence and the entire video as bimodal representations and projects them into a common low-dimensional latent space, as shown in Figure 1(d). Moreover, the gloss sequence part in the bimodal representations will be added to random Gaussian noise, which we name the partial noising process. Secondly, inspired by the idea of [7], DDA conducts the Denoising-Diffusion Autoencoder by adding a decoder in the auxiliary condition diffusion to denoise the partial noisy bimodal representations via the proposed DDA loss. In this denoising process, the proposed Autoencoder can naturally model the global temporal context of video to denoise the global

gloss context in the common space by the DDA loss, which implicitly makes the video representation learn and further present the global gloss context. Furthermore, the DDA loss also transfers the knowledge of the globally aligned representation to the video representation. This alignment approach reliably guides each video clip representation of video to re-establish the global temporal context between them based on the global text context. In particular, the Denoising-Diffusion Autoencoder can be optimized by the DDA loss in a self-supervised manner, which effectively back-propagates the learned alignment knowledge to refine the video representations. In summary, our contributions are three-fold:

- Orthogonal to the cross-modality alignment schemes commonly used in CSLR, we propose a novel Denoising-Diffusion Alignment (DDA) to perform the global temporal context alignment, which is proposed for the first time in CSLR.
- A Denoising-Diffusion Autoencoder, which is a self-supervised manner, and a loss function is proposed to denoise the partial noisy bimodal representations and back-propagate the alignment knowledge to enhance the video representations.
- Experiments show that our method achieves state-of-the-art recognition performance, enhances the generalization of video representations, and has plug-and-play scalability.

2 RELATED WORK
2.1 Continuous Sign Language Recognition
The CSLR aims to recognize signs in a video corresponding to several glosses, where the order of glosses is consistent with the signs. Due to weak sentence-level annotation (lacking segmentation ground-truth for each sign) and small-scale data being available in current CSLR benchmarks, many state-of-the-art methods [13, 25, 28, 43] exploit the connectionist temporal classication (CTC) [10] to accurately map each video clip to the corresponding gloss by maximizing the probabilities of all alignment paths between video clips and glosses. To enhance the above cross-modality mapping, some methods exploit pre-captured pose heatmaps [48], body keypoints [9], or model movements trajectories [17, 18] to dynamically emphasize sign movements. However, these methods ignore the problem of the CTC conditional independence assumption, which only achieves video clip→individual gloss mapping, and lacks gloss context learning. To solve this problem, [8, 30, 45] further exploit the language model or VAE model or provide extensive gloss context supervision, subsequently maps the visual representation to the high-dimensional gloss space, and employ constraints to close their distributions. Meanwhile, C^{2}ST [44] recurrently fuses gloss representations from all previous time steps with the current time visual representation, which conducts multiple hybrid spaces to inject the context of two modalities. However, these above methods only focus on guiding the video clip representation to learn gloss context, ignoring guiding each video clip to learn the global temporal context alignment among them. In this work, our method aims to perform global temporal context alignment via denoising partial noised bimodal representations.

2.2 Corss-modality alignment in CSLR

In CSLR, due to the weak annotation of benchmarks, almost all methods adopt the CTC loss function to map the visual representations to textual gloss space for cross-modality alignment [12, 13, 18, 25, 28, 30, 31]. Based on the CTC, some methods also employ the cross-attention operation [12, 31], dynamic time warping (DTW) [30, 31], and contrastive learning method [30] to provide gloss context supervision. Current effective cross-modal alignment is at the core of many cross-modal tasks [1, 20], and the main methods that enable advanced cross-modal alignment are contrastive learning methods [24] as well as diffusion models [2, 3, 19, 41]. [3, 41] come to the same conclusion that the latent space features of diffusion models are indeed able to present the image, and DiffDis [19] also employs the powerful diffusion framework to perform superior image-text alignment. Due to the contrastive learning is data-hungry [32], which is not suitable for the data-limited CSLR. Therefore, we adopt the diffusion model to achieve the global cross-modal alignment.

2.3 Denoising Diffusion Model

The denoising diffusion model (DDM) incorporates a forward Gaussian diffusion noising process and a reverse denoising generation process, which can refine the generated objects starting from Gaussian noise iteratively. The denoising diffusion model is impressive and powerful for cross-modalities generation such as text-to-visual generation: Stable Diffusion [34], LGD [38], UniDiffuser [2] or Auto sequence-to-Video sequence mutual generation: MM-Diffusion [35]. [3, 41]has presented that the latent space of the DDM can present the cross-modalities alignment. However, [7] demonstrates that these DDMs are generation-oriented, and their representations are not robust enough for recognition. [7] transforms the DDM to a denoising autoencoder to perform self-supervised learning for recognition-oriented representations, and concludes that the representation capability of DDM is mainly gained by the denoising process, rather than a diffusion process. In this work, we also formulate the diffusion model as a denoising-diffusion autoencoder to achieve global temporal context alignment.

3 METHOD

3.1 General CSLR Framework

Formally, given a T frames sign language video $X = \{x_i\}_{i=1}^T$, and its corresponding gloss sequence with L glosses $G = \{g_i\}_{i=1}^L$, where g_i denotes the i-th gloss. The dominant continuous sign language recognition (CSLR) framework [12, 17, 44, 45] embraces the paradigm that comprises a video encoder, a classification module, and a cross-modality alignment function.

Video encoder. The video encoder Φ_{VEnc} contains a spatial perception module and a temporal perception module. Specifically, the spatial perception module first extracts spatial features $F_{sp} = \{f_{sp}^i\}_{i=1}^T$ from X. Subsequently, the temporal perception module learns sign-specific knowledge and contextual correlation to extract video representations $F_V = \{f_V^i\}_{i=1}^{T'} \in \mathbb{R}^{(T') \times d}$, which is the video encoder's output. Moreover, F_V will be fed into the classifier Φ_Z to predict corresponding logits $Z_V = \{z_V^i\}_{i=1}^{T'}$. Finally, the cross-modality alignment function learns the mapping $p(g_i|clip; \theta)$.

$clip = \{x_i\}_{i=1}^{T'}, T' < T$ is the video clip in a video, and θ indicates video encoder's parameters.

Gloss Sequence encoder. Additionally, given the sign language gloss sequence with L glosses $G = \{g^i\}_{i=1}^L$, the sign gloss sequence representation $F_G \in \mathbb{R}^{L \times d}$ is extracted by a mBART model pre-trained by the sign language data [9, 45].

Connectionist Temporal Classification. Due to the weak sentence-level annotation, the Connectionist Temporal Classification [10] $\mathcal{H}$ is employed in recent SOTA methods [9, 12, 17, 44, 45], which can mapping unsegmented video clips $\{x_i\}_{i=1}^{T'}$ and gloss sequence G by summing the probabilities of all feasible alignment paths π:

$$\mathcal{H} = -\log p(G|x_i; \theta) = -\log \left(\sum_\pi p(\pi|x_i; \theta) \right), \quad (1)$$

where $p(\pi|x_i; \theta)$ is calculated by CTC: $p(\pi|x_i; \theta) = \prod_i p(\pi_i|x_i; \theta)$, the probabilities $P_\theta = softmax(Z_V)$ can be calculated via a *softmax* function to the video encoder's logits Z_V. The CTC has been validated to achieve superior video clip→individual gloss alignment.

Baseline method. It is worth noticing that the video encoder equipped the classifier and the CTC alignment function $\mathcal{H}$ standing for the baseline in this work, which is the same as the other CSLR methods [12, 17, 44].

3.2 Revisiting the CSLR.

Formally, given a T frames sign language video $X = \{x_i\}_{i=1}^T$, and its corresponding gloss sequence with L glosses $G = \{g_i\}_{i=1}^L$, we introduce the video representation F_V (see Sec. 3.1) as latent variables z to help model the conditional probability of CSLR:

$$p(G|X) = \int_z p(G, z|X) dz = \int_z p(G|z, X) p(z|X) dz \quad (2)$$

From Equation 2, latent variables z play a crucial role in aligning video X and gloss sequence G. Notice that, z can be drawn from the posterior distribution $p_\theta(z|X, G)$, which can present bi-modal contextual information of both video and gloss sequence. And the desired posterior distribution $p_\theta(z|X, G)$ can be modeled by the global video-gloss sequence alignment learning. Therefore, through the global alignment, the prior $p_\theta(z|X)$ can effectively achieve global temporal context alignment by capturing bi-modal contextual information from $q_\phi(z|X, G)$, and finally enhances the CSLR model's recognition performance.

3.3 Denoising-Diffusion Alignment

Based on the observations of Sec. 3.2, in this section, we propose a novel Denoising-Diffusion Alignment (DDA) to formulate the global video-gloss sequence alignment learning to facilitate the global temporal context alignment. The DDA consists of the denoising-diffusion autoencoder and the DDA loss function. The workflow of the DDA is illustrated in Figure 2.

Auxiliary condition diffusion. The current CSLR methods achieve video clip→gloss mapping in the high-dimensional space (illustrated in Figure 1(b) and (c)). However, this mapping manner usually faces a large bias towards the ground-truth gloss sequence [6]. To solve this problem and ensure the effective global context guidance of textual gloss modality to video, we propose the auxiliary
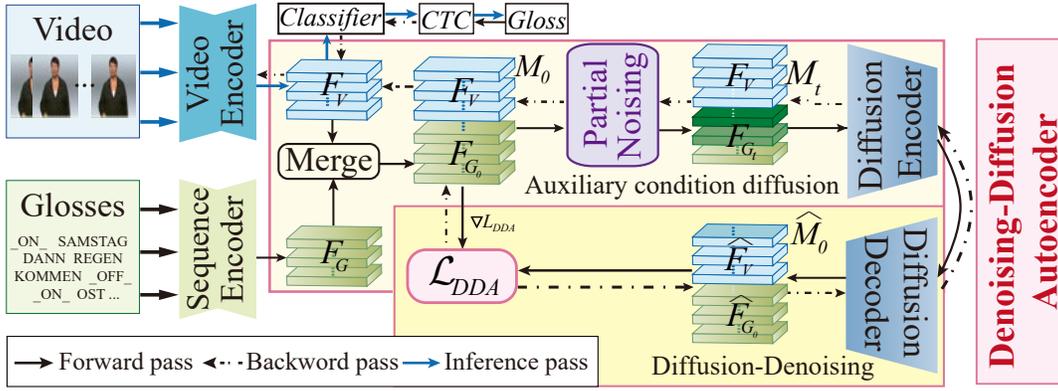
Figure 2: Illustration of the workflow of the proposed Denoising-Diffusion Alignment (DDA). Given a pair of visual representations and gloss sequence representations, the two are merged in sequence dimension to get the bimodal representations M_0, and only the gloss sequence part of it will be gradually perturbed with Gaussian noise in the partial noising process, obtaining M_t. Subsequently, the diffusion encoder will embed M_t to learn both the video global context and the noised gloss sequence context and output the latent representations F_H. Consequently, the diffusion decoder will encode F_H to the denoising output $\hat{M}_0$. Finally, the $\mathcal{L}_{DDA}$ loss function will constrain the predicted output $\hat{M}_0$ to approximate the clean bimodal representation M_0 and guide F_V to learn the knowledge of global temporal context alignment.

condition diffusion. As shown in Figure 2, given the pair of visual representation F_V and gloss sequence representation F_G, they are mapped into a common latent space (as shown in Figure 1d), and F_G used as prompt, is concatenated to F_V to generate a bimodal representation $M \in \mathbb{R}^{(L+\hat{L}) \times d}$. Where $\hat{L}$ and L denote the sequence length of visual representation and gloss sequence representation, respectively. In this common space, the consistency between two different modalities [48] can be enhanced and easily transfer global contextual knowledge and fine-grained knowledge of modalities [6]. Then the bimodal representation M will be projected into a low-dimensional latent space. Due to the low dimensional, the two modalities' semantics can be abstracted in this space to reduce the bias towards the ground-truth sequence. Therefore, M can be regarded as incorporating the gloss sequence's global context into the approximate posterior $q_\phi(z|X, G)$.

$$F_G = Projg(F_G), F_V = Projv(F_V),$$
$$M = [f_G^1, ..., f_G^T, f_V^1, ..., f_V^L] \quad (3)$$
$$M = Projm(M)$$

Furthermore, we adopt the DDM [16, 37] forward process, which incrementally adds multi-level Gaussian noise to the gloss part in bimodal representation M with a Markov chain manner $M_0, ..., M_T$, $M_0 = M$ (the partial noising process). The single step (from M_{t-1} to M_t) is presented as follows:

$$q(M_t|M_{t-1}) = \mathcal{N}\left(M_t; \sqrt{1 - \beta_t} M_{t-1}, \beta_t I\right), \quad (4)$$

Consequently, given M_0, M_t can be empirically redefined as:

$$M_t = \sqrt{\overline{\alpha}_t} M_0 + \sqrt{1 - \overline{\alpha}_t} \epsilon, \quad (5)$$

where ϵ stands for Gaussian noises, we empirically set all diffusion steps including T, $\beta_t \in (0, 1)_{t=1}^t$, $\alpha_t := 1 - \beta_t$, $\overline{\alpha}_t := \prod_s \alpha_s$, and the linear noise schedule [7, 16] is adopted. Since only gloss part in M is imposed noising in T timesteps, such as $M_t = [f_V^1, ..., f_V^{\hat{L}}, f_{G_t}^1, ..., f_{G_t}^L]$,

f_{G_t} stands for the noisy visual representation, $F_{G_0} = F_G$, $\hat{L}$ denotes the video sequence length. As a result, M_t denotes the noisy gloss sequence representation with Gaussian distribution connected with the video representation.

Denoising-diffusion autoencoder. Our preliminary objective is to formulate the video-gloss sequence global alignment as a DDM paradigm to effectively transfer the global gloss context to the global temporal context. However, the powerful representation of DDM is gained by the denoising process [7], this process works at the inference, which cannot provide alignment knowledge back to the visual representation. Therefore, we conduct a denoising-diffusion autoencoder procedure [7] to achieve a self-supervised learner to perform the global alignment.

To efficiently preserve the semantic context of gloss sequence, and spontaneously learn the global semantic correlation between the video and gloss modalities, we adopt a Transformer architecture, an mBART language model pre-trained on sign language [8], as diffusion encoder Φ_{DEnc} and diffusion decoder Φ_{DDec}. Inspired by [7], both the encoder and decoder have the same architecture, and the decoder is deeper than the encoder. The diffusion encoder $\Phi_{DEnc}(M_t, t)$ is adopted to embed the part-noisy representation to learn the context of both the noised gloss sequence representation and the visual representation and output the latent features F_H. Φ_{DEnc} is optimized to retain the most relevant information between the initial bimodal representation M_0 and the denoising result. Consequently, the diffusion decoder $\Phi_{DDec}(M_t, t)$ will encode F_H to the denoising output $\hat{M}_0 = \Phi_{DDec}(M_t, t)$ by optimized to make the tractable variational lower-bound $\mathcal{L}_{VLB} \leq \mathbb{E}[-log p_\theta(M_0)]$ to effectively perform denoising M_t to approximate the clean initial bimodal representation M_0. In this work, we follow [7, 16] to simplify the $\mathcal{L}_{VLB} \leq \mathbb{E}[-log p_\theta(M_0)]$ to the mean squared error objective to train the denoising-diffusion autoencoder.

The simplified objective is defined as follows:

$$\mathcal{L}_{DDA} = \mathbb{E}_{F_{G_0} \sim q(F_{G_0}), \hat{F}_{G_0}, t \sim [1,T]} \left\| F_{G_0} - \hat{F}_{G_0} \right\|^2 + KL(F_V, \hat{F}_V)$$
$$KL(F_V, \hat{F}_V) = KL\left(F_V || \hat{F}_V\right) \tag{6}$$

Different from DDPM for noising ϵ prediction, our DDA predicts the clean initial bimodal representation M_0. Where only the gloss sequence part F_G in M will compute the loss. In addition, $KL(F_V, \hat{F}_V)$ denotes the KL divergence acting on F_V and $\hat{F}_V$ to further explicitly transfer the learned global temporal context and gloss context from $\hat{F}_V$ to the video representation F_V. Therefore, the diffusion-denoising autoencoder ensures an effective capturing of the global context within and between two modalities and can back-propagate the learned global temporal alignment to refine the video encoder.

3.4 Objective

The objective of our method contains the objective function $\mathcal{L}_{DDA}$ and $\mathcal{L}_{CTC}$ to achieve both the video clip-textual gloss local alignment, the video-gloss sequence global alignment, and the global temporal context alignment. Therefore, our objective $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \mathcal{L}_{CTC} + \gamma_1 \mathcal{L}_{DDA}, \tag{7}$$

where $\mathcal{L}_{CTC}$ denotes the connectionist temporal classification (CTC) loss function, γ_1 is hyperparameter for balance the contribution. According to the experimental results in Table 7, it is set to 10.

4 EXPERIMENTS

4.1 Datasets and Evaluation

PHOENIX-2014 [23]. This benchmark was recorded from public sign language interpreters of weather forecasts, and it delivers 6,842 sentences interpreted by 9 signers, composed of 1,295 sign language gloss vocabulary. Additionally, the PHOENIX-2014 dataset is officially divided into the train set, dev set, and test set with 5,672, 540, and 629 videos.

PHOENIX-2014T[4]. This benchmark is widely utilized for both continuous sign language recognition (CSLR) task and sign language translation (SLT) task. It contains 1,085 sign language gloss vocabulary for the CSLR task, and all videos are officially divided into 7,096, 519, and 642 videos for the training set, dev set, and test set, respectively.

CSL-Daily[47]. This benchmark is a large Chinese CSLR dataset for both continuous sign language recognition (CSLR) task and sign language translation (SLT) task. Specifically, the CSLR task records 2000 sign language vocabulary and it is officially divided into 18,401, 1,077, and 1,176 videos for the training, dev, and test sets.

Evaluation metric. In this work, the word error rate (WER) metric is adopted for the CSLR evaluation. The WER belongs to the edit distance, which measures the minimum number of substitutions (#sub), deletions (#del), and insertions (#ins) operation needed to convert the predicted gloss sequence to the associated reference gloss sequence. The WER calculation method is as follows:

$$WER = \frac{\#sub + \#del + \#ins}{L}, \tag{8}$$

where $\#sub, \#del, \#ins$ are the number of substitutions, deletions, and insertions operation, respectively. Therefore, lower WER values imply better recognition performance.

4.2 Implementation Details

Network architecture. Specifically, in the video encoder, the Resnet50 [15] pre-trained by CLIP [33] is leveraged to be the spatial perception module, and the temporal perception module is empirically set to TLP [25] equipped with a two-layer Bi-LSTM module. The feature dimensions of the TLP and Bi-LSTM are set to 1024, and for diffusion encoder, and diffusion decoder are set to 512.

Parameter setting. For auxiliary condition diffusion in Sec. 3.3, following DDIM [37] and [7], the diffusion noising timestep T is set to 1000, and we set β_t from $\beta_1 = 0.0001$ to $\beta_T = 0.99$, which provides noise to the gloss sequence representation linearly. The covariance Σ_θ is fixed and defined as $\Sigma_\theta = \beta_t \mathbb{I}$. The hyperparameter γ_1 in Eq. 7 are set to 10.0, and we also conduct ablation studies to evaluate the impact of different γ_1.

Training and inference process. We train the DDA with a batch size of 4, using the Adam optimizer [22] with an initial learning rate of $1e - 5$ for Resnet50, and $1e - 4$ for others, a weight decay factor of $1e - 4$, and momentum as 0.9 and 0.99 for 80 epochs. And the learning rate decays (0.2) at 31 and 61 epochs. All experiments are implemented in PyTorch and on one A100 GPU. Specifically, the denoising-diffusion autoencoder is dropped in inference. The inference process of DDA begins by feeding the test video $\mathcal{X}$ into the video encoder, where the spatial perception model (ResNet50) first extracts spatial features. Following this, the temporal perception model (TLP+BLSTM) models local-global temporal context, generating video representations. These representations are then input into the classifier to obtain classification probability scores $logsoftmax(Z_V)$. Finally, CTC beam search is employed to decode the output and generate the predicted gloss sequence $\mathcal{Y}_p = \{y_p^i\}_{i=1}^L$. The beam width is set to 10.

4.3 Comparison with state-of-the-art methods

We present the experiments comparison with several state-of- the-art approaches on three public benchmarks. As shown in Table 1, Table 2, and Table 3, we can observe that our proposed DDA only models the RGB cue of sign language video and accompanying gloss sequence, which achieves state-of-the-art recognition performances on three public benchmarks, demonstrating its effectiveness. Remarkably, our DDA outperforms the keypoints supervised TwoStream-SLR [9] by 1.5% and 1.5% WERs on the dev and test set of PHOENIX-2014 and even surpasses it by 0.7% and 0.8% WERs on the dev and test set of PHOENIX-2014T. Furthermore, our DDA also gains a significant improvement compared to other multi-cue methods (marked as ∗), which employ the pre-captured hands, face, keypoints, or heartmaps as supervision. It shows the effectiveness of text guidance by denoising-diffusion autoencoder. In particular, C^2ST enforces injecting the gloss semantic context into the label prediction to refine the video clip-gloss mapping, which also demonstrates the effectiveness of semantic context learning. Our DDA also exceeds the C^2ST by 0.6% and 0.5% WERs on the dev and test set of PHOENIX-2014, even surpasses it by 0.3% and 0.4%

Table 1: Compatibility to other stat-of-the-art CSLR methods on the PHOENIX-2014. The entries denoted by "∗" used extra cues (keypoints). "Group1-4" corresponds to the cross-modality alignment paradigm shown in Figure 1(a), (b), (c) and (d), respectively.

Groups	Methods	Dev (%) ↓		Test (%) ↓	
		del/ins	WER	del/ins	WER
Group1	VAC[28]	7.9/2.5	21.2	8.4/2.6	22.3
	SEN [18]	5.8/2.6	19.5	7.3/4.0	21.0
	TLP[25]	6.3/2.8	19.7	6.1/2.9	20.8
	C^2SLR^*[48]	-	20.5	-	20.4
	SGN [42]	5.1/3.1	19.5	5.3/2.8	20.2
	RadialCTC [29]	6.5/2.7	19.4	6.1/2.6	20.2
	CoSign [21]	-	19.7	-	20.1
	CorrNet[17]	5.6/2.8	18.8	5.7/2.3	19.4
	TwoStream-SLR*[9]	-	18.4	-	18.8
Group2	CMA[30]	7.3/2.7	21.3	7.3/2.4	21.9
	CVT-SLR [45]	6.4/2.6	19.8	6.1/2.3	20.1
	CTCA[12]	6.2/2.9	19.5	6.1/2.6	20.1
Group3	Align-iOpt[31]	12.9 / 2.6	37.1	13.0 / 2.5	36.7
	C^2ST[44]	4.2/3.0	17.5	4.2/3.0	17.7
Group4	DDA(Ours)	4.0/2.5	**16.9**	4.2/2.8	**17.3**

WERs on the dev and test set of PHOENIX-2014T, and by 0.3% and 0.5% on both the dev and test sets on the CSL-Daily.

In particular, Table 1 also illustrates the performance of different cross-modality alignment paradigms (shown in Figure 1(a), (b), (c) and (d), respectively) based methods. We can observe that CTC performs significant recognition, furthermore, when employing other cross-modality alignment paradigms, the performance can be further improved, it is shown that improved cross-modality alignment technology can effectively embrace recognition gain. Moreover, our DDA, which conducts the global temporal context alignment achieves the best performance, demonstrating its effectiveness.

4.4 Ablation Studies

Ablation on the proposed DDA. Table 4 ablates the ablation studies of the denoising-diffusion alignment (DDA) on the PHOENIX-2014 benchmark. The baseline method (See 3.1) obtains WERs of 18.8% and 19.2% on both dev and test sets, and remarkably, the proposed DDA gains significant improvement, achieving 16.9% and 17.3% on both dev and test sets. The introduction of DDA promoting the baseline model demonstrates the superiority of the global temporal context alignment via diffusion autoencoder generation. Besides, the partial noising process and the denoising-diffusion autoencoder are contained in DDA, and the autoencoder consists of the diffusion encoder and decoder. When DDA removes the diffusion decoder ("w/o Auto-Dec"), only remains the diffusion encoder, its training objective is to predict the added noise in M_t at each timestep, according to the encoder output. We can observe that the recognition performance drops 0.9% and 0.7% (WERs raising) on both dev and test sets. Furthermore, DDA removes the

Table 2: Comparison (%) with baseline methods on the PHOENIX-2014T. The entries denoted by "∗" used extra cues (keypoints).

Methods	WER	
	Dev%	Test%
V-L Mapper [8]	21.9	22.5
TLP [25]	19.4	21.2
SEN [18]	19.3	20.7
CorrNet [17]	18.9	20.5
C^2SLR^* [48]	20.2	20.4
CVT-SLR [46]	19.4	20.3
CTCA (2023) [12]	19.3	20.3
CoSign [21]	19.5	20.1
TwoStream-SLR* [9]	17.7	19.3
C^2ST[44]	17.3	18.9
DDA (Ours)	**17.0**	**18.5**

Table 3: Comparison (%) with baseline methods on the CSL-Daily. The entries denoted by "∗" used extra cues (keypoints).

Methods	Dev%		Test%	
	del/ins	WER	del/ins	WER
BN-TIN+Transf [47]	13.9/3.4	33.6	13.5/3.0	33.1
SLT [5]	10.3/4.4	33.1	9.6/4.1	32.0
SEN [18]	-	31.1	-	30.7
CorrNet [17]	-	30.6	-	30.1
CTCA [12]	9.2/2.5	31.3	8.1/2.3	29.4
CoSign [21]	-	28.1	-	27.2
TwoStream-SLR* [9]	-	**25.4**	-	**25.3**
C^2ST[44]	9.3/2.7	25.9	9.0/2.7	25.8
DDA (Ours)	9.1/2.8	25.6	9.0/2.1	**25.3**

diffusion encoder ("w/o Auto-Enc"), the M_t will be fed into the remaining diffusion decoder, subsequently, the decoder will be optimized by $\mathcal{L}DDA$ to denoise M_t to approximate M_0, causing trivial performance degradation (0.6% and 0.5% on both dev and test sets). The two experiments demonstrate the superiority of the diffusion decoder and also validate that the denoising-driven process can enforce significant representation learning, as well as proved in [7]. Additionally, "FN" denotes we design a full noising process, which adds noise to both video and gloss sequence parts in the bimodal representation M. As shown in Table 4, when DDA replaces "PN" (indicates the partial noising process) with "FN", the performance decreases 0.7% and 0.8% WERs on both dev and test sets, which demonstrates the effectiveness of the partial noising process to conduct nature modalities context leaning and condition to promote the video-gloss sequence alignment. Besides, Table 4 also depicts that the optimization of the DDA relies more on the Diffusion Decoder than the Diffusion Encoder.

Table 4: Ablation study on the DDA on the PHOENIX-2014. "Auto-Enc" and "Auto-Dec" denote the Encoder and Decoder of the denoising-diffusion autoencoder, respectively. "FN" stands for the full noising process.

Variants	Dev (%) ↓		Test (%) ↓	
	del/ins	WER	del/ins	WER
DDA (Ours)	4.0/2.5	16.9	4.2/2.8	17.3
Baseline	5.6/2.8	18.8 (1.9%↑)	5.4/2.7	19.2 (1.9%↑)
DDA w/o Auto-Dec	4.6/2.9	17.8 (0.9%↑)	4.8/2.8	18.0 (0.7%↑)
DDA w/o Auto-Enc	4.4/2.6	17.5 (0.6%↑)	4.6/2.7	17.8 (0.5%↑)
DDA w/ FN	4.4/2.9	17.6 (0.7%↑)	4.9/2.8	18.1 (0.8%↑)

Table 5: Performance comparison of distinct distribution alignment methods on the PHOENIX-2014.

Methods	Alignment	Dev (%) ↓	Test (%) ↓
	-	18.8	19.2
	MMD [11]	19.0	19.2
	JMMD [26]	18.8	18.8
Baseline	NCE [33]	17.9	18.1
	SimMIM [40]	17.7	17.8
	MAE [14]	17.3	17.6
	DDA (Ours)	16.9	17.3

Ablation on distinct distribution alignment objects. We investigate the superiority of the proposed DDA by comparing other distribution alignment methods, which can achieve global video-global sequence alignment. As shown in Table 5, considering the semantic context of text modality, DDA achieves a more significant performance than MMD [11], JMMD [26], NCE loss [33], SimMIM [40], and MAE [14]. In particular, the NCE loss achieves distributions closer to the common sentence, and distributions apart from the different, however, limited by the amount of training data as well as computational resources, the performance of NCE cannot be fully utilized. Furthermore, although both SimMIM and MAE are generative methods similar to DDM, they cannot learn multi-level mask ratios simultaneously, compared to the multi-level Gaussian noise addition of DDM. These experiments demonstrate that formulating the global temporal context alignment in a denoising-diffusion autoencoder is feasible and achieves promising alignment capacity.

Table 6: Performance comparison of distinct networks for the autoencoder on the PHOENIX-2014.

Methods	Networks	Dev (%) ↓	Test (%) ↓
	pre-trained mBART	16.9	17.3
Autoencoder	mBART	17.3	17.7
	BLSTM	17.7	18.1
	1D U-Net	17.5	17.8

Table 7: Ablation study on γ_1 factor on the PHOENIX-2014.

γ_1	1.0	5.0	7.0	**10**	20	30
Dev (%) ↓	17.5	17.4	17.1	16.9	17.2	17.8
Test (%) ↓	18.0	17.8	17.5	17.3	17.8	18.2

Ablation on distinct networks for the denoising-diffusion autoencoder. We investigate the influence of different networks of diffusion encoder and decoder. "pre-trained mBART" denotes the mBART language model pre-trained on sign language [8]. "mBART" indicates the mBART language model. As shown in Table 6, we observe that the pre-trained mBART can model better gloss sequence context than others, resulting in the best performance.

Ablation on the distinct γ_1 factors. Table 7 delivers the loss weight of the DDA in Eq. 7. We observe that the performance gradually increases as γ_1 increases until $\gamma_1 = 10$, then decreases after a certain value, the optimal weight for it is 10.

4.5 Other Evaluations

Evaluation on DDA generalizing other CSLR methods. As shown in Figure 4, we can see that with the DDA optimization, all methods have a consistent performance and generalizability boost. Specifically, the VAC [28] and TLP [25] gain remarkable performance improvement and generalizability enhancement. In particular, both TLP and SRN achieve remarkable recognition results (18.6% and 18.8% WERs) on the phoenix-2014 test set. These experiments validate the superior generalizability of the DDA as a plug-and-play optimization.

Method efficiency comparison. Similar to [44], we employ the THOP [27] tool to evaluate the parameters and GFLOPs of CSLR methods in the inference process and adopt the Throughout (videos/s). We adopt the 140 frames as the default. As shown in Table 8, our DDA achieves a balance between recognition performance and inference speed.

Table 8: Efficiency comparison between our DDA and other SOTA CSLR methods on the PHOENIX-2014. All experiments are measured on a A100 GPU with batch size 1.

Methods	Param	GFLOPs	Throughout	Dev (%) ↓	Test (%) ↓
VAC	34.3	567	17.0	21.2	22.3
TLP	59.5	573	17.0	19.7	20.2
SEN	34.5	578	15.5	19.5	21.0
C^2ST_{SW+WE}	78.2	1368	4.4	17.6	18.3
DDA (Ours)	70.1	1655	10.5	**16.9**	**17.3**

Evaluation on the generalization capability of DDA. We employ the compression of information stored in weights (IIW) [39] to quantitatively measure the generalization capability of DDA. According to the information bottleneck theory, the IIW value of the robust model should conform to a trend of rapid increase followed by a slow decrease. Our DDA is more consistent with this trend than the baseline.
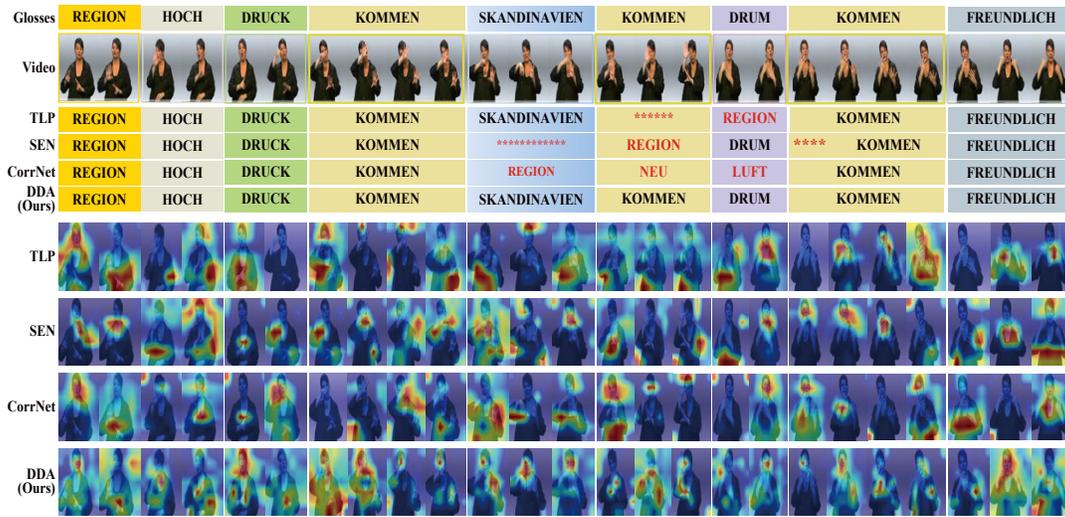
Figure 3: Visualizing the recognition results and the Grad-CAMs [36] results of TLP [25], SEN [18], CorrNet [17] and the proposed DDA on a PHOENIX-2014 test video. Glosses with red symbols denote the wrongly predicted gloss. The shades of color of the regions (blue, yellow, red, dark red) represent the weak to strong attention of the model to the sign spatial regions.
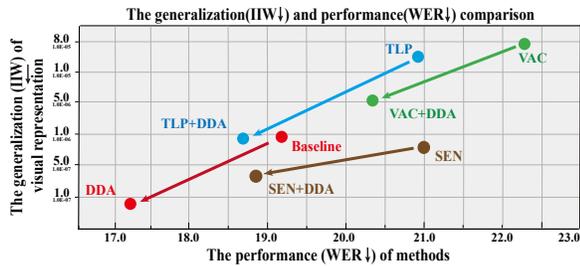

Figure 4: Evaluation for the generalization and recognition accuracy of DDA over other SOTA CSLR methods on the PHOENIX-2014 test set.
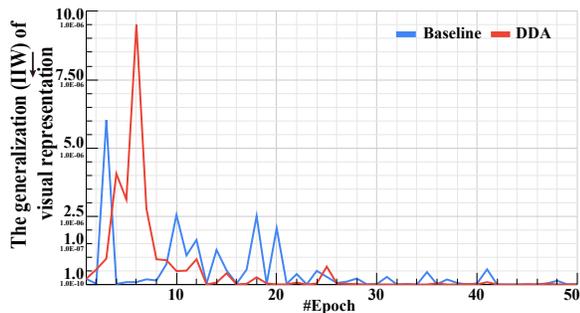

Figure 5: Evaluation on the generalization capability of DDA. IIW [39] (the compression of information stored in weights). The low IIW values denote the high generalization capability.

To further qualitatively evaluate the capability of our DDA to capture sign language movements. In Figure 3, we visualize the recognition results and the Class Activation Mapping (CAM) [36] of

TLP [25], SEN [18], CorrNet [17], and our DDA for an example video on the test set of PHOENIX-2014. For the gloss recognition, the verb gloss "KOMMEN" appears three times within the sentence, each time conveying different meanings corresponding to "came", "come from", and "will come". Therefore, the interpretation of "kommen" within a sentence varies, necessitating a refined global temporal context informed by the gloss global context for full understanding. Our DDA can recognize all "kommen", but TLP, SEN, and CorrNet fail to comprehend the second occurrence because they neglect to learn this refined global temporal context. For the Class Activation Mapping, we can observe that our DDA is more sensitive to sign movements and can accurately capture the sign movements of sign happened areas than other methods, which shows its powerful recognition ability.

5 CONCLUSION

This study investigated the effect of global temporal context alignment on continuous sign language recognition (CSLR). We propose a novel **D**enoising-**D**iffusion global **A**lignment (DDA), which consists of a denoising-diffusion autoencoder and DDA loss function, to maintain the diffusion alignment presentations gained by the denoising-driven process. The denoising-diffusion autoencoder is a self-supervised paradigm that performs the global temporal context alignment by denoising the partial noised bimodal representations. The DDA loss promotes the denoising process and facilitates the aligned knowledge transfer for the video representations. The excellent recognition performance on three publicly available CSLR benchmarks not only confirms the effectiveness of DDA but also corroborates the strong potential of the denoising-diffusion model for visual representation learning.

Limitations. Compared with the baseline method, about 8 minutes per epoch training time overhead will be incurred for DDA, which

faces excessive computational overhead. We may still need long-term research to achieve the denoising-diffusion model to work robustly in visual feature extraction and replace the alignment process in CSLR.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multi-modal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.

[2] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. 2023. One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale. In *ICML*, Vol. 202. 1692–1717.

[3] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2022. Label-Efficient Semantic Segmentation with Diffusion Models. In *ICLR*.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *CVPR*.

[5] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*.

[6] Shiming Chen, Ziming Hong, Guosen Xie, Wenhan Wang, Qinmu Peng, Kai Wang, Jian jun Zhao, and Xinge You. 2022. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. *CVPR*, 7602–7611.

[7] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. 2024. Deconstructing Denoising Diffusion Models for Self-Supervised Learning. *ArXiv*.

[8] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation. In *CVPR*. https://doi.org/10.1109/CVPR52688.2022.00506

[9] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie LIU, and Brian Mak. 2022. Two-Stream Network for Sign Language Recognition and Translation. In *NIPS*.

[10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*.

[11] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alex Smola. 2012. A Kernel Two-Sample Test. *J. Mach. Learn. Res.* 13 (2012), 723–773.

[12] Leming Guo, Wanli Xue, Qing Guo, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Shengyong Chen. 2023. Distilling Cross-Temporal Contexts for Continuous Sign Language Recognition. In *CVPR*.

[13] Aiming Hao, Yuecong Min, and Xilin Chen. 2021. Self-Mutual Distillation Learning for Continuous Sign Language Recognition. In *ICCV*.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*. 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NIPS*.

[17] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. Continuous Sign Language Recognition With Correlation Network. In *CVPR*.

[18] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023. Self-Emphasizing Network for Continuous Sign Language Recognition. In *AAAI*.

[19] Runhu Huang, Jianhua Han, Guansong Lu, Xiaodan Liang, Yihan Zeng, Wei Zhang, and Hang Xu. 2023. DiffDis: Empowering Generative Diffusion Model with Cross-Modal Discrimination Capability. *ICCV*, 15667–15677.

[20] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A Survey on Contrastive Self-Supervised Learning. *Technologies* 9 (2021).

[21] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. 2023. CoSign: Exploring Co-occurrence Signals in Skeleton-based Continuous Sign Language Recognition. In *ICCV*.

[22] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[23] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* (2015).

[24] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *ACL*.

[25] Zekang Liu Lianyu Hu, Liqing Gao and Wei Feng. 2022. Temporal Lift Pooling for Continuous Sign Language Recognition. In *ECCV*.

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. 2017. Deep Transfer Learning with Joint Adaptation Networks. In *ICML*, Vol. 70. 2208–2217.

[27] Lyken17, HaoKang-Timmy, lvmingzhe, and ttumiel. [n. d.]. Thop: Pytorch-opcounter. https://github.com/Lyken17/pytorch-OpCounter,2019.8

[28] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. 2021. Visual Alignment Constraint for Continuous Sign Language Recognition. In *ICCV*.

[29] Yuecong Min, Peiqi Jiao, Yanan Li, Xiaotao Wang, Lei Lei, Xiujuan Chai, and Xilin Chen. 2022. Deep Radial Embedding for Visual Sequence Learning. In *ECCV*.

[30] Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. 2020. Boosting continuous sign language recognition via cross modality augmentation. In *ACMMM*.

[31] Junfu Pu, Wengang Zhou, and Houqiang Li. 2019. Iterative alignment network for continuous sign language recognition. In *CVPR*.

[32] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2023. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *ICML*. 28223–28243.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

[34] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR* (2022), 10674–10685.

[35] Ludan Ruan, Y. Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. 2022. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. *ArXiv* abs/2212.09478 (2022).

[36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*. https://doi.org/10.1109/ICCV.2017.74

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

[38] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. 2023. Loss-Guided Diffusion Models for Plug-and-Play Controllable Generation. In *ICML*.

[39] Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Jimeng Sun, Xi Chen, and Yefeng Zheng. 2022. PAC-Bayes Information Bottleneck. In *ICLR*.

[40] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. SimMIM: A Simple Framework for Masked Image Modeling. In *CVPR*. 9653–9663.

[41] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. In *CVPR*.

[42] Wanli Xue, Ze Kang, Leming Guo, Shourui Yang, Tiantian Yuan, and Shengyong Chen. 2023. Continuous Sign Language Recognition for Hearing-Impaired Consumer Communication via Self-Guidance Network. *TCE* (2023). https://doi.org/10.1109/TCE.2023.3342163

[43] Wanli Xue, Jingze Liu, Siyi Yan, Yuxi Zhou, Tiantian Yuan, and Qing Guo. 2023. Alleviating data insufficiency for Chinese sign language recognition. *Visual Intelligence* 1 (2023), 1–9.

[44] Huaiwen Zhang, Zihang Guo, Yang Yang, Xin Liu, and De Hu. 2023. C2ST: Cross-Modal Contextualized Sequence Transduction for Continuous Sign Language Recognition. In *ICCV*.

[45] Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z. Li. 2023. CVT-SLR: Contrastive Visual-Textual Transformation for Sign Language Recognition With Variational Alignment. In *CVPR*.

[46] Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z. Li. 2023. CVT-SLR: Contrastive Visual-Textual Transformation for Sign Language Recognition With Variational Alignment. In *CVPR*.

[47] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving Sign Language Translation with Monolingual Data by Sign Back-Translation. In *CVPR*.

[48] Ronglai Zuo and Brian Mak. 2022. C2SLR: Consistency-Enhanced Continuous Sign Language Recognition. In *CVPR*.