Signer Diversity-driven Data Augmentation for Signer-Independent Sign Language Translation

Honghao Fu[1,3,4*] **Liang Zhang**[2,3,4*] **Biao Fu**[2,3,4] **Rui Zhao**[2,3,4]
Jinsong Su[1,2,3,4] **Xiaodong Shi**[1,2,3,4] **Yidong Chen**[1,2,3,4†]
[1]Institute of Artificial Intelligence, Xiamen University, China
[2]School of Informatics, Xiamen University, China
[3]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China
[4]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China
{fuhonghao, lzhang}@stu.xmu.edu.cn, ydchen@xmu.edu.cn

Abstract

The primary objective of sign language translation (SLT) is to transform sign language videos into natural sentences. A crucial challenge in this field is developing signer-independent SLT systems which requires models to generalize effectively to signers not encountered during training. This challenge is exacerbated by the limited diversity of signers in existing SLT datasets, which often results in suboptimal generalization capabilities of current models. Achieving robustness to unseen signers is essential for signer-independent SLT. However, most existing method relies on signer identity labels, which is often impractical and costly in real-world applications. To address this issue, we propose the Signer Diversity-driven Data Augmentation (SDDA) method that can achieve good generalization without relying on signer identity labels. SDDA comprises two data augmentation schemes. The first is data augmentation based on adversarial training, which aims to utilize the gradients of the model to generate adversarial examples. The second is data augmentation based on diffusion model, which focuses on using the advanced diffusion-based text guided image editing method to modify the appearances of the signer in images. The combination of the two strategies significantly enriches the diversity of signers in the training process. Moreover, we introduce a consistency loss and a discrimination loss to enhance the learning of signer-independent features. Our experimental results demonstrate our model significantly enhances the performance of SLT in the signer-independent setting, achieving state-of-the-art results without relying on signer identity labels.

1 Introduction

Sign languages are an indispensable communication medium for individuals who are deaf or hearing-impaired, utilizing the combination of handshapes, facial expressions, and body movements to convey information (Sutton-Spence and Woll, 1999). Converting sign language into spoken language sentences, known as Sign Language Translation (SLT), is an essential bridge that connects the deaf community with the hearing world (Camgoz et al., 2018; Yin et al., 2021), thus receiving increasing attention and leading to significant advancements by the research community in recent years (Camgoz et al., 2020a,b; Zhou et al., 2021a,b; Chen et al., 2022b,c; Zhang et al., 2023; Fu et al., 2023; Yu et al., 2023).

Despite these progresses, a major hurdle remains the limited signer diversity within the datasets used for training SLT models. For example, PHOENIX-2014T (Camgoz et al., 2018), a widely used dataset for German Sign Language, includes data from only nine different signers. This lack of signer diversity leads to a significant decrease in the performance of SLT models when confronted with data from unseen signers, a common occurrence in real-world applications (Jin and Zhao, 2021). The critical need for SLT systems that can generalize to unseen signers has led to the emergence of signer-independent SLT (Jin and Zhao, 2021) as a distinct and more challenging research focus.

Jin and Zhao (2021) propose a contrastive disentangled meta-learning method (CDM) to improve the ability of the model to generalize to unseen signers by disentangling signer-specific features from the sign language content. However, the effectiveness of CDM relies heavily on the availability of signer identity labels. As illustrated in Table 1, the generalization ability of CDM significantly diminishes in the absence of signer identity labels. This reliance limits the practical application of CDM, as acquiring such detailed signer information is often impractical and costly in real-world scenarios.

[*] Equal contribution
[†] Corresponding author

Method	signer 3			signer 4			signer 7			signer 8			Average		
	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R
CDM	41.11	16.86	41.70	44.52	19.18	45.29	40.37	15.94	41.29	42.84	17.70	43.30	42.21	17.42	42.90
w/o. id	39.71	15.57	40.49	41.12	17.80	43.30	37.03	14.81	39.71	40.85	16.63	42.70	39.68	16.20	41.55

Table 1: Comparing the translation performance of CDM w/. and w/o. id. **id** denotes signer identity labels. In the signer-independent setting, the PHOENIX-2014T (Camgoz et al., 2018) dataset can be divided into 4 situations, that is, signers 3, 4, 7, and 8 have not been seen, respectively.

To this end, we propose the **S**igner **D**iversity-driven **D**ata **A**ugmentation (**SDDA**) method to improve the generalization of the model to unseen signers without relying on signer identity labels. SDDA consists of two main components: Data Augmentation based on Adversarial Training (DAAT) and Data Augmentation based on Diffusion Model (DADM). **Firstly**, DAAT utilizes the gradients of the model to generate adversarial examples to enhance the robustness of the model to changes in the signers. Different from vanilla adversarial training methods (Goodfellow et al., 2014; Wang et al., 2022) that perturb the whole image indiscriminately, we only perturb the sign language non-critical regions of the image by introducing a keypoint masking. Since gestures and expressions contain rich information of sign language, we regard these parts as critical regions and other parts as non-critical regions. In this way, we can improve the model's robustness to signers without losing the semantics of the sign language. **Secondly**, motivated by the recent advancements in diffusion-based image generation models, DADM applies a variety of elaborate prompts to guide the diffusion-based text-guided image editing model (Rombach et al., 2022; Meng et al., 2021) in modifying the appearance of the signer for each video frame, thereby significantly increasing the diversity of signers. When combined, DAAT and DADM provide comprehensive augmentation for signers. DAAT ensures that the model is robust to small changes and noise occurring in the signer, while DADM significantly increases the diversity of signer appearances that the model is exposed to during training.

To effectively learn signer-independent representations, we introduce a consistency loss and a discrimination loss into our model training. The former minimizes the KL divergence between the output distributions of the original and augmented samples to ensure that the augmented data does not deviate semantically from the original data. The latter trains a discriminator to distinguish the features from the original and augmented samples, ensuring that the model's feature extraction is robust to variations in signer appearance.

We conduct a variety of experiments on the PHOENIX-2014T (Camgoz et al., 2018) benchmark to verify the effectiveness of our model. Experimental results indicate that SDDA effectively enhances the performance of SLT in the signer-independent setting without relying on signer identity labels, achieving state-of-the-art results.

2 Related Work

2.1 Sign Language Translation.

SLT seeks to convert raw videos into spoken language sentences. Camgoz et al. (2018) firstly introduce an end-to-end neural SLT model that fuses Convolutional Neural Networks (CNNs) and the attention-based sequence-to-sequence model. Their goal is to jointly learn the alignment and translation processes from sign videos to spoken language sentences. However, the advancement of SLT is hampered by data scarcity. To address this issue, Camgoz et al. (2020b) simultaneously train SLR and SLT, aiming to regularize the translation encoder. Camgoz et al. (2020a) and Zhou et al. (2020) propose a multi-channel transformer architecture to utilize multiple visual cues in sign language. Li et al. (2020) introduce a hierarchical sign video feature learning method, which use a temporal semantic pyramid network to learn more discriminative features. Zhou et al. (2021a) design a data augmentation method that uses gloss as pivot to generate more visual features from text. Fu et al. (2023) propose a token-level contrastive learning framework to improve token representation effectiveness. Chen et al. (2022b) propose a multi-modal pretraining approach to cope with the data scarcity issue for SLT.

The aforementioned works belong to conventional SLT methods, which do not take into account the model's generalization ability to unseen signers. Jin and Zhao (2021) first introduce the

task of signer-independent SLT. They propose a framework called contrastive disentangled meta-learning, which relies heavily on signer identity to learn signer-independent feature. In contrast to them, we have designed two data augmentation methods to enhance the model's generalization to unseen signers without relying on signer identity.

2.2 Domain Generalization

Domain generalization (DG) aims to train models on known domains that can generalize well to unseen domains. Over the past decades, a variety of DG algorithms have been proposed. Shao et al. (2019) propose a multi-adversarial discriminative deep domain generalization framework, aiming to learn a generalized feature space. Dai et al. (2021) propose the relevance-aware mixture of experts, which utilize an effective voting-based mixture mechanism. This dynamic approach leverages diverse characteristics from source domains, thus enhancing the model's generalization capabilities. Lv et al. (2022) introduce a general structural causal model, providing a formalized framework for addressing the challenges within DG.

However, the above methods require domain labels that are not available in many real-world scenarios. To solve this problem, Huang et al. (2020) introduce Representation Self-Challenging, a technique that discards dominant features activated iteratively during training, compelling the network to activate remaining features correlated with labels. Chen et al. (2022a) present Compound Domain Generalization via Meta-knowledge ENcoding (COMEN), a two-step approach that autonomously discovers and models latent domains, eliminating the need for explicit domain labels. Qu et al. (2022) leverage hypernetworks, taking vectors as input to generate experts' weights. This unique approach enables the sharing of useful meta-knowledge among experts and facilitate exploration of experts' similarities in a low-dimensional vector space. Vidit et al. (2023) leverage a pre-trained vision-language model to introduce semantic domain concepts via textual prompts, providing an innovative avenue for domain generalization without explicit domain labels.

Signer-independent SLT can be viewed as a domain generalization task, where different signers with varying appearances are treated as different domains, and signer identity serves as the domain label. However, obtaining domain labels, in this case, signer identity, is often expensive in real-life scenarios, as recent domain generalization methods have highlighted. Therefore, we propose a novel signer-diversity driven data augmentation for this task, eliminating the need for relying on signer identity.

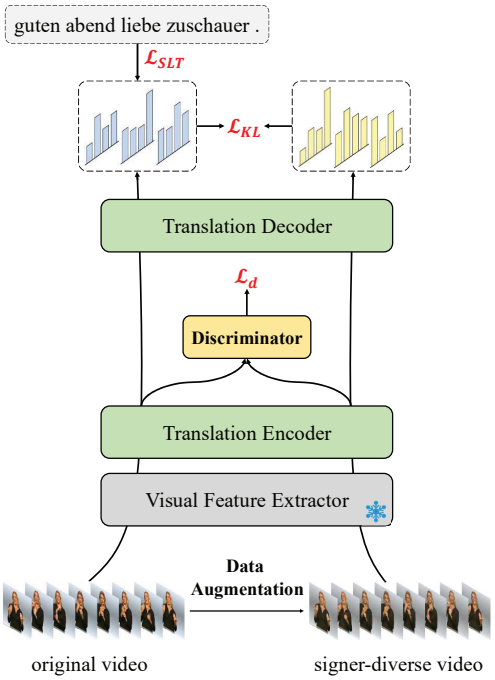

Figure 1: The overall framework of SDDA. With data augmentation based on adversarial training and data augmentation based on diffusion model, the original video transforms to signer-diverse video, which has same sign language semantics but different signer. Subsequently, discriminator determines whether the hidden state belongs to the original video or the synthetic video.

3 Approach

A typical SLT corpus contains video-sentence pairs, which can be denoted as $\mathcal{D}^{\text{SLT}} = \{(\mathbf{x}, \mathbf{y})\}$. Here $\mathbf{x} = (x_1, \cdots, x_{T_x})$ denotes a sign video with T_x frames and $\mathbf{y} = (y_1, \cdots, y_{T_y})$ is the corresponding spoken sentence with T_y token. SLT systems aim to translate sign video $\mathbf{x}$ to the spoken sentence $\mathbf{y}$. The training objective of SLT is the cross-entropy loss defined as follows:

$$\mathcal{L}_{\text{SLT}} = -\log p_\theta(\mathbf{y}|\mathbf{x}). \tag{1}$$

where θ is the parameters of model.

In the signer-independent setting, the SLT model is trained on the signer group $\mathbf{g}$ and subsequently tested on another signer group $\mathbf{g}'$, where $\mathbf{g} \cap \mathbf{g}' = \emptyset$. However, the limited availability of signers in current SLT datasets restricts the ability of the SLT model to generalize effectively to unseen signers.

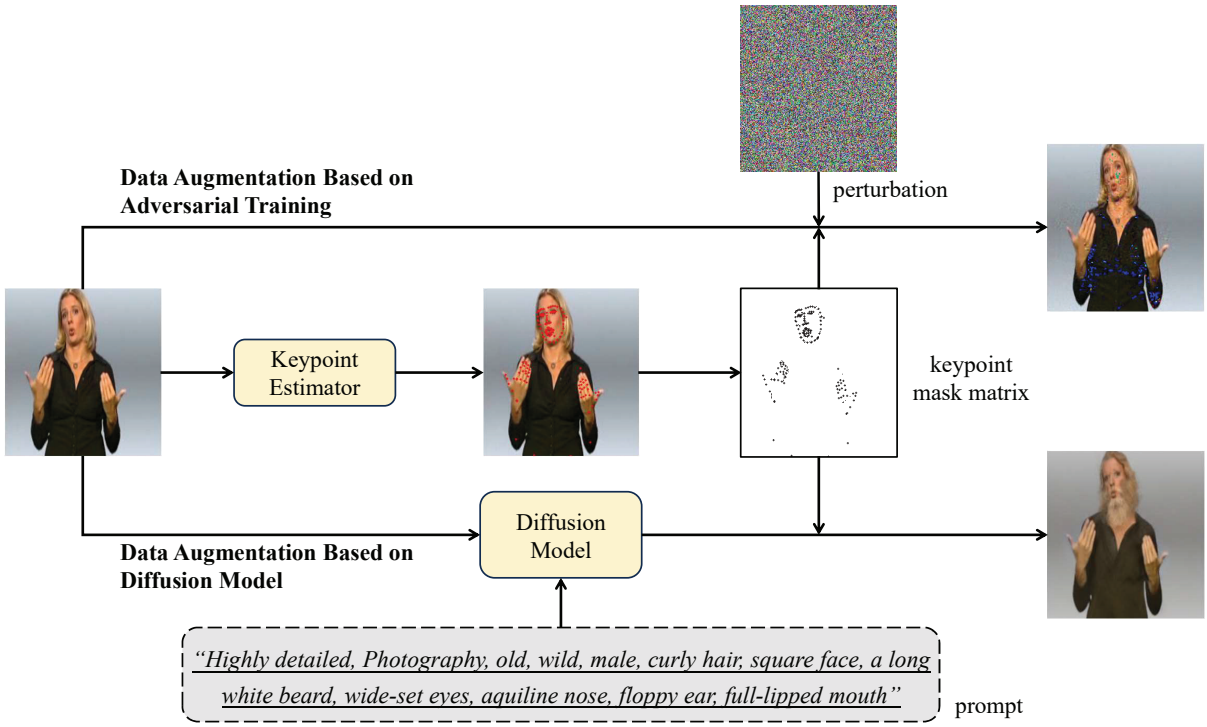

Figure 2: Detailed process of our two data augmentations.

To address this limitation, we propose a novel signer diversity-driven data augmentation method, which consists of data augmentation based adversarial training (Secction 3.1) and data augmentation based on diffusion model (Secction 3.2). The overall framework of SDDA is illustrated in Figure 1.

3.1 Data Augmentation Based on Adversarial Training

To enhance the model's robustness against variations in signer identity, we employ an adversarial strategy to generate signer gradient-perturbed images, as shown in the upper panel of Figure 2. Given a sign video-sentence pair $(\mathbf{x}, \mathbf{y})$, we add a perturbation $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_{T_x}] \in \mathbb{R}^{T_x \times C \times H \times W}$ to the sign video $\mathbf{x}$, such that its conditional likelihood is minimized as follows:

$$\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}, \tag{2}$$

$$\boldsymbol{\delta} = \underset{\boldsymbol{\delta}, \|\boldsymbol{\delta}\|_2 \leq \epsilon}{\arg\min} \log p_\theta(\mathbf{y} \mid \mathbf{x} + \boldsymbol{\delta}). \tag{3}$$

Following Goodfellow et al. (2014), the minimization of the conditional log likelihood with respect to $\boldsymbol{\delta}$ can be approximated as:

$$\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \boldsymbol{g}, \tag{4}$$

where $\boldsymbol{g} = \nabla_{\mathbf{x}} \mathcal{L}_{\text{SLT}}$ and ϵ is a scalar controlling the perturbation magnitude. Considering that both the face and the hand of the signer contain rich information, we add perturbations exclusively to regions beyond these critical areas. To achieve this, we introduce a keypoint mask matrix M. Consequently, Eq.(4) is modified as:

$$\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \boldsymbol{g} \odot M, \tag{5}$$

where $\odot$ denotes element-wise multiplication, ensuring that perturbations are applied selectively based on the mask matrix M. To obtain the keypoint mask matrix, we first employ an off-the-shelf keypoint estimator (Wang et al., 2020) to generate keypoint sequences, then set the mask values in the regions corresponding to the keypoints of the hands and face to 0 and the others to 1. After applying the masking operation, the perturbations are restricted to non-critical regions, maintaining the semantic integrity of the sign language content in the perturbed videos.

3.2 Data Augmentation Based on Diffusion Model

Diffusion models have demonstrated remarkable ability in generating high-quality, diverse, and customized samples that are perceptually similar to real data (Ramesh et al., 2022; Rombach et al., 2022). Inspired by this, we propose a data augmentation strategy based on the diffusion model, as

illustrate in the lower part of Figure 2. Leveraging the strengths of diffusion-based text-guided image editing techniques[1] (Rombach et al., 2022; Meng et al., 2021), our method carefully modify each image in the SLT dataset to expand the diversity of signers, thus enhancing the model's generalization to unseen signers.

Specifically, we first craft a series of prompts that depict various human facial features, such as eyes, mouth, and nose. Then, we utilize these descriptive prompts to guide the diffusion model in modifying the images to yield a variety of signer appearances. In parallel, we employ the previously established keypoint mask M to ensure that modifications are confined to non-critical areas of the sign video, thereby preserving the semantic integrity of the sign language information. By applying this approach, we obtain a new dataset in which the signer in each video frame feature a unique face. This not only expands the size of the SLT dataset but also effectively increases the variety of signers. We present some augmented examples and their corresponding prompts in the Appendix A.

3.3 Training Objective

To learn signer-independent features, we introduce a consistency loss and a discrimination loss to align the features of the original and augmented videos.

Consistency Loss Since the augmented sample expresses the same semantics as the original sample, we regularize the output predictions of the original and augmented samples by minimizing the Kullback-Leibler (KL) Divergence between their output distributions. Given the original sample $\mathbf{x}$ and the augmented sample $\tilde{\mathbf{x}}$, the consistency loss is defined as:

$$\mathcal{L}_{\text{KL}} = \sum_{t=1}^{T_y} \text{KL}(p_\theta(y_t|y_{<t}, \mathbf{x})||p_\theta(y_t|y_{<t}, \tilde{\mathbf{x}})) \quad (6)$$

This loss encourages the model to produce consistent token predictions for both original and synthetic data, thus enabling the model to learn more robust visual features for the signers.

Discrimination Loss We also introduce a discriminator to distinguish the augmented samples from the original samples, which facilitates the model to learn consistent global contextual representations for different signers expressing the same sign language semantics. For the hidden states $\mathbf{h}$ and $\tilde{\mathbf{h}}$ of the original and augmented samples output by the encoder, the discriminator aims to distinguish whether the hidden state is from the original or augmented samples. Thus, the discrimination loss is defined as:

$$\mathcal{L}_d = \log p(1|D(\mathbf{h})) + \log p(0|D(\tilde{\mathbf{h}})) \quad (7)$$

where D denotes the discriminator. By incorporating this discriminator, we ensure that the encoder representations from original and augmented samples become indistinguishable during training.

Our discriminator consists of a gradient reversal layer (Ganin and Lempitsky, 2015), followed by a mean pooling function, a two-layer feed-forward network and the softmax operation.

Finally, the overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{SLT}} + \alpha\mathcal{L}_{\text{KL}} + \beta\mathcal{L}_d, \quad (8)$$

where α, β are hyperparameters which control the importance of each loss.

4 Experiments

4.1 Dataset and Metrics

Datasets. We assess the performance of our model on the PHOENIX-2014T benchmark dataset (Camgoz et al., 2018), which comprises sign language videos, gloss annotations, and spoken language translations sourced from the German Weather Forecast. This dataset is labeled for 9 different signers. To comprehensively evaluate the effectiveness of our model, we employ four distinct experimental settings, wherein the data of signers 3, 4, 7, and 8 constitute the test set, while the remaining data serve as the training or validation set. Given the relatively limited data for signers 2, 6, and 9, we allocate their corresponding data to the validation set. The detailed statistical results are listed in Table 7.

Evaluation metrics. To fairly evaluate the effectiveness of our SDDA, we use BLEU-N (Ngrams ranges from 1 to 4) (Papineni et al., 2002) and ROUGE-L (Lin and Och, 2004) as the evaluation metrics. BLEU-N measures precision up to n-grams, while ROUGE-L calculates the F1 score based on the longest common sub-sequences between predictions and ground-truth translations.

4.2 Implementation Details

Both the encoder and decoder of Transformer have 12 layers. The size of the word embedding and the

[1] https://huggingface.co/stabilityai/stable-diffusion-2-1

Method	signer 3			signer 4			signer 7			signer 8			Average		
	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R
Neural-SLT (Camgoz et al., 2018)	38.70	13.54	39.52	40.35	14.84	41.06	36.30	12.89	38.56	38.02	13.65	39.08	38.34	13.73	39.56
TSPNet (Li et al., 2020)	40.96	15.35	41.17	42.64	17.52	43.85	38.92	14.48	40.27	41.74	15.59	42.21	41.07	15.74	41.88
Joint-SLRT (Camgoz et al., 2020b)	40.66	15.17	41.29	42.39	17.76	43.91	38.62	14.34	39.94	41.70	15.41	42.02	40.84	15.67	41.79
CDM (w/o. id) (Jin and Zhao, 2021)	39.71	15.57	40.49	41.12	17.80	43.30	37.03	14.81	39.71	40.85	16.63	42.70	39.68	16.20	41.55
MMTLB (Chen et al., 2022b)	44.95	19.48	42.62	50.15	25.82	50.07	39.07	16.11	40.67	44.92	20.06	44.64	44.77	20.37	44.50
SDDA (Ours)	**46.40**	**20.89**	**44.47**	**52.28**	**27.77**	**52.03**	**41.09**	**17.16**	**41.55**	**46.45**	**21.43**	**45.99**	**46.56**	**21.81**	**46.01**

Table 2: In comparison to SLT methods in signer-independent setting without signer identity labels, where B@1, B@4, and R represent BLEU-1, BLEU-4, and ROUGE-L, respectively.

Method	signer 3			signer 4			signer 7			signer 8			Average		
	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R	B@1	B@4	R
SDDA	46.40	**20.89**	**44.47**	**52.28**	**27.77**	**52.03**	**41.09**	**17.16**	**41.55**	**46.45**	**21.43**	**45.99**	**46.56**	**21.81**	**46.01**
w/o. DAAT	45.03	20.16	43.20	50.87	26.72	51.15	40.16	16.50	41.36	46.47	20.60	45.48	45.63	21.00	45.30
w/o. DADM	**46.47**	20.33	43.86	51.99	26.68	51.16	40.09	16.68	40.59	46.02	20.70	44.86	46.14	21.10	45.12

Table 3: Ablation study of SDDA for singer-independent SLT

hidden is 1024. We use 16 attention heads for each layer. The network parameters are initialized with Kaiming (He et al., 2015), and a shared weight matrix is employed for the input and output word embeddings in the decoder during training. We adopt the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and Cosine Annealing learning rate schedule. The visual feature extractor (Chen et al., 2022b) first performs SLR pre-training on a dataset without unseen signers. Our model is trained with batch size 32 and initial learning rate 1e-5. The dropout rate is 0.3. We set α, β to 1.0, 1e-4. ϵ is 2.0, 3.0, 3.0, 1.0 for signer 3, 4, 7, 8 respectively. SDDA requires to be trained on 1 NVIDIA TITAN RTX GPU with 24 GB memory for 30 hours.

4.3 Comparison Results

We compare SDDA with several state-of-the-art SLT methods, Neural-SLT (Camgoz et al., 2018), TSPNet (Li et al., 2020), Joint-SLRT (Camgoz et al., 2020b), CDM (Jin and Zhao, 2021), MMTLB (Chen et al., 2022b), in a signer-independent setting without using signer identity.

As presented in Table 2, SDDA achieves state-of-the-art results. Previous methods, which did not account for generalization to unseen signers, focused solely on extracting cues specific to signers in the training set, resulting in lower scores. In comparison to previous methods, CDM exhibits limited performance improvement. This is attributed to CDM's primary reliance on signer identity labels as supervision to enhance the model's generalization to unseen signers. On the other hand, MMTLB, which is pre-trained on a substantial amount of sign language-related data, outperforms these methods. Through signer diversity-driven data augmentation, we have alleviated data scarcity to a certain extent and increased the types of signers in the dataset. By aligning the features of original data and synthetic data, SDDA learns signer-independent features. Therefore, the generalization of SDDA has been further improved.

4.4 Ablation Study

To assess the effectiveness of all contributions, we conduct a comprehensive evaluation by comparing SDDA against a series of ablation models with various settings. As indicated in Table 3, **w/o. DAAT** represents the model without data augmentation based adversarial training and **w/o. DADM** represents the model without diffusion model based data augmentation.

From Table 3, we consistently observe that SDDA outperforms **w/o. DAAT** in terms of BLEU and ROUGE. This improvement can be attributed to SDDA's alignment of features between original samples and adversarial samples generated by adversarial training. Through applying gradient perturbations to sign language non-critical parts of the image, this augmentation method produces adversarial examples that preserve the key semantics of sign language and reduce the translation accuracy of the sign language model. By aligning features of original and adversarial samples, the model improves its robustness to changes in signers.

Comparing SDDA with **w/o. DADM**, we observe that performance is notably poorer in the absence of diffusion model based data augmentation,

highlighting its effectiveness. By using various prompts to guide the diffusion model to modify the image, this method generates a variety of data from signers. This data nicely simulates real-world scenarios in which different people perform sign language. By aligning the features of the original images with those of these synthesized images, SDDA improves generalization to unseen signers.

4.5 Comparison with other data augmentation methods

To further validate the effectiveness of signer diversity-driven data augmentation, we compare it with three prominent data augmentation schemes (Cubuk et al., 2019, 2020; Müller and Hutter, 2021), which effectively enhance the accuracy in the task of image classification. These methods use reinforcement learning to combine various data augmentation operations, modifying the color, brightness, contrast, and other properties of the image. However, as depicted in Table 4, these augmentation solutions prove ineffective in enhancing the model's generalization to unseen signers. This highlights the limited impact of conventional data augmentation methods on signer-independent SLT. In contrast, our data augmentation method focuses on enhancing the diversity of signers in the dataset, thereby improving the model's generalization to unseen signers.

4.6 Further analysis of data augmentation based on adversarial training

To further analyze the contribution of each component in data augmentation based on adversarial training, we compared the performance of SDDA under various settings of this approach. As shown in Table 5,**w/o. keypoint mask** represents data augmentation based on adversarial training without keypoint mask and **w/o. discriminator** represents our model without discriminator. **Impulse noise** and **Gaussian noise** mean utilizing other perturbation, Impulse noise and Gaussian noise, instead of gradient perturbation during training. Compared to **w/o. keypoint mask**, which adds perturbations to all pixels, SDDA achieves a higher score. This indicates that applying perturbations to all pixels in the image does not yield qualified adversarial samples. Instead, it adversely affects the sign language information in the image, hindering the effective improvement of model generalization. When compared with **w/o. discriminator**, SDDA demonstrates improved performance. As the discriminator aims to differentiate the temporal mean-pooling hidden representation of the original sample from that of the augmented sample, and the translation model seeks to fool the discriminator, our model learns global contextual representations for different signers. To fully demonstrate the effectiveness of gradient perturbation, we replaced it with two common perturbations (Chantry et al., 2022). In Table 5, compared with these alternatives, SDDA effectively enhances the generalization of the model. This improvement is attributed to gradient perturbation, which is generated through adversarial training. Pictures with such perturbations are included to better simulate scenarios where changes in the signer result in reduced translation performance.

Method	B@1	B@4	R
Auto (Cubuk et al., 2019)	50.49	25.87	50.01
Rand (Cubuk et al., 2020)	50.04	25.28	49.49
Trival (Müller and Hutter, 2021)	49.61	25.55	50.04
Ours	**52.28**	**27.77**	**52.03**

Table 4: Replace signer diversity-driven data augmentation with other data augmentation methods. It can be observed that the performance of the model has not been effectively improved. Note that we conduct experiments based on using signer4 as the unseen signer.

Method	B@1	B@4	R
SDDA	**52.28**	**27.77**	**52.03**
w/o. keypoint mask	51.89	26.86	50.82
w/o. discriminator	52.01	27.24	51.38
perturbation → Impulse noise	42.37	25.85	50.55
perturbation → Gaussian noise	51.48	26.43	51.00

Table 5: Further analysis of data augmentation based on adversarial training. Note that we conduct experiments based on using signer 4 as the unseen signer.

4.7 Qualitative Analysis

We present the translation quality of SDDA in this section, showcasing translation samples in Table 6. Due to space constraints, we exclusively provide results for CDM (w/o. id) and SDDA, alongside the ground truth translations serving as references. Given that the annotations in the PHOENIX-2014T dataset are in German, the generated sentences and their English translations are shared. A comparison reveals that, owing to the data augmentation method proposed in this paper, our model performs well even on unseen signers. When compared with CDM (w/o. id), our model accurately translates key information in the reference. As shown in the third example, our model got the correct transla-

Reference	am mittwoch und donnerstag bleibt es häufiger trüb örtlich etwas sprühregen stellenweise zeigt sich die sonne.
	(Wednesday and Thursday it will often be cloudy, with some drizzle in places and the sun will appear in places.)
CDM (w/o. id)	am donnerstag ist es teils wolkig oder neblig trüb teils freundlich.
	(Thursday it will be partly cloudy or foggy, partly friendly.)
SDDA	am mittwoch und donnerstag verbreitet trübes wetter gebietsweise regnet es etwas teilweise zeigt sich die sonne.
	(Wednesday and Thursday will have widespread cloudy weather with scattered rain and occasional sunshine in some areas.)
Reference	dann morgen von osten schon wieder trockener.
	(then tomorrow it will be drier again from the east.)
CDM (w/o. id)	morgen bleibt es meist trocken und trocken.
	(Tomorrow it will mostly stay dry and dry.)
SDDA	morgen bleibt es im nordosten noch trocken.
	(Tomorrow it will still be dry in the northeast.)
Reference	daran ändert sich am dienstag in der nordhälfte nur wenig.
	(Little will change in the northern half on Tuesday.)
CDM (w/o. id)	am dienstag ist es im norden und auch im norden bleibt es recht kühl.
	(On Tuesday it will be in the north and it will also remain quite cool in the north.)
SDDA	am dienstag ändert sich an diesem wetter im norden wenig.
	(On Tuesday there will be little change in this weather in the north.
Reference	am alpenrand kann es länger anhaltend regnen.
	(It can rain for a long time on the edge of the Alps.)
CDM (w/o. id)	in den alpen regnet es gebietsweise kräftig.
	(In the Alps it rains heavily in some areas.)
SDDA	an den alpen kann es längere zeit regnen.
	(It can rain for a long time in the Alps.)
Reference	morgen fünf grad im allgäu bis elf an rhein elbe und saale.
	(Tomorrow five degrees in the Allgäu until eleven on the Rhine Elbe and Saale.)
CDM (w/o. id)	am tag fünf grad am tag fünf grad am niederrhein und fünf grad am niederrhein.
	(on the day five degrees on the day five degrees on the Lower Rhine and five degrees on the Lower Rhine.)
SDDA	am tag fünf grad im allgäu und fünf grad an rhein und main.
	(on the day five degrees in the Allgäu and five degrees on the Rhine and Main.)

Table 6: Qualitative Results of SDDA

tion: "ändert wenig" (changes little), but CDM (w/o. id) didn't. Besides, Comparing the translations of the two models, it is obvious that our model translates the whole sentence more completely and smoothly. In the first example, CDM (w/o. id) misleads "mittwoch" (wednesday) but our model's translation results include it. Lastly, from these examples, we can see that our model generates fewer under-translation sentences.

5 Conclusion

In this work, we propose SDDA, a signer diversity-driven augmentation for signer-independent SLT. SDDA comprises two data augmentation methods. The first is data augmentation based on adversarial training, which focuses on using the gradient of the model to generate adversarial samples. The second is data augmentation based on the diffusion model, which focuses on using the advanced diffusion based text guide image editing method to edit the signers in the picture, alleviating the problem of scarcity of signer diversity. By employing the two data augmentation methods, each frame in the sign language video can be transformed into a signer-diverse image. To learn signer-independent features, we introduce a consistency loss and a discrimination lass to align the features of the original and augmented videos. Through our method, the model learns more robust visual features and consistent global contextual representations for different signers, thus improving the generalization ability of the model to unseen signers. Experimental results on the benchmark PHOENIX-2014T affirm the effectiveness of SDDA.

6 Limitation

Our methods involve data augmentation based on adversarial training and data augmentation based on the diffusion model. However, our approach faces three limitations. Firstly, Our method requires a long training time. Due to adversarial training, the model computes the same sample twice. Furthermore, the diffusion model's high computational complexity results in prolonged data synthesis times. Secondly, we concentrate on designing prompts to enhance signer diversity but have not explored how to design prompts that make synthesized pictures closer to real pictures. Future work will focus on how to efficiently generate realistic and signer-diverse data. Thirdly, our method does not take into account how to deal with the dynamic nature between different signers, such as variations in performance styles.

7 Acknowledgements

We would like to thank all the anonymous reviewers for their insightful and valuable comments. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62076211 and in part by the University-Industry Cooperation Programs of Fujian Province of China under Grant No. 2023H6001.

References

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

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 301–319. Springer.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Madeline Chantry, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. 2022. Robustness analysis of video-language models against visual and language perturbations. *Advances in Neural Information Processing Systems*, 35:34405–34420.

Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. 2022a. Compound domain generalization via meta-knowledge encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7129.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022b. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5130.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie LIU, and Brian Mak. 2022c. Two-stream network for sign language recognition and translation. In *Advances in Neural Information Processing Systems*, volume 35, pages 17043–17056. Curran Associates, Inc.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. 2021. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16145–16154.

Biao Fu, Peigen Ye, Liang Zhang, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2023. A token-level contrastive framework for sign language translation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. 2020. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer.

Tao Jin and Zhou Zhao. 2021. Contrastive disentangled meta-learning for signer-independent sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5065–5073.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. 2020. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8046–8056.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Samuel G Müller and Frank Hutter. 2021. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 774–782.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jingang Qu, Thibault Faney, Ze Wang, Patrick Gallinari, Soleiman Yousef, and Jean-Charles de Hemptinne. 2022. Hmoe: Hypernetwork-based mixture of experts for domain generalization. *arXiv preprint arXiv:2211.08253*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. 2019. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10031.

Rachel Sutton-Spence and Bencie Woll. 1999. *The linguistics of British Sign Language: an introduction*. Cambridge University Press.

Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. 2023. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3219–3229.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364.

Ruibin Wang, Yibo Yang, and Dacheng Tao. 2022. Art-point: Improving rotation robustness of point cloud classifiers via adversarial rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14371–14380.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Pei Yu, Liang Zhang, Biao Fu, and Yidong Chen. 2023. Efficient sign language translation with a curriculum-based non-autoregressive decoder. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5260–5268. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Biao Zhang, Mathias Müller, and Rico Sennrich. 2023. SLTUNET: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations*.

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

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2020. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016.

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

A Augmented examples

Here are some diffusion model based augmented examples and corresponding prompts shown in Fig 3.

B Augmented examples

We list the statistical results of the signer-independent settings in Table 7.

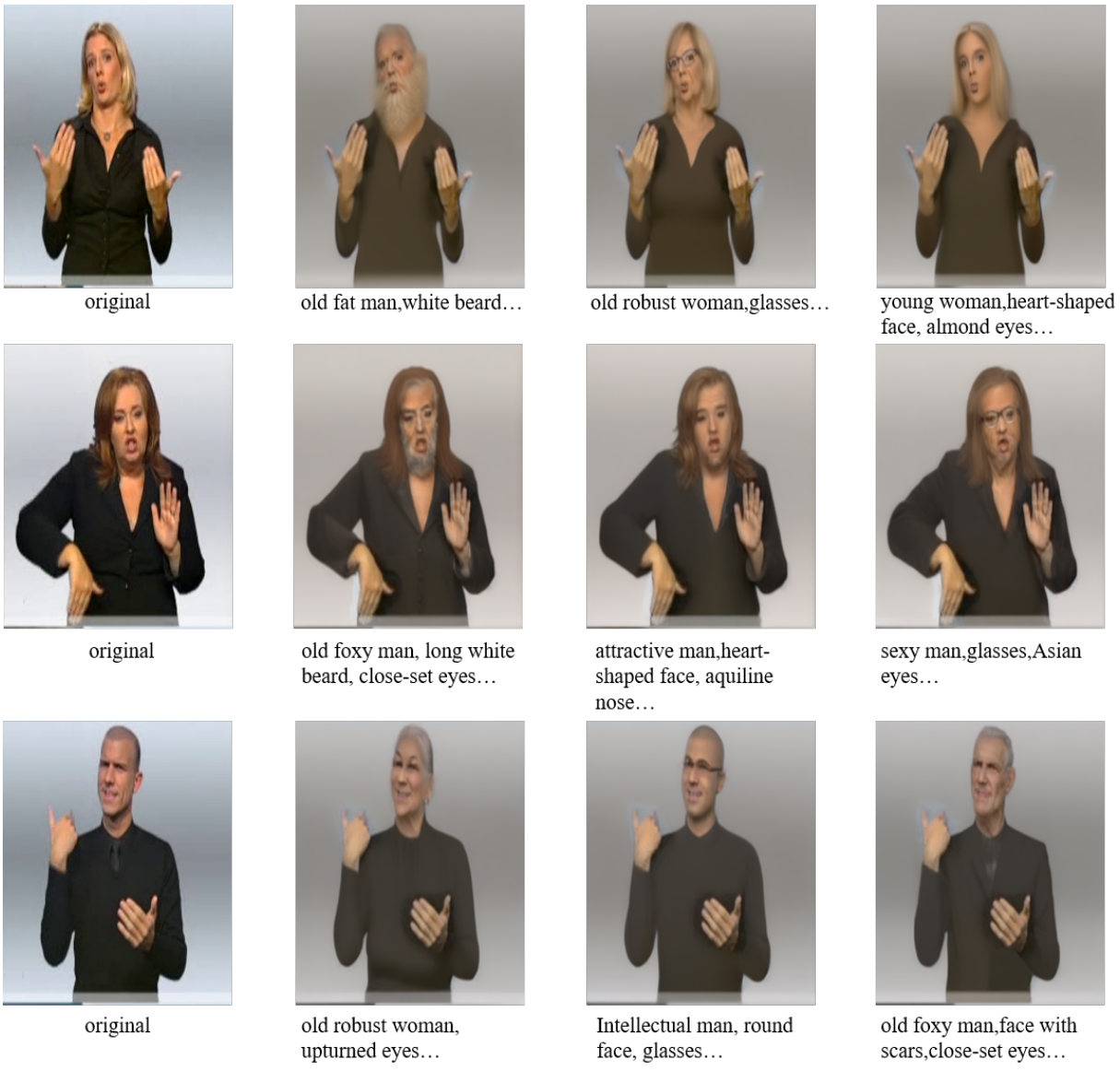

Figure 3: Some diffusion model based augmented examples and corresponding prompts.

Type	Split 1		Split 2		Split 3		Split 4	
	Signers	Samples	Signers	Samples	Signers	Samples	Signers	Samples
Train	1, 4, 5, 7, 8	7163	1, 3, 5, 7, 8	6639	1, 3, 4, 5, 8	6980	1, 3, 4, 5, 7	6880
Dev	2, 6, 9	411	2, 6, 9	411	2, 6, 9	411	2, 6, 9	411
Test	3	683	4	1207	7	866	8	966

Table 7: The statistical results of the signer-independent settings.