Diffusion Models for Sign Language Video Anonymization

Zhaoyang Xia[1], **Yang Zhou**[1], **Ligong Han**[1], **Carol Neidle**[2], **Dimitris N. Metaxas**[1]
[1] Rutgers University, [2] Boston University
[1] 110 Frelinghuysen Road, Piscataway, NJ 08854,
[2] Boston University Linguistics, 621 Commonwealth Ave., Boston, MA 02215
zx149@rutgers.edu, eta.yang@rutgers.edu, lh599@scarletmail.rutgers.edu,
carol@bu.edu, dnm@cs.rutgers.edu

Abstract

Since American Sign Language (ASL) has no standard written form, Deaf signers frequently share videos in order to communicate in their native language. However, this does not preserve privacy. Since critical linguistic information is transmitted through facial expressions, the face cannot be obscured. While signers have expressed interest, for a variety of applications, in sign language video anonymization that would effectively preserve linguistic content, attempts to develop such technology have had limited success and generally require pose estimation that cannot be readily carried out in the wild. To address current limitations, our research introduces DiffSLVA, a novel methodology that uses pre-trained large-scale diffusion models for text-guided sign language video anonymization. We incorporate ControlNet, which leverages low-level image features such as HED (Holistically-Nested Edge Detection) edges, to circumvent the need for pose estimation. Additionally, we develop a specialized module to capture linguistically essential facial expressions. We then combine the above methods to achieve anonymization that preserves the essential linguistic content of the original signer. This innovative methodology makes possible, for the first time, sign language video anonymization that could be used for real-world applications, which would offer significant benefits to the Deaf and Hard-of-Hearing communities.

Keywords: Sign Language Anonymization, Diffusion Model, Text-to-Video Editing, ASL

1. Introduction

American Sign Language (ASL), the predominant language used by the Deaf Community in the US and parts of Canada, is a full-fledged natural language. It employs manual signs in parallel with non-manual elements, including facial expressions and movements of the head and upper body, to convey linguistic information. The non-manual elements are crucial for conveying many types of lexical and adverbial information, as well as for marking syntactic structures (e.g., negation, topics, question status, and clause types (Baker-Shenk, 1985; Kacorri and Huenerfauth, 2016; Neidle et al., 2000; Coulter, 1979; Valli and Lucas, 2000)). Consequently, in video communications, e.g., on the Web, involving sensitive subjects such as medical, legal, or controversial matters, obscuring the face for purposes of anonymity would result in significant loss of essential linguistic information.

Despite the fact that several writing systems have been developed for ASL (Arnold, 2009), the language has no standard written form. While ASL signers could use written English in order to preserve privacy, that is frequently not their preference, as signers generally have greater ease and fluency in their native language, ASL, than in English.

Many Deaf signers have shown interest in a mechanism that would maintain the integrity of linguistic content in ASL videos while disguising the identity of the signer, as discussed in several recent studies (Lee et al., 2021). There are many potential applications of such a tool. For example, this could enable anonymous peer review for academic submissions in ASL. This could also ensure impartiality in various multimodal ASL-based applications, e.g., enabling production of neutral definitions for ASL dictionaries, not tied to the identity of the signer producing them. It could also enable maintenance of neutrality in interpretation scenarios. Additionally, such a tool could increase signers' willingness to contribute to video-based AI datasets (Bragg et al., 2019b), which hold significant research value.

For these reasons, privacy preservation for ASL videos has been explored (Isard, 2020). However, most of these approaches suffer from limitations with respect to preservation of linguistic meaning, and they generally achieve only a limited degree of anonymity. They also require accurate pose estimation, and some require substantial human labor. These limitations significantly reduce the potential for practical applications of such technologies.

To overcome the limitations of existing anonymization tools, we introduce DiffSLVA, a novel anonymization approach leveraging large-scale pre-trained diffusion models, notably Stable Diffusion (Rombach et al., 2022). DiffSLVA is designed to tackle text-guided sign language anonymization. Through a text prompt, it generates a new video in which the original linguistic meaning is retained, but

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 395–407
25 May 2024.

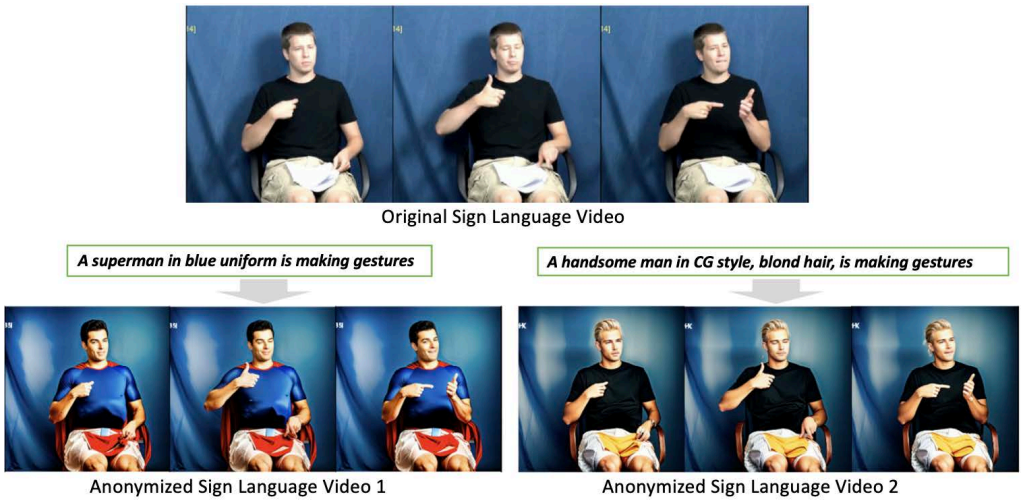

Figure 1: **Text-guided Sign Language Video Anonymization.** We introduce DiffSLVA, an innovative approach that leverages the capabilities of diffusion models to achieve text-guided sign language video anonymization. This method is capable of anonymizing sign language videos with a single text prompt, effectively masking the identity of the original signer while preserving the linguistic content and nuances.

the identity of the signer is altered. Figure 1 illustrates the method. Unlike traditional methods that require skeleton extraction, our approach uses the Stable Diffusion model enhanced with ControlNet (Zhang et al., 2023) to process language videos with Holistically-Nested Edge (HED) (Xie and Tu, 2015), which can more easily and robustly process videos in the wild. To adapt the image-based Stable Diffusion for video, we follow Yang et al. (2023), but modify the methods. We replace the self-attention layer in U-Net with a cross-frame attention layer and implement an optical-flow-guided latent fusion for consistent frame generation. Additionally, to capture fine-grained facial expressions, we have developed a specialized facial generation module using a state-of-the-art image animation model (Zhao and Zhang, 2022) fine-tuned on our mixed dataset (see Section 4.1). The outcomes are integrated via a face segmentation technique (Yu et al., 2018). Our results show substantial promise for anonymization applications, which would be invaluable for the Deaf and Hard-of-Hearing communities.

Our work makes several key contributions to the field of sign language video anonymization:

(1) We propose text-guided sign language anonymization. The anonymized videos are based on computer-generated humans, transforming the original signer's appearance to that of a computer-generated individual.

(2) We have developed a specialized module dedicated to improving facial expression transformation. Our ablation studies show that this significantly enhances the preservation of linguistic meaning.

(3) Our approach relies solely on low-level image features, such as edges, enhancing the potential for practical applications.

(4) Our anonymization can accommodate a diverse range of target humans. The anonymized signers can have any ethnic identity, gender, clothing, or facial style, a feature many ASL signers want; this simply requires changing the text input.

2. Related Work

2.1. Video Editing with Diffusion Models

Diffusion models (Ho et al., 2020; Song et al., 2020) have shown exceptional performance in the field of generative AI. Once trained on large-scale datasets (e.g., LAION (Schuhmann et al., 2022)), text-guided latent diffusion models (Rombach et al., 2022), e.g., Stable Diffusion, are capable of producing diverse and high-quality images from a single text prompt. Additionally, ControlNet (Zhang et al., 2023) presents a novel enhancement. It fine-tunes an additional input pathway for pre-trained latent diffusion models, enabling them to process various modalities, including edges, poses, and depth maps. This innovation significantly augments the spatial control capabilities of text-guided models.

Image-based diffusion models can also be used for video generation or editing. There have been efforts to modify image-based diffusion models for consistent generation or editing across frames. Tune-A-Video (Wu et al., 2023) inflates a pre-trained image diffusion model, modified with pseudo 3D convolution and cross-frame attention and then fine-tuned on a given video sequence. During the inference stage, with the DDIM inversion noises (Song et al., 2020) as the starting point, the fine-tuned model is able to generate videos with similar motions but varied appearance. Edit-A-Video (Shin et al., 2023), Video-P2P (Liu et al., 2023), and vid2vid-zero (Wang et al., 2023) utilize Null-Text Inversion (Mokady et al., 2023) for improved reconstruction of video frames, which

provides better editing results. Fine-tuning or optimization based on one or more input video sequences is required by these methods. Moreover, the detailed motion in the video cannot be captured properly without having a negative impact on the editing abilities. Therefore, they are not suitable for the sign language video anonymization task.

Other methods use the cross-frame attention mechanism or latent fusion to achieve the video editing or generation ability of image-based diffusion models. Text2Video-Zero (Khachatryan et al., 2023) modifies the latent codes and attention layer. FateZero (Qi et al., 2023) blends the attention features based on the editing masks detected by Prompt-to-Prompt (Hertz et al., 2022). Pix2Video (Ceylan et al., 2023) aligns the latent features between frames for better consistency. Rerender-A-Video (Yang et al., 2023) utilizes a cross-frame attention mechanism and cross-frame latent fusion to improve the consistency of style, texture, and details. It can also be used with ControlNet for spatial guidance. However, these methods cannot accurately transfer facial expressions from the original videos. Therefore, they lose a significant amount of the linguistic meaning from the original video. Our approach is based on the Rerender-A-Video (Yang et al., 2023) method, without the post video processing, to best capture manual signs. To overcome the loss of linguistically important non-manual information, we designed a specialized facial expression translation module (Zhao and Zhang, 2022), which we combine with the rest of the anonymized body using a face parser model (Yu et al., 2018).

2.2. Sign Language Video Anonymization

Various strategies have been explored for privacy preservation in ASL video communication (Isard, 2020). Early approaches used graphical filters, such as a tiger-shaped filter (Bragg et al., 2019b), to disguise the face during signing. However, these filters often lead to a loss of critical facial expressions, thereby hindering comprehension. Alternatives like blocking parts of the face (Bleicken et al., 2016) also result in significant information loss. Approaches involving re-enacting signed messages with actors (Isard, 2020) or using virtual humans for anonymous sign language messaging (Heloir and Nunnari, 2016; Efthimiou et al., 2015) are labor-intensive, challenging, and time-consuming.

Some approaches to avatar generation for sign language, e.g., that of Bragg (2019a), use cartoon-like characters to replace signers. Cartoonized Anonymization (Tze et al., 2022b) proposes use of pose estimation models (Li et al., 2018; Xiu et al., 2018; Lugaresi et al., 2019) to automatically enable the avatars to sign. Yet, these methods often lead to unrealistic results (Kipp et al., 2011).

Deep-learning approaches, such as AnonySign (Saunders et al., 2021) or Neural Sign Reenactor (Tze et al., 2022a), leverage GAN-based methods for photo-realistic sign language anonymization using skeleton keypoints for accurate image generation. The results are encouraging. However, they require accurate skeleton keypoints and face landmarks. In sign language videos, rapid hand movements can lead to blurring in the video frames. Occlusions of the face by the hands also occur frequently. For these reasons, the performance of existing human pose estimation models is often inadequate when applied to sign language videos, which leads to errors in the anonymized video.

Recent work (Lee et al., 2021) applies the facial expression transfer method of Siarohin et al. (2019b) for sign language anonymization. This method involves replacing the signer's face in the video with another individual's face, while transferring the facial expressions to the new face. As a result, this approach successfully preserves the linguistic meanings conveyed by facial expressions and alters the identity of the signer in the video. However, in Lee et. al (2021), the extent of the anonymization is not complete, since only the face is replaced, while the arms, torso, and hands remain the same as in the original video. Another method (Xia et al., 2022) uses an unsupervised image animation method (Siarohin et al., 2021; Ren et al., 2020) with a high-resolution decoder and loss designed for the face and hands to transform the identity of a signer to that of another signer from the training videos. The results are promising. However, this method can work well only in the training data domain with limited signer identities and is hard to adapt to sign language videos in the wild.

To address the above limitations, we propose Diff-SLVA, a method that is based on the modification of large-scale diffuson models and ControlNet for consistent high-fidelity video generation, which can be used to achieve effective sign language video anonymization in the wild. Our approach is a text-guided sign language video anonymization, as shown in Figure 1. For the anonmyization of signers' body, arms and hands, we use large-scale diffusion models, which do not rely on the use of sign language video data for training and can perform zero-shot sign language video anonymization. With the help of ControlNet, we use low-level features instead of accurate skeleton data as signal for generation guidance, so that the results are not adversely affected by inaccurate skeleton estimations. To further improve the facial expression translation, we designed a specialized model for facial expression enhancement and combine it with the model that anonymizes the rest of the body using a face parser model. Our method can anonymize sign language videos based on a single text prompt. The

anonymized video is based only on a wide range of computer-generated humans. Our anonymization technique thereby offers great promise for applications that would benefit the Deaf community.

3. Methodology

In this section, we introduce our method for text-guided sign language video anonymization. The process is structured as follows: Given a sign language video with N frames $\{I_i\}_{i=0}^{N}$, we use a pre-trained latent diffusion model, augmented with ControlNet, to execute the anonymization. A text prompt c_p serves as guidance for the desired anonymization identity or style. Our goal is to generate an altered video sequence, represented by $\{I_i'\}_{i=0}^{N}$, that conceals the identity of the original signer while preserving the linguistic content.

In 3.1, we introduce the text-guided latent diffusion models and the ControlNet, which serve as the foundation for text-guided image generation. Section 3.2 details the methods for adapting the text-to-image method for consistent video editing. To ensure preservation of linguistic meaning through accurate facial expression translation, we introduce a specialized facial enhancement module in 3.3. Figure 2 shows an overview of our method.

3.1. Latent Diffusion Models

Latent diffusion models operate in the latent space for faster image generation. The input image I is first input to an encoder ε to obtain its latent features $x_0 = \varepsilon(I)$. The following diffusion forward process adds noise to the latent features:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \quad (1)$$

where $t = 1, ..., T$ is the time step indicating the level of noises added; $q(x_t|x_{t-1})$ is the conditional probability of x_t given x_{t-1}; and α_t are hyper-parameters that adjust the noise level across the time step t. Leveraging the property of Gaussian noise, we can also sample x_t at any time step by the following equation:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \quad (2)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

In the diffusion backward process, a U-Net ϵ_θ is trained to estimate the above added noise to recover x_0 from x_T. For the conditional diffusion model, ϵ_θ takes the conditional information c_p as input to guide the generation process. After ϵ_θ has been trained, the x_{t-1} can be sampled by strategies such as DDIM sampling (Song et al., 2020):

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t, c_p), \quad (3)$$

where $\epsilon_\theta(x_t, t, c_p)$ is the predicted noise at time step t. For the DDIM sampler, we can estimate the final clear output $\hat{x}_0$ at each time step t. $\hat{x}_0$ can also be represented as the following equation:

$$\hat{x}_0 = (x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t, c_p))/\sqrt{\bar{\alpha}_t}, \quad (4)$$

During inference, for a Gaussion noise x_T, we can sample a clear latent x_0 with the DDIM Sampler and decode it to the generated image $I' = D(x_0)$. Our methodology also incorporates ControlNet, introducing an additional signal to the text-guided latent diffusion models. This structure makes it possible for the text-guided diffusion model to take diverse inputs like edges, human poses, and segmentation maps for more spatial constraints. Consequently, with incorporation of an additional input c_n, the predicted noise at each time step t is represented as $\epsilon_\theta(x_t, t, c_p, c_n)$. This approach enhances the alignment of the final outputs with the spatial features specified by the input condition c_n.

3.2. Consistent Video Generation

Although Stable Diffusion models exhibit outstanding performance in image generation, application to videos is challenging. Directly applying Stable Diffusion to videos gives rise to significant frame inconsistency issues. To address this, we adapt text-to-image diffusion models for video editing tasks, drawing upon the framework established by Yang et al. (2023). Our approach begins by encoding and sampling the original frames $I_i, i = 1, \ldots, N$, of the sign language video into noisy latents $x^i{}_t, i = 1, \ldots, N$, serving as starting points for the generation of anonymized video frames, following the method described by Meng et al. (2021). An anchor frame I_a is selected from the sequence $I_i, i = 1, \ldots, N$. The corresponding latent feature x_t^a, along with the Holistically-Nested Edge, is processed through ControlNet to create the transformed anchor frame I_a', which constrains the global consistency in general. Empirically, we find that selecting the anchor frame from the middle of the video, where both hands of the signer are visible, yields optimal results. For each frame I_i, the previously generated frame I_{i-1}' and the anchor frame I_a' provide cross-frame attention control during the generation of I_i', as detailed in Section 3.2.1. A two-stage optical-flow-guided latent fusion, described in Section 3.2.2, is applied during the generation process. Finally, a specialized facial expression enhancement module, outlined in Section 3.3, is used to refine the results.

3.2.1. Cross-Frame Attention Consistency

In the Stable Diffusion model, there are two kinds of attention mechanisms used in the U-Net. The cross-attention retrieves the information from the text embedding. The self-attention helps define the layout and style of the generated images. In order to achieve consistent generation across frames

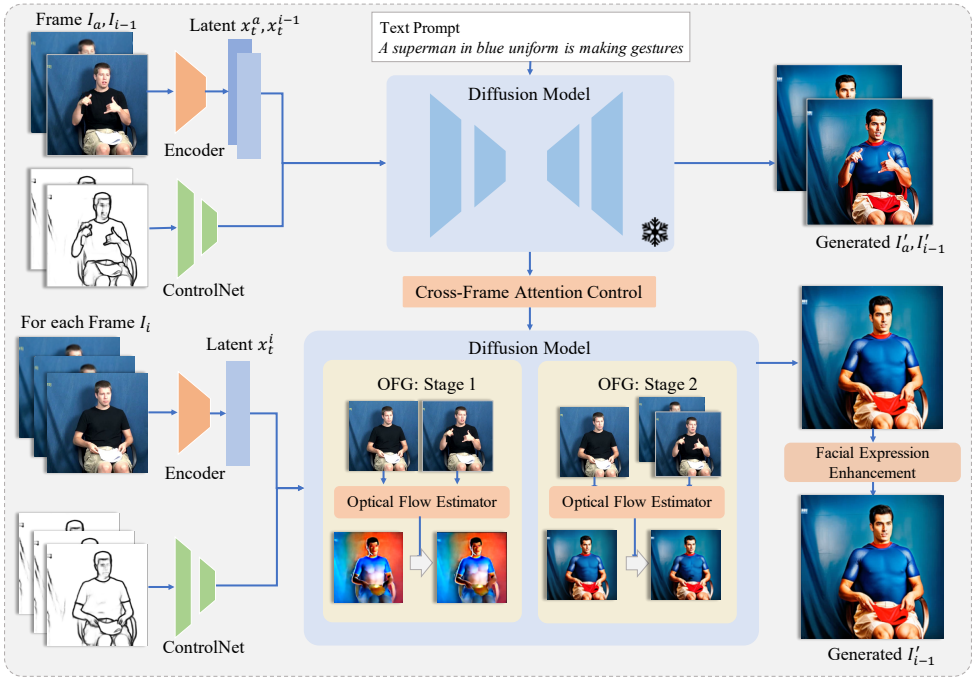

Figure 2: **Method Overview.** The original frames $\{I_i\}, i = 1, ..., N$ in the sign language video are encoded and sampled as noisy latent features $\{x^i\}_t, i = 1, ..., N$. An anchor frame I_a and its Holistically-Nested Edge are used to generate the I'_a with ControlNet, which will constrain the global style consistency. For each frame I_i, the previous generated frame I'_{i-1} and the anchor-generated frame I'_a provide cross-frame attention control during the generation process of I'_i. A two-stage optical-flow-guided latent fusion is applied. A specialized facial expression enhancement module is used to update I'_i for the final result.

in the sign language video sequence, the self-attention layers are replaced with cross-frame attention layers. The self-attention layer of the U-Net used in Stable Diffusion is represented as follows:

$$Q = W^Q v_i, K = W^K v_i, V = W^V v_i, \quad (5)$$

where v_i is the latent features input to the self-attention layer when generating I'_i. W^Q, W^K, and W^V are the weights for project v_i to the query, key, and value in the attention mechanism, respectively. The attention map SA is calculated as following:

$$SA(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}})V \quad (6)$$

where d is the dimension of K. To obtain consistent generation across frames, we replace K and V with $K_{a,i-1}$ and $V_{a,i-1}$, which are the combination of keys and values when generating the selected anchor frame I_a and previous frame I_{i-1}. The cross-frame attention layer is represented as:

$$\begin{aligned} K_{a,i-1} &= W^K[v_a; v_{i-1}], \quad Q = W^Q v_i \\ V_{a,i-1} &= W^V[v_a; v_{i-1}], \end{aligned} \quad (7)$$

where v_a, v_{i-1} are the latent features obtained when generating frame I'_a and I'_{i-1}. The cross-attention map CA is calculated as:

$$CA(Q, K_{a,i-1}, V_{a,i-1}) = \text{Softmax}(\frac{QK^T_{a,i-1}}{\sqrt{d}})V_{a,i-1} \quad (8)$$

The cross-frame attention mechanism is designed to foster consistency in image generation across frames by directing the current generation process to reference patches in both the generated anchor frame and the previous frame.

3.2.2. Optical-Flow-Guided Cross-Frame Latent Fusion

Following Yang et al. (2023), we use 2-stage latent fusion guided by optical flow: OFG stages 1 and 2.

• OFG stage 1: In the early stage of the diffusion backward process, the optical flow w_a^i and occlusion mask M_a^i are estimated from I_a to I_i to wrap and fuse the estimated latent of I'_a and I'_i. This latent wrap and fusion is performed when the denoising step t is large, to prevent distortion of results. At time step t, the predicted $\hat{x}_0$ is updated by:

$$\hat{x}_0^i = M_a^i \hat{x}_0^i + (1 - M_a^i) w_a^i(\hat{x}_0^a), \quad (9)$$

where $\hat{x}_0^i$ and $\hat{x}_0^a$ are the predicted clear outputs for I'_i and I'_a at denoising time step t, from equation 4.

• OFG stage 2: At the second stage, the generated anchor frame I'_a and previous generated frame I'_{i-1} are used to further enhance consistency during the late stages of the diffusion backward process. The

optical flow and occlusion mask are also estimated. We obtain a reference image $\bar{I'}_i$ by wrapping and fusing with the previous generated images:

$$\begin{aligned}\bar{I'}_i =& M_a^i(M_{i-1}^i \hat{I'}_i + (1 - M_{i-1}^i) w_{i-1}^i (I'_{i-1})) \\ &+ (1 - M_a^i) w_a^i I'_a,\end{aligned} \tag{10}$$

After obtaining this reference-estimated image $\bar{I'}_i$, we can update the sampling process for generating I'_i using the following equation:

$$x_{t-1}^i = M_i x_{t-1}^i + (1 - M_i)\bar{x}_{t-1}^i, \tag{11}$$

where $M_i = M_a^i \cap M_{i-1}^i$, and $\bar{x}_{t-1}^i$ is the sampled x_{t-1} from reference image $\bar{I}'_i$. We use the same strategy as the fidelity-oriented image encoding in Yang (2023) to encode $\bar{I}'_i$ to avoid information loss when repeatedly encoding and decoding latents.

To maintain coherent color throughout the whole process, we also apply AdaIN (Huang and Belongie, 2017) to $\hat{x}_0^i$ with $\hat{x}_0^a$ at time step t during the late stage of the diffusion backward process. This mitigates the color drift problem with diffusion models.

3.3. Facial Expression Enhancement

Facial expressions convey important linguistic information in signed languages. However, current methods cannot transfer meaningful facial expressions; see the ablation study discussed in Section 4.6. ControlNet and Stable Diffusion usually fail to produce faces with the same expressions as the original signer. To address this issue, we propose an additional module to enhance the face generation based on an image-animation model. See Figure 3 for an overview of this module.

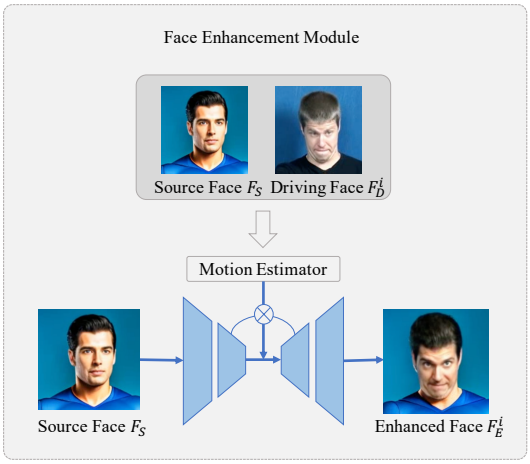

Figure 3: **Face Enhancement Module.** The motion estimator obtains dense motion and multi-resolution occlusion maps between the source face F_s and the driving face. The output along with a U-Net is applied to generate the enhanced face F_E^i

When generating the first frame I'_1 , we crop the result face and use it as the source face F_s for the image animation module from Zhao and Zhang (2022). The facial images in the original videos are also cropped and aligned to formalize the driving face set $[F_d^i], i = 1...N$. A motion estimation module will estimate the dense motion W_i and multi-resolution occlusion maps M_i between the source face F_s and the driving face set $[F_d^i], i = 1...N$.

The obtained optical flow and occlusion maps are input to a U-Net to generate new face images that match the identity of the source face F_s but have the same facial expression as F_d^i. The input image F_s is processed through the encoder, and optical flow W_i is applied to wrap the feature map at each level. This adjusted feature map is then combined with the occlusion mask M_i^f that matches its resolution. Subsequently, it is merged into the decoder through a skip connection. The feature map is then input to the next upsampling layer. Finally, the enhanced face image F_E^i is produced at the last layer.

A face parser model (Yu et al., 2018) is applied on F_E^i to segment the face area and obtain a mask M_i^f. Then, the mask and enhanced face image are aligned with the face location in I'_i. Finally, I'_i is updated by the following equation:

$$I'_i = M_i^f F_E^i + (1 - M_i^f) I'_i. \tag{12}$$

4. Experiments and Results

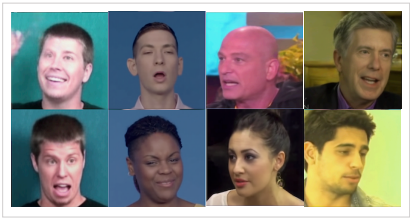

Figure 4: **Example Images** from the mixed dataset. We sampled more images from ASL videos for a balanced dataset.

4.1. Dataset

We implemented our method on video datasets distributed through the American Sign Language Linguistic Research Project (ASLLRP): https://dai.cs.rutgers.edu/dai/s/dai (Neidle et al., 2018, 2022b). Each test sample was limited to a maximum of 180 video frames. Example results are presented in Figure 5. We also produce a mixed dataset for fine-tuning the facial expression module, as illustrated in Section 4.3.

4.2. Models

Our experiments utilized Stable Diffusion models version 1.5 and other customized models. The ControlNet version 1.0 was employed, producing optimal results with HED as a conditional input. Optical flow estimation was performed using the model from Xu et al. (2022).

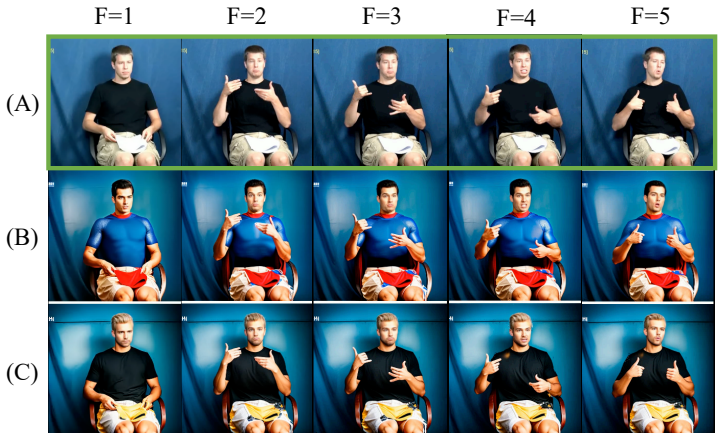

Figure 5: **Anonymization Result Examples.** Row (A) contains some frames from the original ASL video (taken from ASLLRP file Cory_2013-6-27_sc115, Utterance 22, meaning 'If friends play Frisbee, I will join them.'). Rows (B) and (C) show anonymization from different prompts: (B) *a Superman in blue uniform is making gestures* (C) *a man in CG style, blond hair, is making gestures.*

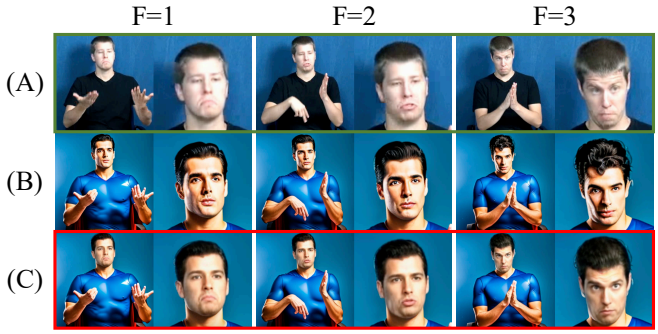

Figure 6: **Ablation Study of Facial Expression Enhancement**. The frames in Row (A) are taken from ASLLRP file Cory_2013-6-27_sc114, Utterance 102. Row (B) is the result without the facial enhancement module. Row (C) is the final result of our method.

4.3. Fine-tuning Facial Expression Model

State-of-the-art facial reenactment models are usually trained on large-scale speaking head datasets such as Voxceleb (Nagrani et al., 2017). The rich identity information contained in such datasets makes it possible to generalize on face images in the wild. However, the speaking head videos lack linguistically important facial expressions. In contrast, the face images cropped from ASL videos contain linguistic information, but lack diversity of identities, which impacts the model's ability to generalize. To address this, we propose to mix these two datasets and apply a balance sampling strategy in training in order to maintain the model's generalization ability and enable generation of facial expressions carrying linguistic meanings. Figure 4 shows example face images for this mixed dataset. We fine-tune the pre-trained model from Zhao and Zhang (2022) on this mixed dataset for 40 epochs.

4.4. Qualitative Evaluation

To our knowledge, this is the first instance of text-guided sign language anonymization capable of generating an unlimited array of diverse anonymized videos. Methods like Cartoonized Anonymization (CA) (Tze et al., 2022b) cannot generate photorealistic results and rely on skeleton estimation for accurate anonymization. Methods that can generate photorealistic results, e.g., AnonySign (Saunders et al., 2021), SLA (Xia et al., 2022), and Neural Sign Reenactor (NSR) (Tze et al., 2022a), require accurate skeleton estimation or have very limited choices of anonymization identites.

Our initial results are encouraging. Our method can generate clear handshapes with high fidelity to the original signer's handshapes and hand/arm movements. Most generated facial expressions are good; further refinements to fully preserve subtle linguistic expressions are underway. Effectiveness for complete disguise of identity, transmission of linguistic content, and production of natural-looking signing remains to be confirmed through user studies, to be carried out soon. In the very near future, we will also validate our results by processing our anonymized videos through our independent system for sign recognition from video (Zhou et al., 2024, under review), to confirm that the anonymized versions are correctly recognized

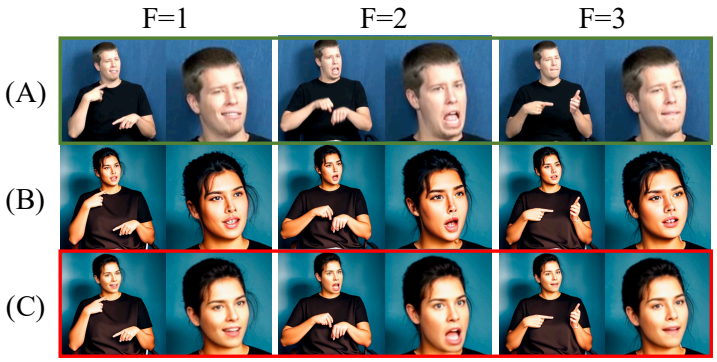

Figure 7: **Ablation Study of Facial Expression Enhancement**. The frames in (A) are taken from ASLLRP file Cory_2013-6-27_sc107, Utterance 14. Row (B) shows the result without the facial enhancement module. (C) shows the final result of our method.

as the originally produced sign. Figure 5 shows that our method can produce computer-generated signers with varying identities: Text prompts allow for varying anonymized versions of ASL videos. The results underscore the practical potential of our approach. Video examples can be seen at https://github.com/Jeffery9707/DiffSLVA2.

4.5. Quantitative Evaluation

We use an identity classifier (Schroff et al., 2015; Cao et al., 2018) to check whether our method successfully changes the identity of the original signer. In particular, we calculate the cosine similarity between face embeddings of multiple images of the same signer and of anonymized signers. See Table 1. Cosine similarity close to 1 or 0 means the faces are from the same person or an unrelated person, respectively.

	Original	Anonymized
Signer A	0.7740	0.1273
Signer B	0.8917	0.0566
Signer C	0.8566	-0.0165

Table 1: Anonymization Analysis for the Face. Each column contains the cosine similarity between faces of the same signer and anonymized signers.

From the table, we can see that our anonymized face has a cosine similarity close to 0 with the original face. Therefore, our method has successfully anonymized the signers to a unrelated identity.

4.6. Ablation Study

Our ablation study focused on the facial expression enhancement module. Results are shown in Figures 6 & 7. Using this module significantly improves preservation of linguistic meaning. (The examples shown include topic and wh-question marking.)

The Stable Diffusion model does not do well with accurate generation of varied facial expressions for ASL anonymization. Instead of producing diverse expressions, the model tends to replicate a uniform expression across frames, resulting in loss of linguistic information. This limitation highlights the importance of applying facial expression enhancement module for ASL video anonymization.

5. Conclusion and Discussion

We introduce DiffSLVA, a novel approach using large-scale pre-trained diffusion models for text-guided ASL video anonymization. Our approach could be applied to various use cases. It could enable signers to share sensitive information while preserving privacy. It could enable anonymous peer review for ASL-based academic submissions, thereby ensuring unbiased academic review. It could bring neutrality to multimodal ASL tools, e.g., for anonymized definitions for ASL dictionaries. Furthermore, our approach could enhance neutrality in interpreting scenarios in digital communications, such as messaging, enabling maintenance of confidentiality in ASL communications. The implementation of DiffSLVA could also increase participation in video-based AI databases, enriching AI research with diverse ASL data.

This approach does not address the possibility that even anonymized signers could be recognized by those who know them very well, based on signing style. Furthermore, our current method has some limitations. It may encounter challenges in cases where the face is occluded by one or both hands or where there is blurring due to rapid movements in ASL videos. In addition, as is a known issue for Stable Diffusion Models, artifacts of various types sometimes appear in our anonymized videos. We aim to address these issues in our future work. We are also working on further refinements to improve the facial transformation module. However, overall, DiffSLVA shows substantial promise for anonymization applications, which could offer invaluable tools for the Deaf and Hard-of-Hearing communities.

6. Acknowledgments

We are grateful to the many, many people who have helped with the collection, linguistic annotation, and sharing of the ASL data upon which we have relied for this research. In particular, we are endebted to the many ASL signers who have contributed to our database; to Gregory Dimitriadis at the Rutgers Laboratory for Computer Science Research, the principal developer of SignStream®, our software for linguistic annotation of video data (https://www.bu.edu/asllrp/SignStream/3/); to Matt Huenerfauth and his team for data collection at RIT; to DawnSignPress for sharing video data; to the many who have helped with linguistic annotations (especially Carey Ballard and Indya Oliver); and to Augustine Opoku, for development and maintenance of our Web-based database system for providing access to the linguistically annotated video data (https://dai.cs.rutgers.edu/dai/s/dai). This work was supported in part by NSF grants #2235405, #2212302, #2212301, and #2212303, although any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. Bibliographical References

Robert W Arnold. 2009. *A proposal for a written system of American Sign Language*. Gallaudet University.

Charlotte Baker-Shenk. 1985. The facial behavior of deaf signers: Evidence of a complex language. *American Annals of the Deaf*, 130(4):297–304.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*.

Julian Bleicken, Thomas Hanke, Uta Salden, and Sven Wagner. 2016. Using a language technology infrastructure for German in order to anonymize German Sign Language corpus data. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3303–3306, Portorož, Slovenia. European Language Resources Association (ELRA).

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019a. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.

Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2019b. Exploring collection of sign language datasets: Privacy, participation, and model performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14.

Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE.

Chen Cao, Qiming Hou, and Kun Zhou. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):1–10.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942.

Geoffrey Restall Coulter. 1979. *American Sign Language typology*. Ph.D. thesis, University of California, San Diego.

DawnSign Press. 2022. DawnSign-Press (2022) About Us. *DawnSignPress Website*.

Eleni Efthimiou, Stavroula-Evita Fotinea, Theodore Goulas, and Panos Kakoulidis. 2015. User friendly interfaces for sign retrieval and sign synthesis. In *International Conference on Universal Access in Human-Computer Interaction*, pages 351–361. Springer.

Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. 2012. The dictasign Wiki: Enabling web communication for the deaf. In *International Conference on Computers for Handicapped Persons*, pages 205–212. Springer.

Eleni Efthimiou, Stavroula-Evita Fotinea, Christian Vogler, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and Jérémie Segouat. 2009. Sign language recognition, generation, and modelling: a research effort with applications in deaf communication. In *International Conference on Universal Access in Human-Computer Interaction*, pages 21–30. Springer.

Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional multi-person pose estimation. In *ICCV*.

Jakub Fišer, Ondřej Jamriška, David Simons, Eli Shechtman, Jingwan Lu, Paul Asente, Michal Lukáč, and Daniel Sỳkora. 2017. Example-based synthesis of stylized facial animations. *ACM Transactions on Graphics (TOG)*, 36(4):1–11.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. 2020. MarioNETte: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10893–10900.

Alexis Heloir and Fabrizio Nunnari. 2016. Toward an intuitive sign language animation authoring system for the deaf. *Universal Access in the Information Society*, 15(4):513–523.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.

Amy Isard. 2020. Approaches to the anonymisation of sign language corpora. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 95–100, Marseille, France. European Language Resources Association (ELRA).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.

Hernisa Kacorri and Matt Huenerfauth. 2016. Continuous profile models in ASL syntactic facial expression synthesis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2084–2093.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*.

Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 107–114.

Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2021. American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA. Association for Computing Machinery.

Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. 2018. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2023. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

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*.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Carol Neidle, Judy Kegl, Benjamin Bahan, Dawn MacLaughlin, and Robert G Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. MIT press.

Carol Neidle, Augustine Opoku, Carey M. Ballard, Konstantinos M. Dafnis, Evgenia Chroni, and Dimitris Metaxas. 2022a. Resources for computer-based sign recognition from video, and the criticality of consistency of gloss labeling across multiple large ASL video corpora. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 165–172, Marseille, France. European Language Resources Association (ELRA).

Carol Neidle, Augustine Opoku, Gregory Dimitriadis, and Dimitris Metaxas. 2018. New shared & interconnected ASL resources: SignStream® 3 software; DAI 2 for web access to linguistically annotated video corpora; and a sign bank. In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 147–154, Miyazaki, Japan. European Language Resources Association (ELRA).

Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022b. ASL Video Corpora & Sign Bank: Resources available through the American Sign Language Linguistic Research Poject (ASLLRP). *arXiv preprint arXiv:2201.07899*.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. On aliased resizing and surprising subtleties in gan evaluation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11400–11410.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*.

Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, and Ran He. 2019. Make a face: Towards arbitrary high fidelity face manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10042.

Jian Ren, Menglei Chai, Sergey Tulyakov, Chen Fang, Xiaohui Shen, and Jianchao Yang. 2020. Human motion transfer from poses in the wild. In *European Conference on Computer Vision*, pages 262–279.

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*, pages 10684–10695.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. AnonySIGN: Novel human appearance synthesis for sign language video anonymisation. *arXiv preprint arXiv:2107.10685*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. 35:25278–25294.

Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. 2023. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019a. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019b. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32:7137–7147.

Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662.

Jenny L Singleton, Gabrielle Jones, and Shilpa Hanumantha. 2014. Toward ethical research practice with deaf participants. *Journal of Empirical Research on Human Research Ethics*, 9(3):59–66.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Christina O Tze, Panagiotis P Filntisis, Athanasia-Lida Dimou, Anastasios Roussos, and Petros Maragos. 2022a. Neural sign reenactor: Deep photorealistic sign language retargeting. *arXiv preprint arXiv:2209.01470*.

Christina O Tze, Panagiotis P Filntisis, Anastasios Roussos, and Petros Maragos. 2022b. Cartoonized anonymization of sign language videos. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE.

Clayton Valli and Ceil Lucas. 2000. *Linguistics of American Sign Language: An introduction*. Gallaudet University Press.

Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. 2003. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018a. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018b. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.

Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. 2023. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*.

Olivia Wiles, A Koepke, and Andrew Zisserman. 2018. X2Face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633.

Zhaoyang Xia, Yuxiao Chen, Qilong Zhangli, Matt Huenerfauth, Carol Neidle, and Dimitris Metaxas. 2022. Sign language video anonymization. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 202–211, Marseille, France. European Language Resources Association (ELRA).

Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403.

Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. 2018. Pose Flow: Efficient online pose tracking. In *BMVC*.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. 2022. GMFlow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130.

Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM SIGGRAPH Asia Conference Proceedings*.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Jian Zhao and Hui Zhang. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666.

Yang Zhou, Zhaoyang Xia, Yuxiao Chen, Carol Neidle, and Dimitris N. Metaxas. 2024. A Multimodal Spatio-temporal GCN Model with Enhancements for Isolated Sign Recognition. In *Proceedings of the LREC2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*.